UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

ANA PAULA DE CASTRO SILVA

ANÁLISE DE DADOS COMPOSICIONAIS

JUIZ DE FORA
2021

ANA PAULA DE CASTRO SILVA

ANÁLISE DE DADOS COMPOSICIONAIS

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a colação do grau em Bacharel em Estatística.

Orientador: Tiago Maia Magalhães.

JUIZ DE FORA

2021

FICHA CATALOGRÁFICA

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

Castro Silva, Ana Paula de .

Análise de Dados Composicionais / Ana Paula de Castro Silva. --2021.

91 p. : il.

Orientador: Tiago Maia Magalhães

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2021.

1. Dados Composicionais. 2. Transformações Composicionais. 3. Regressão Ridge. 4. Regressão Composicional. I. Magalhães, Tiago Maia, orient. II. Título.

ANA PAULA DE CASTRO SILVA

ANÁLISE DE DADOS COMPOSICIONAIS

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a colação do grau em Bacharel em Estatística, aprovada em 02/09/2021.

BANCA EXAMINADORA

Tiago Maia Magalhães

Doutor em Estatística – USP – Orientador

Camila Borelli Zeller

Doutora em Estatística - UNICAMP

Clécio da Silva Ferreira

Doutor em Estatística - USP

Para minha família.

AGRADECIMENTOS

À Deus, pela minha vida, por me dar forças e perseverança para chegar até aqui e por fazer Sua presença visível em todas as situações. À minha família, minha base, que sempre me apoiou e me aconselhou, ao meu pai de maneira racional, à minha mãe de maneira emocional e ao meu irmão de maneira objetiva. Aos meus amigos da vida, que sempre me ouviram e me encorajaram em cada desânimo passado. À Rebecca, companheira das emoções e decisões. À Gabriela, companheira de casa e conselheira. À Jéssica, companheira de todas as disciplinas e de vida. Ao Calvin, companheiro das atividades e risadas. Ao Carlinhos, companheiro de encontros e de comemorações. Aos meus amigos de graduação. Obrigada, por tanto.

Agradeço ao professor Tiago por aceitar ao desafio de me orientar, por me auxiliar nas dificuldades e por todo conhecimento compartilhado. À Universidade Federal de Juiz de Fora (UFJF), pela experiência maravilhosa nesses anos e por fornecer condições para um bom aprendizado. Aos professores do departamento de Estatística, pelos ensinamentos, lições, dedicação e paciência ao longo da minha graduação. Aos professores da banca por aceitarem o convite e agregarem ao Trabalho. Muito obrigada!

"Mantenha o foco. Nunca desvie seus olhos, porque se uma abertura surge, mesmo nosso poder insignificante pode ser suficiente para determinar o destino do mundo. É por isso que todos devem ficar alertas e prontos para reagir a qualquer momento!"
(Shikamaru Nara)

"A vida é como um lápis que certamente se esgotará, mas deixará uma bela escrita."
(Nami)

RESUMO

Os Dados Composicionais (ou *Compositional Data, CoDa*, acrônimo inglês) são descrições quantitativas das partes de um todo, que transmitem informações de forma relativa ao total. O grande diferencial, nesse tipo de situação, é que as componentes dos dados apresentam soma constante, 1 - para proporções e 100 - para porcentagens. Devido a esta restrição, os dados composicionais possuem uma estrutura própria, denominada *Simplex* (espaço natural para tais dados), no qual as operações realizadas podem não ter correspondência evidente com métodos usuais do espaço real. Este fato culmina no uso de algumas transformações, que permitem uma equivalência entre o espaço euclidiano (real) e o espaço *simplex*, sendo possível usufruir das facilidades usuais da estatística e depois retornar ao espaço com restrições. O trabalho abordou as transformações composicionais, exibiu os princípios da análise composicional, enunciou as operações em composições, demonstrou a análise exploratória apropriada e, por fim, sugeriu alternativas para adequação de modelos, além de explicitar simulações e aplicações práticas, que demonstraram opções de análise de regressão utilizando os métodos composicionais em detrimento aos métodos usuais.

Palavras-chave: Dados Composicionais. Transformações Composicionais. Regressão Ridge. Regressão Composicional.

ABSTRACT

Compositional Data (CoDa) are quantitative descriptions of the parts of a whole, which convey information relative to the total. The big difference, in this type of situation, is that the data components present a constant sum, 1 - for proportions and 100 - for percentages. Due to this restriction, compositional data has its own structure, called *Simplex* (natural space for such data), in which the operations performed may not have a clear correspondence with usual methods of real space. This fact culminates in the use of some transformations, which allow an equivalence between the Euclidean (real) space and the *simplex* space, making it possible to take advantage of the usual facilities of statistics and then return to the space with restrictions. The work addressed compositional transformations, exhibited the principles of compositional analysis, enunciated the operations in compositions, demonstrated the appropriate exploratory analysis and, finally, suggested alternatives for adapting models, in addition to explaining simulations and practical applications, which demonstrated analysis options of regression using the compositional methods in detriment to the usual methods.

Keywords: Compositional Data. Compositional Transformations. Regression Ridge. Compositional Regression.

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTO E PROBLEMA

Os Dados Composicionais (ou *Compositional Data, CoDa*, acrônimo inglês), em termos estatísticos, são descrições quantitativas das partes do todo que transmitem informações de forma relativa ao total. As medições que envolvem probabilidades ou proporções podem ser pensadas como dados composicionais. A peculiaridade, nesse tipo de situação, é que as componentes dos dados apresentam soma constante (1 - para proporções e 100 - para porcentagens). Apesar dessa estrutura ser facilmente encontrada em várias áreas de estudo, a análise de dados composicionais foi sistematizada, somente, na década de 80 por Aitchison (1986) e desde então, várias técnicas e métodos têm sido desenvolvidos para modelagem dos dados composicionais. Coenders e Pawlowsky-glahn (2020) fazem uma revisão bibliográfica exemplificando a aplicabilidade de um *CoDa* como, por exemplo, nas Geociências, na Medicina e no Turismo.

Por apresentarem a restrição de soma constante, os dados composicionais possuem uma estrutura própria, denominada *Simplex.* O *Simplex* é um espaço amostral natural para tais dados, porém, provou ser um espaço amostral complicado para ser tratado estatisticamente devido à ausência de classes paramétricas satisfatórias (PAWLOWSKY-GLHAN; BUCCIANTI, 2011). Buchanan *et al.* (2012) definiram esse espaço complexo como uma representação geométrica de atributos, em que uma composição de *D* partes é representada por um número mínimo de vértices correspondente ao número de dimensões.

O espaço amostral de dados composicionais de *D* partes, designado por *D−simplex*, e denotado por S^D, é definido por:

$$S^D = \left\{ x; x_i > 0, i = 1, \ldots, D; \sum_{i=1}^{D} x_i = c \right\}$$

Karl Pearson (1897) e Aitchison (1986) alertaram sobre a correlação espúria em dados com restrições, isto é, a existência de uma relação estatística entre duas ou mais variáveis, em que não existe nenhuma explicação lógica ou significado teórico. Sendo assim, devido a este problema, métodos usuais da Análise Multivariada são incapazes de interpretar os coeficientes

de correlação entre as componentes dos dados. Esse fato ocorre com frequência quando se lida com dados em que a soma das componentes é constante.

Aitchison (1986) concluiu que todas as análises das partes que compõem um todo poderiam ser realizadas em termos de razões das partes da composição. E, visto que a transformação do logaritmo das razões entre as variáveis (log-razão) é uma correspondência biunívoca em $\mathbb{R}$, o tratamento matemático de um quociente é mais simples em termo de seu logaritmo. Dessa forma, Aitchison (1986) propôs metodologias baseadas em vários tipos de transformações log-razões. Essas transformações permitiram a aplicação de procedimentos de Análise Multivariada sobre os dados transformados, traduzindo, em seguida, as conclusões extraídas em termos de dados originais (PAWLOWSKY-GLAHN et. Al., 2015).

De forma geral, a regressão linear visa propor e estimar um modelo, de modo que a variável resposta (caso univariado) ou as variáveis respostas (caso multivariado) dependam linearmente de uma ou mais covariáveis. No caso dos dados composicionais, as componentes podem aparecer como variáveis dependentes ou independentes em modelos lineares. Para que todas as restrições e peculiaridades das composições sejam respeitadas na modelagem, é necessário propor um método de regressão apropriado, que descreva a estrutura composicional, a fim de estimar parâmetros de interesse e realizar predições.

Assim sendo, em consonância com Aitchison (1986), os métodos usuais não são apropriados para a análise dos dados composicionais, dessa forma, apresentamos as técnicas necessárias para a análise destes dados: como transformações e operações. Além disso, foram realizadas simulações que objetivaram observar o comportamento de três diferentes transformações na Regressão Composicional. Por fim, realizamos três aplicações em bancos de dados composicionais, a fim de compararmos o desempenho entre a regressão Clássica e a regressão Composicional.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo geral desta monografia é estudar a metodologia de dados composicionais, a fim de indicar técnicas que evidenciem a importância de abordar as particularidades das

restrições no que concerne a soma constante. Ademais, busca-se observar o comportamento em modelos quando as covariáveis são composicionais.

1.2.2 Objetivos Específicos

O primeiro objetivo específico é apresentar os métodos adequados para a correspondência entre o espaço euclidiano e o espaço *simplex,* e, a partir disso, as possíveis análises descritivas. O segundo objetivo específico é simular diferentes modelos de regressão composicional, a fim de obter a comparação dos resultados com a regressão usual. O terceiro objetivo específico, consequentemente, é apresentar o desempenho com a aplicação da metodologia composicional e com a aplicação da metodologia usual.

1.3 JUSTIFICATIVA

Quando se utilizam métodos usuais para análises em dados com restrição, conforme dito na Seção 1.1, problemas de correlações espúrias podem ocorrer. Nesse sentido, faz-se necessária a implementação de transformações que permitam a utilização de técnicas adequadas (KARL PEARSON, 1897; AITCHISON, 1986).

Na existência de um conjunto de variáveis, pode-se utilizar a transformação composicional para que qualquer subconjunto dos dados originais possa ser considerado uma subcomposição. Tal transformação é denominada fechamento dos dados, ou operação de fecho $C(x_i)$, isto é, cada valor original (x_i) é dividido pelo somatório de todos os outros valores da amostra:

$$C[x_i] = \frac{x_i}{\sum_{i=1}^{D} x_i}. \quad (1)$$

Apesar disso, a transformação composicional desloca os dados de um espaço real para um espaço com restrições, no qual as operações realizadas podem não ter correspondência evidente com métodos usuais.

O trabalho se justifica por abordar tais transformações supracitadas, por exibir os princípios da análise composicional, por mostrar as operações em composições, por trabalhar a

análise exploratória apropriada e por adequar os modelos, além de explicitar o funcionamento de tal técnica por meio de aplicações práticas, que, futuramente, poderão ser úteis para comparações da eficácia e da confiabilidade de estudos utilizando os métodos usuais e os métodos composicionais.

1.4 APLICAÇÕES EM PROBLEMAS COMPOSICIONAIS

No início, as principais aplicações de *CoDa* estavam nas Geociências, por exemplo em Grantham e Velbel (1988), em Kuhn et al. (2000) e em Tolosama-Delgado e Eynatten (2010). Porém, houve um avanço em outras áreas, por exemplo, Chastin et al. (2021) investigaram a relação entre a mortalidade de um indivíduo e suas atividades cotidianas, como tempo gasto com exercícios e duração do sono. Algumas das possíveis aplicações da metodologia composicional são apresentadas a seguir.

- Turismo

Um dono de hotel é capaz, dentre outras coisas, de decidir o que é mais rentável para si e mais agradável para seu hóspede. Supondo que o tempo gasto na estadia pelo turista é um todo dividido em partes, por exemplo: academia, piscina, restaurante e atividades de lazer. Pode-se realizar uma análise composicional para inferir o que é mais satisfatório para os clientes. Além disso, podem surgir questões como quais as origens e quais os destinos concentram a maior parte dos fluxos turísticos.

- Estudos Comportamentais

Os dados composicionais também podem ser considerados em questões mais subjetivas como o comportamento. Por exemplo, menos tempo sedentário pode ser associado a uma menor adiposidade (% de gordura corporal). É possível estudar as associações entre a adiposidade e uma duração fixa de atividade, conceituando as atividades diárias dos indivíduos como composições (tempo gasto no sono, comportamento sedentário e atividade física leve), que são restringidos por uma soma constante diária de, por exemplo, alguns minutos.

- Análises de Orçamento de Tempo

A regressão com resposta composicional, ou mesmo regressão entre partes de uma composição, é frequentemente empregada nas ciências sociais. Entre outras aplicações

possíveis, também pode ajudar a revelar características interessantes na análise de alocação de tempo. Como as atividades individuais representam contribuições relativas à quantidade total de tempo, o processamento estatístico de dados brutos (frequentemente representado como proporções ou porcentagens) pode ser realizado com a abordagem composicional.

- Qualidade de Alimentos

Na tentativa de melhorar a qualidade do leite das vacas, por exemplo, um experimento pode ser realizado. O leite de um grupo de vacas pode ser avaliado pela composição antes e depois de uma dieta estritamente controlada e um regime hormonal durante um certo período. Sabendo que o leite é composto por proteína, gordura do leite, carboidrato, cálcio, sódio, proporções de potássio por peso do conteúdo alimentar total, com o experimento é possível determinar se o novo regime produziu qualquer mudança na composição do leite.

2 METODOLOGIA

2.1 ESPAÇO DAS COMPOSIÇÕES

As composições integram uma estrutura em que as propriedades do espaço euclidiano não são adequadas. Desse modo, esta seção se dedica a exemplificar as mudanças existentes no espaço das composições. Algumas outras notações e conceitos relacionados a espaços vetoriais podem ser consultados em Lawson (1997), por exemplo.

2.1.1 Princípios da Análise Composicional

Aitchison (1986) indicou 4 princípios sobre os quais se devem reger as técnicas adequadas de análise de dados composicionais, dentre estes:

1. Invariância de Escala:

Quando um problema é composicional, deve-se reconhecer que o valor absoluto das partes que compõem as amostras é irrelevante, uma vez que composições equivalentes contêm essencialmente a mesma informação. Por exemplo: as composições [12, 3, 4], [12/17,3/17, 4/17] são equivalentes.

2. Invariância por Perturbação:

Quando ocorre mudança de unidades, as composições continuam equivalentes. Exemplo: Não há diferença de quilogramas para gramas.

3. Invariância por Permutação:

Os resultados de qualquer análise não dependem da ordem em que os componentes se apresentam nos dados.

4. Coerência Subcomposicional:

As subcomposições são composições com apenas uma parte das componentes. Usualmente se trabalha com subcomposições, já que dificilmente se analisam todos os possíveis componentes de uma amostra. Sobre estas, pode-se dizer que as distâncias entre duas composições serão maiores do que as distâncias entres as respectivas subcomposições. Além disso, a dispersão total entre um conjunto de dados composicionais é maior do que a dispersão

para um conjunto das respectivas subcomposições. Ademais, o estudo de uma subcomposição não pode conduzir a resultados contraditórios com os obtidos a partir da composição total.

2.1.2 Operações com Composições

Existem duas operações puramente composicionais conhecidas na literatura por perturbação e potenciação, introduzidas por Aitchison (1986), que permitem conferir ao *simplex* de *D* observações a estrutura de um espaço vetorial e, deste modo, definir bases, linha retas e outros operadores no simplex.

-Perturbação ou soma

Considerando duas composições $x = (x_1, x_2, \dots, x_d)^T \; e \; y = (y_1, y_2, \dots, y_d)^T \in S^D$. A perturbação de x por y é dada por:

$$x \oplus y = C(x_1 y_1, x_2 y_2, \dots, x_d y_d),$$

em que *C(.)* é a operação de fecho, definida na equação (1).

-Potenciação ou Produto por uma Constante

Considerando uma composição $x = (x_1, x_2, \dots, x_d)^T \in S^D$ e um escalar $\alpha \in \mathbb{R}$. A potenciação de x por α é dada por:

$$x \odot \alpha = C(x_1^\alpha, x_2^\alpha, \dots, x_d^\alpha),$$

em que *C(.)* é a operação de fecho, equação (1).

Além dessas duas operações, pode-se também definir o produto interno de Aitchison:

$$\langle x, \mathrm{y} \rangle_A = \sum_{i=1}^{D} \left(ln \frac{x_i}{g(x)} \times ln \frac{y_i}{g(y)} \right),$$

em que $g(x)$ é a média geométrica de x, definida da seguinte forma:

$$g(x) = \sqrt[n]{x_1 \times x_2 \times \dots \times x_D}.$$

2.2 ANÁLISE DESCRITIVA

A estatística descritiva composicional se diferencia da tradicional e baseia-se na proporção entre os elementos. Esta seção se dedica a expor as formas adequadas de retirar informações e produzir visualizações das composições.

2.2.1 Diagrama Ternário

Uma forma de visualizar a distribuição da amostra pelas componentes é dada pelo diagrama com formato de triângulo, como exemplo, ver a Figura 1. O diagrama pode ser interpretado como:

i. as (sub) composições se concentram num vértice: indica a dominância da parte associada a esse vértice;

ii. as (sub) composições se distribuem ao longo de uma aresta: indica a dominância das partes associadas a essa aresta;

iii. as (sub) composições se concentram em torno do baricentro do *simplex*: indica que as partes representadas têm proporções aproximadamente iguais;

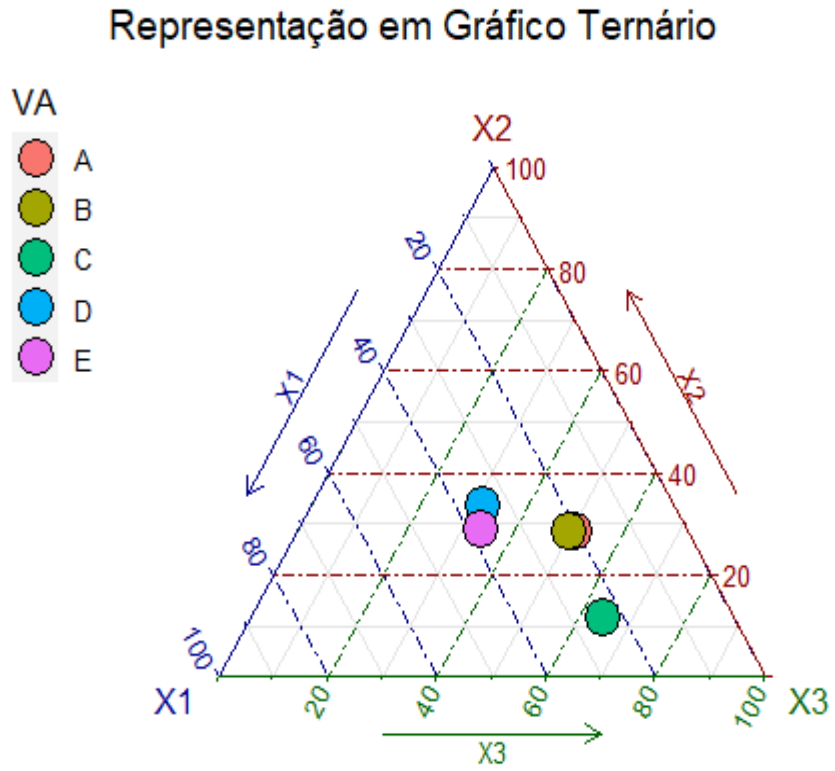
iv. as (sub) composições formam um padrão linear paralelo a um dos lados: indica que as proporções da parte associada ao vértice oposto nas (sub) composições são (aproximadamente) constantes;

v. as (sub) composições formam um padrão linear (aproximadamente) perpendicular a um dos lados: indica que as parte associadas a esse lado são (aproximadamente) proporcionais (reduzida variabilidade relativa);

vi. as (sub) composições estão dispersas no simplex: indica que as partes apresentam elevada variabilidade relativa entre si.

A partir da Figura 1, pode-se exemplificar tal ferramenta. Considera-se que as variáveis $A, B\ e\ C$ apresentam maiores influências da característica X_3, enquanto as variáveis $D\ e\ E$ possuem proporções aproximadamente iguais para as três características: $X_1, X_2\ e\ X_3$, já que se encontram aproximadamente no baricentro do diagrama.

Figura 1 - Representação de Diagramas Ternários



Fonte: A Autora (2021).

2.2.2 Gráfico de Radar

Outra forma de representar tais dados pode ser descrita pelos gráficos de radar, também conhecidos como gráficos de teia, gráficos de aranha, gráficos de estrela, polígonos irregulares, gráficos polares, ou diagramas Kiviat. Tal gráfico é equivalente a um gráfico de coordenadas paralelas em coordenadas polares e fácil de interpretar, contendo os seguintes elementos:

Ponto central: núcleo de um gráfico de radar, a partir do qual diferentes eixos são desenhados.

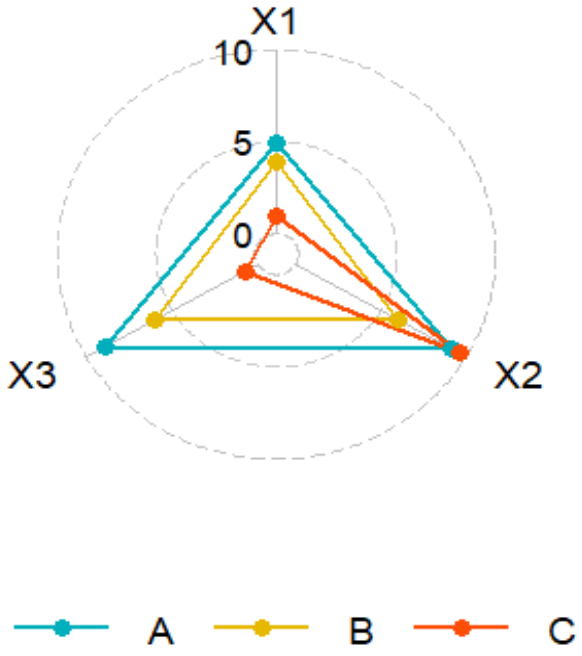
Eixo: cada eixo indica uma variável.

Grades: quando os eixos são conectados em um gráfico de radar, ele divide o gráfico completo em grades distintas que ajudam a constituir os fatos de uma maneira mais elevada.

Valores: Uma vez que o gráfico é desenhado, constituem-se vários valores em cada eixo e o gráfico é traçado através da atribuição de cores.

A Figura 2 apresenta uma possível representação dos gráficos de setores, pode-se afirmar, a partir de tal representação, que a variável C, por exemplo, apresenta maiores características de X_2 e que a variável A apresenta maiores características de X_2 e de X_3 que de X_1.

Figura 2 - Representação de Gráficos de Radar



Fonte: A Autora (2021).

Os conjuntos de dados composicionais podem ser representados de diversas formas: diagramas de dispersão com dois componentes, diagramas ternários (apresentados anteriormente), diagramas de dispersão em escalas logarítmicas para as componentes, ou sequência de barras, porém, as duas primeiras formas são mais utilizadas (BOOGAART; TOLOSANA-DELGADO, 2013). As outras técnicas de visualização são descritas em, por exemplo, Aitchison (1986) e Egozcue (2005).

2.2.3 Estatísticas Composicionais Descritivas

As estatísticas convencionais não são adequadas para Dados Composicionais, pois as composições são quantidades relativas. Em particular, a média aritmética e a variância ou desvio padrão das componentes não se adequam como expressões de tendência central e dispersão.

- Média Composicional

Para obter o valor médio de um conjunto de dados composicionais são usadas as operações de soma ou de perturbação e a operação de potência, assim, o centro ou média composicional de um conjunto Y com n composições e D componentes é a composição:

$$\bar{Y} = \frac{1}{n} \odot C \oplus Y_i.$$

A interpretação da média composicional é análoga ao da média aritmética.

- Matriz de Variância

A matriz de variação descreve a dispersão em um conjunto composicional. Quanto menor o elemento, maior é a proporcionalidade entre os dois componentes, além disso, para melhores interpretações, Aitchison (1986) sugere analisá-la como um coeficiente de correlação. Essa matriz $\pi_{ij}, DxD,$ tem os elementos definidos da seguinte forma:

$$\pi_{ij} = var\left(ln\frac{x_i}{x_j}\right).$$

2.3 TRANSFORMAÇÕES COMPOSICIONAIS

Qualquer subconjunto dos dados originais pode ser considerado uma subcomposição com a aplicação da transformação composicional, porém, tal transformação transfere os dados de um espaço real para um espaço com restrições, no qual as operações realizadas podem não ter correspondência evidente com métodos usuais da análise multivariada, conforme visto na Seção 1.3. Existem algumas transformações que permitem uma equivalência entre o espaço euclidiano e o espaço *simplex*. Dentre elas, serão enumeradas a Transformação Logaritmo da

Razão Aditiva, a Transformação Logaritmo da Razão Centralizada e a Transformação Isométrica do Logaritmo da Razão.

2.3.1 Transformação Logaritmo da Razão Aditiva

A transformação Logaritmo da Razão Aditiva, *Additive Log Ratio, ALR,* em inglês, permite reduzir a perturbação e a potenciação a operações comuns de adição e multiplicação no espaço $\mathbb{R}^{D-1}$.

Seja X uma composição de D partes no simplex S^D. A transformação de log-razões aditiva de X é definida como:

$$alr(x_i) = ln\left(\frac{x_i}{x_D}\right), i = 1, \ldots, D-1.$$

E a inversa descrita como:

$$alr(x_i)^{-1} = \exp\left(ln\left(\frac{x_i}{x_D}\right)\right) x_D\,, i = 1, \ldots, D-1.$$

Com a *ALR*, é possível sair do espaço simplex e ir para o espaço dos números reais, além de ser possível inverter e retornar às variáveis originais. Apesar disso, existem possíveis problemas com divisões por zero e a *ALR* leva a dados assimétricos (AITCHISON ET AL., 2002).

2.3.2 Transformação Logaritmo da Razão Centralizada

A transformação logaritmo da Razão Centralizada, *Centered Log Ratio*, *CLR,* em inglês, permite que uma composição se associe a um vetor muldimensional, e que a transformação inversa volte ao domínio composicional. Além disso, a perturbação e a potenciação em S^D correspondem à soma e ao produto no espaço real $\mathbb{R}$, e, considerando que na transformação *CLR* o denominador é a média geométrica das partes, então a análise de dados composicionais em coordenadas *clr*-transformadas satisfaz o princípio de invariância sob permutação, além de levar a variáveis simétricas. Apesar disso, ainda podem ocorrer problemas de divisão por zero.

Seja X uma composição de D partes no simplex S^D. A transformação de log-razões centralizada de X é definida como:

$$clr(x_i) = ln\left(\frac{x_i}{g(x)}\right),$$

$i = 1, \ldots, D$ e $g(x)$ é a média geométrica.

E a inversa descrita como:

$$clr(x_i)^{-1} = \exp\left(ln\left(\frac{x_i}{g(x)}\right)\right) g(x)\,, i = 1, \ldots, D.$$

2.3.3 Transformação Isométrica do Logaritmo da Razão

A transformação Isométrica do Logaritmo da Razão ou *Isometric Log Ratio*, *ILR*, em inglês, provê uma identificação entre o espaço euclidiano $\mathbb{R}^{\mathbb{D}-\mathbb{1}}$ e o espaço simplex S^D ao representar uma composição em função de uma base ortonormal de S^D. Essa transformação é dada por:

$$ilr(x) = (x_i^*, x_2^*, \ldots, x_{D-1}^*),$$

em que $x_i^* = \langle x, e_i \rangle$, isto é,

$x_i^* = \langle clr(x), clr(e_i) \rangle\,,\ i = 1, \ldots, D-1;$

$clr(.)$ é a transformação definida na Seção 2.3.2;

e_i é uma base ortonormal.

A base ortonormal deve ser uma base ortogonal que possui vetores unitários. Ser ortogonal indica que o produto interno entre pares de vetores distintos é igual a zero, ou seja, $\langle u_i, v_j \rangle = 0$, em que $\langle u_i, v_j \rangle = u_1 \times v_1 + u_2 \times v_2 + \cdots + u_i \times v_j$ e ter vetores unitários indica que $\left|\left|u_i\right|\right| = 1$, em que $\left|\left|u\right|\right| = \sqrt{\langle u, u \rangle}$.

Para se transformar uma base ortogonal em uma base ortonormal, no espaço euclidiano, basta fazer com que o conjunto de seus vetores tenham módulo igual a 1, para isso, pode-se normalizar os dados. O processo de normalização consiste em dividir cada vetor pelo seu respectivo módulo (norma), ou seja:

$\widehat{u_i} = \frac{u_i}{||u_i||}$, para *i = 1, 2...,*

em que $\hat{u}$ é um vetor unitário. A base formada por $\hat{u}_1, \hat{u}_2$, ..., será uma base ortonormal.

No caso do espaço *Simplex,* as bases ortonormais podem ser obtidas pelo procedimento de Gram-Schmidt, uma vez que um conjunto independente de composições $D-1$ é dado em S^D.

O primeiro passo consiste em encontrar uma expressão simplificada para a ortogonalidade de duas composições. Para isso, pode-se utilizar a transformação *clr* (Seção 2.3.2). Essa transformação atribui a cada composição de S^D um vetor linha $a = clr(x)$ em $\mathbb{R}$, satisfazendo $\sum_{i=1}^{D} a_i = 0$. Os vetores que satisfazem tal condição constituem um espaço de $\mathbb{R}$, $(D-1)$ – dimensional, denotado como V_S. Assim, observa-se que, ao usar a transformação *clr,* as partes da composição com a média geométrica são convenientes, porque a *clr* fornece uma imagem neutra, isto é, $e = C[1,1,\ldots,1]$ é $clr(e) = [0,0,\ldots,0]$, que é o elemento neutro em $\mathbb{R}$, isto é necessário para se definir o isomorfismo dos espaços lineares entre S^D e V_S.

Então, o produto interno de Aitchison, definido em 2.1.2, pode ser reescrito no espaço da *clr* transformado como segue:

Para $a = clr(x)$ e $b = clr(y)$

$$\langle x, \mathrm{y} \rangle_A = \frac{1}{D} \sum_{i<j} (a_i - a_j)(b_i - b_j) = \frac{1}{D} a M b^T, \tag{2}$$

em que M é uma matriz DxD:

$$M = \begin{pmatrix} D-1 & -1 & -1 & \ldots & -1 \\ -1 & D-1 & -1 & \ldots & -1 \\ -1 & -1 & D-1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & \cdots & D-1 \end{pmatrix}.$$

O lado direito da equação (2) define o produto interno degenerado em $\mathbb{R}$, cuja matriz métrica é *M.* É fácil verificar que *M* é degenerada por possuir dois autovalores: 0 com multiplicidade 1 e *D* com multiplicidade $D-1$, consequentemente *M* é semidefinida positiva. O espaço próprio associado ao autovalor *D* é precisamente o subespaço linear V_S, definido sob a condição $\sum_{i=1}^{D} a_i = 0$, e consequentemente, $clr(y)\, M = D\, clr(y)$. Isso significa que o resultado do produto interno não depende do valor da escala utilizado na transformação *clr*. Tal resultado é condizente com o fato de que uma composição é apenas uma representante de uma

classe de equivalência, conforme citado em Barceló-Vidal (2001), Martín-Fernández (2001) e Pawlowsky-Glahn (2001).

Com o objetivo de criar a base ortonormal, seleciona-se um conjunto de $D-1$ vetores nesse subespaço. Sejam $v_1, v_2, \ldots, v_{D-1}$ definidos como $v_i = [0, \ldots, 0, 1, -1, 0, \ldots, 0]$, o primeiro elemento sendo colocado na i-ésima coluna. Para quaisquer dois vetores a e b em V_S, o produto interno Euclidiano comum é $\langle a, b\rangle = D^{-1} aMb^{`}$, como mencionado. Portanto, o procedimento de Gram-Schimdt com respeito ao produto interno Euclidiano comum, pode ser aplicado a $v_1, v_2, \ldots, v_{D-1}$ e obtém-se o que se segue:

Sejam $u_i \in \mathbb{R}, i = 1, 2, \ldots, D-1$, os vetores:

$$u_i = \sqrt{\frac{i}{i+1}}\left[\frac{1}{i}, \ldots, \frac{1}{i}, -1, 0, \ldots, 0\right],$$

em que i são os elementos.

Os vetores u_i são ortonormais com respeito ao produto interno Euclidiano comum em $\mathbb{R}$ e constituem uma base de $(D-1)-$ dimensões no subespaço linear V_S.

Após encontrar os vetores ortonormais para o espaço Euclidiano comum, pode-se encontrar a base ortonormal no espaço *Simplex* aplicando a transformação clr^{-1} aos vetores u_i.

Sejam $e_i, i = 1, 2, \ldots, D-1$, as composições em S^D

$$e_i = C(\exp u_i) = C\left[\exp\left(\sqrt{\frac{1}{i(i+1)}}, \ldots, \sqrt{\frac{1}{i(i+1)}}, -\sqrt{\frac{1}{i+1}}, 0, \ldots, 0\right)\right],$$

em que *C(.)* é a operação de fecho, equação (1).

As bases $e_i, i = 1,2, \ldots, D-1$, assim definidas, são ortonormais com respeito ao produto interno de Aitchison no espaço S^D.

Tem-se que:

i. $\langle e_i, e_j\rangle_A = 0$, para $i \neq j$;

ii. $||e_i|| = 1, i = 1,2 \ldots, D-1$.

De acordo com Pawlowsky-Glahn et al. (2007), as bases ortonormais no *simplex* podem ser definidas por diferentes formas, o critério de escolha é o impacto na qualidade da interpretação dos dados.

A correspondência da métrica no espaço à custa da métrica no espaço euclidiano, permite a aplicação de técnicas usuais de análise multivariada aos dados composicionais em termos de suas coordenadas *ilr*-transformadas.

A inversa da ILR é definida por Egozcue e Pawlowsky (2003) como:

$$ilr(x)^{-1} = C(\exp(x\psi)),$$

em que:

$C(.)$ é a operação de fecho, equação (1);

ψ é a matriz que representa as coordenadas.

2.4 ANÁLISE DE REGRESSÃO

A Análise de regressão é um método estatístico que permite examinar a relação entre duas ou mais variáveis, de modo a identificar quais variáveis têm maior impacto diante de um tema de interesse, e de gerar conhecimento sobre este. Neste trabalho, serão abordadas 3 formas distintas de ajustes de regressão.

2.4.1 Regressão Linear Clássica (caso univariado)

A regressão linear é utilizada para investigar e modelar a relação entre variáveis, isto é, moldurar a dependência de uma variável *Y* com respeito à outras variáveis *X*. Esta regressão gera uma equação que descreve a relação estatística entre uma ou mais variáveis preditoras e a variável resposta. A regressão linear encontra a linha que melhor representa as variáveis de entrada com a variável de saída. Tal modelagem é realizada da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i,$$

em que,

$i = 1, \ldots, n$;

$Y_i, i = 1, \ldots, n$, é a i-ésima variável dependente;

$X_{ki}, i = 1, \ldots, p$, é a i-ésima variável independente ou explicativa;

β_0 é o intercepto, e pode ser interpretado como o valor médio de Y_i quando $X_i = 0$;

$\beta_k, k = 1, \ldots, p$, é k-ésimo coeficiente angular, ou seja, o acréscimo esperado na variável resposta, quando nós aumentamos uma unidade em x_i;

$\epsilon_i \sim N(0, \sigma^2)$, em que N indica uma distribuição normal dos dados;

E assumimos que $n > (p + 1)$;

Sua forma matricial pode ser representada por:

$$y = X\beta + \epsilon,$$

em que, para uma amostra de tamanho *n*, tem-se:

$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, vetor de variáveis respostas;

$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}$, matriz de variáveis regressoras;

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$, vetor de parâmetros desconhecidos;

$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$, vetor de componentes aleatórios.

A realização de modelos de regressão visa estimar os parâmetros desconhecidos. Na análise frequentista, o Método dos Mínimos Quadrados e o de Máxima Verossimilhança são os mais utilizados, este último é o mais interessante por permitir uma análise inferencial, como estimação intervalar e testes de hipóteses. Para ambos os casos, o estimador de Mínimos Quadrados (*EMQ)* e o estimador de Máxima Verossimilhança (*EMV)* de β coincidem e são iguais, podendo ser descritos da seguinte forma:

$$\hat{\beta} = (X^T X)^{-1} X^T y, \tag{3}$$

em que a matriz *X* e o vetor *y* foram definidos anteriormente.

2.5 MULTICOLINEARIDADE DE VARIÁVEIS

Os dados composicionais apresentam a restrição de soma constante, ou seja, uma estrutura relacionada que pode influenciar nos resultados procurados, esta estrutura relacionada pode ser interpretada como multicolinearidade. A multicolinearidade consiste em um problema comum em regressões, no qual as variáveis independentes possuem relações lineares entre si Freund (2006), Reund (2006) e Wilson (2006). As principais consequências da multicolinearidade em uma regressão são a de erros-padrão elevados, no caso de multicolinearidade moderada ou severa, e, até mesmo, a impossibilidade de qualquer estimação, se a multicolinearidade for perfeita. Segundo HAIR (2005), a multicolinearidade pode ter sérias influências nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo estimado.

Tal característica pode ser proveniente de quatro fontes (MONTGOMERY; PECK; VINING, 2006):

- Forma de coleta dos dados (Apenas uma subamostra coletada);

- Restrições do modelo ou população (Dados Composicionais);

- Especificações do modelo (Modelo de regressão polinomial);

- Modelo super definido (Dados com mais covariáveis que observações).

A multicolinearidade pode ser detectada a partir da matriz de correlações, se algum par apresentar correlação alta, haverá um forte indício de multicolinearidade, porém, se mais de dois regressores estão envolvidos na regressão linear, então a matriz de correlação não é tão eficiente. Em concordância a (MONTGOMERY; PECK; VINING, 2006), os autovalores da matriz de correlação podem ser utilizados para diagnosticar a multicolinearidade, caso um autovalor seja pequeno em relação aos demais indica um mau condicionamento da matriz.

Além disso, a partir do Fator de Inflação da Variância (*VIF*), também pode-se diagnosticar a dependência entre as variáveis. Supondo que as variáveis estão centradas e padronizadas, tem-se que $R = (X^T X)^{-1}$ e que os elementos diagonais da matriz R são

chamados de fatores de inflação da variância e que representam o incremento da variância devida à presença de multicolinearidade (MONTGOMERY; PECK; VINING, 2006).

O VIF pode ser calculado da seguinte maneira:

$$VIF_i = \frac{1}{1 - R_j^2}, i = 1, 2, \dots, p,$$

em que:

P é quantidade de variáveis preditoras;

R_j^2 é o coeficiente de determinação múltipla resultante da regressão entre X_i e as demais covariáveis.

Um $VIF > 5$ indica forte indício de multicolinearidade, e, consequentemente, os regressores com valores grandes de *VIF* foram, provavelmente, mal estimados.

Ademais, através do número de condição, é possível identificar informações sobre os potenciais empecilhos a serem encontrados em cálculos baseados na matriz $X'X$. Quanto maior tal número, maior será o mau condicionamento da matriz, seu cálculo é dado da seguinte maneira:

$$\eta = \frac{\lambda_{máx}}{\lambda_{mín}},$$

em que $\lambda_1, \lambda_2, \dots, \lambda_n$ são os autovalores da matriz $X'X$.

Geralmente $\eta < 100$ não indica grandes problemas relacionados a multicolinearidade; $100 < \eta < 1000$ indica problemas moderados com multicolinearidade e $\eta > 1000$ indica fortes evidências de multicolinearidade.

Por fim, caso o interesse seja identificar quais variáveis estão envolvidas na multicolinearidade, pode-se calcular o índice de condição:

$$k_m = \frac{\lambda_{máx}}{\lambda_m}, m = 1,2, \dots, p.$$

O número de casos em que $k_m > 100$, dá a ideia de quantas variáveis estão moderadamente dependentes, enquanto o número de casos em que $k_m > 1000$ dá a ideia de quantas variáveis estão linearmente dependentes.

Diversas metodologias têm sido propostas para remediar o problema causado pela multicolinearidade. É possível transformar as variáveis explicativas, coletar informações extras, excluir covariáveis do modelo ou utilizar metodologias alternativas (regressão composicional).

2.5.1 Regressão Ridge

Considerando o modelo de regressão definido na Seção 2.5.1, e, baseando em um mau condicionamento da matriz de variáveis regressoras - que significa existência de correlação das variáveis independentes e que gera multicolinearidade – tem-se que um método alternativo de regressão, para controlar os problemas de multicolinearidade, é a regressão Ridge.

A regressão Ridge é um método de regularização do modelo, que tem como principal objetivo suavizar atributos que sejam relacionados uns aos outros e que aumentam os ruídos, a partir da retirada de determinados atributos do modelo. Para tal retirada de atributos, é realizado um mecanismo de penalização, que adiciona um parâmetro viesado na matriz de correlações e reduz os valores de betas até não nulos. Dessa forma, os atributos que contribuem menos para o poder preditivo do modelo são levados para valores próximos a irrelevância.

São várias as propostas existentes para a escolha do parâmetro viesado λ, apesar de não haver um consenso sobre qual a melhor alternativa. Hoerl e Kennard (1970b) sugeriram selecionar o valor de λ baseado no traço Ridge, um gráfico com as estimativas de $\hat{\beta}^{ridge}$ (no eixo das ordenadas) baseada em λ (no eixo das abscissas), de tal forma que o valor de λ é aquele em que as estimativas começam a convergir para algum valor.

De acordo com Hoerl (1970), a penalização Ridge pode ser escrita como:

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} + X^T y,$$

em que:

X é a matriz de variáveis regressoras padronizadas utilizando a escala de tamanho unitária, nesta padronização, a matriz $X^T X$ geram as correlações entre as covariáveis e das covariáveis com a resposta;

I é a matriz identidade de ordem k.

2.5.2 Regressão Composicional

De forma geral, a regressão linear visa propor e estimar um modelo a partir de dados que dependem linearmente de uma ou mais variáveis. No caso dos dados composicionais, as composições podem surgir como variáveis dependentes ou independentes em modelos lineares. A maioria dos métodos para modelos lineares clássicos tem um análogo próximo para os modelos lineares composicionais.

Segundo Spracklen (2018), existem duas regras para as composições na análise de regressão: as composições podem ser as covariáveis (preditoras) ou as variáveis resposta. De todo modo, o método utilizado para performar a análise de regressão é o mesmo, transforma-se a composição dentro do espaço apropriado e a análise é performada nos dados transformados. Para o caso em que as covariáveis são composicionais, tem-se os seguintes modelos de regressão:

$$Y_i = \beta_0 + \beta(alr(X_i))^T + \epsilon_i, \tag{4}$$

ou

$$Y_i = \beta_0 + \beta(clr(X_i))^T + \epsilon_i, \tag{5}$$

ou

$$Y_i = \beta_0 + \beta(ilr(X_i))^T + \epsilon_i. \tag{6}$$

No caso da *ilr,* as componentes transformadas podem ser transformadas de volta ao *simplex* da seguinte forma:

$$Y_i = \beta_0 + \langle ilr(\beta), ilr(X_i) \rangle + \epsilon_i.$$

As transformações nas covariáveis descritas nas equações de 4 a 6 permitem que as estimações e que os processos inferenciais sejam os mesmos da regressão tradicional linear. Isto é, após a aplicação das transformações *ALR*, *CLR* ou *ILR*, o processo para o ajuste do modelo de regressão composicional é o mesmo utilizado para o ajuste linear, descrito na Equação (3).

2.6 MEDIDAS DE COMPARAÇÃO DE AJUSTE DE MODELOS

Na existência de mais de um modelo ajustado para as mesmas variáveis, faz-se necessária a escolha do modelo que melhor represente o cenário real. Uma das medidas mais utilizadas para tal escolha é o Critério de Informação de Akaike (*AIC*) (AKAIKE, 1974). As principais vantagens do *AIC* residem na possibilidade de comparar modelos com diferentes famílias de distribuições (modelos não aninhados) e por não precisar de mais inferências sobre o modelo para corroborar em seu resultado (BURNHAM; ANDERSON, 2004). Além disso, é de fácil avaliação, pois o melhor modelo é o que possuir o menor valor do *AIC*. Este critério é um dos mais encontrados em diferentes áreas do conhecimento, como regressão espacial, pesquisa operacional, séries temporais, entre outras (BURNHAM; ANDERSON, 2002).

Outra medida muito utilizada é o Critério de Informação Bayesiano (*BIC)*, proposto por Schwarz (1978), e é um critério de avaliação de modelos em termos de probabilidades a posteriori. Tanto o *AIC* quanto o *BIC* se baseiam na verossimilhança, apesar de imporem diferentes penalizações, além disso, o *BIC* é mais punitivo. O melhor modelo, é aquele apresenta os menores valores para ambas as medidas.

Até o presente momento, não se encontrou nenhum indicativo de qual seria a melhor medida para a avaliação de modelos composicionais, isto é, não há informações sobre qual a forma sugerida para medir a qualidade de um modelo que seja composicional. Assim sendo, utilizamos o *AIC* e o *BIC* por serem as métricas mais utilizadas na comparação de modelos.

3 ESTUDO DE SIMULAÇÃO

3.1 SIMULAÇÃO DE DADOS COMPOSICIONAIS

É possível simular composições. A distribuição natural e bem estruturada para os dados no espaço *Simplex* é a Dirichlet, por ser uma distribuição discreta multivariada com um vetor de parâmetros α não negativo e real Aitchison (1896). Sua função de probabilidade é definida da seguinte forma:

$$f(x_1, x_2, \ldots, x_p; \alpha_1, \alpha_2, \ldots, \alpha_p) = \frac{1}{\beta(\alpha)} \sum_{i=1}^{p} x_i^{\alpha_i - 1},$$

em que $x_i \geq 0$, $\sum_{i=1}^{p} x_i = 1$, e $x_i > 0$.

A normalização é uma função beta, que pode ser expressa nos termos da função gama:

$$\beta(\alpha) = \frac{\prod_{i=1}^{p} \gamma(\alpha_i)}{\gamma\left(\sum_{i=1}^{p} \alpha_i\right)}.$$

A distribuição de Dirichlet é a generalização da distribuição beta. Se uma variável $Y = \left(Y_1, Y_2, \ldots, Y_p\right)^T \sim Dirichlet(\alpha))$, tem-se que:

$$E(Y_i) = \frac{\alpha_i}{\sum_{i=1}^{p} \alpha_i}, i = 1, \ldots, p.$$

Outras propriedades desta distribuição podem ser consultadas em Pereira e Stern (2008).

Utilizando a distribuição gama, pode-se simular uma distribuição Dirichlet como demonstrado em Luc (1986). Para simular uma estrutura de dados composicionais a partir de uma Dirichlet, usa-se o seguinte algoritmo:

1. Fixa-se $\alpha = \left(\alpha_1, \alpha_2, \ldots, \alpha_p\right)$;
2. Simulam-se n valores de $y_i \sim$Gama$(\alpha_i, 1)$ com $i = 1,2, \ldots, p$;
3. Calculam-se $x_i = \frac{y_i}{\sum_{i=1}^{p} y_i}$ com $i = 1,2, \ldots, p$.

3.2 SIMULAÇÃO DE MODELOS

Supondo um modelo de Regressão Linear da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \beta_3 X_{3_i} + \epsilon_i,$$

em que $X_i \sim Dirichilet(\alpha)$ e $\epsilon_i \sim N(0,4)$.

Será realizado um estudo de simulação Monte Carlo com 5.000 réplicas e tamanhos amostrais $n = \{10, 20, 40, 80, 160\}$, a fim de verificar o comportamento das estimativas de $\beta_0, \beta_1, \beta_2$ e β_3 através do viés absoluto e do erro quadrático médio. Todas as simulações realizadas no software estatístico R – 2021 e os códigos realizados podem ser consultados no Apêndice B. É importante salientar que neste trabalho as composições foram consideradas apenas no caso de serem covariáveis.

O viés absoluto é a distância entre a média do conjunto de estimativas e o único parâmetro a ser estimado, de modo que quanto menor o viés absoluto, maior a acurácia das estimativas. Para um parâmetro θ, o viés absoluto é:

$$Viés\ Absoluto = \left|E(\hat{\theta}) - \theta\right|,$$

em que $E(\hat{\theta})$ é o valor esperado do estimador.

O erro quadrático médio (*EQM*), por sua vez, é usado para indicar o quão distante, em média, o conjunto de estimativas está do parâmetro a ser estimado, de forma que valores menores de erros quadráticos médios significam menos dispersão das estimativas. Para um parâmetro θ, o erro quadrático médio é:

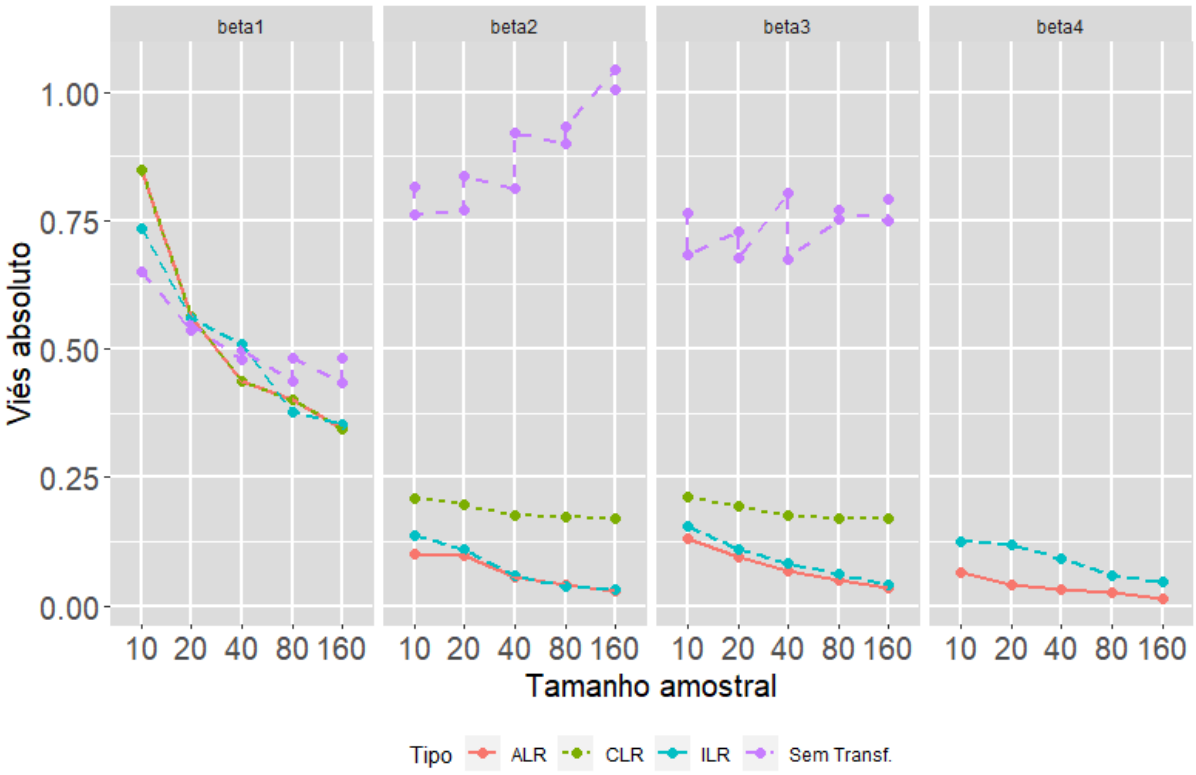
$$EQM(\hat{\theta}) = E\left[\left(\hat{\theta}(X) - \theta\right)^2\right].$$

Os dados provenientes de uma Dirichlet (simulados utilizando o algoritmo descrito em 3.1), são naturalmente composicionais, e, com o objetivo de avaliar o comportamento das estimativas, ajustou-se os modelos com a regressão clássica e com a regressão composicional, considerando as três transformações descritas em 2.5.3. Os valores originais de β são: $1;\ 0{,}33;\ 0{,}33;\ 0{,}34$, e, fora o intercepto, obedecem a restrição de soma 1. Os valores de α_i da Dirichlet podem ser considerados como *pseudocontagens.*

Se todos os α_i são iguais, a distribuição é simétrica. Se $\alpha_i < 1$, pode-se pensar em anti-pesos que empurram x_i para extremos, enquanto valores altos de α_i atraem x_i para algum valor central, no sentido de que todos os pontos estão concentrados em torno deste valor, ou seja, é simetricamente central. Na simulação realizada, não se observou comportamentos distintos dos vieses absolutos e dos erros quadráticos médios a partir de diferentes valores de α_i. Dessa forma, os valores de α_i foram fixados em: 1, 2 e 3.

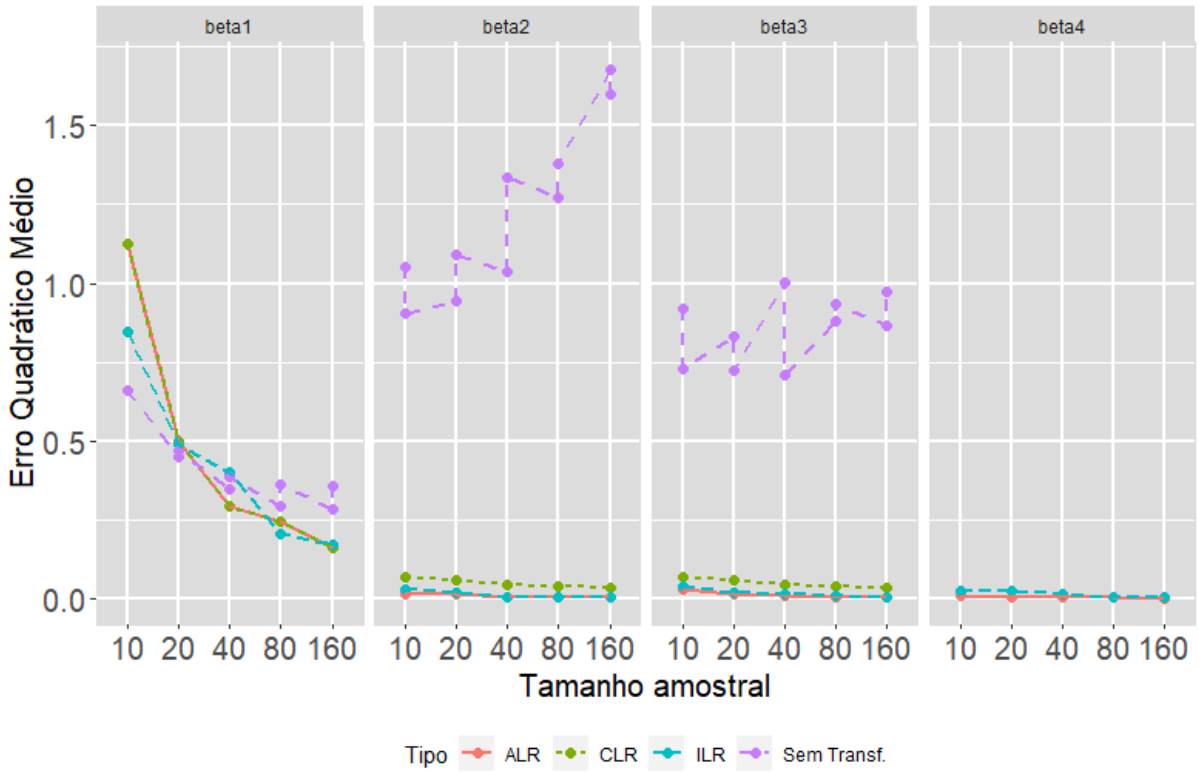
As Figuras 3 e 4, demonstram, respectivamente, o comportamento do viés absoluto e do erro quadrático médio para a simulação da regressão clássica e da regressão composicional, com as transformações *ALR*, *CLR* e *ILR* nas covariáveis. Neste primeiro caso considerou-se o intercepto.

Figura 3 - Viés Absoluto das Estimativas de Beta para Simulação com Intercepto



Fonte: A Autora (2021).

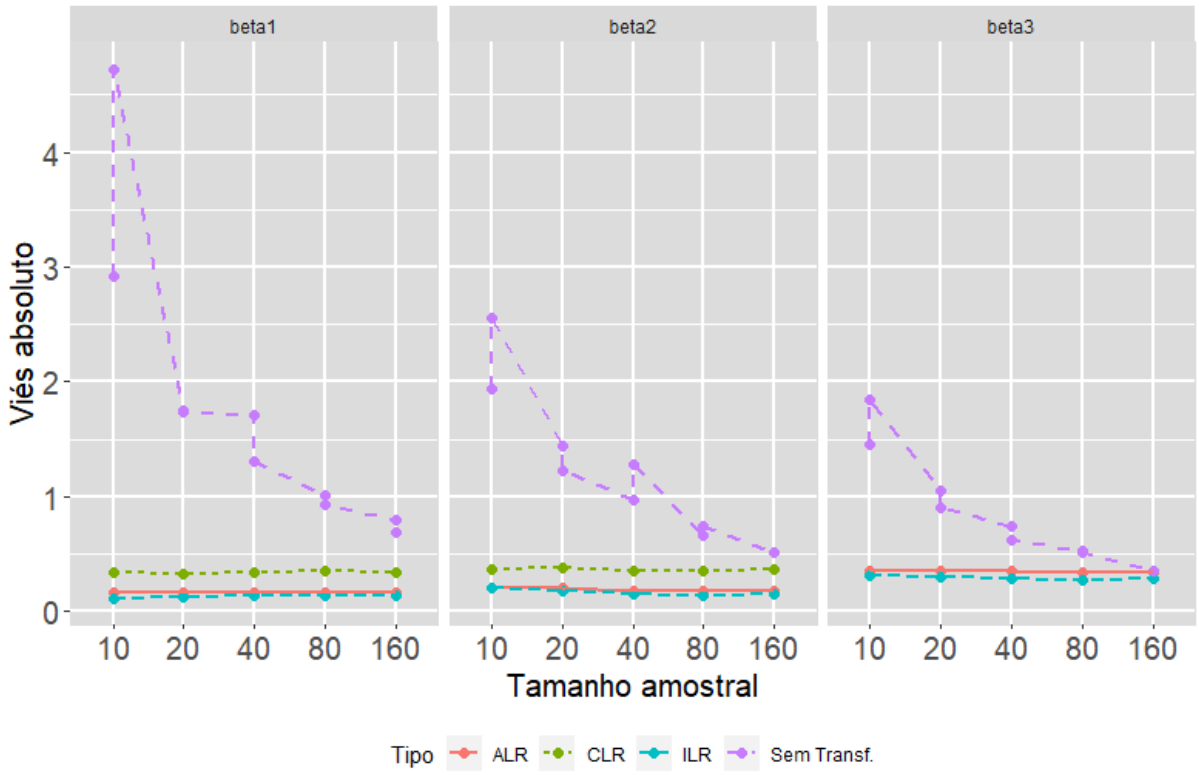
Figura 4 – Erro Quadrático Médio das Estimativas de Beta para Simulação com Intercepto



Fonte: A Autora (2021).

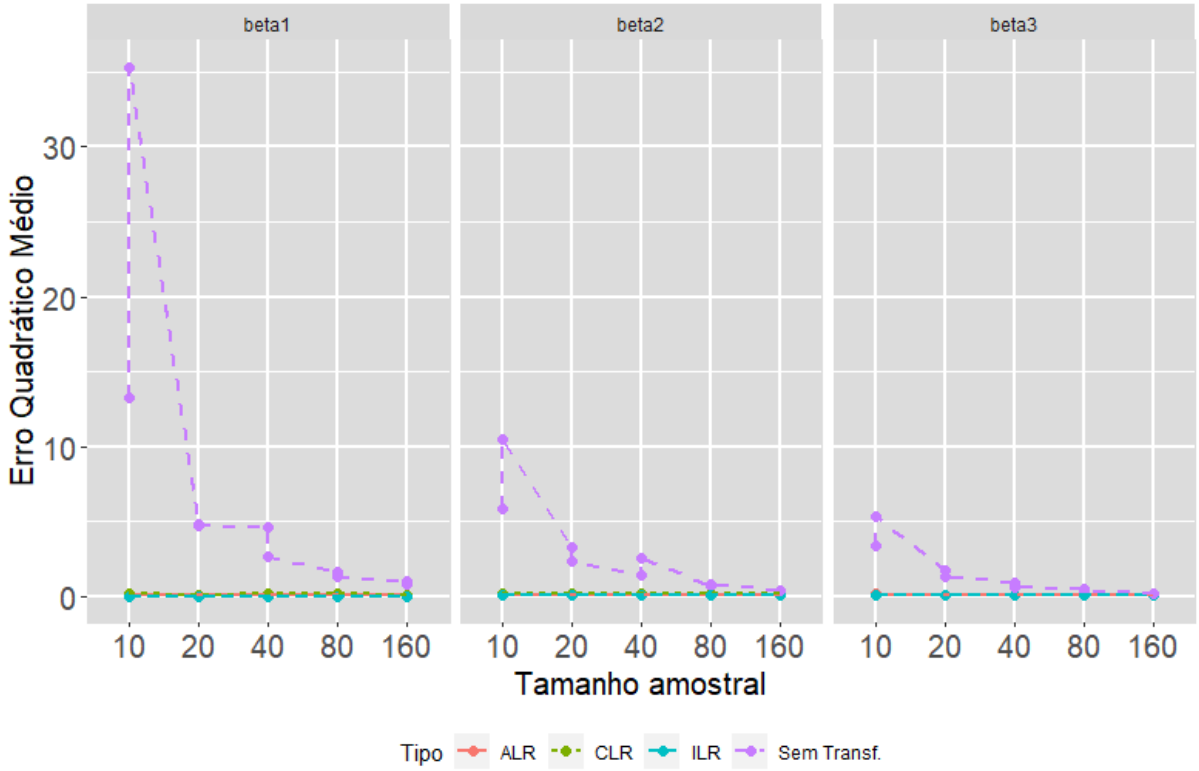
As Figuras 5 e 6, por sua vez, apresentam o comportamento do viés absoluto e do erro quadrático médio para a simulação da regressão clássica e da regressão composicional, com as transformações *ALR*, *CLR* e *ILR,* respectivamente, sem considerar o intercepto.

Figura 5 – Viés Absoluto das Estimativas de Beta para Simulação sem Intercepto



Fonte: A Autora (2021).

Figura 6 – Erro Quadrático Médio das Estimativas de Beta para Simulação sem Intercepto



Fonte: A Autora (2021).

Em todos os casos, observa-se que a regressão composicional, independentemente da transformação utilizada – *ALR*, *CLR* ou *ILR*, e da existência ou não do intercepto no modelo, resulta, tanto em menores valores de vieses absolutos, quanto em menores erros quadráticos médios, quando comparada à regressão clássica - que não considera a natureza dos dados composicionais e não se vale das transformações.

Além disso, quando o ajuste foi realizado sem transformações o último valor do parâmetro Beta não foi estimado, o que pode ser considerado como uma limitação da regressão clássica para dados em composição, observa-se também este comportamento com o uso da transformação *CLR*.

Assim sendo, pelo estudo de simulação performado, pode-se afirmar que, na presença de dados composicionais, é recomendado utilizar a regressão composicional, pelo menos com as transformações *ALR* e *ILR*, visto que estas estimam todos os parâmetros originais e com menores vieses.

4 BASES DE DADOS E ANÁLISES

Para a aplicação da teoria discutida, serão analisadas 3 bases de dados composicionais – uma compilação de análises químicas de sedimento de solo, uma composição dos sedimentos do Lago Stanwell-Fletcher e, finalmente, a proporção de tempo gasto por um estatístico acadêmico durante um dia. Em todos estes casos as composições são as covariáveis.

Todas as análises foram realizadas no software estatístico R – 2021 e um *script* de como foram performadas estas análises podem ser encontradas no Apêndice B.

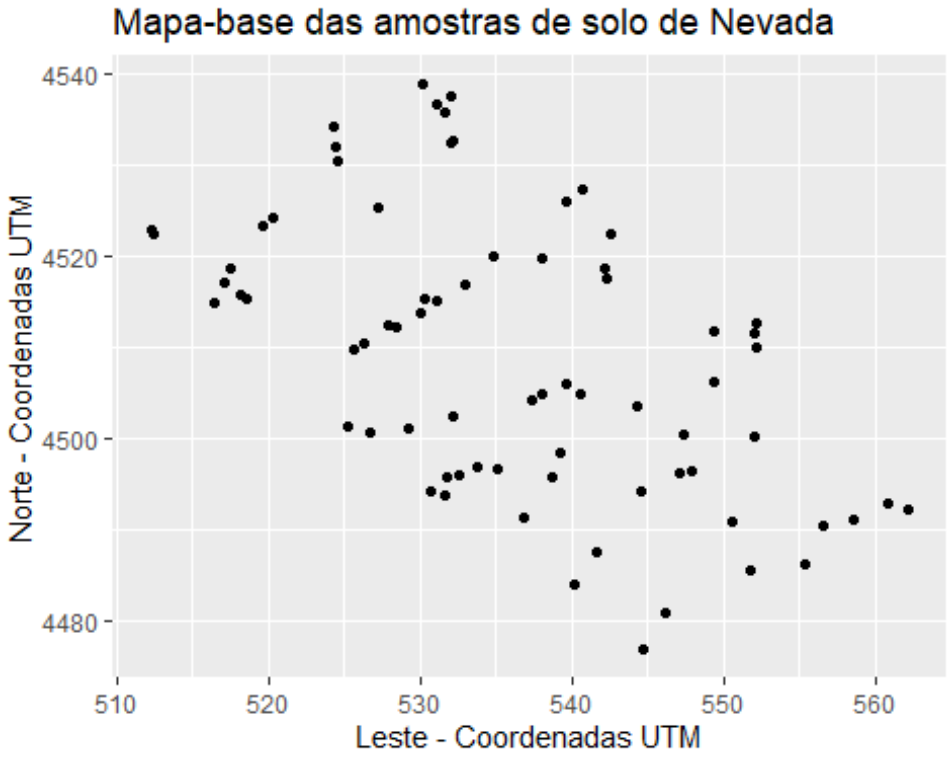
4.1 BASE 1

O primeiro banco de dados é relativo a uma compilação de análises químicas de mais de 10.200 amostras de sedimento e de solo, originalmente coletadas durante o programa de Hidrogeoquímica e Reconhecimento de Sedimentos de Corrente (HSSR) do Departamento de Energia pelo US Geological Survey (USGS). Os locais de amostragem estão no sudeste do Oregon, sudoeste de Idaho, nordeste da Califórnia e, principalmente, no norte de Nevada - EUA. Essas amostras foram coletadas de 1977 a 1983, antes do desenvolvimento da maior parte da infraestrutura de mineração em grande escala atual no norte de Nevada. Os dados foram obtidos do US Geological Survey, Coombs et al. (2002).

Originalmente, o banco possui 52 variáveis de tamanho 75, tais como *ID*, Coordenadas, Latitude e Longitude, além das variáveis da composição do solo: Prata, Alumínio, Arsênio, Ouro, Bário, Berílio, Bismuto, Cálcio, Cádmio, Monóxido de Carbono, Cromo, Cobre, Ferro, Gálio, Potássio, Lítio, Magnésio, Manganês, Molibdênio, Sódio, Nióbio, Fósforo, Chumbo, Antimônio, Estanho, Estrôncio, Tório, Titânio e Zinco.

Uma visualização da localização da retirada dos dados é dada na Figura 7.

Figura 7 - Mapa-Base das Amostras de Solo Colhidas no Nordeste de Nevada.



Fonte: A Autora (2021).

Para análise a ser realizada, elegeu-se usar uma subcomposição do conjunto todo e, pela coerência composicional, Seção 2.1.1, é sabido que os resultados nas subcomposições devem ser coerentes com a composição total. Porém, para que a subcomposição também seja composicional, ou seja, mantenha a restrição de soma constante, faz-se necessária a aplicação da operação de fechamento de dados – Seção 1.3.

Para exemplificação do estudo, apresentar-se-á as observações dos seguintes metais: Arsênio (As), Nióbio (Nb), Antimônio (Sb) e Ouro (Au).

As primeiras 6 observações dos 4 metais do solo podem ser vistas na Tabela 1.

Tabela 1- Descrição dos Metais

As	Nb	Sb	Au
−3	11	1,43	0,001
6	9	2,93	0,006
5	10	1,87	0,003
9	8	2,44	0,004
−3	9	2,22	0,003
19	9	2,04	0,002

Fonte: A Autora (2021).

Com o objetivo futuro de ajustar um modelo de regressão, consideraremos uma subcomposição formada por Arsênio (As), Nióbio (Nb) e Antimônio (Sb), a variável Ouro (Au) será mantida nos valores originais. Para que esta subcomposição conservasse a natureza composicional, realizou-se a operação de fechamento de dados, de tal forma que a Tabela 1 culminou na Tabela 2. É importante salientar que os valores negativos são desconsiderados e que, agora, as linhas possuem a restrição de soma constante em 1.

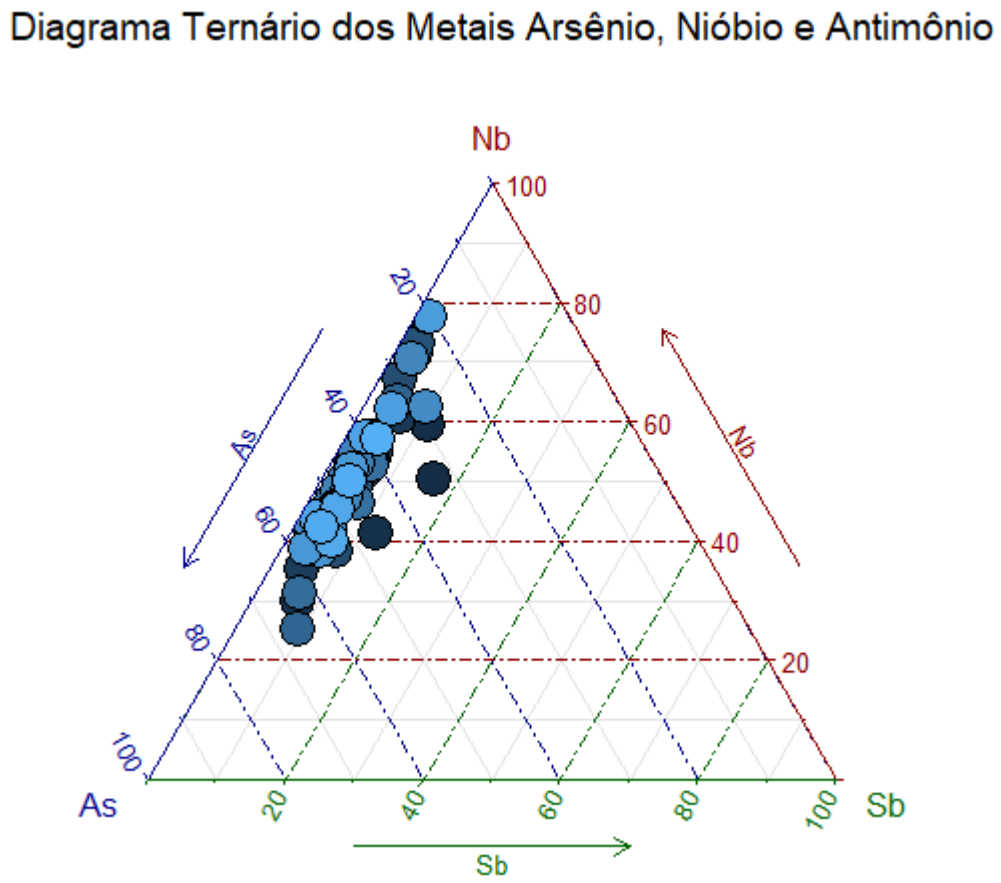
Tabela 2 - Demonstração dos Metais Transformados

As	Nb	Sb
−0,2413	0,8849	0,1150
0,3347	0,5019	0,1634
0,2964	0,5928	0,1108
0,4630	0,4115	0,1255
−0,2674	0,8021	0,1978
0,6325	0,2996	0,0679

Fonte: A Autora (2021).

A fim de observar o comportamento das componentes do solo, as Figuras 8 e 9 foram criadas.

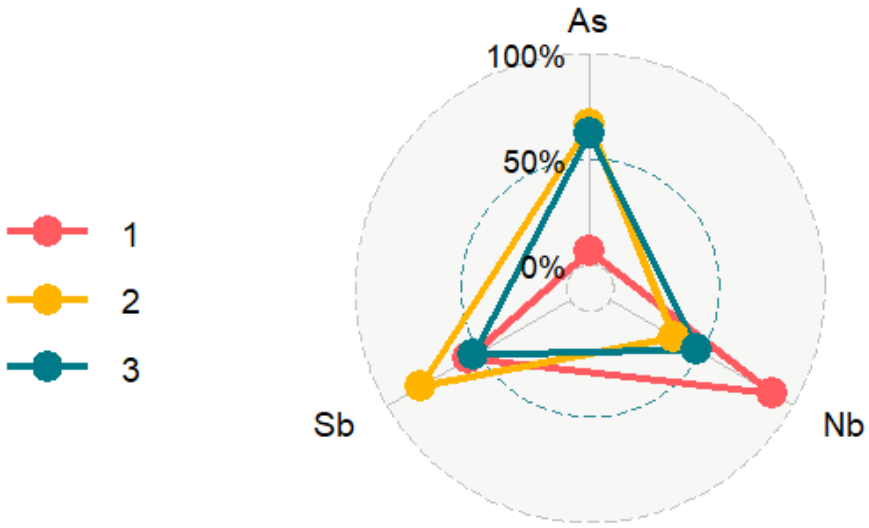
Figura 8 - Diagrama Ternário para os Metais Arsênio, Nióbio e Antimônio.



Fonte: A Autora (2021).

Pelo diagrama ternário, pode-se afirmar que o Antimônio (Sb) não é um elemento com grande proporção na composição, além disso, as componentes Arsênio (As) e Nióbio (Nb) apresentam proporções semelhantes na composição. Pelo gráfico de radar, Figura 9, chega-se à conclusão de que, para a primeira amostra, a proporção de Nióbio (Nb) é maior, já para as duas conseguintes, as proporções das 3 componentes são semelhantes.

Figura 9 - Gráfico de Radar para as Três primeiras Amostras de Arsênio, Nióbio e Antimônio



Fonte: A Autora (2021).

Pela Tabela 3, é possível observar a média composicional dos metais, que demonstra que o metal com maior valor médio é o Nióbio (Nb), seguido pelo Arsênio (As) e, por fim, pelo Antimônio (Sb). Para fins de comparação, a Tabela 4 apresenta a média aritmética (soma dos valores das observações divido pela quantidade das observações) destes metais. Apesar da ordem média ser a mesma, observa-se que existem discrepâncias entre os dois resultados.

Tabela 3 - Média Composicional dos Metais

As	Nb	Sb
0,4496	0,5041	0,0460

Fonte: A Autora (2021).

Tabela 4 - Média Aritmética dos Metais

As	Nb	Sb
0,3794	0,5433	0,0535

Fonte: A Autora (2021).

Por fim, pela Tabela 5, que apresenta a Matriz de Variância dos Metais, pode-se destacar maior correspondência entre as componentes Nióbio (Nb) e Antimônio (Sb), seguido por Nióbio (Nb) e Arsênio (As), e por fim, Antimônio (Sb) e Arsênio (As).

Tabela 5 - Matriz de Variância dos Metais

	As	Nb	Sb
As	0,0000	0,2168	0,2149
Nb	0,2168	0,0000	0,3094
Sb	0,2149	0,3094	0,0000

Fonte: A Autora (2021).

4.1.1 Regressão para a Base 1

Os dados composicionais possuem a restrição de soma constante, tal fato culmina em uma estrutura relacionada, que pode gerar multicolinearidade das variáveis. Com o objetivo de avaliar os ajustes de regressão em dados com restrição, primeiramente será feita a verificação de multicolinearidade dos mesmos. Para o ajuste em questão, optou-se por seguir Braga (2020) e considerar a subcomposição: Arsênio (As), Nióbio (Nb) e Antimônio (Sb), já retratada na Seção 4.1. Desse modo, esta nova composição segue a soma constante em 1 e será a variável explicativa, enquanto o Teor de Ouro será a variável resposta.

O modelo para análise será descrito da seguinte maneira:

$$Au = \beta_0 + \beta_1 As + \beta_2 Nb + \beta_3 Sb + \epsilon$$

Para a verificação da multicolinearidade das variáveis explicativas, foram utilizadas as medidas descritas na Seção 2.4. A Tabela 6 apresenta os resultados respectivos.

Tabela 6 - Medidas de Multicolinearidade das Variáveis Explicativas para a Base 1

	As	Nb	Sb
VIF	34,95	31,18	3,06
Autovalores	2,05	0,93	0,01
K_m	1,00	2,21	138,94

Fonte: A Autora (2021).

Dois valores de VIF são maiores que 5 - indicativos de multicolinearidade, um autovalor é pequeno em relação aos demais, também se configura como indício de relação linear, além disso, o número de condição estimado é de 138,94, e valores de condição maiores que 100 significam que existe pelo menos uma multicolinearidade moderada das variáveis. Devido a isso, pode-se dizer que as variáveis explicativas em questão são, pelo menos, moderadamente multicolineares. Por fim, por k_m (índice de condição), confirma-se a relação de dependência.

Como visto que há relação linear entre as variáveis explicativas, serão realizados os ajustes pela Regressão Clássica, Seção 2.5.1, pela Regressão Ridge, Seção 2.5.2, que geralmente é utilizada para contornar os problemas de multicolinearidade e, posteriormente, pela Regressão Composicional, Seção 2.5.3.

4.1.1.1 Regressão Linear Clássica para a Base 1

O ajuste linear obtido para o modelo descrito na Seção 4.1.1 resultou nas estimativas apresentadas na Tabela 7:

Tabela 7 - Coeficientes Estimados pela Regressão Linear para a Base 1

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	$-0{,}0042$	$0{,}0055$	$-0{,}7590$	$0{,}4506$
As	$0{,}0015$	$0{,}0006$	$2{,}7150$	$0{,}0083\ **$
Nb	$-0{,}0001$	$0{,}0006$	$-0{,}2700$	$0{,}7880$
Sb	$0{,}0228$	$0{,}0057$	$4{,}0170$	$0{,}0001\ ***$

Fonte: A Autora (2021).

As medidas da Tabela 7 podem ser interpretadas da seguinte maneira:

O Erro Padrão representa a distância média em que os valores observados caem da linha de regressão, pode-se dizer que menores valores nos erros padrão significam um melhor ajuste, pois indicam que as observações estão mais próximas da linha ajustada.

O Valor T, por sua vez, é utilizado para determinar se um coeficiente é significativo. Sob suposições do modelo de regressão, esta estatística t segue uma distribuição t-student com $(n-p)$ graus de liberdade:

$$t = \frac{\hat{\beta}_m - \beta_m}{\sqrt{QMRes\ c_{mm}}},$$

em que:

$m = 1,2, \dots, p$ e

c_{mm}` é o (m, m')-ésimo elemento da matriz $C = (X^T X)^{-1}$

e $QMRes$ é o quadrado médio do resíduo.

Haverá significância do parâmetro se, para um nível de significância (α) definido pelo pesquisador, $|t|$ for maior que o quantil $100(1-\alpha/2)\%$ de uma $t(n-p)$.

Por fim, o Valor P testa, para cada termo, a hipótese de que os coeficientes são iguais a zero, ou seja, sem efeito. Um valor P menor que α indica significância da variável preditora, isto é, as alterações na variável preditora são relacionadas a alterações na variável resposta.

Desse modo, supondo nível de significância de 5%, e considerando o Valor P, observa-se que a variável Arsênio (As) e Antimônio (Sb) são estatisticamente significantes para explicar o teor de Ouro (Au).

Além disso, para a Regressão Linear Clássica, encontrou-se as seguintes medidas de qualidade de ajuste (Tabela 8):

Tabela 8 - Medidas de Qualidade do Ajuste com Regressão Clássica para a Base 1

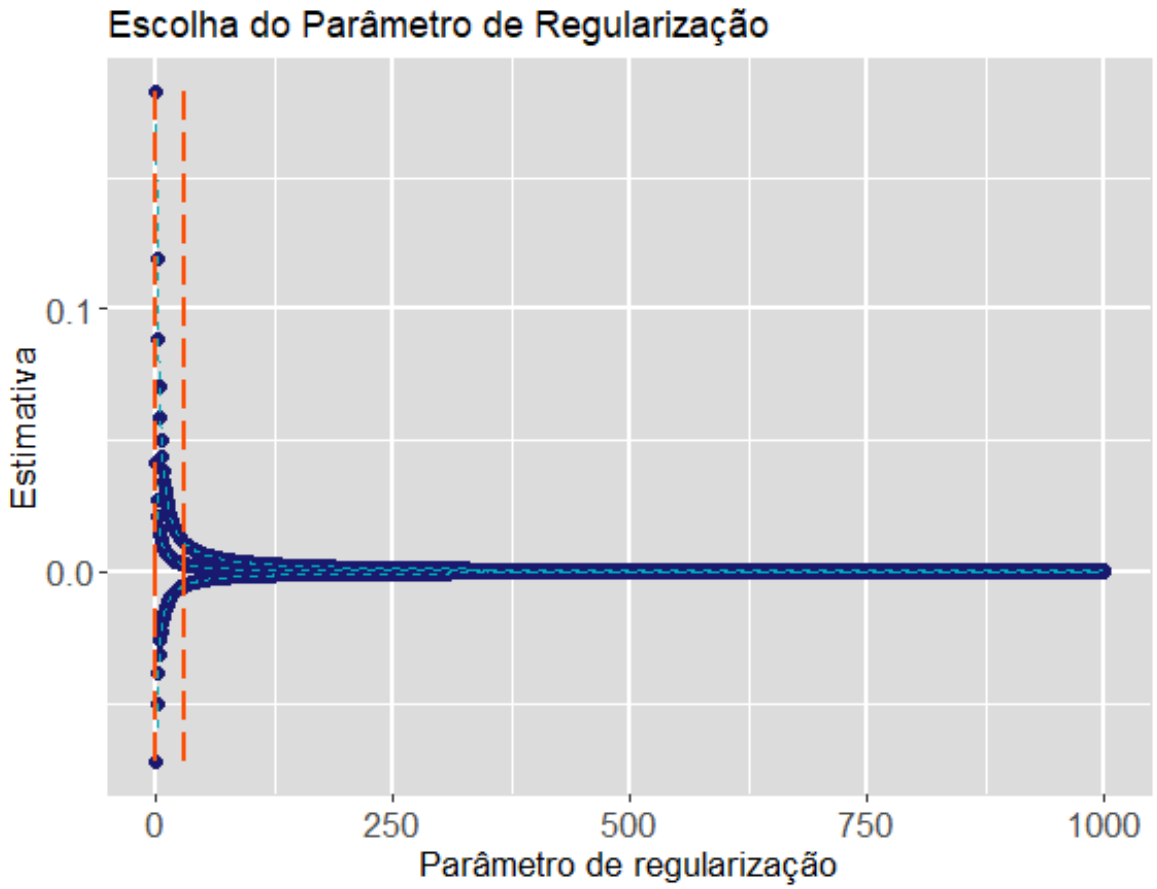
AIC	BIC
$-750,44$	$-738,85$

Fonte: A Autora (2021).

4.1.1.2 Regressão Ridge para a Base 1

A Regressão Ridge é conhecida por controlar os problemas de multicolinearidade, assim sendo, ajustou-se o modelo definido na Seçã0 4.1.1 com tal regressão. Primeiramente, pela Figura 10, observa-se que o parâmetro de regularização λ é aproximadamente 30, pois é a partir desse valor de λ que as estimativas começam a se estabilizar.

Figura 10 - Escolha do Parâmetro de Regularização para a Base 1



Fonte: A Autora (2021).

O ajuste da regressão Ridge, com parâmetro de regularização $\lambda = 30$, resultou nas estimativas demonstradas na Tabela 9.

Tabela 9 - Estimativas pela Regressão Ridge para a Base 1

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	0,0012	NA	NA	NA
As	0,0004	0,0001	1,3270	0,1843
Nb	$-0{,}0009$	0,0001	1,7420	0,0816
Sb	0,0148	0,0001	3,5240	$0{,}0004\ **$

Fonte: A Autora (2021).

Pode-se observar, a 5% de significância, que somente o Antimônio (Sb) é estatisticamente significativo para explicar o Teor de Ouro (Au). Com o ajuste realizado pela regressão Ridge, não foi possível calcular as medidas de comparação *AIC* e *BIC*, desse modo, considerou-se apenas as estimativas do modelo para futuras comparações.

4.1.1.3 Regressão Composicional para a Base 1

Com o objetivo de avaliar a transformação composicional, e, sabendo que os componentes de solo são composicionais, realizou-se o ajuste da Regressão Composicional com aplicação das três transformações enunciadas na Seção 2.3.

4.1.1.3.1 Regressão Composicional com Transformação *ALR* para a Base 1

Primeiramente, aplicou-se a Transformação *ALR* nas variáveis explicativas, visto que através desta transformação é possível sair do espaço *simplex* e ir para o espaço dos números reais. Porém, como mencionado em 2.3.1, divisões por zero podem gerar valores infinitos, o que, de fato, ocorreu.

Desse modo, para o ajuste em questão, a transformação *ALR* não foi efetiva para explicar o Teor de Ouro (Au), pela composição de Arsênio (As), Nióbio (Nb) e Antimônio (Sb).

4.1.1.3.2 Regressão Composicional com Transformação *CLR* para a Base 1

A segunda transformação descrita é a *CLR*, que permite que a composição se associe a um vetor muldimensional. Assim sendo, aplicando esta transformação nas variáveis explicativas do modelo, encontrou-se as estimativas da Tabela 10, destacando-se que as duas

primeiras variáveis transformadas são estatisticamente significantes para explicar o Teor de Ouro (Au), a um nível de 5% de significância e que o último parâmetro não foi estimado, o que também havia ocorrido na simulação realizada na Seção 3. A sigla *NA* se refere a valores faltantes.

Tabela 10 - Coeficientes Estimados pela Regressão Composicional com *CLR* no Espaço *Simplex* para a Base 1

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	0,0046	0,0009	5,0480	$3{,}24X10^{-6}$ ***
*CLR(X)*1	−0,0014	0,0004	3,0650	0,0030 **
*CLR(X)*2	−0,0023	0,0003	2,1290	0,0366 *
*CLR(X)*3	*NA*	*NA*	*NA*	*NA*

Fonte: A Autora (2021).

Com a aplicação da inversa da *CLR*, pode-se exibir os coeficientes no espaço real – Tabela 11.

Tabela 11 - Coeficientes Estimados pela Regressão Composicional com *CLR* no espaço real para a Base 1.

Coeficientes	As	Nb	Sb
Valores Estimados	0.5001	0.4998	*NA*

Fonte: A Autora (2021).

Por fim, para a possível comparação desse ajuste com os demais, a Tabela 12 apresenta as medidas de qualidade do ajuste.

Tabela 12 - Medidas de Qualidade do Ajuste para a Regressão Composicional com *CLR* para a Base 1

AIC	*BIC*
$-749{,}55$	$-740{,}27$

Fonte: A Autora (2021).

4.1.1.3.3 Regressão Composicional com Transformação *ILR* para a Base 1

Por fim, aplicou-se a Transformação *ILR* nas variáveis explicativas, visto que esta transformação permite que a regressão composicional possa ser resolvida como uma regressão múltipla.

Dessa forma, para o mesmo modelo de regressão, encontrou-se as estimativas para os coeficientes na Tabela 13; pode-se observar que além do intercepto, somente uma variável transformada é estatisticamente significante para explicar o Teor de Ouro (Au), a um nível de 5% de significância. É importante salientar que a aplicação da *ILR* nas variáveis explicativas converte os dados para o espaço *simplex.*

Tabela 13 - Coeficientes Estimados pela Regressão Composicional com *ILR* no Espaço *Simplex* para a Base 1

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	0,0046	0,0009	5,0480	$3{,}24X10^{-6}$ ***
*ILR(X)*1	$-0{,}0006$	0,0004	$-1{,}3300$	0,1878
*ILR(X)*2	0,0015	0,0005	3,1320	0,0025 **

Fonte: A Autora (2021).

Para que os coeficientes retornassem ao espaço euclidiano, aplicou-se a inversa da *ILR*, assim sendo, a Tabela 14 apresenta as estimativas dos coeficientes estimados no espaço real.

Tabela 14 - Coeficientes Estimados pela Regressão Composicional com *ILR* no espaço real para a Base 1

Coeficientes	As	Nb	Sb
Valores Estimados	0,3333	0,3330	0,3337

Fonte: A Autora (2021).

Por fim, para a verificação de qualidade do ajuste, a Tabela 15 apresenta as medidas descritas na Seção 2.6.

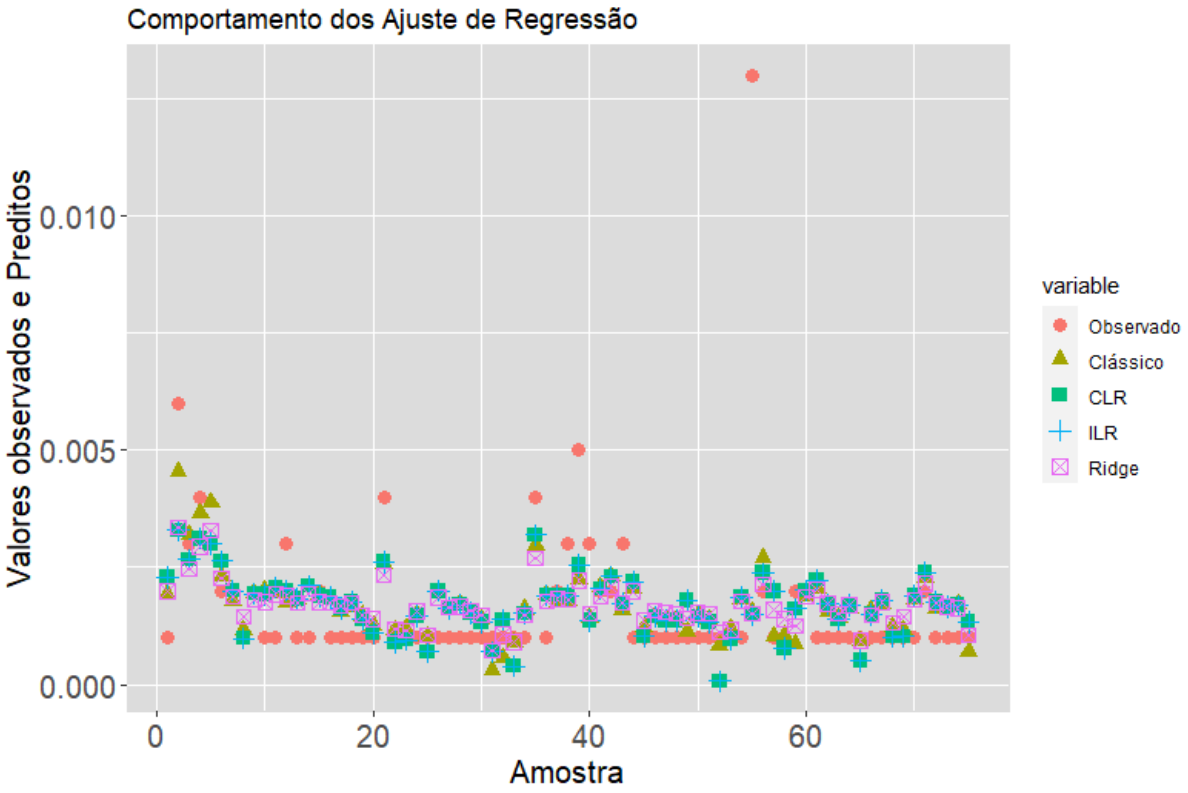
Tabela 15 - Medidas de Qualidade do Ajuste para a Regressão Composicional com *ILR* para a Base 1

AIC	*BIC*
−749,55	−740,27

Fonte: A Autora (2021).

Com o propósito final de comparação do comportamento dos ajustes aplicados, a Figura 11 foi plotada. Esta apresenta os valores originais e os valores preditos para cada ajuste performado. A partir dela, é possível dizer que a Regressão Clássica e a Regressão Composicional com aplicação da *CLR* e da *ILR* (mesmos valores previstos), foram as que apresentaram predições mais próximas dos valores observados do Teor de Ouro, enquanto a Regressão Ridge subestimou estes valores.

Figura 11 – Comportamento dos Ajustes de Regressão Performados para a Base 1



Fonte: A Autora (2021).

4.2 BASE 2

O segundo banco de dados utilizado é relativo a uma composição formada por três componentes na presença de uma covariável, apresentado primeiramente por Coakley e Rust (1968) e posteriormente revisado por Aitchison (1986). Se trata da composição dos sedimentos do Lago Stanwell-Fletcher, localizado na Ilha Somerset, no arquipélago Ártico do Canadá. Esse conjunto, na literatura, é mais conhecido como Lago Ártico, e é composto por partes de Areia, Lodo e Argila para 39 profundidades, em metros, do lago. As primeiras 6 linhas do conjunto de dados são apresentadas na Tabela 16, é importante destacar que a soma dos sedimentos é constante em 1.

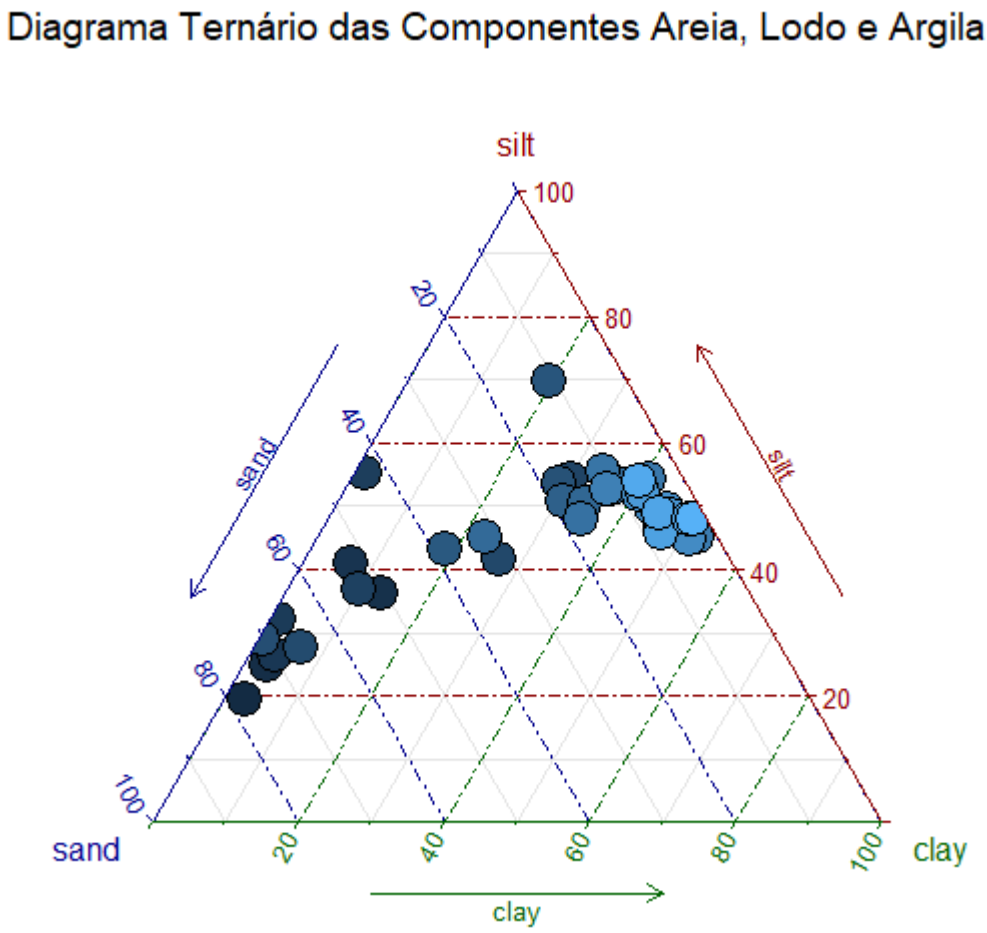
Tabela 16 - Demonstração das Componentes do Lago Ártico.

Areia	Lodo	Argila	Profundidade
0,775	0,195	0,030	10,4
0,719	0,249	0,032	11,7
0,507	0,361	0,132	12,8
0,522	0,409	0,066	13,0
0,700	0,265	0,035	15,7
0,665	0,322	0,013	16,3

Fonte: A Autora (2021).

Para obter as informações descritivas sobre o lago Ártico, foi plotado o Diagrama ternário das 39 amostras das componentes: Areia (*sand*), Lodo (*silt*) e Argila (*clay*), conforme demonstrado na Figura 12.

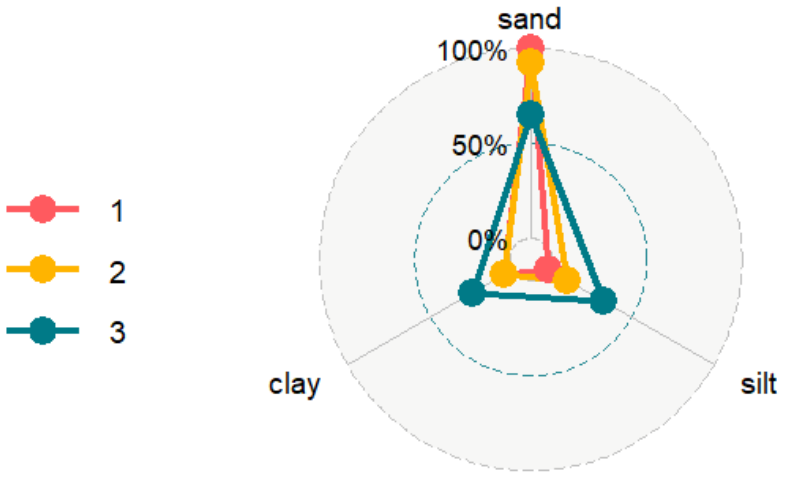
Figura 12 - Diagrama Ternário das Componentes Areia Lodo e Argila.



Fonte: A Autora (2021).

Pelo diagrama ternário, pode-se afirmar que as amostras estão distribuídas aproximadamente nos entornos do baricentro, o que culmina na conclusão de que as componentes possuem proporções aproximadas. Pelo gráfico de radar, Figura 13, que representa o comportamento das 3 primeiras amostras das componentes, chega-se à conclusão de que o teor de Areia (*sand*) é mais presente na composição, além disso, para as três primeiras amostras, as componentes Argila (*clay*) e Lodo (*silt*) não são tão presentes.

Figura 13 - Gráfico de Radar das Três Primeiras Amostras do Lago Ártico



Fonte: A Autora (2021).

Pela Tabela 17, observa-se que a maior média composicional é referente à componente Lodo (*silt*), seguido pela Argila (*clay*) e, por fim, pela Areia (*sand*). Pela Tabela 18, observa-se que a média aritmética resulta em distintos valores médios.

Tabela 17 - Média Composicional das Componentes do Lago Ártico

Areia	Lodo	Argila
0,1779	0,5637	0,2582

Fonte: A Autora (2021).

Tabela 18 - Média Aritmética das Componentes do Lago Ártico

Areia	Lodo	Argila
0,2420	0,4569	0,3012

Fonte: A Autora (2021).

Por fim, a Tabela 19 que apresenta a Matriz de Covariâncias das Componentes do Lago Ártico, permite destacar maior correspondência entre as variáveis Areia e Argila, seguida por Areia e Lodo, e uma menor correspondência entre Argila e Lodo.

Tabela 19 - Matriz de Variância dos Componentes do Lago Ártico

	Areia	Lodo	Argila
Areia	0,0000	1,6412	4,7315
Lodo	1,6412	0,0000	1,0350
Argila	4,7315	1,0350	0,0000

Fonte: A Autora (2021).

4.2.1 Regressão para a Base 2

Os dados composicionais da composição do Lago Ártico possuem a restrição de soma constante, que, consequentemente, culmina em uma estrutura relacionada que pode gerar multicolinearidade das variáveis. Assim sendo, será feita a verificação de multicolinearidade dos mesmos. Para o ajuste em questão, o modelo de interesse é descrito da seguinte maneira:

$$Profundidade = \beta_0 + \beta_1 Areia + \beta_2 Silte + \beta_3 Argila + \epsilon$$

Para a verificação da multicolinearidade das variáveis explicativas, serão utilizadas as medidas descritas na Seção 2.6. A Tabela 20 apresenta os resultados respectivos.

Tabela 20 - Medidas de Multicolinearidade das Variáveis Explicativas para a Base 2

	Areia	Lodo	Argila
VIF	6.3273,0692	$1{,}0747X10^{4}$	$3{,}0893X10^{4}$
Autovalores	2,5846	$4{,}1538X10^{-1}$	$9{,}5319X10^{-6}$
K_m	1,0000	6,2222	$2{,}7115X10^{5}$

Fonte: A Autora (2021).

Todos os valores de *VIF* são maiores que 5, indicando multicolinearidade e, como um autovalor é pequeno em relação aos demais, também há indício de relação linear. Além disso, o número de condição estimado é de $2{,}71X10^5$, e, valores de condição maiores que 1000 significam que existe forte multicolinearidade das variáveis. Devido a isso, pode-se dizer que as variáveis explicativas em questão são multicolineares. Por fim, por k_m, observa-se a relação dependente das variáveis.

Como visto que há uma relação linear entre as variáveis explicativas, serão realizados os ajustes pela Regressão Clássica, pela Regressão Ridge e pela Regressão Composicional.

4.2.1.1 Regressão Linear Clássica para a Base 2

O ajuste linear obtido para o modelo descrito na Seção 4.2.1 resultou nas estimativas apresentadas na Tabela 21:

Tabela 21- Coeficientes Estimados pela Regressão Linear para a Base 2

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	2.759,60	2.603,89	1,06	0,29
Areia	−27,46	26,05	−1,05	0,30
Lodo	−27,66	26,03	−1,06	0,29
Argila	−26,00	26,04	−0,99	0,32

Fonte: A Autora (2021).

Supondo um nível de significância de 5%, e considerando o Valor P, observa-se que, além do intercepto, nenhuma variável é estaticamente significante para explicar a profundidade do lago.

Além disso, para a Regressão Linear Clássica, encontrou-se as seguintes medidas de qualidade de ajuste (Tabela 22):

Tabela 22 - Medidas de Qualidade de Ajuste com Regressão Clássica para a Base 2

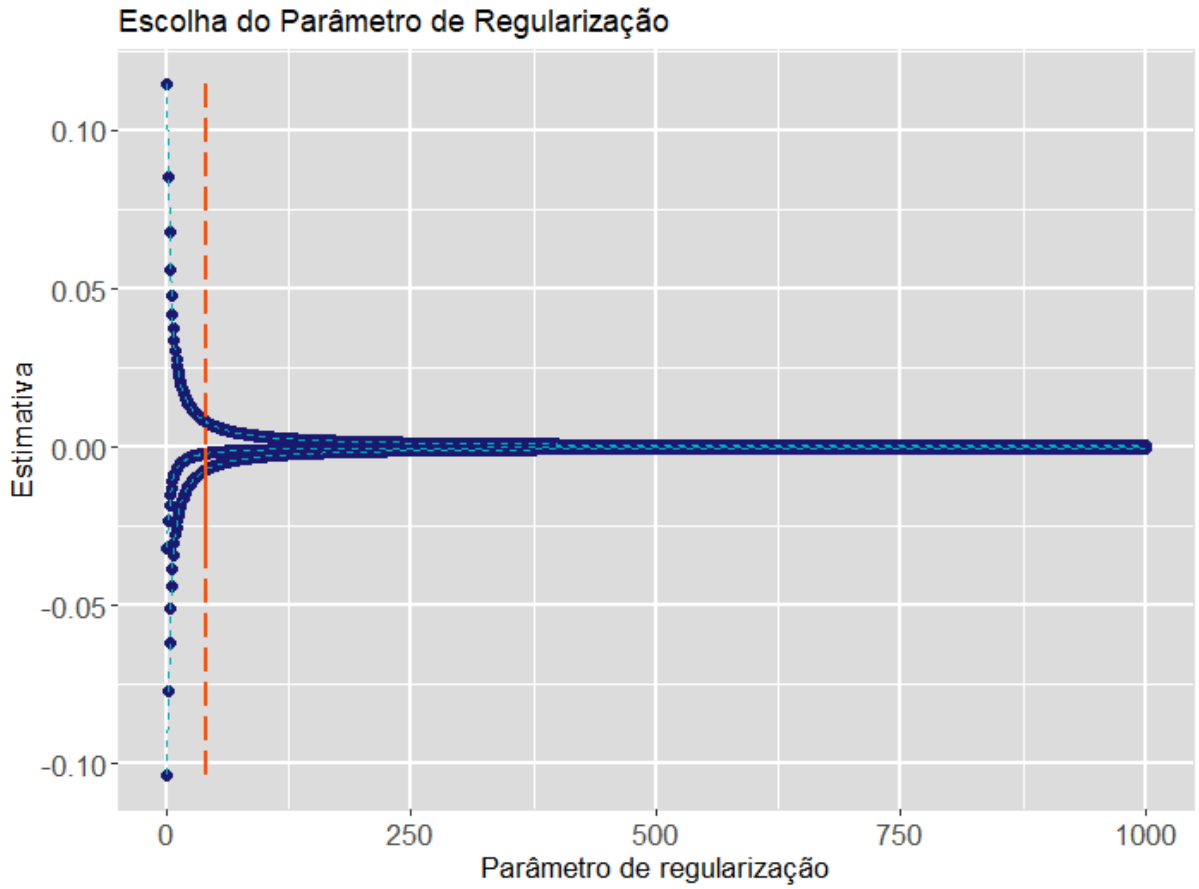
AIC	*BIC*
331,08	339,39

Fonte: A Autora (2021).

4.2.1.2 Regressão Ridge para a Base 2

A Regressão Ridge controla problemas de multicolinearidade, assim sendo, o modelo definido na Seção 4.2.1 foi ajustado com esta metodologia. Para a definição do parâmetro de regularização, pode-se observar, pela Figura 14, que as estimativas começam a se estabilizar em aproximadamente $\lambda = 40$, desse modo, este foi o parâmetro de regularização utilizado.

Figura 14 - Escolha do Parâmetro de Regularização para a Base 2



Fonte: A Autora (2021).

O ajuste da regressão Ridge, com parâmetro de regularização $\lambda = 40$, resultou nas estimativas demonstradas na Tabela 23.

Tabela 23 - Estimativas pela Regressão Ridge para a Base 2

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	34,4687	NA	NA	NA
Areia	$-0{,}2193$	0,7595	7,0860	$1{,}38X10^{-12}$ ***
Lodo	0,1526	1,0243	1,5080	0,132
Argila	0,3954	0,9317	7,2800	$3{,}34X10^{-13}$ ***

Fonte: A Autora (2021).

Pode-se concluir, com o ajuste da regressão Ridge, supondo nível de significância de 5%, e considerando o Valor P, que apenas o intercepto e o lodo não são estatisticamente significantes para explicar a Profundidade do Lago. Com o ajuste realizado pela regressão Ridge não foi possível calcular as medidas de comparação *AIC* e *BIC*, desse modo, considerou-se apenas as estimativas do modelo para futuras comparações.

4.2.1.3 Regressão Composicional para a Base 2

Novamente, com o objetivo de avaliar as alternativas composicionais, e, valendo-se de que as componentes do Lago Ártico possuem restrição de soma constante, realizou-se o ajuste da Regressão Composicional. Para tal ajuste, aplicou-se as Transformações *ALR, CLR* e *ILR* nas variáveis explicativas – Areia, Lodo e Argila.

4.2.1.3.1 Regressão Composicional com Transformação *ALR* para a Base 2

Aplicou-se a Transformação *ALR* nas variáveis explicativas para que estas tenham correspondência com o espaço dos números reais. A Tabela 24 apresenta as estimativas encontradas. Observa-se que tanto o intercepto, quanto as duas variáveis transformadas são estatisticamente significantes, considerando um nível de significância de 5%, para explicar a profundidade do lago.

Tabela 24 - Coeficientes Estimados pela Regressão Composicional com *ALR* no Espaço *Simplex* para a Base 2

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	30,785	7,285	4,226	$0{,}0001\ ***$
*ALR(X)*1	$-16{,}553$	3,365	$-4{,}920$	$1{,}92X10^{-5}\ ***$
*ALR(X)*2	14,208	7,194	1,975	$0{,}0559\ *$

Fonte: A Autora (2021).

Para que os coeficientes retornassem ao espaço euclidiano, aplicou-se a inversa da *ALR*, assim sendo, a Tabela 25 apresenta as estimativas dos coeficientes estimados no espaço real.

Tabela 25 - Coeficientes Estimados pela Regressão Composicional com *ALR* no Espaço Real para a Base 2

Coeficientes	Areia	Lodo	Argila
Valores Estimados	$4{,}37X10^{-14}$	1,0000	$6{,}75X10^{-7}$

Fonte: A Autora (2021).

Por fim, a fim de realizar comparações com os outros modelos ajustados, a Tabela 26 apresenta as medidas descritas em Seção 2.6.

Tabela 26 - Medidas de Qualidade do Ajuste para a Regressão Composicional com *ALR* para a Base 2

AIC	*BIC*
333,48	340,13

Fonte: A Autora (2021).

4.2.1.3.2 Regressão Composicional com Transformação *CLR* para a Base 2

A segunda transformação descrita é a *CLR*, que permite que a composição se associe a um vetor dimensional. A aplicação desta transformação nas variáveis explicativas do modelo descrito na Seção 4.2.1, resultou nas estimativas da Tabela 27. Observa-se que o intercepto e a primeira variável transformada são estatisticamente significantes para explicar a profundidade do Lago a um nível de 5 % de significância. Porém, novamente, o último parâmetro não foi estimado, o que ocorreu na simulação realizada na Seção 3 e no ajuste do modelo da Base 1, Seção 4.1.4.2.

Tabela 27 - Coeficientes Estimados pela Regressão Composicional com *CLR* no Espaço *Simplex* para a Base 2

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	30,785	7,285	4,226	$0{,}0001$ ***
*CLR(X)*1	$-18{,}898$	2,606	$-7{,}251$	$1{,}54X10^{-8}$ ***
*CLR(X)*2	11,864	11,318	1,048	0,3015
*CLR(X)*3	*NA*	*NA*	*NA*	*NA*

Fonte: A Autora (2021).

A Tabela 28 apresenta os coeficientes no espaço real, após a aplicação da inversa da *CLR*.

Tabela 28 - Coeficientes Estimados pela Regressão Composicional com *CLR* no Espaço Real para a Base 2

Coeficientes	Areia	Lodo	Argila
Valores Estimados	$4{,}36X10^{-14}$	1,0000	*NA*

Fonte: A Autora (2021).

Por fim, a Tabela 29 explicita a medidas de comparação de ajuste.

Tabela 29 - Medidas de Comparação de Qualidade do Ajuste para a Regressão Composicional com *ALR* para a Base 2

AIC	*BIC*
333,48	340,13

Fonte: A Autora (2021).

4.2.1.3.3 Regressão Composicional com Transformação *ILR* para a Base 2

Por fim, para o mesmo modelo, o último ajuste performado foi o da regressão composicional com transformação *ILR*. Desse modo, encontrou-se as estimativas para os coeficientes na Tabela 30, observa-se que todas as variáveis explicativas são estatisticamente significantes para explicar a profundidade do lago, a um nível de 5% de significância. Salienta-se que a aplicação da *ILR* nas variáveis explicativas converte os dados para o espaço *simplex.*

Tabela 30 - Coeficientes Estimados pela Regressão Composicional com *ILR* no Espaço *Simplex* para a Base 2

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	30,7850	7,2850	4,2260	0,0001 ***
$ILR(X)1$	21,7520	7,3560	2,9570	0,0054 **
$ILR(X)2$	2,8720	5,1890	0,5530	0,5834

Fonte: A Autora (2021).

Para uma representação dos coeficientes no espaço euclidiano, aplicou-se a inversa da *ILR*, os resultados são demonstrados na Tabela 31.

Tabela 31 - Coeficientes Estimados pela Regressão Composicional com *ILR* no Espaço Real para a Base 2

Coeficientes	Areia	Lodo	Argila
Valores Estimados	$4{,}36X10^{-14}$	1,0000	$7{,}03X10^{-6}$

Fonte: A Autora (2021).

Por fim, para a verificação de qualidade do ajuste, a Tabela 32 apresenta as medidas descritas na Seção 2.6.

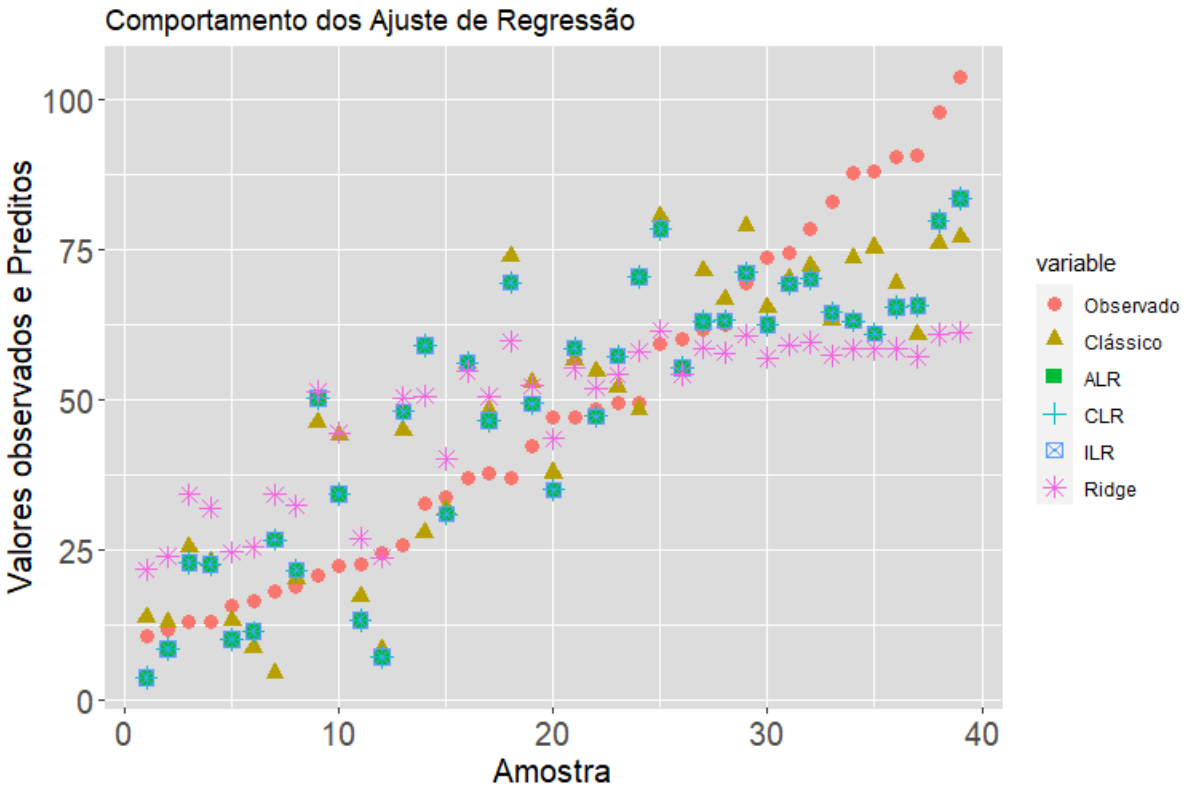
Tabela 32 - Medidas de Comparação de Qualidade do Ajuste para a Regressão Composicional com *ILR* para a Base 2

AIC	*BIC*
333,48	340,13

Fonte: A Autora (2021).

Novamente, com o propósito final de comparação do comportamento dos ajustes aplicados, a Figura 15 foi plotada. Esta apresenta os valores originais e os valores preditos para cada ajuste performado. A partir dela, é possível dizer que a Regressão Clássica e a Regressão Composicional com aplicação da *ARL*, *CLR* e da *ILR* (mesmos valores previstos), foram as que apresentaram predições mais próximas dos valores observados de Profundidade do Lago Ártico, enquanto a Regressão Ridge, novamente, subestimou os valores observados.

Figura 15 – Comportamento dos Ajustes de Regressão Performados para a Base 2



Fonte: A Autora (2021).

4.3 BASE 3

O terceiro banco de dados utilizado é relativo a um problema de orçamento de tempo, mais especificamente, como um dia ou período de trabalho é dividido em diferentes atividades. Para ilustrar tal problema, retirou-se de Aitchison (1986) um conjunto de dados relativos a seis atividades diárias de um estatístico acadêmico: ensino, consulta, administração, pesquisa, outras atividades de vigília e sono. As proporções das 24 horas dedicadas a cada atividade foram registradas em 20 dias, selecionadas aleatoriamente de dias úteis em semanas alternadas, de modo a evitar possíveis efeitos colaterais, como um dia de sono curto sendo compensado por sono de reposição no dia seguinte. As seis atividades podem ser divididas em duas categorias: Trabalho – ensino, consulta, administração e pesquisa; e Lazer – outras atividades e sono. As primeiras 6 linhas do conjunto de dados são apresentadas na Tabela 33, é importante destacar que a soma das proporções das atividades é constante em 1, salvo erros de arredondamento.

Tabela 33 - Demonstração das Proporções de Tempo Gasto por um Estatístico Acadêmico

Ensino	Consulta	Administração	Pesquisa	Outras	Sono
0,144	0,091	0,179	0,107	0,263	0,217
0,162	0,079	0,107	0,132	0,265	0,254
0,153	0,101	0,131	0,138	0,209	0,267
0,177	0,087	0,140	0,132	0,155	0,310
0,158	0,110	0,139	0,116	0,258	0,219
0,165	0,079	0,113	0,113	0,275	0,255

Fonte: A Autora (2021).

Com o objetivo futuro de ajustar um modelo de regressão, considerar-se-á uma subcomposição formada por Ensino, Administração e Pesquisa. Para que esta subcomposição mantivesse a natureza composicional, realizou-se a operação de fechamento de dados, de tal forma que a Tabela 33 culminou na Tabela 34.

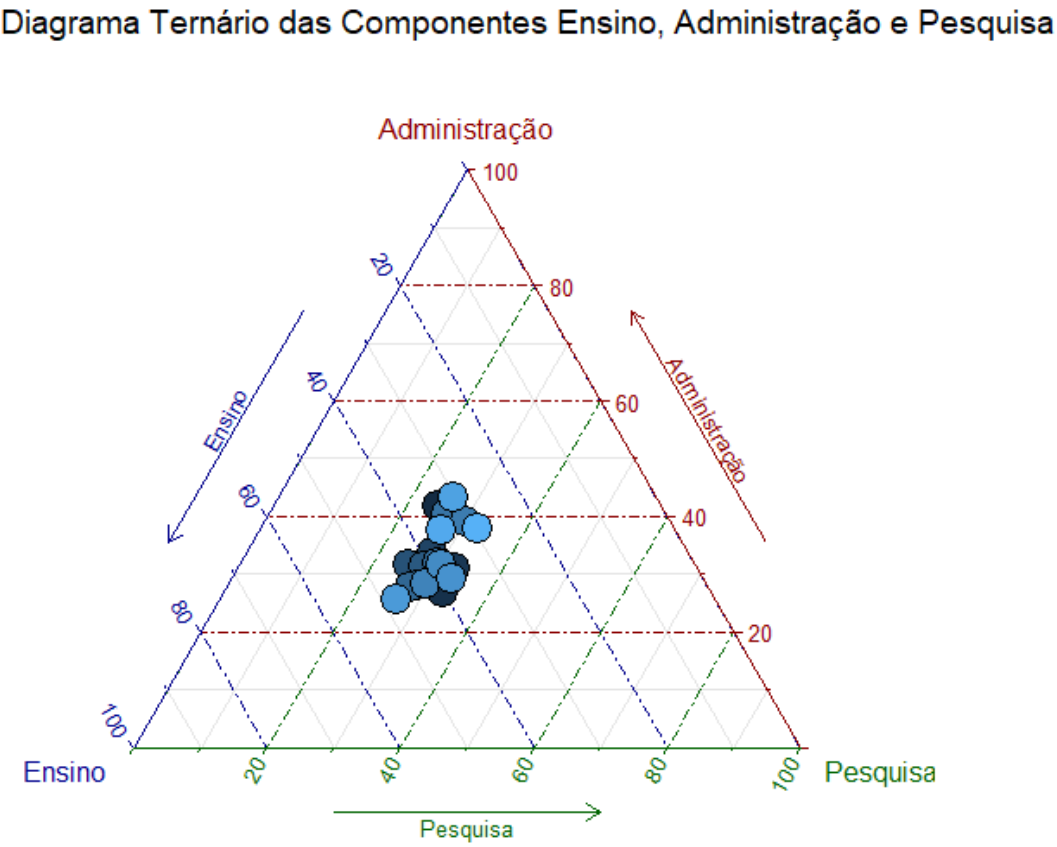
Tabela 34 - Subcomposição da Proporção de Tempo Gasto por Estatístico Acadêmico

Ensino	Administração	Pesquisa
0,3349	0,4163	0,2488
0,4040	0,2668	0,3292
0,3625	0,3112	0,3270
0,3942	0,3112	0,2940
0,3826	0,3366	0,2809
0,4220	0,2890	0,2890

Fonte: A Autora (2021).

Para obter as informações descritivas sobre o conjunto de dados, foi plotado o Diagrama ternário das 20 amostras das componentes: Ensino, administração e Pesquisa, conforme demonstrado na Figura 16.

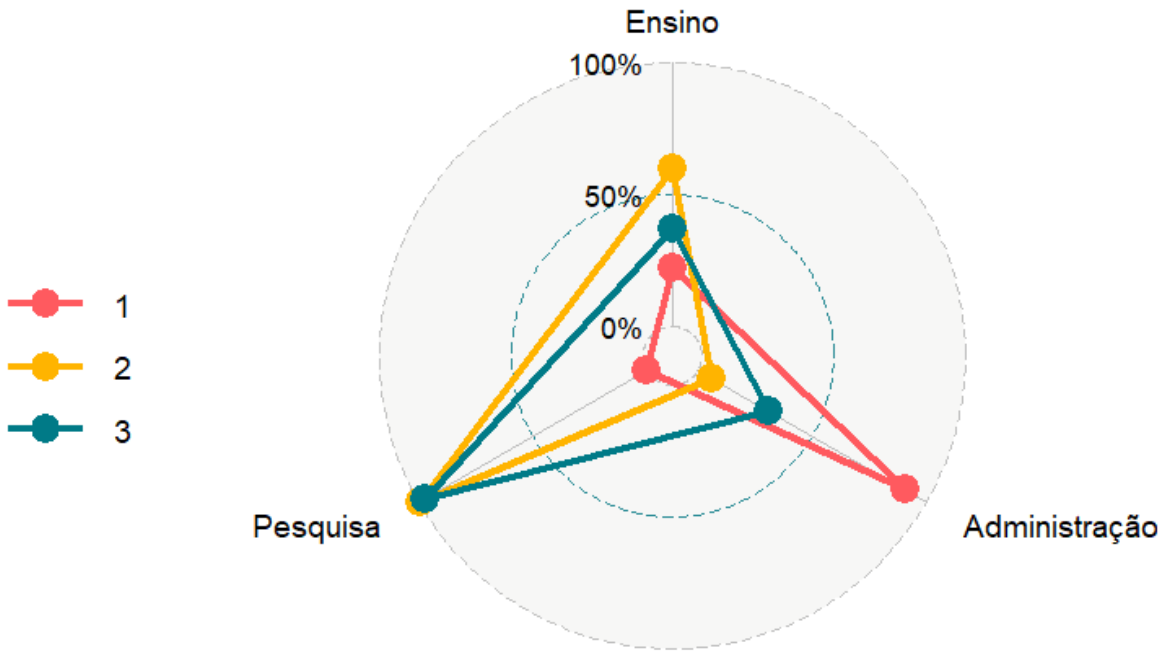
Figura 16 - Diagrama Ternário das Componentes Ensino, Administração e Pesquisa



Fonte: A Autora (2021).

Pelo diagrama ternário, pode-se afirmar que as amostras estão distribuídas majoritariamente nos entornos do baricentro, o que culmina na conclusão de que as componentes possuem proporções aproximadas. Pelo gráfico de radar, Figura 17, que representa o comportamento das 3 primeiras amostras das componentes, chega-se à conclusão de que a Pesquisa é variável que mais demanda tempo, além de que Administração e Ensino demandam proporções aproximadas de tempo.

Figura 17 - Gráfico de Radar das Três primeiras Amostras das Componentes Ensino, Administração e Pesquisa



Fonte: A Autora (2021).

Pela Tabela 35, observa-se que a maior média composicional é referente à componente Ensino, seguido por Administração e Pesquisa, porém, como observado no diagrama ternário, o tempo gasto é aproximadamente simétrico para estas 3 atividades.

Tabela 35 - Média Composicional das Componentes da Base 3

Ensino	Administração	Pesquisa
0,3824	0,3268	0,2908

Fonte: A Autora (2021).

Por fim, a Tabela 36, que apresenta a Matriz de Variância das componentes, permite destacar maior correspondência entre as variáveis Pesquisa e Ensino, seguida por Administração e Pesquisa, e uma menor correspondência entre Administração e Ensino.

Tabela 36 - Matriz de Variância das Componentes da Base 3

	Ensino	Administração	Pesquisa
Ensino	0,0000	0,0813	0,0288
Administração	0,0813	0,0000	0,0403
Pesquisa	0,0288	0,0403	0,0000

Fonte: A Autora (2021).

4.3.1 Regressão para a Base 3

Os dados composicionais possuem a restrição de soma constante, que pode gerar multicolinearidade das variáveis. Assim sendo, será feita a verificação de multicolinearidade dos mesmos. Para o ajuste em questão, optou-se por definir o tempo gasto em Sono como variável resposta e a subcomposição: Ensino, Administração e Pesquisa como variável explicativa, a fim de se observar o comportamento do sono em função das horas gastas em trabalho.

O modelo para análise será descrito da seguinte maneira:

$$Sono = \beta_0 + \beta_1 Ensino + \beta_2 Administração + \beta_3 Pesquisa + \epsilon$$

Para a verificação da multicolinearidade das variáveis explicativas, serão adotas as medidas descritas na Seção 2.4. A Tabela 37 apresenta os resultados respectivos.

Tabela 37 - Medidas de Multicolinearidade das Variáveis Explicativas da Base 3

	Ensino	Administração	Pesquisa
VIF	63.273,06	10.747,72	30.892,8
Autovalores	1,9018	1,0982	0,0000
K_m	1,0000	1,7317	Inf

Fonte: A Autora (2021).

Todos os valores de *VIF* são maiores que 5, indicando multicolinearidade, além disso, o número de condição tende a infinito, e, já que valores de condição maiores que 1000

significam que existe forte multicolinearidade das variáveis, pode-se dizer que as variáveis explicativas em questão são multicolineares. Por fim, por k_m, observa-se que a Pesquisa é linearmente dependente do Ensino.

Como visto que há uma relação linear entre as variáveis explicativas, serão realizados os ajustes pela Regressão Clássica, pela Regressão Ridge e pela Regressão Composicional.

4.3.1.1 Regressão Linear Clássica para a Base 3

O ajuste linear obtido para o modelo descrito na Seção 4.3.1 resultou nas estimativas apresentadas na Tabela 38.

Tabela 38 - Coeficientes Estimados pela Regressão Linear para a Base 3

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	0,2079	0,3413	0,60900	0,550
Ensino	$-0{,}0499$	0,5040	$-0{,}099$	0,922
Administração	0,2830	0,4831	0,5860	0,5660
Pesquisa	NA	NA	NA	NA

Fonte: A Autora (2021).

Supondo um nível de significância de 5% e considerando o Valor P, observa-se que não somente o intercepto, como todas as variáveis explicativas não são estatisticamente significantes para explicar o Tempo de Sono, com o ajuste clássico. Além disso, não há estimação para o último parâmetro – Pesquisa.

Além disso, para a Regressão Linear Clássica, encontrou-se as seguintes medidas de qualidade de ajuste (Tabela 39):

Tabela 39 - Medidas de Comparação de Qualidade do Ajuste.

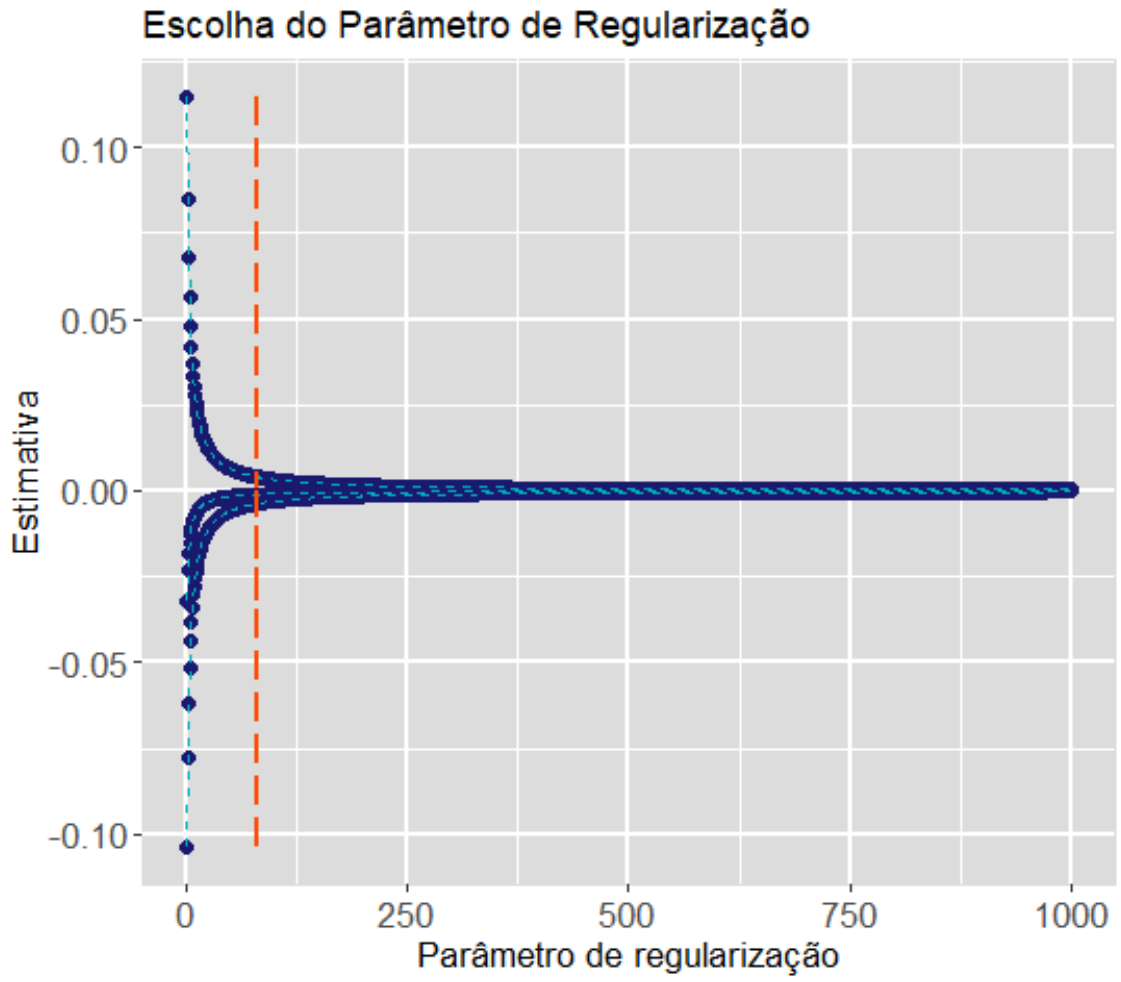
AIC	*BIC*
$-56{,}82$	$-52{,}83$

Fonte: A Autora (2021).

4.3.1.2 Regressão Ridge para a Base 3

A Regressão Ridge é uma alternativa para o controle da multicolinearidade, assim sendo, o modelo definido na Seção 4.3.1 foi ajustado com esta metodologia. Para a definição do parâmetro de regularização, pode-se observar, pela Figura 18, que as estimativas começam a se estabilizar em aproximadamente $\lambda = 80$, desse modo, este foi o parâmetro de regularização utilizado.

Figura 18 - Escolha do Parâmetro de Regularização para a Base 3



Fonte: A Autora (2021).

O ajuste da regressão Ridge, com parâmetro de regularização $\lambda = 80$, resultou nas estimativas demonstradas na Tabela 40.

Tabela 40 - Estimativas pela Regressão Ridge para a Base 3

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	0,2981	NA	NA	NA
Ensino	−0,0772	0,0029	1,3190	0,1870
Administração	0,0814	0,0028	1,490	0,1360
Pesquisa	−0,0472	0,0036	0,3290	0,7420

Fonte: A Autora (2021).

Pode-se concluir com o ajuste da regressão Ridge, supondo nível de significância de 5%, e considerando o Valor P, que o intercepto e as variáveis explicativas não são estatisticamente significantes para explicar o Tempo de Sono. Com o ajuste realizado pela regressão Ridge, não foi possível calcular as medidas de comparação *AIC* e *BIC*, desse modo, considerou-se apenas as estimativas do modelo para futuras comparações.

4.3.1.3 Regressão Composicional para a Base 3

Novamente, com o objetivo de avaliar as alternativas composicionais, e, valendo-se de que as componentes da subcomposição possuem restrição de soma constante, realizou-se o ajuste da Regressão Composicional. Para tal ajuste, aplicou-se as Transformações *ALR, CLR* e *ILR* nas variáveis explicativas – Tempo de Estudo, Administração e Pesquisa.

4.3.1.3.1 Regressão Composicional com Transformação *ALR* para a Base 3

Aplicou-se a Transformação *ALR* nas variáveis explicativas para que estas tenham correspondência com o espaço dos números reais. A Tabela 41 apresenta as estimativas encontradas. Observa-se que somente o intercepto é estatisticamente significante, considerando um nível de significância de 5%, para explicar o tempo de sono.

Tabela 41 - Coeficientes Estimados pela Regressão Composicional com *ALR* no Espaço *Simplex* para a Base 3

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	0,2894	0,0247	11,7300	$1{,}42X10^{-9}$ ***
*ALR(X)*1	−0,0568	0,0709	−0,8010	0,4340
*ALR(X)*2	0,0671	0,0599	1,1190	0,2790

Fonte: A Autora (2021).

Para que os coeficientes retornassem ao espaço euclidiano, aplicou-se a inversa da *ALR*, assim sendo, a Tabela 42 apresenta as estimativas dos coeficientes estimados no espaço real.

Tabela 42 - Coeficientes Estimados pela Regressão Composicional com *ALR* no Espaço Real para a Base 3

Coeficientes	Areia	Lodo	Argila
Valores Estimados	0,3134	0,3548	0,3318

Fonte: A Autora (2021).

Por fim, a fim de realizar comparações com os outros modelos ajustados, a Tabela 43 apresenta as medidas descritas em 2.4.

Tabela 43 –Medidas de Qualidade do Ajuste para a Regressão Composicional com *ALR* para a Base 3

AIC	*BIC*
−57,00	−53,02

Fonte: A Autora (2021).

4.3.1.3.2 Regressão Composicional com Transformação *CLR* para a Base 3

A segunda transformação descrita é a *CLR*, que permite que a composição se associe a um vetor muldimensional. A aplicação desta transformação nas variáveis explicativas do

modelo descrito na Seção 4.3.1, resultou nas estimativas da Tabela 44, é importante destacar que o intercepto é estatisticamente significante para explicar o tempo de sono, porém, novamente, o último parâmetro não foi estimado, o que ocorreu na simulação realizada na Seção 3, no ajuste do modelo da Base 1 e no ajuste da Base 2.

Tabela 44 - Coeficientes Estimados pela Regressão Composicional com *CLR* no Espaço *Simplex* para a Base 3

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	0,2894	0,0247	11,7300	$1{,}42X10^{-9}$ ***
*CLR(X)*1	−0,0465	0,1635	−0,2840	0,7800
*CLR(X)*2	0,0774	0,1498	0,5170	0,6120
*CLR(X)*3	*NA*	*NA*	*NA*	*NA*

Fonte: A Autora (2021).

A tabela 45 apresenta os coeficientes no espaço real, após a aplicação da inversa da *CLR*.

Tabela 45 - Coeficientes Estimados pela Regressão Composicional com *CLR* no Espaço Real para a Base 3

Coeficientes	Ensino	Administração	Pesquisa
Valores Estimados	0,4690	0,5309	*NA*

Fonte: A Autora (2021).

Por fim, a Tabela 46 explicita a medidas de comparação de ajuste.

Tabela 46 - Medidas de Qualidade do Ajuste para a Regressão Composicional com *CLR* para a Base 3

AIC	*BIC*
−57,00	−53,02

Fonte: A Autora (2021).

4.3.1.3.2 Regressão Composicional com Transformação *ILR* para a Base 3

Por fim, para o mesmo modelo de regressão descrito na Seção 4.3.1, o último ajuste performado foi o da regressão composicional com transformação *ILR*. Desse modo, encontrou-se as estimativas para os coeficientes na Tabela 47, observa-se que somente o intercepto é estatisticamente significante para explicar o tempo de sono, a um nível de 5% de significância. Salienta-se que a aplicação da *ILR* nas variáveis explicativas converte os dados para o espaço *simplex.*

Tabela 47 - Coeficientes Estimados pela Regressão Composicional com *ILR* no Espaço *Simplex* para a Base 3

Coeficientes	Valores Estimados	Erro Padrão	Valor T	Valor P
Intercepto	0,2894	0,0247	11,7300	$1{,}42X10^{-9}$ ***
*ILR(X)*1	0,0876	0,0597	1,4680	0,1600
*ILR(X)*2	−0,0126	0,1233	−0,1020	0,9200

Fonte: A Autora (2021).

Para uma representação dos coeficientes no espaço euclidiano, aplicou-se a inversa da *ILR*, os resultados são demonstrados na Tabela 48.

Tabela 48 - Coeficientes Estimados pela Regressão Composicional com *ILR* no espaço real para a Base 3

Coeficientes	Ensino	Administração	Pesquisa
Valores Estimados	0,3145	0,3560	0,3295

Fonte: A Autora (2021).

Por fim, para a verificação de qualidade do ajuste, a Tabela 49 apresenta as medidas descritas na Seção 2.6.

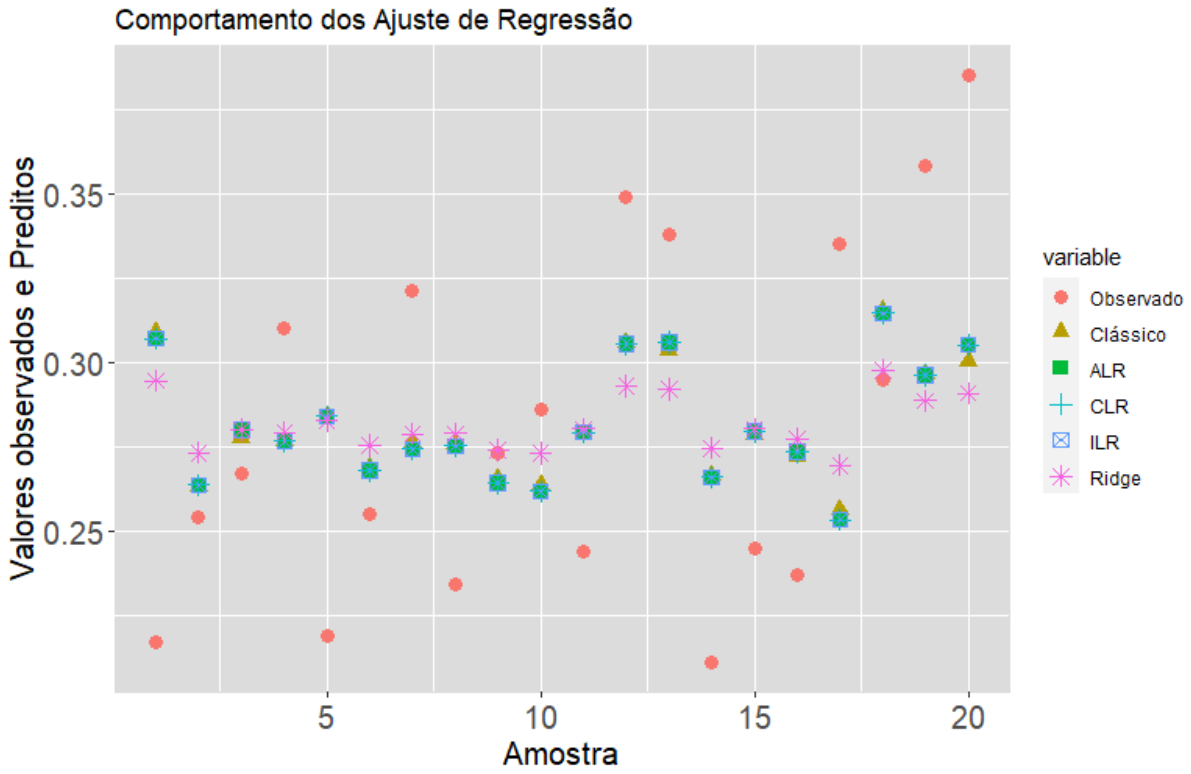
Tabela 49 - Medidas de Qualidade do Ajuste para a Regressão Composicional com *ILR* para a Base 3

AIC	*BIC*
−57,00	−53,02

Fonte: A Autora (2021).

Finalmente, com o propósito derradeiro de comparação do comportamento dos ajustes aplicados, a Figura 19 foi plotada. Esta apresenta os valores originais e os valores preditos para cada ajuste performado. A partir dela, é possível dizer que a Regressão Clássica e a Regressão Composicional com aplicação da *ARL*, *CLR* e da *ILR* (mesmos valores previstos), resultaram em previsões muito análogas e mais próximas dos valores observados do Tempo de Sono, enquanto a Regressão Ridge subestimou os valores observados.

Figura 19 – Comportamento dos Ajustes de Regressão Performados para a Base 3



Fonte: A Autora (2021).

4.4 SÍNTESE DA APLICAÇÃO

A partir da análise dos conjuntos de dados, algumas questões devem ser levantadas. A fim de buscar um consenso sobre qual modelo de regressão melhor se adequa à restrição dos dados, considerou-se também a regressão Ridge – comumente utilizada para contornar problemas em dados com restrições. Dessa forma, serão realizados paralelos entre as medidas de comparação de qualidade de ajuste de modelos, e entre a performance de valores preditos pelos 3 modelos de regressão aplicados. As Tabelas 50, 51 e 52 retratam, respectivamente, o *AIC* e o *BIC* para os ajustes (exceto para a regressão Ridge) da Base 1, da Base 2 e, por fim, da Base 3.

Tabela 50 – Compilado de Medidas de Ajuste para a Base 1

	AIC	*BIC*
Regressão Clássica	$-750{,}44$	$-738{,}85$
Regressão Comp. – *CLR*	$-749{,}55$	$-740{,}27$
Regressão Comp. – *ILR*	$-749{,}55$	$-740{,}27$

Fonte: A Autora (2021).

Tabela 51 – Compilado de Medidas de Ajuste para a Base 2

	AIC	*BIC*
Regressão Clássica	$331{,}08$	$339{,}34$
Regressão Comp. – *ALR*	$333{,}48$	$340{,}13$
Regressão Comp. – *CLR*	$333{,}48$	$340{,}13$
Regressão Comp. – *ILR*	$333{,}48$	$340{,}13$

Fonte: A Autora (2021).

Tabela 52 – Compilado de Medidas de Ajuste para a Base 3

	AIC	*BIC*
Regressão Clássica	$-56{,}82$	$-52{,}83$
Regressão Comp. – *ALR*	$-57{,}00$	$-53{,}02$
Regressão Comp. – *CLR*	$-57{,}00$	$-53{,}02$
Regressão Comp. – *ILR*	$-57{,}00$	$-53{,}02$

Fonte: A Autora (2021).

Analisando as 3 tabelas supracitadas, observa-se majoritariamente dois fatos:

- Os valores de *AIC* e de *BIC* são mesmos para as regressões Composicionais, independentemente da transformação utilizada.
- Os valores de *AIC* e de *BIC* são muito próximos para a regressão Clássica e para a regressão Composicional.

Apesar desta semelhança, e considerando o *BIC* – mais punitivo, pode-se afirmar que, para todas as bases, a regressão Composicional obteve melhores comportamentos - quanto menor o *BIC*, melhor o ajuste do modelo.

As Figuras 11, 15 e 19, apresentaram os comportamentos dos valores ajustados pelos três métodos de regressões performados, para os três conjuntos de dados. Em todas foi possível observar que os valores preditos pela regressão Composicional, independentemente da transformação aplicada, foram análogos. Além disso, a regressão Ridge tendeu a subestimar os valores observados; por fim, as predições com a regressão Clássica e com a regressão Composicional foram semelhantes.

5 CONCLUSÃO

Esta Seção é dedicada a debater os resultados encontrados ao longo do trabalho. O objetivo esperado era observar os contrastes ao considerar a metodologia composicional em dados com a restrição de soma constante. Dessa forma, estas metodologias foram aplicadas em diferentes bases de dados e, também, na simulação performada.

Primeiramente, pela análise descritiva composicional, foi possível observar as diferentes formas de visualização dos dados, nas quais evidencia-se as correspondências entre as componentes das variáveis, as restrições ocasionadas pelas transformações, e as maneiras de contornar tais restrições. Também foram observáveis as diferenças em não considerar a análise composicional, como por exemplo na média aritmética ou na média composicional dos dados.

Sobre a simulação, foi bastante destoante a performance entre os modelos de regressão Clássica e os modelos de regressão Composicional, visto que, além de maiores vieses absolutos e maiores erros padrões, a regressão Clássica também foi ineficiente para estimar o último parâmetro tratado em questão – situação análoga para os casos de modelos com ou sem intercepto. Apesar disso, a regressão Composicional, com o uso da Transformação Logaritmo Centralizada - *CLR*, de mesma forma, não estimou o último parâmetro requisitado. A partir disso, é possível afirmar que, na presença de dados composicionais, os ajustes de regressão mais apropriados são aqueles que considerarem a transformação logaritmo aditiva - *ALR* e a transformação logarítmica isométrica – *ILR*, desde que não existam valores nulos nas componentes, caso isto seja um fato, somente a transformação *ILR* é indicada, já que a *ALR* pode levar a valores infinitos.

Pela análise da aplicação aos conjuntos de dados reais, é possível concluir que a regressão Composicional com uso da Transformação Logaritmo Isométrica da Razão – *ILR* pode ser dita como a mais apropriada para produzir informação sobre dados composionais. Isto é afirmado pelos motivos enumerados a seguir:

1 – Produz menores erros padrões e menores vieses absolutos que a regressão Clássica;

2 – Não leva a valores infinitos como a regressão Composicional com a transformação *ALR*;

3 – Estima todos os parâmetros de interesse, diferentemente da regressão Clássica e da regressão Composicional com a transformação *CLR*;

4 – É menos trabalhosa que, por exemplo, a regressão Ridge.

Além disso, mesmo que a regressão clássica tenha entregado resultados semelhantes aos da regressão composicional, o ajuste clássico aos dados composicionais não respeita os princípios da regressão clássica de não existência de relação linear entre as covariáveis.

Por fim, pode-se salientar que, de alguma forma, os resultados para as três transformações composicionais foram análogos, necessitando maiores investigações.

Assim sendo, a partir de todas as considerações descritas, conclui-se que é importante lançar um olhar específico para a análise de dados que possuem a restrição de soma constante. Os dados composicionais estão nas mais variadas áreas e carecem de uma metodologia particular, doravante qual os resultados são mais condizentes com a realidade.

O estudo realizado foi pautado nas composições como variáveis preditoras. Como motivação futura, aconselha-se considerar situações em que as composições apareçam como variáveis respostas multivariadas.

Recomenda-se, também, ajustar os modelos de regressão Composicionais com outras transformações existentes (Transformação Esférica, Transformação Log-Central), a fim de comparar o desempenho destas com o ajuste da *ILR*, por exemplo.

6 REFÊNCIAS BIBLIOGRÁFICAS

AITCHISON, J. 1986 **The Statistical Analysis of Compositional Data**. London: Chapman and Hall.

AITCHISON et al. (2002) J AITCHISON, C. BARCELÓ-VIDAL, JJ EGOZCUE e V PAWLOWSKT-GLAHN. **A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis**. Proceedings of IAMG, 2, 387-392.

AKAIKE, H**. A new look at the statistical model identification.** IEEE Transactions on Automatic Control, 1974. 19 (6): 716–723.

BARCELÓ-VIDAL, C., MARTÍN-FERNINDES, J. A., e PAWLOWSKY-GLAHN, V., 2001, **Mathematical foundations of compositional data analysis: in G**. Ross, ed., Proceedings of the sixth annual conference of the International Association for Mathematical Geology Cancun (Mexico) CD-Rom.

BRAGA, LUIS PAULO VIEIRA. **Análise de dados composicionais: Uma abordagem aplicada e computacional no ambiente R / Luis Paulo Vieira Braga.** - 1. ed. - Rio de Janeiro: E-papers, 2020.

BOOGAART, K.G. van den; TOLOSANA-DELGADO, R. **Analyzing Compositional Data with R**. New York: Springer, 2013. 258 p.

BURNHAM, K. P.; ANDERSON, D. R. **Multimodel inference: understanding aic and bic in model selection. Sociological Methods and Research**. Beverly Hills, v.33, n.2, p.261-304, May 2004.

BURNHAM, K. P.; ANDERSON, D. R. **Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach**, 2 ed. New York: Springer-Verlag, 2002.

CHASTIN, Sebastien et al. **Joint association between accelerometry-measured daily combination of time spent in physical activity, sedentary behaviour and sleep and all-cause mortality: a pooled analysis of six prospective cohorts using compositional analysis**. British Journal of Sports Medicine, 2021.

COENDERS.G and PAWLOWSKY-GLAHN, V. **On interpretations of tests and effect sizes in regression models with a compositional predictor.** SORT-Statistics and Operations Research Transactions, 44(1):201–220, Jun. 2020.

EGOZCUE, J.J., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G. & BARCELÓ-VIDAL, C., (2003). **Isometric logratio transformations for compositional data analysis.** Mathematical Geology, 35 (3), 279-300.

EGOZCUE, J.J., PAWLOWSKY-GLAHN, V. (2005) **Groups of parts and their balances in compositional data analysis.** Mathematical Geology, 37(7), 795–828.

FREUND, R. J.; WILSON, W. J.; SA, P. **Regression analysis – Statistical Modeling of a response variable**. Elsevier, Inc., San Diego, 459p, 2006.

GRANTHAM and M. A. Velbel. **The influence of climate and topography on rock-fragment abundance in modern fluvial sands of the southern blue ridge mountains, north carolina.** Journal of Sedimentary Research, 58:219–227, 1988.

HAIR, JR., J. H.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. trad. Adonai Schlup Sant'Ana e Anselmo Chaves Neto. **Análise Multivariada de Dados**. 5 ed. Porto Alegre: Bookman. 2005.

HOERL, A. E., KENNARD, R. W. E BALDWIN, K. F., **'Ridge regression: some simulations'**, Communications in Statistics - Theory and Methods 4(2), 105–123, 1975.

KUHN, T. et al. **Volcanic and hydrothermal history of ridge segments near the rodrigues triple junction (central indian ocean) deduced from sediment geochemistry**. Marine Geology, 169:391–409, 2000.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. John, Wiley and Sons, Inc., New York, 612p, 2006.

PAWLOWSKY-GLAHN, V., EGOZCUE, J.J., TOLOSANA-DELGADO, R. (2015) **Modeling and Analysis of Compositional Data**. John Wiley & Sons.

PAWLOWSKY-GLAHN, V., BUCCIANTI, A., **Compositional Data Analysis: Theory and Applications**. Wiley, 2011, 400 p.

PEARSON, Karl (1897). **The Chances of Death and Other Studies in Evolution**, 2 Vol. London: Edward Arnold.

PEREIRA C. A. B. and STERN J. M. **Special characterizations of standard discrete models**. REVSTAT - Statistical Journal, 6(3):199–230, 2008.

R Core Team (2020). R: A language and environment for statisticalb computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

SCHWARZ, G. **Estimating the dimensional of a model**. Annals of Statistics, Hayward, v.6, n.2, p.461-464, Mar. 1978.

SPRACKLEN, C. **Compositional observations and their role in regression**. University of Cape Town, 2018.

TOLOSAMA-DELGADO, R and EYNATTEN, H. **Simplifying compositional multiple regression: Application to grain size controls on sediment geochemistry.** Computers & Geosciences, 36:577–589, 2010.

APÊNDICE A – CÓDIGO EM LINGUAGEM R PARA SIUMLAÇÃO DE MODELOS COMPOSICIONAIS

```
# EXEMPLO PARA A SIMULAÇÃO CONSIDERANDO O INTERCEPTO NO MODELO COM COVARIÁVEIS COMPOSICIONAIS.

# PACOTES NECESSÁRIOS

library("dplyr")
library("compositions")
library("ggplot2")
library("MCMCpack")
library("reshape")

#-----------------------------------------------------------------------------------------------------------
# PREPARAÇÃO
#-----------------------------------------------------------------------------------------------------------

set.seed(123)
iRpl <- 5000 # - Number of Monte Carlo's replications
vN <- c(10, 20, 40, 80, 160) #- Sample size
iN <- length(vN)

vAlpha <- c(1, 2, 3)
vBeta <- matrix(c(1, 0.33, 0.33, 0.34), ncol = 1)
iP <- dim(vBeta)[1]
dSigma <- 2 #- True sigma
vParameters <- c(vBeta, rep(dSigma^2, 2))

mEstimation <- array(NA, dim = c(iP+2, 6, length(vN)),
                     dimnames = list(
                       c(paste(expression(beta), 1:iP, sep = ""),
                         "QMRes", "Sigma2MV"),
                       c("True", "Estimated", "SE",
```

```
                  "Bias", "|Bias|", "MSE"),
                vN))

#-------------------------------------------------------------------------------------------------------
# AMOSTRAS COMPOSICIONAIS
#-------------------------------------------------------------------------------------------------------

for(i in 1:iN)
{
 mX <-  cbind(1, rdirichlet(vN[i], vAlpha))
 dMu <- mX %*% vBeta
 mTheta <- matrix(NA, nrow = iP + 2, ncol = iRpl) # Betas + sigma2
 rownames(mTheta) <- c(paste(expression(beta), 1:iP, sep = ""),
                  "QMRes", "sigma2MLE")

#-------------------------------------------------------------------------------------------------------
# REPLICAÇÕES COM USO DA TRANSFORMAÇÃO ILR NAS COVARIÁVEIS
#-------------------------------------------------------------------------------------------------------

 for(j in 1:iRpl)
 {
  vY <- rnorm(vN[i], mean = dMu, sd = dSigma)
  mData <- data.frame(vY, mX[, -1])
  mModel <- lm(vY ~ ilr(mX[, -1]), data = mData) # vY ~ ilr(X sem intercepto) -> mX[-1]
  mTheta[1:iP, j] <- c(coef(mModel)[1],
                   ilrInv(coef(mModel)[-1], orig = mData))
  mTheta[iP+1, j] <- (summary(mModel)$sigma^2)# QMRes     (verificar se transforma algo
na var)
  mTheta[iP+2, j] <- ((vN[i]-iP)/vN[i])*(summary(mModel)$sigma^2) #EMV
 }
 mEstimation[,, i] <- cbind(vParameters,
                       apply(mTheta, 1, mean),
                       apply(mTheta, 1, sd),
                       apply((mTheta - vParameters), 1, mean),
```

```
                    apply(abs(mTheta - vParameters), 1, mean),
                    apply((mTheta - vParameters)^2, 1, mean))
}

mSizes <- apply(mEstimation, 3L, c)
colnames(mSizes) <- paste("n", vN, sep = "")
mParam <- expand.grid(dimnames(mEstimation)[1:2])
iLen <- dim(mParam)[1]
mParam <- cbind(mParam, rep("ILR", times = iLen))
colnames(mParam) <- c("Parametro", "Medida", "Tipo")
mTable <- data.frame(mParam, mSizes)

#-------------------------------------------------------------------------------------------------------
# REPLICAÇÕES SEM TRANSFORMAÇÕES NAS COVARIÁVEIS
#-------------------------------------------------------------------------------------------------------

 for(j in 1:iRpl)
 {
  vY <- rnorm(vN[i], mean = dMu, sd = dSigma)
  mData <- data.frame(vY, mX[, -1])
  mModel <- lm(vY ~ ., data = mData)
  mTheta[1:iP, j] <- coef(mModel)
  mTheta[iP+1, j] <- (summary(mModel)$sigma^2) #QMRes
  mTheta[iP+2, j] <- ((vN[i]-iP)/vN[i])*(summary(mModel)$sigma^2) #EMV
 }
 mEstimation[,, i] <- cbind(vParameters,
                    apply(mTheta, 1, mean),
                    apply(mTheta, 1, sd),
                    apply((mTheta - vParameters), 1, mean),
                    apply(abs(mTheta - vParameters), 1, mean),
                    apply((mTheta - vParameters)^2, 1, mean))
}

mSizes <- apply(mEstimation, 3L, c)
```

```
colnames(mSizes) <- paste("n", vN, sep = "")
mParam <- expand.grid(dimnames(mEstimation)[1:2])
iLen <- dim(mParam)[1]
mParam <- cbind(mParam, rep("Sem Transf.", times = iLen))
colnames(mParam) <- c("Parametro", "Medida", "Tipo")
mTableAux <- data.frame(mParam, mSizes)

mTable <- bind_rows(mTable, mTableAux)

#-------------------------------------------------------------------------------------------------------
# VIÉS ABSOLUTO
#-------------------------------------------------------------------------------------------------------

mAux <- mTable[mTable$Medida == "|Bias|", ]
mAux <- mAux[, -2]
mData <- melt(mAux, id = c("Parametro", "Tipo"))
levels(mData$variable) <- vN

iAux <- rep(rep(c(TRUE, FALSE), times = c(iP, 2)),
            times = iN*2)


#-------------------------------------------------------------------------------------------------------
# GRÁFICO DO VIÉS ABSOLUTO
#-------------------------------------------------------------------------------------------------------

pTS <- ggplot(data = mData[iAux, ],
              aes(x = variable, y = value, group = Tipo)) +
  geom_point(aes(color = Tipo), size = 2) +
  geom_line(aes(linetype = Tipo, color = Tipo), size = 0.9) +
  facet_grid(. ~ Parametro) +
  labs(x = "Tamanho amostral", y = "Viés absoluto") +
  theme(legend.position = "bottom",
        axis.text = element_text(size = 15),
```

```
      axis.title = element_text(size = 15),
      panel.background = element_rect(fill = "#DCDCDC",
                                      colour = "#DCDCDC",
                                      size = 0.5, linetype = "solid"),
      panel.grid.major = element_line(size = 0.9, linetype = 'solid',
                                      colour = "white"),
      panel.grid.minor = element_line(size = 0.5, linetype = 'solid',
                                      colour = "white"))

ggsave(filename = "Bias.pdf",
     plot = pTS, width = 9.0, height = 6.0)

#-------------------------------------------------------------------------------------------------------------
# ERRO QUADRÁTICO MÉDIO
#-------------------------------------------------------------------------------------------------------------

mAux <- mTable[mTable$Medida == "MSE", ]
mAux <- mAux[, -2]
mData <- melt(mAux, id = c("Parametro", "Tipo"))
levels(mData$variable) <- vN

iAux <- rep(rep(c(TRUE, FALSE), times = c(iP, 2)),
          times = iN*2)

#-------------------------------------------------------------------------------------------------------------
# GRÁFICO DO ERRO QUADRÁTICO MÉDIO
#-------------------------------------------------------------------------------------------------------------

pTS <- ggplot(data = mData[iAux, ],
            aes(x = variable, y = value, group = Tipo)) +
  geom_point(aes(color = Tipo), size = 2) +
  geom_line(aes(linetype = Tipo, color = Tipo), size = 0.9) +
  facet_grid(. ~ Parametro) +
  labs(x = "Tamanho amostral", y = "Erro Quadrático Médio") +
```

```
 theme(legend.position = "bottom",
       axis.text = element_text(size = 15),
       axis.title = element_text(size = 15),
       panel.background = element_rect(fill = "#DCDCDC",
                                       colour = "#DCDCDC",
                                       size = 0.5, linetype = "solid"),
       panel.grid.major = element_line(size = 0.9, linetype = 'solid',
                                       colour = "white"),
       panel.grid.minor = element_line(size = 0.5, linetype = 'solid',
                                       colour = "white"))

ggsave(filename = "MSE.pdf",
      plot = pTS, width = 9.0, height = 6.0)
```

APÊNDICE B – CÓDIGO EM LINGUAGEM R PARA ANÁLISE E REGRESSÃO COMPOSICIONAL

```
# EXEMPLO PARA A BASE 2. AS ANÁLISES SÃO ANÁLOGAS PARA TODOS AS BASES.

# PACOTES NECESSÁRIOS

library(compositions)
library(ggradar)
library(ggtern)
library(ggplot2)
library(dplyr)
library(scales)

# BANCO

data(ArcticLake)

####### CÓDIGOS PARA ANÁLISES DESCRITIVAS COMPOSICIONAIS

artico <- data.frame(ArcticLake)

#-------------------------------------------------------------------------
# diagrama ternário
#-------------------------------------------------------------------------

amostras <- 1:39
plot <- ggtern(data = data.frame(ArcticLake), aes(x = sand, y = silt, z = clay )) +
  ggplot2::geom_point(aes(fill = amostras),
                      size = 6,
                      shape = 21,
                      color = 'black') +
```

```
  ggtitle('Diagrama Ternário das Componentes Areia, Lodo e Argila') +
  labs(fill = 'Amostras') +
  theme_rgbw() +
  theme(legend.position = c(0,2),
        legend.justification = c(1, 1))

#---------------------------------------------------------------------------
# gráfico de radar
#---------------------------------------------------------------------------

artico_radar <- artico[,-4] %>%
  as_tibble(rownames = "group") %>%
  mutate_at(vars(-group), rescale) %>%
  head(3)

ggradar(artico_radar)

#---------------------------------------------------------------------------
# média composicional
#---------------------------------------------------------------------------

work <- acomp(artico[,-4])
valores <- summary(work)
media_comp <- valores$mean

#---------------------------------------------------------------------------
# matriz de variancia
#---------------------------------------------------------------------------

valores$variation


####### CÓDIGOS PARA REGRESSÕES COMPOSICIONAIS
```

```
artico <- data.frame(artico)
X <- cbind(artico$sand, artico$silt, artico$clay)
Y <- artico$depth

# ---------------- ALR

modelo_alr <- lm(Y ~ alr(X))
summary(modelo_alr)
alrInv(coef(modelo_alr)[-1])
AIC(modelo_alr)
BIC(modelo_alr)

# --------------- CLR

modelo_clr <- lm(Y~clr(X))
summary(modelo_clr)
clrInv(coef(modelo_clr)[-1])
AIC(modelo_clr)
BIC(modelo_clr)


# --------------- ILR

modelo_ilr <- lm(Y~ilr(X))
summary(modelo_ilr)
ilrInv(coef(modelo_ilr)[-1])
AIC(modelo_ilr)
BIC(modelo_ilr)
```