Received: 12 October 2018 | Accepted: 24 May 2019

DOI: 10.1111/2041-210X.13234

REVIEW

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression

Jacob C. Douma[1,2] | James T. Weedon[3,4]

[1]Centre for Crop Systems Analysis, Wageningen University, Wageningen, The Netherlands

[2]Laboratory of Entomology, Wageningen University, Wageningen, The Netherlands

[3]Department of Biology, University of Antwerp, Antwerpen, Belgium

[4]Department of Ecological Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Correspondence
Jacob C. Douma
Email: bob.douma@wur.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 863.14.018 & 016.171.089

Handling Editor: David Warton

Abstract

1. Proportional data, in which response variables are expressed as percentages or fractions of a whole, are analysed in many subfields of ecology and evolution. The scale-independence of proportions makes them appropriate to analyse many biological phenomena, but statistical analyses are not straightforward, since proportions can only take values from zero to one and their variance is usually not constant across the range of the predictor. Transformations to overcome these problems are often applied, but can lead to biased estimates and difficulties in interpretation.
2. In this paper, we provide an overview of the different types of proportional data and discuss the different analysis strategies available. In particular, we review and discuss the use of promising, but little used, techniques for analysing continuous (also called non-count-based or non-binomial) proportions (e.g. percent cover, fraction time spent on an activity): beta and Dirichlet regression, and some of their most important extensions.
3. A major distinction can be made between proportions arising from counts and those arising from continuous measurements. For proportions consisting of two categories, count-based data are best analysed using well-developed techniques such as logistic regression, while continuous proportions can be analysed with beta regression models. In the case of >2 categories, multinomial logistic regression or Dirichlet regression can be applied. Both beta and Dirichlet regression techniques model proportions at their original scale, which makes statistical inference more straightforward and produce less biased estimates relative to transformation-based solutions. Extensions to beta regression, such as models for variable dispersion, zero-one augmented data and mixed effects designs have been developed and are reviewed and applied to case studies. Finally, we briefly discuss some issues regarding model fitting, inference, and reporting that are particularly relevant to beta and Dirichlet regression.
4. Beta regression and Dirichlet regression overcome some problems inherent in applying classic statistical approaches to proportional data. To facilitate the adoption of these techniques by practitioners in ecology and evolution, we present detailed, annotated demonstration scripts covering all variations of beta and Dirichlet regression discussed in the article, implemented in the freely available language for statistical computing, R.

KEYWORDS
beta regression, Dirichlet regression, fractions, non-count proportions, one augmented, proportions, transformations, zero augmented

1 | INTRODUCTION

Many types of observations in ecology and evolution can be most conveniently expressed and compared as fractions (a part of a whole). It has been estimated that over a third of publications in ecology analyse some kind of proportional data (Warton & Hui, 2011). Examples can be found in a variety of sub-fields, for example the analysis of proportional cover of a given plant functional type in vegetation survey quadrats (Defries, Hansen, Townshend, Janetos, & Loveland, 2000); the proportion of time spent by animals in a certain activity (Cotgreave & Clayton, 1994); percentages of biomass allocated to different plant organs (Poorter et al., 2012); or numbers of eggs hatching from a cohort under varying environmental conditions (De Majo, Montini, & Fischer, 2017).

Statistical analysis of proportions can present numerous difficulties. By definition, the observations are limited to numerical values between, and including, 0 and 1, and the variability in the observed proportions usually varies systematically with the mean of the response. These properties likely violate two important assumptions of standard statistical techniques that assume that the error term is normal and has constant variance. Moreover, when a whole is partitioned into more than two parts (e.g. the relative proportions of particle size classes in a soil; sand, silt, clay), analysis and interpretation become even more complex, since the response variable is now expressed as a vector of several interdependent fractional values. These properties of proportional data mean that the standard techniques of statistical analysis familiar to biologists (i.e. linear regression and ANOVA and their extensions) are usually not appropriate.

The issues related to analysing proportional data have long been recognized (Bartlett, 1936; Sokahl & Rohlf, 1995) and several analysis strategies are available to deal with them. For proportions that are derived from discrete counts, logistic or binomial regression are appropriate techniques which are well treated in most introductory biostatistics textbooks (Quinn & Keough, 2002; Zuur, Ieno, Walker, Saveliev, & Smith, 2009). For proportions not derived from counts, agreement on the most appropriate techniques is less established. A common recommendation is to apply a data transformation and proceed with ordinary linear models (Crawley, 2012; Quinn & Keough, 2002; Sokahl & Rohlf, 1995) – a solution that has important drawbacks with respect to interpretability and the validity of the resulting inference.

More recently, methods to model continuous proportions that are easier to interpret and more flexible than transformation-based solutions have become widely available, namely beta (Cribari-Neto & Zeileis, 2010; Ferrari & Cribari-Neto, 2004) and Dirichlet regression (Hijazi & Jernigan, 2009; Maier, 2014).

With the ongoing wide adoption of the open-source statistical programming language R by ecologists (R Core Team, 2013), these techniques are increasingly within reach of non-specialists. Despite the availability of these methods, their adoption in ecology and evolution is relatively low. To illustrate in relation to beta regression: if we combine Warton and Hui's (2011) estimate of 14% of ecology papers involving data based on non-count proportions, with the 156 Web of Science articles within the domain Ecology (from 2004 to 2018) that cite the key beta regression references (Cribari-Neto & Zeileis, 2010; Ferrari & Cribari-Neto, 2004; Grün, Kosmidis, & Zeileis, 2012; Smithson & Verkuilen, 2006), we arrive at a rough estimate of only 0.5% of studies using these techniques when they are potentially suitable. This suggests the timeliness and utility of a user guide that describes in non-technical terms the various possible applications and extensions of these methods for analysing proportional data derived from continuous observations.

In this article, we: (a) review the types of proportional data and identify the cases for which beta and Dirichlet regression are appropriate; (b) give brief, non-technical overviews of the principles underlying the techniques; (c) discuss some of the issues that a practitioner will encounter when applying beta and Dirichlet regression; and (d) describe the extensions to these techniques which are most relevant to researchers in ecology and evolution. In addition, we present three case studies and include annotated accompanying R code and detailed discussions in Supplementary material.

2 | TYPES OF PROPORTIONAL DATA

Proportional representations of data are commonly used when the relative amounts of two or more categories of observation are more biologically meaningful than their absolute quantities. What is usually referred to as 'proportional' or 'fractional' data in ecology is often referred to as 'compositional' data in the statistical literature (Aitchison, 1986). Proportional data can be formally understood as a division of a total W (e.g. counts, area, time, mass) into C parts or categories. If we then designate the measurements of each of the categories on a given observational unit as w_i, it follows that $W = (w_1 + \cdots + w_{C-1} + w_C)$, and that the proportions p of each category relative to the total is calculated as $p_i = \frac{w_i}{w_1+\cdots+w_{C-1}+w_C} = \frac{w_i}{W}$; and $\sum p_i = 1$. The proportions $\boldsymbol{p}$ are therefore scale independent, and are used as the response variable in subsequent modelling (see third distinction in the following paragraphs).

Proportional data can be obtained from a variety of different underlying data types, a fact that has implications for the choice of procedure used in their analysis. Therefore, before providing guidance on the statistical methods, we provide a brief classification of proportions (Figure 1). We discuss three major categorizations that can be used to subdivide all proportional data.

First, a distinction can be made between proportions arising from counts or proportions arising from continuous measurements (Warton & Hui, 2011; van den Boogaart & Tolosana-Delgado, 2013). Count-based proportions arise when the observed variables w_i that are used to calculate the proportions p_i are themselves each discrete, countable quantities that can take only non-negative integer values. For example, in plant biology, the number of pollinated inflorescences out of the total set of inflorescences observed; or in population genetics, the numbers of individuals in a sample belonging to each of various genotypes. In such cases, the calculated proportions p_i can take only a limited set of values determined by the total number of observations across all categories within a given observational unit. In contrast, continuous proportions arise when the measurements of each category that are used to compute the proportions take *continuous* non-negative values. For example, the proportion of plant biomass allocated to fruits relative to the total plant biomass. Both biomass measurements can theoretically take any real positive values, and are thus treated as continuous quantities, and in contrast to count-based proportions, p_i can take any value on the unit interval (allowing for the sum to 1 constraint with respect to other categories.) For the remainder of this article, we will focus on proportions derived from continuous (continuous) observations, i.e. the lower branch of Figure 1. Methods for analysing count-based proportions data are extensively treated in many ecological and biostatistics textbooks (Quinn & Keough, 2002; Zuur et al., 2009).

The second way in which proportional data sets can be subdivided is by the number of categories. Historically, the development

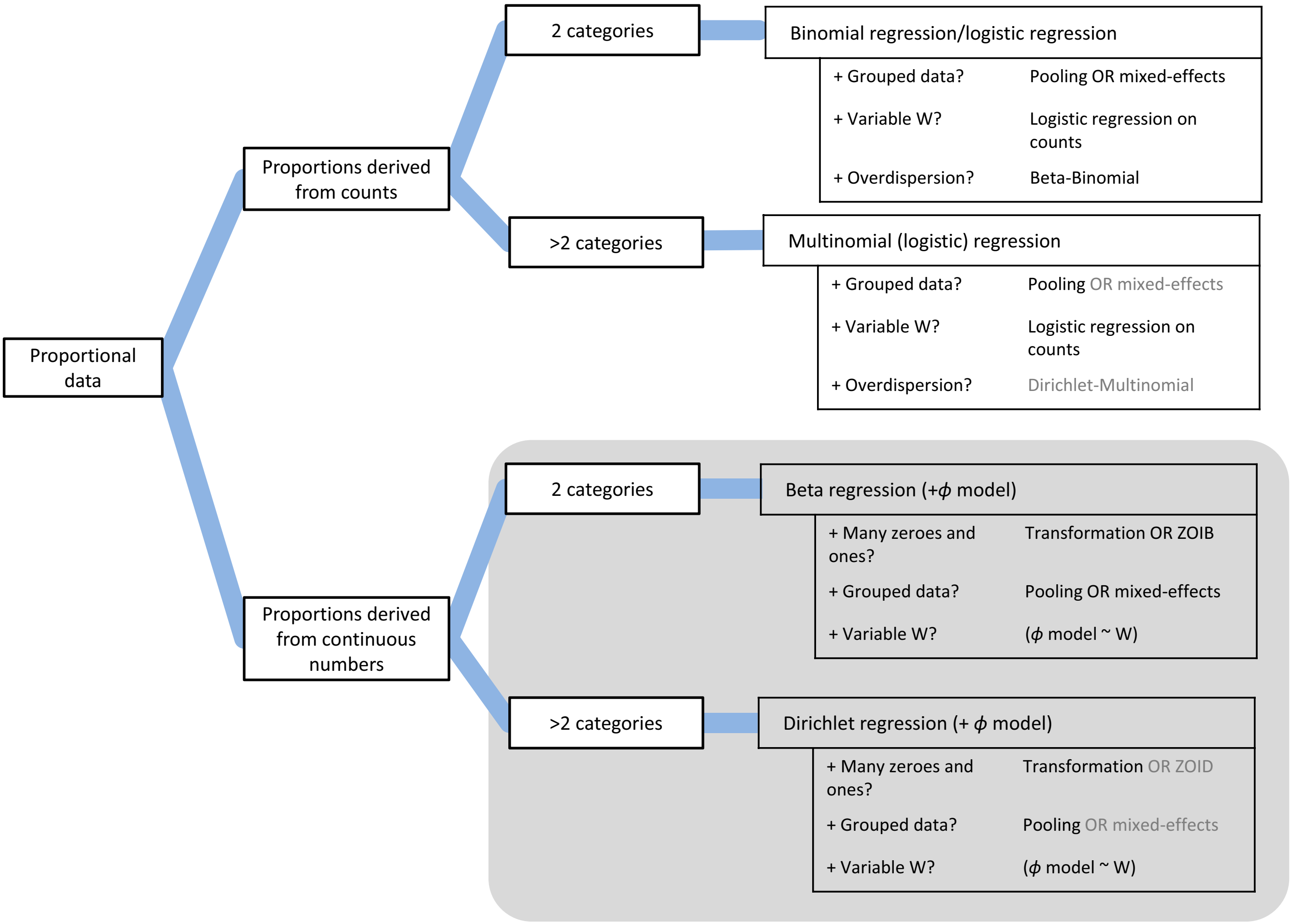

Proportional data
Proportions derived from counts
Proportions derived from continuous numbers
2 categories
>2 categories
2 categories
>2 categories
Binomial regression/logistic regression
+ Grouped data? Pooling OR mixed-effects
+ Variable W? Logistic regression on counts
+ Overdispersion? Beta-Binomial
Multinomial (logistic) regression
+ Grouped data? Pooling OR mixed-effects
+ Variable W? Logistic regression on counts
+ Overdispersion? Dirichlet-Multinomial
Beta regression (+ϕ model)
+ Many zeroes and ones? Transformation OR ZOIB
+ Grouped data? Pooling OR mixed-effects
+ Variable W? (ϕ model ~ W)
Dirichlet regression (+ ϕ model)
+ Many zeroes and ones? Transformation OR ZOID
+ Grouped data? Pooling OR mixed-effects
+ Variable W? (ϕ model ~ W)

FIGURE 1 Decision tree to determine the type of analysis for proportional observations based on properties of the data. The two most important branches are concerned with whether i) the proportions originate from discrete (counts) or continuous measurements, ii) 2 or >2 categories are modelled. At the end nodes, the type of analysis is given. Depending on the structure of the data (i.e. grouping, overdispersion etc.) extensions are available. Analyses in the grey box are the focus of this article. Grey coloured text are extensions that are currently not implemented in the software R. *W* refers to the total of the measurements from which proportions were derived, ϕ refers to a model for the variance. ZOIB and ZOID refer to 'zero/one augmented' beta and Dirichlet regression respectively. See main text for further explanations. Table 1 provides information on the possible analyses and their associated R packages

of methods for analysing proportional data has focused on two-category datasets, leading to binary proportions (e.g. percent cover in Case Study 1 below). In most cases, the proportion is presented in terms of a single category, with the complementary category merely implied (e.g. non-cover). In contrast, many ecological datasets are concerned with aggregated observations of $C > 2$ categories, e.g. leaf, stem and root mass fractions in plant biology (see Case Study 2 below). Although models for two-category data are a special case of those for more general C-category datasets, historically the development of methods have focused on either one or the other, a convention we will follow in this article.

The third distinction relates to the nature of the total measurement W of which each category w_i is a part. In some cases the value of W is fixed, for example, when plant cover is estimated from a quadrat of fixed size. In other cases, because of the nature of the system or the experimental design, the total W is variable between observational units. For example, when conducting animal behavioural studies, in which the time spent by a subject on various activities is recorded, the observed subject may move out of sight after an uncontrolled period of time, such that data p_i on time spent on different activities are computed from different total observation times (W) for different subjects. Variation in W can also arise when combining data from different studies, for example, if each study used different 'fixed' quadrat sizes, subject counts, or observation intervals.

Importantly, the spatial-temporal scale at which the variable of interest is measured is likely to affect the variability in the observed outcomes. In the extreme case when the observation interval is chosen infinitesimally small, the observed variation will be very large: vegetation is measured as present or absent, or an animal either exhibits a particular behaviour or not. Thus, by reducing the scale of observation the data can go from continuous proportions to observations resembling count-based proportions. The converse can also apply: for count-based proportions if the number of counts is extremely large, one practically moves from count-based proportion to continuous proportions since the sampling error will be low (van den Boogaart & Tolosana-Delgado, 2013). Even between these two extremes, variation in W can affect the precision with which the proportions are estimated, it can therefore sometimes usefully be incorporated into the model specification (see Section 4.1).

3 | ANALYSING PROPORTIONS ORIGINATING FROM CONTINUOUS MEASUREMENTS

Methods for analysing continuous, proportional data are less widely adopted by ecologists than those for data arising from counts and the remainder of this article will therefore focus on them. In the following sections, we present methods for analysing continuous proportions, starting with simple (two category) proportions and extending to multi-category proportions. In both cases, we briefly review traditionally used strategies for dealing with these data types, before presenting more recently developed modelling techniques.

3.1 | Analysing proportions with two continuous categories

3.1.1 | Traditional approaches – transformations

The most widely used and recommended approaches to model continuous proportions with two categories are as follows: to apply ordinary least squares techniques without transformations (Kieschnick & McCullough, 2003); or apply the arcsine transformation (Sokahl & Rohlf, 1995) or logit transformation followed by nonlinear least squares regression on the transformed variables (Warton & Hui, 2011). Modelling proportions with models that assume a normal distribution may give problems in estimation and predictions because the normal distribution allows values over the full range −∞ to ∞ and assumes constant variance. Therefore, transformations are applied to the data to meet the requirements of the statistical model. The arcsine transformation is defined as the arcsine of the square root of p: $\arcsin(\sqrt{p})$ (Quinn & Keough, 2002; Sokahl & Rohlf, 1995); while the logit transformation is defined as the natural logarithm of the odds: $\log(\frac{p}{1-p})$.

Warton and Hui (2011) showed that the logit transformation is to be preferred over the arcsine transformation because the coefficients of the logit transformation are more readily interpretable, and the arcsine leads to problems in case of extrapolation beyond the fitted range (Warton & Hui, 2011). All transformations have in common that a model for the mean proportion is estimated on the transformed scale, which is subsequently back-transformed to proportions for reporting and interpretation. As the relationship between the original and transformed proportions is usually *non*-linear, issues arise due to Jensen's inequality: for a nonlinear function $f(.)$ and a random variable x, with an average of $\bar{x}$ the average of $f(x)$ is not equal to $f(\bar{x})$ (Ruel & Ayres, 1999). This implies that parameter estimates on the transformed data will be biased when interpreted on the original untransformed scale (Cribari-Neto & Zeileis, 2010; Schmid et al., 2013, see case study 3). Furthermore, it follows from Jensen's inequality that transformations will lead to the least bias in the regions of x where the function is close to linear. In the case of a logit-transformation, this implies that bias in estimates will increase as the observations approach the asymptotic values of zero and one. In addition, the degree of bias is affected by the variation around the mean proportion. With increasing variance, the bias gets larger because the observations are spread over a larger part of the non-linear curve (see Appendix S1). It is therefore advisable to model proportional data on the original (untransformed) scale of the observations whenever possible.

3.1.2 | Beta regression

Beta regression is a technique that has been proposed for modelling of data for which the observations are limited to the open interval (0, 1) (Ferrari & Cribari-Neto, 2004; Smithson & Verkuilen, 2006). Some recent examples of beta regression in ecological contexts include analyses of: the contribution of food derived from different energy pathways (Child & Moore, 2015; Fukumori, Yoshizaki, Takamura, & Kadoya, 2016); forest simulation output in terms of proportions of tree basal area belonging to focal tree species (Ameztegui, Coll, & Messier, 2015); and the degree of leaf damage due to a leaf pathogen (Busby et al., 2013). Although we present beta regression in this review as a method for analysing continuous-based proportions, there are also many examples for using this techniques to analyse data from derived indices that are bound between 0 and 1 (e.g. an evenness index in Nogueira, González-Troncoso, and Tolimieri (2016); or a straightness index in Shimada et al. (2016)). In addition, it can be applied to variables that are constrained to the interval *a* and *b* as they can be rescaled to [0, 1] through $(y - a)/(b - a)$.

Beta regression consists of the same three components as generalized linear models (GLMs) (Bolker et al., 2009; McCullagh & Nelder, 1989), and those familiar with GLM will recognize the most important aspects of beta regression (the distinction between the two arises from the non-orthogonality of the model parameters, see below). Here, we briefly review these three elements: the random component (the beta distribution and its implied mean–variance relationship), the systematic component (the linear predictor) and the link function (specifying the link between the random and systematic component). We refer readers to Ferrari and Cribari-Neto (2004) for a more comprehensive explanation.

Beta regression begins with the assumption that the data-generating process can reasonably be modelled by a beta probability distribution (Balakrishnan & Nevzorov, 2003). The beta distribution is a member of the exponential family (Kieschnick & McCullough, 2003), and is defined by two parameters for values on the open interval (0, 1). Two parameterizations for the beta distribution are available, but the mean-precision parameterization, with μ (for the expected value) and ϕ (as a measure of 'precision', or the inverse of dispersion), is most commonly used in the context of beta regression (see Box 1). The variance can be related to the mean (μ) by $\frac{\mu(1-\mu)}{1+\phi}$ and is therefore proportional to the variance of the binomial distribution for one trial, $\mu(1 - \mu)$, by a factor of $\frac{1}{1+\phi}$.

Depending on the choice of values for the two parameters a large range of shapes can be obtained including symmetrical, skewed, uniform, roughly bell-shaped and bimodal. This flexibility, combined with the limitation to values between 0 and 1, make the beta distribution a particularly useful model for continuous proportional data. In addition, fitting a beta distribution gives increasingly less biased estimates of the mean compared to transformation-based approaches when observations get closer to zero and one and/or their variance is large (see Appendix S1).

Once a beta distribution has been chosen, the next step is the specification of the systematic component of the model relating the expected value of the response variable to one or more (continuous or categorical) predictor variables. In the familiar case of ordinary linear regression this dependence is specified through the regression equation $\mu = E[Y|X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ where β are regression parameters to be estimated based on observed values of *y* and corresponding matrix *X* of *p* predictors. Given that such a linear predictor function can potentially range between $-\infty$ and $+\infty$ it cannot be used to specify the mean for distributions, such as the beta, that are restricted to a particular interval. For this reason, a link function must be specified to convert between the linear predictor model and the conditional mean on the scale of observations (Zuur et al., 2009). The model relating the values of the covariates, and the expected value of the response therefore becomes $g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, implying that $\mu = g^{-1}(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$, where $g(.)$ and $g^{-1}(.)$ are an appropriate link function and its inverse, respectively. In the case of beta regression, several link functions are potentially applicable, with the logit function (see Box 1) being the most common choice. The inverse of this function ensures that any value from the linear predictor will fall between 0 and 1. Appendix S2 presents a simple example of beta regression to make these concepts more concrete.

Estimates of β and ϕ that lead to a model that best fit the observed data are obtained by maximum likelihood estimation. The probability model (including covariates) and data are combined to define a likelihood function (see Bolker, 2008, for an accessible treatment), for which the maximum is obtain by numerical optimization, leading to maximum likelihood estimates for the parameters of the mean model (the βs) as well as for the precision parameter ϕ. The estimation procedure also produces values for the standard error of each parameter, allowing the application of the usual inference tools such as significance testing and confidence intervals. Here, it should be noted that for beta regression, the parameters β and ϕ are not orthogonal (Ferrari & Cribari-Neto, 2004), complicating estimation for β when ϕ is not known, or incorrectly specified. This non-orthogonality between the mean model parameters and the error parameter is also the reason why beta regression with unknown ϕ is not strictly a GLM (Huang & Rathouz, 2017). Moreover, two important properties of the beta distribution, not shared by common GLMs, are that the maximum likelihood estimate for μ can be different from the corresponding sample mean, in particular for small sample sizes; and that changes in the way that ϕ is modelled (see Section 4.1 below) can have consequences for the maximum likelihood estimates of μ. This can lead to bias in estimation and inference (see Section 5.1 and Case Study 1, below).

Although best suited for continuous proportions, beta regression has also been successfully employed to analyse count-based proportions in cases where the numbers of observed units is large (e.g Bennett, Nimmo, & Radford, 2014; Briand, Schwilk, Gauthier, & Bergeron, 2015). However, a few cautionary remarks should be made. The standard error of a sample proportion decreases with the number of trials, *n*, according to $\frac{\sqrt{p(1-p)}}{n}$. Thus, with a small number of trials random selection error of the trials may be important. In addition, when employing beta regression to count-based proportions one loses information on the number of trials that was used to calculate the proportion. Effectively, each proportion is given equal weight, which can be problematic if the number of trials varies across samples. Two potential alternatives in this case would be to apply 'beta-binomial regression' models (Skellam, 1948), or the use of an 'observation-level random intercept' (Harrison, 2015). These

BOX 1 Mathematical details of beta and Dirichlet regression

Beta regression

Definitions of the beta distribution usually employ a parameterization using the symbols α and β such that the probability density function for a beta-distributed response variable y is given by the following:

$$f(y|\alpha,\beta)=\frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}$$

where

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

And $\Gamma(.)$ is the gamma function.

The corresponding expectation and variance of the distribution are given by the following:

$$E[y]=\frac{\alpha}{\alpha+\beta}$$

$$\mathrm{var}[y]=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

When using the beta distribution for modelling data, it is usually more convenient to use an alternative parameterization with μ and ϕ, such that the expectation of the distribution is simply $E[y] = \mu$, and the variance is given by the following:

$$\mathrm{var}[y]=\frac{\mu(1-\mu)}{1+\phi}$$

In beta regression, the conditional model for the mean μ of the response given covariates X is usually assumed to be linear on the logit transformed scale:

$$y\sim Beta(\mu,\phi)$$

$$\mathrm{logit}(\mu)=\log\left(\frac{\mu}{1-\mu}\right)=\eta=X\beta$$

where η is known as the linear predictor, β is a vector of parameters to be estimated and X is the design matrix of covariate values. To obtain estimates from a beta regression fit that are interpretable on the scale of observations (0, 1) the values from the linear predictor therefore need to be back-transformed with the inverse logit function:

$$\mathrm{logit}^{-1}(\eta)=\frac{e^{\eta}}{1+e^{\eta}}$$

As discussed in the text, ϕ can either be estimated as a single value for all observations, or modelled as a function of covariates with design matrix Z, corresponding regression parameters γ, and linear predictor ζ – in which case a log link is appropriate:

BOX 1 (Continued)

$$\log(\phi)=\zeta=Z\gamma$$

Dirichlet regression

The traditional parameterization results in the following probability density function for a vector valued ***p***.

$$f(x_1,\ldots,x_C|\alpha_1,\ldots\ \alpha_C)=\frac{1}{B(\boldsymbol{\alpha})}\prod_{c=1}^{C}x_c^{(\alpha_c-1)}$$

where $\boldsymbol{x}$ and $\boldsymbol{\alpha}$ are both vectors (of observations and model parameters respectively) of length C, and $B()$ is the multinomial Beta function.

In this parameterization, there is a parameter αc for each of the C components. If we define α0 as the sum of all elements of α then the expected value of any given component xc is given by $E[x_c]=\frac{\alpha_c}{\alpha_0}$, with the associated variance:

$$\mathrm{var}[p_c]=\frac{\alpha_c(\alpha_0-\alpha_c)}{\alpha_0^2(\alpha_0+1)}$$

The value of α_0 is therefore interpretable as a precision parameter.

An alternative parameterization is realized by representing the expected values of each of the components as a vector $\boldsymbol{\mu}$ where each μ_c is between 0 and 1, and the sum of all elements of $\boldsymbol{\mu}$ is 1. Additionally, we define a precision parameter ϕ. Conversion between the two paramaterizations is therefore possible with $\alpha_c = \mu_c\phi$ and $\alpha_0 = \phi$, leading to the follow expressions for expected value and variance of each component:

$$E[p_c]=\mu_c$$

$$\mathrm{var}[p_c]=\frac{\mu_c(1-\mu_c)}{\phi+1}$$

The main contrast with the traditional parameterization is therefore that the group means and the precision parameter are explicitly modelled rather than indirectly in terms of the values of $\boldsymbol{\alpha}$. In the case of the traditional paramaterization, a regression model should be fitted for each of the C values of α_c. An appropriate link function for the corresponding regression model is the log-function.

$$\log(\alpha_c)=\eta_c=X_c\beta_c$$

Under the alternative paramaterization, both the expected values $\boldsymbol{\mu}$ and the precision parameter ϕ are modelled as a function of covariates. One of the c components is defined as the base category b (all regression coefficients = 0). The sum constraint implies that the appropriate link function for the regression models for μ is the multinomial logit function. In this way, given a separate linear predictor for each component $\eta_c = \mathbf{X}_c\beta_c$, these can be converted into the corresponding values of μ with the expressions:

$$\mu_c = \frac{e^{\eta_c}}{\sum_{a=1}^{C} e^{\eta_a}}$$

$$\mu_b = \frac{1}{\sum_{a=1}^{C} e^{\eta_a}}$$

Note that the expected value of the baseline component is implictly modelled through the models for the other $C - 1$ components. As for beta regression, a model for ϕ is typically specified with the log–link function.

methods are particularly suited to situations where count proportions show more variance than can be modelled by binomial regression, i.e. when count proportions are *overdispersed*. Overdispersion is more likely to occur when observations are grouped in space or time or when W becomes very large. In a beta-binomial regression model, the probability of success for a given level of the covariate is not fixed, but comes from a beta distribution. In the binomial model with observation-level random intercept each observation gets a random intercept. These random intercepts come from a Gaussian distribution at logit scale. A simulation study showed that in general a beta-binomial is preferred and lead to least biased estimates (Harrison, 2015).

3.2 | Analysing compositions with >2 continuous categories

3.2.1 | Historical approaches

Various approaches have historically being used to analyse continuous proportions with $C > 2$ categories. The most straightforward approach is to ignore the sum constraint of the proportions and model the proportions of each category separately (e.g. Poorter, Vijver, Boot, & Lambers, 1995). The second, more sophisticated, approach is to recognize the sum to one constraint of the C categories and transform the category proportions relative to the proportion of a reference category. The additive log-ratio, centred log-ratio and the isometric log-ratio transformations are commonly suggested (Aitchison, 1986; van den Boogaart & Tolosana-Delgado, 2013), although the latter two are, to the best of our knowledge, not often applied. The transformed values are subsequently modelled by assuming that they follow a multivariate normal distribution (Aitchison, 1986; Billheimer, Guttorp, & Fagan, 2001; van den Boogaart & Tolosana-Delgado, 2013). As with the logit and the arcsine transformations described above, these transformation procedures lead to issues arising from Jensen's inequality.

3.2.2 | Dirichlet regression

An extension of beta regression to cases where proportional data are distributed over more than two categories is provided by Dirichlet regression (Camargo, Stern, & Lauretto, 2012; Gueorguieva, Rosenheck, & Zelterman, 2008; Hijazi & Jernigan, 2009) – not to be confused with Dirichlet processes, or Dirichlet mixture models. Although less commonly used than beta regression, Dirichlet regression has also recently been applied to several data analysis problems in ecology. To give an idea of the breadth of data types amenable to this type of analysis: Regular et al. (2014) analysed time budgets of seabirds divided over four different activities; Acevedo-Trejos, Brandt, Merico, and Smith (2013) modelled the relative proportions of different phytoplankton size fractions to total biomass as environmental data; and Sánchez and Dos Santos (2015) examined spatial patterns in the dietary compositions of two species of bat.

The Dirichlet distribution (Box 1; Balakrishnan & Nevzorov, 2003) is a generalization of the beta distribution to any number of categories i.e. it models a vector-valued observation $[p_1, p_2, ..., p_C]$ subject to the constraint $\sum_{c=1}^{C} p_c = 1$, that the sum of all categories must equal unity.

As with beta regression, in practice two alternative parameterizations are used in Dirichlet regression. They differ mostly in whether the variance of the categories is modelled explicitly or not, and lead to quite different interpretations of model parameters after estimation. The so-called 'alternative' parameterization (Maier, 2014, see Box 1) is probably most useful for ecologists and analogous to the μ and ϕ (mean and precision) parameterization for the beta distribution.

Under the alternative parameterization, the vector of expected values $\boldsymbol{\mu}$ is modelled as a function of covariates, and a precision parameter ϕ is also estimated from the data. Since the values of $\boldsymbol{\mu}$ are constrained to sum to 1, a separate model for each category is overdetermined so in practice only $C - 1$ models are fitted, and the C th category is treated as a baseline category and modelled implicitly as the 'residual' category remaining after the others are accounted for. Importantly, there is no requirement to use the same covariates for each category. Given the strict sum constraint the multinomial logit function is used as a link function to convert between the linear predictors and the vector $\boldsymbol{\mu}$ (Box 1). For the precision parameter ϕ, which is strictly positive, an appropriate link is the log function.

Dirichlet regression uses maximum likelihood estimation to determine the values of the parameters $\boldsymbol{\beta}$ and ϕ that best fit the observed data $\boldsymbol{p}$. The dependence of any given μ_c on the predicted values of the other μ's makes direct interpretation of $\boldsymbol{\beta}$ difficult. In particular, and somewhat counterintuitively, individual categories can show negative relationships with covariates on the proportional scale even when the corresponding best-fit regression parameter for that category is positive. The best method for interpreting the results of Dirichlet regression is therefore to produce plots of predicted values under a range of covariate values that are relevant to the question of interest (see Case Study 2, Figure 4).

4 | EXTENSIONS TO DIRICHLET AND BETA REGRESSION

4.1 | Independent model for precision ϕ in beta and Dirichlet regression

In many situations, it is useful to allow for the possibility that precision varies across observations as a function of one or more covariates. Such a variable ϕ model may be appropriate when, for example, certain treatment levels display more variance at a given predicted mean (Case Study 1), or if there is systematic increase or decrease in variance as one of the covariates changes. It also allows for modelling situations where the variance of a response changes without any systematic change in the mean value. An additional use is for cases when differences in variability are a result of the sampling or experimental design, i.e. in cases where the observation interval cannot be fixed a priori. For example, in animal behavioural studies, the observation interval may be determined by how long the subject is within sight. In such cases, the precision of the observation can be expected to partly depend on the scale of the observation unit. Intuitively, an observation of 50% of time spent on a certain activity has a different evidential value when it is based on a 20 s total observation time, when compared to the same observed proportion derived from a 5 min observation interval. In such a situation, it may be appropriate to incorporate information about the length (or size) of the observation unit by including it in a model for the precision term ϕ.

Fitting beta and Dirichlet regression models with variable ϕ parameters is accomplished by extending the model and associated likelihood function to include dependence of ϕ on pre-specified covariates (Simas, Barreto-Souza, & Rocha, 2010; Smithson & Verkuilen, 2006). Since ϕ is defined to be always positive, a link function (usually the log function) is used to relate the continuous linear predictor to the ϕ scale. The fitting procedure will return maximum likelihood estimates and associated standard errors for the parameters of the model for ϕ which can be interpreted in the same way as the parameters for μ, after taking into account the different link functions.

4.2 | Hierarchical data structures

Many experimental designs in ecology and evolution lead to observations that are grouped in some way. Common examples are multiple observations within an experimental plot or repeated observations of the same experimental unit through time. Such experimental designs lead to non-independence of observations, violating assumptions of most statistical techniques and potentially leading to incorrect inference. Mixed effect models can account for non-independence of observations and can be implemented as extensions to Generalized Linear Models (Bolker, 2015; Zuur et al., 2009) and have been developed for beta distributed variables as well (Brooks et al., 2017). Case study 1 includes a section on the fitting and interpretation of such a mixed-effects model in the context of beta regression. At the time of writing, mixed effect models for Dirichlet regression have not yet been implemented in standardized software to the best of our knowledge, although see Regular et al. (2014) for an example of a customized analysis using Markov Chain Monte Carlo techniques to analyse seabird time budgets.

5 | ISSUES WITH BETA AND DIRICHLET REGRESSION

5.1 | Bias in estimation

Beta and Dirichlet regression use maximum likelihood methods for parameter estimation. However, in many cases, the methods are known to lead to bias – a systematic deviation of an estimated parameter value from the true value – particularly in cases with small sample size (Firth, 1993). Biased estimators can lead to erroneous inference, and methods for bias–reduction and bias–correction are an active research area in statistics and in beta regression in particular (Grün et al., 2012; Kosmidis, 2014; Kosmidis & Firth, 2010; Ospina, Cribari-Neto, & Vasconcellos, 2006; Simas et al., 2010). In particular, the precision parameter in beta regression models is prone to overestimation bias (Kosmidis & Firth, 2010) which leads directly to underestimation of the width of confidence intervals for other model parameters. Two main types of solutions are available to reduce bias: bias correction and bias reduction (Firth, 1993). Bias correction methods correct for bias in a separate step following maximum likelihood estimation, while bias reduction methods modify the maximum likelihood estimation procedure such that the resulting estimator is less biased. See Appendix S3 accompanying Case study 1 for a demonstration of bias correction and bias reduction and a bootstrap-based technique for assessing the degree of bias for any given data-model combination (Kosmidis, 2014).

5.2 | Dealing with 0 and 1 in observations

The proportions modelled by the beta and Dirichlet distributions are defined on the interval (0, 1). However, for some combinations of parameters, the probability density function is zero or infinity at the boundaries, which precludes a meaningful calculation of the likelihood. Zeros and ones may occur in the data because the true (non-zero) value is below the detection limit of the measurement device or method. Conversely, true zeros may occur when a given category is absent from the sample e.g. no vegetation is present in a sampled quadrat. Regardless of their source, observations of zero or one will lead to a failure of the fitting algorithms in both beta and Dirichlet regression. Note that this problem is not unique to these two techniques – both logit or additive log-ratio transformations cannot be applied to datasets containing zeros and ones.

Here, we focus on the case of observations of zero, but the same advice and techniques can be applied to datasets containing observations of 1, or both. Several solutions are available depending on the origin of the zero. When the zero arises because of the detection limit of the observation method, a simple workaround is to replace all observed zeros with a small term prior to computing p for each sample, taking care to include ε in the new demoninator. ε can be chosen equal to the detection limit or to the smallest non-zero observation. Warton and Hui (2011) recommend exploring the sensitivity of results to the value of ε.

Alternatively, the data can be transformed according to the following equation:

$$p^{*}=\frac{p(n-1)+\frac{1}{C}}{n} \tag{1}$$

with p being the proportion of a category, n the total number of observations in the dataset, and C the number of categories (Maier, 2014; Smithson & Verkuilen, 2006).

Note that for a fixed value of n this is a linear transformation, and does not lead to issues arising from Jensen's inequality.

In case of exact zeros, as an alternative to the transformations above, it is possible to apply the so-called zero-inflated beta regression (Fang & Kong, 2015; Ospina & Ferrari, 2012), perhaps more properly referred to as zero-augmented beta regression (Wright, Irvine, Warren, & Barnett, 2017), given the independence of the two processes. This type of regression assumes that the data-generating process involves two linked stochastic processes: first a Bernoulli process (with or without covariates) describes the probability of observing a non-zero; and subsequently a beta regression model is specified for the value of the proportion itself for all non-zero observations. It can therefore be thought of as a special case of two-component finite mixture modeling. This approach is also available for one-augmented beta regression (Ospina & Ferrari, 2012) or zero-and-one augmented beta regression (Fang & Kong, 2015). See Joseph, Preston, and Johnson (2016) for an application of zero-one-augmented beta regression to vegetation cover, and Wright et al. (2017) for a further extension of zero-augmented beta regression that leverages repeated observations to separately model true absences from apparent (observation-error related) absences in vegetation survey data. A somewhat different method, involving modification of the likelihood function, has recently been proposed for zero-augmented Dirichlet regression (Tsagris & Stewart, 2018).

The added value of a zero-inflated/augmented models will be larger when both zeros and relatively high proportions are observed within replicates of the same treatment – evidence that two data-generating processes are operating. Moreover, the separation of the data-generating process into two independent components allows for additional inference about processes underlying presence/absence, as distinct from processes determining the observed proportions (see e.g. Keim, DeWitt, Fitzpatrick, & Jenni, 2017; Wright et al., 2017). Given these advantages, and the relative ease of working with zero-augmented models within existing software packages (see Table 1), we suggest that zero/one-augmented models should in general be used whenever there is a reasonable a priori expectation of zero or one values in the dataset, e.g. for species cover data. On the other hand, for certain data types (e.g. biomass partitioning in Case Study 2, below), zero or one values in the dataset will be absent in most cases.

Since the application of regular beta regression to data with zeros (and/or ones) requires transformation of the data, formal model selection criteria such as AIC or Bayesian Information Criterion (BIC) cannot be applied to compare the fit of a beta regression model fitted to a transformed response to zero-and/or-one inflated beta regression fit to an untransformed response. Therefore, model selection needs to be based on other criteria such as visual inspection of residuals and comparison of model predictions with observations. In case study 1, we compare the conclusions drawn from beta regression on transformed variables, and zero-augmented beta regression (see also Appendix S3).

6 | MODEL INFERENCE AND EXAMPLE ANALYSES

Below, we present three examples of analysing continuous proportions as an illustration of the methods discussed above and to demonstrate the major steps in applying these techniques for inference. We have chosen two existing datasets to represent two commonly arising forms of continuous proportions: fractional cover (case study 1) and biomass partitioning among plant organs (case study 2). In Appendices S4 and S5, we use these case studies to provide detailed step-by-step demonstrations of all variations of beta and Dirichlet regression discussed in this paper using the popular statistical software package R. A number of R packages with which continuous and count-based proportions can be modelled are listed in Table 1. In addition, we use a simulation approach to compare transformation-based analyses with beta regression, and illustrate the effects of varying link functions (case study 3 and Appendix S5).

The steps to be taken to fit models to continuous proportions are similar as for any other type of regression analysis (Zuur & Ieno, 2016; Zuur et al., 2009). Here, we highlight a few points that warrant particular attention for analyses of this type.

In the data exploration phase, it is advisable to explore how the variation in the proportions vary as function of the covariates. An appropriate model for ϕ can improve estimates of the other model parameters. Additionally, the choice of the link function may affect model fit when at least one of the predictors is continuous.

Alternative link functions to the logit are possible, and in theory, any function which is invertible and that maps the unbounded linear predictor to the appropriate domain of the corresponding parameter ((0, 1) for μ and >0 for ϕ) could be used. In case of a continuous covariate, we recommend testing several link functions (see case study 3). Alternative link functions for μ besides the standard logit are the probit (inverse of the cumulative distribution function of the standard normal distribution), the complementary log–log (clog–log(θ) = log(−log(1 − p))), and the Cauchit function (inverse of the cumulative distribution function of the Cauchy distribution). Link functions that do not map the real line to (0, 1), such as the log or identity link, can also be used, albeit with care (Marschner & Gillett, 2011). For example, the log link is commonly used in binomial GLM to assess relative risks, and can be used in a beta regression setting as well. However, a log link can lead to fitting problems when the log-likelihood function is maximized near the boundary of the parameter space, e.g log(μ) = $\beta X \approx 0$, and may be practically impossible when

TABLE 1 Examples of R packages that implement models for the types of proportional data mentioned in Figure 1. The most commonly used packages with user–friendly interfaces are mentioned. Other, more general purpose modelling frameworks can also be used, but may require customized code

Type of model	Count-based	Non-count based	2 categories	>2 categories	Overdispersion	Grouping	Zero/one augmentation	R packages
GLM with binomial error model	✓		✓					stats, brms
Beta-binomial model	✓		✓		✓			bblme, brms
GLMM with binomial error model	✓		✓			✓		lme4, glmmTMB, MCMCglmm, glmmADMB, brms
Beta binomial GLMM	✓		✓		✓	✓		glmmTMB, brms
GLM with multinomial error model	✓			✓				brglm[a], mlogit, nnet
Dirichlet-Multinomial model	✓			✓	✓			dirmult[a]
Dirichlet-Multinomial mixed effect model	✓			✓	✓	✓		[b]
Zero-, one-augmented Dirichlet-Multinomial mixed effect model	✓			✓	✓	✓	✓	[b]
Beta regression		✓	✓					betareg, brms
Beta regression variable phi		✓	✓		✓			betareg, brms
Mixed effect beta regression		✓	✓			✓		glmmTMB, glmmADMB[c], brms
Mixed effect beta regression variable phi		✓	✓		✓	✓		glmmTMB, brms
Zero-, one-augmented mixed effect beta regression		✓	✓		✓	✓	✓	glmmTMB[d], glmmADMB, brms, zoib
Dirichlet regression		✓		✓				DirichletReg
Dirichlet regression variable phi				✓	✓			DirichletReg
Zero-, one-augmented Dirichlet mixed effect model				✓	✓	✓	✓	[b]

[a] Not allowing for covariates.

[b] Not implemented in R to the best of our knowledge.

[c] Not allowing for variable ϕ.

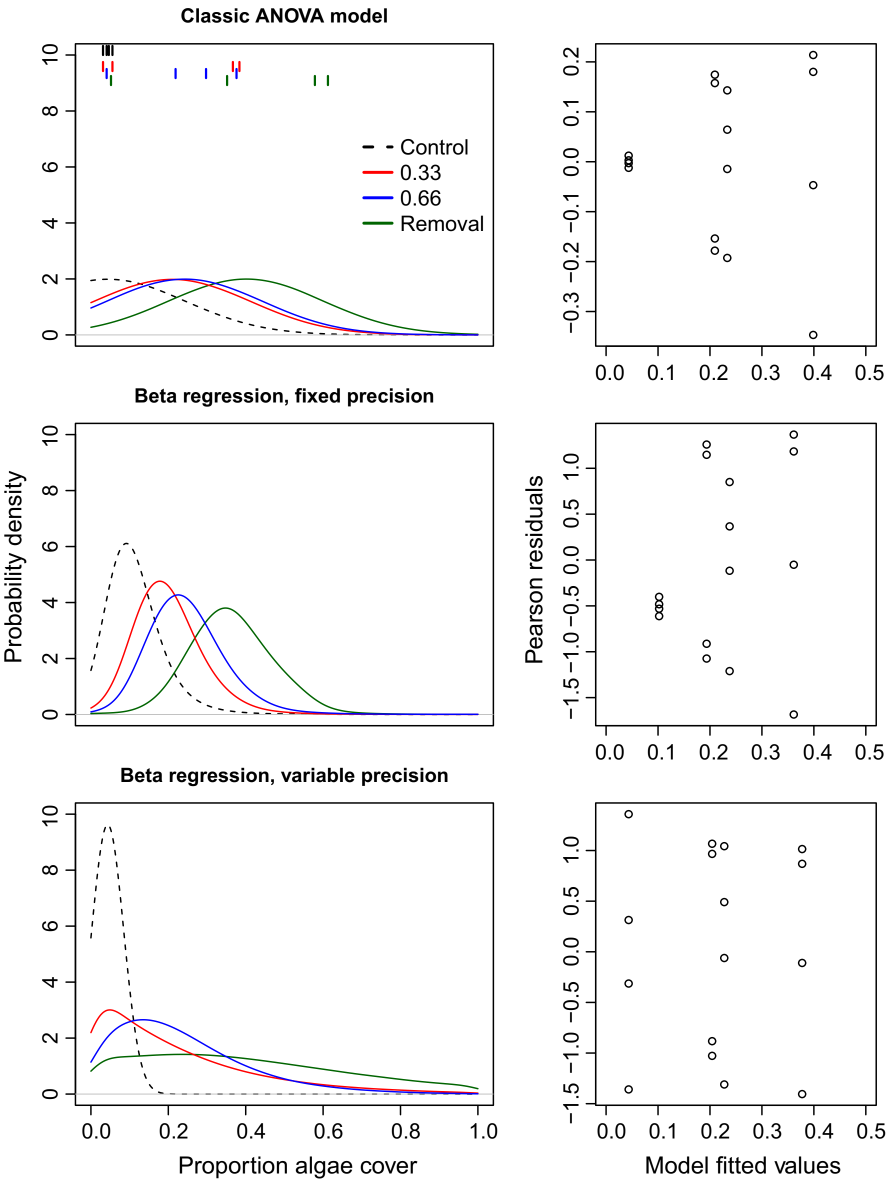
[d] Not allowing for one-augmentation.

using Bayesian MCMC methods. Similarly, even when stable solutions are obtained, confidence and prediction intervals may include non-sensical parameter values. In some contexts, these issues can be avoided by reparameterization of the model-data combination, see Case Study 3 below for an example.

Standard model selection criteria such as the likelihood ratio test (LRT), AIC or BIC can be used to compare among models with different link functions or variable ϕ, although these will be most reliable at larger sample sizes, so visual assessment of model fits should also be carried out for smaller datasets.

It is also recommended to determine whether the data contains a large number of zeros/ones, and if their presence in the dataset varies systematically with potential predictor variables. If this is the case, a zero-and/or-one augmented beta regression model may be more appropriate than transformations that remove zero or one observations from the dataset.

Once a model has been fitted, inspecting the standardized residuals may help in assessing any remaining pattern in the data that were not captured by the covariates. It is important to avoid inspecting raw residuals on the response scale, since the expected variance of observations is related to the fitted response. Plots of standardized residuals (e.g. Pearson) against fitted values, and/or available covariates should ideally not show any systematic pattern in either spread or location. In particular, a systematic pattern of variation in the spread of residuals along the range fitted values or covariates indicates the need for a separate model for the precision parameter ϕ (see e.g. Figure 2 in Case Study 1 below). For beta regression, a method of computing residuals that account for observation leverage has been proposed, which can more clearly identify atypical observations compared to the common standardized residuals (Espinheira, Ferrari, & Cribari-Neto, 2008, and see Appendix S3). Calculation of these residuals is the default in the betareg R


Classic ANOVA model

Beta regression, fixed precision

Beta regression, variable precision

FIGURE 2 Probability density plots of posterior predictions of proportion algae cover under each grazer removal treatment, and Pearson residual plots for three different modeling approaches. Top row: classical ANOVA model, middle row: beta regression with fixed precision across all levels of treatment; bottom: beta regression with variable precision. Vertical lines (|) in the uppermost panel represent the observed data, vertically staggered by treatment for legibility. Pearson residuals are presented to allow comparison between classical and beta regression models. The alternative weighted residuals advocated by Espinheira et al. (2008) are to be preferred when making comparisons among beta regression specifications. Note that data were pooled per patch and transformed according to Equation (1) prior to analyses, see Figure 3 for an analysis that accounts for nested structure and zero-observations

package, but is not universally implemented in more general modeling packages. See Espinheira et al. (2008) for a range of expressions for the calculation of standardized residuals, and a detailed discussion of their relative merits. To our knowledge, a comparable analysis of residuals has not yet been undertaken for Dirichlet regression.

Influential observations can be identified using measures such as generalized leverage or Cook's distance (for beta regression see Ferrari & Cribari-Neto, 2004). These measures can be applied in the same manner as for the classical linear model; i.e. to identify specific observations that substantially change the model fit as candidates for further examination or exclusion from the dataset.

To further assess the fit of the model we recommend plotting the model predictions, either as posterior predictive densities, or simulations from the model, and comparing them with plots of the observed data. This can identify aspects of the original data that are inadequately captured by the model. This is particularly useful for Dirichlet regression where inspection of the parameter estimates themselves may not be very insightful because other categories are most likely also changing as a function of a given covariate.

6.1 | Case Study 1 Percent cover in quadrats

The first example involves experimental manipulation of the density of the sea urchin *Centrosthepanus rodgersii* to investigate its effect of grazing on the colonization of filamentous algae (Andrew & Underwood, 1993). Algae colonization was measured by percent cover in five 0.25 m^2 quadrats randomly located within larger patches subject to one of four levels of grazer removal treatment. Andrew and Underwood (1993) analysed this data (reanalysed by Quinn & Keough, 2002) using a nested ANOVA to account for the within-patch replication. They concluded that treatments did not significantly affect the cover of filamentous algae.

In Appendix S3, we provide a detailed, step-by-step analysis of this dataset using different versions of beta regression, with accompanying code. We approach the analysis in two ways: first to illustrate the basic ideas we compare classical ANOVA to beta regression with and without a model for varying precision. To focus on the comparison of these basic model types, and the role of residuals in diagnostics, we use data pooled per patch and transform the data according to Equation (1) to initially avoid issues with nested observations, and the presence of zeroes, respectively. Secondly, we perform the analysis on the original data, retaining the hierarchical structure and comparing the results of models with and without zero-augmentation. This approach is what we would recommend in a ‘real-world’ analysis, since it incorporates *a priori* information about the presence of zeros and the structure of the experimental design.

Figure 2 and Table 2 compare the results of classical ANOVA (assuming normal errors), beta regression with a fixed ϕ, and beta regression with ϕ dependent on removal treatment. Model selection based on AIC clearly favours the variable ϕ model (Table 2). The improved fit is also evident from comparison of the residual plots (Figure 2, second column) – the first two models show a strong relationship between values of the standardized residuals and the fitted values. This is due to overestimation of the amount of variance in the control treatment plots, a fact also visible when comparing the posterior predictive densities of the first two models with the observed data (Figure 2, first column). Interestingly, there is not a large difference in the estimates for the mean of each group (Table 2), but the classical ANOVA model has much broader confidence intervals (leading to non-significant pairwise comparisons, see Appendix S3) than the beta regression, and moreover predicts values outside the possible range of (0, 1). Pairwise comparisons of the groups based on the variable ϕ beta regression model indicated that the control treatment differed significantly from the other treatments, but the other treatments did not differ significantly from each other, both in terms of mean response and precision.

To illustrate the use of mixed-effects and zero-augmented beta regression, we analyse the original dataset in which replicate quadrats (observational units) are observed within each patch (experimental units). Given the non-independence of quadrats within each patch a mixed-effects model is fit with patch as the grouping variable and ϕ dependent on treatment. The predicted distributions for each treatment (Figure 3a) are broadly similar to those obtained from the variable ϕ model on the pooled data (Figure 2), however, the higher number of data points seems to have increased the precision of the estimates for the non-control treatments. Incorporating a zero-augmented component to the model, where the probability of observing a zero is modelled as a function of removal treatment, leads to only slightly adjusted posterior predictions for the mean of each group (Figure 3b,c). As expected, the main added value of the zero-augmented model is a much more accurate prediction of zero observations in the dataset (Figure 3d). The inability of models without zero-augmentation to reproduce this important feature of the dataset would limit their usefulness for making further predictions.

As for the analysis of the pooled data, the conclusions from both the mixed-effects and zero-augmented mixed effects models are that any form of any degree of sea urchin removal from this environment leads to an increase in algal cover. This finding contrasts with the conclusions in the original analyses of Andrew and Underwood (1993) and the reanalysis by Quinn and Keough (2002), both of which concluded that there was no significant difference in percentage cover of filamentous algae between treatments. We would argue that by choosing a more realistic model for the response variable, allowing the dispersion to vary between treatments, and explicitly modelling the occurence of zeroes, a beta regression model better captures the features of the dataset, and therefore provides a more reliable basis for inference.

6.2 | Case study 2 Biomass partitioning in plants

The second dataset comes from a study testing whether differences in growth parameters between fast- and slow-growing plant species at optimal nitrogen supply persisted at low nitrogen supply (Poorter & Sack, 2012; Poorter et al., 1995). Two species, *Deschampsia flexuosa* (slow growing) and *Holcus lanatus*

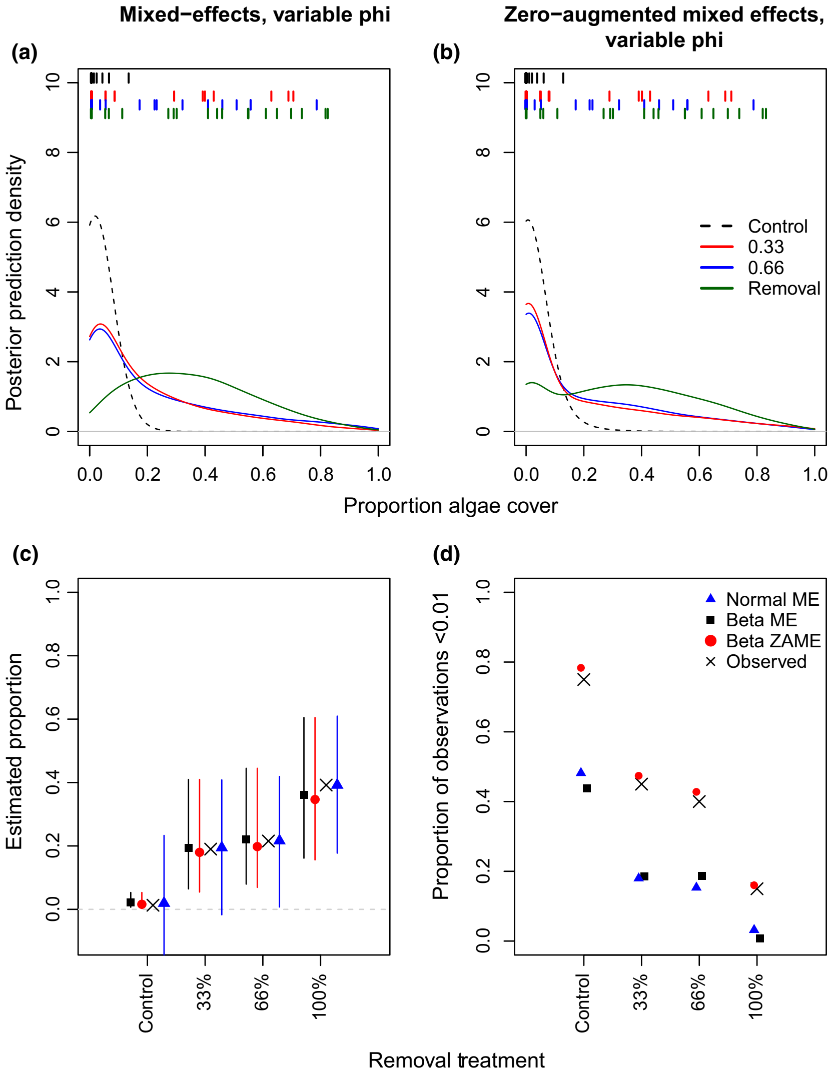


FIGURE 3 Summary of models fitted to nested data. Top row: posterior predictive distributions according per removal treament derived from (a) mixed effects beta regression model (ME), and (b) zero-augmented hierarchical beta regression (ZAME). In both top panels vertical lines (|) represent the observed data, vertically staggered by treatment for visibility. Bottom row: (c) estimates and 95% credible intervals for μ per removal treatment for ME and ZAME models, with observed means marked for reference and predictions from a normal mixed-effects model (normal ME) added for comparison; (d) posterior predictive estimates of the proportion of observed zeroes (effectively <0.01 due to transformation in the mixed model case) for each model and observed proportions

(fast growing) were grown under low and high nitrate supply for a maximum of 49 days. Replicate individuals were harvested at regular intervals for determination of biomass in roots, stems and leaves. This case study is an illustration of proportions that arise in a situation where the size of the observational unit (total biomass) is not fixed, and where there are more than two continuous categories (biomass of stems, leaves and roots). Dirichlet regression (Appendix S4) was used, as the generalization of beta regression for situations where proportions are calculated for more than two categories.

The response variables were vectors of the proportions of total plant biomass in leaves (LMF), roots (RMF) and stems (SMF).

TABLE 2 Parameter estimates (and their associated 95% confidence intervals) of the beta regression model explaining proportion algae cover averaged within patches by treatment (with bias reduction). Three different models were tested: 1) the mean cover not dependent on treatment and with fixed precision, 2) mean proportion cover dependent on treatment and fixed precision, and 3) mean proportion cover and precision dependent on treatment. The last model was the most parsimonious model based on AIC. Estimates and intervals are reported on the original scale of the observations

Model	Control	33% removal	66% removal	Removal	AIC
Classical ANOVA	0.04	0.21	0.23	0.40	−4.7
	(−0.15–0.24)	(0.02–0.40)	(0.04–0.43)	(0.21–0.59)	
Beta regression, fixed ϕ	0.10	0.19	0.23	0.36	−13.2
	(0.04–0.24)	(0.09–0.36)	(0.12–0.42)	(0.21–0.55)	
Beta regression, variable ϕ	0.04	0.20	0.23	0.38	−21.6
	(0.04–0.05)	(0.09–0.41)	(0.12–0.39)	(0.19–0.61)	

TABLE 3 Maximum likelihood parameter estimates and their associated standard errors (in parentheses) of the Dirichlet regression model explaining leaf mass fraction (LMF), root mass fraction (LMF) and stem mass fraction (LMF). Significant parameters (p < 0.05) are shown in bold. The most parsimonious model based on AIC is presented. Log-likelihood 1990, n = 500, 25 parameters estimated, AIC-3930, logit link on mean models, log link on precision models. S, T and D refer to Species, Treatment and Day respectively

Component	Intercept	Species (*H. lanatus*)	Treatment (low)	Day (scaled)	Day (scaled)^2	Total biomass	S × T	S × D	T × D	S × T × D
LMF	–	–	–	–	–	–	–	–	–	–
RMF	**−0.914**	**0.210**	0.05	0.03	**−0.03**	0.03	**−0.160**	−0.023	−0.004	**0.143**
	(0.02)	**(0.05)**	(0.03)	(0.03)	**(0.02)**	(0.02)	**(0.06)**	(0.06)	(0.04)	**(0.07)**
SMF	**−0.354**	**0.397**	**0.628**	**−0.057**	**−0.057**	**0.072**	−0.062	−0.004	**0.276**	**−0.267**
	(0.02)	**(0.04)**	**(0.03)**	**(0.03)**	**(0.01)**	**(0.01)**	(0.05)	(0.05)	**(0.03)**	**(0.05)**
Precision	**5.429**	**−0.277**	**−0.478**	**0.130**				**0.311**		
	(0.08)	**(0.09)**	**(0.10)**	**(0.06)**				**(0.31)**		

These proportions were modelled as function of species identity and nitrate levels. In contrast to Poorter et al. (1995), we included time as a covariate to investigate the temporal dynamics of biomass partitioning. We refer to Poorter and Sack (2012) for other options regarding the analysis of biomass fractions. The most parsimonious model (based on AIC) explained the mean RMF and SMF as a function of time, a quadratic transformation of time, species, nitrate supply, total biomass, the interactions between species and nitrate supply, species and time, nitrate supply and time, and a three way interaction between species, nitrate supply and time. In addition, the precision was modelled as a function of species, nitrate supply, time and the interaction between species and time (Table 3).

How the different fixed effects determine LMF, SMF and RMF is difficult to infer from direct inspection of the parameter estimates of the best fitted model because the different fractions are interrelated. We therefore displayed the predicted values of this model in Figure 4. The predicted fractions show that in the high nitrate supply treatment, both species changed their allocation to shoots, roots and leafs in a similar fashion, while under low nitrate supply *H. lanatus* and *D. flexuosa* allocate root and shoot biomass differently over time. This explains the significant three way interaction between

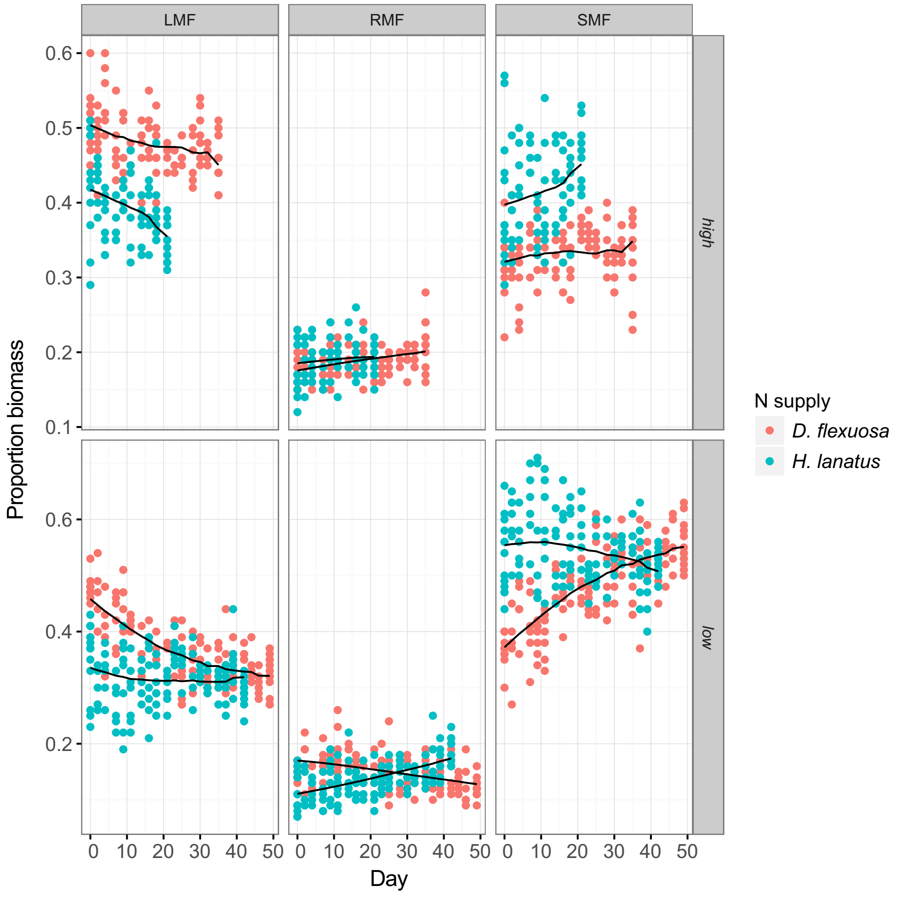


FIGURE 4 The observed (dots) and predicted (lines) proportion of biomass invested into leaves (LMF), roots (RMF) and stems (SMF) over time for *Deschampsia flexuosa* (red) and *Holcus lanatus* (blue) for two levels of nitrogen supply (rows; high and low). The predicted values come from a Dirichlet regression model and are computed for a smoothed average plant biomass at each sampling date (see main text for details)

time, species and nitrate supply. Excluding the species term led to an increase in AIC of 412, emphasizing the importance of species–specific effects on biomass partitioning. No obvious pattern remained in Pearson residuals when plotted against the fitted values or the covariates.

Another way to gain insight into the species effect is to compute the ratio of organ biomass of the two species. This ratio can be calculated from the predicted biomass fractions of the Dirichlet regression model, and can be thought of as a measure of effect size expressing the relative partitioning of biomass invested in leaves, stems and roots in *H. lanatus* compared to *D. flexuosa* (Figure 5). For example, early on in development, it is predicted that *H. lanatus* invest up to 1.5 times more in roots than *D. flexuosa*, while at harvest the investment in roots is similar. Despite this pattern, the 95% prediction interval of the investment ratios, taking the variation of individual replicates into account, shows that the variation in investment ratio within species is substantial.

6.3 | Case study 3 Percent cover and comparison of beta regression and transformations-based approaches

In this case study, beta regression and transformation-based approaches are compared to illustrate the mismatch between observations and predictions that can arise when using transformation-based approaches or when choosing an inappropriate link function within beta regression. A synthetic dataset was used to compare the performance of various approaches against true ground cover.

We created a dataset of tree cover using a stochastic, two dimensional spatial model where tree density is modeled as a function of mean annual precipitation (mm/year) following findings of (Hirota, Holmgren, Nes, & Scheffer, 2011; Staver, Archibald, & Levin, 2011). Importantly, the underlying data-generating process is not directly related to any of the model specifications we are comparing. Projected ground-cover of a range of 20 forests was simulated that varied in the mean annual rainfall received (ranging from 125 to 2,500 mm/year). Trees were positioned randomly in the area by drawing coordinates from a uniform distribution within the grid, and the size of the (circular) individuals was simulated by sampling values of crown diameter from a lognormal distribution. Percent cover on 1 ha plots was ‘estimated’ by simulating 15 randomly positioned non-overlapping quadrats of 10 × 10 m^2 (Figure 6a and Appendix S5 for details). We then used these simulated samples to estimate the relationship between mean percent cover and mean annual precipitation using one of five methods: log transformation or logit transformation followed by an ordinary least squares linear regression model, or a beta regression with either a cloglog, logit, or log link. To avoid fitting problems in the beta model with log link, we fitted a regression model on the proportion of non-cover (i.e. 1 − cover) and set the intercept at 1. This

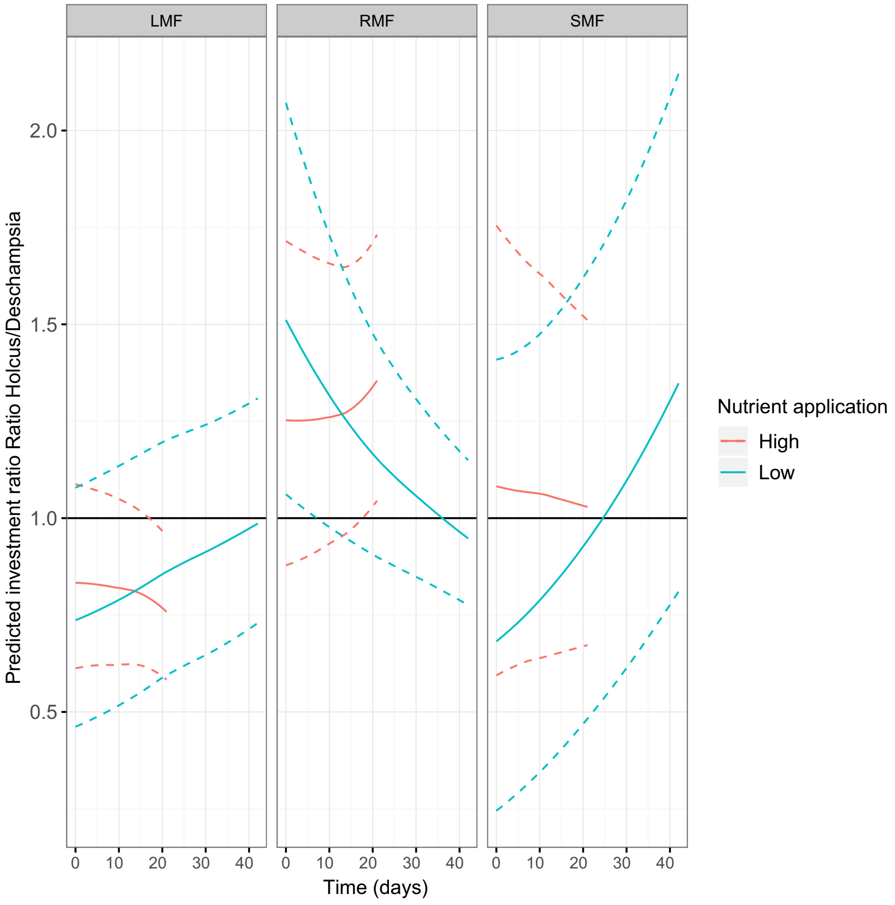


FIGURE 5 The predicted ratio of the proportion of biomass invested by *Holcus lanatus* in leaves (LMF), roots (RMF) and shoots (SMF) compared to *Deschampsia flexuosa* (solid line). The colors represent the two levels of nutrient application. The dotted lines represent the 95% prediction interval of the investment ratio of future observations

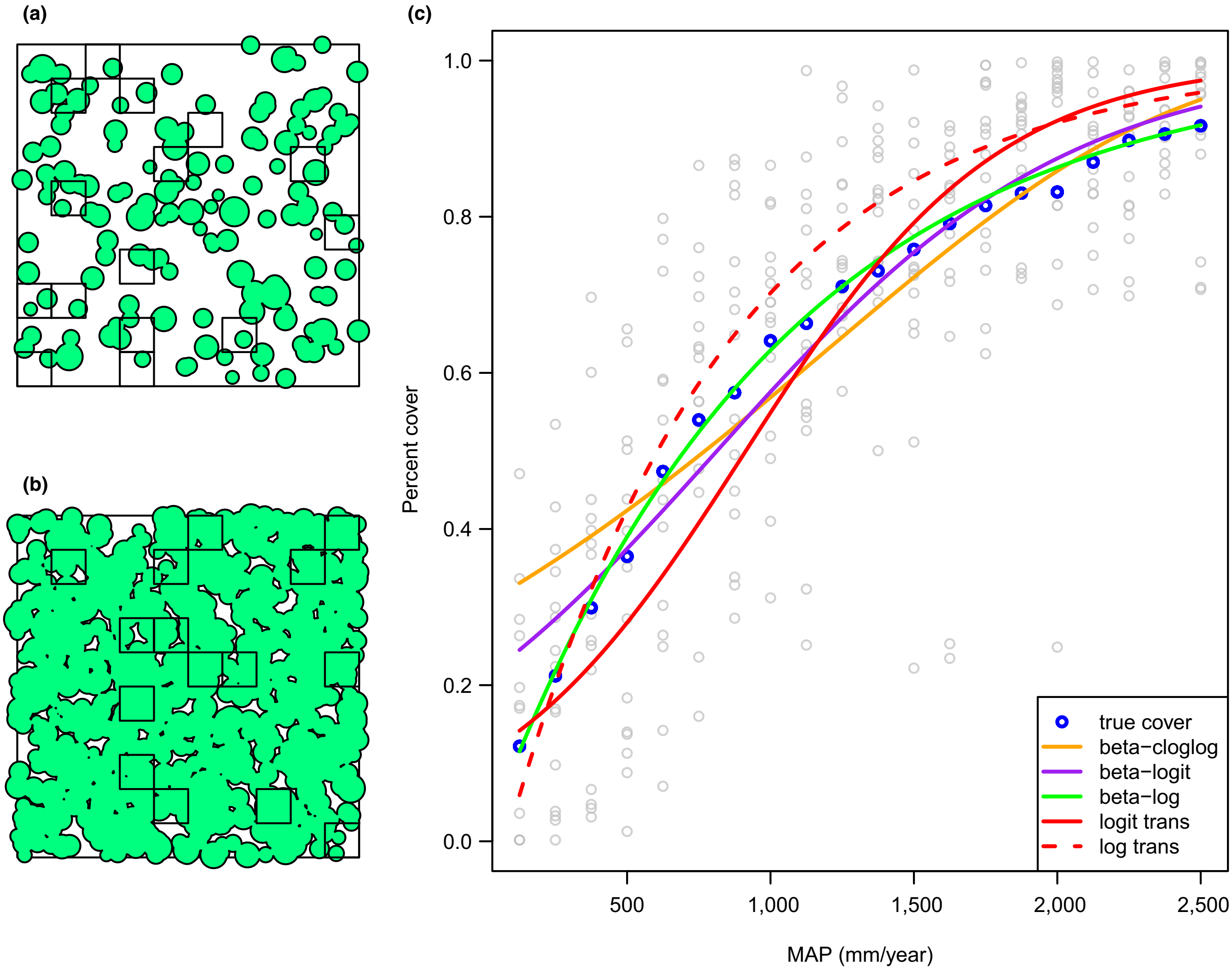


FIGURE 6 Two simulated types of forest with low and high tree project ground cover (green blobs) and 15 quadrats positioned in the area (squares 10 × 10 m^2; panel a & b). The relationship between mean annual precipitation and the projected ground cover of trees as measured in quadrats (panel c; grey dots), and over the whole area (blue dots). Lines represent the fitted relationships between number of individuals and percent cover with beta regression using three different link functions (cloglog (orange), logit (purple), log (green)), and a normal regression model on logit transformed data (red) and log transformed data (red dotted). The beta regression model with log-link fit the data best

constrains the model to return zero cover at values of zero annual precipitation, which is biologically plausible in this case.

The beta regression with log link best fitted the data (Figure 6). The root mean squared error (RMSE) of the model predictions versus true tree cover was on average 0.058, 0.073 and 0.099 for the log, logit and cloglog link respectively, and 0.069 and 0.663 for the logit and log transformation respectively (average of 50 simulations). In all cases, the RMSE decreased with increasing quadrat size while keeping the area sampled constant to 16% of the total area. However, the RMSE for the logit transformation decreased much faster compared to the beta regression models, but always stayed above the RMSE of the beta regression model with log link. The stronger reduction in RMSE with increasing quadrat size compared to the decrease in RMSE in beta regression models illustrates the point that larger variance in the observed proportions increases the mismatch between observations and the predictions of transformation-based approaches (see Appendix S5). Furthermore, the choice of the link function substantially affected model fit (Figure 6). Thus, the simulations show that beta regression is better able to predict tree cover compared to transformation-based approaches, provided that the link function is chosen carefully.

7 | CONCLUSIONS

Given the high prevalence of proportional data in ecology and evolution, appropriate techniques for their statistical modelling are an important component of the methodological toolbox of the modern biologist. When proportional data are derived from continuous measurements, beta (2 categories) or Dirichlet (>2 categories) regression can be used for modelling and inference, and avoid some issues related to bias and interpretation that arise when using traditional transformation-based techniques. Extensions to basic beta regression in the last decade such as variable ϕ models, bias correction, hierarchical models and zero-one augmented models, mean that most commonly encountered data structures can now be effectively analysed with these techniques. Further gains could be made if these techniques are implemented in the Dirichlet regression framework. We encourage scientists in the ecological research community to adopt these methods for their own analyses.

ACKNOWLEDGEMENTS

J.C.D. and J.T.W. were both supported by NWO project no. 863.14.018 and 016.171.089 respectively. J.T.W. was also partially supported by a postdoctoral fellowship from the Flemish Science Foundation (FWO). We thank Hendrik Poorter for providing the data on the biomass fractions, F. Bongers and R. Bevans for constructive criticism on a draft of this article, M. Joseph, the associate editor and two anonymous reviewers for valuable feedback on an earlier submitted version and Zishen Wang for providing a Chinese abstract.

AUTHORS' CONTRIBUTION

Both authors contributed equally to all aspects of the work presented in this paper.

DATA AVAILABILITY STATEMENT

Data and source files are available on https://doi.org/10.5281/zenodo.3234670

ORCID

Jacob C. Douma https://orcid.org/0000-0002-8779-838X

James T. Weedon https://orcid.org/0000-0003-0491-8719

REFERENCES

Acevedo-Trejos, E., Brandt, G., Merico, A., & Smith, S. L. (2013). Biogeographical patterns of phytoplankton community size structure in the oceans. *Global Ecology and Biogeography*, *22*(9), 1060–1070. https://doi.org/10.1111/geb.12071

Aitchison, J. (1986). *The statistical analysis of compositional data, volume 25 of Monographs on statistics and applied probability*. London: Chapman and Hall.

Ameztegui, A., Coll, L., & Messier, C. (2015). Modelling the effect of climate-induced changes in recruitment and juvenile growth on mixed-forest dynamics: The case of montane–subalpine pyrenean ecotones. *Ecological Modelling*, *313*, 84–93. https://doi.org/10.1016/j.ecolmodel.2015.06.029

Andrew, N., & Underwood, A. (1993). Density-dependent foraging in the sea urchin centrostephanus rodgersii on shallow subtidal reefs in new south wales, australia. *Marine Ecology Press Series*, *99*, 89–98. https://doi.org/10.3354/meps099089

Balakrishnan, N., & Nevzorov, V. (2003). *A primer on statistical distributions*. Hoboken, NJ: John Wiley and Sons.

Bartlett, M. S. (1936). The square root transformation in analysis of variance. *Supplement to the Journal of the Royal Statistical Society*, *3*(1), 68–78. https://doi.org/10.2307/2983678

Bennett, A. F., Nimmo, D. G., & Radford, J. Q. (2014). Riparian vegetation has disproportionate benefits for landscape-scale conservation of woodland birds in highly modified environments. *Journal of Applied Ecology*, *51*(2), 514–523. https://doi.org/10.1111/1365-2664.12200

Billheimer, D., Guttorp, P., & Fagan, W. F. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, *96*(456), 1205–1214. https://doi.org/10.1198/016214501753381850

Bolker, B. M. (2008). *Ecological models and data in* R*. Princeton, NJ: Princeton University Press.*

Bolker, B. M. (2015). Linear and generalized linear mixed models. G. Fox, S. Negrete-Yankelevich, & V. Sosa (Eds.), *Ecological statistics: contemporary theory and application* (pp. 309–333). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199672547.003.0014.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Briand, C. H., Schwilk, D. W., Gauthier, S., & Bergeron, Y. (2015). Does fire regime influence life history traits of jack pine in the southern boreal forest of Quebec, Canada? *Plant Ecology*, *216*(1), 157–164.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., … Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Busby, P. E., Zimmerman, N., Weston, D. J., Jawdy, S. S., Houbraken, J., & Newcombe, G. (2013). Leaf endophytes and populus genotype affect severity of damage from the necrotrophic leaf pathogen, drepanopeziza populi. *Ecosphere*, *4*(10), art125. https://doi.org/10.1890/ES13-00127.1

Camargo, A. P., Stern, J. M., & Lauretto, M. S. (2012). Estimation and model selection in Dirichlet regression. AIP Conference Proceedings 31st, volume 1443, pp. 206–213. AIP.

Child, A. W., & Moore, B. C. (2015). Effects of hypolimnetic oxygenation on the dietary consumption of methane-oxidizing bacteria by chironomus larvae in dimictic mesotrophic lakes. *Freshwater Science*, *34*(4), 1293–1303. https://doi.org/10.1086/683242

Cotgreave, P., & Clayton, D. H. (1994). Comparative-analysis of time spent grooming by birds in relation to parasite load. *Behaviour*, *131*, 171–187. https://doi.org/10.1163/156853994X00424

Crawley, M. J. (2012). *The* R *book* (2nd ed.). Chichester, UK: Wiley Publishing.

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, *34*(2), 1–24.

De Majo, M. S., Montini, P., & Fischer, S. (2017). Egg hatching and survival of immature stages of aedes aegypti (diptera: Culicidae) under natural temperature conditions during the cold season in buenos aires, argentina. *Journal of Medical Entomology*, *54*(1), 106–113. https://doi.org/10.1093/jme/tjw131

Defries, R. S., Hansen, M. C., Townshend, J. R. G., Janetos, A. C., & Loveland, T. R. (2000). A new global 1-km dataset of percentage tree cover derived from remote sensing. *Global Change Biology*, *6*(2), 247–254. https://doi.org/10.1046/j.1365-2486.2000.00296.x

Espinheira, P. L., Ferrari, S. L. P., & Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, *35*(4), 407–419. https://doi.org/10.1080/02664760701834931

Fang, L., & Kong, Y. (2015). zoib: An R package for bayesian inference for beta regression and zero/one inflated beta regression. *The R Journal*, *7*(2), 34–51.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. https://doi.org/10.1093/biomet/80.1.27

Fukumori, K., Yoshizaki, E., Takamura, N., & Kadoya, T. (2016). Detritivore diversity promotes a relative contribution rate of detritus to the diet of predators in ponds. *Ecosphere*, *7*(3), e01211. https://doi.org/10.1002/ecs2.1211

Grün, B., Kosmidis, I., & Zeileis, A. (2012). Extended beta regression in r: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software, Articles*, *48*(11), 1–25. https://doi.org/10.18637/jss.v048.i11

Gueorguieva, R., Rosenheck, R., & Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational Statistics & Data Analysis*, *52*(12), 5344–5355. https://doi.org/10.1016/j.csda.2008.05.030

Harrison, X. A. (2015). A comparison of observation-level random effect and beta-binomial models for modelling overdispersion in binomial data in ecology & evolution. *PeerJ*, *3*, e1114. https://doi.org/10.7717/peerj.1114

Hijazi, R., & Jernigan, R. (2009). Modeling compositional data using dirichlet regression models. *Journal of Applied Probability & Statistics*, *4*(1), 77–91.

Hirota, M., Holmgren, M., Van Nes, E. H., & Scheffer, M. (2011). Global resilience of tropical forest and savanna to critical transitions. *Science*, *334*(6053), 232–235. https://doi.org/10.1126/science.1210657

Huang, A., & Rathouz, P. J. (2017). Orthogonality of the mean and error distribution in generalized linear models. *Communications in Statistics – Theory and Methods*, *46*(7), 3290–3296. https://doi.org/10.1080/03610926.2013.851241

Joseph, M. B., Preston, D. L., & Johnson, P. T. (2016). Integrating occupancy models and structural equation models to understand species occurrence. *Ecology*, *97*(3), 765–775.

Keim, J. L., DeWitt, P. D., Fitzpatrick, J. J., & Jenni, N. S. (2017). Estimating plant abundance using inflated beta distributions: Applied learnings from a lichen–caribou ecosystem. *Ecology and Evolution*, *7*(2), 486–493. https://doi.org/10.1002/ece3.2625

Kieschnick, R., & McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Statistical Modelling*, *3*(3), 193–213. https://doi.org/10.1191/1471082X03st053oa

Kosmidis, I. (2014). Bias in parametric estimation: Reduction and useful side-effects. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(3), 185–196. https://doi.org/10.1002/wics.1296

Kosmidis, I., & Firth, D. (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, *4*, 1097–1112. https://doi.org/10.1214/10-ejs579

Maier, M. (2014). DirichletReg: Dirichlet regression for compositional data in R. Report, University of Economics and Business - Institute for Statistics and Mathematics. WU Vienna University of Economics and Business, Vienna.

Marschner, I. C., & Gillett, A. C. (2011). Relative risk regression: Reliable and flexible methods for log-binomial models. *Biostatistics*, *13*(1), 179–192. https://doi.org/10.1093/biostatistics/kxr030

McCullagh, P., & Nelder, J. (1989). *Generalized linear models, volume 37 of monographs on statistics and applied probability* (2nd ed.). New York: Chapman and Hall.

Nogueira, A., González-Troncoso, D., & Tolimieri, N. (2016). Changes and trends in the overexploited fish assemblages of two fishing grounds of the Northwest Atlantic. *ICES Journal of Marine Science*, *73*(2), 345–358. https://doi.org/10.1093/icesjms/fsv172

Ospina, R., Cribari-Neto, F., & Vasconcellos, K. L. P. (2006). Improved point and interval estimation for a beta regression model. *Computational Statistics & Data Analysis*, *51*(2), 960–981. https://doi.org/10.1016/j.csda.2005.10.002

Ospina, R., & Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

Poorter, H., Niklas, K. J., Reich, P. B., Oleksyn, J., Poot, P., & Mommer, L. (2012). Biomass allocation to leaves, stems and roots: Meta-analyses of interspecific variation and environmental control. *New Phytologist*, *193*(1), 30–50. https://doi.org/10.1111/j.1469-8137.2011.03952.x

Poorter, H., & Sack, L. (2012). Pitfalls and possibilities in the analysis of biomass allocation patterns in plants. *Frontiers in Plant Science*, *3*, 259. https://doi.org/10.3389/fpls.2012.00259

Poorter, H., van de Vijver, C. A. D. M., Boot, R. G. A., & Lambers, H. (1995). Growth and carbon economy of a fast-growing and a slow-growing grass species as dependent on nitrate supply. *Plant and Soil*, *171*(2), 217–227. https://doi.org/10.1007/BF00010275

Quinn, G., & Keough, M. (2002). *Experimental design and data analysis for biologists*. Cambridge: Cambridge University Press.

R Core Team. (2013). R: *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Regular, P. M., Hedd, A., Montevecchi, W. A., Robertson, G. J., Storey, A. E., & Walsh, C. J. (2014). Why timing is everything: Energetic costs and reproductive consequences of resource mismatch for a chick-rearing seabird. *Ecosphere*, *5*(12), art155. https://doi.org/10.1890/ES14-00182.1

Ruel, J. J., & Ayres, M. P. (1999). Jensen's inequality predicts effects of environmental variation. *Trends in Ecology & Evolution*, *14*(9), 361–366. https://doi.org/10.1016/S0169-5347(99)01664-X

Sánchez, M. S., & Dos Santos, D. A. (2015). Understanding the spatial variations in the diets of two sturnira bats (chiroptera: Phyllostomidae) in Argentina. *Journal of Mammalogy*, *96*(6), 1352–1360. https://doi.org/10.1093/jmammal/gyv144

Schmid, M., Wickler, F., Maloney, K. O., Mitchell, R., Fenske, N., & Mayr, A. (2013). Boosted beta regression. *PLoS ONE*, *8*(4), 1–15. https://doi.org/10.1371/journal.pone.0061623

Shimada, T., Limpus, C., Jones, R., Hazel, J., Groom, R., & Hamann, M. (2016). Sea turtles return home after intentional displacement from coastal foraging areas. *Marine Biology*, *163*(1), 8. https://doi.org/10.1007/s00227-015-2771-0

Simas, A. B., Barreto-Souza, W., & Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, *54*(2), 348–366. https://doi.org/10.1016/j.csda.2009.08.017

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society Series B (Methodological)*, *10*(2), 257–261. https://doi.org/10.1111/j.2517-6161.1948.tb00014.x

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. https://doi.org/10.1037/1082-989x.11.1.54

Sokahl, R., & Rohlf, F. (1995). *Biometry* (3rd ed.). New York: W. H. Freeman.

Staver, A. C., Archibald, S., & Levin, S. A. (2011). The global extent and determinants of savanna and forest as alternative biome states. *Science*, *334*(6053), 230–232.

Tsagris, M., & Stewart, C. (2018). A dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, *39*(3), 398–412. https://doi.org/10.1134/s1995080218030198

van den Boogaart, K., & Tolosana-Delgado, R. (2013). *Analyzing compositional data with R*. Heidelberg, Germany: Springer.

Warton, D. I., & Hui, F. K. C. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology*, *92*(1), 3–10. https://doi.org/10.1890/10-0340.1

Wright, W. J., Irvine, K. M., Warren, J. M., & Barnett, J. K. (2017). Statistical design and analysis for plant cover studies with multiple sources of observation errors. *Methods in Ecology and Evolution*, *8*(12), 1832–1841. https://doi.org/10.1111/2041-210X.12825

Zuur, A. F., & Ieno, E. N. (2016). A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*, *7*(6), 636–645. https://doi.org/10.1111/2041-210x.12577

Zuur, A., Ieno, E., Walker, N., Saveliev, A., & Smith, G. (2009). *Mixed effects models and extensions in ecology with* R. *New York: Springer.*

SUPPORTING INFORMATION

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

How to cite this article: Douma JC, Weedon JT. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods Ecol Evol*. 2019;10:1412–1430. https://doi.org/10.1111/2041-210X.13234