Received: 24 October 2023 | Revised: 28 December 2023 | Accepted: 18 January 2024

DOI: 10.1002/ecy.4256

STATISTICAL REPORT

ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

Estimation and interpretation problems and solutions when using proportion covariates in linear regression models

Denis Valle[1] | Jeffrey Mintz[2] | Ismael Verrastro Brack[1]

[1]School of Forest, Fisheries, and Geomatics Sciences, University of Florida, Gainesville, Florida, USA

[2]School of Natural Resources and Environment, University of Florida, Gainesville, Florida, USA

Correspondence
Denis Valle
Email: drvalle@ufl.edu

Funding information
NSF Division of Biological Infrastructure (DBI), Grant/Award Number: 2040819; National Institute of Food and Agriculture, Grant/Award Number: 1005163

Handling Editor: Elise F. Zipkin

Abstract

Proportion variables, also known as compositional data, are very common in ecology. Unfortunately, few scientists are aware of how compositional data, when used as covariates, can adversely impact statistical analysis. We describe here how proportion covariates result in multicollinearity and parameter identifiability problems. Using simulated data on bird species richness as a function of land use, we show how these problems manifest when fitting a wide range of models in R, both in a frequentist and Bayesian framework. In particular, we show that similar models can often generate substantially different parameter estimates, leading to very different conclusions. Dropping a covariate or the intercept from the model can solve the multicollinearity and parameter identifiability problems. Unfortunately, these solutions do not fix the inherent challenges associated with interpreting parameter estimates. To this end, we propose focusing the interpretation on the difference of slope parameters to avoid the inherent unidentifiability of individual parameters. We also propose conditional plots with two *x*-axes and marginal plots as visualization techniques that can help users better interpret their modeling results. We illustrate these problems and proposed solutions using empirical data from the North American Breeding Bird Survey. The practical and straightforward approaches suggested in this article will help the fitting of linear models and interpretation of its results when some of the covariates are proportions.

KEYWORDS
compositional covariates, conditional plot, inference, linear model, marginal plot, multicollinearity, parameter identifiability, parameter interpretation

INTRODUCTION

Proportion variables (also known as compositional variables) abound in the biological sciences. Examples of these variables include species composition (e.g., Gloor et al., 2017), soil mineral composition (e.g., Czechowski et al., 2022), the fraction of biomass in different parts of plants (e.g., Poorter et al., 2012), the proportion of wildlife behaviors throughout the day or the year (e.g., Cullen et al., 2022, 2023), and the proportion of land use/land cover (LULC) surrounding the location in which observations were made (e.g., Fink et al., 2020). These compositional variables are often used as covariates (also known as independent, explanatory, or predictor variables) when attempting to understand different types of phenomena. For example, LULC change is the major driver of biodiversity decline and ecosystem integrity loss throughout the world (Díaz et al., 2019; Tilman et al., 2017) and, as a result, there has been considerable interest in understanding how the surrounding LULC influences a wide range of phenomena, such as animal behavior (e.g., Gallo et al., 2022; Giroux et al., 2023; Noonan et al., 2022; Paviolo et al., 2018; Zeller et al., 2016), occurrence/abundance of species (e.g., Canibe et al., 2022; Fink et al., 2020; Miller et al., 2019; Valle et al., 2022), water quality and pollutant concentration (Hoek et al., 2008; Piffer et al., 2021), perception of environmental problems (Suchy et al., 2023), and disease incidence (e.g., Machado et al., 2023; Valle & Tucker Lima, 2014).

Although considerable work has focused on how to analyze compositional data as the response variable (Aitchison, 1981; Douma & Weedon, 2019; Gloor et al., 2017; Greenacre, 2021; Jackson, 1997), there has been less attention on how compositional data, when used as covariates, can adversely impact statistical analysis. Critically, while scientists are typically aware of the impact that multicollinear covariates can have on regression models, as this is a topic that is often covered in introductory courses and textbooks in statistics (Agresti, 2002; James et al., 2013; McElreath, 2020), it is not widely recognized that variables that have sum constraints are inherently multicollinear. Variables with sum constraints are variables whose sum is always equal to a given quantity. Proportions are one type of variable that is sum-constrained (i.e., they sum to one) but other types of sum-constrained variables also exist. For example, if a buffer area is created around each sampling point and the area under each LULC category is calculated, the sum of these areas will always be equal to the overall buffer area. Similarly, if the number of hours in a day that is devoted to mutually exclusive behaviors is tracked, then these variables will also be sum-constrained because they sum to 24 h.

Our goal with this article is to raise awareness of the problems that occur when proportion covariates (as an example of sum-constrained covariates) are included in regression models and to discuss some straightforward solutions to these problems. In particular, we show how substantially different conclusions can be reached if one is not careful with how to interpret parameter estimates associated with these proportion covariates and how to visualize the corresponding relationships. We start by explaining why variables with sum constraints are inherently multicollinear and how this is detrimental to statistical analysis. Subsequently, we describe how this problem manifests itself when fitting a wide range of models, both in a frequentist and Bayesian framework. Then, we describe some solutions to overcome the estimation and interpretation problems that result from using proportions as covariates. Finally, we illustrate the issues related to parameter estimation when using proportional covariates and the proposed solutions by modeling how the richness of native birds is influenced by LULC in the United States using Breeding Bird Survey data. In this article, all analyses and figures were created in R (R Core Team, 2020) and the corresponding code is available at https://doi.org/10.6084/m9.figshare.25024478.v1 (Valle et al., 2024).

MULTICOLLINEARITY AND PARAMETER IDENTIFIABILITY

It can be intuitive that a negative correlation arises from variables that always sum to a given constant (i.e., an increase in one variable has to necessarily decrease one or more of the other variables to ensure that the sum remains constant) (Aitchison, 1981; Chayes, 1960; Greenacre, 2021; Jackson, 1997). However, it is not necessarily obvious that these variables will always result in a multicollinearity problem that is severe enough to preclude the estimation of some of the parameters in linear models and that will require careful interpretation of the modeling results. The goal of this section is to clarify why such a severe multicollinearity problem arises.

Multicollinearity arises when one covariate can be expressed as a linear combination of the other covariates. Let the j-th covariate for the i-th observation be denoted by x_{ij} ($i = 1, \ldots, n$ and $j = 1, \ldots, p$). By definition, variables that are sum-constrained can be written as:

$$\sum_{j=1}^{p} x_{ij} = K,$$

where K is a constant. In the case of proportions, K is equal to 1. As a result, multicollinearity arises because any given covariate x_{ij}, can be written as a linear combination of the remaining covariates:

$$x_{ij'} = K - \sum_{j \neq j'}^{p} x_{ij},$$

where $j \neq j'$ indicates that the sum is performed over all j except for j'.

A set of parameters is deemed to be jointly unidentifiable when the likelihood does not change with changes in parameter values. In other words, multiple sets of parameters are equally likely. This identifiability problem is a direct result of multicollinearity and is important because the interpretation of statistical modeling results fundamentally depends on these parameter values (e.g., Is there a statistically discernible relationship between the response and a focal predictor variable? Is this relationship positive or negative?). Here we show how identifiability problems will inevitably occur when proportions are included in regression models.

To help explain how sum-constrained variables (e.g., proportions) impact parameter estimation, we will focus in this section on an example in which we want to understand how the surrounding landscape characteristics (e.g., area of forest and agricultural land) influence the species richness of birds. More specifically, we assess the species richness of birds (denoted as R) using point sampling and we calculate the area covered by forests and agriculture within a circular buffer around each sampling location (denoted as A_{for} and A_{agri}, respectively). Assuming that forests and agriculture correspond to all the LULC classes in the surrounding landscape, then the sum of their areas will be equal to the buffer area and the proportion of forests and agriculture (denoted as $p_{\text{for}} = \frac{A_{\text{for}}}{A_{\text{for}} + A_{\text{agri}}}$ and $p_{\text{agri}} = \frac{A_{\text{agri}}}{A_{\text{for}} + A_{\text{agri}}}$, respectively) will sum to one. In a standard multiple regression, we assume that:

$$E[R] = \beta_0 + \beta_1 p_{\text{for}} + \beta_2 p_{\text{agri}}, \tag{1}$$

where β_0 is the intercept and β_1 and β_2 are the slope parameters.

Parameter unidentifiability arises because the likelihood remains the same if we increase the slope parameters by a factor Δ while simultaneously decreasing the intercept by the same amount, as shown in Equation (2):

$$\begin{aligned} E[R] &= (\beta_0 - \Delta) + (\beta_1 + \Delta) p_{\text{for}} + (\beta_2 + \Delta) p_{\text{agri}} \\ &= (\beta_0 - \Delta) + \Delta\left(p_{\text{for}} + p_{\text{agri}}\right) + \beta_1 p_{\text{for}} + \beta_2 p_{\text{agri}} \end{aligned}. \tag{2}$$

Because $p_{\text{for}} + p_{\text{agri}} = 1$, we can write this expression as:

$$E[R] = \beta_0 + (-\Delta + \Delta) + \beta_1 p_{\text{for}} + \beta_2 p_{\text{agri}},$$

which is equal to our original Equation (1) despite the fact that the slope parameters were increased and the intercept was decreased by Δ.

To provide a geometric intuition for this problem, we display simulated data in Figure 1. Notice that all the data lie in a single "data" plane (gray surface) because p_{for} and p_{agri} have to sum to one (Figure 1A). Furthermore, each distinct parameter set $\{\beta_0, \beta_1, \beta_2\}$ defines a distinct plane. For example, the blue surface in Figure 1B arises by setting the slope of p_{agri} to zero and estimating the remaining parameters. Conversely, the red surface in Figure 1B arises by setting the intercept to zero and estimating all the other parameters. As a result, the parameter identifiability problem described above means

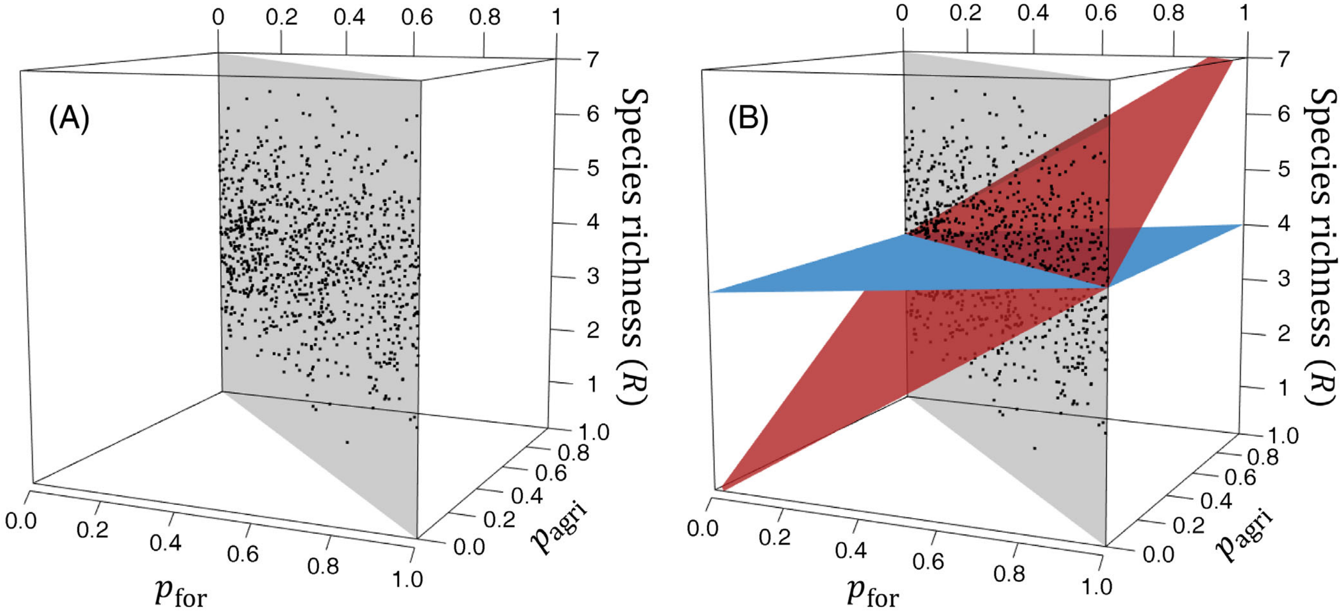

FIGURE 1 Panel (A) displays the distribution of the simulated data. Notice that all observations fall in the gray surface (i.e., "data" plane) because of the condition $p_{\text{for}} + p_{\text{agri}} = 1$. Panel (B) shows an example of two different planes that go through the mean of the data. These planes were created by setting the slope of p_{agri} to zero (blue surface) or the intercept to zero (red surface) and estimating the remaining parameters.

that multiple planes are able to go through the mean of the data equally well.

It is important to emphasize that we restricted this example to two proportion covariates to enable the visualization of the estimated planes in Figure 1 but multicollinearity and identifiability problems will be present even if more than two proportion covariates are used, as exemplified below.

NUMERICAL EXAMPLE TO ILLUSTRATE THE PROBLEMS WHEN USING PROPORTION COVARIATES

The goal of this section is to illustrate how the multicollinearity and identifiability problem described above will impact a wide range of linear models (e.g., standard linear models [LM], generalized linear models [GLM], generalized linear mixed models [GLMM], and generalized additive models [GAM]) using a numerical example. We simulate a dataset with 1000 observations for which bird species richness was a function of the proportion of forest, agriculture, and wetlands (p_{for}, p_{agri}, and p_{wet}, respectively). More specifically, we rely on a Poisson regression model:

$$R \sim \text{Poisson}\left(\exp\left(\beta_0 + \beta_1 p_{\text{for}} + \beta_2 p_{\text{agri}} + \beta_3 p_{\text{wet}}\right)\right),$$

where β_0, β_1, β_2 and β_3 were set to 3, −1, 0, and 0.5, respectively. We draw the p_{for}, p_{agri}, and p_{wet} covariates from a Dirichlet($\mathbf{1}$) distribution. The Dirichlet distribution is a multivariate distribution that is used to model nonnegative positive numbers that sum to one (e.g., probabilities or proportions). When the parameters of this distribution are all equal to 1, then it is equivalent to a uniform distribution within the area restricted by the sum-to-one constraint (i.e., the simplex). This dataset can be modeled in a straightforward fashion using a GLM but another common approach is to log-transform the response variable (assuming none of the response variables is equal to 0) and use a LM. We also include a region identifier in our dataset because this variable allows us to include region-specific random effect intercepts as part of a GLMM. Finally, we include spatial coordinates x and y (randomly generated from independent uniform distributions) because these variables allow us to use bivariate splines within GAMs to accommodate for potential spatial correlation (e.g., Toh et al., 2021). Both the region identifier and the spatial coordinates were unrelated to the response variable.

A standard procedure before fitting any model is to check for multicollinearity problems by calculating the correlation between all continuous covariates (e.g., Feng et al., 2019). The resulting correlation matrix shown in Table 1 reveals that none of the pairwise correlations are particularly high, incorrectly suggesting that multicollinearity is not a problem. Despite the widespread use of pairwise correlations to assess multicollinearity, this problem can be better diagnosed using other approaches, such as through the use of the variance inflation factor (VIF) (James et al., 2013). However, in extreme situations with perfect multicollinearity, in which one covariate is exactly equal to a linear combination of a subset of the other covariates, VIF cannot be calculated. The inability to calculate VIF is a telltale sign of perfect multicollinearity and this is precisely the case for our numerical example.

TABLE 1 Estimated correlation matrix for the continuous covariates in the simulated data.

Covariates	p_{for}	p_{agri}	p_{wet}	x	y
p_{for}	1	−0.51	−0.51	−0.02	−0.01
p_{agri}	−0.51	1	−0.48	0.02	−0.05
p_{wet}	−0.51	−0.48	1	0	0.06
x	−0.02	0.02	0	1	−0.02
y	−0.01	−0.05	0.06	−0.02	1

We fit multiple models in R to this simulated dataset. More specifically, we fit a standard LM using the "lm" function in base R in which the response variable was log species richness. We also fit a GLM ("glm" function in base R) and a GLMM with region-specific random intercepts (function "glmer" in the R package *lme4*; Bates et al., 2015). Finally, we fit a GAM with a bivariate spline for the spatial coordinates using the function "gam" in the R package *mgcv* (Wood, 2017). All of these models were fitted in a frequentist framework. Our last model consisted of a GLM, fit in a Bayesian framework using JAGS (Plummer, 2003). The LM assumes a Gaussian likelihood whereas the GLMs, GLMM, and GAM relied on a Poisson likelihood.

Our results show that, even though no model selection was performed, the p_{wet} covariate is automatically dropped (i.e., removed from the model by the algorithm without any input from the user) in models 1, 2, and 3 and the corresponding slope parameters are set to missing (Table 2). Conversely, the slope of p_{for} (and its standard error) is set to zero for GAM and is almost zero for the Bayesian model, having an effect that is similar to dropping p_{for} (instead of p_{wet}) from the model. These results are surprising because the only covariate with a slope of zero was p_{agri} in our simulation and, as a result, one would expect that p_{agri} would be the dropped covariate, not p_{wet} or p_{for}. Importantly, the order in which the

TABLE 2 Parameter estimates (and standard errors) based on different models fit to the same simulated dataset.

Model	Function (package/software)	Type of model	Estimated parameters (standard error)	Shifted parameters[a]
1	lm (*base R*)	Gaussian linear model (response variable is log(R))	$\hat{\beta}_0 = 3.47\,(0.03)$	$\beta_0 = 3.47 - 0.47 = 3.00$
			$\hat{\beta}_1 = -1.55\,(0.04)$	$\beta_1 = -1.55 + 0.47 = -1.08$
			$\hat{\beta}_2 = -0.44\,(0.04)$	$\beta_2 = -0.44 + 0.47 = 0.03$
			$\hat{\boldsymbol{\beta}}_{\mathbf{3}} = \mathbf{NA}\,(\mathbf{NA})$	$\beta_3 = 0 + 0.47 = 0.47$
2	glm (*base R*)	Poisson regression model (GLM)	$\hat{\beta}_0 = 3.48\,(0.02)$	$\beta_0 = 3.48 - 0.48 = 3.00$
			$\hat{\beta}_1 = -1.49\,(0.04)$	$\beta_1 = -1.49 + 0.48 = -1.01$
			$\hat{\beta}_2 = -0.43\,(0.03)$	$\beta_2 = -0.43 + 0.48 = 0.05$
			$\hat{\boldsymbol{\beta}}_{\mathbf{3}} = \mathbf{NA}\,(\mathbf{NA})$	$\beta_3 = 0 + 0.48 = 0.48$
3	glmer (*lme4*)	Poisson regression mixed model (GLMM) with region-specific random effects	$\hat{\beta}_0 = 3.48\,(0.02)$	$\beta_0 = 3.48 - 0.48 = 3.00$
			$\hat{\beta}_1 = -1.49\,(0.04)$	$\beta_1 = -1.49 + 0.48 = -1.01$
			$\hat{\beta}_2 = -0.43\,(0.03)$	$\beta_2 = -0.43 + 0.48 = 0.05$
			$\hat{\boldsymbol{\beta}}_{\mathbf{3}} = \mathbf{NA}\,(\mathbf{NA})$	$\beta_3 = 0 + 0.48 = 0.48$
4	gam (*mgcv*)	Generalized additive model (GAM) with bivariate splines for spatial coordinates	$\hat{\beta}_0 = 1.99\,(0.03)$	$\beta_0 = 1.99 + 1.01 = 3.00$
			$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = \mathbf{0}\,(\mathbf{0})$	$\beta_1 = 0 - 1.01 = -1.01$
			$\hat{\beta}_2 = 1.06\,(0.04)$	$\beta_2 = 1.06 - 1.01 = 0.05$
			$\hat{\beta}_3 = 1.49\,(0.04)$	$\beta_3 = 1.49 - 1.01 = 0.48$
5	jags (*JAGS*)	Poisson regression model (GLM)	$\hat{\beta}_0 = 2.07\,(1.63)$	$\beta_0 = 2.07 + 0.93 = 3.00$
			$\hat{\beta}_1 = -0.08\,(1.63)$	$\beta_1 = -0.08 - 0.93 = -1.01$
			$\hat{\beta}_2 = 0.99\,(1.63)$	$\beta_2 = 0.99 - 0.93 = 0.06$
			$\hat{\beta}_3 = 1.41\,(1.63)$	$\beta_3 = 1.41 - 0.93 = 0.48$

Note: We highlight in bold the parameters that were either set to zero or were removed from the model (denoted by NA). All models were fitted in R.

[a]The parameter values used to simulate data were $\beta_0 = 3, \beta_1 = -1, \beta_2 = 0, \beta_3 = 0.5$.

covariates are specified in R impacts which covariate ultimately gets dropped in models 1 through 4. These results reveal the arbitrariness regarding which covariates are dropped from the model. Importantly, these discrepant parameter estimates can lead to substantially different inferences (e.g., the proportion of agriculture is found to be negatively associated with species richness for models 1–3 but positively associated with species richness for models 4 and 5). Unfortunately, few functions provide warnings to alert the user that changes have been automatically implemented to ensure parameter identifiability, highlighting the importance of raising awareness regarding these issues. For example, among the five different functions used in Table 2, only "glmer" issued a warning. Despite different parameter estimates, we note that all models estimate the true parameters well ($\beta_0 = 3, \beta_1 = -1, \beta_2 = 0, \beta_3 = 0.5$) once the intercept and the slopes are appropriately shifted by the corresponding constant Δ as described in Equation (2) (see column "shifted parameters" in Table 2). Unfortunately, we can only deduce the constant Δ for a given model because we know the true parameter values used to simulate the data but this constant is not estimable with real data.

Importantly, for the Bayesian model, the standard errors for the slope parameters are very large, resulting in very wide 95% credible intervals and no covariate being statistically discernible from zero (Table 2). This is a direct result of the model having unidentifiable parameters. Furthermore, although our Markov Chain Monte Carlo (MCMC) algorithm converged in this case, our experience has been that more complicated Bayesian models can often fail to converge if they have unidentifiable parameters.

We emphasize that predictions of the mean will be very similar for these different models as long as these predictions are made within the data plane (i.e., the gray surface in Figure 1). In other words, in our example, predictions will be the same if $p_{\text{for}} + p_{\text{agri}} + p_{\text{wet}} = 1$ but will differ if $p_{\text{for}} + p_{\text{agri}} + p_{\text{wet}} \neq 1$. This can be easily shown by calculating the mean $\exp\left(\beta_0 + \beta_1 p_{\text{for}} + \beta_2 p_{\text{agri}} + \beta_3 p_{\text{wet}}\right)$ using the parameter coefficients given in Table 2 and selecting values for p_{for}, p_{agri} and p_{wet} that either satisfy or do not satisfy the sum constraint. A consequence of different models having similar predictions is that model selection procedures (e.g., using information criteria such as Akaike Information Criterion [AIC] or Bayesian Information Criterion [BIC]) can lead to misleading results when applied to regression models with proportion covariates. For example, by repeatedly simulating the data as described above and evaluating models with all possible combinations of p_{for}, p_{agri}, and p_{wet}, both AIC and BIC selected models with p_{agri} approximately 85% of the times despite the fact that this covariate has a slope of zero in the simulated data (Appendix S1). When performing model selection with more covariates, a better approach would be to add or drop the whole set of compositional covariates instead of adding or dropping individual compositional terms.

SOLVING THE MULTICOLLINEARITY AND PARAMETER IDENTIFIABILITY PROBLEMS

The most common approach to avoid these multicollinearity and identifiability problems is to drop one of the covariates; this is often automatically done in statistical software (e.g., see Table 2). While this approach solves the identifiability problem, an important limitation is that there are no guidelines for which variable should be dropped and different software may drop different variables (see Table 2). Furthermore, even if one proportion variable is dropped, multicollinearity problems can still persist if the sum of the remaining proportion covariates is still approximately constant. For example, dropping p_{wet} will avoid exact multicollinearity. However, if wetlands are relatively rare in the landscape, then p_{wet} will be small and dropping them will not avoid the (approximate) collinearity between p_{for} and p_{agri} given that their sum is still close to 1.

A better solution to the multicollinearity and identifiability problem might be to drop the intercept instead of removing one of the covariates. This solution avoids the subjectivity associated with deciding which covariate to drop. Some might argue that this is an overly drastic solution because it forces the mean to be a particular value (0, 0.5, or 1, depending on the link function) when all covariates are zero, irrespective of the data. However, in cases where part of the covariates are proportions that sum to one, selecting a zero intercept is less controversial because it is impossible for all covariates to be zero given the sum constraint. Constraining the intercept to be zero leads to consistency among models. For example, when we fit the different models described in Table 2 without an intercept, we find that all of them yield approximately the same parameter estimates (Appendix S2). Also, different from the results in Table 2, the Bayesian model does not have very large standard errors once the intercept is removed, a strong indication that the parameter identifiability problem has been solved. Finally, in simulation studies in which the true parameter values are known, the original parameter values can be recovered using the estimated parameters and by setting Δ to $-\beta_0$ in Equation (2).

Other approaches focused on transforming the compositional data also exist. For example, there is a rich history associated with log-ratio transformations of compositional data (see early proposals in Aitchison, 1981, 1982, 1984). Unfortunately, interpreting the slope parameters when proportion covariates are log-ratio transformed can be very challenging, a problem that is compounded when the denominators in these ratios are not the same for the different variables (Coenders & Pawlowsky-Glahn, 2020). Furthermore, if some of the compositional data are equal to 0, then some of these ratios might be undefined because of a zero denominator. For both of these reasons, we refrain from further exploring transformations of the compositional data.

ADDRESSING INTERPRETATION USING DIFFERENCES IN SLOPES AND CONDITIONAL AND MARGINAL PLOTS

A fundamental issue associated with both solutions to the multicollinearity and identifiability problems discussed above (i.e., dropping a covariate or dropping the intercept) is that, even if the compositional data are not transformed, interpretation of the estimated parameters is still not straightforward. Some might even argue that it does not make sense to include proportions as explanatory variables in a regression model because this invalidates the usual interpretation of the slope parameters (i.e., a change in the mean of the response variable due to a one-unit increase in the corresponding covariate while all remaining covariates are held constant). When sum constraints are present, it is impossible to hold all other variables constant while changing a single variable. For instance, we cannot increase p_{for} while holding p_{agri} and p_{wet} constant. However, we believe that the removal of all proportion covariates from regression models is an overly restrictive solution given that there are numerous examples of other regression models for which this interpretation of slope parameters also does not hold (e.g., when regression models include quadratic terms, interaction terms, or splines). In these cases, modelers often resort to graphical approaches to understand the estimated relationship between the response variable and each covariate.

Dropping a covariate is similar to what regression models typically do when factor covariates are used. In this situation, one of the levels of the factor variable is omitted and serves as the baseline against which all the other levels are compared. However, in the case of proportions, one has to interpret results carefully because the left-out variable will implicitly change once the other proportion covariates remaining in the model change. For example, if p_{wet} is dropped (as in model 2 in Table 2), then β_1 should be interpreted as the change of the mean (on the log scale) as p_{for} increases while all the other covariates are kept constant and p_{wet} decreases. In other words, an increase of a proportion covariate (while keeping the other proportion covariates constant) necessarily has to come at the expense of decreasing the left-out proportion covariate. This concern is particularly relevant when standard model selection procedures are adopted because it is easy to overlook the proportion covariate that was excluded.

We propose three approaches to improve the interpretation of model results: (1) focusing on the difference in slope parameters; (2) displaying conditional plots with two horizontal axes; and (3) using marginal plots to display model results. In relation to (1), we propose that the focus should be on the interpretation of the difference in slope parameters because, as shown by Equation (2), while the individual parameters are not identifiable, the difference in parameters is identifiable. In other words, if the estimated parameters are shifted by Δ (i.e., $\widehat{\beta_i} = \beta_i + \Delta$ and $\widehat{\beta_j} = \beta_j + \Delta$), then we can still reliably estimate $\beta_i - \beta_j$ with $\widehat{\beta_i} - \widehat{\beta_j}$ without having to worry about Δ as this quantity disappears from the expression. But how do we interpret the difference in slope parameters? The quantity $\widehat{\beta_i} - \widehat{\beta_j}$ can be interpreted as how much the response variable changes as we increase x_i by one unit while decreasing x_j by one unit. Finally, notice that this approach also works for models that include an intercept and for which one of the compositional covariates was dropped as long as we assume that the slope of the dropped-out covariate is equal to 0.

We also propose the display of conditional relationships using graphs that contain two x-axes to make explicit which two proportion covariates are simultaneously changing. More specifically, one x-axis should depict the values of the focal proportion covariate while the other x-axis should show the value of the proportion covariate that was left out. This graphical approach is also useful if the intercept is dropped but, in this case, one must explicitly decide which other proportion covariate decreases as the focal proportion covariate increases. We call these figures conditional graphs because they display the estimated relationship between a focal proportion covariate (and the other proportion variable that co-varies with it) and the response variable while fixing the other proportion covariates to zero. An example of this type of graph is provided in our case study.

Another approach for depicting the relationship between the response variable and a focal proportion covariate is to use what we call marginal plots. Marginal plots show predictions for a wide range of values of the proportion covariates, instead of fixing the remaining proportion covariates to zero. More specifically, we generate random samples of the proportion covariates while ensuring that $p_{\text{for}} + p_{\text{agri}} + p_{\text{wet}} = 1$ (e.g., using the Dirichlet distribution) and then make predictions based on these covariate values. Importantly, this graph does not depend on which covariate was dropped (or if the intercept was dropped) because it relies on model predictions within the area defined by $p_{\text{for}} + p_{\text{agri}} + p_{\text{wet}} = 1$. Note that, while we rely on "flat" Dirichlet distributions (i.e., a Dirichlet($\mathbf{1}$)) to create these compositional covariates, other options are also possible (e.g., using Dirichlet distributions with parameters based on the means of each compositional covariate in the dataset). We illustrate both the conditional and marginal plots in our case study. We have included an R tutorial to show how both the conditional and marginal plots can be created using either base R or *ggplot* (Appendix S3).

CASE STUDY USING THE BREEDING BIRD SURVEY DATA

We illustrate the issues related to parameter estimation when using proportional covariates and the proposed solutions with a case study relating the richness of native birds to LULC variables in the United States. To this end, we used a database derived from the North American Breeding Bird Survey (BBS; Pardieck et al., 2017) for 2010–2014 in the United States (Knowles & Flather, 2021) and the LULC map of 2013 provided by the US National Land Cover Database (NLCD; Dewitz & USGS, 2021). The BBS is a continental-scale bird monitoring program in which trained volunteers conduct annual point count surveys on established roadside routes. Routes are approximately 39.4 km long and point counts are carried out every 0.8 km within a time interval of 3 min. The processed BBS database we used provides richness estimates for different groups of bird species and the BBS routes centroids (Knowles & Flather, 2021). We filtered the estimates of native bird richness for 2013 and selected data with both richness estimates and route centroids ($N = 2237$). For each route, we calculated the proportion of each LULC class (excluding unclassified pixels) within a 20-km radius buffer around the route centroid. For this, we grouped the NLDC LULC classes into six variables: (1) Crop/Pasture, (2) Developed, (3) Forest, (4) Open habitat, (5) Water bodies/Snow, and (6) Wetland. No correlation between the LULC variables was greater than 0.5 in absolute value.

We modeled the relationship between bird richness and LULC classes using four different GLMs with a negative binomial distribution for the response variable: (1) a model with an intercept and all the six LULC covariates; (2) a model similar to model (1) but without the Wetland class; (3) a model similar to model (1) but without the Open habitat class; and (4) a model similar to model (1) but without the intercept. Note that we relied on a negative binomial regression model, instead of the Poisson regression model used for the simulated data, to account for the potential overdispersion of bird richness. Models were fit in a Bayesian framework using JAGS (Plummer, 2003), accessed from R with the package *jagsUI* (Kellner, 2024). For each model, we ran three parallel MCMC chains consisting of 1000 iterations in the adaptive phase, followed by 12,000 iterations, from which the first 4000 were excluded (burn-in). We used vague priors on all parameters. Parameter estimates (mean and 95% credibility intervals) and predictions were calculated using the 24,000 resulting samples of the posterior distributions. Algorithm convergence was assessed using trace plots and the *R*-hat statistic (i.e., values of *R*-hat ≤1.1 were used as an indicator of MCMC convergence).

As expected, our MCMC algorithm did not converge for model 1, a sign that parameters are not identifiable in this model, whereas the remaining models converged successfully. Interestingly, models 2–4 yielded very different parameter estimates (Table 3), with important implications for the conclusions that are drawn from these results. For example, when Wetland is left out (i.e., model 2), the parameter associated with Developed is estimated to be negative, suggesting that species richness decreases with the proportion of the surrounding area that is developed. However, when Open habitat is left out (i.e., model 3) or when the intercept is removed (i.e., model 4), we reach the opposite conclusion (i.e., species richness increases with the proportion of the surrounding area that is developed). Despite these discrepant parameter estimates, it is important to note that the difference between parameter estimates is consistent, a useful feature when attempting to reconcile these contrasting results. For example, these three models revealed that we expect that average species richness will increase by approximately 43% (i.e., $\frac{\exp\left(\widehat{\beta}_0 + \widehat{\beta}_{\text{forest}} \times 1\right)}{\exp\left(\widehat{\beta}_0 + \widehat{\beta}_{\text{developed}} \times 1\right)} = \exp(0.1 - (-0.26)) = 1.43$) as we move from an area that is 100% developed to an area that is 100% forested.

Recall that we propose conditional plots with two *x*-axes to make explicit which covariate was removed or is being co-varied together with the focus variable. We illustrate these plots by depicting the estimated relationships between species richness and each LULC covariate for the two models with a left-out covariate (models 2 and 3). Despite the very contrasting relationships depicted in Figure 2, they are not as puzzling if one realizes that different baseline variables (i.e., the variable that was excluded) were used and, as a result, predictions for very different types of landscapes are being made. For example, Figure 2B shows that the average bird richness diminishes in a landscape where developed areas increase and wetlands decrease while Figure 2F shows that average richness increases slightly when developed areas increase and open habitats decrease.

We can also use marginal plots to interpret model results. Recall that marginal plots show the relationship between species richness and a particular focal covariate while allowing the other covariates to take on different values, subject to the constraint that all proportion

TABLE 3 Slope estimates (and 95% credible intervals [CI]) for three different models relating bird species richness to land-use/land-cover (LULC) covariates.

Parameters	**Model 2 (no wetland)**	**Model 3 (no open habitat)**	**Model 4 (no intercept)**
Crop	**−0.16 (−0.26; −0.06)**	**0.27 (0.22; 0.33)**	**4.18 (4.14; 4.22)**
Developed	**−0.26 (−0.46; −0.07)**	**0.17 (0.01; 0.34)**	**4.07 (3.91; 4.24)**
Forest	0.10 (−0.01; 0.20)	**0.53 (0.48; 0.58)**	**4.44 (4.40; 4.47)**
Open	**−0.44 (−0.53; −0.34)**	…	**3.90 (3.87; 3.93)**
Water	0.03 (−0.17; 0.23)	**0.46 (0.30; 0.62)**	**4.36 (4.21; 4.52)**
Wetland	…	**0.44 (0.35; 0.54)**	**4.35 (4.26; 4.44)**

Note: We do not report the results for model 1 because this model did not converge. Parameter estimates were judged to be significant if their corresponding 95% CI did not include zero. Significant parameters are emphasized in bold.

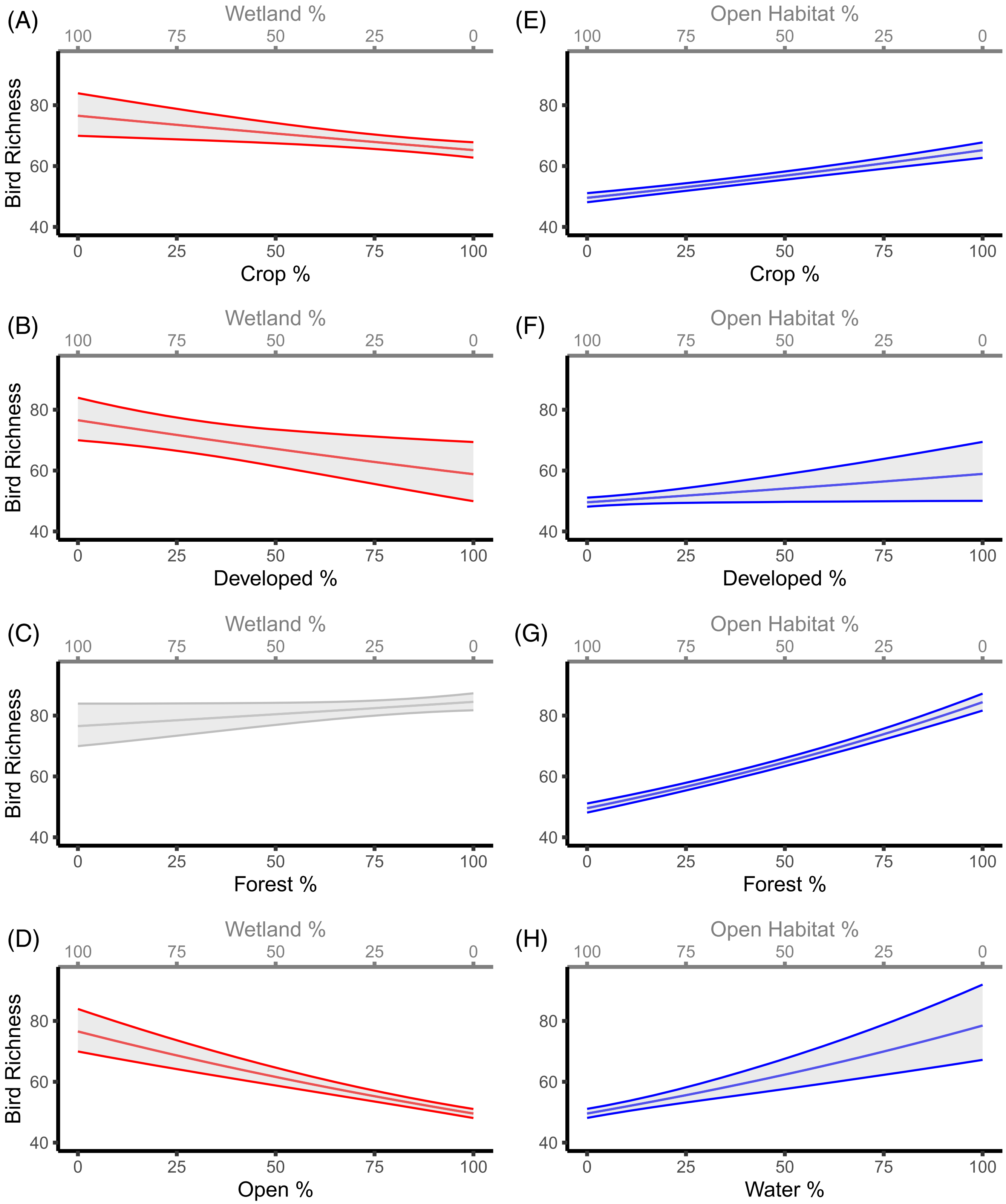

FIGURE 2 Conditional predicted relationships between native bird richness and land-use/land-cover classes from the Breeding Bird Survey of 2013 in the United States. Left panels (A–D) are based on the model without the Wetland variable (model 2) while right panels (E–H) are based on the model without the Open habitat variable (model 3). The lower and upper lines depict the 95% credible interval while the center line represents the median. Significant positive and negative relationships are shown in blue and red, respectively, whereas nonsignificant results are shown in gray.

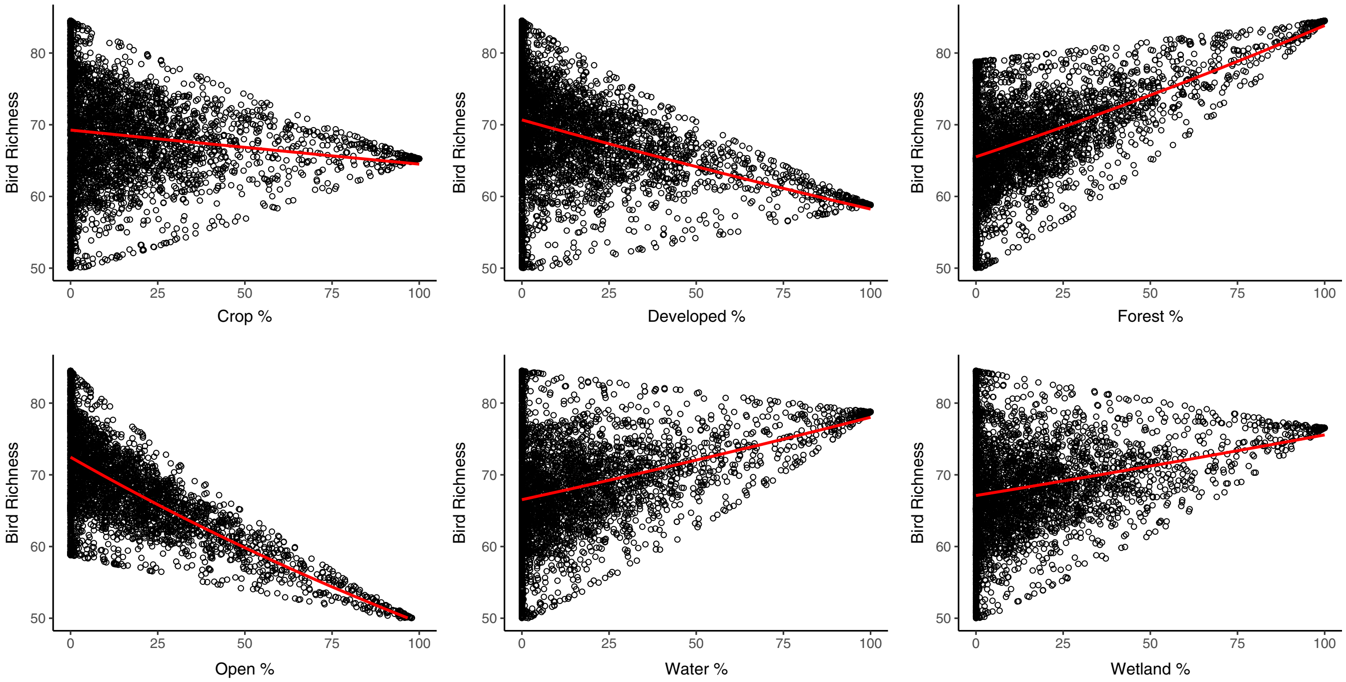

FIGURE 3 Marginal plots with the predicted relationship between the mean richness of native birds and land-use/land-cover classes from the Breeding Bird Survey of 2013 in the United States. Each point represents a model prediction based on coefficient estimates and a given set of values for the predictor variables. The red lines represent the global log–linear tendency.

covariates sum to one. Importantly, the predictions depicted in these plots are the same regardless of the model that is used to create the predictions. These plots show that, as the proportion of open habitat increases, there tends to be a decline in species richness, whereas as the proportion of forest increases there tends to be an increase in species richness (Figure 3). Conversely, there is considerably more variability regarding how species richness is influenced by the other LULC variables. Notice that marginal plots typically have a triangular shape, with a substantial scatter on the left of the plot (i.e., when the focal proportion covariate is equal to 0) and no scatter at all on the right of the plot (i.e., when the focal covariate is equal to 1). The reason for this shape is that, when the focal proportion covariate is equal to zero, there is substantial variability in predictions because the other proportion covariates can take on a wide range of values. Conversely, when the focal covariate is equal to 1 (i.e., 100% of a given LULC class), there is no variability in predictions associated with the value of the remaining proportion covariates (i.e., they are all equal to 0).

CONCLUSION

The use of sum-constrained variables is ubiquitous across multiple fields, including ecology and environmental sciences. In this article we describe how the use of these variables as covariates within regression models can have adverse effects on inference, a problem that many are unaware of. We show how this problem can be identified and discuss common solutions to this problem. Importantly, while these solutions solve the parameter identifiability problem, the interpretation of model results is still challenging even without applying compositional data transformations (e.g., log-ratio transformations; Aitchison, 1981, 1982, 1984; Coenders & Pawlowsky-Glahn, 2020). We propose to improve the interpretation of model results by focusing on the difference in slope parameters as well as visualizing results using conditional and marginal plots. By applying the proposed solutions to both simulated data and a case study, we have demonstrated that failing to properly acknowledge this problem can result in misleading conclusions. Finally, we have also described how extra care is needed when performing model selection (i.e., the whole set of compositional covariates should be added or dropped instead of adding or dropping individual compositional terms) and creating predictions (i.e., predictions for compositional covariate combinations that do not sum to one should be avoided) in the presence of compositional covariates.

It is important to note that we have purposefully focused on relatively simple regression models with few covariates to best illustrate and provide intuition for the identifiability and multicollinearity problems associated with using proportion covariates. However, statistical models are often substantially more complicated. For example, categorical covariates and/or random effects can be added to the model, splines can be included to allow for nonlinear relationships for the proportion covariates, and interaction terms can be included involving proportion covariates and other covariates that are not sum-constrained. In these more complex models, the identifiability issues discussed here remain important but additional restrictions might be needed to avoid parameter identifiability problems. For example, if additional categorical variables are included in a model with compositional covariates but without an intercept, it will be critical to exclude the first level of each of these categorical variables from the model to ensure identifiability. Likewise, some of the proposed solutions to improve interpretability might be more challenging to implement if splines or interaction terms are included in the model.

Despite the challenges described above, we believe that the practical and straightforward approaches proposed here will be of wide use for fitting GLMs and for the proper interpretation of its results when some covariates are compositional or are sum-constrained.

ACKNOWLEDGMENTS

We would like to thank Ana Meiga, Henrique Shirai, Simon Riley, and Edzard van Santen for providing helpful comments and suggestions on an earlier draft of this article. This work was supported by the US Department of Agriculture National Institute of Food and Agriculture McIntire–Stennis project 1005163 and by the NSF Division of Biological Infrastructure (DBI) award 2040819 to DV.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data and code (Valle et al., 2024) are available in Figshare at https://doi.org/10.6084/m9.figshare.25024478.v1. Our example analysis uses a database derived from the North American Breeding Bird Survey data from 2010 to 2014 in the United States (Knowles & Flather, 2021; https://doi.org/10.2737/RDS-2021-0001) and the land-use/land-cover (LULC) map of 2013 provided by the US National Land Cover Database (Dewitz & USGS, 2021; https://doi.org/10.5066/P9KZCM54).

ORCID

Denis Valle https://orcid.org/0000-0002-9830-8876
Jeffrey Mintz https://orcid.org/0000-0003-4345-366X
Ismael Verrastro Brack https://orcid.org/0000-0003-2988-9811

REFERENCES

Agresti, A. 2002. *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons.

Aitchison, J. 1981. "A New Approach to Null Correlations of Proportions." *Mathematical Geology* 13: 175–189.

Aitchison, J. 1982. "The Statistical Analysis of Compositional Data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 44: 139–177.

Aitchison, J. 1984. "Reducing the Dimensionality of Compositional Data Sets." *Mathematical Geology* 16: 617–635.

Bates, D., M. Machler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67: 1–48.

Canibe, M., N. Titeux, J. Dominguez, and A. Regos. 2022. "Assessing the Uncertainty Arising from Standard Land-Cover Mapping Procedures when Modelling Species Distributions." *Diversity and Distributions* 28: 636–648.

Chayes, F. 1960. "On Correlation between Variables of Constant Sum." *Journal of Geophysical Research* 65: 4185–93.

Coenders, G., and V. Pawlowsky-Glahn. 2020. "On Interpretations of Tests and Effect Sizes in Regression Models with a Compositional Predictor." *SORT (Statistics and Operations Research Transactions)* 44: 201–220.

Cullen, J., C. Poli, R. J. Fletcher, Jr., and D. Valle. 2022. "Identifying Latent Behavioral States in Animal Movement with M4, a Non-parametric Bayesian Method." *Methods in Ecology and Evolution* 13: 432–446.

Cullen, J. A., N. Attias, A. L. J. Desbiez, and D. Valle. 2023. "Biologging as an Important Tool to Uncover Behaviors of Cryptic Species." *PeerJ* 11: e14726.

Czechowski, P., M. de Lange, M. Knapp, A. Terauds, and M. I. Stevens. 2022. "Antarctic Biodiversity Predictions through Substrate Qualities and Environmental DNA." *Frontiers in Ecology and Evolution* 20: 550–57.

Dewitz, J., and USGS. 2021. "National Land Cover Database (NLCD) 2019 Products (version 2.0, June 2021)." U.S. Geological Survey Data Release. https://doi.org/10.5066/P9KZCM54.

Díaz, S., J. Settele, E. S. Brondízio, H. T. Ngo, J. Agard, A. Arneth, P. Balvanera, et al. 2019. "Pervasive Human-Driven Decline of Life on Earth Points to the Need for Transformative Change." *Science* 366: 1327.

Douma, J. C., and J. T. Weedon. 2019. "Analysing Continuous Proportions in Ecology and Evolution: A Practical Introduction to Beta and Dirichlet Regression." *Methods in Ecology and Evolution* 10: 1412–30.

Feng, X., D. S. Park, Y. Liang, R. Pandey, and M. Papes. 2019. "Collinearity in Ecological Niche Modeling: Confusions and Challenges." *Ecology and Evolution* 9: 10365–76.

Fink, D., T. Auer, A. Johnson, V. Ruiz-Gutierrez, W. M. Hochachka, and S. Kelling. 2020. "Modeling Avian Full Annual Cycle Distribution and Population Trends with Citizen Science Data." *Ecological Applications* 30: e02056.

Gallo, T., M. Fidino, B. Gerber, A. A. Ahlers, J. L. Angstmann, M. Amaya, A. L. Concilio, et al. 2022. "Mammals Adjust Diel Activity across Gradients of Urbanization." *eLife* 11: e74756.

Giroux, A., Z. Ortega, N. Attias, A. L. J. Desbiez, D. Valle, L. Borger, and L. G. R. Oliveira-Santos. 2023. "Activity

Modulation and Selection for Forests Help Giant Anteaters to Cope with Temperature Changes." *Animal Behavior* 201: 191–209.

Gloor, G. B., J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. 2017. "Microbiome Datasets Are Compositional: And this Is Not Optional." *Frontiers in Microbiology* 8: 02224.

Greenacre, M. 2021. "Compositional data analysis." *Annual Review of Statistics and Its Application* 8: 271–299.

Hoek, G., R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs. 2008. "A Review of Land-Use Regression Models to Assess Spatial Variation of Outdoor Air Pollution." *Atmospheric Environment* 42: 7561–78.

Jackson, D. A. 1997. "Compositional Data in Community Ecology: The Paradigm or Peril of Proportions?" *Ecology* 78: 929–940.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Kellner, 2024. "jagsUI: A Wrapper Around 'rjags' to Streamline 'JAGS' Analyses. R Package Version 1.6.2." https://CRAN.R-project.org/package=jagsUI.

Knowles, M. S., and C. H. Flather. 2021. *Data Supporting an Examination of Estimating the Potential Capacity of Ecosystems to Support Biodiversity in the Prediction of Realized Avian Diversity*. Fort Collins: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2021-0001.

Machado, C. A. L., D. Valle, M. C. Horta, A. Y. Y. Meiga, and A. P. Seva. 2023. "Patterns and Drivers of Human Visceral Leishmaniasis in Pernambuco (Brazil) from 2007 to 2018." *PLoS Neglected Tropical Diseases* 17: e0011108.

McElreath, R. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton: Chapman & Hall/CRC.

Miller, D. A. W., K. Pacifici, J. S. Sanderlin, and B. J. Reich. 2019. "The Recent Past and Promising Future for Data Integration Methods to Estimate species' Distributions." *Methods in Ecology and Evolution* 10: 22–37.

Noonan, M. J., F. Ascensao, D. R. Yogui, and A. L. J. Desbiez. 2022. "Roads as Ecological Traps for Giant Anteaters." *Animal Conservation* 25: 182–194.

Pardieck, K. L., D. J. Ziolkowski, M. Lutmerding, K. Campbell, and M. A. R. Hudson. 2017. "North American Breeding Bird Survey Dataset 1966–2016, Version 2016.0." U. S. Geological Survey, Patuxent Wildlife Research Center. https://www.pwrc.usgs.gov/bbs/RawData/.

Paviolo, A., P. Cruz, M. E. Iezzi, J. M. Pardo, D. Varela, C. de Angelo, S. Benito, et al. 2018. "Barriers, Corridors or Suitable Habitat? Effect of Monoculture Tree Plantations on the Habitat Use and Prey Availability for Jaguars and Pumas in the Atlantic Forest." *Forest Ecology and Management* 430: 576–586.

Piffer, P. R., L. R. Tambosi, S. F. D. B. Ferraz, J. P. Metzger, and M. Uriarte. 2021. "Native Forest Cover Safeguards Stream Water Quality under a Changing Climate." *Ecological Applications* 31: e02414.

Plummer, M. 2003. "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." In *3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, March 20–22, Vienna, Austria.

Poorter, H., K. J. Niklas, P. B. Reich, J. Oleksyn, P. Poot, and L. Mommer. 2012. "Biomass Allocation to Leaves, Stems and Roots: Meta-Analyses of Interspecific Variation and Environmental Control." *New Phytologist* 193: 30–50.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Suchy, A. K., E. C. Anderson, M. L. Fork, L. Lin, D. H. Locke, P. M. Groffman, J. M. Grove, S. L. LaDeau, and E. J. Rosi. 2023. "More Green, Fewer Problems: Landcover Relates to Perception of Environmental Problems." *Frontiers in Ecology and the Environment* 21: 124–130.

Tilman, D., M. Clark, D. R. Williams, K. Kimmel, S. Polasky, and C. Packer. 2017. "Future Threats to Biodiversity and Pathways to their Prevention." *Nature* 546: 73–81.

Toh, K. B., D. Valle, and N. Bliznyuk. 2021. "Improving National Level Spatial Mapping of Malaria through Alternative Spatial and Spatio-Temporal Models." *Spatial and Spatio-temporal Epidemiology* 36: 100394.

Valle, D., Y. Jameel, B. Betancourt, E. T. Azeria, N. Attias, and J. Cullen. 2022. "Automatic Selection of the Number of Clusters usingBayesian Clustering and Sparsity-Inducing Priors." *Ecological Applications* 32: e2524.

Valle, D., J. Mintz, and I. Brack. 2024. "Valle, et al., 2024, Ecology. Proportional Covariates: Data and Code." Figshare. https://figshare.com/articles/software/Valle_et_al_2024_Ecology_Proportional_covariates_data_and_code_/25024478.

Valle, D., and J. M. Tucker Lima. 2014. "Large-Scale Drivers of Malaria and Priority Areas for Prevention and Control in the Brazilian Amazon Region Using a Novel Multi-Pathogen Geospatial Model." *Malaria Journal* 13: 443.

Wood, S. N. 2017. *Generalized Additive Models: An Introduction with R*. Boca Raton: CRC Press.

Zeller, K. A., K. McGarigal, S. A. Cushman, P. Beier, T. W. Vickers, and W. M. Boyce. 2016. "Using Step and Path Selection Functions for Estimating Resistance to Movement: Pumas as a Case Study." *Landscape Ecology* 31: 1319–35.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Valle, Denis, Jeffrey Mintz, and Ismael Verrastro Brack. 2024. "Estimation and Interpretation Problems and Solutions When Using Proportion Covariates in Linear Regression Models." *Ecology* 105(4): e4256. https://doi.org/10.1002/ecy.4256