Regression for compositional data

Guillaume Franchi

1 Some problems related to compositional data

We present in this section some problems leading to regression with compositional data.

Definition 1.1. Compositional data consist of vectors whose components are the proportions of some whole. Mathematically, those vectors are elements of the **simplex**

$$\mathcal{S}_{d-1} = \left\{ (x_1, \dots, x_d) \in]0;1[^d \;\middle|\; \sum_{i=1}^{d} x_i = 1 \right\}. \tag{1}$$

The following regression problem comes from Aitchison and Aitchison (1986)

Example 1.2 (Arctic Lake)**.** In sedimentology, specimens of sediments are traditionally separated into three mutually exclusive and exhaustive constituents: sand, silt and clay.

Table 4 records the compositions of 39 sediment samples at different water depths in an Arctic lake. A first question we could we could ask is: « Does the composition of a sediment depends on the depth of this sediment, and how can we model it ?»

When $d = 2$, there is only one proportion to study, which leads to a regression problem for real values. However, the constraints due to the proportional nature of the values must be taken into account. The next example is derived from Douma and Weedon (2019).

Example 1.3 (Forest Cover)**.** Consider a forest represented here by a square of one hectare in area. We focus here on the ground cover of the forest explained by the annual rainfall.

Twenty forests are simulated, each with a mean annual precipitation (MAP) ranging from 125 to 2,500 mm per year. In practice, it is impossible to measure the ground cover of the whole forest. Thus for each forest, the percentage of ground cover is calculated for fifteen randomly positioned and non-overlapping quadrats of 10×10 m^2 (See Figure 1). This leads to 300 observations of ground cover percentage stored in Table 5.

(a) Ground cover of a forest with a MAP of 375 mm. (b) Ground cover of a forest with a MAP of 2000 mm.

Figure 1 – Examples of forests with different annual rainfalls. The green circles correspond to the threes, and the squares correspond to the quadrats.

2 Regression models

2.1 General framework

One can immediately guess why applying any classical regression model on the raw data is a bad idea for compositional problems. It is indeed impossible to ensure that the constraints on the simplex will be satisfied for the predicted values.

Assume we have a sample of N compositional data $y^{(1)}, \dots, y^{(N)}$ in $\mathbb{R}^d$:

$$\forall i \in \{1, \dots, N\}, \; y^{(i)} = (y_1^{(i)}, \dots, y_d^{(i)}),$$

that we want to explain with K covariates $x_1, \dots, x_K$. Applying an ordinary least square regression would ensure that the predicted value $\hat{y} = (\hat{y}_1, \dots, \hat{y}_d)$ satisfies:

$$\sum_{j=1}^{d} \hat{y}_j = 1. \tag{2}$$

However, it is possible that for some j, $\hat{y}_j \notin [0; 1]$. This would present a huge problem of interpretation when working with data that represent proportions.

Applying a non-linear model to this problem could resolve this issue. Nevertheless, it is mandatory that the predicted value $\hat{y}$ satisfies the sum constraint (2).

These remarks lead to the first model we will present here, which is the one recommended by Aitchison and Aitchison (1986).

2.2 Data transformation

The general idea here is to apply a one-to-one mapping

$$g : \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^k \tag{3}$$

and to apply any classical regression method on the transformed data $z^{(1)} = g(y^{(1)}), \dots, z^{(N)} = g(y^{(N)})$, with covariates $x_1, \dots, x_K$.
Then the prediction $\hat{y}$ is obtained with $\hat{y} = g^{-1}(\hat{z})$ where $\hat{z}$ is the prediction on the space of transformed data.

Although there is a multitude of possible transformations, Aitchison and Aitchison (1986) suggests the **additive logratio**:

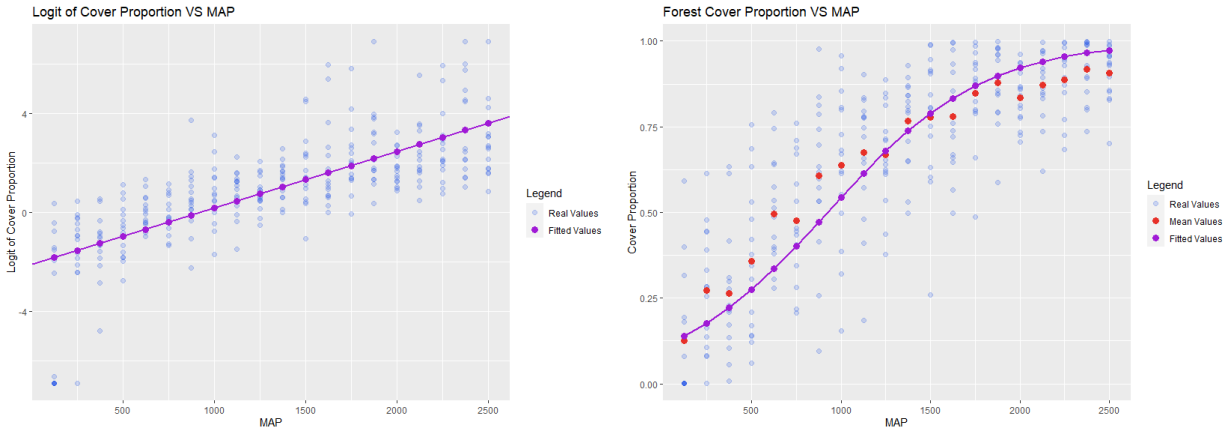
$$\begin{aligned} alr : \; & \mathcal{S}_{d-1} \longrightarrow \mathbb{R}^{d-1} \\ & y \longmapsto \left(\log\left(\frac{y_1}{y_d}\right), \dots, \log\left(\frac{y_{d-1}}{y_d}\right) \right) \end{aligned} \tag{4}$$

with inverse transform:

$$alr^{-1}(z_1, \dots, z_{d-1}) = \left(\frac{\exp(z_1)}{1 + \sum_{j=1}^{d-1} \exp(z_j)}, \dots, \frac{\exp(z_{d-1}}{1 + \sum_{j=1}^{d-1} \exp(z_j)}, \frac{1}{1 + \sum_{j=1}^{d-1} \exp(z_j)} \right). \tag{5}$$

Note that when $d = 2$, i.e. we study one proportion, the alr transform is actually a logit transform.

Example 2.1 (Forest cover)**.** Let us consider Example 1.3, where an ordinary least square regression is applied to the data transformed with a logit function (See Figure 2a). The fitted values are subsequently backtransformed to obtain Figure 2b.



(a) Fitted values on the transformed scale.

(b) Fitted values on the initial scale.

Figure 2 – Regression for the forest cover proportion with a logit transform

We present below an example of regression with a transformation of data for more than one proportion.

Example 2.2 (Arctic Lake)**.** Consider Example 1.2 the question of interest is to predict the composition (Sand, Silt and Clay) of a sediment from the depth at which it was taken.

Here we apply the *alr* transform to the data stored in Table 4. The transformed data are vectors in $\mathbb{R}^2$ denoted $z^{(i)} = (z_1^{(i)}, z_2^{(i)})_{1 \leqslant i \leqslant 39}$. Figure 3 present the scatterplots of the transformed data.

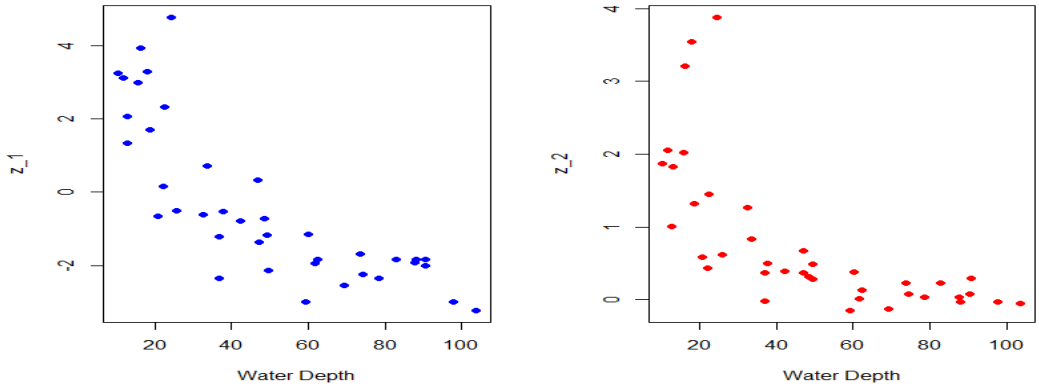


Figure 3 – Scatterplots of the transformed data in the Arctic Lake Regression.

Due to the shape of the scatterplots, we decide to model these data by a polynomial of order 2. We thus apply a non-linear least square regression on the transformed data.
We finally get back to the original scale by applying the inverse transform of the *alr* on the fitted values in the transformed scale. Figure 4 presents the estimated compositions in comparison to the true ones.

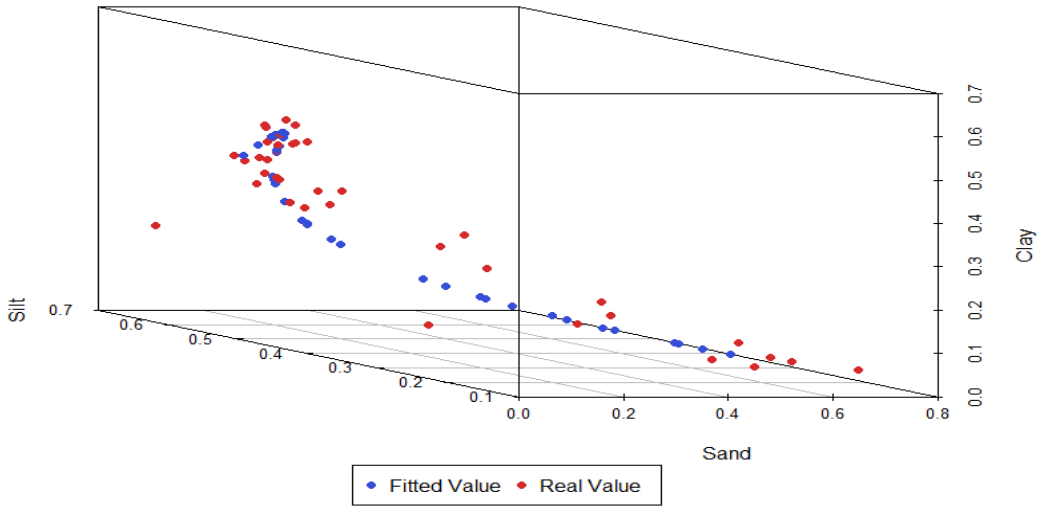


Figure 4 – Regression for the Arctic Lake Sediments compositions with an *alr* transformation.

The main issue of this technique is that, even if the estimates on the transformed scale are unbiased, it might not be the case on the original scale (See Douma and Weedon (2019)). This issue arises due to Jensen's inequality. Since the back-transformation is, in general, not linear, the bias on the original scale might be greater than the one on the transformed scale, resulting sometimes in major discrepancies in the fitted values.

In order to avoid this problem, we propose in the next section a different approach to deal with regression for compositional data.

2.3 The «Stay in the simplex» approach

The principle of this approach is to assume that each observation $y^{(i)}$ is our sample is a realization of a random variable with a specific distribution supported by the simplex.

2.3.1 Beta Regression

We first consider the case $d = 2$, where there is actually only one proportion to study.

In Beta regression, we will assume that each random variable $Y^{(i)}$ folllows a Beta distribution $B(p_i, q_i)$. Varying values of p_i and q_i allows indeed to obtain very different densities (See Figure 5), which makes the Beta distribution very flexible and well suited for modelization.

Definition 2.3. The **Beta distribution** $B(p, q)$ is the distribution whose density is given by:

$$f_{(p,q)}(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \mathbb{1}_{[0;1]}(x) x^{p-1}(1-x)^{q-1} \tag{6}$$

where $\Gamma(\cdot)$ is the gamma function.

Remark 2.4. Recall that the Gamma function is defined on $]0; +\infty[$ by:

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} \exp(-t) \mathrm{d}t \tag{7}$$

and satisfies for all $x > 0$:

$$\Gamma(x+1) = x\Gamma(x). \tag{8}$$

The Beta function is defined for all $a, b > 0$ by:

$$\beta(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1} \mathrm{d}t \tag{9}$$

and satisfies the relation:

$$\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{10}$$

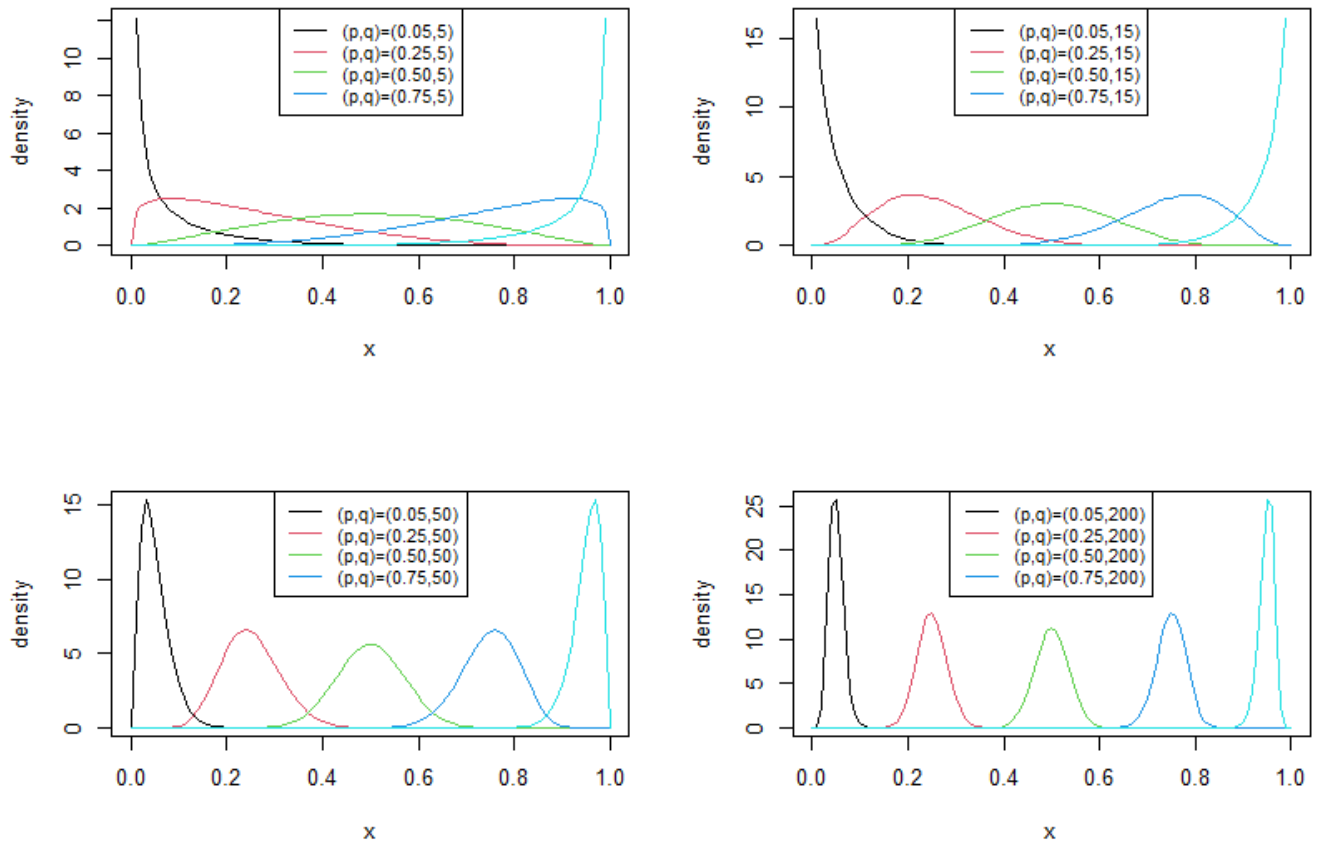


Figure 5 – Beta densities for different combinations of (p, q).

Proposition 2.5. *Let Y be a random variable with distribution $B(p,q)$. The mean and variance of Y are given by:*

$$\mathbb{E}(Y) = \frac{p}{p+q} \tag{11}$$

and

$$\mathrm{Var}(Y) = \frac{pq}{(p+q)^2(p+q+1)}. \tag{12}$$

Proof. We have indeed:

$$\begin{aligned}
\mathbb{E}(Y) &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^1 y^p (1-y)^{q-1} \mathrm{d}y \\
&= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \times \beta(p+1,q) \\
&= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \times \frac{\Gamma(p+1)\Gamma(q)}{\Gamma(p+q+1)} \\
&= \frac{\Gamma(p+q) \times p\Gamma(p)\Gamma(q)}{\Gamma(p)\Gamma(q) \times (p+q)\Gamma(p+q)} \\
&= \frac{p}{p+q}
\end{aligned}$$

and:

$$\begin{aligned}
\mathbb{E}(Y^2) &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^1 y^{p+1} (1-y)^{q-1} \mathrm{d}y \\
&= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \times \beta(p+2,q) \\
&= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \times \frac{\Gamma(p+2)\Gamma(q)}{\Gamma(p+q+2)} \\
&= \frac{(p+1)p}{(p+q+1)(p+q)}.
\end{aligned}$$

Thus:

$$\mathrm{Var}(Y) = \frac{p(p+1)}{(p+q+1)(p+q)} - \frac{p^2}{(p+q)^2} = \frac{pq}{(p+q)^2(p+q+1)}.$$

■

For regression purpose, the parameter of interest is often the mean of the response variable. In Beta regression, we propose a different parametrization of the Beta density given in (6), so that the model we build contains focuses on the mean of the response with a dispersion parameter.

Proposition 2.6. *Let Y be a random variable with distribution $\beta(p,q)$ and denote:*

$$\mu = \frac{p}{p+q} \quad \textit{and} \quad \phi = p+q. \tag{13}$$

We have:

$$\mathbb{E}(Y) = \mu \quad \textit{and} \quad \mathrm{Var}(Y) = \frac{\mu(1-\mu)}{1+\phi} \tag{14}$$

and the density of Y can be written with this new parametrization:

$$f_{(\mu,\phi)}(y) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \mathbb{1}_{]0;1[}(y) \tag{15}$$

where $0 < \mu < 1$ and $\phi > 0$.

Remark 2.7. Equation (15) allows us to interprete ϕ as a dispersion parameter, since $\mathrm{Var}(Y)$ decreases as ϕ increases.

The model we build in Beta regression is obtained by assuming that each independent variable $Y^{(t)}$ follows the density given by (15) with mean μ_t and unknown dispersion ϕ.

Furthermore, we assume that the mean μ_t satisfies the equation:

$$g(\mu_t) = \sum_{i=1}^{k} x_{t,i}\beta_i := \eta_t \tag{16}$$

where $\beta = (\beta_1, \ldots, \beta_k)'$ is a vector of unknown regression parameters of $\mathbb{R}^k$, $x_{t,1}, \ldots, x_{t,k}$ are observations of k covariates, and g is a **link function**, strictly monotonic and twice differentiable, mapping $]0;1[$ into $\mathbb{R}$.

Remark 2.8. There are of course many possibilities in the choice of the link function g. Among the most popular are the logit function $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ or the probit function $g(\mu) = \Phi^{-1}(\mu)$ where Φ is the cumulative distribution function of a standard normal distribution.

In the case of the logit function we can write:

$$\mu_t = \frac{\exp(x_t' \cdot \beta)}{1 + \exp(x_t' \cdot \beta)} \tag{17}$$

where $x_t' = (x_{t,1}, \ldots, x_{t,k})$, so that the regression parameters have an important interpretation.
Suppose that the value of the i^{th} regressor is modified by c units, and all other independent variables remain unchanged. Let ν_t denote the mean of $Y^{(t)}$ under the new covariate values, whereas μ_t denotes the mean of $Y^{(t)}$ under the original covariate values.
One can show that:

$$\exp(c\beta_i) = \frac{\nu_t/(1-\nu_t)}{\mu_t/(1-\mu_t)}, \tag{18}$$

meaning that $\exp(c\beta_i)$ equals the odds ratio.
According to equation (17), we have indeed:

$$\begin{aligned}
\frac{\nu_t}{1-\nu_t} &= \frac{\exp\left(\sum_j \beta_j x_{t,j} + c\beta_i\right)}{1 + \exp\left(\sum_j \beta_j x_{t,j} + c\beta_i\right)} \times \left(1 - \frac{\exp\left(\sum_j \beta_j x_{t,j} + c\beta_i\right)}{1 + \exp\left(\sum_j \beta_j x_{t,j} + c\beta_i\right)}\right) \\
&= \mathrm{e}^{c\beta_i} \cdot \exp\left(\sum_j \beta_j x_{t,j}\right) \\
&= \mathrm{e}^{c\beta_i} \cdot \frac{\mu_t}{1-\mu_t}.
\end{aligned}$$

An example of this interpretation will be given in Example 2.10.

The regression parameter $\beta = (\beta_1, \ldots, \beta_k)$, as well as the dispersion parameter ϕ are estimated with their maximum likelihood estimators $\hat{\beta}$ and $\hat{\phi}$.

Remark 2.9. Note that $\hat{\beta}$ and $\hat{\phi}$ do not have a closed form, hence they are obtained by numerical optimization algorithms such as a Newton algorithm.

Example 2.10 (Forest cover)**.** We consider once again Example 1.3.
This time, we assume that the response variable $Y^{(t)}$, i.e. the proportion of ground cover follows a beta distribution with mean μ_t satisfying:

$$\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta x_t + \alpha \tag{19}$$

where x_t is the mean annual precipitation, in mm per year. In our model, an intercept is added in the covariates.
The maximum likelihood estimators obtained are $\hat{\alpha} = -1,4697$, $\hat{\beta} = 0,0016$ and $\hat{\phi} = 4,5377$. In Figure 6 we can visualize the fitted values with Beta regression, compared to the ones obtained by data transformation in Example 2.1.

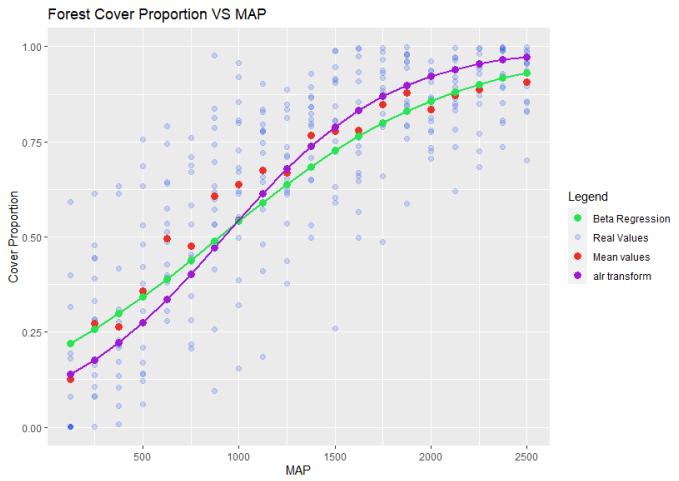


Figure 6 – Comparison of regression models on the Forest Cover example.

In this example, the goodness of fit for both models is quite similar, as the MSE in Table 1 suggest. Note that this MSE are computed with the mean values and the predicted values in each case. However, the Beta regression model seems to fit slightly better the observed mean cover proportion.

	Data Transformation	Beta Regression
MSE	0.0054	0.0035

Table 1 – MSE of the different models in the Forest Cover example.

2.3.2 Dirichlet Regression

We now consider the multivariate case where there are more than two proportions to study: $d \geqslant 3$. A generalization of the Beta distribution is the Dirichlet distribution.

Definition 2.11. The **Dirichlet distribution** $\mathrm{Dir}(\alpha_1, \dots, \alpha_d)$ is the distribution $\mu_\alpha^{(d)}$ such that for any Borel set in $\mathbb{R}^d$ we have:

$$\mu_\alpha^{(d)}(A) = \int \cdots \int \mathbb{1}_A(x_1, \dots, x_d) \frac{\Gamma(\sum_{i=1}^d \alpha_i)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_d)} \prod_{i=1}^d x_i^{\alpha_i - 1} \mathbb{1}_{B_{d-1}}(x_1, \dots, x_{d-1}) \; \delta_{\left\{1 - \sum_{i=1}^{d-1} x_i\right\}}(\mathrm{d}x_d) \mathrm{d}x_1 \dots \mathrm{d}x_{d-1} \quad (20)$$

where $B_{d-1} = \left\{(x_1, \dots, x_{d-1}) \in \mathbb{R}_+^{d-1} \;\middle|\; 0 < \sum_{i=1}^{d-1} x_i < 1\right\}$.

The Dirichlet distribution is indeed a generalization of the Beta distribution, in the sense of Proposition 2.12.

Proposition 2.12. *Let $Y = (Y_1, \dots, Y_d)$ a random vector with distribution $\mathrm{Dir}(\alpha_1, \dots, \alpha_d)$, and let $\phi = \sum_{i=1}^d \alpha_i$. Then, for all $1 \leqslant i \leqslant d$, we have $X_i \sim \beta(\alpha_i, \phi - \alpha_i)$.*
In particular $\mathbb{E}(X_i) = \dfrac{\alpha_i}{\phi}$ and $\mathrm{Var}(X_i) = \dfrac{\alpha_i(\phi - \alpha_i)}{\phi^2(\phi + 1)}$.

Proof. For simplification purpose, we will prove this result for Y_{d-1}. Consider any measurable function $h : \mathbb{R} \longrightarrow \mathbb{R}_+$, we have:

$$\begin{aligned}
\mathbb{E}\left(h(Y_{d-1})\right) &= \frac{\Gamma(\phi}{\Gamma(\alpha_1) \dots \Gamma(\alpha_d)} \int \int h(y_{d-1}) \prod i = 1^d y_i^{\alpha_i - 1} \mathbb{1}_{B_{d-1}}(y_1, \dots y_{d-1}) \delta_{\left\{1 - \sum_{i=1}^{d-1} y_i\right\}}(\mathrm{d}y_d) \mathrm{d}y_1 \dots \mathrm{d}y_{d-1} \\
&= \frac{\Gamma(\phi}{\Gamma(\alpha_1) \dots \Gamma(\alpha_d)} \int_{B_{d-1}} h(y_{d-1}) \prod_{i=1}^{d-1} y_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{d-1} y_i\right)^{\alpha_d - 1} \mathrm{d}y_1 \dots \mathrm{d}y_{d-1} \\
&= \frac{\Gamma(\phi}{\Gamma(\alpha_1) \dots \Gamma(\alpha_d)} \int_0^1 h(y_{d-1}) y_{d-1}^{\alpha_{d-1} - 1} \int_A \left(1 - \sum_{i=1}^{d-2} y_i - y_{d-1}\right)^{\alpha_d - 1} \prod_{i=1}^{d-2} y_i^{\alpha_i - 1} \mathrm{d}y_1 \dots \mathrm{d}y_{d-2} \mathrm{d}y_{d-1}
\end{aligned}$$

with $A = \left\{(y_1, \dots y_{d-2}) \;\middle|\; \sum_{i=1}^{d-2} y_i < 1 - y_{d-1}\right\}$. It yields:

$$\mathbb{E}((h(Y_{d-1})) = \frac{\Gamma(\phi}{\Gamma(\alpha_1) \dots \Gamma(\alpha_d)} \int_0^1 h(y_{d-1}) y_{d-1}^{\alpha_{d-1} - 1} (1 - y_{d-1})^{\alpha_d - 1} \int_A \prod_{i=1}^{d-2} y_i^{\alpha_i - 1} \left(1 - \frac{\sum_{i=1}^{d-2} y_i}{1 - y_{d-1}}\right)^{\alpha_d - 1} \mathrm{d}y_1 \dots \mathrm{d}y_{d-2} \mathrm{d}y_{d-1}.$$

We thus focus on:

$$I = \int_A \prod_{i=1}^{d-2} y_i^{\alpha_i - 1} \left(1 - \frac{\sum_{i=1}^{d-2} y_i}{1 - y_{d-1}}\right)^{\alpha_d - 1} \mathrm{d}y_1 \dots \mathrm{d}y_{d-2}$$

and consider:

$$\begin{aligned} \Phi : \ & \mathbb{R}^{d-2} \longrightarrow \mathbb{R}^{d-2} \\ & (y_1, \dots y_{d-2} \longmapsto \left(\frac{y_1}{1 - y_{d-1}}, \dots \frac{y_{d-2}}{1 - y_{d-1}}\right). \end{aligned}$$

One can show that Φ is a $\mathcal{C}^1$-diffeomorphism from A to B_{d-2}, with jacobian:

$$J_\Phi(y_1, \dots y_{d-2}) = (1 - y_{d-1})^{2-d}.$$

It yields:

$$\begin{aligned} I &= \int_{B_{d-2}} (1 - y_{d-1})^{\sum_{i=1}^{d-2} \alpha_i - d + 2} \left(1 - \sum_{i=1}^{d-2} u_i\right)^{\alpha_d - 1} \prod_{i=1}^{d-2} u_i^{\alpha_i - 1} (1 - y_{d-1})^{d-2} \, \mathrm{d}u_1 \dots \mathrm{d}u_{d-2} \\ &= (1 - y_{d-1})^{\sum_{i=1}^{d-2} \alpha_i} \cdot \frac{\prod_{i=1}^{d-2} \Gamma(\alpha_i)}{\Gamma(\phi - \alpha_{d-1}}. \end{aligned}$$

Thus:

$$\mathbb{E}(h(Y_{d-1})) = \int_0^1 h(y_{d-1}) y_{d-1}^{\alpha_{d-1} - 1} (1 - y_{d-1})^{\phi - \alpha_{d-1} - 1} \cdot \frac{\Gamma(\phi)}{\Gamma(\alpha_{d-1})\Gamma(\phi - \alpha_{d-1})} \mathrm{d}y_{d-1}$$

i.e. $Y_{d-1} \sim B(\alpha_{d-1}, \phi - \alpha_{d-1})$.
The law of the other marginals can be determined the same way. ■

By analogy with the Beta regression, our parameter of interest will be the mean vector of the response variable, along with a dispersion parameter. Thus, we will propose a different parametrization of the Dirichlet distribution to fulfill those requirements.

Proposition 2.13. *Let $Y = (Y_1, \dots, Y_d)$ be a random variable with distribution $\mathrm{Dir}(\alpha_1, \dots \alpha_d)$ and denote:*

$$\mu = (\mu_1, \dots \mu_d) = \left(\frac{\alpha_1}{\phi}, \dots, \frac{\alpha_d}{\phi}\right) \quad \textit{and} \quad \phi = \sum_{i=1}^{d} \alpha_i. \tag{21}$$

We have:

$$\mathbb{E}(Y) = \mu \quad \textit{and} \quad \mathrm{Var}(Y_i) = \frac{\mu_i(-\mu_i)}{\phi + 1} \tag{22}$$

and the density of Y can be written with this new parametrization:

$$f_{(\mu,\phi)}(y_1, \dots y_d) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu_1) \dots \Gamma(\phi\mu_d)} \prod_{i=1}^{d} y_i^{\phi\mu_i - 1} \mathbb{1}_{B_{d-1}}(y_1, \dots, y_{d-1}) \tag{23}$$

where $0 < \mu_i < 1$ for all $1 \leqslant i \leqslant d$ and $\phi > 0$.

Following the steps of the Beta regression, the model we build in Dirichlet regression is obtained by assuming that each independent variable $Y^{(t)}$ follows the density given by (23), with mean μ_t and unknown dispersion ϕ.
Furthermore, by analogy with the multinomial regression, we assume that the mean μ_t satisfies the equations:

$$\forall i \in \{1, \dots, d-1\}, \ \mu_{t,i} = \frac{\exp(\beta^{(i)} \cdot x_t)}{1 + \sum_{j=1}^{d-1} \exp(\beta^{(j)} \cdot x_t)} \tag{24}$$

and

$$\mu_{t,d} = \frac{1}{1 + \sum_{j=1}^{d-1} \exp(\beta^{(j)} \cdot x_t)} \tag{25}$$

where the $\beta^{(i)}$'s are vectors of unknown regression parameters of $\mathbb{R}^k$ and x_t is the observation of k covariates.

Remark 2.14. Similarly to the Beta regression, the vectors $\beta^{(i)}$ have an interpretation in terms of odds ratio.

The regression vectors $\beta^{(i)}$ and the dispersion parameter ϕ are estimated with their maximum likelihood estimators $\widehat{\beta^{(i)}}$ and $\hat{\phi}$.

Remark 2.15. Once again, $\widehat{\beta^{(i)}}$, $1 \leqslant i \leqslant d$ and $\hat{\phi}$ have no closed form, hence they are obtained by a numerical optimization algorithm.

Example 2.16 (Arctic Lake)**.** Return to Example 1.2, and assume that the composition $Y^{(t)}$ of the t-th sediment follows a Dirichlet distribution with mean μ_t satisfying for $i \in \{1, 2\}$:

$$\mu_{t,i} = \frac{\exp(\alpha^{(i)} + \beta_1^{(i)} x_t + \beta_2^{(i)} x_t^2)}{1 + \sum_{j=1}^{2} \exp(\alpha^{(j)} + \beta_1^{(j)} x_t + \beta_2^{(j)} x_t^2)} \tag{26}$$

and :

$$\mu_{t,3} = \frac{1}{1 + \sum_{j=1}^{2} \exp(\alpha^{(j)} + \beta_1^{(j)} x_t + \beta_2^{(j)} x_t^2)} \tag{27}$$

where x_t is the water depth of the sediment, in meters.
Here, we actually consider two covariates in addition to an intercept: the water depth and the square of the water depth. It produced indeed better fitted values than considering only one covariate.

The maximum likelihood estimators obtained are given in Table 2.

$\widehat{\alpha^{(1)}}$	$\widehat{\alpha^{(2)}}$	$\widehat{\beta_1^{(1)}}$	$\widehat{\beta_2^{(1)}}$	$\widehat{\beta_1^{(2)}}$	$\widehat{\beta_2^{(2)}}$	$\hat{\phi}$
4,1566	2,4091	-0,1552	0,0010	-0,0602	0,0040	19,0410

Table 2 – Maximum Likelihood Estimators for the Dirichlet Regression in the Arctic Lake example

In Figure 7, we can visualize the fitted values with Dirichlet regression compared to the ones obtained by data transformation in Example 2.2

We also propose in Table 3 the values of the MSE for the two regression models.

	Data Transformation	Dirichlet Regression
MSE	0.0254	0.0244

Table 3 – MSE of the different models in the Arctic Lake Example

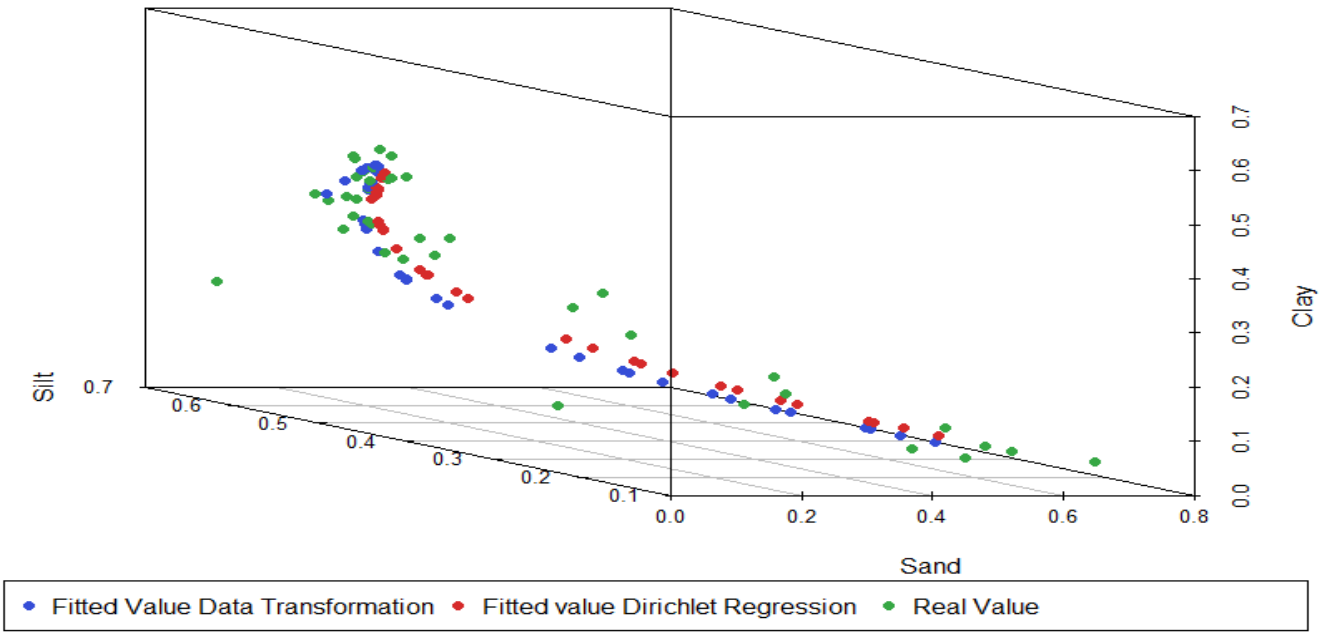


Figure 7 – Comparison of regression models for the Arctic Lake Example.

2.3.3 A visual diagnostic

In both our models, we assume that each response variable Y_j follows a Beta distribution. In the case of the Dirichlet model:

$$Y_j \sim B(\alpha_j, \phi - \alpha_j).$$

If the model is correct, then by denoting F_j the cumulative distribution function of $B(\alpha_j, \phi - \alpha_j)$, we have:

$$F_j(Y_j) \sim \mathcal{U}(]0;1[). \tag{28}$$

and:

$$R_j := \Phi^{-1}\left(F_j(Y_j)\right) \sim \mathcal{N}(0,1) \tag{29}$$

where Φ^{-1} is the inverse cumulative distribution function of a standard gaussian distribution.

Definition 2.17. The random variables R_j are called the **pseudo-residuals** of the regression.

Example 2.18 (Arctic Lake)**.** For the Arctic Lake example, we plotted in Figure 8 the normal QQ-plots of the pseudo-residuals. Those residuals are calculated with the estimates obtained in Example 2.16.

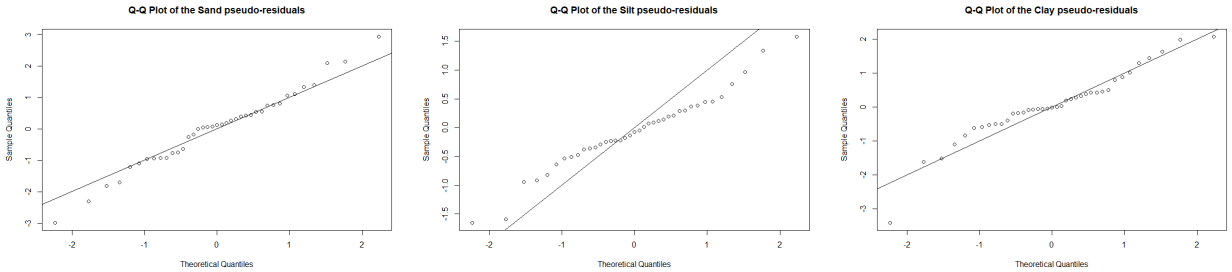


Figure 8 – Normal QQ-Plots of the pseudo residuals in the Arctic Lake Regression.

No particular discrepancy is detected here, except may be for the composition of Silt.

2.4 Main Drawbacks of the methods

With the transformation of the data, we already mentioned that issues can arise due to Jensen's inequality. Indeed, the back-transformation could induce some bias, hence producing discrepancies in the fitted values.
Furthermore, interpretation of the coefficient estimates is difficult, since the estimation is done on the transformed scale.

On the other hand, interpretation is very easy with the Beta or Dirichlet regression. However, the Dirichlet distribution implies a very strong structure of independence in the compositions, as mentioned in Proposition 2.19.

Proposition 2.19. *Let $\alpha_1, \dots, \alpha_d$ be positive real numbers, and $U_1, \dots, U_d$ be independent random variables with gamma distributions:*

$$U_i \sim \gamma(\alpha_i, 1).$$

Then, if we denote $V = \sum_{i=1}^{d} U_i$, we have:

$$\left(\frac{U_1}{V}, \dots, \frac{U_d}{V}\right) \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_d).$$

Remark 2.20. Let us recall that a random variable U follows the gamma distribution $\gamma(a,p)$ if it admits the density:

$$f(u) = \mathbb{1}_{\mathbb{R}_+}(u)\frac{p^a}{\Gamma(a)}\mathrm{e}^{-pu}u^{a-1}. \tag{30}$$

Hence, it is possible that the Dirichlet Regression might have some issues in modeling some data with a very strong dependence structure.

3 Appendices

3.1 Arctic Lake Data

Sediment	Sand	Silt	Clay	Water Depth
1	0.78	0.20	0.03	10.40
2	0.72	0.25	0.03	11.70
3	0.51	0.36	0.13	12.80
4	0.52	0.41	0.07	13.00
5	0.70	0.26	0.04	15.70
6	0.66	0.32	0.01	16.30
7	0.43	0.55	0.02	18.00
8	0.53	0.37	0.10	18.70
9	0.15	0.54	0.30	20.70
10	0.32	0.41	0.27	22.10
11	0.66	0.28	0.06	22.40
12	0.70	0.29	0.01	24.40
13	0.17	0.54	0.29	25.80
14	0.11	0.70	0.20	32.50
15	0.38	0.43	0.19	33.60
16	0.11	0.53	0.36	36.80
17	0.18	0.51	0.31	37.80
18	0.05	0.47	0.48	36.90
19	0.16	0.50	0.34	42.20
20	0.32	0.45	0.23	47.00
21	0.10	0.54	0.37	47.10
22	0.17	0.48	0.35	48.40
23	0.10	0.55	0.34	49.40
24	0.05	0.54	0.41	49.50
25	0.03	0.45	0.52	59.20
26	0.11	0.53	0.36	60.10
27	0.07	0.47	0.46	61.70
28	0.07	0.50	0.43	62.40
29	0.04	0.45	0.51	69.30
30	0.07	0.52	0.41	73.60
31	0.05	0.49	0.46	74.40
32	0.04	0.48	0.47	78.50
33	0.07	0.52	0.41	82.90
34	0.07	0.47	0.46	87.70
35	0.07	0.46	0.47	88.10
36	0.06	0.49	0.45	90.40
37	0.06	0.54	0.40	90.60
38	0.02	0.48	0.49	97.70
39	0.02	0.48	0.50	103.70

Table 4 – Sediments compositions in an arctic lake.

3.2 Forest Cover Data

	MAP	Quadrant	Cover.m
1	125	Quadrant.1	0.001
2	250	Quadrant.1	0.332
3	375	Quadrant.1	0.217
4	500	Quadrant.1	0.420
5	625	Quadrant.1	0.575
6	750	Quadrant.1	0.346
7	875	Quadrant.1	0.838
8	1000	Quadrant.1	0.728
9	1125	Quadrant.1	0.720
10	1250	Quadrant.1	0.814
11	1375	Quadrant.1	0.791
12	1500	Quadrant.1	0.602
13	1625	Quadrant.1	0.498
14	1750	Quadrant.1	0.825
15	1875	Quadrant.1	0.999
16	2000	Quadrant.1	0.962
17	2125	Quadrant.1	0.973
18	2250	Quadrant.1	0.872
19	2375	Quadrant.1	0.892
20	2500	Quadrant.1	0.852
21	125	Quadrant.2	0.080
22	250	Quadrant.2	0.613
23	375	Quadrant.2	0.173
24	500	Quadrant.2	0.210
25	625	Quadrant.2	0.278
26	750	Quadrant.2	0.281
27	875	Quadrant.2	0.533
28	1000	Quadrant.2	0.498
29	1125	Quadrant.2	0.903
30	1250	Quadrant.2	0.710
31	1375	Quadrant.2	0.841
32	1500	Quadrant.2	0.990
33	1625	Quadrant.2	0.668
34	1750	Quadrant.2	0.931
35	1875	Quadrant.2	0.793
36	2000	Quadrant.2	0.868
37	2125	Quadrant.2	0.864
38	2250	Quadrant.2	0.889
39	2375	Quadrant.2	0.999
40	2500	Quadrant.2	0.934
41	125	Quadrant.3	0.001
42	250	Quadrant.3	0.137
43	375	Quadrant.3	0.264
44	500	Quadrant.3	0.633
45	625	Quadrant.3	0.645
46	750	Quadrant.3	0.711
47	875	Quadrant.3	0.533
48	1000	Quadrant.3	0.154
49	1125	Quadrant.3	0.658
50	1250	Quadrant.3	0.535
51	1375	Quadrant.3	0.649
52	1500	Quadrant.3	0.752
53	1625	Quadrant.3	0.647
54	1750	Quadrant.3	0.882
55	1875	Quadrant.3	0.962
56	2000	Quadrant.3	0.936
57	2125	Quadrant.3	0.965
58	2250	Quadrant.3	0.684
59	2375	Quadrant.3	0.993
60	2500	Quadrant.3	0.856
61	125	Quadrant.4	0.195
62	250	Quadrant.4	0.284
63	375	Quadrant.4	0.055
64	500	Quadrant.4	0.122
65	625	Quadrant.4	0.745
66	750	Quadrant.4	0.454
67	875	Quadrant.4	0.696
68	1000	Quadrant.4	0.807
69	1125	Quadrant.4	0.747
70	1250	Quadrant.4	0.835
71	1375	Quadrant.4	0.864
72	1500	Quadrant.4	0.631
73	1625	Quadrant.4	0.975
74	1750	Quadrant.4	0.660
75	1875	Quadrant.4	0.943
76	2000	Quadrant.4	0.864
77	2125	Quadrant.4	0.996
78	2250	Quadrant.4	0.928
79	2375	Quadrant.4	0.799
80	2500	Quadrant.4	0.833
81	125	Quadrant.5	0.001
82	250	Quadrant.5	0.164
83	375	Quadrant.5	0.311
84	500	Quadrant.5	0.685
85	625	Quadrant.5	0.641
86	750	Quadrant.5	0.407
87	875	Quadrant.5	0.619
88	1000	Quadrant.5	0.921
89	1125	Quadrant.5	0.795
90	1250	Quadrant.5	0.886
91	1375	Quadrant.5	0.894
92	1500	Quadrant.5	0.259
93	1625	Quadrant.5	0.567
94	1750	Quadrant.5	0.985
95	1875	Quadrant.5	0.790
96	2000	Quadrant.5	0.734
97	2125	Quadrant.5	0.893
98	2250	Quadrant.5	0.886
99	2375	Quadrant.5	0.853
100	2500	Quadrant.5	0.990
101	125	Quadrant.6	0.001
102	250	Quadrant.6	0.478
103	375	Quadrant.6	0.277
104	500	Quadrant.6	0.169
105	625	Quadrant.6	0.790
106	750	Quadrant.6	0.207
107	875	Quadrant.6	0.787
108	1000	Quadrant.6	0.387
109	1125	Quadrant.6	0.780
110	1250	Quadrant.6	0.612
111	1375	Quadrant.6	0.929
112	1500	Quadrant.6	0.590
113	1625	Quadrant.6	0.761
114	1750	Quadrant.6	0.487
115	1875	Quadrant.6	0.844
116	2000	Quadrant.6	0.842
117	2125	Quadrant.6	0.736
118	2250	Quadrant.6	0.997
119	2375	Quadrant.6	0.806
120	2500	Quadrant.6	0.955
121	125	Quadrant.7	0.317
122	250	Quadrant.7	0.080
123	375	Quadrant.7	0.416
124	500	Quadrant.7	0.756
125	625	Quadrant.7	0.400
126	750	Quadrant.7	0.445
127	875	Quadrant.7	0.812
128	1000	Quadrant.7	0.321
129	1125	Quadrant.7	0.834
130	1250	Quadrant.7	0.714
131	1375	Quadrant.7	0.798
132	1500	Quadrant.7	0.844
133	1625	Quadrant.7	0.997
134	1750	Quadrant.7	0.938
135	1875	Quadrant.7	0.759
136	2000	Quadrant.7	0.725
137	2125	Quadrant.7	0.843
138	2250	Quadrant.7	0.995
139	2375	Quadrant.7	0.989
140	2500	Quadrant.7	0.896
141	125	Quadrant.8	0.001
142	250	Quadrant.8	0.082
143	375	Quadrant.8	0.614
144	500	Quadrant.8	0.345
145	625	Quadrant.8	0.336
146	750	Quadrant.8	0.760
147	875	Quadrant.8	0.571
148	1000	Quadrant.8	0.614
149	1125	Quadrant.8	0.572
150	1250	Quadrant.8	0.641

	MAP	Quadrant	Cover.m		MAP	Quadrant	Cover.m		MAP	Quadrant	Cover.m
151	1375	Quadrant.8	0.824	201	125	Quadrant.11	0.001	251	1375	Quadrant.13	0.808
152	1500	Quadrant.8	0.803	202	250	Quadrant.11	0.001	252	1500	Quadrant.13	0.916
153	1625	Quadrant.8	0.783	203	375	Quadrant.11	0.104	253	1625	Quadrant.13	0.660
154	1750	Quadrant.8	0.873	204	500	Quadrant.11	0.139	254	1750	Quadrant.13	0.966
155	1875	Quadrant.8	0.960	205	625	Quadrant.11	0.512	255	1875	Quadrant.13	0.588
156	2000	Quadrant.8	0.906	206	750	Quadrant.11	0.434	256	2000	Quadrant.13	0.813
157	2125	Quadrant.8	0.621	207	875	Quadrant.11	0.257	257	2125	Quadrant.13	0.851
158	2250	Quadrant.8	0.982	208	1000	Quadrant.11	0.682	258	2250	Quadrant.13	0.949
159	2375	Quadrant.8	0.997	209	1125	Quadrant.11	0.410	259	2375	Quadrant.13	0.929
160	2500	Quadrant.8	0.955	210	1250	Quadrant.11	0.736	260	2500	Quadrant.13	0.829
161	125	Quadrant.9	0.399	211	1375	Quadrant.11	0.842	261	125	Quadrant.14	0.001
162	250	Quadrant.9	0.391	212	1500	Quadrant.11	0.905	262	250	Quadrant.14	0.445
163	375	Quadrant.9	0.208	213	1625	Quadrant.11	0.738	263	375	Quadrant.14	0.299
164	500	Quadrant.9	0.531	214	1750	Quadrant.11	0.800	264	500	Quadrant.14	0.305
165	625	Quadrant.9	0.380	215	1875	Quadrant.11	0.894	265	625	Quadrant.14	0.392
166	750	Quadrant.9	0.533	216	2000	Quadrant.11	0.705	266	750	Quadrant.14	0.217
167	875	Quadrant.9	0.415	217	2125	Quadrant.11	0.909	267	875	Quadrant.14	0.096
168	1000	Quadrant.9	0.854	218	2250	Quadrant.11	0.850	268	1000	Quadrant.14	0.553
169	1125	Quadrant.9	0.776	219	2375	Quadrant.11	0.998	269	1125	Quadrant.14	0.645
170	1250	Quadrant.9	0.847	220	2500	Quadrant.11	0.937	270	1250	Quadrant.14	0.652
171	1375	Quadrant.9	0.497	221	125	Quadrant.12	0.125	271	1375	Quadrant.14	0.651
172	1500	Quadrant.9	0.724	222	250	Quadrant.12	0.283	272	1500	Quadrant.14	0.807
173	1625	Quadrant.9	0.908	223	375	Quadrant.12	0.633	273	1625	Quadrant.14	0.995
174	1750	Quadrant.9	0.891	224	500	Quadrant.12	0.141	274	1750	Quadrant.14	0.997
175	1875	Quadrant.9	0.857	225	625	Quadrant.12	0.535	275	1875	Quadrant.14	0.978
176	2000	Quadrant.9	0.938	226	750	Quadrant.12	0.399	276	2000	Quadrant.14	0.803
177	2125	Quadrant.9	0.939	227	875	Quadrant.12	0.599	277	2125	Quadrant.14	0.782
178	2250	Quadrant.9	0.942	228	1000	Quadrant.12	0.620	278	2250	Quadrant.14	0.957
179	2375	Quadrant.9	0.989	229	1125	Quadrant.12	0.186	279	2375	Quadrant.14	0.922
180	2500	Quadrant.9	0.701	230	1250	Quadrant.12	0.436	280	2500	Quadrant.14	0.929
181	125	Quadrant.10	0.181	231	1375	Quadrant.12	0.531	281	125	Quadrant.15	0.001
182	250	Quadrant.10	0.254	232	1500	Quadrant.12	0.911	282	250	Quadrant.15	0.107
183	375	Quadrant.10	0.008	233	1625	Quadrant.12	0.933	283	375	Quadrant.15	0.135
184	500	Quadrant.10	0.060	234	1750	Quadrant.12	0.795	284	500	Quadrant.15	0.481
185	625	Quadrant.10	0.305	235	1875	Quadrant.12	0.981	285	625	Quadrant.15	0.428
186	750	Quadrant.10	0.582	236	2000	Quadrant.12	0.776	286	750	Quadrant.15	0.688
187	875	Quadrant.10	0.742	237	2125	Quadrant.12	0.832	287	875	Quadrant.15	0.633
188	1000	Quadrant.10	0.670	238	2250	Quadrant.12	0.772	288	1000	Quadrant.15	0.957
189	1125	Quadrant.10	0.778	239	2375	Quadrant.12	0.967	289	1125	Quadrant.15	0.481
190	1250	Quadrant.10	0.618	240	2500	Quadrant.12	0.960	290	1250	Quadrant.15	0.626
191	1375	Quadrant.10	0.870	241	125	Quadrant.13	0.592	291	1375	Quadrant.15	0.708
192	1500	Quadrant.10	0.947	242	250	Quadrant.13	0.443	292	1500	Quadrant.15	0.989
193	1625	Quadrant.10	0.720	243	375	Quadrant.13	0.230	293	1625	Quadrant.15	0.853
194	1750	Quadrant.10	0.917	244	500	Quadrant.13	0.377	294	1750	Quadrant.15	0.788
195	1875	Quadrant.10	0.847	245	625	Quadrant.13	0.487	295	1875	Quadrant.15	0.979
196	2000	Quadrant.10	0.858	246	750	Quadrant.13	0.670	296	2000	Quadrant.15	0.779
197	2125	Quadrant.10	0.920	247	875	Quadrant.13	0.977	297	2125	Quadrant.15	0.971
198	2250	Quadrant.10	0.895	248	1000	Quadrant.13	0.799	298	2250	Quadrant.15	0.724
199	2375	Quadrant.10	0.735	249	1125	Quadrant.13	0.828	299	2375	Quadrant.15	0.895
200	2500	Quadrant.10	0.986	250	1250	Quadrant.13	0.377	300	2500	Quadrant.15	0.999

Table 5 – Forest Cover Proportion.

References

Aitchison, J., & Aitchison, J. (1986). *The statistical analysis of compositional data*. Springer Netherlands. https://books.google.fr/books?id=RHKmAAAAIAAJ

Douma, J. C., & Weedon, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression. *Methods in Ecology and Evolution*, *10*(9), 1412–1430.