Bayesian delta normal spatio-temporal model for zero inflated biological data

Simona Arcuti[1,*], Alessio Pollice[1], Nunziata Ribecco[1] and Angelo Tursi[2]

[1] *Dipartimento di Scienze economiche e metodi matematici, Università degli studi di Bari Aldo Moro, Largo Abbazia Santa scolastica 53, 70124 Bari, ITALY; simona.arcuti@gmail.com, alessio.pollice@uniba.it, nunziata.ribecco@uniba.it*
[2] *Dipartimento di Biologia, Università degli studi di Bari Aldo Moro, Via E. Orabona 4, 70125 Bari, ITALY; angelo.tursi@uniba.it*
[*] *Corresponding author*

Abstract. *The changes in the population density of a demersal species of the North-Western Ionian Sea over spatial and temporal scales and in response to anthropogenic and environmental factors is evaluated. Biological data were collected during trawl surveys carried out from 1995 to 2006 as part of the international project MEDITS (MEDiterranean International Trawl Surveys). The density (N/km^2) index was computed for a number of hauls identified by time, depth, geographic coordinates and geographical sector. A bayesian delta normal model was estimated to evaluate the spatio-temporal changes in the density of a particular species of crustacean known as deep-water rose shrimp, Parapenaeus longirostris. Model estimation was implemented using the JAGS software, specifying the zero-inflated likelihood of the population density by the "zeros trick" method. The study highlights nonlinear spatial and temporal effects and the significant influence of the Winter North Atlantic Oscillations index on the deep-water rose shrimp density distribution.*

Keywords. *Zero-inflation; Bayesian additive model; JAGS; Penalized spline regression; Biological population dynamics*

1 Introduction

The presence of high counts of zeroes is a common feature in the analysis of ecological and biological data. Zero-inflated distributions can result from various mechanisms such as biological processes, sampling limitations, observer effects, etc. Such distributions are involved in the study of the dynamics of many Mediterranean species, due to their adaptation to variable environmental conditions. In this work we focus on the deep-water rose shrimp, Parapenaeus longirostris (Lucas, 1846), widespread throughout the whole Mediterranean Sea at depths between 20 and 700 m. Aspects of the distribution and population

biology of this species are reported in [3] and [1]. Here we propose a Bayesian delta normal additive model to analyze the spatio-temporal distribution of the density of the deep-water rose shrimp in the North-Western Ionian Sea for the period 1995-2010.

2 Data collection

Biological data were collected during 12 experimental trawl surveys conducted from 1995 to 2006 in the North-Western Ionian Sea as part of the international MEDITS (MEDiterranean International Trawl Surveys) program. Here we analyze samples from 247 hauls carried out during day-light hours in the spring season (May- June). The study area has a total surface of 16.350 km^2, depths between 10 and 800 m and is divided into three geographical sectors: Apulia; North-Calabria and South-Calabria. A random-stratified sampling design was adopted, with allocation of hauls proportional to the area of each depth range within geographical sectors.

3 Spatio-temporal model

In this work additive models are used in order to capture the nonlinear influence of some predictors on the response variable, including spatial and temporal effects. Model estimation is implemented in the Bayesian framework, calling the JAGS software [6] from `R` using the `rjags` [8] and `R2jags` [9] packages. JAGS is a general program for the estimation of Bayesian models using Markov chain Monte Carlo (McMC) methods. Gibbs sampling generates a sequence of samples from the joint distribution of multiple random variables using their full conditional distributions. When full conditionals are non-log-concave JAGS utilizes the Adaptive Rejection Metropolis sampler [4].

We assume that response variable follows a *delta normal* distribution, i.e. that the normalized density of the deep water rose shrimp at the *i*-th haul has the following distribution:

$$Y_i = \begin{cases} 0 & \text{with probability} \quad \psi_i \\ y_i & \text{with probability density} \quad (1-\psi_i) f(y_i;\theta_i) \end{cases} \tag{1}$$

where $f(y_i|\theta_i)$ is the probability density function of a generic normal distribution. Let $W \sim Bernoulli(1-\psi_i)$, then the joint probability density function of n observed values of the response variable is given by:

$$f(y,w|\psi,\sigma,\mu) = \prod_{i=1}^{n} (1-\psi_i)^{w_i} \psi_i^{1-w_i} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(y_i-\mu_i)^2}{2\sigma^2} \right) \right]^{w_i}. \tag{2}$$

Here, the two parameter vectors $\mu = \mu_1, \ldots, \mu_n$ and $\psi = \psi_1, \ldots, \psi_n$ are specified as follows:

$$E(y) = g(\mu) = \dot{X}\beta + \sum_{j=1}^{k} f_j(X_j) \qquad\qquad logit(\eta) = \dot{X}^*\beta^* + \sum_{j=1}^{k^*} f_j^*(X_j^*) \tag{3}$$

where $g(\cdot)$ is the identity link function, $\eta = 1-\psi$, $\dot{X}$ and $\dot{X}^*$ are the design matrices for the linear effects, β and β^* are linear coefficients to be estimated, X_j and X_j^* are the explanatory variables corresponding to the j-th non-linear effect and f_j and f_j^* are smooth functions to be estimated. Following [2] penalized spline regression by low-rank thin-plate splines was used for smoothing. The non-standard likelihood in (2) and (3) was implemented using the so-called "*zeros trick*" [5]: say $l_i = log f(y_i|\theta)$, then the model likelihood can be written as $\prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} \exp(l_i)$, i.e. as the product of the densities of n Poisson distributed pseudo-random variables with mean equal to $-l_i$ and all observed values equal to zero.

4 Results

In order to obtain more regular distributions, normalization by the fourth root was considered for the density data. A backward elimination approach was used to select a final model minimizing the deviance information criterion (DIC) and including only those covariates with coefficients having strictly positive or negative 95% credibility intervals. In each step of this procedure, models were preliminarily estimated considering 2 chains and 5000 iterations with 1000 iterations as burn-in and thinning by 30. Chain convergence was checked by visual inspection of the trace plots. The resulting selected model has the following expression:

$$E(y) = Annual\ MO + Temperature + f(log, lat) + f(year)$$

$$logit(\eta) = Intercept + f^*(log, lat) + f^*(year)$$

Bayesian model estimate was first obtained considering 2 chains and 20000 iterations with 300 iterations as burn-in and thinning by 150. Later, chains were initialized considering the previous estimates as initial values and this procedure was repeated until convergence. In order to check chain convergence we used the diagnostic procedures implemented in the `coda R` package [7], specifically: trace plots inspection, autocorrelation plots, Gelman-Rubin and Geweke tests.
Table (1) summarizes the parameter estimates of the delta normal model for the density of *Parapenaeus longirostris*. The annual Mediterranean Oscillation index and the sea surface temperature show a positive relationship with the density, corresponding to hydrologic conditions known to be favorable for the species. Semi-parametric smooth terms are estimated considering splines with 25 and 3 knots for the spatial and temporal effect respectively. The spatial effect on the density of *Parapenaeus longirostris* over the study area is shown in Fig. 1 (left). Yellow coloured areas show stronger spatial effect corresponding to higher values of the density of *Parapenaeus longirostris*. In general, we can report a global decrease of the spatial effect with depth. This result is consistent with the known inverse relation between the density and depth. The temporal effect on the density of the deep-water rose shrimp (Fig. 1, center) is characterized by smaller values in 1999-2000, corresponding to the presence of specific hydrological conditions due to the eastern Mediterranean Transient that affected the Ionian Sea at the end of the '80s. In the last plot in Fig. 1 we compare the model estimates of the zero-probabilities ψ_i for the zero (right) and non-zero (left) observed values.

Linear terms	*Mean (sd)*
Annual MO	7.12(1.34)
Sea Temperature	0.19(0.05)

Table 1: Bayesian model estimates of linear effects on the normalized density of P. logirostris. Standard errors in brackets.

5 Conclusions

We introduce a Bayesian approach for zero-inflated continuous biological data, as a valid alternative to standard non-Bayesian methods. Two major advantages of the Bayesian approach are its ease of interpretation and flexibility to use an arbitrarily complex likelihood function. This is particularly relevant for zero-inflated continuous data, a special case of a mixture structure that is a natural candidate for Bayesian

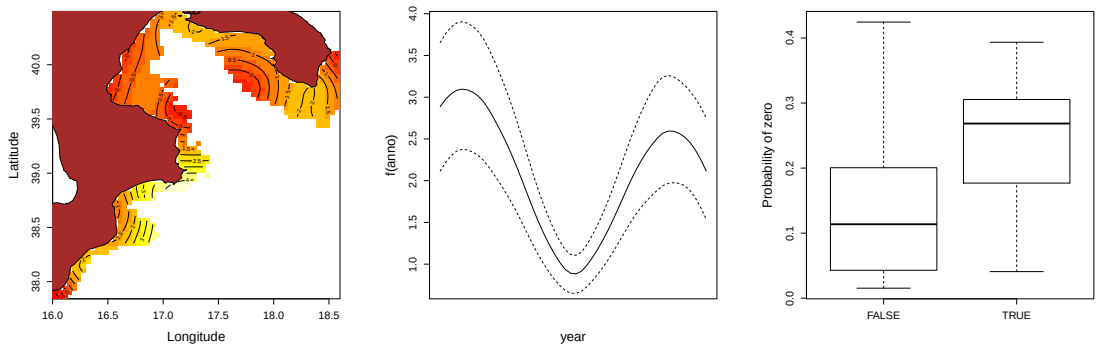

Figure 1: Spatial (left) and temporal (center) smooth components of the model fitted to the *P. longirostris* normalized density data. Boxplots (right) of the estimated zero-probabilities for the zero (TRUE) and non-zero (FALSE) data.

analysis. Here we propose a Bayesian hierarchical delta-normal model to allow easily handling of zero-inflated data using a binary variable for zeroes and a continuous variable for positive measurements. Bayesian analysis offers flexible posterior checks that can be informative about the overall model fit as well as on specific aspects of the data as how well the model fits the zero values. The recent increasing availability of powerful user-friendly software (e.g. JAGS and the others products of the BUGS family) opens the door to easy implementation of estimation, inference, and prediction in the Bayesian framework.

References

[1] Abelló P., Abella A., Adamidou A., Jukic-Peladic S., Maiorano P., Spedicato M.T. (2002). Geographical patterns in abundance and population structure of Nephrops norvegicus and Parapenaeus longirostris (Crustacea: Decapoda) along the European Mediterranean coasts. *Scientia Marina*, **66**(2), pp. 125-141.

[2] Crainiceanu C.M., Ruppert D., Wand M.P. (2005). Bayesian Analysis for Penalized Spline Regression Using WinBUGS. *Journal of Statistical Software*, **14**(14).

[3] D'Onghia G., Maiorano P., Matarrese A., Tursi A. (1998). Distribution, Biology, and Population dynamics of Aristaeomorpha foliacea (Risso, 1827) (Decapoda, Natantia, Aristeidae) in the North-Western Ionian Sea (Mediterranean Sea). *Crustaceana*, **71**(5), pp. 518-544.

[4] Gilks W.R., Wild P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics* **41**(2), 337-348.

[5] Lunn D., Jackson C., Best N., Thomas A., Spiegelhalter D. (2013). The BUGS Book - A Practical Introduction to Bayesian Analysis. CRC Press, Taylor & Francis Group.

[6] Plummer M.(2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

[7] Plummer M.(2008). coda: Output Analysis and Diagnostics for MCMC in R. *R package version 0.13-3.*

[8] Plummer M.(2013). rjags: Bayesian graphical models using MCMC. *R package version 3-10.*

[9] Yu-Sung S., Masanao Yajima J. (2012). R2jags: A Package for Running jags from R. *R package version 0.03-08.*