UNIVERSIDADE FEDERAL DE JUIZ DE FORA
DEPARTAMENTO DE ESTATÌSTICA
CURSO DE BACHARELADO EM ESTATÌSTICA

MATHEUS RAMOS SIQUEIRA NUNES

APLICAÇÃO DA REGRESSÃO DIRICHLET NA MODELAGEM DO DESEMPENHO DOS CAMPEÕES DOS PRINCIPAIS CAMPEONATOS DE FUTEBOL

JUIZ DE FORA
2024

MATHEUS RAMOS SIQUEIRA NUNES

APLICAÇÃO DA REGRESSÃO DIRICHLET NA MODELAGEM DO DESEMPENHO DOS CAMPEÕES DOS PRINCIPAIS CAMPEONATOS DE FUTEBOL

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Bacharel em Estatística.

Orientador: Prof. Tiago Maia Magalhães

JUIZ DE FORA
2024

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Nunes, Matheus Ramos Siqueira .
Aplicação da Regressão de Dirichlet na Modelagem de Desempenho dos Campeões dos Principais Campeonatos de Futebol / Matheus Ramos Siqueira Nunes. -- 2024.
63 f.

Orientador: Tiago Maia Magalhães
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2024.

1. Dados Composicionais. 2. Regressão de Dirichlet. 3. Simulação de Monte Carlo. I. Magalhães, Tiago Maia, orient. II. Título.

MATHEUS RAMOS SIQUEIRA NUNES

APLICAÇÃO DA REGRESSÃO DIRICHLET NA MODELAGEM DO DESEMPENHO DOS CAMPEÕES DOS PRINCIPAIS CAMPEONATOS DE FUTEBOL

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Bacharel em Estatística.

Aprovado em ___ de _____ de _____

BANCA EXAMINADORA:

Tiago Maia Magalhães
Doutor em Estatística – USP – Orientador

Bárbara da Costa Campos Dias
Doutora em Estatística - UFMG

Clécio da Silva Ferreira
Doutor em Estatística - USP

Para toda minha família.

RESUMO

Dados composicionais descrevem parte de um todo, geralmente somando um, no caso de proporções, ou cem, no caso de porcentagens. Os métodos tradicionais de análise multivariada não se mostram adequados para a manipulação deste tipo de dados. Nesta monografia são abordados os métodos para análises de dados composicionais, desde o espaço amostral adequado, chamado Simplex, operações composicionais, as transformações log-razão até a visualização de dados. Em particular, é abordado o caso em que as composições são as variáveis preditoras de uma análise de regressão. Assim, foi discutido o modelo de regressão Dirichlet e o processo de estimação pelo Método da Máxima Verossimilhança. Para avaliar a eficácia deste modelo de regressão, são feitas simulações através do método de Monte Carlo para a estimação de parâmetros de regressão, em que se verificou tendência de aproximação das estimativas aos parâmetros conforme aumento do tamanho da amostra. Posteriormente, são feitas aplicações nas bases de dados do Lago Ártico, de campeões brasileiros de futebol e de campeões nacionais de futebol no ano de 2022. Nestas aplicações são empregadas as técnicas de visualização, estatísticas descritivas e regressão adequadas aos dados. Para os dois últimos conjuntos de dados, foram escolhidas a composição de vitórias, empates e derrotas como variáveis resposta e ajustado um modelo de regressão Dirichlet, sendo possível perceber o comportamento desta composição nos campeões nacionais de futebol.

Palavras-chave: dados composicionais; regressão Dirichlet; simulação de Monte Carlo.

ABSTRACT

Compositional data describe parts of a whole, usually summing to one in the case of proportions, or one hundred in the case of percentages. Traditional multivariate analysis methods are not suitable for handling this type of data. This thesis addresses methods for analyzing compositional data, covering topics such as the appropriate sample space, known as the Simplex, compositional operations, log-ratio transformations, and data visualization. In particular, it focuses on cases where compositions are the predictor variables in a regression analysis. The Dirichlet regression model and the estimation process through the Maximum Likelihood Method are discussed. To evaluate the effectiveness of this regression model, simulations were performed using the Monte Carlo method for parameter estimation, revealing a tendency for the estimates to converge toward the parameters as the sample size increased. Subsequently, applications were made to the Arctic Lake dataset, Brazilian football champions, and national football champions in 2022. In these applications, visualization techniques, descriptive statistics, and regression appropriate for compositional data were employed. For the latter two datasets, the composition of wins, draws, and losses was chosen as the response variable, and a Dirichlet regression model was fitted, allowing for insights into the behavior of this composition among national football champions.

Keywords: compositional data; Dirichlet regression; Monte Carlo simulation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A aquisição, análise e interpretação de dados são fundamentais em diversas áreas da ciência e setores industriais. Entretanto, à medida que a capacidade de coletar informações continua a crescer exponencialmente, surge uma complexidade adicional: a presença de dados composicionais. Eles descrevem partes de um todo, que são comumente apresentados como vetores de proporções, concentrações, frequências ou porcentagens (PAWLOWSKY-GLAHN; EGOZCUE; TOLOSANA-DELGADO, 2015). São dados não-negativos que, geralmente, somam um ou 100. É comum encontrar aplicações nas mais variadas áreas da ciência como Geologia, Nutrição, Química e Medicina, por exemplo. Seguem alguns exemplos de problemas composicionais para ilustrar aplicações práticas.

Imagine que uma pessoa anote quanto tempo ela passa em atividades específicas durante um dia, como dormir, comer, viajar, trabalhar e relaxar, até que tenha contabilizado um total de 24 horas - agora ela terá um conjunto de dados composicionais. Essa pessoa pode repetir esse processo ao longo de vários dias, obtendo um conjunto diferente de valores que quantificam suas atividades a cada dia, até que tenha uma tabela ou matriz desses conjuntos de observações, geralmente salva em uma planilha. As linhas dessa tabela representarão os dias e as colunas representarão as diferentes atividades. A característica importante desses dados é que a soma de cada linha sempre será igual a 24 horas: essa propriedade de soma constante para cada conjunto de valores torna os dados composicionais. Por que não tentar fazer isso ao longo de algumas semanas? Se o fizer, terá sua própria tabela de dados composicionais que poderá analisar, o que permitirá que ela compreenda melhor seu padrão diário de comportamento (GREENACRE, 2018).

Um outro exemplo é um experimento, em que foi observado o leite de trinta vacas com o intuito de melhorar a qualidade do mesmo. O leite foi avaliado quanto à composição da dieta antes e depois de um regime dietético e hormonal rigorosamente controlado ao longo de um período de oito semanas. Embora variações sazonais na qualidade do leite pudessem ter sido consideradas negligenciáveis durante esse período, foi decidido ter um grupo de controle de trinta vacas mantidas nas mesmas condições, mas sob um regime regular estabelecido. As sessenta vacas foram alocadas aleatoriamente aos grupos de controle e tratamento. O conjunto completo de composições de leite antes e depois para as sessenta vacas mostra as proporções, em peso, de proteínas, gordura do leite, carboidratos, cálcio, sódio e potássio em relação ao conteúdo dietético total. O objetivo do experimento era determinar se o novo regime produziu alguma mudança significativa na composição do leite, portanto, é essencial ter uma ideia clara de como a mudança nos dados composicionais é caracterizada por alguma operação significativa. Uma questão fundamental aqui é, portanto, como formular hipóteses de mudança nas composições e, de fato, como podemos investigar toda a gama dessas hipóteses (AITCHISON, 2005).

Além desses, outros exemplos são geógrafos que estudam o uso da terra podem encontrar uma proporção maior de terras agrícolas dedicadas a uma cultura específica onde há mais terra disponível. E cientistas políticos que estudam as porcentagens de votos para partidos políticos também estão interessados em quantos votos foram lançados no total. Da mesma forma, em ecologia comunitária, bem como em botânica, são feitas amostras de quantidades físicas iguais (por exemplo, volumes iguais em biologia marinha ou áreas iguais em botânica, chamadas "quadrats"), e as espécies presentes em cada amostra são identificadas e contadas. Os ecologistas estão interessados tanto na composição das amostras, ou seja, qual proporção de espécies A, B, C, etc., é encontrada em uma amostra em relação ao total encontrado, quanto se o total encontrado na amostra está relacionado com a composição, por exemplo, se a espécie A é proporcionalmente mais comum em amostras que são mais abundantes em geral. De interesse adicional no mesmo contexto está a diversidade de cada amostra, ou seja, quantos tipos diferentes de espécies estão presentes, o que também afeta seus valores composicionais (PAWLOWSKY-GLAHN; EGOZCUE; TOLOSANA-DELGADO, 2015).

Apesar de estar nas mais diversas áreas do conhecimento, esse tipo de dados é particularmente complexo de trabalhar-se, estatisticamente falando, pois apresenta relação entre duas ou mais variáveis, sem nenhuma explicação lógica ou significado teórico, o que Pearson em 1897 chamou de correlação espúria. Por isso, as análises multivariadas tradicionalmente usadas não dão uma interpretação correta das correlações encontradas.

Então, na década de 1980, o escocês John Aitchison desempenhou um papel fundamental na introdução do conceito de composição de dados e no desenvolvimento de técnicas estatísticas essenciais para sua análise. Ele introduziu abordagens que se baseiam em diversas transformações log-razão. Essas transformações possibilitaram a utilização de técnicas tradicionais de Análise Multivariada nos dados transformados, o que, por sua vez, permitiu que as conclusões obtidas fossem reinterpretadas em relação aos dados originais. Desde então, cada vez mais conhecimento sobre o tema é produzido e consequentemente aplicado em análises de variadas áreas.

1.1 OBJETIVO

Esta monografia tem como objetivo realizar uma análise de regressão Dirichlet de dados composicionais, passando pelos métodos mais clássicos de análise composicional até um estudo de simulação, avaliando o desempenho na estimação de parâmetros de regressão e aplicação em três bases de dados reais.

1.2 JUSTIFICATIVA

Conforme posto na Introdução, quando são utilizadas técnicas tradicionais de análise estatística em dados composicionais, ocorrem erros causados pela correlação espúria, de-

finida por Pearson. Assim, o trabalho se justifica por mostrar o processo de trabalhar de forma congruente os dados composicionais no espaço amostral adequado, desde operações básicas, transformações e visualizações no mesmo, além de apresentar como funciona um modelo de regressão composicional multivariado através de teoria, simulação e aplicações práticas.

2 DADOS COMPOSICIONAIS

Conforme apresentado no capítulo anterior, os dados composicionais apresentam uma correlação espúria e por isso requerem métodos adequados para suas aplicações. Assim, neste capítulo são apresentados o seu espaço amostral, as operações, propriedades e transformações das composições, além de medidas descritivas e gráficos para dados composicionais.

2.1 ESPAÇO SIMPLEX

O espaço amostral de dados composicionais é o Simplex, definido da seguinte forma:

$$S^D = \left\{ \mathbf{y} = [y_1, y_2, \ldots, y_D] \Big| y_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} y_i = c \right\},$$

em que $\mathbf{y}$ é o vetor com D partes, cada uma delas com informações reais positivas e com informação relativa de um todo. Além disso, c é uma constante qualquer, mas frequentemente são apresentadas como um ou cem. Isso ocorre porque existe uma equivalência de composições, ou seja, é possível multiplicar y e c pela mesma constante positiva λ para obter um vetor composicional equivalente. Assim, essa propriedade é muito útil não só para apresentar os resultados como porcentagem ou frequência, mas para fazer equivalência entre diferentes unidades de medição. É o que permite facilmente comparar composições de minerais em gramas e kilogramas, por exemplo.

2.1.1 Operações Composicionais

Muitas vezes os pesquisadores não estão interessados em todos os elementos da composição. Assim, é possível fazer subcomposições apenas com as variáveis de interesse. Deste modo, é muito comum a utilização de subcomposições, as quais apresentam as mesmas propriedades da composição. O modo usual de transformar a composição em uma subcomposição é através da chamada operação de fecho, definida a seguir:

$$C(\mathbf{y}) = \left[\frac{ky_1}{\sum_{i=1}^{D} y_i}, \frac{ky_2}{\sum_{i=1}^{D} y_i}, \ldots, \frac{ky_D}{\sum_{i=1}^{D} y_i} \right], \tag{2.1}$$

em que o resultado da operação de fecho é uma reescala do vetor inicial e a constante k é a soma dos seus componentes. Assim, é possível selecionar os atributos desejados e realizar a operação de fecho para trabalhar com alguns dos elementos do vetor original, mantendo as informações mais importantes. Essa operação também pode ser realizada para apresentar os dados numa escala desejada.

Já a operação de soma no espaço Simplex, neste caso também conhecida como perturbação. A soma de dois vetores composicionais $\mathbf{y}$ e $\mathbf{z}$ é dada por:

$$\mathbf{y} \oplus \mathbf{z} = C(y_1 z_1, y_2 z_2, \ldots, y_d y_z),$$

em que $C(.)$ é a operação de fecho definida na Equação (2.1). Existe também o produto por escalar no Simplex, definida como:

$$\mathbf{y} \otimes k = C(y_1^k, y_2^k, \ldots, y_D^k),$$

em que $C(.)$ é a operação de fecho e k é uma constante real qualquer. Já o produto interno de Aitchison é definido como:

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{D} \sum_{i>j}^{D} y_i \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

Assim, a norma de um vetor composicional pode ser definida:

$$\|\mathbf{y}\|_A = \sqrt{\langle \mathbf{y}, \mathbf{y} \rangle_A}.$$

A distância de Aitchison entre dois vetores composicionais é:

$$d_A(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{x}, \mathbf{y}\|_A = \sqrt{\frac{1}{D} \sum_{i-j}^{D} \sum_{i>j}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}$$

.

Define-se a média geométrica de um vetor composicional como:

$$g(\mathbf{y}) = \sqrt[D]{y_1 . y_2 \ldots y_D}.$$

2.1.2 Propriedades de Dados Composicionais

Três condições devem ser completamente atendidas por qualquer método estatístico que for aplicado a composições: invariância de escala, invariância de permutação e coerência subcomposicional (AITCHISON, 1982).

A característica mais importante de dados composicionais é que eles carregam apenas informações relativas (PAWLOWSKY-GLAHN; EGOZCUE; TOLOSANA-DELGADO, 2015). Brevemente citada na seção anterior, a invariância de escala pode ser matematicamente definida como: seja $f(.)$ uma função qualquer definida em $\mathbb{R}_+^D$. Esta função possui invariabilidade de escala se para qualquer valor real positivo $\lambda \in \mathbb{R}_+$ e para qualquer composição $\mathbf{y} \in S^D$ satisfaz $f(\lambda \mathbf{y}) = f(\mathbf{y})$. Isto é, ela produz o mesmo resultado para todos os vetores composicionalmente equivalentes. Assim, este princípio diz que qualquer alteração na escala dos dados originais não tem efeito. Portanto, se os dados originais

forem multiplicados por qualquer fator de escala k, como uma mudança de unidades, os dados composicionais permanecerão inalterados após a operação de fecho.

A invariância de permutação significa que os resultados não dependem da ordem em que as partes aparecem em uma composição. Embora em um conjunto de dados composicionais as partes estejam todas ordenadas da mesma forma para cada amostra, as partes devem poder ser reordenadas em todo o conjunto de dados, ou seja, as colunas do conjunto de dados são permutadas sem afetar os resultados.

A coerência subcomposicional significa que os resultados obtidos para um subconjunto de partes de uma composição, ou seja, uma subcomposição, devem permanecer os mesmos que na composição completa. Na subcomposição cada elemento permanece na mesma proporção em relação aos demais quando comparada ao vetor composicional original.

2.2 TRANSFORMAÇÕES COMPOSICIONAIS

Considere uma composição $\mathbf{y} = [y_1, y_2, \ldots, y_D] \in S^D$ e os coeficientes $k_i \in \mathbb{R}$ para todo $i = 1, 2, \ldots, D$. Logcontraste ou lograzão é uma função:

$$f(\mathbf{y}) = \sum_{i=1}^{D} k_i \ln y_i, \text{ com } \sum_{i=1}^{D} k_i = 0.$$

Em algumas aplicações, logcontrastes são facilmente interpretadas. Um exemplo típico é o equilíbrio químico. Considere uma composição química $D-$parte, denotada por y e expresso em proporções molares. Uma reação química envolvendo quatro elementos pode ser

$$k_1 y_1 + k_2 y_2 \rightleftharpoons k_3 y_3 + k_4 y_4,$$

em que outras partes não estão envolvidas. Os $k_i's$, chamados coeficientes estequiométricos, são normalmente conhecidos. Se a reação é conservacionista de matéria, então $k_1 + k_2 = k_3 + k_4$. A função

$$\ln \frac{y_1^{k_1} y_2^{k_2}}{y_3^{k_3} y_4^{k_4}},$$

é um logcontraste porque $k_1 + k_2 - k_3 - k_4 = 0$. Sempre que essa reação química está em equilíbrio, essa logcontraste deve ser constante e, portanto, desvios dele pode ser interpretado como desvios do equilíbrio (PAWLOWSKY-GLAHN; EGOZCUE; TOLOSANA-DELGADO, 2015, apud Boogaart et al., 2013).

Na década de 1980, passou a ser conhecida como análise de lograzão para problemas de dados composicionais, devido à compreensão da importância do princípio da invariância de escala. Isso, juntamente com a consciência de que os logaritmos de razões são matematicamente mais tratáveis do que as próprias razões, levou à defesa de uma técnica de transformação envolvendo lograzões dos componentes. Havia dois concorrentes

óbvios para isso. Então, a chamada transformação de lograzão aditiva (em inglês *additive logratio transformation*, daí a sigla alr) $alr : S^D \to \mathbb{R}^{D-1}$ é definida por:

$$\mathbf{z} = alr(\mathbf{y}) = \left[\log\left(\frac{y_1}{y_D}\right), \log\left(\frac{y_2}{y_D}\right), \ldots, \log\left(\frac{y_{D-1}}{y_D}\right) \right],$$

em que os primeiros $D-1$ componentes são divididos pelo último componente y_D. A função inversa, $alr^{1-} : \mathbb{R}^{D-1} \to S^D$ é definida por:

$$\mathbf{y} = alr^{-1}(\mathbf{z}) = C(\exp z_1, \exp z_2, \ldots, \exp z_{D-1}),$$

em que C é a operação de fecho anteriormente definida. Essa transformação leva a composição para todo o espaço $\mathbb{R}^{D-1}$, o que permite usar análises multivariadas padrão sem restrições nos dados transformados. Devido à natureza biunívoca da transformação, pode-se transferir quaisquer inferências de volta para o Simplex e para os componentes da composição.

A transformação alr é assimétrica em relação às partes, e às vezes é conveniente tratar as partes de maneira simétrica. Isso pode ser alcançado pela chamada *centred logratio transformation* (clr):

$$\mathbf{z} = clr(\mathbf{y}) = \left[\log\left(\frac{y_1}{g(x)}\right), \log\left(\frac{y_2}{g(x)}\right), \ldots, \log\left(\frac{y_D}{g(x)}\right) \right],$$

em que $g(x)$ é a média geométrica. A função inversa é, portanto,

$$\mathbf{y} = clr^{-1}(\mathbf{z}) = C(\exp z_1, \exp z_2, \ldots, \exp z_{D-1}).$$

Também existe a transformação *isometric log-ratio transformation* (ilr), que é uma transformação que fornece as coordenadas de qualquer composição em relação a uma base ortonormal específica. Seja $[\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_{D-1}}]$ uma base ortonormal de S^D e considere a matriz $\mathbf{\Psi}$ de tamanho $D-1, D$ na qual as linhas são o vetor $\mathbf{\Psi}_i = clr(\mathbf{e_i}), i = 1, 2, \ldots, D-1$. Esta matriz é chamada de matriz de contrastes associada à base ortonormal . Cada linha é denominada um contraste. Uma base ortonormal satisfaz $\langle \mathbf{e_i}, \mathbf{e_j} \rangle_A = \delta_{ij}$, em que δ_{ij} equivale ao delta de Kronecker, que é nulo para $i \neq j$ e um para $i = j$. Isso implica que

$$\mathbf{\Psi}\mathbf{\Psi}^T = \mathbf{I}_{D-1}, \mathbf{\Psi}\mathbf{\Psi}^T = \mathbf{I}_D - \left(\frac{1}{D}\right)\mathbf{1}_D^T\mathbf{1}_D,$$

em que $\mathbf{1}_{D-1}, \mathbf{1}_D$ são as matrizes identidade das dimensões $D-1$ e D respectivamente. $\mathbf{1}_D$ é o vetor de uns de tamanho D. Escolhida a base ortonormal, a composição $y \in S^D$ é expresso como

$$y_i^* = \langle \mathbf{y}, \mathbf{e_i} \rangle_A.$$

Assim, a transformação ilr pode ser definida como

$$\mathbf{y}^* = ilr(\mathbf{y}) = clr(\mathbf{y}).\mathbf{\Psi}^T.$$

E a sua função inversa é

$$ilr^{-1}(\mathbf{y}^*) = C(exp(\mathbf{y}^*\mathbf{\Psi})) = \mathbf{y}.$$

2.3 ANÁLISE DESCRITIVA DE DADOS COMPOSICIONAIS

No caso dos dados composicionais, as estatísticas descritivas tradicionais não são muito informativas. De acordo com (PAWLOWSKY-GLAHN; EGOZCUE; TOLOSANA-DELGADO, 2015), "em particular, a média aritmética e a variância ou o desvio padrão das partes individuais não se encaixam na geometria de Aitchison como valor de tendência central e medidas de dispersão". Sobre essas medidas, Boogart e Tolosana-Delgado (2013) afirmam que:

> Aitchison (1986) destacou que estatísticas significativas de tendência central, de dispersão e de codependência devem possuir algumas propriedades de invariância em relação às manipulações mais razoáveis do conjunto de dados:
>
> - Traduzir (perturbar com uma composição constante) um conjunto de dados não deve alterar a estrutura de dispersão ou de codependência de nenhuma forma, apenas traduzir o centro.
>
> - Redimensionar (por exemplo, alterar as unidades de) um conjunto de dados deve simplesmente redimensionar o centro, enquanto tem uma influência quadrática na dispersão e na codependência (porque as variâncias têm as unidades do conjunto de dados, mas ao quadrado).

Assim, o centro ou média composicional é definido por:

$$\bar{y} = clr^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} clr(\mathbf{y}_n)\right) = C\left[\exp\left(\frac{1}{N}\sum_{n=1}^{N}\ln(\mathbf{y}_n)\right)\right],$$

em que $\bar{y}$ representa o centro de um vetor $\mathbf{Y}$ com N observações e D partes composicionais. Assim, $\mathbf{y}_n$ representa a enésima observação, na enésima linha de $\mathbf{Y}$ e $C(.)$ é a operação de fecho.

Já como uma medida de variabilidade dos dados composicionais, não existe um equivalente como no caso do centro, que pode ser construído diretamente para as composições originais. Em vez disso, a atenção é tradicionalmente direcionada para a informação de origem em uma composição, para razões logarítmicas par a par. Isso resulta na chamada matriz de variação (Aitchison 1986), que é formada pelas variâncias de todas as razões logarítmicas par a par. A matriz de variação é definida como:

$$\mathbf{T}(y) = [t_{ij}] = \left[var\left\{ \log\left(\frac{y_i}{y_j}\right)\right\}\right].$$

em que $\mathbf{T}(y)$ é uma matriz simétrica, com diagonal principal de zeros e não pode ser expressa como uma matriz de covariância padrão de um vetor. Além disso, é possível notar que cada entrada dessa matriz não depende da escala, uma vez que ela á cancelada ao se tomar as razões.

Uma medida de dispersão global de uma amostra composicional é a variância total, dada por:

$$totvar[\mathbf{Y}] = \frac{1}{2D}\sum_{i,j=1}^{D} var\left(\ln\frac{y_i}{y_j}\right) = \frac{1}{2D}\sum_{i,j=1}^{D} t_{ij}$$

2.4 VISUALIZAÇÃO DE DADOS COMPOSICIONAIS

Saber como visualizar dados composicionais é especialmente útil para compreender e interpretar suas propriedades. A restrição de soma fixa nos dados composicionais leva a uma representação geométrica especial das composições em um espaço Simplex. A forma mais simples de um Simplex é um triângulo, que contém composições com três partes. Composições com quatro partes são contidas em um tetraedro tridimensional. Simplex de dimensões mais altas, que não podem ser visualizados diretamente, contêm composições com mais de quatro partes (GREENACRE, 2018).

No entanto, antes de passar ao espaço Simplex é possível visualizar os dados de forma preliminar e simples através de gráficos de barras. As Figuras 1 e 2 referem-se ao conjunto de dados *Vegetables*, disponível no pacote *easyCODA* do R. Neste conjunto estão presentes os percentuais de carboidrato, gordura e proteína de dez vegetais (fonte: *US Department of Agriculture, https://ndb.nal.usda.gov/ndb/nutrients/index*).

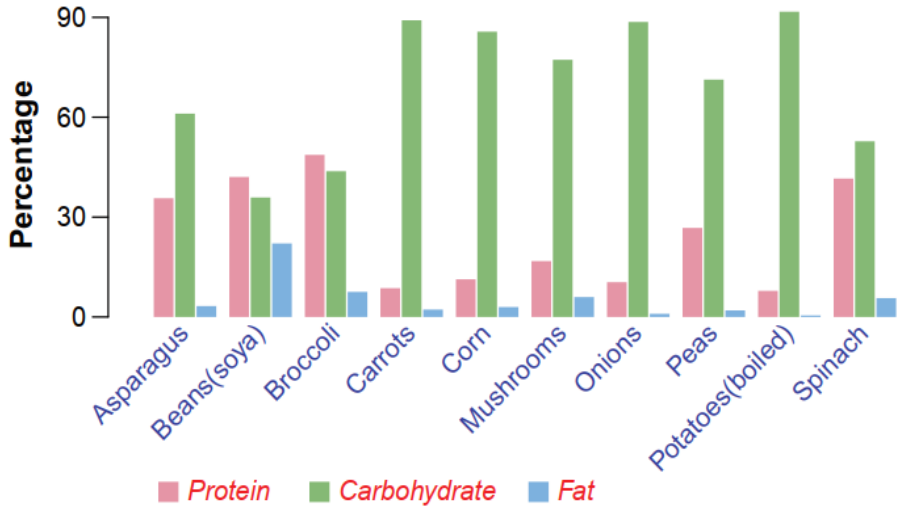

Figura 1 – Gráfico de Barras

Fonte: Greenacre, Michael (2019, p. 10)

Apesar de não estar incorreto, não é possível perceber pela Figura 1 que tratam-se de dados composicionais. Já pela Figura 2 tanto tem-se a ideia de composição quanto é mais fácil a comparação entre os percentuais de cada vegetal.

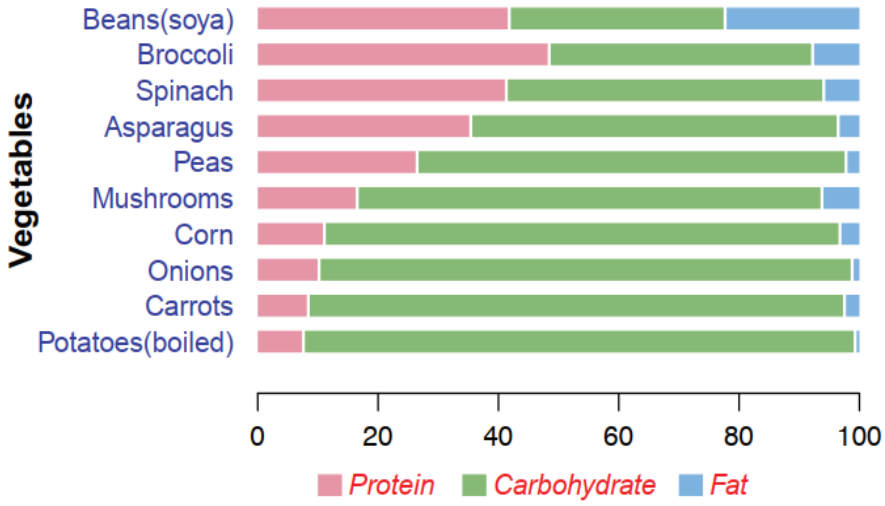

Figura 2 – Gráfico de Barras Horizontais Estacadas

Fonte: Greenacre, Michael (2019, p. 10)

Uma outra forma de visualizar dados composicionais são os diagramas ternários, que são semelhantes a gráficos de dispersão, porém com uma forma triangular, adequada às composições de três partes. Esse tipo de gráfico permite observar três variáveis bidimen-

sionalmente. A Figura 3 ilustra um diagrama ternário. Os dados são de geoquímica de sedimentos glaciais (fonte: Tolosana-Delgado and von Eynatten, 2010).

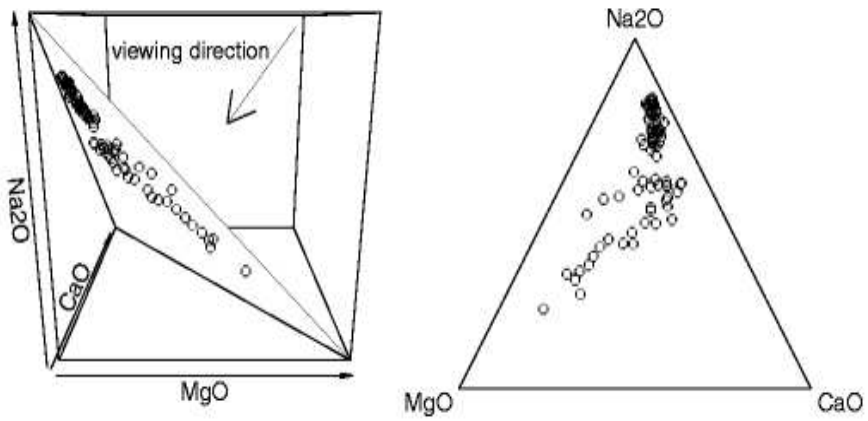

Figura 3 – Diagrama Ternário no Espaço Tridimensional Original

Fonte: van den Boogaart e Tolosana-Delgado (2013, p. 26)

Já a Figura 4 ilustra a leitura e legenda adequadas a um diagrama ternário.

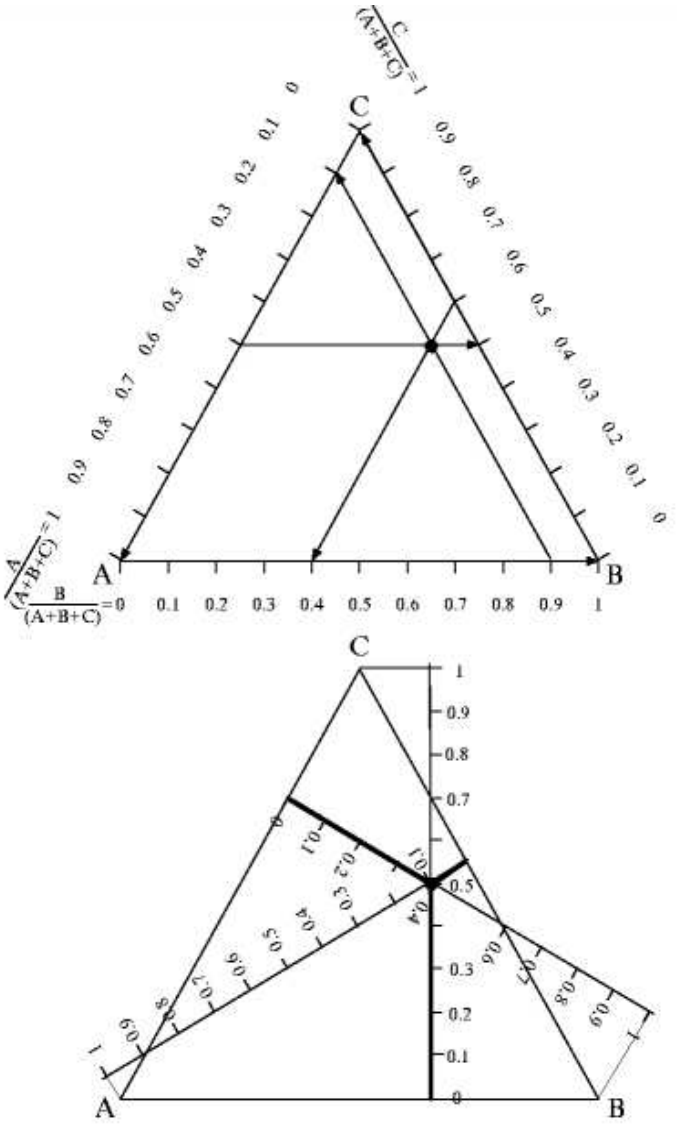

Figura 4 – Interpretação do Diagrama Ternário

Fonte: van den Boogaart e Tolosana-Delgado (2013, p. 27)

Os gráficos de radar também podem ser utilizados para a visualização de dados composicionais. Neste gráfico, é possível comparar duas ou mais observações através de eixos que formam um polígono. No caso de composições, cada eixo representa uma parte composicional. A Figura 5 é um exemplo de gráfico de radar aplicado à dados composicionais. Os dados são das composições de medalhas de ouro, prata e bronze conquistadas pelo Brasil nas Jogos Olímpicos do Rio de Janeiro, Tóquio e Paris, realizados em 2016, 2021 e 2024, respectivamente. (ZALCMAN, 2024).

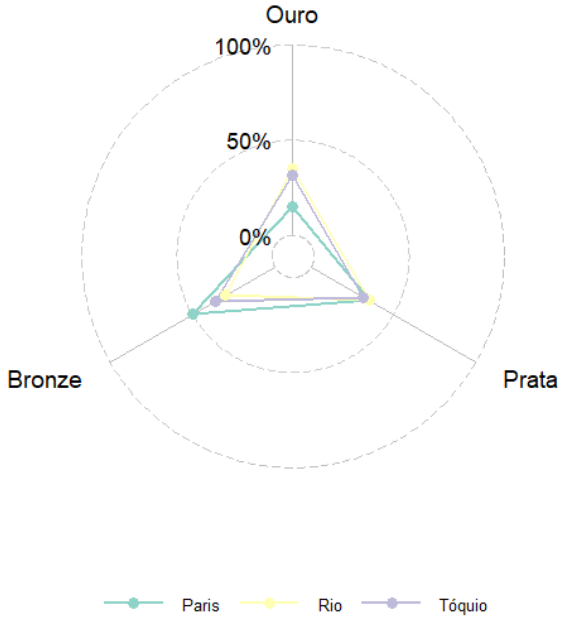

Figura 5 – Exemplo de Gráfico de Radar

Fonte: O Autor 2024

3 REGRESSÃO

Análise de regressão é uma técnica estatística que investiga e modela a relação entre variáveis (MONTGOMERY; PECK; VINING, 2012). Esta técnica é frequentemente utilizada para diversos propósitos como descrição de dados, estimação de parâmetros, predição e controle. Em particular, é tratado o caso em que as variáveis resposta formam uma composição através da chamada Regressão Dirichlet.

3.1 DISTRIBUIÇÃO DE DIRICHLET

Uma alternativa para trabalhar com dados composicionais não transformados é assumir uma estrutura que considere o espaço Simplex. Por exemplo, assumir que o vetor aleatório $\boldsymbol{Y} = (Y_1, \ldots, Y_D)^\top$ segue uma distribuição de Dirichlet com vetor de parâmetros $(\lambda_1, \ldots, \lambda_D)^T$, $\lambda_i \in (0, 1)$, $i = 1, \ldots, D$, com $D \geq 2$ e $\sum_{i=1}^{D} \lambda_i = 1$, e sua função densidade de probabilidade pode ser escrita como

$$f(\boldsymbol{y}|\boldsymbol{\lambda}) = \frac{\Gamma(\sum_{i=1}^{D} \lambda_i)}{\Pi_{i=1}^{D} \Gamma(\lambda_i)} \left[\prod_{i=1}^{D} y_i^{\lambda_i - 1} \right],$$

em que $\Gamma(\cdot)$ é a função gama, isto é, $\Gamma(\lambda) = \int_0^\infty v^{\lambda-1} e^{-v} dv$. Se $\boldsymbol{Y} \sim \text{Dirichlet}(\boldsymbol{\lambda})$, e seja $\lambda_0 = \sum_{i=1}^{D} \lambda_i$ tem-se que

$$\begin{aligned} \mathbb{E}(Y_i) &= \frac{\lambda_i}{\lambda_0}, \ i = 1, \ldots, D; \\ Var(Y_i) &= \frac{\lambda_i(\lambda_0 - \lambda_i)}{\lambda_0^2(\lambda_0 + 1)}; \\ Cov(Y_i, Y_j) &= \frac{-\lambda_i \lambda_j}{\lambda_0^2(\lambda_0 + 1)}; \ i \neq j. \end{aligned}$$

A distribuição de Dirichlet é a generalização da distribuição Beta para duas ou mais variáveis resposta. Assim como na distribuição Beta, existe uma segunda parametrização possível na distribuição de Dirichlet. Pode-se estabelecer relação entre as duas parametrizações com a igualdade $\boldsymbol{\lambda} = \boldsymbol{\mu}\phi$, em que $\mu_j \in (0, 1)$, para $j = 1, \ldots, D$ e $\phi > 0$. Deste modo, a função densidade de probabilidade na parametrização alternativa é escrita como

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \phi) = \frac{\Gamma(\sum_{i=1}^{D} \mu_i \phi)}{\Pi_{i=1}^{D} \Gamma(\mu_i \phi)} \left[\prod_{i=1}^{D} y_i^{\mu_i \phi - 1} \right].$$

$$\mathbb{E}(Y_i) = \mu_i,\ i = 1, \ldots, D;$$
$$Var(Y_i) = \frac{\mu_i(1-\mu_i)}{\phi+1};$$
$$Cov(Y_i, Y_j) = \frac{-\mu_i\mu_j}{\phi+1};\ i \neq j.$$

Na parametrização alternativa, tem-se a facilidade de interpretação tanto dos parâmetros μ, esperança, e ϕ, precisão, quanto dos coeficientes de regressão. Já na parametrização tradicional apresentada é mais flexível para a seleção de modelos e pode ser adequada para tarefas de modelagem complexas (MAIER, 2014).

3.1.1 Regressão Dirichlet

A seguir, será proposta uma estrutura de regressão que permite a modelagem de relações entre vetores com distribuição de Dirichlet e um conjunto de variáveis explicativas. O modelo proposto é definido estabelecendo relações entre os parâmetros que indexam a distribuição de Dirichlet e preditores lineares de variáveis explicativas. Assim, seja um conjunto de variáveis tais que $\boldsymbol{Y_i} \sim \text{Dirichlet}(\boldsymbol{\lambda_i})$ para $i = 1, \ldots, D$, o modelo de regressão Dirichlet pode ser definido como

$$g(\boldsymbol{\lambda_i}) = \boldsymbol{\eta_i} = \boldsymbol{X\beta_i},$$

em que $g(.)$ é a função de ligação log para este modelo, uma vez que $\lambda_i > 0$. O preditor linear é η_i, $\boldsymbol{\beta_i} = (\beta_{1i}, \ldots, \beta_{pi})$ é o vetor de coeficientes de regressão desconhecidos para o componente i e $\boldsymbol{X}$ é a matrix das n observações das p covariáveis conhecidas.

Na parametrização alternativa, uma estrutura de regressão também pode ser incorporada da seguinte forma: $\boldsymbol{Y_i} \sim \text{Dirichlet}(\boldsymbol{\mu_i}, \phi)$, com

$$g_\mu(\boldsymbol{\mu_i}) = \boldsymbol{\eta_{\mu i}} = \boldsymbol{X\beta_i},$$
$$g_\phi(\phi) = \boldsymbol{\eta_\phi} = \boldsymbol{Z\gamma},$$

em que $g_\mu(.)$ e $g_\phi(.)$ são as duas funções de ligação. Como os valores da média estão no intervalo (0,1), a função logito é escolhida como função de ligação. Já para a variância é escolhida a função logarítmica, pois $\phi > 0$. Nesta parametrização, adota-se a estratégia da multinomial logito (Bunch, Louviere e Anderson, 2006), ou seja, um dos componentes como base e todos seus coeficientes de regressão assumem valor zero. Neste trabalho será escolhido o primeiro componente. Assim,

$$\mu_1 = \frac{1}{\sum_{j=1}^{D} \exp(\boldsymbol{X\beta_j})},\ \mu_i = \frac{\exp(\boldsymbol{X\beta_1})}{\sum_{j=1}^{D} \exp(\boldsymbol{X\beta_j})},\ \text{para } i = 2, \ldots, D,$$

em que $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \ldots, \beta_{pi})^T, i = 2, \ldots, D$.

3.1.2 Método da Máxima Verossimilhança

O princípio da Máxima Verossimilhança indica que deve-se escolher o valor do parâmetro desconhecido que maximiza a probabilidade de obter-se a amostra particular observada estar sob as suposições do modelo estatístico. Assim, define-se a função de Máxima Verossimilhança de n variáveis $X_1, X_2, \ldots, X_n$ como a densidade conjunta das n variáveis aleatórias, a qual é uma função do parâmetro desconhecido θ. O Estimador de Máxima Verossimilhança deste parâmetro θ é o ponto que maximiza a função de Máxima Verossimilhança. Deste modo, para dado um conjunto de observações constituídas de n composições Y_i e n observações de p covariáveis $\boldsymbol{X}_j (j = 1, \ldots, n)$ a correspondente função de log-verossimilhança da distribuição de Dirichlet é

$$\ell(\boldsymbol{\theta}) = \log \sum_{j=1}^{n} \Gamma(\sum_{i=1}^{D} \lambda_i) - \sum_{j=1}^{n} \sum_{i=1}^{D} \log \Gamma(\lambda_i) + \sum_{j=1}^{n} \sum_{i=1}^{D} (\lambda_i - 1) \log y_i$$

Seja $\boldsymbol{\theta}$ a forma vetorizada da matriz dos parâmetros, a função escore tem a forma

$$\boldsymbol{U}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{X}^T \boldsymbol{C},$$

em que $\boldsymbol{X}$ é a matriz de variáveis explicativas desta regressão de tamanho $n \times p$ e $\boldsymbol{C}$ é a matriz de tamanho $n \times D$ com o ij-ésimo elemento $c_{ij} = g_i^{'}(\eta_{ij})\{\psi(\lambda_0) - \psi(\lambda_{ij}) + \log(Y_{ij})\}$, em que $g_i^{'}$ é a primeira derivada de g_i com respeito ao seu argumento e ψ representa a primeira derivada do log da função gama. O estimador de máxima verossimilhança de $\boldsymbol{\theta}$ pode ser obtido igualando essas pD derivadas a zero e resolvendo o sistema resultante. Ajustar uma distribuição de Dirichlet com parâmetros constantes é direto, e existem pacotes numéricos para a realização de tal tarefa. No entanto, a dificuldade surge quando tenta-se estender a estimativa para a regressão Dirichlet. Como não é possível encontrar expressões fechadas, métodos numéricos são necessários para calcular as Estimativas de Máxima Verossimilhança. Valores iniciais e políticas de regularização devem ser escolhidos com cuidado para que o algoritmo de otimização convirja. Assim, Hijazi e Jernigan (2009) propuseram o seguinte método para escolher valores iniciais para a etapa de otimização:

1. Retirar r amostras com reposição, cada uma de tamanho $m (m < n)$ de X e Y.

2. Para cada amostra, ajustar um modelo de Dirichlet com parâmetros constantes e calcule a média das covariáveis correspondentes. Isso resultará em matrizes $\boldsymbol{A}$ de tamanho $r \times D$ e $\boldsymbol{W}$ de tamanho $r \times p$, onde $\boldsymbol{A}$ representa as estimativas de Máxima Verossimilhança para as r amostras e a linha w_i representa as médias das covariáveis na amostra i.

3. Ajustar por mínimos quadrados D modelos da forma $A_{i,j} = \alpha_j(\boldsymbol{w}_i) = \sum_P^{k=1} \beta_{jk} w_{ik}$.

4. Usar os coeficientes ajustados $\beta_{i,j}$ como valores iniciais.

Em Melo, Vasconcellos e Lemonte (2009) é demonstrado que a matrix de informação $\boldsymbol{J}(\boldsymbol{\theta})$ pode ser obtida como

$$\boldsymbol{J}(\boldsymbol{\theta}) = (\boldsymbol{I}_p \otimes \boldsymbol{X})^T \boldsymbol{M} (\boldsymbol{I}_p \otimes \boldsymbol{X}),$$

em que $\boldsymbol{I}_p$ é matriz identidade de tamanho $p \times p$ e $\otimes$ representa o produto de Kronecker. Também, $\boldsymbol{M}$ é uma matriz $np \times np$ particionada, definida como

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_{11} & \dots & \boldsymbol{M}_{1p} \\ \vdots & \ddots & \vdots \\ \boldsymbol{M}_{p1} & \dots & \boldsymbol{M}_{pp} \end{bmatrix},$$

em que cada $\boldsymbol{M}_{ab}$, com $a = 1, 2, \dots, p$ e $b = 1, 2, \dots, p$, é uma matriz diagonal com o iésimo elemento da diagonal dado por

$$m_i^{(ab)} = \begin{cases} -g_a^{''}(\eta_{ia}[\psi(\phi_i) + A_{ia}^*] - g_a^{'}(\eta_{ia})^2[\psi^{'}(\phi_i) - \psi^{'}(\lambda_{ia})] & , a = b; \\ -g_a^{'}(\eta_{ia}) g_b^{'}(\eta_{ib}) \psi^{'}(\phi_i) & , a \neq b; \end{cases}$$

com $A_{ia}^* = \log(y_i a) - \psi(\lambda_{ia})$. Assim, seja $\boldsymbol{K} = E(\boldsymbol{J})$ a matriz de informação esperada de Fisher para θ, tem-se que

$$\boldsymbol{K} = (\boldsymbol{I}_p \otimes \boldsymbol{X})^T \boldsymbol{W} (\boldsymbol{I}_p \otimes \boldsymbol{X}),$$

em que $\boldsymbol{W}$ é uma matriz $np \times np$ definida na forma particionada tal qual a matriz $\boldsymbol{M}$ onde cada $\boldsymbol{W}_{ab}, a = 1, \dots, p$ e $b = 1, \dots, p$, é uma matriz diagonal com o iésimo elemento na diagonal dado por

$$w_i^{(ab)} = \begin{cases} -g^{'}(\eta_{ia}^2[\psi^{'}(\phi_i) - \psi^{'}(\lambda_{ia})] & , a = b; \\ -g^{'}(\eta_{ia}) g^{'}(\eta_{ib}) \psi^{'}(\phi_i) & , a \neq b. \end{cases}$$

Melo et al. (2022) mostram que, quando n é grande, $\hat{\theta} \approx N_{dp}(\theta, \boldsymbol{K}^{-1})$.

3.1.3 Testes de Hipóteses

Teste de hipóteses é um procedimento estatístico que tem como objetivo definir a tomada de decisão baseada em uma estatística. No contexto da regressão Dirichlet, para saber se um parâmetro β_i é significante para o modelo de regressão, tem-se o interesse de testar as hipóteses $H_0 : \beta_i = 0$ contra $H_1 : \beta_i \neq 0$. De forma prática, quer dizer que se a hipótese nula não for rejeitada, deve-se retirar aquele parâmetro do modelo, pois o mesmo não influencia significativamente no modelo de regressão.

A estatística do teste da razão de verossimilhança é dada por

$$\omega = 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^0)],$$

em que $\hat{\boldsymbol{\theta}}$ é o modelo irrestrito do estimador de máxima verossimilhança e $\boldsymbol{\theta}^0$ o modelo restrito do estimador de máxima verossimilhança, ou seja, os parâmetros assumem o valor zero. Com a estatística ω calculada, o próximo passo é comparar o valor calculado com um valor crítico de uma distribuição χ^2 com p - 1 graus de liberdade para um determinado nível de significância pré-estabelecido.

Se a estatística for maior que o valor crítico, então rejeita-se a hipótese nula. Ou seja, o parâmetro testado é significativo para o modelo. Caso contrário, não se rejeita a hipótese nula, o que representa que β_i não é significante e deve ser removido do modelo.

4 SIMULAÇÃO

4.1 MÉTODO DE MONTE CARLO

O método de Monte Carlo envolve a geração de objetos ou processos aleatórios por meio de computação. Esses objetos podem ser originados de maneira natural ao modelar sistemas do mundo real, como uma rede rodoviária complexa, a trajetória de nêutrons em transporte ou a dinâmica do mercado de ações. No entanto, em muitos casos, os elementos aleatórios em técnicas de Monte Carlo são introduzidos de forma artificial para resolver problemas puramente determinísticos. Nesse contexto, a simulação de Monte Carlo consiste principalmente na amostragem aleatória de certas distribuições de probabilidade. Independentemente de serem naturais ou artificiais, a essência das técnicas de Monte Carlo é repetir o experimento muitas vezes (ou executar uma simulação por um período prolongado) para obter várias quantidades de interesse. Isso é feito utilizando princípios como a Lei dos Grandes Números e outros métodos de inferência estatística (KROESE et al., 2014)

A amostragem, a estimação e a otimização são algumas das principais utilizações do método de Monte Carlo. Na amostragem o objetivo é coletar informações sobre um objeto aleatório observando muitas realizações desse objeto. Um exemplo é a modelagem de simulação, onde um processo aleatório imita o comportamento de algum sistema da vida real, como uma linha de produção ou rede de telecomunicações. Outro exemplo ocorre na estatística bayesiana, onde a técnica de Cadeia de Markov Monte Carlo é frequentemente usada para amostrar de uma distribuição posteriori. Na estimação o foco está na estimativa de quantidades numéricas relacionadas a um modelo de simulação. Um exemplo no contexto natural das técnicas de Monte Carlo é a estimativa do processamento esperado em uma linha de produção. Um exemplo no contexto artificial envolve a avaliação de integrais multidimensionais por meio de técnicas de Monte Carlo, tratando a integral como a expectativa de uma variável aleatória. No contexto de otimização, o método de Monte Carlo é uma ferramenta poderosa para otimizar de funções objetivas complexas. Em muitas aplicações, essas funções são determinísticas, e a aleatoriedade é introduzida artificialmente para buscar eficientemente o domínio da função objetiva. As técnicas de Monte Carlo também são usadas para otimizar funções ruidosas, onde a função em si é aleatória, como no caso do resultado de uma simulação de Monte Carlo.

4.2 ESTUDO DE SIMULAÇÃO

Deseja-se conhecer a eficiência da estimação de parâmetros de regressão Dirichlet através da simulação de Monte Carlo. Para isso, foram geradas matrizes de dados a partir

de números aleatórios e dos parâmetros verdadeiros definidos e "retiradas" amostras de tamanho 10, 20, 40, 80 e 160, realizada a regressão Dirichlet a partir das matrizes de dados e das variáveis respostas geradas aleatoriamente a partir de uma distribuição normal multivariada. Para cada tamanho de amostra o processo foi repetido 5.000 vezes. Para realizar esta simulação foi utilizado o software R (R Core Team, 2023), que é uma linguagem de programação e ambiente de computação estatística de código aberto amplamente utilizado em análise de dados, estatísticas e visualização. Além das funções básicas oferecidas pelo R, será utilizado o pacote DirichletReg, desenvolvido por Marco J. Maier, em 2013. Este pacote oferece diversas facilidades para realizar a regressão Dirichlet, como por exemplo operação de fecho, cálculo de coeficientes de regressão e testes de significância.

Para avaliar o quão próximo as estimativas ficaram dos parâmetros fornecidos, serão utilizadas duas métricas: o Viés Absoluto e o Erro Quadrático Médio. O Viés Absoluto é distância entre a média do conjunto de estimativas e o parâmetro a ser estimado, de modo que quanto menor o viés absoluto, maior a acurácia das estimativas. Seja θ um parâmetro qualquer, o viés absoluto é:

$$\text{Viés Absoluto} = |\mathbb{E}(\hat{\theta}) - \theta|$$

Já o Erro Quadrático Médio (EQM) é usado para indicar o quão distante, em média, o conjunto de estimativas está do parâmetro a ser estimado. Quanto menores os valores dos erros quadráticos médios, menor a dispersão das estimativas. Para um parâmetro θ qualquer, o EQM é:

$$EQM = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

4.2.1 Simulação 2x2

No gráfico abaixo tem-se o viés de cada parâmetro estimado através da simulação realizada. Temos como parâmetros verdadeiros $\boldsymbol{\beta} = \begin{bmatrix} \beta_{11} = 1,5 & \beta_{12} = 1,2 \\ \beta_{21} = 1,3 & \beta_{22} = 1,0 \end{bmatrix}$

Tabela 1 – Viés absoluto da simulação com duas covariáveis e dois preditores

Parâmetro	n = 10	n = 20	n = 40	n = 80	n = 160
β_{11}	0.735	0.267	0.278	0.207	0.181
β_{12}	0.477	0.434	0.234	0.230	0.186
β_{21}	0.801	0.295	0.307	0.192	0.157
β_{22}	0.518	0.443	0.271	0.194	0.152

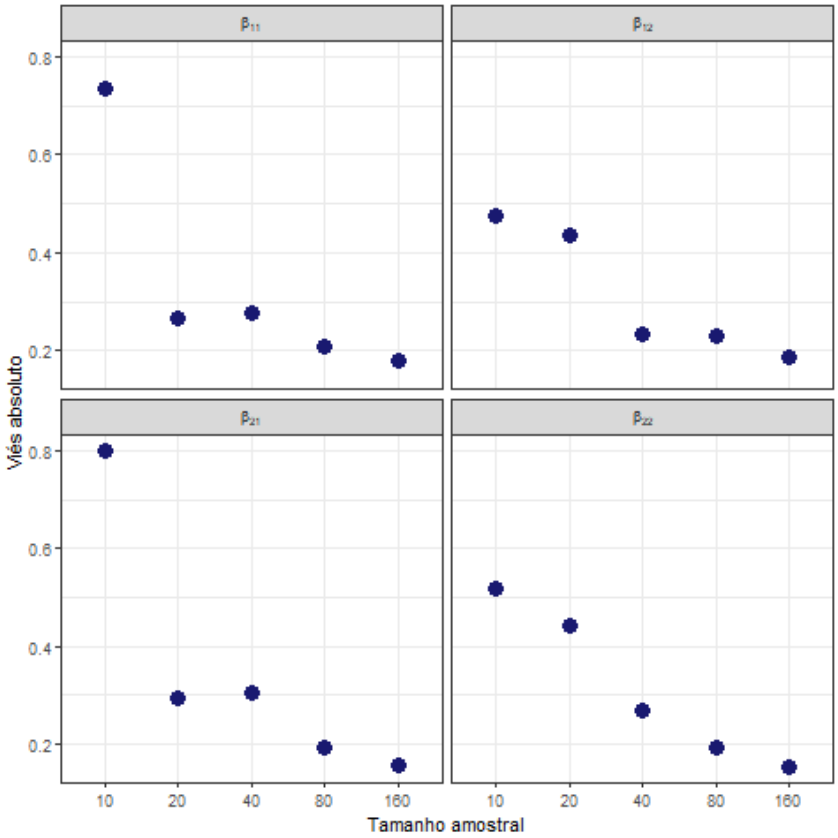

Figura 6 – Viés absoluto dos coeficientes simulação

Fonte: o Autor (2023)

Tabela 2 – EQM da simulação com duas covariáveis e dois preditores

Parâmetro	n = 10	n = 20	n = 40	n = 80	n = 160
β_{11}	1.003	0.112	0.121	0.063	0.047
β_{12}	0.389	0.306	0.087	0.078	0.048
β_{21}	1.167	0.147	0.146	0.058	0.037
β_{22}	0.467	0.315	0.112	0.060	0.034

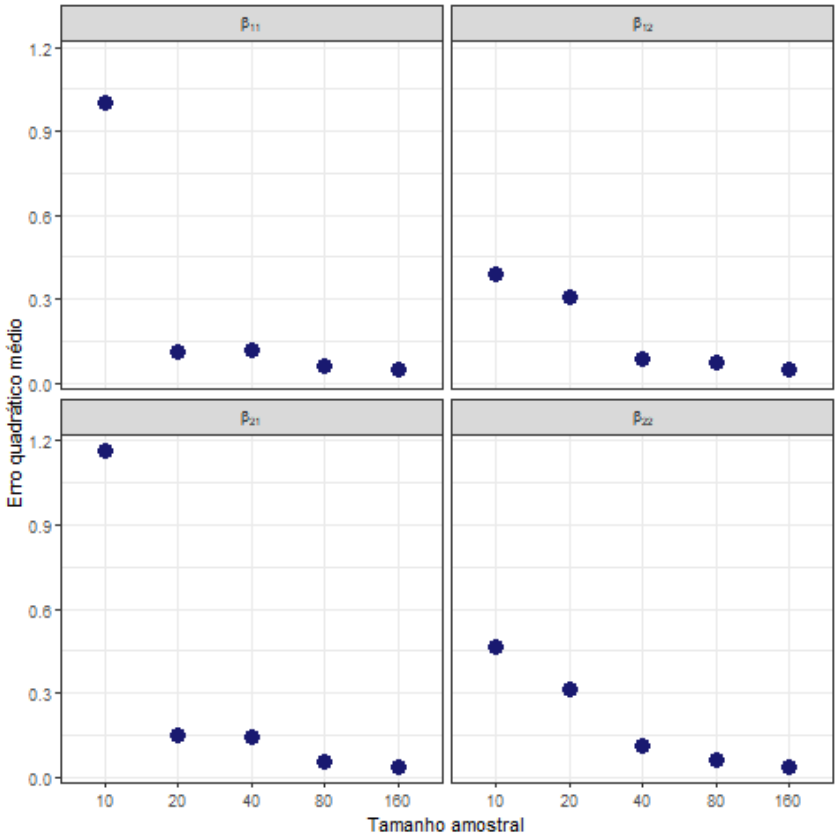

Figura 7 – EQM dos coeficientes simulação

Fonte: o Autor (2023)

A seguir, observa-se o comportamento encontrado quando realizamos a simulação com três preditores e duas variáveis explicativas $\boldsymbol{\beta} = \begin{bmatrix} \beta_{11} = 1,5 & \beta_{12} = 1,2 \\ \beta_{21} = 1,3 & \beta_{22} = 1,0 \\ \beta_{31} = 1,4 & \beta_{32} = 1,1 \end{bmatrix}$

4.2.2 Simulação 3x2

Tabela 3 – Viés absoluto para três preditores e duas covariáveis

Parâmetro	n = 10	n = 20	n = 40	n = 80	n = 160
β_{11}	0.311	0.286	0.160	0.158	0.158
β_{12}	0.568	0.337	0.207	0.174	0.166
β_{21}	0.400	0.338	0.199	0.166	0.146
β_{22}	0.569	0.306	0.223	0.235	0.236
β_{31}	0.355	0.299	0.137	0.100	0.076
β_{32}	0.562	0.296	0.146	0.104	0.087

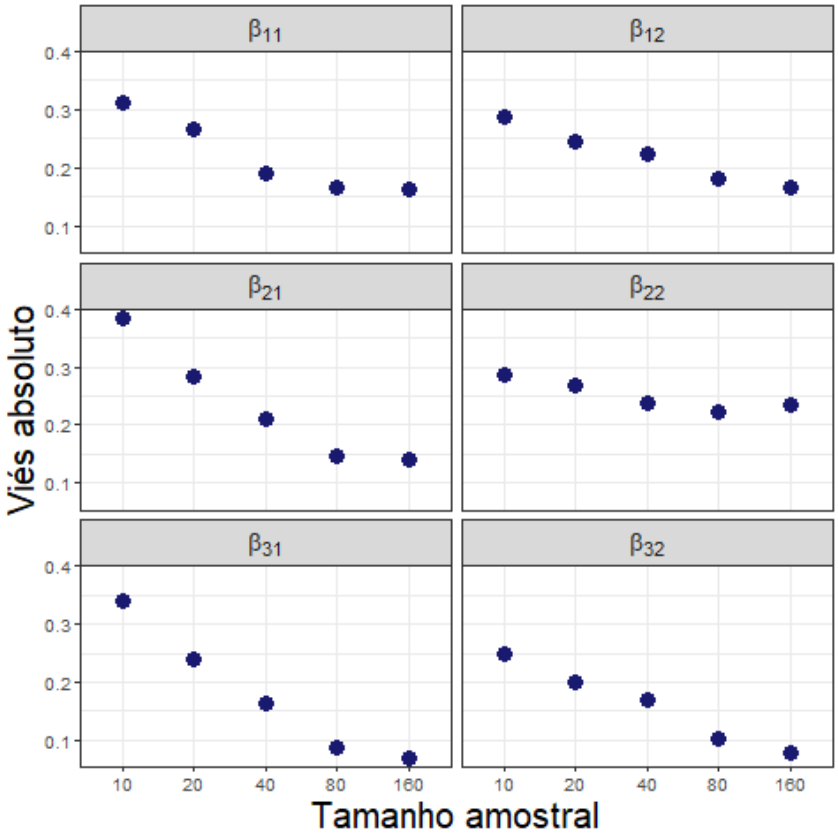

Figura 8 – Viés absoluto para três preditores e duas covariáveis

Fonte: o Autor (2023)

Tabela 4 – EQM para três preditores e duas covariáveis

Parâmetro	n = 10	n = 20	n = 40	n = 80	n = 160
β_{11}	0.153	0.130	0.038	0.035	0.031
β_{12}	0.542	0.179	0.061	0.040	0.034
β_{21}	0.249	0.185	0.056	0.038	0.027
β_{22}	0.526	0.138	0.068	0.067	0.063
β_{31}	0.206	0.146	0.030	0.016	0.009
β_{32}	0.513	0.141	0.033	0.017	0.011

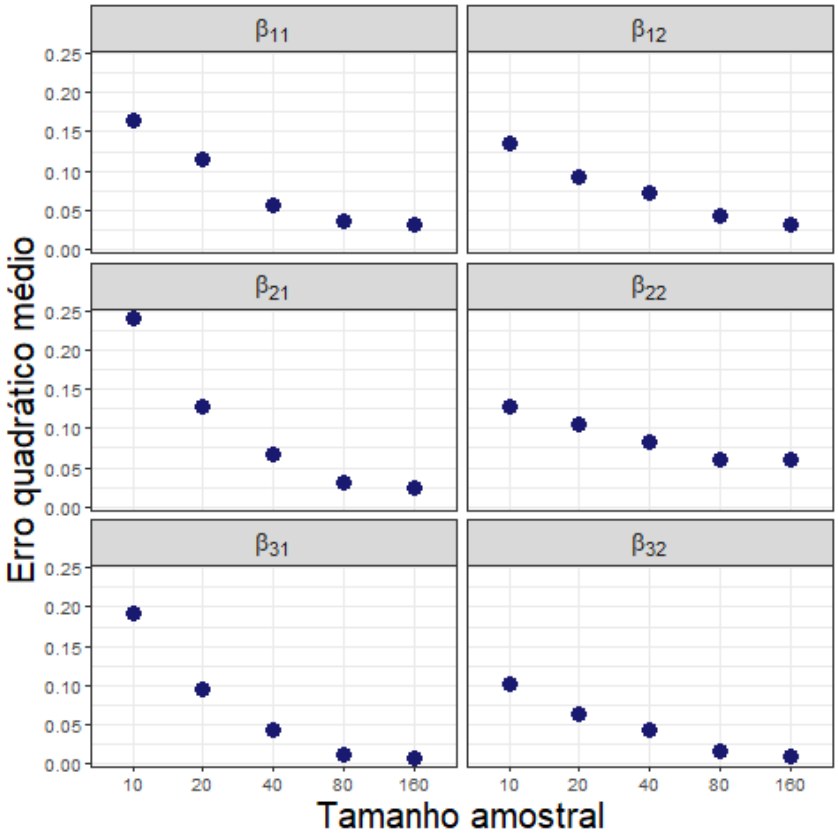

Figura 9 – EQM para três preditores e duas covariáveis

Fonte: o Autor (2023)

4.2.3 Simulação 3x3

A seguir, observa-se o comportamento encontrado quando realizamos a simulação com três preditores e duas variáveis explicativas $\boldsymbol{\beta} = \begin{bmatrix} \beta_{11} = 1,5 & \beta_{12} = 1,2 & \beta_{13} = 1,6 \\ \beta_{21} = 1,3 & \beta_{22} = 1,0 & \beta_{23} = 0,9 \\ \beta_{31} = 1,4 & \beta_{32} = 1,1 & \beta_{33} = 0,8 \end{bmatrix}$

Tabela 5 – Viés absoluto para três preditores e três covariáveis

Parâmetro	n = 10	n = 20	n = 40	n = 80	n = 160
β_{11}	0.374	0.260	0.174	0.147	0.140
β_{12}	0.371	0.326	0.242	0.187	0.158
β_{13}	0.420	0.311	0.260	0.227	0.223
β_{21}	0.431	0.309	0.223	0.175	0.166
β_{22}	0.331	0.267	0.223	0.227	0.243
β_{23}	0.376	0.249	0.187	0.146	0.130
β_{31}	0.398	0.274	0.169	0.106	0.085
β_{32}	0.322	0.278	0.185	0.120	0.080
β_{33}	0.557	0.530	0.547	0.576	0.578

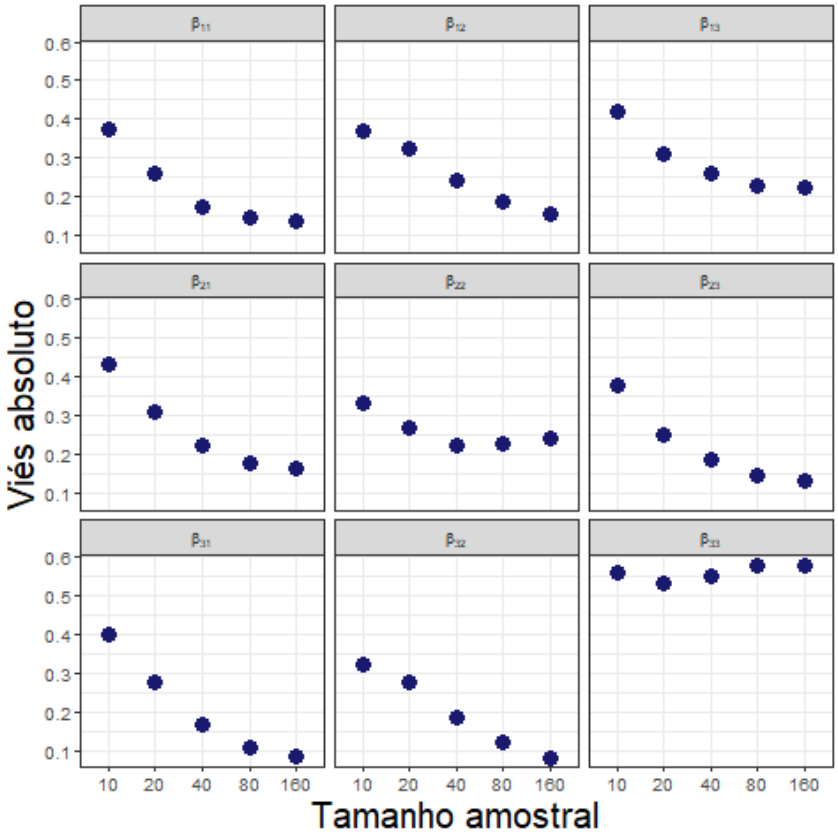

Figura 10 – Viés absoluto para três preditores e três covariáveis

Fonte: o Autor (2023)

Tabela 6 – EQM para três preditores e três covariáveis

Parâmetro	n = 10	n = 20	n = 40	n = 80	n = 160
β_{11}	0.228	0.106	0.048	0.030	0.026
β_{12}	0.212	0.159	0.086	0.047	0.032
β_{13}	0.273	0.141	0.095	0.065	0.058
β_{21}	0.297	0.152	0.071	0.041	0.034
β_{22}	0.165	0.112	0.074	0.066	0.066
β_{23}	0.225	0.098	0.056	0.031	0.023
β_{31}	0.253	0.121	0.044	0.018	0.011
β_{32}	0.168	0.120	0.053	0.022	0.010
β_{33}	0.431	0.339	0.331	0.347	0.342

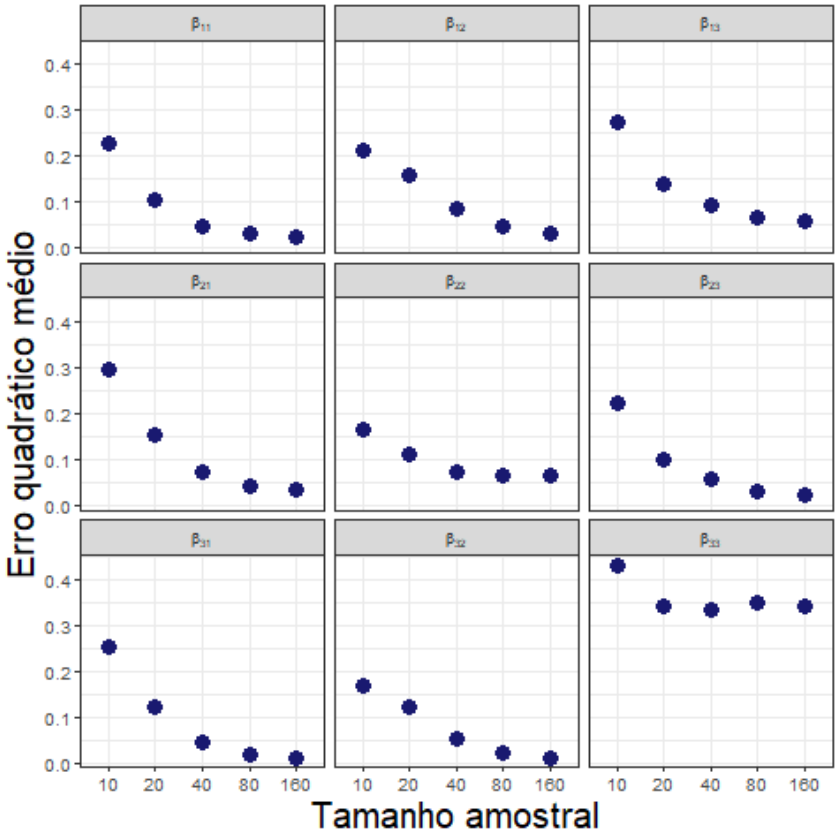

Figura 11 – EQM para três preditores e três covariáveis

Fonte: o Autor (2023)

Para as três configurações de simulação o comportamento foi o mesmo, conforme o tamanho da amostra aumentou tanto o viés absoluto quanto o erro quadrático médio diminuíram, com exceção de um coeficiente na configuração com três preditores e três covariáveis. Ou seja, de modo geral pode-se dizer que através da simulação com a regressão Dirichlet estimou-se bem os parâmetros verdadeiros definidos na geração da matriz de dados.

5 APLICAÇÕES

Apresentadas as teorias e simulações, serão utilizadas três bases de dados para fazer as aplicações. Para ilustração do problema, o primeiro conjunto de dados é de uma área clássica de dados composicionais, a Geologia, e apresenta a composição de sedimentos de um lago ártico. Por fim, os dois últimos conjunto apresentam as composições de vitórias, empates e derrotas de campeões brasileiros de futebol entre 2003 e 2022 e campeões das principais ligas do mundo no ano de 2022. Assim, como as simulações, as análises e gráficos foram feitos com o auxílio do software R (R Core Team, 2023).

5.1 ARCTIC LAKE

O conjunto de dados Arctic Lake trata da composição de areia, lodo e argila (*sand*, *silt* e *clay* , respectivamente) do Lago Stanwell-Fletcher, localizado na Ilha Somerset, no arquipélago Ártico do Canadá. Ao todo são 39 composições desses três materiais em diferentes profundidades de água. Os dados podem ser encontrados em Aitchison (1986). Os valores de cada material está dado em porcentagem, a soma dos três é igual a 100, exceto por erro de arredondamento. Já a profundidade é dada em metros. Assim, seguem as primeiras observações do conjunto de dados

Tabela 7 – Primeiras observações do conjunto de dados Arctic Lake

Fonte: Aitchison (1986)

Sand	Silt	Clay	Depth
77,5	19,5	3,0	10,4
71,9	24,9	3,2	11,7
50,7	36,1	13,2	12,8
52,2	40,9	6,6	13,0
70,0	26,5	3,5	15,7
66,5	32,3	1,3	16,3

Com a Tabela 8 é possível comparar as médias aritmética e composicional, adequada a esse tipo de dados. Nota-se uma predominância maior de lodo (*silt*) na média composicional do que na aritmética. Por consequência, os valores dos demais elemtentos é menor.

Tabela 8 – Médias aritmética e composicional dos materiais do Artic Lake

Fonte: O Autor (2024)

Média	Sand	Silt	Clay
Aritmética	0,242	0,457	0,301
Composicional	0,178	0,564	0,258

Na Tabela 9 é apresentada a matriz de variação do conjunto de dados. Por ela é possível notar que há uma maior correspondência entre areia (*sand*) e argila (*clay*). A partir da Tabela 9 é possível também calcular a variância total, a qual tem valor de 2,469.

Tabela 9 – Matriz de variação dos materiais do Artic Lake

Fonte: O Autor (2024)

	Sand	Silt	Clay
Sand	0,000	1,641	4,731
Silt	1,641	0,000	1,035
Clay	4,731	1,035	0,000

Para observar a distribuição das porcentagens dos materiais, a Figura 12 apresenta os boxplots das mesmas. O lodo (*silt*) apresenta maior mediana entre os materiais e quatro outliers.

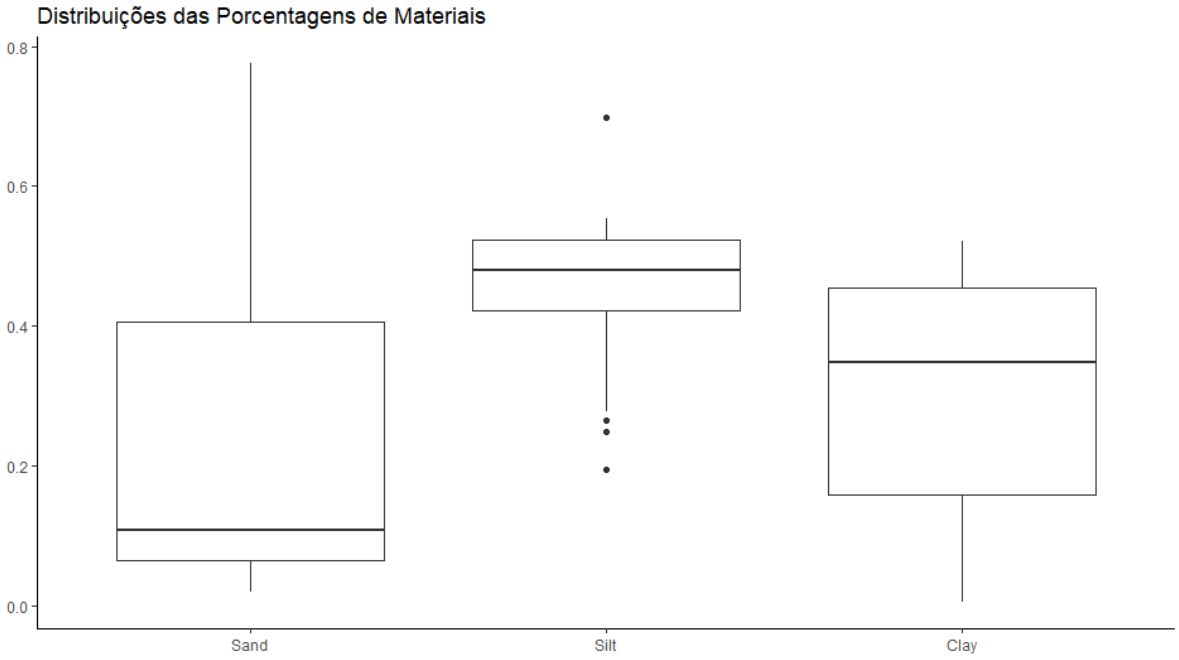

Figura 12 – Boxplots das porcentagens dos materiais

Fonte: O Autor(2023)

As Figuras 13, 14 e 15 mostram a dispersão dos materiais pela profundidade do lago. O percentual de areia parece diminuir conforme aumenta a profundidade. Por outro lado, a parte de argila parece apresentar uma correlação positiva com a profundidade. No entanto, não é possível perceber uma relação clara entre lodo e profundidade.

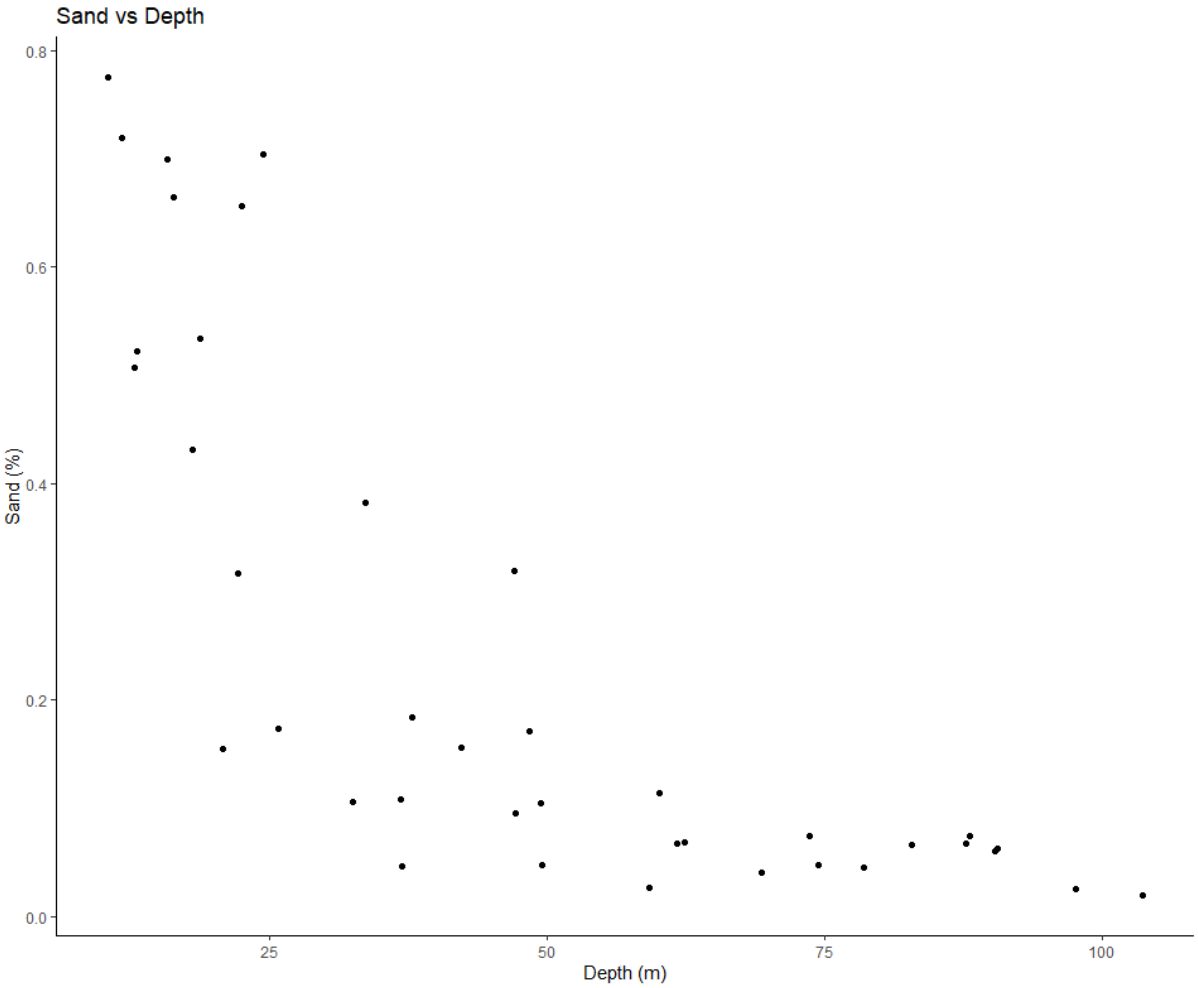

Figura 13 – Dispersão de areia por profundidade

Fonte: O Autor(2023)

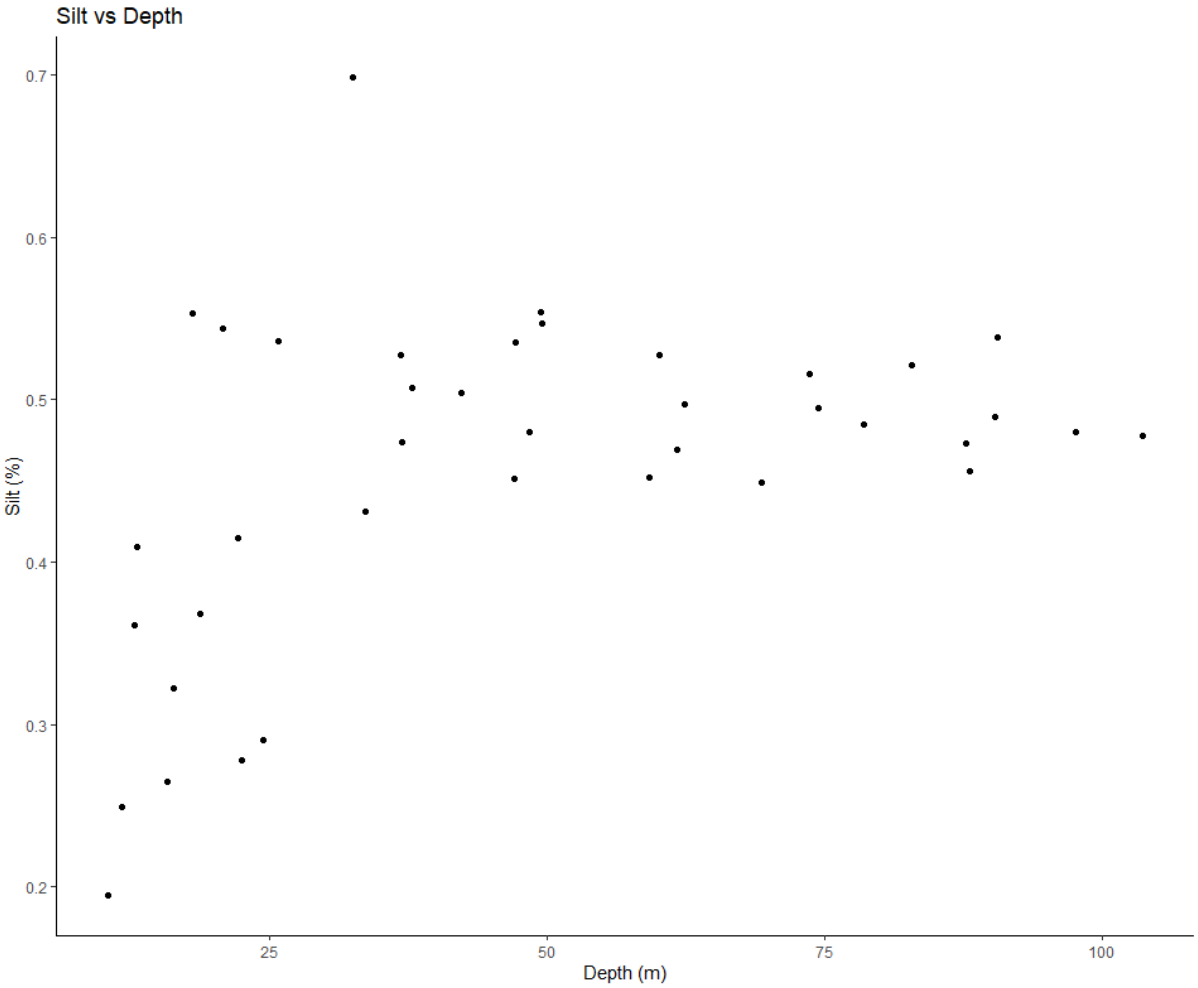

Figura 14 – Dispersão de lodo por profundidade

Fonte: O Autor(2023)

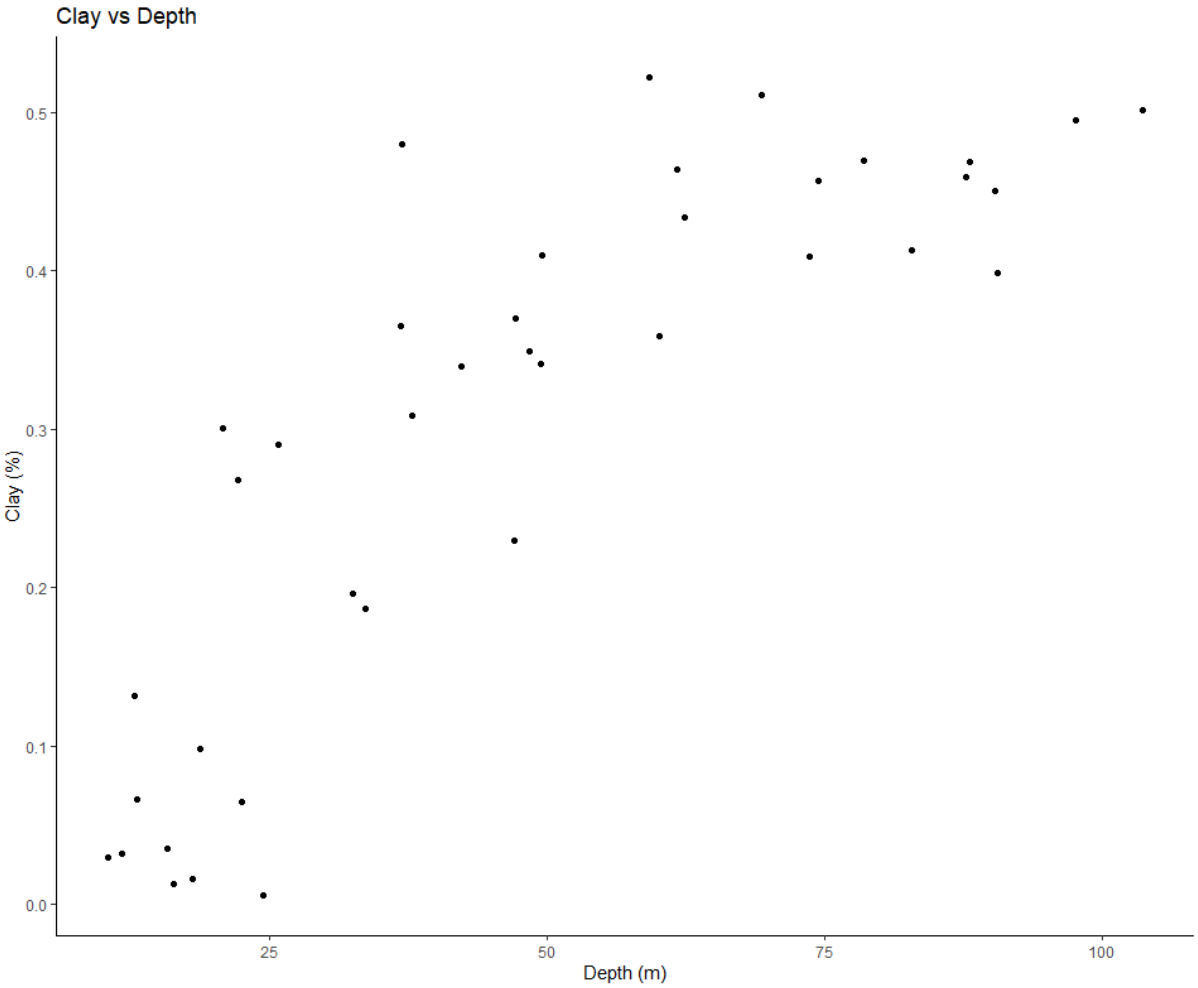

Figura 15 – Dispersão de argila por profundidade

Fonte: O Autor(2023)

Para uma melhor exploração da base de dados serão utilizados os gráficos de barras horizontais estacados e o diagrama ternário, vistos na Seção 2.3. Na Figura 17, foi destacado um ponto em vermelho para exemplificação da leitura e interpretação do diagrama ternário. O ponto em questão é o da profundidade 32,5 metros e tem 10,6% de areia, 69.8% de lodo e 19,6% de argila em sua composição.

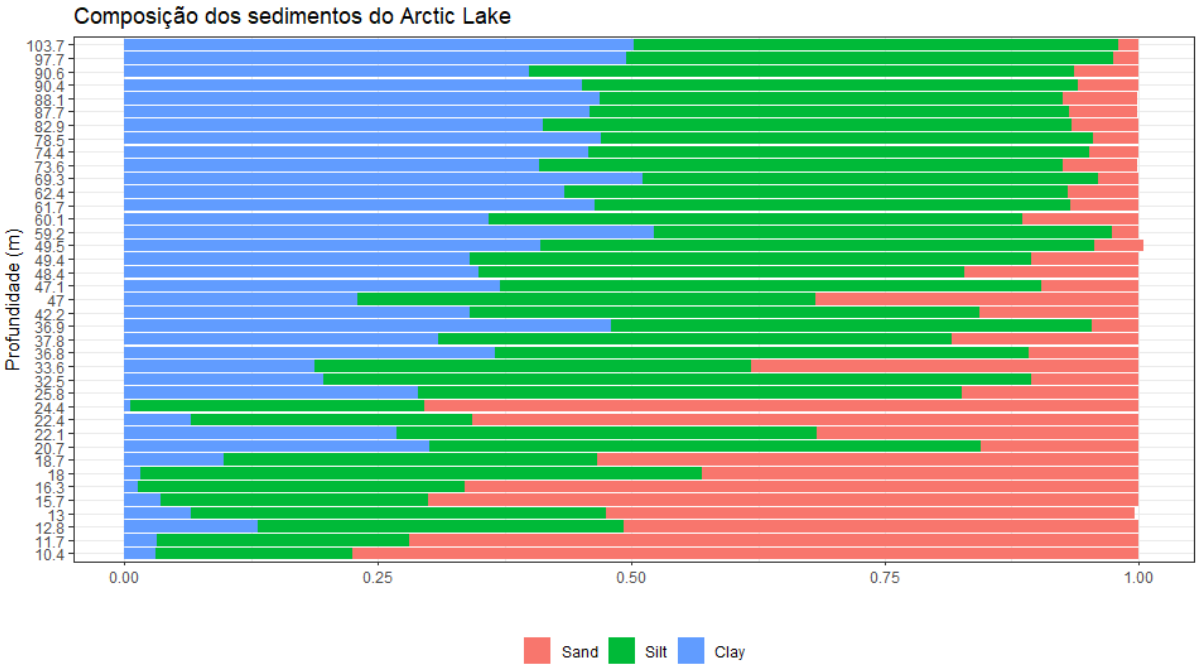

Figura 16 – Gráfico de Barras Horizontais Estacadas de Arctic Lake

Fonte: O Autor(2023)

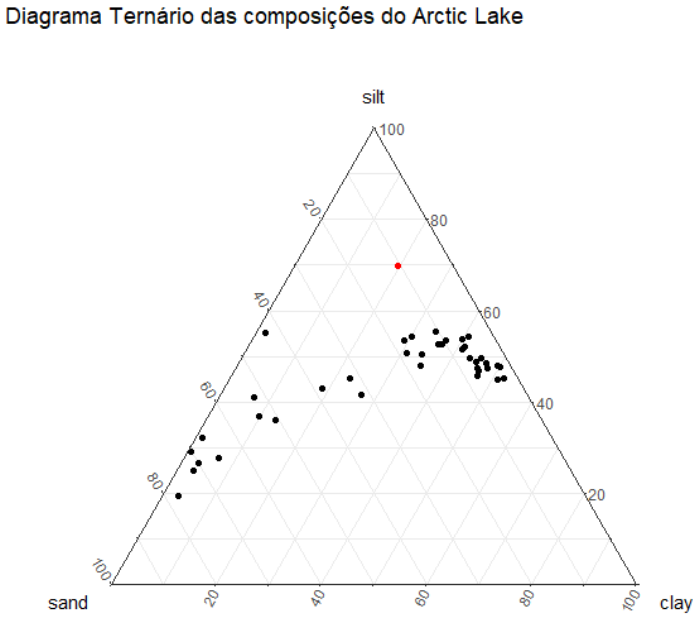

Figura 17 – Diagrama Ternário de Arctic Lake

Fonte: O Autor(2023)

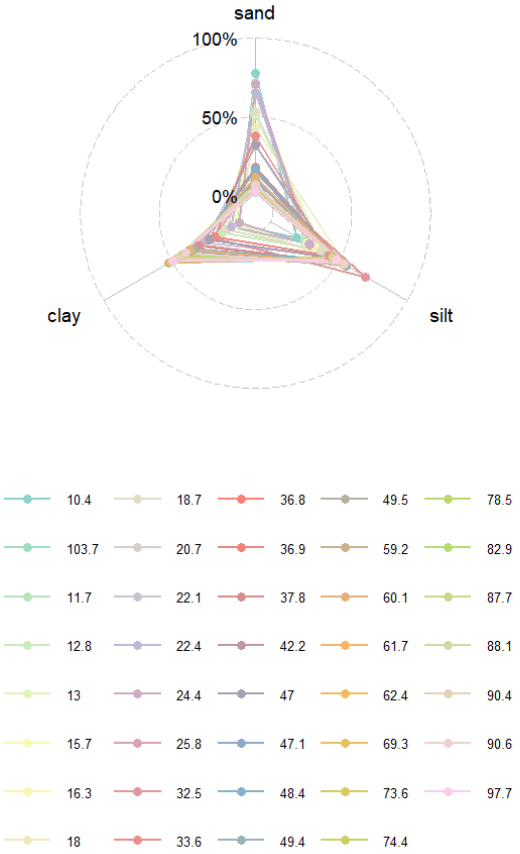

Figura 18 – Gráfico de Radar do Arctic Lake

Fonte: O Autor(2024)

Nota-se pela Figura 16 que conforme a profundidade aumenta, a porcentagem de areia do sedimento diminui. Logicamente, o lodo e a argila aumentam em maiores profundidades.

5.1.1 Regressão Dirichlet para Arctic Lake

Deseja-se predizer a composição de areia, lodo e argila pela profundidade do lago. Deste modo, o modelo de regressão é descrito da seguinte maneira

$$\boldsymbol{\eta} = \boldsymbol{\beta}_0^T + \boldsymbol{\beta}_1^T Profundidade,$$

em que $\boldsymbol{\eta} = (\eta_1 = \text{Areia}, \eta_2 = \text{Lodo}, \eta_3 = \text{Argila})^T$ é o vetor de preditores, $\boldsymbol{\beta}_0^T = (\beta_{10}, \beta_{20}, \beta_{30})$ é o vetor de coeficientes dos interceptos de cada variável resposta e $\boldsymbol{\beta}_1^T = (\beta_{11}, \beta_{21}, \beta_{31})$ é o vetor dos coeficientes de regressão relacionados à variável explicativa.

Tabela 10 – Coeficientes de regressão calculados do conjunto de dados Arctic Lake

Parâmetro	Estimativa	Erro Padrão	z-valor	Pr(>|z|)
β_{10}	$0,117$	$0,409$	$0,035$	$0,776$
β_{11}	$0,023$	$0,008$	$3,133$	$0,002*$
β_{20}	$-0,311$	$0,345$	$-0,901$	$0,368$
β_{21}	$0,056$	$0,006$	$8,730$	$< 2 \times 10^{-16}*$
β_{30}	$-1,152$	0.299	$-3,860$	$0,0001*$
β_{31}	$0,064$	$0,006$	$11,210$	$< 2 \times 10^{-16}*$

Se for considerado um nível de significância de 5%, nota-se que a variável explicativa profundidade é significativa para as três variáveis resposta, enquanto entre os interceptos, apenas o da variável resposta argila é significativo. Assim, o modelo pode ser reescrito como

$$\eta_1 = 0,023 Profundidade$$
$$\eta_2 = 0,056 Profundidade$$
$$\eta_3 = -1,152 + 0,064 Profundidade.$$

Utilizando as profundidades do conjunto de dados e os coeficientes de regressão, é possível predizer os valores das variáveis respostas. Na Figura 19 estão plotados em preto os valores reais da composição observada nos sedimentos do lago e em azul os valores preditos pelo modelo de regressão. É possível notar que o modelo não prevê um percentual alto de areia mesmo nas profundidades mais baixas, ficando distante das observações. No entanto, nas maiores profundidades, onde existe maior concentração de lodo e argila, é possível notar que os valores preditos estão próximos às observações. Já a Figura 20 é o gráfico de barras estacadas dos valores preditos. É possível notar que quando comparado aos dados observados, os valores preditos apresentam mais argila nas profundidades menores e mais areia nas maiores.

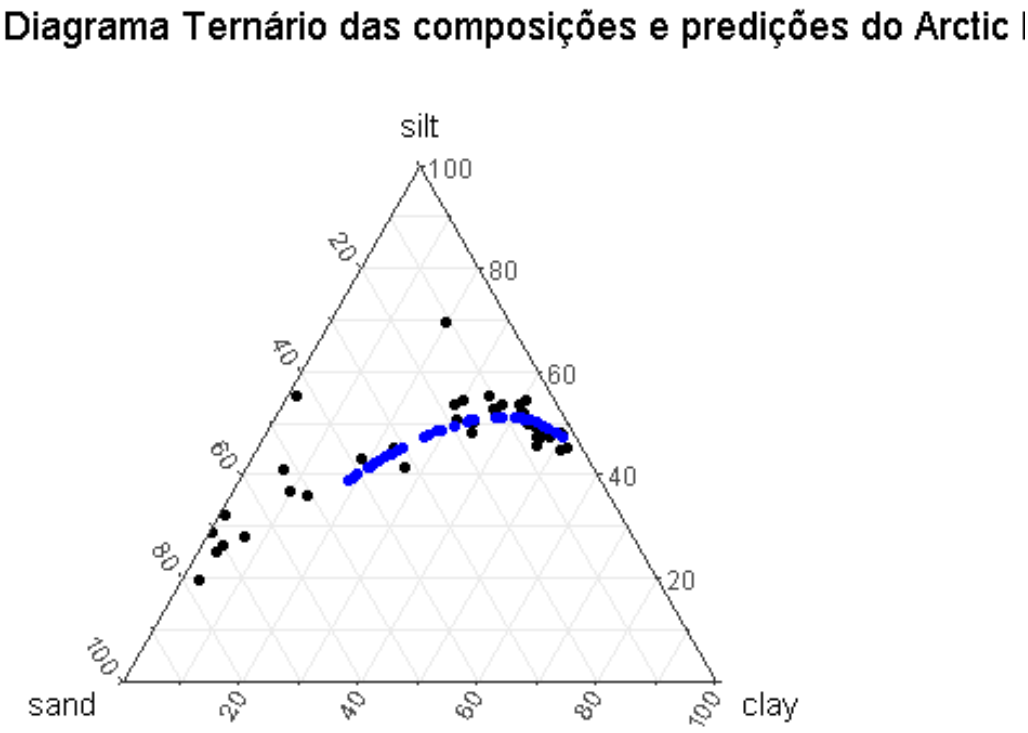

Figura 19 – Diagrama Ternário de Arctic Lake

Fonte: O Autor(2023)

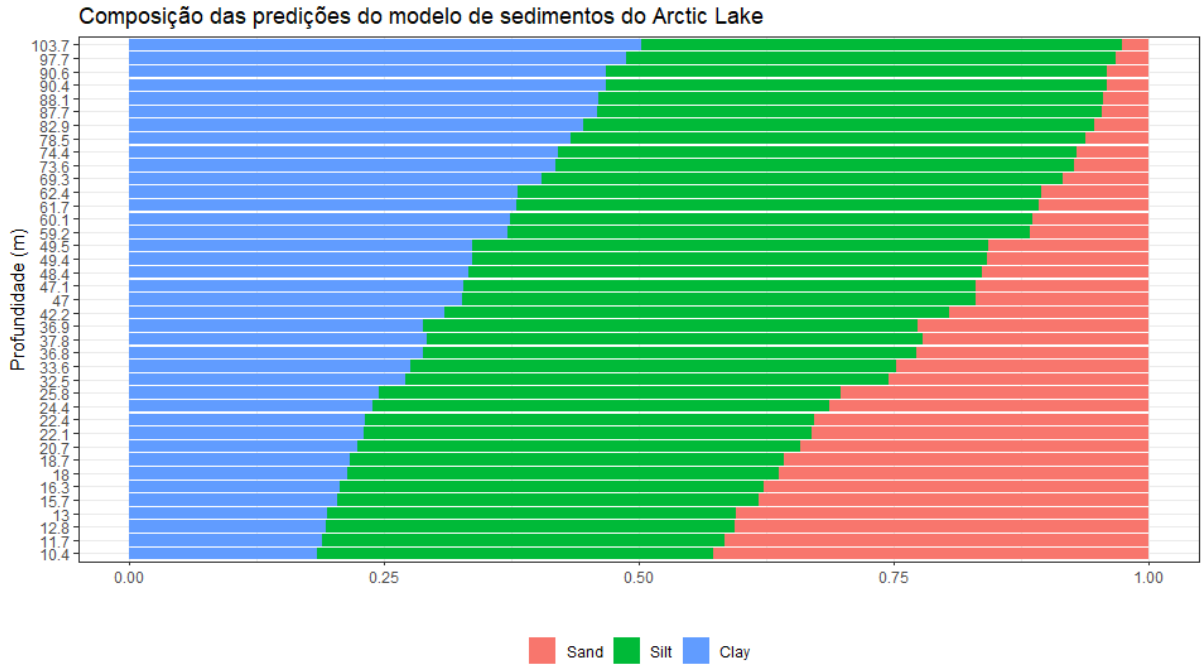

Figura 20 – Gráfico de Barras de Predições de Arctic Lake

Fonte: O Autor(2024)

5.2 CAMPEÕES BRASILEIROS

O conjunto de dados Campeões Brasileiros refere-se a algumas características dos vencedores da primeira divisão do campeonato brasileiro de futebol masculino, atualmente conhecido como Brasileirão Betano 2024, entre os anos de 2003 e 2022. Coletada pelo professor Tiago Magalhães, esta base de dados tem 19 observações com 28 variáveis cada sobre o campeão brasileiro de cada um dos anos. As variáveis são: o ano da competição, time campeão, cidade do time campeão, estado do time campeão, se o time foi ou não campeão estadual, número de clubes brasileiros no final da CONMEBOL Libertadores, um time brasileiro foi ou não campeão da CONMEBOL Libertadores, número de clubes brasileiros nas finais da CONMEBOL Libertadores e CONMEBOL Sudamericana, se o clube possui estádio próprio, Se é ano da FIFA World Cup, quantos clubes jogam no mesmo estádio, número de rivais, se era ou não o atual campeão, número de participantes no campeonato,número de gols na competição, média de gols por partida na competição, pontos conquistados, jogos disputados, número de vitórias, número de empates, número de derrotas, número de gols marcados, número de gols sofridos, saldo de gols, aproveitamento, salário mínimo, reajuste do salário mínimo e taxa do crescimento do PIB. O período de tempo escolhido deve-se à fórmula de disputa de pontos corridos, ou seja, todos os times participantes se enfrentam em partidas de ida e volta. Neste sistema uma vitória vale

três pontos, um empate vale um ponto e uma derrota vale zero pontos. É declarado campeão o clube que conquistou o maior número de pontos ao final do campeonato. Dentre as variáveis estão o número de vitórias, empates e derrotas, que podem ser enxergadas de forma composicional uma vez que há um número fixo de jogos, ou seja, uma soma constante. Algumas das colunas das seis primeiras observações do conjunto de dados serão apresentadas a seguir.

Tabela 11 – Primeiras observações de algumas colunas do conjunto de dados Campeões Brasileiros

Fonte: Magalhães (2023)

Ano	Time	Vitorias	Empates	Derrotas	Atual	GP	GC
2003	Cruzeiro	31	7	8	0	102	47
2004	Santos	27	8	11	0	103	58
2005	Corinthians	24	9	9	0	87	59
2006	São Paulo	22	12	4	0	66	32
2007	São Paulo	23	8	7	1	55	19
2008	São Paulo	21	12	5	1	66	36

Pode-se observar pela Figura 21 que há uma predominância do estado de São Paulo com onze títulos. Rio de Janeiro e Minas Gerais são os outros dois estados que conseguiram conquistar o Campeonato Brasileiro no período analisado com cinco e quatro títulos respectivamente.

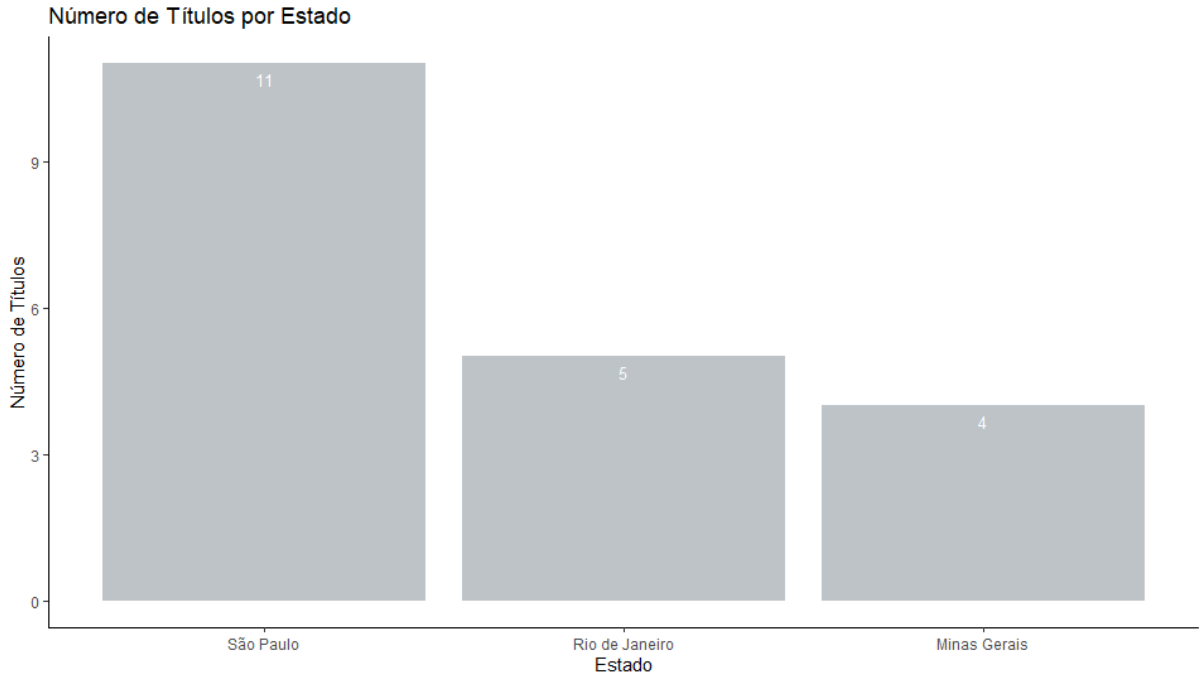

Figura 21 – Número de Títulos por Estado

Fonte: O Autor(2023)

Pela Figura 22, nota-se que em apenas quatro ocasiões o atual campeão conseguiu conquistar o Campeonato Brasileiro novamente. Destaque para o São Paulo, único tricampeão no período observado (2006, 2007, 2008). Conseguiram ser bicampeões o Cruzeiro (2013, 2014) e o Flamengo (2019, 2020). Assim, fica nítido o equilíbrio desse campeonato. Dentre os 20 anos observados, oito clubes diferentes conseguiram conquistar o título.

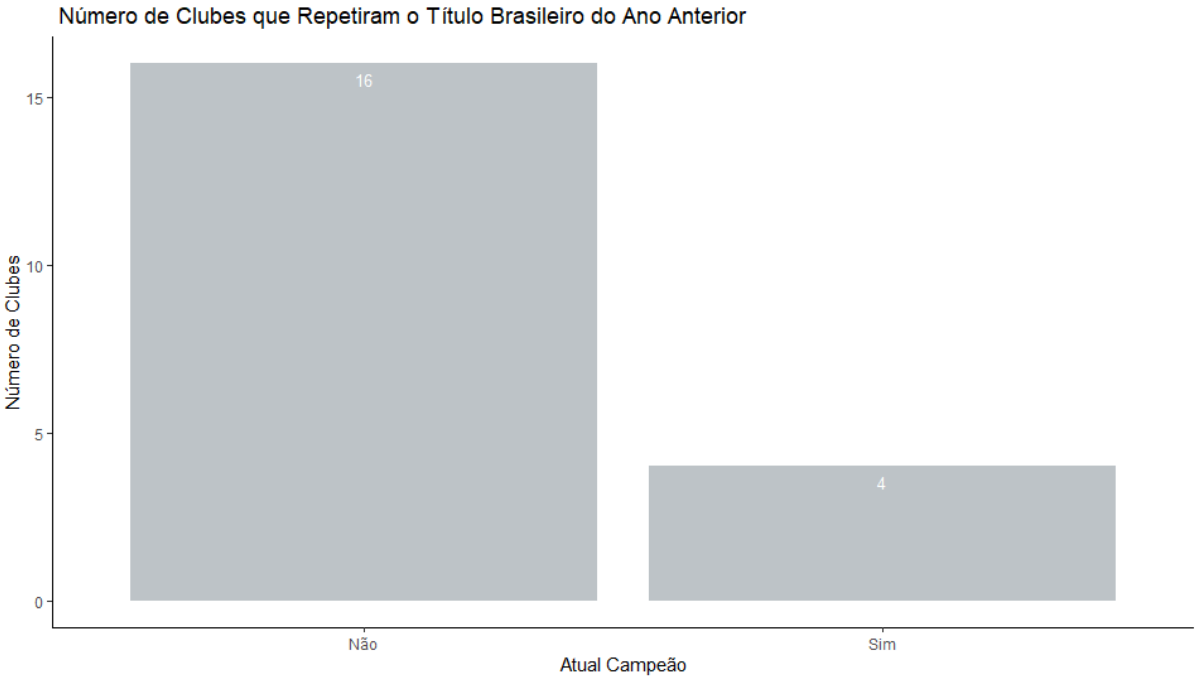

Figura 22 – Campeões Brasileiros que Repetiram o Título do Ano Anterior

Fonte: O Autor(2023)

A Tabela 12 apresenta a média composicional dos dados de resultados. Nota-se uma predominância de Vitórias na composição, o que é esperado para campeões.

Tabela 12 – Média composicional dos Campeões Brasileiros de 2003 a 2022

Fonte: O Autor (2024)

Vitórias	Empates	Derrotas
0,606	0,230	0,164

Na Tabela 13 é apresentada a matriz de variação do conjunto de dados. A partir da Tabela 13 é possível também calcular a variância total, a qual tem valor de 0,156.

Tabela 13 – Matriz de variação dos Campeões Brasileiros de 2003 a 2022

Fonte: O Autor (2024)

	Vitórias	Empates	Derrotas
Vitórias	0,000	0,094	0,143
Empates	0,094	0,000	0,230
Derrotas	0,143	0,230	0,000

Para uma melhor exploração da base de dados serão utilizados os gráficos de barras horizontais estacados ordenado pelo saldo de gols, o diagrama ternário e o gráfico de radar, vistos na Seção 2.3. O ponto destacado em vermelho na Figura 24 refere-se ao Flamengo de 2009, o campeão com menor percentual de vitórias. Este time teve 50% de vitórias, 26,3% de empates e 23,7% de derrotas no Campeonato Brasileiro daquele ano.

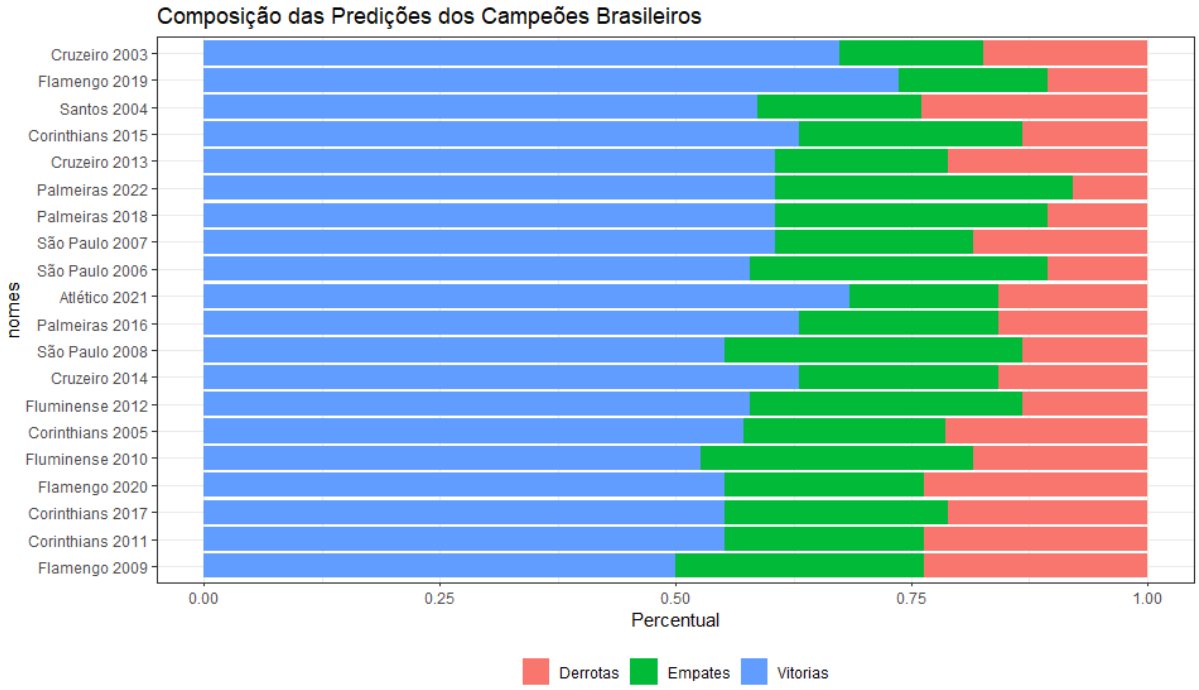

Figura 23 – Gráfico de Barras Horizontais Estacadas de Campeões Brasileiros

Fonte: O Autor(2023)

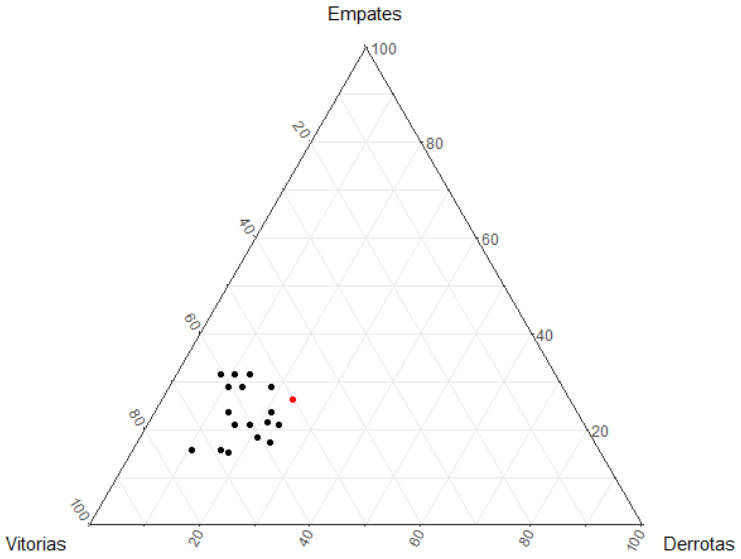

Figura 24 – Diagrama Ternário de Campeões Brasileiros

Fonte: O Autor(2023)

Figura 25 – Gráfico de Radar dos Campeões Brasileiros

Fonte: O Autor(2024)

Como esperado de um time campeão, note-se uma maior concentração das composições em vitórias.

5.2.1 Regressão Dirichlet para Campeões Brasileiros

Para modelar a relação de vitórias, empates e derrotas e outros aspectos da campanha de um campeão brasileiro, adota-se o seguinte modelo

$$\boldsymbol{\eta} = \boldsymbol{\beta}_0^T + \boldsymbol{\beta}_1^T Saldo + \boldsymbol{\beta}_2^T Atual$$

em que $\boldsymbol{\eta} = (\eta_1 = \text{Vitórias}, \eta_2 = \text{Empates}, \eta_3 = \text{Derrotas})^T$ é o vetor de preditores, $\boldsymbol{\beta}_0^T = (\beta_{10}, \beta_{20}, \beta_{30})$ é o vetor de coeficientes dos interceptos de cada variável resposta, $\boldsymbol{\beta}_1^T = (\beta_{11}, \beta_{21}, \beta_{31})$ é o vetor dos coeficientes de regressão relacionados ao saldo de gols e $\boldsymbol{\beta}_2^T = (\beta_{12}, \beta_{22}, \beta_{32})$ é o vetor dos coeficientes de regressão relacionados a informação se o clube foi o campeão da edição anterior. Utilizando a parametrização comum, foram calculados os coeficientes de regressão mostrados na Tabela 14.

Tabela 14 – Coeficientes de regressão calculados do conjunto de dados Campeões Brasileiros

Parâmetro	Estimativa	Erro Padrão	z-valor	Pr(>|z|)
β_{10}	$6,762$	$0,947$	$7,139$	$9,4 \times 10^{-13}*$
β_{11}	$-0,081$	$0,027$	$-3,000$	$0,003*$
β_{12}	$0,0028$	$0,576$	$0,049$	$0,961$
β_{20}	$6,128$	$0,937$	$6,537$	$6,3 \times 10^{-11}$
β_{21}	$-0,090$	$0,026$	$-3,400$	$0,001*$
β_{22}	$-0,032$	$0,578$	$-0,056$	$0,955$
β_{30}	$6,434$	$0,984$	$6,541$	$6,1 \times 10^{-11}$
β_{31}	$-0,112$	0.029	$-3,883$	$0,0001*$
β_{32}	$0,054$	$0,575$	$0,093$	$0,925$

Deste modo, considerando um nível de significância de 5%, os interceptos e o saldo de gols foram significativos para as três variáveis resposta, enquanto a variável atual campeão não foi significativa para nem para vitórias, nem empates e nem derrotas. Então, o modelo reajustado sem a variável atual campeão pode ser reescrito como

$$\eta_1 = 6,607 - 0,077 Saldo$$
$$\eta_2 = 5,968 - 0,086 Saldo$$
$$\eta_3 = 6,273 - 0,107 Saldo$$

Utilizando as variáveis explicativas do conjunto de dados e os coeficientes de regressão, é possível predizer os valores das variáveis respostas. Na Figura 26 estão plotados em preto os valores reais das composições observadas nos campeonatos brasileiros e em azul os valores preditos pelo modelo de regressão.

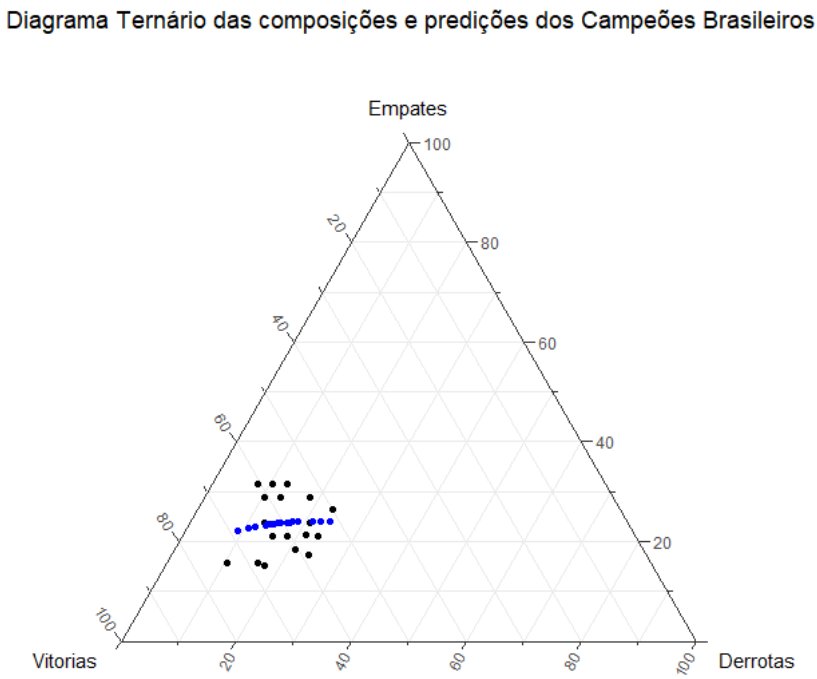

Figura 26 – Diagrama Ternário de Predições de Campeões Brasileiros

Fonte: O Autor(2023)

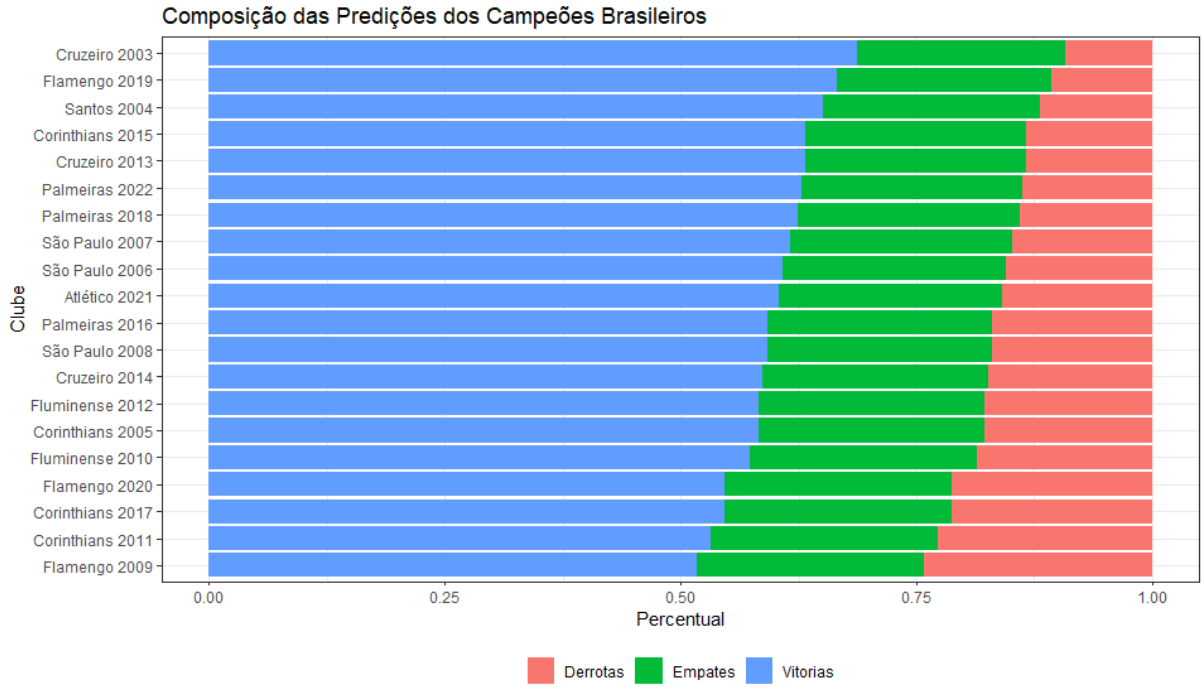

Figura 27 – Composições de Predições de Campeões Brasileiros

Fonte: O Autor(2024)

O saldo de gols é uma variável que é a diferença entre os gols marcados e sofridos por um time. Assim, a seguir iremos separar essas duas informações para ajustar um modelo para a média das variáveis resposta com a parametrização alternativa, com a variável empates como referência. O modelo pode ser escrito como

$$\begin{aligned}\boldsymbol{\eta} =&(\beta_{10}, \beta_{20}, \beta_{30})^T + (\beta_{11}, \beta_{21}, \beta_{31})^T GolsMarcados \\ &+ (\beta_{12}, \beta_{22}, \beta_{32})^T GolsSofridos\end{aligned}$$

O resultados estão apresentados na Tabela 15

Tabela 15 – Coeficientes de regressão calculados do conjunto de dados Campeões Brasileiros com o modelo alternativo de base Empates

Parâmetro	Estimativa	Erro Padrão	z-valor	Pr(>|z|)
β_{10}	$0,196$	$0,265$	$0,740$	$0,460$
β_{11}	$0,015$	$0,005$	$2,953$	$0,003*$
β_{12}	$-0,007$	$0,007$	$-0,999$	$0,318$
β_{30}	$-1,141$	$0,354$	$-3,223$	$0,001*$
β_{31}	$-0,005$	$0,007$	$-0,740$	$0,459*$
β_{32}	$0,031$	$0,010$	$3,283$	$0,001*$

Assim, sob um nível de significância de 5%, quando comparados aos coeficientes de empates, o número de gols marcados é significativo para as vitórias e o intercepto e o número de gols sofridos são significantes para as derrotas. No entanto, se for utilizado o mesmo modelo, porém a variável resposta base for vitórias, não só os coeficientes mudam, mas suas significâncias, como pode ser observado na Tabela 16.

Tabela 16 – Coeficientes de regressão calculados do conjunto de dados Campeões Brasileiros com o modelo alternativo de base Vitórias

Parâmetro	Estimativa	Erro Padrão	z-valor	Pr(>|z|)
β_{20}	$-0,196$	$0,265$	$-0,740$	$0,460$
β_{21}	$-0,015$	$0,005$	$-2,953$	$0,003*$
β_{22}	$0,007$	$0,007$	$0,999$	$0,318$
β_{30}	$-1,337$	$0,354$	$-3,223$	$1,1 \times 10^{-5}*$
β_{31}	$-0,020$	$0,007$	$-0,740$	$0,001*$
β_{32}	$0,039$	$0,010$	$3,283$	$1,8 \times 10^{-6}*$

Nota-se que as estimativas de vitórias com empates como referência são iguais em módulo às estimativas de empates com vitórias como base. Entretanto, não só as estimativas mudam para a variável resposta derrota, mas a variável explicativa gols marcados passa a ser significativa quando a vitória é a variável de referência.

5.3 CAMPEÕES NACIONAIS DE 2022

No conjunto de dados anterior, foi vista a composição de vitórias, empates e derrotas no Campeonato Brasieliro em diversos anos. No conjunto a seguir será analisada a mesma composição, mas agora de diversos campeonatos no mesmo ano.

O IFFHS (*International Federation of Football History & Statistics*) é uma organização que administra e divulga recordes e estatísticas de futebol. Dentre suas publicações, está um ranking com as melhores ligas nacionais de futebol masculino em 2022, IFFHS (2023). A partir desta publicação, foram colhidos pelo autor dados sobre os campeões nacionais das vinte melhores ligas segundo o instituto. São essas ligas, em ordem, o Brasileirão Assaí - Série A (Brasil), a Premier League (Inglaterra), a La Liga (Espanha), a Bundesliga (Alemanha), a Serie A TIM (Itália), a League 1 Uber Eats (França), a Liga Portugal Bwin (Portugal), a Eredivisie (Países Baixos), a Copa Binance (Argentina), a Copa de Primera TIGO-Visión Banco 2022 Clausura (Paraguai), a LigaPro Betcris 2022 (Equador), a Liga BetPlay DIMAYOR 2022 Clausura (Colômbia), a Egyptian Premier League (Egito), a Jupiler Pro League (Bélgica), a Spor Toto Süper Lig (Turquia), a Primera Divisón Profesional de Uruguay (Uruguai), a Mozzart Bet SuperLiga (Sérvia), a Hana 1Q K League 1 (Coréia do Sul), a Superliga (Romênia) e a Ligue 1 (Argélia). Deste modo, são observados os vinte campeões nacionais e foram colhidos dados sobre o desempenho

de cada um nos seus respectivos campeonatos. Assim, a base de dados é composta pelo nome do clube, país de origem, continente, número de vitórias, empates e derrotas, se é ou não o atual campeão, gols marcados, gols sofridos, cartões amarelos recebidos e cartões vermelhos recebidos. Como nem todos os campeonatos apresentaram o mesmo número de jogos, tanto os gols feitos e sofridos e quanto o número de cartões vermelhos e amarelos recebidos foram divididos pelo número de jogos realizados no campeonato para as análises realizadas. Também é válido ressaltar que nem todos os campeonatos têm como fórmula de disputa os pontos corridos. Nos campeonatos em que o campeão não é definido por pontos corridos os dados coletados referem-se apenas a fase regular do campeonato, ou seja, sem contar os jogos eliminatórios.

Tabela 17 – Primeiras observações de algumas colunas do conjunto de dados Campeões Nacionais de 2022

Fonte: O Autor (2023)

Clube	Vitórias	Empates	Derrotas	Atual	GP	GC	CA	CV
Palmeiras	23	12	3	1	66	27	72	3
Manchester City	29	6	3	1	99	26	42	1
Real Madrid	26	8	4	0	80	31	76	0
Bayern	24	5	5	1	97	37	36	2
Milan	26	8	4	0	69	31	71	1
PSG	26	8	4	0	90	36	78	4

Dentre as vinte melhores ligas, pode-se observar pela Figura 28 que há uma predominância de campeonatos europeus, seguidos pelo sul-americanos com 6 ligas.

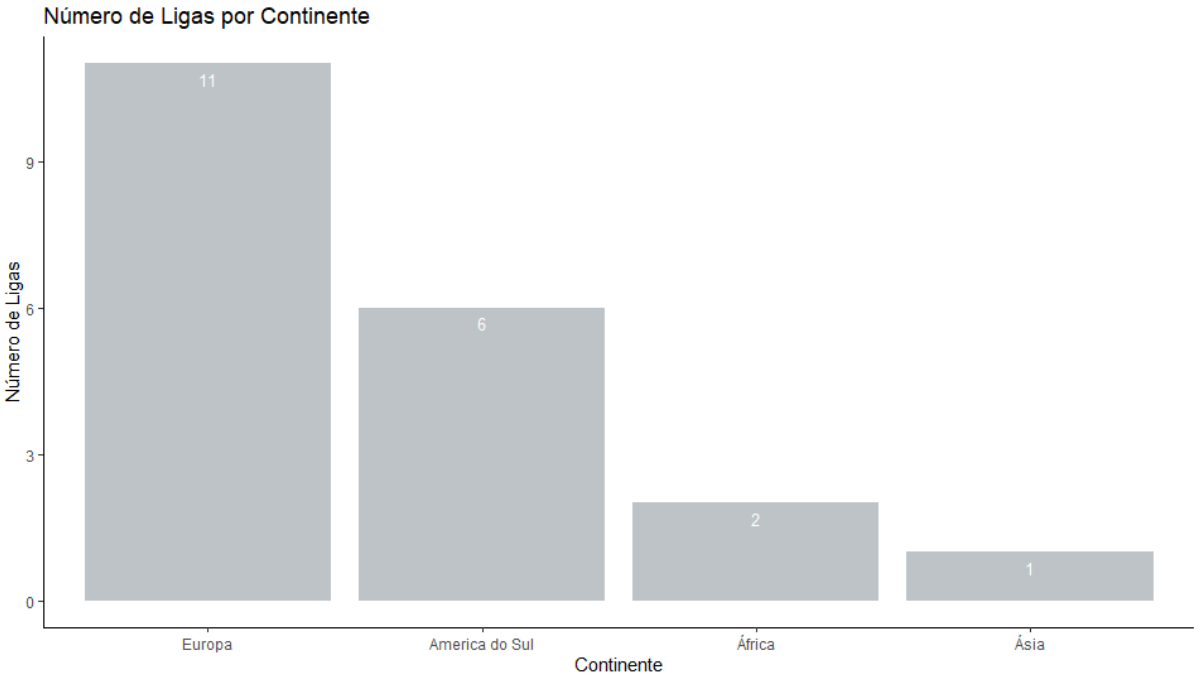

Figura 28 – Número de Ligas por Continente

Fonte: O Autor(2023)

Pela Figura 29, nota-se que uma 40% dos campeões das vinte melhores ligas de 2022 também foram campeões do último campeonato realizado.

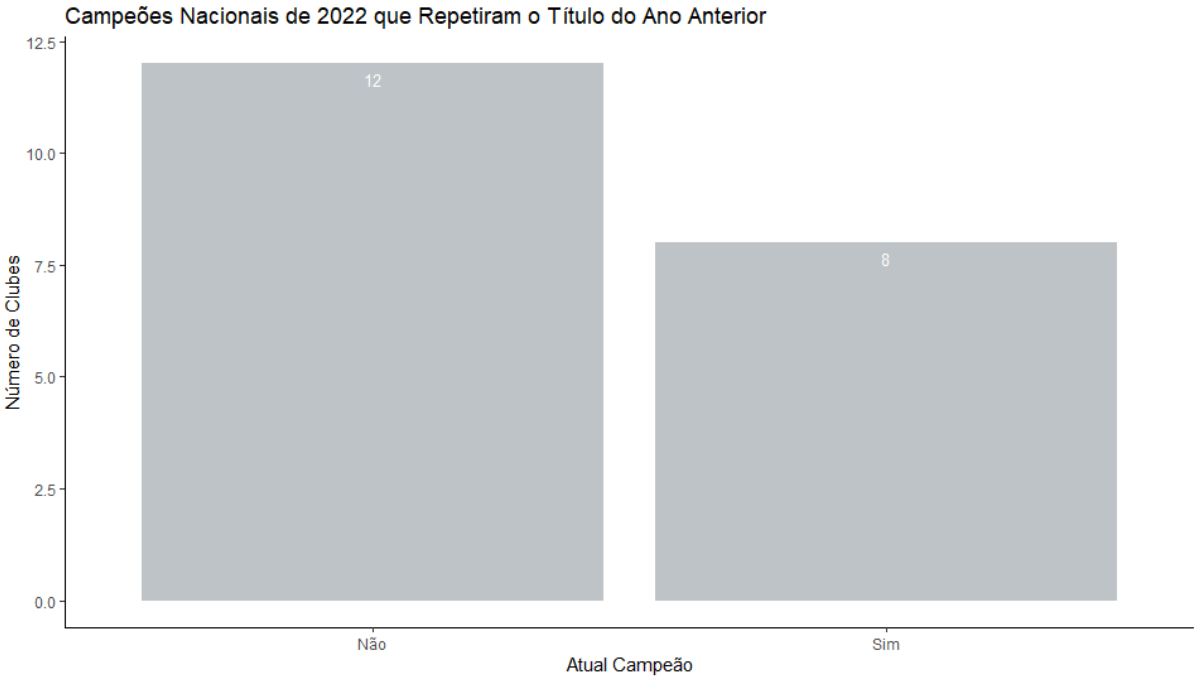

Figura 29 – Campeões Nacionais de 2022 que Repetiram o Título do Ano Anterior

Fonte: O Autor(2023)

A Tabela 18 apresenta a média composicional dos dados de resultados. Nota-se uma predominância de Vitórias na composição, o que é esperado para campeões. Quando comparado à Tabela 12, percebe-se uma maior proporção de vitórias nos campeões nacionais de 2022 do que os vencedores do Campeonato Brasileiro de 2003 a 2022.

Tabela 18 – Média composicional dos Campeões Nacionais de 2022

Fonte: O Autor (2024)

Vitórias	Empates	Derrotas
0,695	0,197	0,108

Na Tabela 19 é apresentada a matriz de variação do conjunto de dados. A partir da Tabela 13 é possível também calcular a variância total, a qual tem valor de 0,356.

Tabela 19 – Matriz de variação dos Campeões Nacionais de 2022

Fonte: O Autor (2024)

	Vitórias	Empates	Derrotas
Vitórias	0,000	0,228	0,502
Empates	0,228	0,000	0,340
Derrotas	0,502	0,340	0,000

Para uma melhor exploração da base de dados serão utilizados os gráficos de barras horizontais estacados ordenados pelos gols sofridos, o diagrama ternário e o gráfico de radar, vistos na Seção 2.3. O ponto em destaque de vermelho na Figura 31 é o Palmeiras, que tem 60,5 % de vitórias, 31,6% de empates e 7,9% de derrotas.

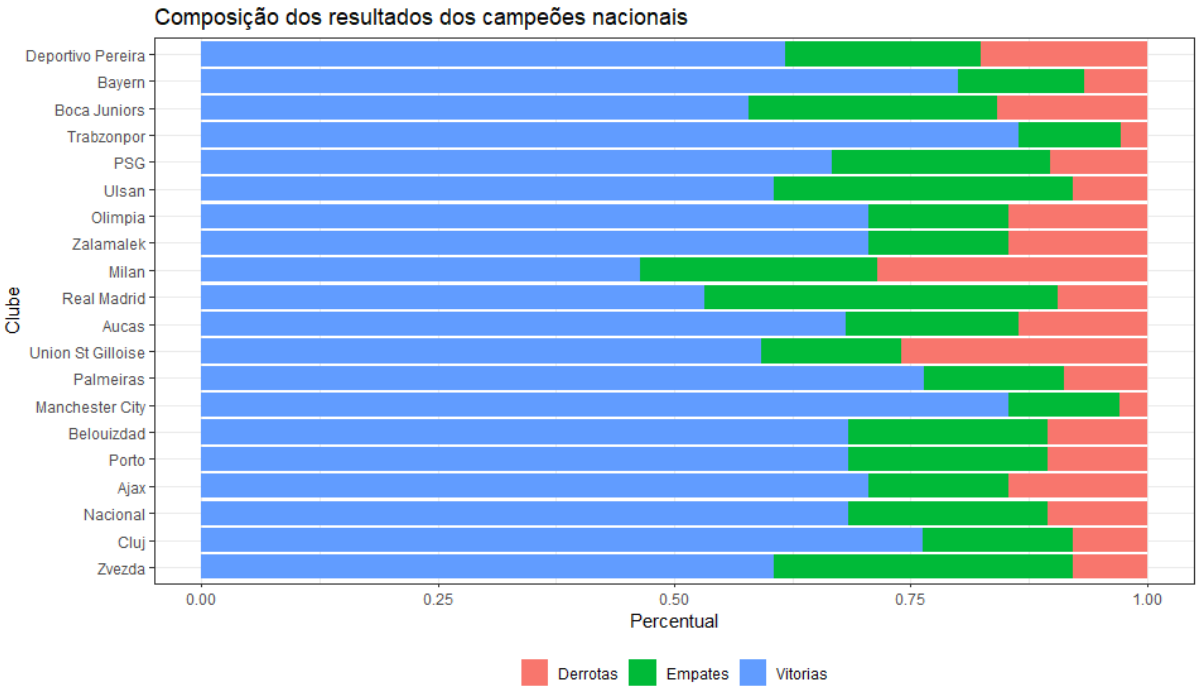

Figura 30 – Gráfico de Barras Horizontais Estacadas de Campeões Nacionais de 2022

Fonte: O Autor(2023)

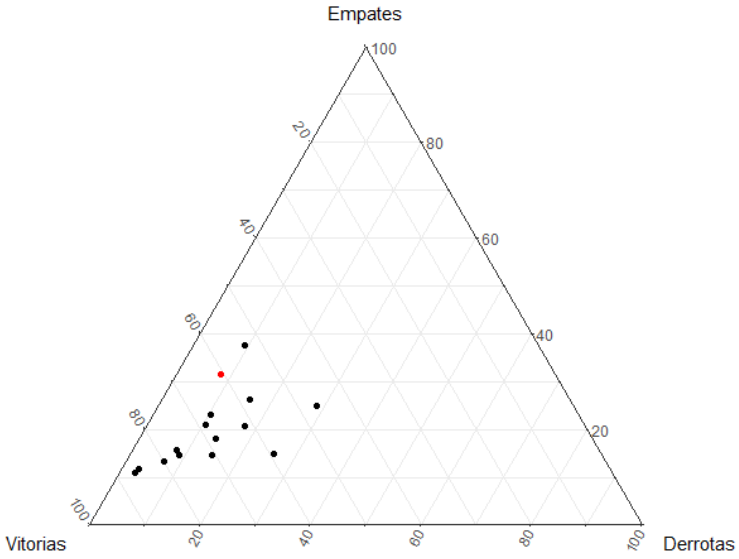

Figura 31 – Diagrama Ternário de Campeões Nacionais de 2022

Fonte: O Autor(2023)

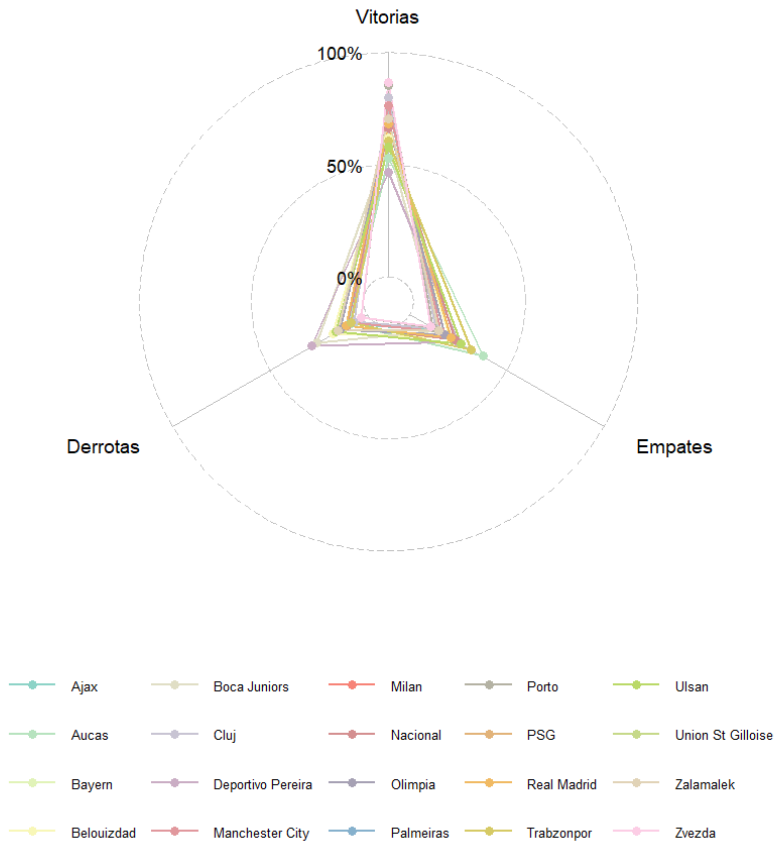

Figura 32 – Gráfico de Radar dos Campeões Nacionais de 2022

Fonte: O Autor(2024)

Como esperado e visto pela média composicional, nota-se a predominância de vitórias nas composições. Destacam-se o Porto e o Zvezda que sofreram apenas uma derrota em suas respectivas ligas.

5.3.1 Regressão Dirichlet para Campeões Nacionais de 2022

Para modelar a relação de vitórias, empates e derrotas e outros aspectos da campanha de um campeão nacional em 2022, adota-se o seguinte modelo de regressão

$$\boldsymbol{\eta} = \boldsymbol{\beta}_0^T + \boldsymbol{\beta}_1^T GolsMarcados + \boldsymbol{\beta}_2^T GolsSofridos + \boldsymbol{\beta}_3^T CV$$

em que $\boldsymbol{\eta} = (\eta_1 = \text{Vitórias}, \eta_2 = \text{Empates}, \eta_3 = \text{Derrotas})^T$ é o vetor de preditores, $\boldsymbol{\beta}_0^T = (\beta_{10}, \beta_{20}, \beta_{30})$ é o vetor de coeficientes dos interceptos de cada variável resposta, $\boldsymbol{\beta}_1^T = (\beta_{11}, \beta_{21}, \beta_{31})$ é o vetor dos coeficientes de regressão relacionados aos gols marcados, $\boldsymbol{\beta}_2^T = (\beta_{12}, \beta_{22}, \beta_{32})$ é o vetor dos coeficientes de regressão relacionados aos gols sofridos

e $\boldsymbol{\beta}_3^T = (\beta_{13}, \beta_{23}, \beta_{33})$ é o vetor dos coeficientes de regressão relacionados aos cartões vermelhos recebidos. Utilizando a parametrização comum, foram calculados os coeficientes de regressão mostrados na Tabela 20.

Tabela 20 – Coeficientes de regressão calculados do conjunto de dados Campeões Nacionais de 2022

Parâmetro	Estimativa	Erro Padrão	z-valor	Pr(>|z|)
β_{10}	$-1,613$	$2,595$	$-0,622$	$0,534$
β_{11}	$1,610$	$0,677$	$2,378$	$0,017*$
β_{12}	$4,059$	$2,010$	$2,020$	$0,434*$
β_{13}	$-0,307$	$0,159$	$-1,935$	$0,530$
β_{20}	$-2,111$	$2,477$	$-0,852$	$0,394$
β_{21}	$1,041$	$0,659$	$1,581$	$0,114$
β_{22}	$4,613$	$1,951$	$2,364$	$0,018*$
β_{23}	$-0,317$	$0,165$	$-1,916$	$0,055$
β_{30}	$-4,006$	$2,604$	$-1,538$	$0,124$
β_{31}	$1,014$	$0,699$	$1,449$	$0,147$
β_{32}	$6,468$	$2,094$	$3,089$	$0,002*$
β_{33}	$-0,356$	$0,154$	$-2,311$	$0,021*$

Sob um nível de significância de 5%, o intercepto não se mostrou significante para nenhum preditor. Por outro lado, o número de gols sofridos foi significante para as três variáveis resposta, enquanto o número de gols marcados foi significante nas vitórias e os cartões vermelhos significantes nas derrotas. Assim, o modelo pode ser reescrito como

$$\begin{aligned}
\eta_1 &= 1,610 GolsMarcados + 4,059 GolsSofridos \\
\eta_2 &= 4,613 GolsSofridos \\
\eta_3 &= 6,468 GolsSofridos - 0,356 CartoesVermelhos
\end{aligned}$$

Utilizando as variáveis explicativas do conjunto de dados e os coeficientes de regressão, é possível predizer os valores das variáveis respostas. Na Figura 33 estão plotados em preto os valores reais das composições observadas nos campeonatos brasileiros e em azul os valores preditos pelo modelo de regressão. Na Figura 34 está o gráfico de barras estacadas dos valores preditos do modelo de regressão ajustado.

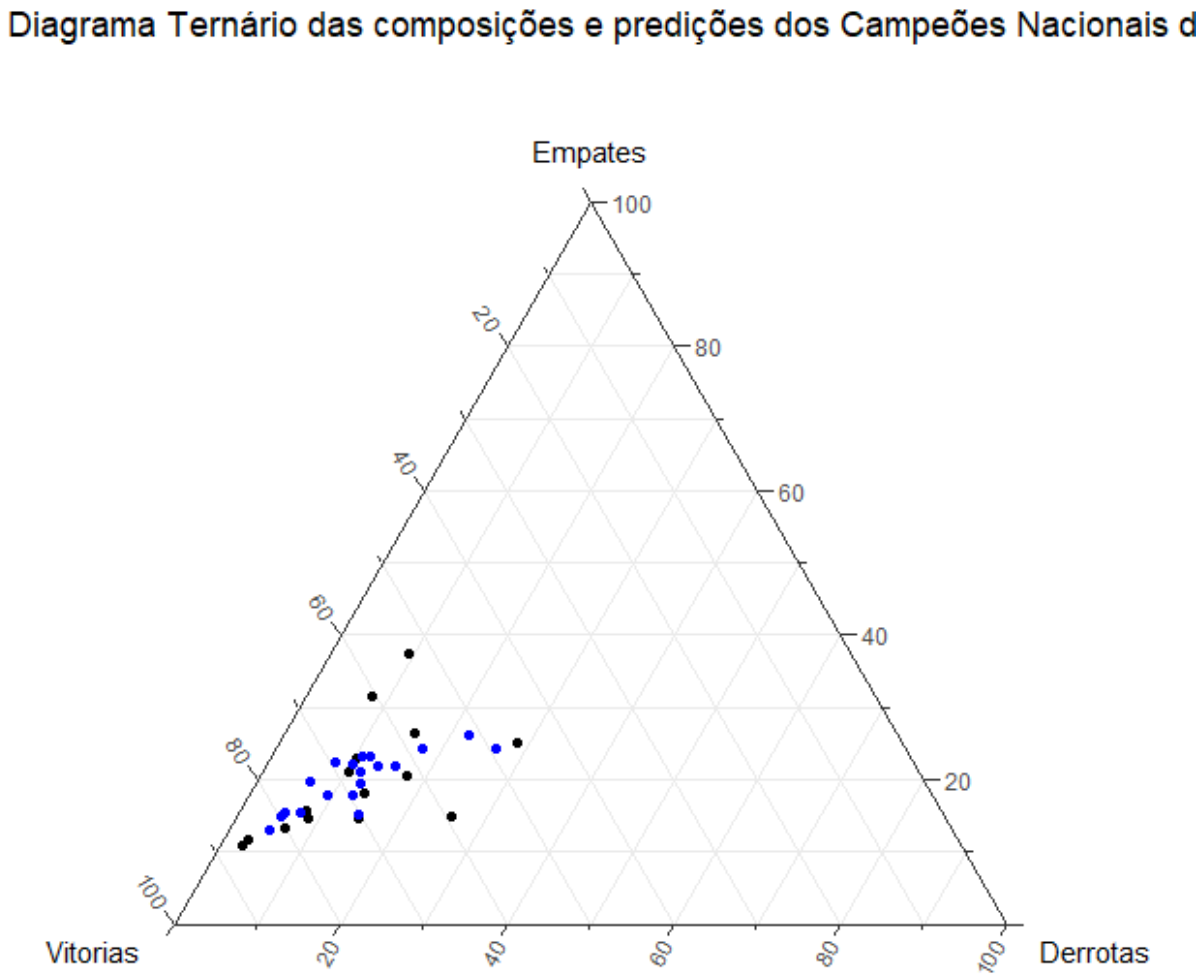

Figura 33 – Diagrama Ternário de Campeões Nacionais de 2022

Fonte: O Autor(2023)

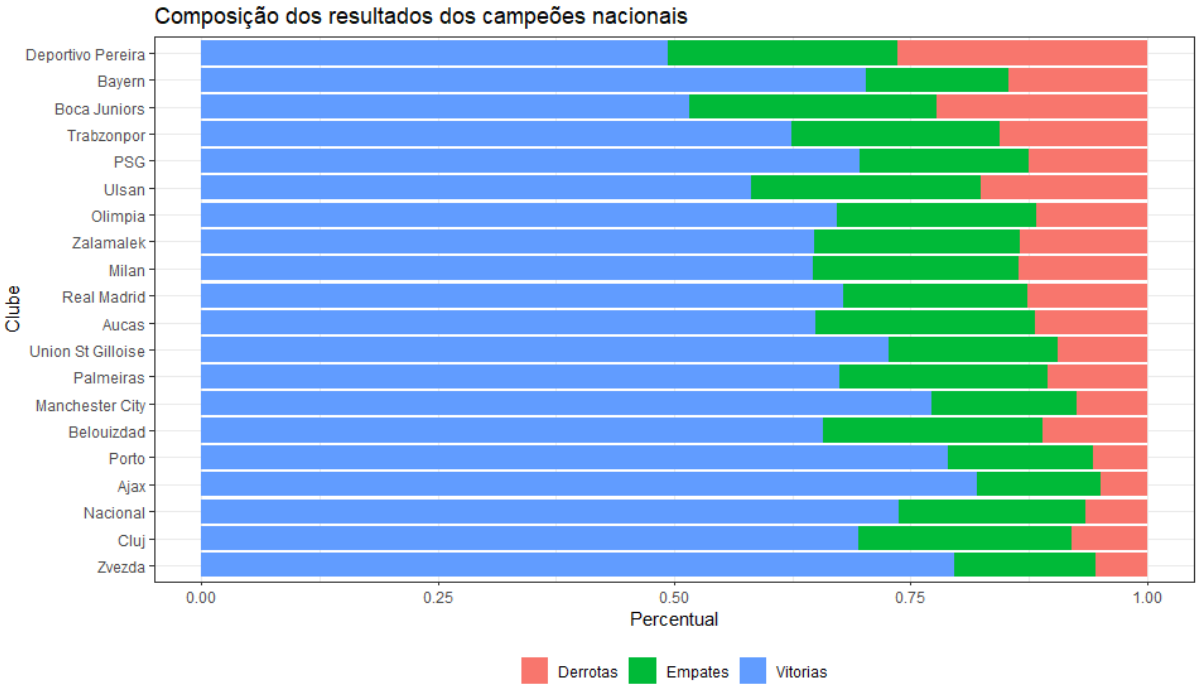

Figura 34 – Gráfico de Barra de predições de Campeões Nacionais de 2022

Fonte: O Autor(2024)

6 CONCLUSÃO

Neste trabalho, foi possível entender grande parte do processo de uma análise de regressão de dados composicionais. O estudo sobre a regressão Dirichlet mostrou-se bastante útil nas análises e simulações, apesar das limitações e desafios apresentados. Dentre essas limitações e desafios estão o estimador na simulação com três preditores e três covariáveis que tanto o EQM quanto o viés absoluto não diminuíram conforme o tamanho da amostra aumentou, não ter feito simulação com mais covariáveis e preditores e o baixo número de variáveis resposta testadas devido ao tamanho dos conjuntos de dados de futebol (20 observações cada).

De modo preliminar, foi vista a importância e particularidades nas análises de dados composicionais. Com as propriedades, transformações e visualizações apresentadas é possível fazer uma análise inicial desse tipo de dados. A seguir, foi mostrada a distribuição de Dirichlet, uma estrutura adequada ao trabalho com dados composicionais, suas parametrizações e como pode-se construir modelos de regressão a partir desta distribuição.

Apresentada a teoria, nas simulações realizadas observou-se o bom comportamento dos estimadores dos coeficientes de regressão, ou seja, diminuição tanto do viés absoluto quanto do erro quadrático médio conforme o aumento de tamanho das amostras.

Nas aplicações, puderam ser utilizadas as técnicas previamente apresentadas para construir modelos de regressão Dirichlet. O modelo feito para os sedimentos do Arctic Lake, mostrou-se mais eficiente nas profundidades intermediárias. Com o conjunto de dados de campeões brasileiros notou-se as diferenças entre as duas parametrizações da regressão Dirichlet, cuja parametrização alternativa depende da variável de referência. No último conjunto de dados, o dos campeões nacionais de 2022, foi novamente realizada a regressão com a parametrização tradicional e a variável resposta número de gols sofridos mostrou-se significante tanto para vitórias, quanto empates e derrotas.

REFERÊNCIAS

AITCHISON, J. The statistical analysis of compositional data. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 44, n. 2, p. 139–160, 1982.

AITCHISON, J. **The Statistical Analysis of Compositional Data**. [S.l.]: Springer Netherlands, 1986.

AITCHISON, J. A concise guide to compositional data analysis. In: **Compositional Data Analysis Workshop**. [S.l.: s.n.], 2005.

GREENACRE, M. **Compositional data analysis in practice**. [S.l.]: CRC press, 2018.

IFFHS. **IFFHS MEN'S STRONGEST NATIONAL LEAGUE IN THE WORLD 2022**. 2023. Disponível em: https://iffhs.com/posts/2483. Acesso em: 09 nov.2023.

KROESE, D. P. et al. Why the monte carlo method is so important today. **Wiley Interdisciplinary Reviews: Computational Statistics**, Wiley Online Library, v. 6, n. 6, p. 386–392, 2014.

MAIER, M. Dirichletreg: Dirichlet regression for compositional data in r. 2014.

MELO, T. F. et al. Higher-order asymptotic refinements in the multivariate dirichlet regression model. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 51, n. 1, p. 53–71, 2022.

MELO, T. F.; VASCONCELLOS, K. L.; LEMONTE, A. J. Some restriction tests in a new class of regression models for proportions. **Computational statistics & data analysis**, Elsevier, v. 53, n. 12, p. 3972–3979, 2009.

MONTGOMERY, D.; PECK, E.; VINING, G. **Introduction to Linear Regression Analysis**. Fifth. [S.l.]: Wiley series in probability and statistics, 2012. ISBN 9780470542811; 0470542810.

PAWLOWSKY-GLAHN, V.; EGOZCUE, J. J.; TOLOSANA-DELGADO, R. **Modeling and analysis of compositional data**. [S.l.]: John Wiley & Sons, 2015.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. Disponível em: https://www.R-project.org/.

ZALCMAN, F. **Quantas medalhas o Brasil já ganhou em Jogos Olímpicos**. 2024. Disponível em: https://olympics.com/pt/noticias/quantas-medalhas-brasil-ganhou-jogos-olimpicos. Acesso em: 14 set.2024.