Environmetrics 00, 1–23

DOI: 10.1002/env.XXXX

A Bayesian semi-parametric GLMM for historical and newly collected presence-only data: an application to species richness of Ross Sea Mollusca

C. Carota[a*] , C. R. Nava[a], C. Ghiglione[b] and S. Schiaparelli[b]

Summary: Historical datasets from vast and relatively inaccessible areas are sources of potentially unique information still valuable for biodiversity studies today. In many research fields, ranging from climate change to projection of species loss, great efforts have been made to integrate historical datasets with recent data to create databases that are as complete as possible. Unlocking the information contained in presence-only data, largely prevalent in such databases, presents a challenge for statistical modeling because of insidious observational errors due to the opportunistic nature of the data gathering process. In this article we propose an appropriate statistical method for the joint analysis of historical and newly collected presence-only data, i.e. a Bayesian semi-parametric generalized linear mixed model (GLMM) with Dirichlet process random effects. The potential of the method is illustrated by considering the Ross Sea section of the SOMBASE, an international compilation of Southern Ocean Mollusc distributional records, from 1899 to 2004 and beyond. Despite the presence of sampling bias and non-detection errors, the proposed model draws latent information from the data such that the resulting estimates of the parameters of interest are not only coherent with those obtained in indirectly related studies based on well structured data, but also suggest interesting ideas for further research.

Keywords: Bayesian hierarchical GLMM, Dirichlet process random effects, Opportunistic sampling schemes, Presence-only data, Ross Sea Mollusca, Species richness.

[a] *Department of Economics and Statistics "Cognetti de Martiis", Università degli Studi di Torino, Lungo Dora Siena 100, Torino, 10100, Italy*

[b] *DiSTAV, University of Genova and Italian National Antarctic Museum (section of Genova), Corso Europa 36, Genova, 16132, Italy*

[*] *Correspondence to: C. Carota, Department of Economics and Statistics "Cognetti de Martiis", Università degli Studi di Torino, Lungo Dora Siena 100A, Torino, 10100, Italy. E-mail: cinzia.carota@unito.it*

This paper has been submitted for consideration for publication in *Environmetrics*

1. INTRODUCTION

In biodiversity research a considerable effort has been devoted to the digitization and integration of historical datasets with recent data to create databases that are as complete as possible and make them available to the scientific community. In particular, close attention has been paid to the so-called primary biodiversity data, hereafter also referred to as presence-only data. Pearce and Boyce (2006) define presence-only data as "consisting only of observations of the organism but with no reliable data on where the species was not found. Sources of these data include atlases, museum and herbarium records, species lists, incidental observation databases and radio-tracking studies." Such opportunistic data – i.e. data collected by non-standardized means, with no sampling design and no standardized protocol – are continuously increasing, also because of the recent development of citizen science programs, and currently constitute the main component of the biodiversity information stored in large-scale aggregators like the GBIF[†] or the BioCASE[‡]. A number of statistical packages allow the easy retrieval of biodiversity information from these and other repositories; and, of course, statistical methods that can deal with opportunistic sampling schemes are increasingly required. In this article we propose a method for the joint analysis of historical and newly collected presence-only data and we present an application to the study of species richness of Ross Sea Mollusca.

Presence-only data, i.e. points in space and time where a species has been recorded as being present, are often used to model the distribution of a species as a function of a set of explanatory variables (e.g Warton and Shepherd, 2010; Chakraborty et al., 2011; Dorazio, 2012; Fithian and Hastie, 2013; Renner and Warton, 2013; Giraud et al., 2016). Chakraborty et al. (2011) indicate how to use presence-only data in modeling species richness, that is the number of distinct species in a given area. More precisely, they provide a parametric

[†]Global Biodiversity Information Facility.
[‡]Biological Collection Access Service.

function for the expected richness based on a non-homogeneous Poisson process model for the set of observed presence points. Under the assumption that a set of possible species can be specified, they obtain inferences endowed with measures of uncertainty, thereby improving on the often-used approach with MAXENT (Renner and Warton, 2013). The availability of environmental features able adequately to explain the intensity of each species, however, is a prerequisite of this method not easily satisfied in cases where the set of possible species is large and/or includes species whose characteristics are very different or partially unknown. Moreover, determining what types of observational errors may have occurred with presence-only data and how these errors can be modeled is a challenging issue. Common errors are sampling bias, non-detection and location error, all of which result in biased parameter estimates and predictions. These issues are addressed in an increasing number of articles (e.g. Dorazio, 2012; Fithian et al., 2015; Dorazio, 2014; Warton and Shepherd, 2010; Hefley and Hooten, 2016), as this is a recent, very active area of research. Statistical methods suggested to account for such errors – often in articles focussed on species distribution modeling – consist of a variety of strategies aimed at achieving identifiability of the parameters of interest in appropriate generalized linear models (GLMs) by exploiting auxiliary information provided by data collected in independent, planned surveys. Such strategies range from applying regression calibration in the presence of location error (Hefley and Hooten, 2016) to introducing suitable functions of the original parameters in the presence of unknown sampling effort and detectability (see Giraud et al. (2016), where the interest lies in relative abundances of multiple species on multiple sites). Furthermore, many of these strategies rely on strong or uncheckable assumptions, as, for instance, that observational bias toward some species is the same across different sites (Giraud et al., 2016). See also Dorazio (2014); Renner and Warton (2013). In this regard, Fithian et al. (2015) conclude their paper by saying "[. . .] in our approach, we are forced to assume a functional form for the sampling bias, and if our model is wrong, we will not account correctly for sampling bias. [. . .] in future

work we plan to investigate models that treat the sampling bias nonparametrically, imposing no assumption on its parametric form." To the best of our knowledge, there are currently no methods that try to use presence-only data without using auxiliary information, nor methods that exploit the potential of the Dirichlet process (DP) in modeling sampling bias, detection errors, and, more generally, the very complicated errors resulting from opportunistic data collation[§].

In this article we propose a Bayesian semi-parametric method useful for this purpose. Specifically, we focus on species richness and model presence-only data through a generalized linear mixed model (GLMM) with Dirichlet process (Ferguson, 1973; Antoniak, 1974) random effects. We show that each model belonging to this general family of models is equivalent to a mixture of standard parametric GLMMs with observation-specific random effects grouped in all possible ways. For this reason models belonging to this family capture a wider range of variability in the random component and induce more accurate estimates of the parameters of interest that are given by the coefficients (fixed effects) of a set of explanatory variables. The proposed method is evaluated using presence-only data affected by sampling bias and non-detection, taken from the Ross Sea section of the SOMBASE, an international compilation of Southern Ocean Mollusc distributional records (Griffiths et al., 2003). In this case we specify a Bayesian semi-parametric GLMM with area-specific (spatial) random effects. Interestingly, in the absence of auxiliary information, this model draws latent information from these opportunistic data such that the resulting estimates of the parameters of interest are coherent with the results obtained in indirectly related studies based on data collected in planned surveys. We interpret this result as evidence of the capacity of the proposed model to induce estimates not severely biased by the presence of the above mentioned two types of errors. Moreover, such estimates (and the corresponding predictions) are relatively more

[§]Only in Chakraborty (2010) a hierarchical Dirichlet process is applied to cluster the presence localities of multiple species and, subsequently, to develop measures of range overlap from posterior draws.

accurate than the ones obtained under a parametric GLMM including the same fixed effects and random effects distributed according to a Normal distribution. Overall, such results suggest that Bayesian semi-parametric modeling can open new avenues for application of presence-only data, thereby helping to mobilize a wealth of biodiversity information largely underutilized to date.

For similar reasons, i.e. to adjust for confounding effects of unknown underlying factors, similar models are considered by Gill and Casella (2009) and Kyung et al. (2010, 2011) to model political science data. Dorazio et al. (2008) use a GLMM with DP random effects to model spatial heterogeneity in animal abundance. See also Carota et al. (2015), where a similar approach is used to model population and sample frequencies in a privacy protection problem.

A review of a large number of Bayesian nonparametric models is conducted by Phadia (2013); fundamentals of nonparametric Bayesian inference are discussed at great length in Ghoshal and van der Vaart (2016), while Mueller et al. (2015) focus on nonparametric Bayesian data analysis (see Gelman et al., 2014). Other recent discussions on nonparametric priors include Hjort et al. (2010); Ghoshal (2010); Mueller and Quintana (2004); Mueller and Rodriguez (2013) and Walker (2013).

The article is structured as follows. Section 2 presents material and methods: the structure of the data and their peculiarities and limitations are described in sub-section 2.1; the general features characterizing the proposed model are presented in sub-section 2.2. In Section 3 a Bayesian GLMM with spatial random effects a priori distributed according to a mixture of Dirichlet Processes is applied to data on Ross Sea Mollusca and compared to several simpler parametric GLMMs. There we see that the proposed semi-parametric hierarchical model exhibits a good predictive performance and provides meaningful information about the parameters of interest (fixed effects), thereby providing a useful guide for future research. Finally, Section 4 provides a brief summary of the article and sketchs further research

directions.

2. MATERIALS AND METHODS

2.1. Illustrative dataset and motivation

Our illustrative dataset is a compilation of primary biodiversity records collected in a series of scientific expeditions in the Ross Sea region of Antarctica. This is one of the most pristine ecosystems remaining on the planet (Ainley, 2002, 2010), and it has become an increasingly important hub for biodiversity studies, hosting several permanent research stations and numerous scientific expeditions. Biodiversity data on the region have been published since 1899, when the British Southern Cross Expedition overwintered in Antarctica (see Schiaparelli et al. (2014) for a review); moreover, some of the "historical" sites have never been re-sampled, thereby standing as the only source of biodiversity data for those areas.

The Ross Sea section of the SOMBASE – the most complete database available to date for any biogeographical or biodiversity study of marine molluscs in Antarctica – contains 293 species collected from 1899 to 2004 at different sites (sampling stations) in 619 discrete sampling events. Data collected over this extended period of time present many challenging issues for statistical analysis. Although they are always the results of repeated scientific efforts (rather than being contributed by citizens), such data clearly suffer from sampling bias because of the different sampling protocols applied in different scientific expeditions. Moreover, a severe issue of unknown, and not identifiable, detection probability is created by the lack of useful information about the gears used in data collection, as explained in the next paragraph. Vice versa, they do not suffer from location error: for all 619 sampling events we have perfectly reliable, point observations on latitude, longitude, and, in addition, distance from the nearest scientific station and maximum depth. Maximum depth is the value of depth at the site reported for towed gears and reduces to the value of depth at the

site for the remaining gears.

Indeed, specific mention of the tools used to collect data in each sampling station (site) is often made throughout the dataset (see Figure 1).

[Figure 1 about here.]

Such tools include grab, towed gears and fine-mesh (0.5 mm) towed gears, such as the Rauschert dredge used in the Ross Sea for the first time in 2004 (Schiaparelli et al., 2014). Nonetheless, the database is not always consistent in reporting the mesh size of the towed gears, and, even worse, values of the gear are sometimes reported as "unknown" (gear type not recorded in the expedition logbook and hence not evaluable in terms of sampling performance), "multiple" (when a variety of sampling gears or several copies of the same gear have been used at the same site, making it impossible to distinguish the contribution of each gear/copy to the total number of species/individuals at a sampling site) and "other" (e.g. sampling by hand during a scuba dive). For statistical analysis, it is very difficult to treat these values; therefore in our analysis the gear is not exploited as a covariate so as not to sacrifice part (i.e. 163) of the 619 observations. This choice implies that in modeling species richness we are acutely aware of the existence of underlying unobserved or poorly measured factors that determine hidden clusters in the data (and "gear" is certainly one of them). Not accounting for these factors "[...] means that some systematic component of the data falls to an error term, exacerbating efforts to find parsimonious models with a good fit. Lacking direct covariate information about their effects, we seek here to find help in the data itself by specifying a nonparametric prior that reflects information in the data to help account for underlying structure in the context of the model. Thus this heterogeneity is actually the motivation for our methodological approach [...]" (Gill and Casella, 2009, p. 2)¶.

¶See also Kyung et al. (2011), where data on terrorists and terrorist attacks are successfully modeled by a similar approach. These data are either observed public events, which omit planned but failed or cancelled attacks (i.e. "undetectable" events), or classified information at government agencies that are "inaccessible" to general researches. In other words, as in our case, but for completely different reasons, these data suffer from both non-detection and sampling bias.

Another crucial aspect often omitted in the historical data is the abundance of species observed, which prevents studying biodiversity with measures more complex than species richness. In conclusion, the structure of our dataset is the one described in Table 1, where each cell registers, according to each site (sampling station) j, the presence (✓) of a specific species i. This is a matrix with 293 rows, representing the different species observed in at least one of the 619 sites, recorded in columns, and we are interested in the total number of distinct species along each column, or "species richness", y_j, observed at each site j, $j = 1, \ldots, 619$.

[Table 1 about here.]

Despite the high heterogeneity of these data, here, for the first time, we model all 619 available observations on species richness of Ross Sea Mollusca, y_j, as a function of the above mentioned geographical covariates: latitude, longitude, distance from the nearest scientific station and maximum depth. To date, only samples from single research expeditions have been studied (see Schiaparelli et al., 2006, 2014, e.g.). This choice will be further discussed in Section 3, after a general description of the proposed model in the next sub-section. Here, we just observe that:

(i) a first noticeable consequence of the joint analysis of data from different scientific expeditions is an economic benefit since data collection is an expensive and difficult task in the case of Ross Sea Mollusca and in many other biodiversity studies;

(ii) this illustrative dataset is used to clarify the potential of the proposed model with presence-only data and, more in general, with problematic data.

For richer datasets, for instance datasets obtained by applying well-structured sampling protocols (see Royle and Dorazio (2008) and references therein), such potential can be exploited from within more complex hierarchical models where, for instance, there are levels of the hierarchy explicitly devoted to estimate detection probabilities or to model species

occurrences. Indeed, for very rich datasets "[...] multispecies occupancy models provide an integrated approach to modelling both community features, such as species richness, as well as features of individual species, such as occupancy and habitat use, while accounting for species-specific detection probabilities [...]" (Tobler et al., 2015). When datasets are "poor" such an approach cannot be successfully adopted; there are however different models, like the one described here, that can be suitably deployed to try to retrieve valuable information from problematic data.

2.2. The proposed family of models

The heterogeneity and the latent clustering due to incompleteness of data and meta-data associated with a vast number of sampling events are problems quite common in primary biodiversity databases including historical data. As previously mentioned, the basic reason why we propose a Bayesian nonparametric approach – specifically the family of hierarchical semi-parametric GLMMs described below – is that it allows us to exploit a data-driven clustering of observations where the grouping is done nonparametrically rather than on prefixed criteria. This clustering property is a distinctive feature of this family and proves to be useful in many circumstances. Here we describe this property and suggest exploiting it with presence-only data. In Section 3 we use a specific model belonging to this family to cope with the confounding effects due to sampling bias and non-detection, i.e. the two types of errors affecting the dataset on Ross Sea Mollusca.

At the first stage of the proposed hierarchical model any parametric distribution belonging to the exponential family, say $f(\boldsymbol{y})$, and any link function $h(\boldsymbol{\mu})$ can be specified for the response variable $\boldsymbol{y} = [y_1, \ldots, y_n]$ and its mean $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]$. For instance, in our case,

because the species richness is a count, we assume that:

$$\begin{aligned} &y_j|\mu_j \sim \mathsf{Poisson}(\mu_j),\ \forall\, j = 1, \ldots, n \\ &h(\mu_j) = \log(\mu_j) = \boldsymbol{x}_j'\boldsymbol{\beta} + \phi_j \end{aligned} \tag{1}$$

where $\boldsymbol{x}_j$ represents a $p \times 1$ vector of covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, and ϕ_j denotes a random effect accounting for observation-specific deviations.

Then the distribution function of the random effect ϕ_j, say G, is assumed to be unknown and drawn from a mixture of Dirichlet processes assigned to the family of all possible distribution functions on the real line. The large support of this prior provides great flexibility in estimating G and enhances the ability to account for unobservable or poorly understood sources of heterogeneity in the response variable.

Specifically, G is modelled by a Dirichlet process prior, $\mathcal{D}(\alpha, G_{0(\boldsymbol{\theta})})$, with precision parameter α and base probability measure $G_{0(\boldsymbol{\theta})}$ (Ferguson, 1973), where G_0 and $\boldsymbol{\theta}$ are a fixed distribution function and a vector of hyperparameters, respectively. When the hyperparameters are unknown, the prior assigned to $\boldsymbol{\theta}$ leads to a mixture of DPs (Antoniak, 1974). The general expression of the resulting GLMM is as follows,

$$\begin{aligned} &y_j|\mu_j \sim f(y_j|\mu_j),\ \forall\, j = 1, \ldots, n \\ &h(\mu_j) = \boldsymbol{x}_j'\boldsymbol{\beta} + \phi_j \\ &\phi_j|G \overset{iid}{\sim} G \\ &G|\alpha, G_{0(\boldsymbol{\theta})} \sim \mathcal{D}(\alpha, G_{0(\boldsymbol{\theta})}) \\ &\alpha \sim \mathsf{Gamma}(a_0, b_0); \quad \boldsymbol{\beta} \sim \pi_1(\boldsymbol{\beta}); \quad \boldsymbol{\theta} \sim \pi_2(\boldsymbol{\theta}) \end{aligned} \tag{2}$$

where $\pi_1(\boldsymbol{\beta})$ and $\pi_2(\boldsymbol{\theta})$ are suitable parametric priors on the fixed effects $\boldsymbol{\beta}$ and on the hyperparameters $\boldsymbol{\theta}$, respectively.

Under assumptions (2), the likelihood function uses information in the sample which does not become expressed in a likelihood function corresponding to a GLMM with parametric

random effects, hereafter "parametric GLMM", and this fact improves the fit of the corresponding model. In fact, the likelihood function is given by the sum of the huge number of terms provided in (3), where all possible partitions (clusterings) C of the n observations into c nonempty clusters, $c = 1, \ldots, n$, are automatically considered (Liu, 1996; Lo, 1984):

$$L(\boldsymbol{\beta}, \alpha | \boldsymbol{y}, \boldsymbol{\theta}) = \sum_{c=1}^{n} \sum_{C:|C|=c} \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \alpha^c \prod_{k=1}^{c} \Gamma(n_k) \int f(\boldsymbol{y}_{(k)} | \boldsymbol{\beta}, \phi_k) dG_{0(\boldsymbol{\theta})}(\phi_k), \tag{3}$$

where $\boldsymbol{y} = [y_1, \ldots, y_n]$, n_k $(1 \leq n_k \leq n)$ denotes the number of observations in the k-th cluster $\boldsymbol{y}_{(k)}$,

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha+N)} \alpha^c \prod_{k=1}^{c} \Gamma(n_k) = \mathsf{Pr}\{n_1, \ldots, n_c | C, c\}, \tag{4}$$

and, finally,

$$f(\boldsymbol{y}_{(k)} | \boldsymbol{\beta}, \phi_k) = \prod_{j \in \text{cluster } k} f(y_j | \mu(\boldsymbol{\beta}, \phi_k)). \tag{5}$$

In such a likelihood, in particular in formula (5), we can observe that the same random effect is assigned to all observations belonging to the same cluster $\boldsymbol{y}_{(k)}$. The presence of clusters implies that:

i. to learn about a given observation, information additional to that provided by covariates explicitly introduced in the model is borrowed from observations included in the same cluster, and this happens for each cluster to which such an observation can be assigned in the context of all possible clusterizations into nonempty clusters of the n observations. Therefore, these clusters include the ones representing patterns of association between observations due to spatial and/or temporal correlation not explicitly modeled, the ones due to omitted explanatory variables, and so on;

ii. as a matter of fact, the results obtained under such a Bayesian semi-parametric GLMM

are obtained under an "average model", i.e. under a model whose likelihood is obtained by averaging over all the likelihoods corresponding to parametric GLMMs with the same fixed effects and random effects $\phi_j \overset{iid}{\sim} G_{0(\boldsymbol{\theta})}$ grouped in all possible ways into c nonempty clusters, $c = 1, .., n$. The weights associated with such parametric likelihoods are given by the probabilities $\mathsf{Pr}\{n_1, \ldots, n_c | C, c\} = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \alpha^c \prod_{k=1}^{c} \Gamma(n_k)$. This means that accounting for uncertainty in the specification of a single distribution function for random effects is roughly equivalent to accounting for uncertainty in the specification of a single parametric GLMM for the response variable. This results in more robust inferences to a distributional assumption that cannot be checked given that random effects are unobservable variables.

3. GLMMS FOR SPECIES RICHNESS OF ROSS SEA MOLLUSCA

The literature on species richness usually focuses on estimation methods (see, e.g., the rich lists of references in Gotelli and Colwell (2010), Dorazio et al. (2011) and Gotelli and Chao (2013)). Contributions on direct modelling of species richness as a function of "suitable explanatory variables" include, for instance, Mac Nally and Fleishman (2004); Steinmann et al. (2009); Rahbek et al. (2007); Andrew et al. (2012). An alternative general simulation model for macroecology is proposed in Gotelli et al. (2009), which contains a brief review of the so-called curve fitting approach that has dominated the contemporary analyses of species richness data and a specific mention (see also references therein) of the technical challenges of spatial autocorrelation, nonlinear responses of species richness to environmental variables, effects of spatial scale and problems with contemporary and historical factors influencing species richness that are likely to interact in complex ways.

Our work differs from this literature in two ways. Firstly, we carry out an exploratory data analysis aimed at discovering basic drivers or proxies for the underlying drivers of species

richness; secondly, we do not aggregate point observations within pre-specified areas or, in general, within grid cells. Vice versa, we consider the values of species richness, y_j, observed at each of the 619 sites and study their association with the set of geographical characteristics of these sites: latitude, longitude, distance from the nearest scientific station and maximum depth. In addition to fixed effects – i.e. to coefficients of such covariates observed without error at all sites – we include a spatial random effect as follows. We overlay a 1° latitude $\times$ 1° longitude grid on the Ross Sea area and identify 112 boxes containing at least one of the 619 available point observations; then we assign the same random effect to all point observations in the same box. The analysis is conducted using a semi-parametric Bayesian GLMMs with box-specific random effects whose complete specification is given below

$$
\begin{aligned}
& y_j|\mu_j \sim \mathsf{Poisson}(\mu_j), \ \forall\, j = 1, \ldots, 619 \\
& \log(\mu_j) = \boldsymbol{x}_j'\boldsymbol{\beta} + \phi_j \\
& \boldsymbol{\beta} \sim \mathsf{N}_p(\mu_\beta, V_\beta) \\
& \phi_j = \boldsymbol{z}_j'\boldsymbol{b} \\
& \text{where the design vector } \boldsymbol{z}_j' \text{ and the vector } \boldsymbol{b} \text{ are such that} \\
& \phi_j = b_i \ \forall\, j \in \text{box}_i, \ i = 1, \ldots, 112 \qquad (6) \\
& b_i|G \overset{iid}{\sim} G \\
& G|\alpha, \mathsf{N}_{(m,\Sigma)} \sim \mathcal{D}(\alpha, \mathsf{N}_{(m,\Sigma)}) \\
& \alpha \sim \mathsf{Gamma}(a_0, b_0) \\
& m \sim \mathsf{N}(\mu_m, V_m) \\
& \Sigma \sim \mathsf{IW}(r, T)
\end{aligned}
$$

where IW denotes an Inverse Wishart distribution with expected value is given by $T^{-1}/(r-2)$.

Consequently all possible forms of association between non-empty boxes are automatically

considered and exploited for inference in a way similar to the one illustrated in sub-section 2.2, point (i.). These forms of association can be interpreted as spatial association patterns between groups of observations included in the non-empty boxes.

For comparison, we also consider models for the species richness of Ross Sea Mollusca that are parametric elaborations of a Poisson regression model enriched in various ways to account for over-dispersion, absence of zeros, excess of 1s and other visible violations of the assumptions underlying the Poisson model (see Figure 2).

[Figure 2 about here.]

Initially, we consider parametric GLMMs with observation-specific random effects.

Given that $y_j \sim \mathsf{Poisson}(\mu_j)$, two different distributions are explored for the random effect ϕ_j in $\log(\mu_j) = \boldsymbol{x}_j'\boldsymbol{\beta} + \phi_j$. First, we assume that $e^{\phi_j} \overset{iid}{\sim}$ Gamma(a, b) with $a = b$, so that $E(e^{\phi_j}) = 1$ and $Var(e^{\phi_j}) = 1/a$. This assumption introduces extra-variability on a different scale as ordinary predictors (Agresti (2013), p.556) and leads to the Negative Binomial regression model (Cameron and Trivedi, 2013; Hilbe, 2007):

$$y_j \sim \mathrm{NB}\Big(a, \frac{a}{a + e^{\boldsymbol{x}_j'\boldsymbol{\beta}}}\Big), \;\; j = 1, \ldots, 619. \tag{7}$$

where $E(y_j) = e^{\boldsymbol{x}_j'\boldsymbol{\beta}}$ and $Var(y_j) = e^{\boldsymbol{x}_j'\boldsymbol{\beta}}(1 + e^{\boldsymbol{x}_j'\boldsymbol{\beta}}/a)$. The parameter a is often referred to as the "clumping parameter" (Anscombe, 1948; Young and Youn, 1998) since count data in ecology are often clumped[‖], producing an expected variance that is greater than the mean. Successively, we introduce extra-variability in the Poisson regression model by assuming

$$\phi_j \overset{iid}{\sim} \; N(0, \sigma^2), \tag{8}$$

i.e. a Lognormal distribution for e^{ϕ_j}, implying $E(y_j) = e^{\boldsymbol{x}_j'\boldsymbol{\beta} + \frac{\sigma^2}{2}}$, $\; j = 1, .., 619$. In this case we denote the distribution of y_j, $j = 1, .., 619$, by PL, $y_j \sim PL(t|\boldsymbol{\beta}, \sigma^2)$, $t = 0, 1, 2, \ldots$.

[‖] If the rate of capture of individuals varies randomly.

Finally, we consider a parametric GLMM including box-specific random effects. Under this model

$$\phi_j = \boldsymbol{z}_j'\boldsymbol{b}; \quad b_i \sim \mathsf{N}(0, V_m); \qquad j = 1, \ldots, 619; \; i = 1, \ldots, 112 \tag{9}$$

where the design vector $\boldsymbol{z}_j'$ and the vector $\boldsymbol{b}$ are such that, if observation j belongs to box$_i$, then $\phi_j = b_i$, so that we have a random intercept specific to each box.

The set of models described above was estimated from the data by different methods. As regards the point estimates reported in Table 2, the models in the first two columns (Poisson and Negative Binomial families) were estimated by the maximum likelihood method, while for the remaining models - more directly comparable to our semi-parametric GLMM - a Bayesian view and Markov chain Monte Carlo (MCMC) methods were adopted. In any case, the comparability of point estimates reported in different columns of Table 2 was guaranteed by assuming independent vague normal priors for β_j, $N(0, 10000)$, $j = 1, \ldots, p$, and a vague Inverse Gamma prior for σ^2. The point estimates in column 5 were obtained as a special case of our Bayesian semi-parametric GLMM whose hyperparameters, in turn, are fixed so as to induce vague priors for $\boldsymbol{\beta}$ and α ($a_0 = 0.0001, b_0 = 0.0002, r = 0.1, T = 1, \mu_\beta = 0, V_\beta = 10000, \mu_m = 0, V_m = 1$).

Columns 2 and 3 of Table 2 show the results corresponding to parametric GLMMs with observation-specific random effects. For the sake of completeness, column 1 also provides the results obtained under a simple Poisson regression model, i.e. the special case of (8) for $\sigma^2 \to 0$. Under the latter model, all estimated coefficients turn out to be significant, but the ratio of the residual deviance to the residual degrees of freedom is much greater than 1, thereby suggesting the introduction of extra-variability. On the contrary, except

for the intercept, most fixed effects estimated under the Negative Binomial and Poisson-Lognormal families are not significant**, and the few significant effects do not have a clear biological interpretation, or, more precisely, they are not inscribed in a coherent biological framework. These problems are even more pronounced when, in order to account for the absence of zeros, the Poisson-Lognormal family is substituted by its zero-truncated version (column 4). This is an adaptation of the Poisson-Lognormal family obtained by truncation, that is by assuming that $y_j \sim PL(t|\boldsymbol{\beta}, \sigma^2)/(1 - PL(0|\boldsymbol{\beta}, \sigma^2))$, $t = 1, 2, ...$, independently for $j = 1, .., 619$. As the complexity of the model increases, almost none of the geographical covariates seem to maintain any explanatory power for the species richness, with the latter model representing the extreme case. A slightly better conclusion is suggested by inspection of the results in the right side of Table 2 (column 5) obtained under the parametric GLMM including box-specific random effects.

[Table 2 about here.]

Finally, Figure 3 describes our main result. It compares the 95% credible intervals (CI) corresponding to the Bayesian semi-parametric hierarchical model (in red) to the ones corresponding to the parametric GLMM with box-specific random effects (in black). All models were implemented with 10000 MCMC iterations after a burn-in of the first 2000, and standard diagnostic tools confirmed the convergence of runs. For the semi-parametric model we made use of MCMC methods for nonconjugate priors. More in details: the algorithm 8 of (Neal, 2000) was considered (for a complete description, see Neal (2000) p. 262), and the Metropolis-Hastings algorithm with an iteratively weighted least squares proposal (Gamerman, 1997) was applied to generate the fully conditional distribution for fixed and random effects (see also West (1985)). For the precision parameter of the DP process we used the method described in Escobar and West (1995).

**Note that they cannot be directly compared because they assume different parametrizations by having different contrasts on their estimates (Lee and Nelder, 2004, p. 222).

[Figure 3 about here.]

When comparing the lengths of credible intervals in the two cases, we observe that the semi-parametric GLMM results in uniformly shorter intervals than those corresponding to a GLMM with normal random effects. Moreover, the values of standard measures of the predictive performance of a model like the Deviance Information Criterion (DIC) and the log pseudo marginal likelihood (LPML) (Geisser and Eddy, 1979) improve respectively from 4953.30 and -2657.58, under the parametric GLMM, to 4938.96 and -2655.77, under the semi-parametric one. This means that the richer random effect model is able to remove more extraneous variability and also confirms the presence of extra information in the data fruitfully exploited by our model. Following Kyung et al. (2010), we take this as evidence that the proposed model captures nonparametric information of interest, which will now be exploited in the interpretation of the fixed effects estimated under this model.

We observe that species richness decreases when depth and distance from the nearest scientific station increase: these negative effects can be reasonably explained in terms of varying availability of food at different depths or changes in sediment texture, for the first coefficient, and in terms of different sampling intensities, which are generally higher in areas closest to research stations, for the second coefficient. As regards latitude, it is usually taken as a general proxy for other environmental gradients, particularly in polar areas. However, for the Ross Sea at least, latitude has been shown to be a rather poor predictor of environmental changes for benthic communities (Cummings et al., 2010). That its credible interval is not bounded away from zero is therefore not surprising. Nonetheless such a result is now also confirmed for Ross Sea Mollusca by a significantly enlarged sample from an enlarged area, given that here, for the first time, it is also based on data from the never re-sampled and never analyzed historical sites. Hence, in a sense, this is a new evidence in favor of an existing conjecture based on indirectly related data provided by an independent, planned surveys. Much more intriguing is the interpretation of the longitude effect. In this case we have a

credible interval bounded away from zero under the semi-parametric GLMM, instead of a "not significant" effect (CI including zero) under the parametric one, see also column 5 of Table 2. This suggests that, in order to explain the increase of species richness of Ross Sea Mollusca with longitude, specific environmental covariates need to be taken into account to understand if the observed variation is related to different water masses and/or benthic food availability. The Ross Sea shelf has a variety of possible causal factors that might be related to longitude (see Smith et al. (2007) for a review of Ross Sea features and the general environmental setting). Therefore such an interesting result warrants further analysis using other taxa from the complete dataset to determine whether it holds true only for molluscs or for other groups as well.

Overall, the inferential results that we have obtained can be interpreted as confirming the ability of a Bayesian GLMM with nonparametric random effects to discount the effects of omitted variables, or, in other words, to take them into account indirectly, by purifying from their confounding effects the estimated effects of manifest variables in the model. In addition, our model seems to be able to mobilize the biodiversity information stored in historical and newly collected presence-only data in a useful way that contributes generating new research questions and, probably, new knowledge.

4. CONCLUSIONS

This article has presented a general method for the joint analysis of historical and newly collected presence-only data, and illustrated its potential by means of an application to data on Ross Sea Mollusca. The analysis of data from over a century's worth of scientific expeditions presents particular challenges related to a variety of sampling issues that ultimately lead to incomplete data and meta-data associated with a vast number of

sampling events. This results in two types of observational errors: sampling bias and non-detection. Sophisticated models must therefore be devised to overcome these challenges. In particular, we have focused on a semi-parametric GLMM with DP random effects; this is considered preferable to the standard approaches with parametric GLMMs also explored. The nonparametric nature of DP random effects draws latent information from the data, leading to more accurate inferences which, in turn, provide new answers to old research questions and new questions soliciting further research. On the one hand, we see from data ranging from 1899 to 2004 what has been highlighted in recent studies based on indirectly related data collected in planned surveys, like the lack of significativity of the latitude in the Ross Sea region. On the other, the significant coefficient of longitude indicates specific directions in which to try to discover new variables with a direct impact on the mollusc species richness.

Ultimately, the Bayesian semi-parametric approach that we have proposed in this article emerges as a plausible general framework within which future biodiversity studies will be able to embed presence-only data and, in general, problematic data, without losing a broad comprehensive view, in order to try to produce new knowledge and information.

ACKNOWLEDGEMENTS

We are deeply indebted to the review team for their valuable, substantial suggestions.

REFERENCES

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

Ainley, D.G. (2002). The Ross Sea, Antarctica, where all ecosystem processes still remain for study, but maybe not for long. *Marine Ornithology* **30**(2), 55–62.

Ainley, D.G. (2010). A history of the exploitation of the Ross Sea, Antarctica. *Polar Record* **46**(03), 233–243.

Andrew, M.E., Wulder, M.A., Coops, N.C. and Baillargeon, G. (2012). Beta-diversity gradients of butterflies along productivity axes. *Global Ecology and Biogeography*, **21**(3), 352-364.

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

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

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

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.

Chakraborty, A. (2010), *Modeling point patterns, measurement error and abundance for exploring species distributions*. PhD Thesis. Duke University, Durham.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., and Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **60**(5), 757–776.

Cummings, V.J., Thrush, S.F., Chiantore, M., Hewitt, J.E. and Cattaneo-Vietti R. (2010). Macrobenthic communities of the north-western Ross Sea shelf: links to depth, sediment characteristics and latitude. *Antarctic Science*, **22**(6),793–804.

Dorazio, R. M. (2012). Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, **68**(4), 1303–1312.

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, **23**(12), 1472–1484.

Dorazio, R.M., Gotelli, N.J. and Ellison, A.M. (2011). Modern Methods of Estimating Biodiversity from Presence-Absence Surveys. In Grillo O. and Venora G. (eds.) *Biodiversity Loss in a Changing Planet*. InTech.

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.

Escobar, M.V. and West, M.(1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

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

Fithian, W. and Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, **7**(4), 1917–1939.

Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, **6**(4),424–438.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, **7**, 57–68.

Gill, J. and Casella, G. (2009). Nonparametric priors for Ordinal Bayesians Social Science Models: Specification and Estimation. *Journal of the American Statistical Association*, **104**, 453–464.

Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. *Journal of Am. Statist. Assoc.*, **76**, 153–160.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013). *Bayesian Data Analysis*. CRC Press. Boca Raton.

Ghoshal, S. (2010). *The Dirichlet Process, related priors and posterior asymptotics*. In: Hjort, N.L., Holmes, C., Mueller, P. and Walker, S.G. (eds.).*Bayesian Nonparametrics*. Cambridge University Press, Cambridge, UK.

Ghoshal, S. and van der Vaart, A. (2016). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, Cambridge, UK.

Giraud, C., Calenge, C., Coron, C., and Julliard, R. (2016). Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics*, **72**(2), 649–658.

Gotelli, N.J., Anderson, M.J., Arita, H.T., Chao, A., Colwell, R.K., Connolly, S.R., Currie, D.J., Dunn, R.R., Graves, G.R., Green, J.L., Grytnes, J.A., Jiang, Y.H., Jetz, W., Lyons, S.K., McCain, C.M., Magurran, A.E., Rahbek, C., Rangel, T., Soberon, J., Webb, C.O. and Willig,M.R. (2009). Patterns and causes of species richness: a general simulation model for macroecology. *Ecology Letters*, **12**, 873–886.

Gotelli, N.J. and Chao, A. (2013). *Measuring and estimating species richness, species diversity, and biotic similarity from sampling data*. In Levin S.A. (ed.). *Encyclopedia of Biodiversity*, 2nd edition. Volume 5. Academic Press, Waltham, MA.

Gotelli, N.J. and Colwell, R.K. (2010). *Estimating species richness*. pp. 39-54 in: Magurran, A.E. and McGill, B.J. (eds.). *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press,

Oxford.

Griffiths, H.J., Linse, K. and Crame, A.J. (2003). SOMBASE – Southern Ocean Mollusc Database: a tool for biogeographic analysis in diversity and ecology. *Organisms Diversity & Evolution*, **3**(3), 207–213.

Hefley, T. J. and Hooten, M. B. (2016). Hierarchical species distribution models. *Current Landscape Ecology Reports*, **1**(2), 87–97.

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

Hjort, N.L., Holmes, C., Mueller, P. and Walker, S.G. (2010). *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, UK.

Kyung, M., Gill, J. and Casella, G. (2010). Estimation in Dirichlet random effects models. *Annals of Statistics*, **28**, 979–1009.

Kyung, M., Gill, J. and Casella, G. (2011). New findings from terrorism data: Dirichlet process random-effects models for latent groups. *Journal of the Royal Statistical Society, series C, Applied Statistics*, **60**, 701–721.

Lee, Y. and Nelder, J.A. (2004). Conditional and marginal models: another view. *Statistical Science*, **2**, 219–238.

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

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

MacEachern, S. N. and Muller, P. (1998). Estimating mixture of Dirichlet Process Models. *Journal of Computational and Graphical Statistics*, **7**(2), 223–338.

Mac Nally, R. and Fleishman, E. (2004). A successful predictive model of species richness based on indicator species. *Conservation Biology*, **18**, 646–654.

Mueller, P., Quintana, M.A., Jara, A. and Hanson, J. (2015). *Bayesian Nonparametric Data Analysis*. Springer.

Mueller, P., and Quintana, F.A. (2004). Nonparametric Bayesian Data Analysis *Statistical Science*, **19**, 95–110.

Mueller, P. and Rodriguez, F.A. (2004). Nonparametric Bayesian Inference *IMS-CBMS Lecture Notes*. **IMS**, Beachwood.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Pearce, J. L. and Boyce, M. S. (2006), Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**, 405–412.

Phadia, E.G.(2013). *Prior processes and their applications*. Springer, New York.

Rahbek, C., Gotelli, N.J., Colwell, R.K., Entsminger, G.L., Rangel, T.F.L.V.B., Graves, G.R. (2007). Predicting continental-scale patterns of bird species richness with spatially explicit models. *Proc. R. Soc. B*, **274**, 165–174.

Renner, I. W. and Warton, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**(1), 274–281.

Royle, J.A. and Dorazio, R.M. (2008). *Hierarchical Modeling and Inference in Ecology*. Academic Press, London.

Schiaparelli, S., Lorz, A.N. and Cattaneo-Vietti, R. (2006). Diversity and distribution of mollusc assemblages on the Victoria Land coast and the Balleny Islands, Ross Sea, Antarctica. *Antarctic Science*, **18**(4), 615–631.

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.

Smith, W.O., Ainley, D.G. and Cattaneo-Vietti, R. (2007). Trophic interactions within the Ross Sea continental shelf ecosystem. *Philosophical Transaction of the Royal Society. B*, **362**, 95–111.

Steinmann, K., Linder, H.P., Zimmermann, N.E. (2009), Modelling plant species richness using functional groups. *Ecological Modelling*, **220**, 962–967.

Tobler, M.W., Hartley, A.Z., Carrillo-Percastegui, S.E., Powell, G.V.N. (2015). Spatiotemporal hierarchical modelling of species richness and occupancy using camera trap data. *Journal Applied Ecology*, **52**, 413–421.

Young, L.J and Jerry Youn, J. (1998). *Statistical Ecology*. Springer.

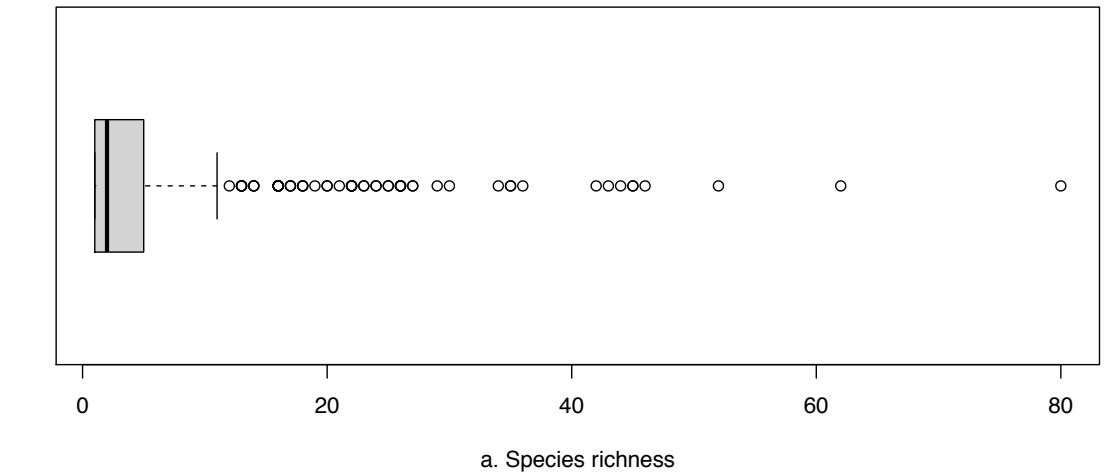
Walker, S. G. (2013). Bayesian Nonparametrics. In: Damien, P., Dellaportas, P., Polson, P. and Stephen, D.A. (eds.). *Bayesian Theory and Applications*, Oxford University Press, Oxford, 249–270.

Warton, D. I. and Shepherd, L. C. (2010). Poisson point process models solve the pseudo-absence problem for presence-only data in ecology. *The Annals of Applied Statistics*, **4**(3), 1383–1402.

West, M. (1985). Generalized linear models: outlier accomodation, scale parameter and prior distributions. In: Bernardo et al (eds.). *Bayesian Statistics 2*, North Holland, Amsterdam, 531-558.

APPENDIX

FIGURES



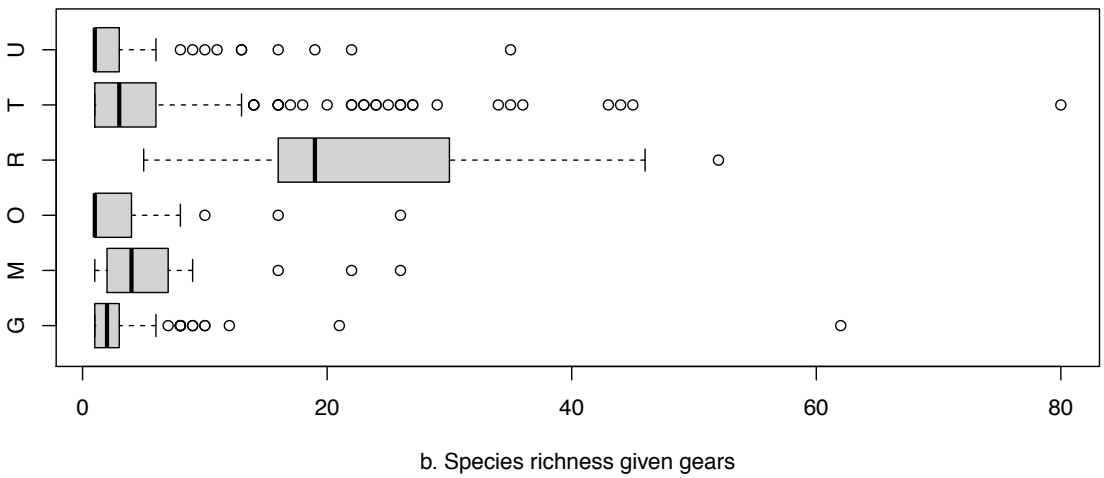


Figure 1. Box plot of the species richness (a.) and of the species richness conditionally on gears (b.). The latter (number of observations associated with each gear in parentheses) are: G=Grab (196), T=Towed gears (242), R=Rauschert (18), U=Unknown (115), M=Multiple (22) and O=Other (26).

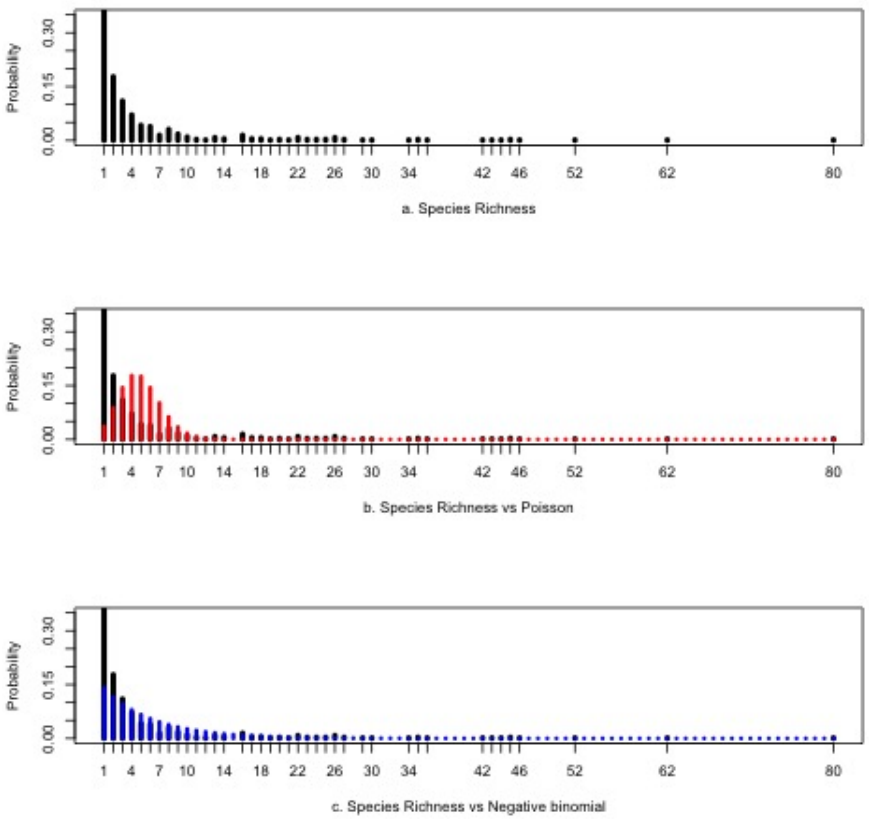


Figure 2. Comparison of the observed species richness, represented in plot **a**, with $y \sim \text{Poisson}(4.932)$, in plot **b**, and with $y \sim \text{NB}(0.977, 0.165)$, in plot **c**. The parameters in the Poisson and Negative Binomial distributions are estimated from the data.

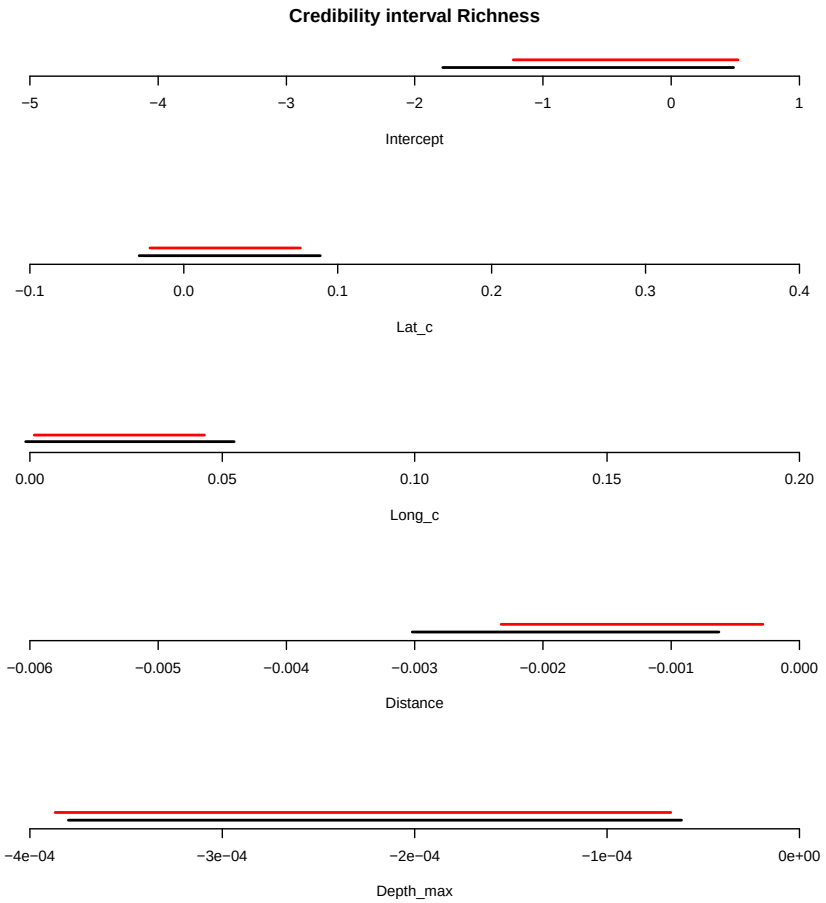


Figure 3. 95% Credible intervals under the Bayesian semi-parametric (in red) and the parametric (in black) GLMM with box-specific random effects.

TABLES

Table 1. Data matrix for the Ross Sea Mollusca

Species	**Sampling Stations**						**Incidence based frequency of species (y_i)**
	1	2	3	4	...	619	
Acirsa antarctica			✓		...		12
Acteon antarcticus					...		3
Adacnarca limopsoides					...		8
Adacnarca nitens	✓				...		79
Adamussium colbecki	✓				...		80
Admete haini					...		2
...	...	...	...	...	...	...	...
Yoldiella sabrina					...		11
Species Richenss (y_j)	26	6	3	4	...	4	

Table 2. Fixed effect estimated under a parametric GLMM with observation-specific random effects and family: Poisson, Negative Binomial, Poisson-Lognormal, zero-truncated Poisson-Lognormal (columns 1-4, left side), and with box-specific random effects (columns 5, right side). Significant estimates are denoted by *. Bayesian point estimates are denoted by * when the corresponding 95% credible intervals are bounded away from zero.

Variable name	Poisson	Negative Binomial	Poisson Lognormal	Zero truncated Poisson Lognormal	GLMM with Gaussian box-specific random effects
Intercept	-4.13656*	-5.58741*	-3.47300*	-5.79178*	-0.41728
Latitude	-0.02378*	-0.02665	-0.01962	-0.02622	-0.03123
Longitude	0.02428*	0.03167*	0.01854*	0.02706	0.02689
Distance	-0.00047*	-0.00053	-0.00028	-0.00045	-0.00178*
Max Depth	-0.00027*	-0.00032*	-0.00020	-0.00027	-0.00024*