Regression on compositional covariates: assessing substrate suitability for vegetation

Francesca Bruno[1], Fedele Greco[1] and Massimo Ventrucci[1*]

[1] *Department of Statistical Sciences, University of Bologna, Via Belle Arti 41, 40126, Bologna, Italy; francesca.bruno@unibo.it, fedele.greco@unibo.it, massimo.ventrucci@unibo.it*
[*] *Corresponding author*

Abstract. *Investigating the relationship between vegetation cover and substrate typologies is important in habitat conservation and management. We focus on a modern ecological survey, where information regarding vegetation cover are derived from digital ground photos taken at different times. The aim is to estimate the effect of different substrate typologies on vegetation cover (substrate suitability). As it is often the case in ground cover imaging, information on substrate typologies are available as compositional data, e.g., the area proportion occupied by a certain substrate. We develop a novel procedure for managing compositional covariates within a Bayesian hierarchical framework and illustrate it with data from a gypsum outcrop located in the Emilia Romagna region, Italy.*

Keywords. *Compositional covariates; GMRF; Spatio-temporal models; Substrate suitability.*

1 Introduction

Modern ecological surveys are increasingly based on high-resolution spatially referenced data, collected by digital ground photos taken several times during sampling campaigns and fieldworks. Typically, the objective of these studies is to understand the complex relationships between ecological outcomes (e.g. species richness, abundance, vegetation cover) and the characteristics of an habitat (e.g. substrate typologies). In the present work, the focus is on modelling vegetation cover as a function of substrate typologies which is crucial to evaluate *substrate suitability*, i.e., substrates' natural ability to support vegetation. Knowledge on substrate suitability is strategic in predicting future developments and reactions to possible environmental changes.

1.1 Data

A sampling campaign was performed on a gypsum outcrop within the Site of Community Importance *IT4050001 Gessi Bolognesi e Calanchi dell'Abbadessa*, located in the Emilia Romagna Region, Italy. The study area is a 1.5 m^2 regular grid, consisting of $S = 900$ grid cells indexed by $s = 1, ..., 900$. Data were collected from April 2012 to late March 2013, but sampling campaigns were adjourned during the dryness period of August and September, when low vegetation cover is expected, and on January because of snow cover. Overall, data were measured at $T = 9$ unequally spaced times, indexed by $t = 1, ..., 9$. At time t, a ground photo was taken and then processed to produce data on vegetation and substrate ground cover. At grid cell s and time t, information collected over $n = 100$ sub-pixels are available; response data (y_{st}) consist in the number of sub-pixels covered by a plant; covariate data ($\boldsymbol{z}_{st} = \{z_{st,d}\}$, $d = 1, ..., D$) consist in the number of sub-pixels covered by each one of $D = 4$ substrate typologies: moss, litter, soil and bare rock. Within each grid cell, substrate counts sum to n giving information about the composition of the cell in terms of substrates.

1.2 Compositional covariates

The compositional nature of substrate information implies several challenges for statistical modelling. First, if substrate compositional parts $\boldsymbol{z}_{st}$ are used as covariates in a simple linear regression model, then the design matrix would be singular due to the sum-to-n constraint. A recent proposal [3] to address this problem is to use transformations that allow the D parts to be expressed in a real unconstrained space $\mathbb{R}^{D-1}$, rather than in the simplex $\mathbb{S}^D$. Though this approach allows model identifiability, determining the effect associated to increasing each compositional part is generally a non trivial issue. Compositional covariates are expressed in terms of coordinates with respect to a given orthonormal basis. Such a representation, introduced by [2], is called *isometric logratio transformation* (*ilr*):

$$ilr(\boldsymbol{z}_{st})_i = \sqrt{\frac{D-i}{D-i+1}} \log \frac{z_{st,i}}{\sqrt[D-i]{\prod_{j=i+1}^{D} z_{st,j}}}; \qquad i = 1, ..., D-1 \tag{1}$$

In what follows we propose a Bayesian hierarchical modelling framework for estimating compositional covariate effects in a spatio-temporal framework.

2 A modelling proposal for estimating substrate suitability

The relationship between substrate typology and vegetation cover is modelled by a hierarchical Bayesian model where we assume a binomial likelihood and a probit link for the vegetation occupancy probability π_{st},

$$y_{st}|\pi_{st} \sim Bin(\pi_{st}, n) \qquad s = 1, ..., S; \;\; t = 1, ..., T \tag{2}$$

$$\Phi^{-1}(\pi_{st}) = \boldsymbol{x}_{st}^{\top}\boldsymbol{\beta}_t + \alpha_t + \theta_s + \delta_{st} \tag{3}$$

where $\boldsymbol{x}_{st} = ilr(\boldsymbol{z}_{st})$ is a vector of *ilr* coordinates of length $D-1$ expressing substrate compositional information, and $\boldsymbol{\beta}_t = \{\beta_{t,1}, ..., \beta_{t,D-1}\}$ are the associated effects which are assumed as time-varying. Parameters α_t, θ_s and δ_{st} are vectors of temporal, spatial and spatio-temporal structured random effects, respectively, which are modelled using Intrinsic Gaussian Markov Random Fields (IGMRF; [4]). An IGMRF prior is also assumed for modelling $\boldsymbol{\beta}_t$, as our assumption, to be verified, is that substrate suitability changes smoothly over time.

One disadvantage of model (3) is that only the first regression coefficients of vector $\boldsymbol{\beta}_t$, i.e. $\beta_{t,1}$, can be interpreted as the effect of substrate 1 on the vegetation occupancy probability; this is because the first *ilr* coordinate $z_{st,1}$ properly summarizes the comparison between part 1 and all the others; note that all the parts, except for part 1, are included in the denominator of (1) when $i = 1$. In contrast, coefficients $\{\beta_{t,2}, ..., \beta_{t,D-1}\}$ do not express the effect of the other parts in an analogous manner, as the associated *ilr* coordinates for $i = 2, ..., D-1$ do not summarize comparisons of one single part with respect to all the others.

In order to obtain meaningful regression coefficients, we propose to exploit a remarkable feature of the ilr transformation: permutation of the parts in the simplex corresponds to a permutation of the components of the orthogonal basis. Given the vector of permuted parts $\boldsymbol{z}_{st}^{(d)} = \{z_{st,d}, z_{st,1}, ..., z_{st,d-1}, z_{st,d+1}, ..., z_{st,D}\}$, we note that $\boldsymbol{x}_{st}^{(d)} = ilr(\boldsymbol{z}_{st}^{(d)}) = M^{(d)}\boldsymbol{x}_{st}$, where $M^{(d)}$ is a symmetric and orthogonal matrix that allows a change of coordinates from $ilr(\boldsymbol{z}_{st})$ to $ilr(\boldsymbol{z}_{st}^{(d)})$. Matrix $M^{(d)}$ is given by $BB^{(d)\mathsf{T}}$, where B is the orthonormal basis associated to coordinates $ilr(\boldsymbol{z}_{st})$ and $B^{(d)}$ is obtained by switching columns 1 and d of B. Therefore, the suitability of a generic substrate d, evaluated at any time t, is calculated as the first elements of vector $M^{(d)}\boldsymbol{\beta}_t$.

3 Results

The approach described above has a crucial advantage: the contribution of a given compositional part with respect to all others can be properly identified and straightforwardly estimated via Monte Carlo Markov Chain (MCMC). Model fitting was performed via Gibbs sampling using the augmented approach proposed in [1]. In each panel of Figure 1, posterior means for suitability of each substrate are displayed at all observed times (filled circles) and prediction times (black solid line), together with credible bands (grey shadowed areas). It can be shown that substrate suitability estimates sum to 0 by construction: therefore they can be interpreted as relative suitability measures, indicating how much a substrate is more (> 0) or less (< 0) suitable than average ($= 0$). Except from litter, which shows constant average suitability, all other substrate suitabilities changes over time. In particular, moss shows positive relative suitability for vegetation all along the study period. Bare rock is less suitable than average and shows a decreasing pattern over time. The relative suitability of soil is approximately zero at the beginning and increases over time reaching positive suitability in winter.

4 Discussion

A novel approach for estimating the effect of the single part of a compositional covariate on a response variable is developed within a Bayesian framework, through a procedure based on appropriate transformations of the regression coefficients associated to *ilr* transformed covariates. One interesting ecological

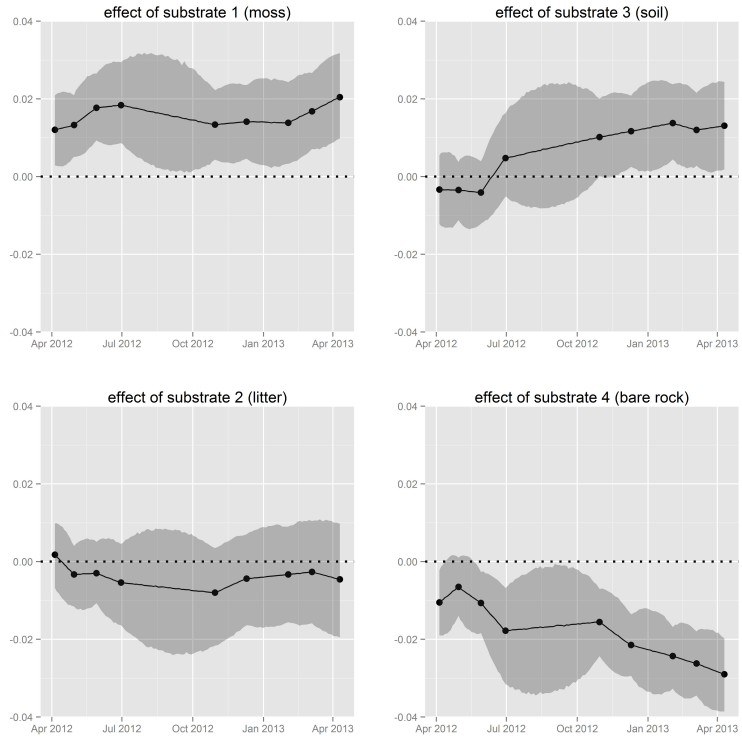

Figure 1: Substrate relative suitability over time

finding of our work is that substrate relative suitability varies over time. A future line of research will focus on extending the proposed model to allow for smooth, rather than simply linear, effects of compositional covariates on ecological responses. Non-linear effects of the proportion of ground covered by a substrate may often be observed on certain ecological responses, such as a biodiversity index, or species richness.

Acknowledgments. The research work underlying this paper was funded by a FIRB 2012 grant (project no. RBFR12URQJ; title: Statistical modeling of environmental phenomena: pollution, meteorology, health and their interactions) for research projects by the Italian Ministry of Education, Universities and Research.

References

[1] Albert J.H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, vol. 88:669-679.

[2] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, vol. 35:279-300.

[3] Hron, K., Filzmoser, P. and Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, vol. 39:1115-1128.

[4] Rue, H. and Held, L. (2005). Gaussian Markov Random Fields. *Chapman and Hall/CRC*.