Deuxième Forum Halieumétrique, Nantes, 1995
*Session II : **Analyse de l'information***

ESTIMATION DE PROBABILITÉS PAR LE MODÈLE LINÉAIRE GÉNÉRALISÉ. APPLICATION AU SAUMON (*Salmo salar* L.)

Jacques Badia, Robert Faivre, Marie-Hélène Charron [a] ; Patrick Prouzet, Jacques Dumas [b] ; François-Xavier Cuende [c]

I - INTRODUCTION

Le modèle linéaire généralisé permet d'analyser la relation de causalité entre une variable à expliquer de la famille exponentielle et des variables explicatives par des techniques de régression. Pour une étude donnée, différentes modélisations sont possibles. Pour les expliciter, nous présentons deux applications : A) l'étude des effets des conditions hydroclimatiques sur les captures par unité d'effort de saumons dans l'Adour, B) l'estimation des probabilités de retour des saumons atlantiques dans la Nivelle. Dans chaque cas, l'objectif est d'estimer des probabilités à partir de données groupées.

II - MATÉRIEL ET MÉTHODES

A - Captures par unité d'effort et facteurs hydroclimatiques

L'étude des carnets de pêche des marins-pêcheurs a permis d'analyser l'influence des conditions hydroclimatiques sur le niveau moyen des captures journalières des saumons prélevés dans l'Adour en 1988 et de 1990 à 1992 (Cuende, 1994). Pour estimer les probabilités de capture on a utilisé le modèle multinomial (McCullagh & Nelder, 1989).
Les captures par unité d'effort (*cpue*), la saison (s), le débit fluvial (df) et le coefficient de marée (cm) sont des descripteurs de l'importance des captures et du contexte dans lequel les données ont été collectées. Ces facteurs ont des statuts différents : s, df et cm sont explicatifs, *cpue* est à expliquer ; ils ont plusieurs modalités : {printemps, été} pour s, {faible, moyen, fort} pour df et cm, {nul (n), faible (f), moyen (m), fort (F)} pour *cpue*. Cette dichotomie précise les conditions dans lesquelles les observations (nombre de pêches) sont faites, elle oriente l'utilisateur dans le

[a]INRA. Biométrie et Intelligence Artificielle. BP 27,31326 Castanet Tolosan.
[b]INRA/IFREMER. Hydrobiologie, St Pée sur Nivelle. BP 3,64310 Ascain.
[c]IMA. Plateau de l'Atalaye, 64200 Biarritz.
Pôle de recherche sur la gestion des ressources aquatiques en environnement sensible

choix d'un modèle (hypothèse distributionnelle, fonction de lien). Conditionnellement à s, df et cm, on a considéré que le nombre de pêches dans une classe de *cpue* est une réalisation d'une variable aléatoire multinomiale. Le modèle multinomial étant choisi, il faut modéliser le lien $g_k(p_i)$ entre le vecteur des probabilités $p_i = (p_{in}, p_{if}, p_{im}, p_{iF})'$ et les caractéristiques de la population i (une combinaison $s \times df \times cm$) soit : $g_k(p_i) = x_i\beta_k$ où x_i est le vecteur ligne des caractéristiques de la population i et β_k le vecteur des paramètres associés aux facteurs explicatifs. Pour montrer comment la prise en compte d'une information influence les résultats d'une analyse, on a choisi d'utiliser les liens logit cumulé et logit généralisé.

B - Probabilités de retour des saumons

Pour étudier le cycle biologique des saumons de la Nivelle, un modèle a été programmé par Charron (1994). Il permet de simuler les évolutions prévisibles de cette population en fonction du niveau de ponte, des taux de séparation en classes de saumons de rivière et de saumons de mer. Pour le mettre en œuvre les probabilités de retour des saumons partis en mer doivent être connues. Pour estimer ces probabilités, on modélise les observations qui se réfèrent aux six années de naissance (1985 à 1990) pour lesquelles le cycle complet des retours a été observé. Du point de vue statistique, chaque année de naissance est une modalité du facteur classe de naissance (cn). Les saumons pouvant rester un ou deux ans en rivière avant de partir en mer et un ou deux ans en mer avant de revenir dans la Nivelle, on note ri et me respectivement les facteurs "temps passé en rivière" et "temps passé en mer". Ces trois facteurs caractérisent l'histoire des saumons et pour chaque histoire on observe le nombre de retours (y) des saumons de mer (n). Chaque y est une réalisation d'une variable aléatoire binomiale Y ($Y \sim B(n,p)$), on peut donc théoriquement estimer les probabilités de retour p en utilisant un modèle logistique linéaire. Cependant, l'ajustement des observations au modèle peut ne pas être satisfaisant : 1) pour des observations indépendantes, si des facteurs importants ne sont pas contrôlés (surdispersion de la déviance résiduelle), il faut modéliser la variabilité des probabilités de réponse. 2) si les observations sont dépendantes, il faut alors modéliser la corrélation entre les réponses binaires (cf. Collet, 1991).

1) Modélisation de la variabilité des probabilités de réponse

On suppose que la probabilité de réponse de la i*ème* observation notée θ_i est une variable aléatoire d'espérance $\mathrm{E}(\theta_i) = p_i$ et de variance $\mathrm{Var}(\theta_i) =$

$\phi p_i(1-p_i)$. $\mathrm{Var}(\theta_i)$ contient un paramètre d'échelle ($\phi \geq 0$), sa forme correspond au choix de la fonction la plus simple qui prend la valeur zéro quand la probabilité p_i est égale à zéro ou à un. Conditionnellement à θ_i, $\mathrm{E}(Y_i|\theta_i) = n_i\theta_i$ et $\mathrm{Var}(Y_i|\theta_i) = n_i\theta_i(1-\theta_i)$. D'après les résultats des probabilités conditionnelles, $\mathrm{E}(Y_i) = n_ip_i$ et $\mathrm{Var}(Y_i) = n_ip_i(1-p_i)[1+(n_i-1)\phi]$. Quand les probabilités de réponses ne varient pas ($\theta_i = p_i$, $\mathrm{Var}(\theta_i) = 0$, $\phi = 0$), on retrouve bien $\mathrm{Var}(Y_i) = n_ip_i(1-p_i)$.

2) Modélisation de la corrélation entre réponses binaires

Supposons que pour le i*ème* ensemble d'observations, on ait y_i succès parmi n_i d'observations et que $R_{i1}, \ldots, R_{in_i}$ sont des variables aléatoires de Bernouilli telles que $R_{ij} = 1$ pour un succès (retour) et $R_{ij} = 0$ pour un échec ($j = 1, \ldots, n_i$). Si p_i est la probabilité de succès, $\mathrm{E}(R_{ij}) = p_i$ et $\mathrm{Var}(R_{ij}) = p_i(1-p_i)$. Le nombre de succès ($y_i$) est la valeur observée de $\sum_{j=1}^{n_i} R_{ij}$. Ainsi,

$$\mathrm{E}(Y_i) = n_ip_i \text{ et } \mathrm{Var}(Y_i) = \sum_{j=1}^{n_i} \mathrm{Var}(R_{ij}) + \sum_{j=1}^{n_i}\sum_{k\neq j} \mathrm{Cov}(R_{ij}, R_{ik}).$$

Si les R_{ij}, $j = 1, \ldots, n_i$ ne sont pas mutuellement indépendantes, $\mathrm{Var}(R_{ij}) = \mathrm{Var}(R_{ik}) = p_i(1-p_i)$ et $\mathrm{Cov}(R_{ij}, R_{ik}) = \rho p_i(1-p_i)$ avec ρ le coefficient de corrélation. On en déduit que $\mathrm{Var}(Y_i) = n_ip_i(1-p_i)[1+(n_i-1)\rho]$.

Dans ces deux situations, les estimateurs de $\mathrm{Var}(Y_i)$ sont identiques dans la forme mais ont des significations différentes. Pour le premier modèle, ϕ ne peut pas être négatif, c'est seulement une mesure de surdispersion. Pour le second modèle, la corrélation ρ peut être négative, on peut donc théoriquement modéliser une sousdispersion. Cependant l'inégalité $1+(n_i-1)\rho > 0$ implique que $-(n_i-1)^{-1} \leq \rho \leq 1$ d'où une borne pour ρ inférieure en général ou égale à zéro sauf si n_i est petit. Williams (1982) a montré que des estimations de ces paramètres sont obtenues par un calcul itératif en égalant la valeur de la statistique χ^2 à l'approximation de son espérance.

III - RÉSULTATS

A - Captures par unité d'effort et facteurs hydroclimatiques

La méthode des moindres carrés généralisés utilisée pour estimer les paramètres, permet d'observer que les tests d'adéquation des modèles étudiés (logits cumulé et généralisé) sont tous deux acceptables, les niveaux de signification des résidus étant respectivement $\Pr(\chi^2_{(36)} > 29.35) = 0.776$ et $\Pr(\chi^2_{(36)} > 42.58) = 0.209$. L'adéquation du modèle qui prend en compte

l'ordre sur le facteur réponse est meilleure. Pour les tests de signification des facteurs, les résultats des analyses sont comparables mais différents dans l'absolu. Au risque de 1*ère* espèce de 5 %, *cm* n'est pas significatif pour le logit cumulé ($\Pr(\chi^2_{(6)} > 12.19) = 0.058$) alors qu'il l'est avec le logit généralisé ($\Pr(\chi^2_{(6)} > 14.60) = 0.024$). Ces deux modèles sont admissibles mais on ne sait pas vraiment lequel choisir, le risque étant de faire un choix sur des critères subjectifs. Pour éviter cela, il faut s'appuyer sur des considérations biologiques et sur le contenu statistique des modèles. Plus précisément, le modèle logit cumulé tient compte de l'ordre des classes du facteur réponse et recherche les facteurs qui influencent le passage d'une classe à la classe voisine. Le modèle logit généralisé ne tient pas compte de cela, il est fondé sur la différence des classes à une classe de référence, il semble moins adapté aux données à analyser. Ainsi, avec le modèle logit cumulé, on peut constater qu'aucun des facteurs *s*, *df* et *cm* n'est significatif, les probabilités des classes de *cpue* sont identiques.

B - Probabilités de retour des saumons

L'analyse du modèle standard le plus complet : $logit(p) = cn + ri + me + cn*ri + cn*me + ri*me$ met en évidence une déviance résiduelle très grande (36.86) par rapport au nombre de degrés de liberté (5). Vraisemblablement, des facteurs importants ne sont pas contrôlés, il faut modéliser la variabilité des probabilités de réponse. Les analyses montrent que le modèle logistique linéaire surdispersé $logit(p) = cn + ri + me$ est acceptable, la déviance résiduelle (14.234) étant proche du nombre de degrés de liberté résiduels (16). C'est ce modèle qui est utilisé pour estimer les probabilités de retour des saumons.

IV - DISCUSSION

Une attitude critique doit être prise vis-à-vis des résultats fournis par la démarche statistique dans l'estimation de ces probabilités. Les données collectées ne sont pas toujours suffisantes, quantitativement et qualitativement, pour résoudre ces problèmes d'estimation par la voie statistique malgré un travail et des coûts importants de suivi d'une population sur de nombreuses années. Utiliser un modèle statistique qui prend en compte des niveaux de dispersion importants dûs à des effets non contrôlés, permet une meilleure évaluation de la précision des estimateurs. En revanche, il n'élimine pas tous les problèmes de définition et d'interprétation de certains paramètres. Dans le cas des probabilités de retour, celles-ci sont estimées

pour servir de point d'entrée du modèle stochastique du cycle biologique des saumons. Pour les saumons de un et deux ans de mer, les probabilités de retour sont établies sur la seule base des taux de départ, elles ne peuvent pas être interprétées comme des probabilités de survie en mer alors que ce sont ces dernières qu'il faudrait estimer. La probabilité de retour après deux années de mer (paramètre estimé par le modèle statistique) résume les paramètres nécessaires au modèle stochastique qui sont les taux de survie la première, puis la seconde année de mer ainsi que les taux d'exploitation par pêches. Cependant, cette approche n'est pas inutile, elle permet d'alimenter la réflexion en fixant des ordres de grandeur pour les taux de retour, d'avancer des hypothèses, de compléter ou de réorienter le choix des variables à observer.

BIBLIOGRAPHIE

CHARRON (M-H.), 1994 - Modélisation stochastique du cycle biologique des salmonidés migrateurs. Application à la modélisation du cycle du saumon atlantique de la Nivelle et de l'Adour. Diplôme d'études supérieures spécialisées, Université Paul Sabatier.

COLLET (D.), 1991 - *Modelling Binary Data.* Chapman & Hall, London.

CUENDE (F-X.), 1994 - *Contribution à l'étude de la pêche professionnelle du saumon (Salmo salar L.) dans le bassin de l'Adour (France).* Thèse de Doctorat de l'INPT, spécialité Sciences Agronomiques.

McCULLAGH (P.) & NELDER (J.A.), 1989 - *Generalized Linear Models. Monographs on Statistics and Applied Probability.* Chapman and Hall, London. 511 p.

WILLIAMS (D.A.), 1982 - Extra-binomial variation in logistic linear models. *Applied Statistics.* **31**, 144-148.