Suppl2

Data analysis for the Generalized Linear Models

The Generalized Linear Models (GLM) were implemented to explain distribution patterns (density and mean size) according to environmental factos, considered two response variables: 1) density and 2) mean size, both *per* replicate. Mean sizes of transects had a positive tail and, for this response variable, the analyses (partial and multiple regressions) were ran with a gamma distribution and a logarithmic link function. Density presented a zero-inflated distribution, typical of fisheries data, with unbalanced CPUE records depicting a point-mass at zero, a positive tail and non-negative values (Figure S1).

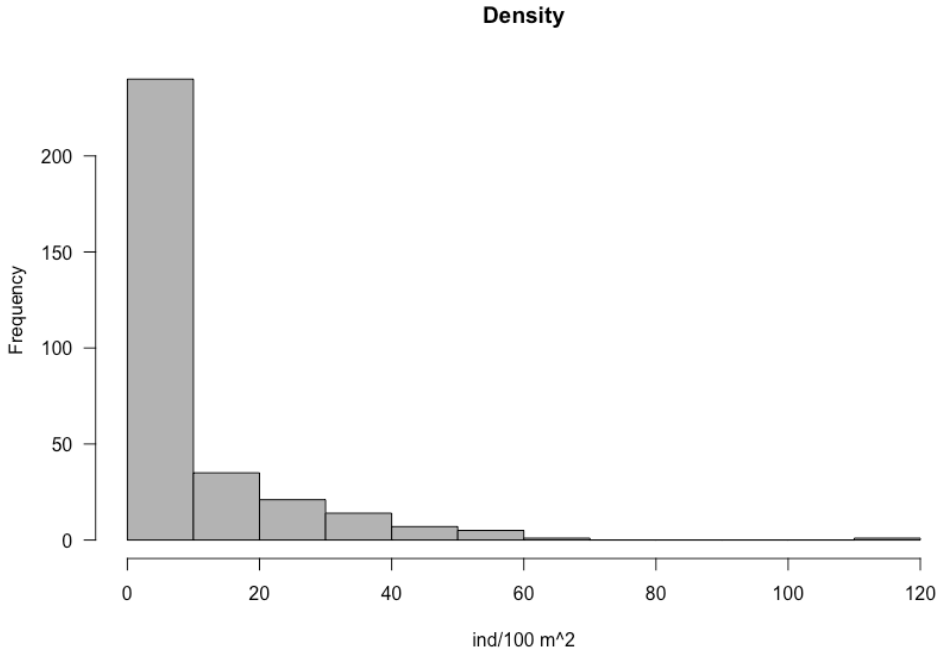

Figure S1. Data distribution of two response variables with a point-mass at zero: density and proportion of individuals on sand substrate, considering the two sample substrates (sand and rock)

For this response variable, a Tweedie distribution was used, which can deal with excessive zeroes (Tweedie, 1984; Shono, 2008). The Tweedie distribution is part of the exponential family of distributions and is defined by a mean (μ) and a variance ($\varphi\mu^p$), in which φ is the dispersion parameter and p is an index parameter (Tweedie, 1984). It can express the Poisson, Gamma or inverse Gaussian distributions if the power-parameter (p) is adjusted to 1, 2, or 3, respectively (Shono, 2008; Ma et al., 2018). For the present case, using abundance data based on non-negative values with a probability mass at zero and highly right-skewed, the power-parameter must be greater than 1 and less than 2 (Wood et al., 2016). The R package mgcv

implements the Tweedie family directly into a *gam()* function that was implemented without smooth terms, via the maximum likelihood method and with a log-link function. For mean size a GLM function was implemented with the same predictor variables, using a Gamma distribution and a log-link function. For both response variables, after fitting the full models, including all variables, a model selection was conducted based on a set of functions from the package MuMIn (Bartoń, 2020): i) run all possible models with the different combinations of predictor variables via *dredge()* function ; ii) rank models according to the corresponding AIC values (Akaike Information Criterion); iii) models meeting the condition $\Delta AIC<2$ were considered for further inference (Burnham and Anderson, 2002). At this point, either one model is selected, with a ΔAIC higher than two regarding the model of the following position or several models are considered to have similar explanation power; iv) for the latter case model averaged inferences via *model.avg()* provided the model-averaged coefficients, for a full and a conditional average. The full model coefficients set terms to zero when they are not included in a given model while averaging, whereas the conditional coefficients ignore the predictors whenever they are not included in a model and only considers them in the models where they are represented. Thus, the full model coefficients are more conservative (Burnham and Anderson, 2002; Nakagawa and Freckleton, 2011). However, this does not mean that the non-significant variables in the full model coefficient output are irrelevant. Although both coefficients often render similar results, there may be relevant variables in the conditional model that are below the nominal significance level in the full model. This means that those particular variables have a lower influence in explaining the response variation. Thus, model averaging makes interpretation using p-values to test the significance of a particular variable more difficult (Grueber et al., 2011). Hence, when full and conditional coefficient outputs show different significant predictors, two additional methods of validation were performed: a) considering the sum of weights of the average model, with the *importance()* function of the MuMIn R package, which vary according to the Akaike weights a variable sums up from each selected model, indicating how many models the variable was considered relevant in explaining the variance of the response; together with the b) confidence intervals of each variable, depending on its range and value, *i.e.* a variable whose confidence interval includes zero can be deemed less informative, whereas those not including zero can be said to have a noticeable effect on the response variable (Grueber et al., 2011). Significant interactions were assessed to determine its type (additive, synergistic or antagonistic) and behaviour (how one predictor responds along the gradient of the other). This evaluation was done considering the signs of the coefficients for the interaction and individual variables and with surface plots. The latter

are two-dimensional contour plots showing the fitted response values (isolines and colour gradient) against a surface defined by both variables of the interaction (Feld et al., 2016). These were plotted using the *curve3d()* function from the emdbook R package (Bolker, 2008).

References - Suppl2

Bartoń, K. (2020). MuMIn: Multi-Model Inference. *R Packag. version 1.43.17.* Available at: https://cran.r-project.org/package=MuMIn [Accessed January 20, 2001].

Bolker, B. M. (2008). *Ecological Models and Data in R*. Princeton, NJ: Princeton University Press doi:10.2307/j.ctvcm4g37.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*. 2nd ed. , eds. K. P. Burnham and D. R. Anderson New York, NY: Springer New York doi:10.1007/b97636.

Feld, C. K., Segurado, P., and Gutiérrez-Cánovas, C. (2016). Analysing the impact of multiple stressors in aquatic biomonitoring data: A 'cookbook' with applications in R. *Sci. Total Environ.* 573, 1320–1339. doi:10.1016/j.scitotenv.2016.06.243.

Grueber, C. E., Nakagawa, S., Laws, R. J., and Jamieson, I. G. (2011). Multimodel inference in ecology and evolution: challenges and solutions. *J. Evol. Biol.* 24, 699–711. doi:10.1111/j.1420-9101.2010.02210.x.

Ma, R., Yan, G., and Hasan, M. T. (2018). Tweedie family of generalized linear models with distribution-free random effects for skewed longitudinal data. *Stat. Med.* 37, 3519–3532. doi:10.1002/sim.7841.

Nakagawa, S., and Freckleton, R. P. (2011). Model averaging, missing data and multiple imputation: a case study for behavioural ecology. *Behav. Ecol. Sociobiol.* 65, 103–116. doi:10.1007/s00265-010-1044-7.

Shono, H. (2008). Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fish. Res.* 93, 154–162. doi:10.1016/j.fishres.2008.03.006.

Tweedie, M. C. K. (1984). "An index which distinguishes between some important exponential families," in *Statistics: Applications and New Directions*, eds. J. K. Ghosh and J. Roy (Calcutta: Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, Indian Statistical Institute), 579–604.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *J. Am. Stat. Assoc.* 111, 1548–1563. doi:10.1080/01621459.2016.1180986.