Modélisation de la distribution de petits échantillons de données d'abondance : exemple des poissons du bassin du Rhône

Lise Vaudor, N. Lamouroux

MODELISATION DE LA DISTRIBUTION DE PETITS ECHANTILLONS DE DONNEES D'ABONDANCE : EXEMPLE DES POISSONS DU BASSIN DU RHONE

Lise Vaudor et Nicolas Lamouroux

Cemagref (DYNAM), 3 bis Quai Chauveau, CP 220, 69336 Lyon Cedex 09

Résumé

On étudie la distribution statistique de données d'abondance pour des poissons d'eau douce dans le bassin du Rhône. On s'appuie sur un jeu de données très grand d'échantillons ponctuels d'abondance collectés par pêche électrique, qui comprend 12 espèces sur 7 tronçons et plus de 20 ans, soit au total 2258 échantillons. Les données d'abondance sont connues pour être généralement asymétriques à droite et surdispersées. Dans notre cas, la plupart des échantillons sont en outre de petite taille (78% des échantillons comptent moins de 30 points) et sont d'abondance moyenne faible (50% des échantillons ont une abondance moyenne inférieure à 0.6 individu par point). On propose quatre modèles différents de la distribution de nos données d'abondance : une distribution de Poisson enflée en zéro, une binomiale négative, une binomiale négative enflée en zéro, et une distribution hurdle de Pareto. Globalement, le modèle le plus fréquemment sélectionné est celui de la binomiale négative (46% des échantillons). Cependant la qualité de l'ajustement dépend beaucoup de certaines caractéristiques des échantillons comme l'abondance moyenne : en particulier, le modèle de Poisson enflé en zéro est sélectionné pour 56% des échantillons dont l'abondance moyenne est inférieure à 0.6 individu par point. Nous illustrons l'importance du choix d'un modèle approprié (ici, une distribution binomiale négative) à travers l'exemple de l'intervalle de confiance pour l'abondance moyenne. On compare deux types d'intervalles de confiance : ceux qui reposent sur une hypothèse de distribution négative binomiale, et qui sont calculés grâce au profil de vraisemblance, et ceux qui reposent sur une hypothèse de distribution gaussienne. Les intervalles de confiance reposant sur l'hypothèse de distribution binomiale négative sont généralement plus larges (92% des échantillons), en particulier quand l'échantillon compte peu d'abondances non nulles, et sont toujours plus longues du côté droit de l'estimation de la moyenne d'abondance.

We study the statistical distribution of count data for freshwater fish abundance in the Rhone basin. We rely on a huge data set of point-abundance samples collected by electrofishing, that comprises 12 species over 7 reaches and more than 20 years, being in total 2258 samples. Count data is known to be generally right-skewed and overdispersed. In our case, most samples are also small-sized (78% of samples represent less than 30 points) and have low mean abundance (50% of samples have a mean abundance weaker than 0.6). We propose four different models for count data : a zero-inflated Poisson, a negative binomial, a zero-inflated negative binomial, and a two-part Pareto distribution models. For each sample, we fit these four models by maximum-likelihood and select one model according to the BIC criterion. Overall, the negative binomial is the most often selected distribution model (46% of samples). However the goodness of fit depends a lot on sample features such as mean abundance: in particular, the zero-inflated Poisson model is selected in 56% of the samples whose mean abundance is weaker than 0.6 individual per point. We illustrate the importance

of choosing an appropriate model (here, a negative binomial distribution) through the example of confidence interval estimation for mean abundance. We compare two kinds of confidence intervals: those based on a binomial negative distributional assumption and calculated through profile-likelihood, and those based on a Gaussian distributional assumption. Binomial negative-based confidence intervals are generally wider (92% of samples), in particular when there are very few non-null counts, and always longer on the right side of mean abundance estimate.

Mots-clés

Abondance, Maximum de vraisemblance, Distribution de Poisson, Distribution binomiale négative, Distribution enflée en zéro, Choix de modèle, BIC, Intervalles de confiance, Profil de vraisemblance

Bibliographie

[1] Anscombe, F. J. (1949) The Statistical Analysis of Insect Counts Based on the Negative Binomial Distribution. *Biometrics* 5(2), 165-173.
[2] Bliss, C. I. et Fisher, R. A. (1953) Fitting the negative binomial distribution to biological data. *Biometrics* 9(2), 176-200.
[3] Brown, L. D., T. T. Cai, et al. (2003) Interval estimation in exponential families. *Statistica Sinica* 13(1): 19-49.
[4] Cai, T. (2005) One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference* 131(1),63-88.
[5] Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics* 3: 22-38.
[6] Di Stefano, J., F. Fidler, et al. (2005) "Effect size estimates and confidence intervals: an alternative focus for the presentation and interpretation of ecological data". In *New Trends in Ecology Research*, 71-102, sous la dir. de A.R. Burk, Nova Science Publishers, New York.
[7] Gray, B. R. (2005) Selecting a distributional assumption for modelling relative densities of benthic macroinvertebrates. *Ecological Modelling* 185(1), 1-12.
[8] Lambert, D. (1992) Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing *Technometrics* 34(1), 1-14.
[9] Martin, T. G., Wintle, B. A., et al. (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters* 8(11), 1235-1246.
[10] McArdle, B. H. and Anderson, M. J. (2004). Variance heterogeneity, transformations, and models of species abundance: a cautionary tale. *Canadian Journal of Fisheries and Aquatic Sciences* 61(7), 1294-1302.
[11] McArdle, B. H., Gaston, K. J., et al. (1990). Variation in the size of animal populations: patterns, problems, and artefacts. *Journal of Animal Ecology* 59(2), 439-454.
[12] McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. Editions Chapman and Hall, New York.
[13] Potts, J. M. et Elith, J. (2006) Comparing species abundance models. *Ecological Modelling* 199(2), 153-163.
[14] Power, J. H. and Moser, E. B. (1999) Linear model analysis of net catch data using the negative binomial distribution. *Canadian Journal of Fisheries and Aquatic Sciences* 56(2), 191-200.
[15] Venzon, D. J. and S. H. Moolgavkar (1988) Method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society. Series C: Applied Statistics* 37(1), 87-94.

[16] Welsh, A. H., Cunningham, R. B., et al. (1996) Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling* 88(1-3), 297-308.