Modelos de contagem considerando superdispersão e excesso de zeros para a estimativa de abundância de espécies pouco frequentes em pescarias comerciais

Fernando de Pol Mayer - PPG Ecologia, UFSC[1]

Resumo: *A captura não intencional de espécies que não são alvo de uma pescaria é chamada de "captura incidental". A estimativa da abundância destas espécies é fundamental para que medidas de manejo adequadas sejam tomadas. Os dados provenientes deste tipo de captura geralmente possuem uma grande quantidade de zeros e superdispersão. Neste trabalho foi realizada uma comparação entre sete modelos de contagem diferentes (Poisson, quase-Poisson, binomial negativo (BN), mistura com Poisson e BN, e condicional com Poisson e BN), utilizando a captura de uma espécie pouco frequente como variável resposta dependente das covariáveis ano, trimestre e área. A comparação entre modelos foi realizada através dos critérios de informação de Akaike (AIC) e de Bayes (BIC). O modelo com o melhor ajuste (i.e. menores valores de AIC/BIC) foi o de mistura considerando a distribuição BN. Devido à grande quantidade de zeros e à elevada superdispersão encontrada, mesmo os modelos de mistura e condicional de Poisson tiveram um desempenho inferior aos modelos BN (simples, mistura e condicional). Além disso, os modelos de mistura parecem ser mais indicados para modelar a abundância de espécies capturadas incidentalmente devido aos processos envolvidos na obtenção de zeros.*

Palavras-chave: *Superdispersão, Excesso de zeros, Captura incidental, MLG.*

1 Introdução

Em pescarias comercias, frequentemente ocorrem capturas de espécies que não são o alvo principal, denominadas de "captura incidental". Esse é caso da pescaria com espinhel de superfície no Oceano Atlântico, onde os alvos são os atuns e o espadarte, mas outras espécies de peixes são também atraídas pelas iscas nos anzóis e capturadas de forma não intencional.

A estimativa da abundância e a identificação de fatores que controlam a distribuição são fundamentais para que medidas de manejo adequadas possam ser tomadas visando a conservação destes recursos. No entanto, como a frequência de captura incidental é baixa, existe sempre uma grande quantidade de zeros em dados coletados a partir das informações de captura das embarcações pesqueiras comerciais. Quando isso ocorre, normalmente também surgem violações às suposições básicas das distribuições de probabilidade. Um destes casos é a superdispersão, que ocorre quando os dados apresentam uma variabilidade maior do que aquela que pode ser

[1] *fernandomayer@gmail.com*

predita pela relação variância-média do modelo. Não considerar esta superdispersão pode levar à subestimativas de erros-padrão e à uma inferência deficiente dos parâmetros da regressão [2].

A regressão de Poisson, por exemplo, é uma candidata natural à análise de dados em forma de contagem, como a quantidade de peixes capturados com um determinado número de anzóis. No entanto, quando existe excesso de zeros e superdispersão ($Var(Y) > E(Y)$), modelos mais apropriados devem ser utilizados. Algumas alternativas incluem os modelos de quase-verosimilhança (e.g. quase-Poisson) e o binomial negativo, que permitem de alguma forma incorporar a superdispersão. Ainda assim, o excesso de zeros pode causar interferência nas estimativas dos parâmetros [3]. Modelos capazes de incluir tanto a superdispersão quanto o excesso de zeros incluem os modelos inflacionados de zeros [3], e os modelos condicionais ou de barreira [7].

O objetivo deste trabalho é verificar qual o melhor modelo para estimativas de abundância de espécies capturadas incidentalmente em pescarias comerciais. Foi realizada uma comparação entre os diferentes modelos para dados de contagem descritos acima, utilizando como exemplo os dados de captura do agulhão branco (*Tetrapturus albidus*) capturado no Atlântico Sul.

2 Metodologia

Os dados de captura em número do agulhão branco são provenientes do banco de dados *Task 2* da *International Commission for the Conservation of Atlantic Tunas* (ICCAT), entidade internacional responsável pelo gerenciamento das pescarias realizadas pelas diversas frotas (países) que capturam atuns e espécies afins no Atlântico. Esta base de dados está disponível no endereço `www.iccat.int`. Foram selecionadas as informações da frota de espinhel de superfície do Japão, devido à grande série temporal disponível (1960–2007). Ao todo, foram analisadas 15247 linhas da base de dados (média de 317 por ano). Cada linha contém informações como a posição geográfica, o trimestre, o esforço de pesca (número de anzóis) e a captura. É importante ressaltar que uma linha normalmente contêm a informação agregada de diversos lances do espinhel (que possui em média cerca de 1000 anzóis).

Os modelos de regressão para dados de contagem podem ser classificados como um tipo particular da classe de Modelos Lineares Generalizados (MLGs) [4]. De maneira geral, os MLGs descrevem a relação entre a variável resposta y_i ($i = 1, \ldots, n$) através de preditores x_i. A distribuição condicional de $y_i|x_i$ deve pertencer à família exponencial. Se esta condição for satisfeita, então a média de y_i condicionada aos preditores x_i é $E(y_i|x_i) = \mu_i$, e existe uma transformação de μ_i de forma que $g(\mu_i) = x_i^T\beta$, onde $g(\cdot)$ é uma função de ligação conhecida, e β é o vetor de parâmetros a ser estimado. A variância de y_i é então dada por $Var(y_i) = \phi V(\mu_i)$, onde ϕ é o parâmetro de dispersão (geralmente constante) e $V(\mu_i)$ é a função variância [4]. Os modelos de contagem utilizados nesta análise são definidos a seguir.

Modelo de Poisson: A distribuição de Poisson é dada pela função densidade de probabilidade (f.d.p.) $f(y;\mu) = e^{-\mu}\mu^y/y!$, onde o parâmetro μ representa o número médio de ocorrências em um intervalo [1]. Uma propriedade desta distribuição é que $E(y) = Var(y) = \mu$, e, por isso, o parâmetro de dispersão é fixo em $\phi = 1$. Agora suponha que y_i seja o número de ocorrências de um evento para um dado número de "exposições" n_i. Nesse caso, a esperança de y_i pode ser escrita como $E(y_i) = \mu_i = n_i\theta_i$. Se θ_i depende de variáveis explicativas, então pode ser modelado através de $\theta_i = e^{x_i^T\beta}$. Portanto, se $y_i \sim \text{Poisson}(\mu_i)$, o MLG é dado por $E(y_i) = \mu_i = n_i e^{x_i^T\beta}$.

Como a função de ligação canônica é a logarítmica, o modelo resultante é $\log \mu_i = \log n_i + x_i^T \beta$. Note que este modelo difere da especificação tradicional do componente linear $x_i^T \beta$, devido à inclusão do termo $\log n_i$ [1]. Este termo (chamado de *offset*) é uma "compensação" nos casos onde a variável resposta é observada em intervalos de comprimento variável, e não fixo.

Modelo quase-Poisson: Uma maneira de lidar com a superdispersão no modelo de Poisson é assumir que o parâmetro ϕ não seja fixo em 1, mas permitir que ele seja irrestrito e estimado a partir dos próprios dados [8]. O modelo de quase-Poisson obtêm as mesmas estimativas dos coeficientes que o modelo de Poisson, porém elas são ajustadas para "acomodar" a superdispersão. Como a distribuição da variável resposta para esse modelo depende apenas das funções média e variância [6], ele não possui verossimilhança totalmente especificada.

Modelo binomial negativo: A distribuição binomial negativa pode surgir quando consideramos um processo de Poisson onde o parâmetro μ também possui uma distribuição de probabilidade. Se essa distribuição for uma gama com parâmetro θ, a combinação destas distribuições é a BN com f.d.p. dada por $f(y;\theta,\mu) = \frac{\Gamma(\theta+y)}{\Gamma(\theta)y!}\frac{\mu^y\theta^\theta}{(\mu+\theta)^{\theta+y}}$. A esperança para esta distribuição é $E(y) = \mu$ e a variância é dada por $Var(y) = \mu + \mu^2/\theta$ [6]. Note que, ao contrário da distribuição de Poisson, aqui a variância é maior do que a média e θ é frequentemente chamado de parâmetro de superdispersão. Devido à inclusão desse parâmetro, esta distribuição tem uma maior capacidade de lidar com dados superdispersos.

Modelos de mistura: Estes modelos surgem quando a variável resposta é modelada como uma mistura de uma distribuição de Bernoulli para a presença/ausência de zeros e uma outra distribuição para os valores positivos. Se ω é a proporção de capturas iguais a zero, então a distribuição de mistura assume a forma geral

$$P(Y = y) = \begin{cases} \omega + (1-\omega)f(0) & y = 0 \\ (1-\omega)f(y) & y > 0 \end{cases} \tag{1}$$

onde $f(\cdot)$ é uma distribuição discreta como a Poisson ou BN, e ω deve ser interpretado como a probabilidade de que uma observação é proveniente desta distribuição [3]. Dessa forma, os parâmetros a serem estimados são ω (através de uma distribuição binomial) e os parâmetros da distribuição específica dada em $f(\cdot)$.

Modelos condicionais: Nesta abordagem, a modelagem da variável resposta é realizada em duas etapas. A primeira é quando nenhum animal ocorre, e a segunda é quando os animais ocorrem em diferentes níveis de abundância. Se ω for o parâmetro de probabilidade de uma distribuição de Bernoulli, então

$$P(Y = y) = \begin{cases} \omega & y = 0 \\ (1-\omega)f(y) & y > 0 \end{cases} \tag{2}$$

onde $f(\cdot)$ é uma distribuição discreta "truncada" (e.g. Poisson ou BN). A diferença para a abordagem anterior é que, aqui, ω é a probabilidade da captura ser igual a zero [7]. Dado que houve captura, o número de animais observados (capturados) pode ser modelado através da distribuição dada em $f(\cdot)$.

Ao todo foram ajustados os sete modelos descritos acima (Poisson, quase-Poisson, BN, mistura Poisson e BN, condicional Poisson e BN), para a variável resposta captura, e com a mesma matriz de covariáveis: ano (fator; 1960–2007), trimestre (fator; 1–4) e área (fator; NW, NE,

SW, SE, em relação ao Atlântico Sul, dividido na latitude 20°S e longitude 20°W). Com estas covariáveis fixas, as diferenças entre as predições deve ser resultado apenas da especificação de cada modelo. Convém ressaltar que em todos eles, o *offset* com o logaritmo do número de anzóis foi utilizado para compensar os valores agregados. A comparação entre esses modelos foi realizada através dos conhecidos critérios de informação de Akaike (AIC) e de Bayes (BIC). O número de zeros predito por cada modelo também foi estimado para uma comparação com o número de zeros observado. Todas as análises foram realizadas através do *software* R 2.9.0 [5], com auxílio dos pacotes MASS [6] e pscl [8].

3 Resultados

A proporção de zeros encontrada para a captura do agulhão branco no Atlântico Sul pela frota de barcos japoneses de espinhel de superfície foi de 63%. Os valores de captura variaram no intervalo [0, 4325]. A média de captura ponderada pelo esforço foi de cerca de $\bar{x} = 73$ peixes, com variância $s^2 = 37051$ para o período analisado (1960–2007). Com isto, pode-se constatar que a variância é cerca de 500 vezes maior do que a média, o que indica uma superdispersão considerável nos dados. Entre os modelos ajustados, o de Poisson foi o que apresentou os maiores valores de AIC e BIC, indicando o pior ajuste (Tabela 1). Os modelos de mistura e condicional ajustados com a distribuição de Poisson para os valores positivos tiveram resultados similares de acordo com os critérios de informação, e apresentaram um ganho pequeno em relação ao modelo de Poisson tradicional. Os modelos ajustados com a distribuição BN (tradicional, de mistura e condicional) apresentaram uma grande redução nos valores de AIC/BIC em relação aos modelos anteriores, o que indica um melhor ajuste. De acordo com esses critérios, o modelo de mistura BN foi o que apresentou o melhor ajuste entre todos (menor AIC/BIC). O número de zeros predito pelos modelos condicionais (Poisson e BN) foi exatamente igual ao número de zeros observado (i.e. 9545). O modelo de mistura BN, com o melhor ajuste de acordo com os critérios de informação, foi o segundo modelo que mais se aproximou do número de zeros observado. O modelo de Poisson subestimou substancialmente o número de zeros.

Tabela 1: Graus de liberdade (GL), logaritmo da verossimilhança ($\log L$), AIC, BIC, e número de zeros preditos (o número de zeros observados foi de 9545) pelos diferentes modelos ajustados.

	Poisson	Quase-Poisson	BN	Poisson (mistura)	BN (mistura)	Poisson (condicional)	BN (condicional)
GL	54	54	55	108	109	108	109
$\log L$	-357997	—	-29272	-313999	-28564	-314698	-29589
AIC	716102	—	58653	628215	57347	629613	59396
BIC	716514	—	59073	629039	58178	630437	60228
$\sum_i \hat{f}_i(0)$	5727	—	9370	9473	9517	9545	9545

4 Discussão

A comparação entre os modelos descrita acima mostra claramente que os modelos que consideraram a distribuição BN tiveram melhores resultados (i.e. melhores ajustes) do que os demais

modelos onde a distribuição de Poisson foi utilizada. Note que mesmo os modelos de mistura e condicional de Poisson, que levam em conta o excesso de zeros, não foram suficientes para que boas estimativas pudessem ser feitas. Isso se deve possivelmente ao fato de que a superdispersão é tão grande (note que a amplitude de valores de captura foi muito elevada), que mesmo que um modelo considere o excesso de zeros, ainda falta considerar essa superdispersão excessiva. É nesse ponto que o modelo BN leva vantagem, pois possui um parâmetro a mais (θ) que é responsável pela "acomodação" dessa grande variabilidade.

No geral, os modelos de mistura demonstraram um melhor desempenho se comparados aos modelos condicionais (com a mesma distribuição). Como os modelos de mistura assumem que os processos que levam à obtenção de zeros e os que levam à valores positivos são distintos [3], esses modelos parecem ser mais apropriados para a modelagem de dados de captura incidental. Isso se justifica pois as capturas iguais a zero são provenientes principalmente dos denominados "falsos zeros", que são, por exemplo, resultado de um lance onde nenhum exemplar de agulhão foi capturado simplesmente por não estar presente naquele local e período de tempo. Além disso, os resultados obtidos demonstram que estimativas de abundância de espécies capturadas incidentalmente devem ser feitas através de modelos que considerem tanto a superdispersão como o excesso de zeros. Caso contrário, subestimativas significativas podem ocorrer e interferir no processo de decisão de manejo.

Referências

[1] DOBSON, A. J., *An introduction to Generalized Linear Models*, London: Chapman & Hall, 2002.

[2] HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: Models and estimation. *Computational Statistics & Data Analysis*, v. 27, p. 151–170, 1998.

[3] LAMBERT, D., Zero-Inflated Poisson regression, with application to defects in manufacturing, *Technometrics*, v. 34, p. 1–14, 1992.

[4] McCULLAGH, P.; NELDER, J. A., *Generalized Linear Models*, London: Chapman & Hall, 1989.

[5] R DEVELOPMENT CORE TEAM, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2009, ISBN 3-900051-07-0, URL `www.R-project.org`.

[6] VENABLES, W. N.; RIPLEY, B. D., *Modern applied statistics with S*, Nova Iorque: Springer, 2002.

[7] WELSH, A. H.; CUNNINGHAM, R. B.; DONNELY, C. F.; LINDENMAYER, D. B., Modelling the abundance of rare species: statistical models for counts with extra zeros, *Ecological Modelling*, v. 88, p. 297–308, 1996.

[8] ZEILEIS, A.; KLEIBER, C.; JACKMAN, S., Regression models for count data in R, *Journal of Statistical Software*, v. 27, p. 1–25, 2008.