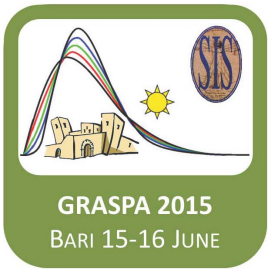

Statistical models for species richness in the Ross Sea

Cinzia Carota[1,*], Consuelo R. Nava[1], Irene Soldani[2], Claudio Ghiglione[3] and Stefano Schiaparelli[3]

[1] *Department of Economics and Statistics "Cognetti de Martiis", University of Torino; cinzia.carota@unito.it, consuelorubina.nava@unito.it*
[2] *aizoOn Technology Consulting; irene.soldani@aizoon.it*
[3] *DiSTAV, University of Genova and Italian National Antarctic Museum (section of Genova); claudio.ghiglione@riftia.eu, stefano.schiaparelli@unige.it*
[*] *Corresponding author*

Abstract. *In recent years, a large international effort has been placed in compiling a complete list of Antarctic mollusc distributional records based both on historical occurrences, dating back to 1899, and on newly collected data. Such dataset is highly asymmetrical in the quality of contained information, due to the variety of sampling gears used and the amount of information recorded at each sampling station (e.g. sampling gear used, sieve mesh size used, etc.). This dataset stimulates to deploy all statistical potential in terms of data representation, estimation, clusterization and prediction.*
In this paper we aim at selecting an appropriate statistical model for this dataset in order to explain species richness (i.e. the number of observed species) as a function of several covariates, such as gear used, latitude, etc.. Given the nature of data, we preliminary implement a Poisson regression model and we extend it with a Negative Binomial regression to manage over-dispersion. Generalized linear mixed models (GLMM) and generalized additive models (GAM) are also explored to capture a possible extra explicative power of the covariates. However, preliminary results under them suggest that more sophisticated models are needed. Therefore, we introduce a hierarchical Bayesian model, involving a nonparametric approach through the assumption of random effects with a Dirichlet Process prior.

Keywords. *Bayesian hierarchical model; Dirichlet Process; GAM; GLMM; Ross Sea.*

1 Introduction

Since many years, an international team of researchers has focused its attention on distributional data of Ross Sea (Antarctica) Mollusca, compiling a large dataset based on revised species identification and classification. The selection of this geographical position is crucial, especially in the light of the effects that climate changes might have on the biodiversity of the area. The dataset is the result of several scientific expeditions, performed with different goals, that span for a temporal timeframe of more than one century, specifically from 1899.

This dataset results to be highly asymmetrical in terms of available information. Expeditions in the last century essentially aimed at making a census of the Antarctic species while recent expeditions apply

balanced sampling designs that enable better statistical analyses and are focused on the study of species spatial and geographical distribution.

Hence, there have been some difficulties in the treatment and the adaption of the data collected before 2004, for instance due to the lack of information about species picked up dead or alive. Moreover, from 1899 to 2004, there is no record of "zero occurrences", i.e. stations that have been properly investigated but where no molluscs were found. This inevitably affects species richness, i.e. the number of different species observed in each sampling unit or station, which is the most used variable in biodiversity studies.

Despite these limitations, collected data remain a precious and unique source of information (see [8]) and several papers are going to be published based on these data, as [12]. Here we focus on species richness: our response variable Y. We also consider covariates such as the tools employed to collect sampling units. They can be grab, towed gears, Rauschert dredge (i.e. a towed dredge with a very fine mesh), or even "unknown" (i.e. where the gear was not recorded in the data log) or "multiple" (i.e. where more gears were deployed at the same station). In addition, geographical variables such as longitude, latitude, depth and distance from the nearest scientific station are taken into account. Successively, a factor geographical covariate, referred to as box[1], has been introduced.

2 Methods and results

We investigate the explanatory and predictive power of a large number of models and methods for count data. The simple Poisson regression model, inadequate because of over-dispersion, absence of zeros and excess of 1s in the data (see Figure 1 for a representation), has been variously enriched and made more flexible [3].

First, we introduce random effects of different nature alongside the effects of the covariates described in Section 1. In particular, we assume that $y_i \sim \text{Poisson}(\mu_i)$, for $i = 1,\dots,n$ and $log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} + \phi_i$, where $\mathbf{x}_i$ represents a $q \times 1$ vector of covariates, $\boldsymbol{\beta}$ is a $q \times 1$ vector of fixed effects, and ϕ_i denotes a random effect accounting for observation specific deviations. In regarding the distribution of ϕ_i, denoted by G, two parametric assumptions are compared: $\phi_i \overset{iid}{\sim} N(0,\sigma^2)$ and $e^{\phi_i} \overset{iid}{\sim} \text{Gamma}(a,b)$ with $a = b$, so that $E(e^{\phi_i}) = 1$ and $Var(e^{\phi_i}) = 1/a$. The latter assumption introduces extra-variability on a different scale as ordinary predictors ([1], p.556) and leads to the Negative Binomial regression model [3, 9]:

$$y_i \sim \text{NB}\left(a, \frac{a}{a + e^{\mathbf{x}_i'\boldsymbol{\beta}}}\right),\ i = 1,\dots,n. \tag{1}$$

where $E(Y_i) = e^{\mathbf{x}_i'\boldsymbol{\beta}}$ and $Var(Y_i) = e^{\mathbf{x}_i'\boldsymbol{\beta}}(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}/a)$. As count data in ecology are often clumped (if the rate of capture of individuals varies randomly), producing an expected variance that is greater than the mean, in such literature [2] the parameter a is often referred to as the *clumping parameter* [2, 13].

We also explicitly consider special generalized linear mixed models, GLMM, where subsets of the n observations are given the same random effect, as for instance observations in the same box.

Given the absence of zeros, we explore truncated versions of the models just described and, for comparisons, we also apply linear mixed models to a log-transformation of the respose variable, a controversial practice very often recommended in the ecological literature.

Moreover, in stations where the gear is unknown, we try to impute its value in order to improve such a covariate.

Then, trying to increase the potential of all available covariates to explain the species richness, we explore more general parametric models such as generalized additive models, GAM [14].

[1] Boxes are defined with 1 degree of latitude and 1 of longitude. In the dataset 112 boxes are identified in which there have been at least the observation of one species.

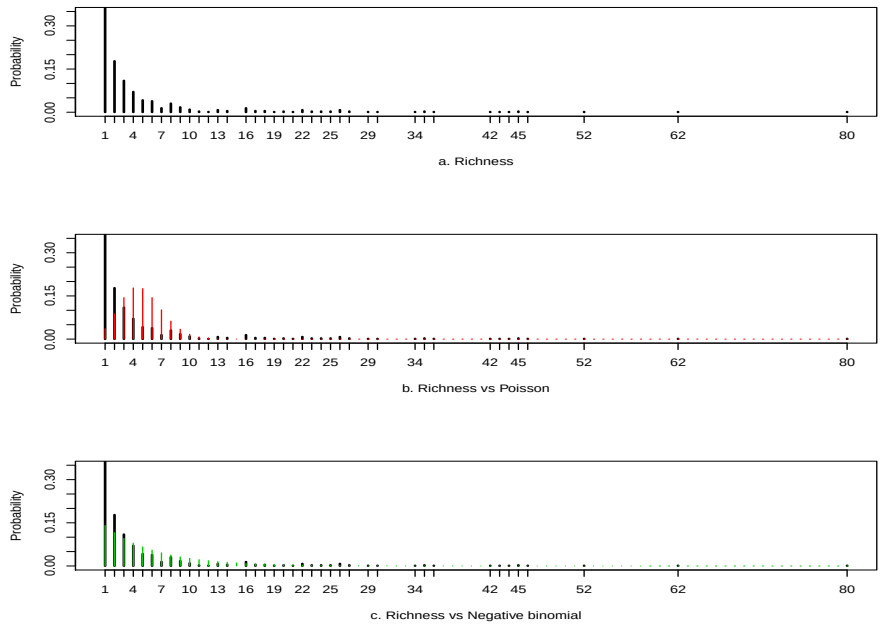

Figure 1: We compare the species richness, represented in plot **a**, respectively with $Y \sim \text{Poisson}(4.932)$ in graph **b** and with $Y \sim \text{NB}(0.977, 0.165)$ in plot **c**. The parameters of the Poisson and Negative Binomial distribution are estimated from the observed data.

Although, in terms of reduction of the residual deviance and AIC, all the strategies illustrated above provide appreciable contributions, their predictive power turns out to be further improvable, precisely because of the clumping of the data. In order to make the model able to capture the multimodal distribution of species richness (see Figure 1.a), we decided to re-interpret the described GLMMs as Bayesian hierarchical models and add the further level described below to the hierarchy.

We relax the assumption on the parametric form of the distribution function of random effects G and we model it by a Dirichlet Process prior $\mathcal{D}$ with base probability measure G_0 and total mass parameter m [7],

$$\phi_i | G \overset{iid}{\sim} G, \qquad G \sim \mathcal{D}(m, G_0),\ m > 0. \tag{2}$$

Considering that $E(G) = G_0$ and m controls the variance of the process, in practice G_0 specifies one's "best guess" about an underlying model of the variation in ϕ, and m identifies the extent to which G_0 holds ([6], p. 638). Within the class of models just defined, we consider specifications of G_0 that lead to direct generalizations of the GLMMs described above, namely $G_0 = N(\alpha, \sigma^2)$ and $G_0 = LG(a,b)$. LG denotes the distribution of ϕ_i, being $e^{\phi_i} \overset{iid}{\sim} \text{Gamma}(a,b)$ with $a = b$ as already discussed. Moreover, vague priors are assumed on β, a and m [4].

Under the previous assumptions, the likelihood function turns out to be a sum of terms where all possible partitions (clusterings) of the n observations into nonempty clusters are considered [10, 11]. This fact implies that:

i. to learn about a given observation/station, additional information to the one provided by covariates is borrowed from observations/stations belonging to the same subset, for each subset to which the observation can be assigned in the context of all possible partitions in nonempty subsets of the n

observations;

ii. the results under a hierarchical semi-parametric model with Dirichlet process random effects can be interpreted as averages over GLMMs, corresponding to all possible clusterizations of the $N(0,\sigma^2)$ or $LG(a,b)$ parametric random effects.

3 Conclusions

The natural implementation of discussed parametric statistical models – Poisson regression, Negative Binomial regression, GAM or GLMM – only partially explain our variable of interest. Multimodality and over-dispersion of species richness can be jointly modeled by adopting a more general non parametric hierarchical Bayesian approach as confirmed by the encouraging preliminary results we have obtained.

References

[1] Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons.

[2] Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35** (3–4), 246–254.

[3] Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, vol. 53. Cambridge University press.

[4] Carota, C., Filippone, M., Leombruni, R. and Polettini, S. (2015) Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *Annals of Applied Statistics* **9**, 525-546.

[5] Clarke, A. (2008). Antarctic marine benthic diversity: patterns and processes. *Journal of Experimental Marine Biology and Ecology*, **366**(1),48–55.

[6] Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L. and Jordan, F. (2008). Modeling Unobserved Sources of Heterogeneity in Animal Abundance Using a Dirichlet Process Prior. *Biometrics* **64**, 635–644.

[7] Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* **1**, 209–230.

[8] Griffiths, H.J., Danis, B. and Clarke, A. (2011). Quantifying Antarctic marine biodiversity: the SCAR-MarBIN data portal. *Deep-Sea Res II* **58**,18–29.

[9] Hilbe, J., (2007). *Negative Binomial Regression*. Cambridge University Press, Cambridge, UK.

[10] Liu, J. S. (1996). Nonparametric Hierarchical Bayes via Sequential Imputations. *Annals of Statistics* **24**, 911–930.

[11] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Annals of Statistics* **12**, 351–357.

[12] Schiaparelli, S., Ghiglione, C., Alvaro, M. C., Griffiths, H. J., and Linse, K. (2014). Diversity, abundance and composition in macrofaunal molluscs from the Ross sea (Antarctica): results of fine-mesh sampling along a latitudinal gradient. *Polar biology* **37**(6), 859–877.

[13] Young, L.J and Jerry Youn, J. (1998). *Statistical Ecology*. Spinger.

[14] Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A. and Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.