Modeling of compositional data: a multilevel approach to benthic cover Abrolhos bank.

Pamela M. Chiroque-Solano

Multidisciplinary Institute, Federal Rural University of Rio de Janeiro and
Institute of Biology and SAGE-COPPE, Federal University of Rio de Janeiro, Brazil.

13th of October, 2021

Framework

Compositional Data

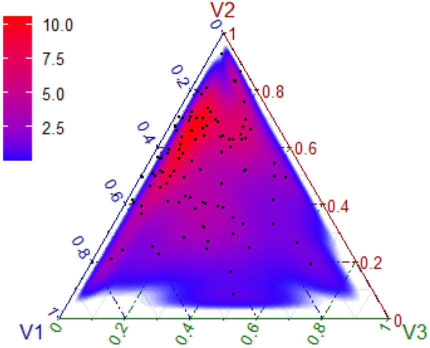
- Multivariate regression with constrained response.
- Challenge:
    - Unbalanced;
    - Lot of missing data;
    - Identificability issues

Objectives

- To model the variability effects including a hierarchical structure;
- To achieve flexibility with the proposed model, so that it can be useful in many settings.

Objective: To study the variability of the process

Case 1: $\log \phi = 1$



Case 3: $\log \phi = -1$

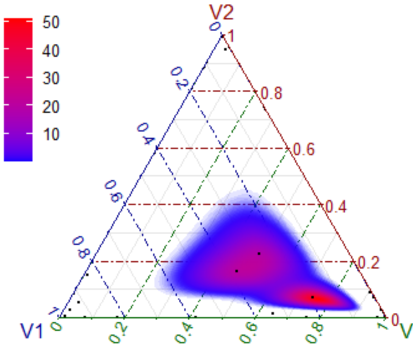


Figure: Consider a three-components (compositional data). The first simplex contains the information for high entropy (case 1), and low entropy (case 3).

Proposal

Filtered information through the decomposition

- of the Dirichlet distribution parameter into two components:
  - level and;
  - precision.
- This decomposition allows us to obtain a flexible proposal.

Notation (Maier, 2014)

Observations

- $y_{ic} \in (0,1)$: The proportion of coverage at observation i of component c.
- $\sum_{c=1}^{C} y_{ic} = 1$: Constraint

Assumptions

- $\boldsymbol{Y} \sim \mathcal{D}(\alpha)$ on $C > 2$-dimensional hyperplane or closed simplex $\mathbb{T}_C(1)$. $\alpha_c > 0$.

Model

Maier (2014) and Holger (2018)

Filtered information through the decomposition of $\boldsymbol{\alpha}$

- $\mathbf{Y}_l \sim D(\mu_l, \phi_l)$ with parameter $\alpha_{cl} = \mu_{cl}\phi_l$
- μ_{cl} : level term
- ϕ_l : precision term

Reference component: $c^\star$

- Alternative parametrization: $c^\star$ should be chosen.
- Stochastic representation for Dirichlet random vector

Sharing information equation

$$
\begin{aligned}
\beta_{cl} &= \beta_c + \epsilon_{\beta_l}, \quad \epsilon_{\beta_l} \sim \mathcal{N}(0, V_\beta) \\
\theta_l &= \theta + \epsilon_{\theta_l}, \quad \epsilon_{\theta_l} \sim \mathcal{N}(0, V_\theta)
\end{aligned}
$$

Inference procedure

Let $\boldsymbol{\Theta} = (\beta, \phi)$ be the vector of parameters

Proper independent prior distribution for the parametric vector $\boldsymbol{\Theta}$ are Normal with zero mean and precision $1/K$ for all effects of the model.

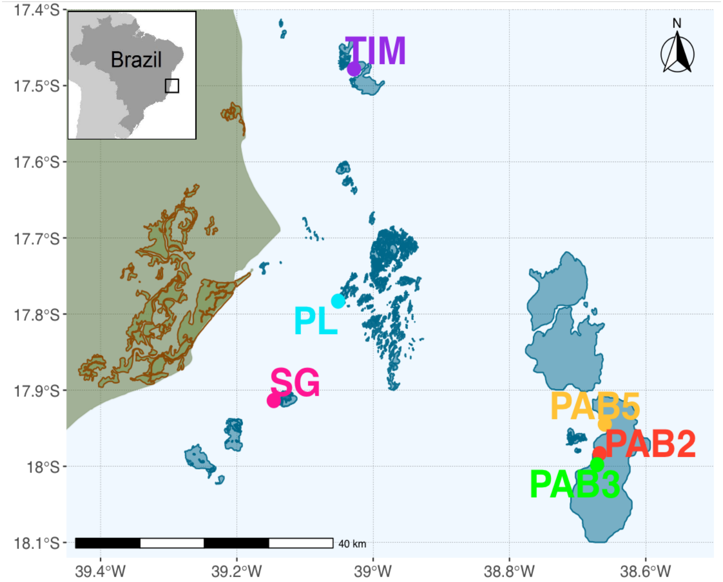
The joint posterior distribution does not have a known closed form

$$\pi(\boldsymbol{\Theta} \mid \mathbf{y}) \propto L(\boldsymbol{\Theta} \mid \mathbf{y}) \prod_{l}^{L} \pi(\phi_l) \prod_{c}^{C} \pi(\beta_{cl}) \tag{1}$$

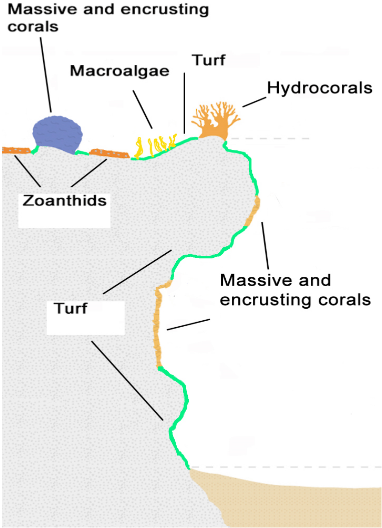
Sampling from the posterior distribution

by Markov chain Monte Carlo (MCMC) via the Stan software.

Application: The Area

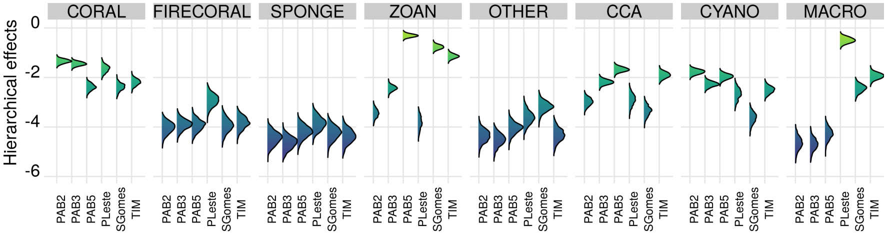


Composition of benthic communities

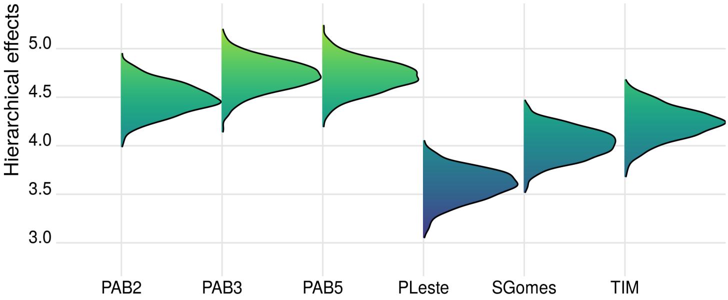


(Teixeira, Chiroque-Solano, et all. 2021)

Results: Posterior density of the β effect for each of the nine components by site and habitat levels.



Results: Posterior density of the θ effect for each of the nine components by site and habitat levels.



The results validate the original hypotheses

Sites near the coast (inshore) are more variable than the offshore sites.

Conclusions and Future Work

Main conclusions

- The proposed model quantifies the heteroscedasticity through precision effects via hierarchical structures by site;
- The method is flexible;
- The reference component has been chosen using objective criteria;
- The proposal allows to obtain adequate predictions.
- This work contributes to the United Nations's Sustainable Development Goal 14 - "Life Under Water".

References

- Gelman, Andrew, and Jennifer Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Analytical Methods for Social Research. Cambridge University Press.
- Holger, and Sennhenn-Reulen. 2018. "Bayesian Regression for a Dirichlet Distributed Response Using Stan."
- Maier, Marco J. 2014. "DirichletReg: Dirichlet Regression for Compositional Data in R." Research Report Series/Department of Statistics and Mathematics 125. Vienna: WU Vienna University of Economics and Business.
- Wang, K., G. Tian, and M Tang. 2011. Dirichlet and Related Distributions: Theory, Methods and Applications. Wiley Series in Probability; Statistics.

Thank you

- Guido A. Moreira, UMinho.
- Marine Biodiversity and Conservation Lab at UFRJ.
- The Fundação Espírito Santense de Tecnologia, FEST.
- The Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro, FAPERJ - E-26/200.016/2021.

Contact

pamela@ufrrj.br