UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-graduação em Estatística

João Paulo Sena Souza

AJUSTE DE MODELO DE REGRESSÃO LINEAR MÚLTIPLA EM DADOS DE δ^{15}N DO SOLO PARA O BIOMA CERRADO

Belo Horizonte

2023

João Paulo Sena Souza

AJUSTE DE MODELO DE REGRESSÃO LINEAR MÚLTIPLA EM DADOS DE δ^{15}N DO SOLO PARA O BIOMA CERRADO

Monografia de especialização apresentada ao Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção de título de Especialista em Estatística.

Área de Ênfase: Estatística

Orientadora: Thais Rotsen Correa

Belo Horizonte

2023

Souza,João Paulo Sena

S729a Ajuste de modelo de regressão linear múltipla em dados de $\delta^{15}N$ do solo para o bioma cerrado [recurso eletrônico] / João Paulo Sena Souza —2023.
1 recurso online (52 f. il, color.): pdf.

Orientador Thais Rotsen Correa.
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística
Referências: 35-37.

1. Estatística. 2. Isótopos – Nitrogênio. 3.Modelo preditivo. II. Correa, Thais Rotsen.. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais – ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 298ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE JOÃO PAULO SENA SOUZA.

Aos vinte e sete dias do mês de junho de 2023, às 14:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **João Paulo Sena Souza**, intitulado: "Ajuste de Modelo de Regressão Linear Múltipla em Dados de δ 15N do Solo para o Bioma Cerrado", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, a Presidente da Comissão, Professora Thais Rotsen Correa – Orientadora, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente o candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 27 de junho de 2023.

Thais Rotsen Correa

Prof.ª Thais Rotsen Correa (Orientadora)
Departamento de Estatística / ICEX / UFMG

gov.br Documento assinado digitalmente
MARCELO AZEVEDO COSTA
Data: 23/08/2023 14:02:59-0300
Verifique em https://validar.iti.gov.br

Prof. Marcelo Azevedo Costa
DEP/UFMG

gov.br Documento assinado digitalmente
GUILHERME LOPES DE OLIVEIRA
Data: 22/08/2023 15:36:22-0300
Verifique em https://validar.iti.gov.br

Prof. Guilherme Lopes de Oliveira
Departamento de Computação - CEFET MG

RESUMO

O nitrogênio desempenha um papel fundamental na produtividade primária e na estrutura dos ecossistemas terrestres. Em alguns biomas, como o Cerrado, o nitrogênio é um nutriente limitante da produtividade. Portanto, é essencial compreender os processos que afetam sua disponibilidade e transformações. Isótopos estáveis de nitrogênio ($\delta^{15}N$) fornecem uma medida integradora desses processos, permitindo a avaliação da contribuição relativa de diferentes fontes de nitrogênio para o ambiente, como a deposição atmosférica e a fixação biológica. Entender a variação do $\delta^{15}N$ no solo fornece uma compreensão abrangente do ciclo do nitrogênio e seus efeitos nos ecossistemas terrestres. Neste contexto, o objetivo do presente trabalho é propor um modelo preditivo e explicativo para o $\delta^{15}N$ do solo do bioma Cerrado. Foram coletadas 97 amostras de solo em diversas localidades do bioma. Para cada localidade, foram extraídos valores de variáveis preditoras climáticas e biofísicas provenientes de dados espaciais secundários; e variáveis edáficas provenientes de análises laboratoriais das amostras de solo. No total, 46 variáveis preditoras do $\delta^{15}N$ foram inseridas no banco de dados. Um modelo de regressão linear múltipla foi ajustado considerando todas as variáveis que apresentaram correlação com a variável dependente. Posteriormente, modelos de regressão foram ajustados seguindo as suposições da regressão. Para isso, matrizes de correlação foram usadas para eliminar variáveis preditoras correlacionadas, evitando multicolinearidade. O modelo final foi ajustado após seleção das melhores variáveis pelo método *stepwise* AIC (Critério de Informação de Akaike). Os resíduos do modelo final foram testados para normalidade, homoscedasticidade, autocorreção e autocorrelação espacial. Para avaliar o desemprenho preditivo, foi realizada uma validação cruzada *Leave-One-Out*. O primeiro modelo (Modelo 1), ajustado com todas as variáveis, obteve o R^2_{adj} de 0,66 e o R^2 de 0,81 após validação cruzada. Apesar do alto poder preditivo, o Modelo 1 não respeita os pressupostos da regressão múltipla, o que impede a interpretação dos coeficientes do modelo. O modelo final seguiu todas as suposições da regressão linear múltipla e obteve o R^2_{adj} de 0,57 e o R^2 de 0,59 após validação cruzada. Apesar do menor coeficiente de determinação, os coeficientes do modelo podem ser interpretados. As variáveis que contribuíram para o ajuste do modelo final foram concentração Ca^{2+}, $\delta^{13}C$ do solo, sazonalidade de temperatura, amplitude anual da temperatura e precipitação pluviométrica do trimestre mais úmido. Todas estas variáveis foram significativas para o modelo final e apresentaram coeficientes coerentes com sua contribuição esperada para a variação do $\delta^{15}N$ do solo. O modelo ajustado traz avanços no estudo do ciclo biogeoquímico do nitrogênio para o bioma Cerrado, pois acrescenta dados edáficos e indica uma influência positiva e ainda pouco explorada do Ca^{2+} no $\delta^{15}N$ do solo. Apesar do avanço, o modelo final pode ter excluído variáveis preditoras importantes na explicação do fenômeno devido aos critérios de seleção de variáveis e ao método utilizado. Futuros trabalhos devem aplicar métodos que identifiquem relações não lineares entre as variáveis.

Palavras-chave: isótopos estáveis de nitrogênio; variáveis edáficas; modelo preditivo

ABSTRACT

Nitrogen plays a fundamental role in primary productivity and the structure of terrestrial ecosystems. In some biomes, such as the Cerrado, nitrogen is a limiting nutrient for productivity. Therefore, it is essential to understand the processes that affect its availability and transformations. Stable nitrogen isotopes ($\delta^{15}N$) provide an integrated measure of these processes, allowing for the assessment of the relative contribution of different nitrogen sources to the environment, such as atmospheric deposition and biological fixation. Understanding the variation of $\delta^{15}N$ in soil provides a comprehensive understanding of the nitrogen cycle and its effects on terrestrial ecosystems. In this context, the objective of this study was to propose a predictive and explanatory model for soil $\delta^{15}N$ in the Cerrado biome. A total of 97 soil samples were collected from various locations within the biome. Values of climatic and biophysical predictor variables were extracted from secondary spatial data for each location, and soil variables were obtained from laboratory analyses. In total, 46 predictor variables of $\delta^{15}N$ were included in the database. A multiple linear regression model was fitted including all predictive variables with some significant correlation with the dependent variable. Subsequently, regression models were adjusted following the assumptions of regression analysis. Correlation matrices were used to eliminate correlated predictor variables, avoiding multicollinearity. The final model was adjusted after selecting the best variables using the stepwise AIC method. The residuals of the final model were tested for normality, homoscedasticity, autocorrelation, and spatial autocorrelation. Leave-One-Out cross-validation was applied to assess predictive performance. The first model (Model 1), adjusted with all variables, achieved an R^2_{adj} of 0,66 and an R^2 of 0,81 after cross-validation. Despite its high predictive power, Model 1 did not meet the assumptions of multiple linear regression, which hinders the interpretation of the model coefficients. The final model followed all assumptions of multiple linear regression and achieved an R^2_{adj} of 0,57 and an R^2 of 0,59 after cross-validation. Despite its lower predictive power, the coefficients of the model can be interpreted. The variables that contributed to the adjustment of the final model were Ca^{2+} concentration, soil $\delta^{13}C$, temperature seasonality, annual temperature range, and precipitation in the wettest quarter. All these variables were significant for the final model and exhibited coefficients consistent with their expected contribution to the variation of soil $\delta^{15}N$. The adjusted model brings advancements in the study of the nitrogen biogeochemical cycle for the Cerrado biome, as it incorporates soil data and indicates a positive and yet underexplored influence of Ca^{2+} on soil $\delta^{15}N$. Despite the progress, the final model may have excluded important predictor variables in explaining the phenomenon due to variable selection criteria and the method used. Future studies should apply methods that identify non-linear relationships among variables.

Keywords: nitrogen stable isotopes; edaphic variables; predictive model

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O nitrogênio (N) é um elemento básico para o funcionamento dos ecossistemas terrestres por ser um dos principais responsáveis pela produtividade primária, afetando toda a teia trófica. O ciclo do nitrogênio em ecossistemas terrestres é complexo, com diferentes caminhos de entrada (deposição seca e úmida, fixação biológica) e saída (lixiviação, volatilização), além de diferentes processos internos mibrobiológicos como mineralização, nitrificação e denitrificação. As atividades humanas estão modificando o ciclo do N, principalmente através da agricultura e queima de combustíveis fósseis (VITOUSEK *et al.*, 1997). Essas atividades modificam a quantidade de N reativo (N_r) nos sistemas naturais, alterando o funcionamento dos ecossistemas e, consequentemente, alterando serviços ecossistêmicos (COMPTON *et al.*, 2011; AUSTIN *et al.*, 2013). Diversos trabalhos apontam os isótopos estáveis de N ($\delta^{15}N$) do solo como uma variável integradora do ciclo de N (AMUNDSON *et al.*, 2003; CRAINE *et al.*, 2015; HOULTON *et al.*, 2015).

Isótopos são elementos que ocupam a mesma posição na tabela periódica, mas têm diferentes massas atômicas devido à presença de um maior número de nêutrons em seus núcleos. Os isótopos que não sofrem decaimento radioativo são considerados estáveis. O isótopo mais leve é mais prevalente na natureza em comparação com o isótopo mais pesado. Nas reações e processos químicos que ocorrem na natureza, as moléculas que contêm o isótopo mais pesado são discriminadas. Isso resulta em um maior enriquecimento do substrato com isótopos mais pesados em relação aos produtos das reações. No caso do N do solo, três processos se destacam pela maior discriminação isotópica: nitrificação, que deixa as moléculas de NH_4^+ enriquecidas na formação de NO_3^-; perda de N gasoso por volatilização de NH_3; nitrificação ou desnitrificação; e absorção do nitrogênio pelos fungos micorrízicos, que preferem o N mais leve.

A variação dos valores de $\delta^{15}N$ do solo depende de como os fatores de estado do ecossistema influenciam o estoque e a ciclagem de N. Na escala global, a variação espacial do $\delta^{15}N$ do solo se dá em função do teor de argila e concentração do carbono (C) orgânico do solo que, por sua vez, são variáveis controladas pelo clima (CRAINE *et al.,* 2015). Craine *et al.,* (2015) inferiram que em ambientes frios e/ou úmidos as perdas gasosas de N tendem a ser menores em comparação com ambientes frios e/ou secos, deixando-os com valores maiores de $\delta^{15}N$ do solo.

O $\delta^{15}N$ do solo é um integrador natural dos processos do ciclo do N. Os valores de $\delta^{15}N$ do solo variam de acordo com as entradas de N, transformações de N no solo e saída de N do

sistema. Entradas de N no sistema podem ocorrer, por exemplo, via fixação biológica de N e decomposição da matéria orgânica. A mineralização de N e a nitrificação são exemplos de transformações que ocorrem no solo e causam fracionamento. O fracionamento isotópico nas perdas de N do sistema ocorre na denitrificação e volatilização do N em formas gasosas (ROBINSON *et al.,* 2001).

Os fatores que influenciam a entrada, as transformações e as saídas de N têm potencial para afetar os valores de δ^{15}N do solo. As características físicas e químicas do solo e o padrão de vegetação exercem influência direta no δ^{15}N do solo. A textura do solo, por exemplo, pode determinar a quantidade de N que se perde para a atmosfera. A vegetação influencia os valores de δ^{15}N do solo por fixação de N e pela decomposição da serapilheira (BUSTAMENTE *et al.,* 2004). Assim, os fatores que influenciam diretamente as características do solo e da vegetação são consideradas fatores indiretos para a distribuição de δ^{15}N do solo, como a topografia e o clima. Esses fatores indiretos afetam a dinâmica da água no solo (SALEMI *et al.,* 2016) e os graus de intemperismo, influenciando variáveis importantes do solo, como pH e capacidade de troca de cátions (CTC).

Os padrões de distribuição de δ^{15}N do solo dependem da integração dos fatores diretos e indiretos. Na escala global, a distribuição dos valores de δ^{15}N do solo estão relacionadas com os fatores climáticos, como média anual de temperatura e precipitação anual (AMUNDSON *et al.,* 2003), e apresenta relação direta com características do solo, como textura (CRAINE *et al.,* 2015). Em pequenas áreas, os processos internos do ciclo do N, como mineralização, nitrificação e denitrificação, são responsáveis diretos pela distribuição dos valores de δ^{15}N. Um estudo recente mostrou que em ambiente de floresta tropical em montanhas a topografia é um fator determinante da distribuição do δ^{15}N do solo (WEINTRAUB *et al.,* 2015). Nesse caso, as variáveis físicas, a estabilidade e a idade do solo favorecem maiores taxas de denitrificação e perdas de N para a atmosfera nas áreas de menor declividade. Por outro lado, nas áreas com maior declividade as perdas sem fracionamento por erosão e lixiviação são favorecidas (WEINTRAUB *et al.,* 2015).

Diante do exposto, fica evidente que modelos que representam o ciclo do N incluindo o δ^{15}N do solo podem ser usados para auxiliar predições sobre mudanças ambientais globais, como problemas de eutrofização de ambientes aquáticos e emissão e sequestro de CO_2. A quantificação do ciclo de N usando δ^{15}N do solo também pode ajudar a diminuir incertezas em modelos globais de mudanças climáticas (HOULTON *et al.,* 2015). Além disso, modelos preditivos de δ^{15}N podem ser usados para rastrear movimento de animais (GARCIA-PEREZ

AND HOBSON, 2014; K. A. HOBSON *et al.*, 2012) e aplicados em estudos forenses (MALLETTE *et al.*, 2016).

Apesar da aplicabilidade potencial, compreender os fatores que causam a variação do δ^{15}N do solo nesses ambientes é difícil devido à complexidade das interações entre os fatores diretos e indiretos do ecossistema. Por isso, existem poucos trabalhos que ajustaram modelos preditivos para o δ^{15}N na escala regional em ambientes savânicos. Em estudo recente, Sena-Souza *et al.,* (2020) ajustaram um modelo de predição do δ^{15}N do solo feito para a América do Sul usando *Random Forest*. Embora consiga representar um recorte para a escala regional do Cerrado, o estudo tem abrangência continental, o que pode mascarar relações regionais (Sena-Souza *et al*., 2020). Neves *et al.,* (2021) demonstraram uma possível abordagem para o Cerrado ao ajustarem um modelo preditivo para isótopos estáveis de carbono (δ^{13}C) no solo para o bioma usando regressão linear múltipla. O modelo proposto permitiu uma análise da variação espacial dos processos que envolvem a dinâmica regional do carbono. A mesma abordagem poderia ser aplicada ao δ^{15}N para ajudar a explicar a dinâmica do nitrogênio em escala regional.

Além disso, a maioria dos modelos preditivos para isótopos estáveis são para fins de mapeamento o que faz com que as variáveis preditoras sejam extraídas de dados secundários espacialmente explícitos (BOWEN 2010). Porém, essa abordagem pode mascarar relações regionais do δ^{15}N do solo com elementos do ambiente que não são considerados em modelagens globais ou continentais (SENA-SOUZA *et al.* 2020). Vale destacar que o ajuste de modelos de regressão para predição geralmente busca um maior coeficiente de determinação e um menor erro médio dos resíduos, sem preocupação com a multicolinearidade de preditoras e com a confiabilidade dos coeficientes da regressão (HAWKINS 2004). Um modelo explicativo para o δ^{15}N do solo pode ajudar a explorar e descrever essas relações regionais, e posteriormente ser utilizado em modelos de predição que possam ser adotados em pesquisas aplicadas.

Considerando o contexto apresentado, o objetivo do presente trabalho é propor um modelo preditivo e explicativo para o δ^{15}N do solo do bioma Cerrado em escala regional, considerando variáveis preditoras edáficas, climáticas, biofísicas e geográficas.

1.1. Questões de pesquisa

1. Quais são os principais fatores responsáveis pela variabilidade do δ^{15}N do solo no Cerrado?
2. De que forma esses fatores influenciam a variabilidade do δ^{15}N do solo?

3. É possível criar um modelo de regressão linear múltipla para o δ^{15}N do solo no Cerrado com base nas amostras e variáveis disponíveis?

2. METODOLOGIA

2.1. Banco de dados

O banco de dados utilizado no presente trabalho é composto por 97 amostras de solo coletadas em fragmentos de vegetação natural ao longo do bioma Cerrado (Figura 1) e 46 variáveis preditivas, todas potencialmente relacionadas com a variável resposta do estudo (δ^{15}N do solo).

A variável resposta são os valores dos isótopos estáveis de nitrogênio (δ^{15}N) do solo, medido em laboratório. O valor isotópico dado na notação delta (δ) seguida por um expoente (x) que indica a massa atômica do isótopo mais pesado, seguido pela sigla do elemento químico (E), conforme Eq. 1.

$$\delta^{x}E = \left(\frac{R_{amostra}}{R_{padrão}} - 1\right) \times 1000 \qquad (1)$$

onde R indica a razão entre a concentração do elemento mais pesado e do elemento mais leve (para o nitrogênio: $R = {}^{15}N/{}^{14}N$); $R_{amostra}$ é a razão R medida na amostra; $R_{padrão}$ é a razão R de uma amostra padrão estabelecida internacionalmente, É um valor constante utilizado para padronizar a interpretação isotópica. O padrão para N é o ar atmosférico, com $R_{padrão} = 0{,}0036765$. Os valores de δ são muito próximos de zero, portanto, para facilitar a interpretação eles são multiplicados por 1000 e expressos em per mil (‰).

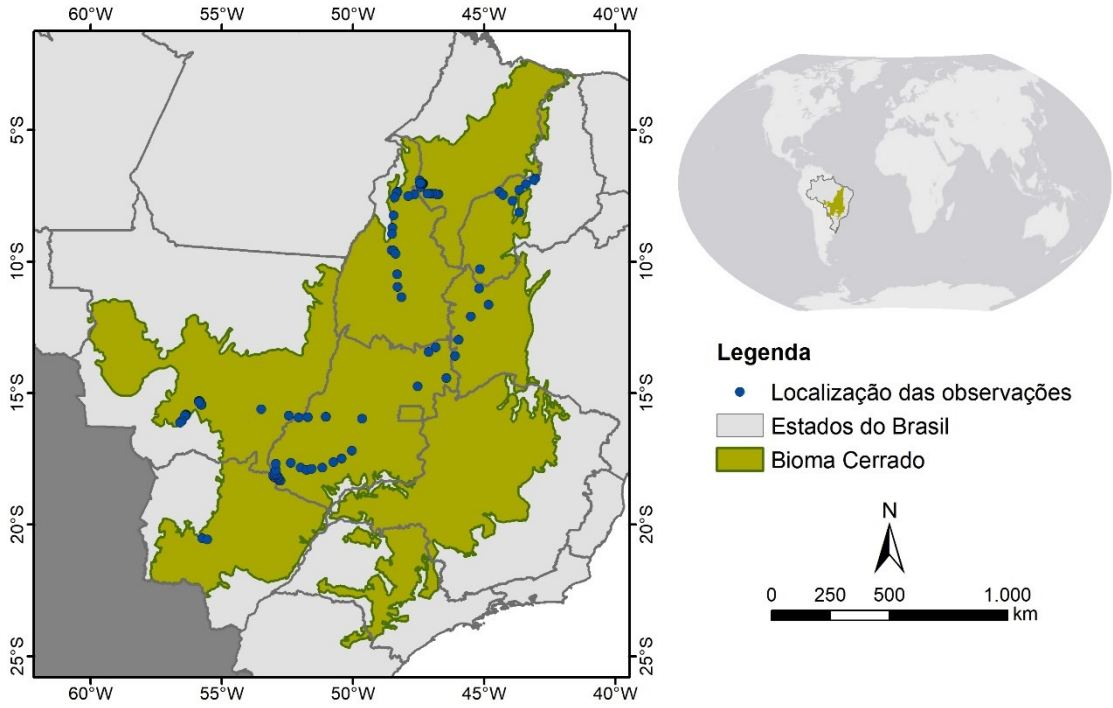

Figura 1. Localização das amostras (pontos azuis) em relação ao bioma Cerrado.

O banco de dados é composto por amostras de solos coletadas no contexto do projeto Origem, financiado pelo Edital CAPES 25/2014 – Pró-Forenses (parte dos dados foram usados em Sena-Souza et al., 2020 e estão disponíveis em https://data.mendeley.com/datasets/rhhp9x6bcd/1). As coletas ocorreram entre 2016 e 2019 em fragmentos de vegetação natural ao longo de rodovias e em unidades de conservação. As amostras foram retiradas do solo em um intervalo de profundidade de 0 a 20 cm usando um trado holandês para amostras deformadas. As amostras foram secas ao ar e peneiradas (malha de 2 mm). Subamostras de 30 a 35 mg de solo foram encapsuladas em estanho e inseridas em analisador elementar (Carlo Erba, Modelo 1110, Milão, Itália). No analisador, os produtos de combustão foram purificados em uma coluna de cromatografia gasosa e introduzidos diretamente em um espectrômetro de massa para análise isotópica (ThermoQuest-Finningan Delta Plus, Finnigan-MAT, California, USA). O erro analítico do laboratório é ± 0,30 ‰. Análises isotópicas das amostras foram realizadas no Laboratório de Ecologia Isotópica no Centro de Energia Nuclear em Agricultura (CENA), Universidade de São Paulo, Piracicaba, Brasil.

2.1.1. Variáveis preditoras

As variáveis preditoras foram divididas em quatro grupos: edáficas, climáticas, biofísicas e geográficas. As variáveis edáficas foram medidas em laboratório a partir das mesmas amostras de solo usadas para medir a variável resposta. Portanto, são dados primários. As variáveis climáticas, biofísicas e geográficas são provenientes de modelos espaciais disponíveis para *download* em formato matricial georreferenciado (raster), e seus valores foram extraídos dos pixels correspondentes à localização de cada amostra. A fonte bibliográfica e as unidades de medida de cada variável preditora estão listadas no APÊNDICE A. As variáveis preditoras estão listadas abaixo com suas respectivas siglas usadas na modelagem:

- Variáveis climáticas: Temperatura média anual (*temp_wc*); Média mensal da variação diária de temperatura (*bio2*); Isotermalidade (*bio3*); Sazonalidade de temperatura (*bio4*); Temperatura máxima do mês mais quente (*bio5*); Temperatura mínima do mês mais frio (*bio6*); Amplitude anual de temperatura (*bio7*); Temperatura média do trimestre mais úmido (*bio8*); Temperatura média do trimestre mais seco (*bio9*); Temperatura média do trimestre mais quente (*bio10*); Temperatura média do trimestre mais frio (*bio11*); Precipitação anual (*pre_wc*); Precipitação do mês mais chuvoso (*bio13*); Precipitação do mês mais seco (*bio14*); Sazonalidade de precipitação (*bio15*); Precipitação do trimestre mais úmido (*bio16*); Precipitação do trimestre mais seco (*bio17*); Precipitação do trimestre quente (*bio18*); Precipitação do trimestre mais frio (*bio19*).
- Variáveis biofísicas: Média de abril (1999-2017) do índice de vegetação de diferença normalizada (*ndvi_apr*); Média de setembro (1999-2017) do índice de vegetação de diferença normalizada (*ndvi_sep*); Produção primária líquida (*npp*); Produção primária bruta (*gpp*); Fração de radiação absorvida por atividade fotossintética (*fapar*); Conteúdo de água no solo a 1500 kpa (*soilwater1500*).
- Variáveis geográficas: Latitude (*lat*); Longitude (*long*); Altitude (*alt*).
- Variáveis edáficas: pH em água (*pH_H2O*); Fóstoro (*P*); Potássio (*K*); Cálcio (*Ca2*); Magnésio (*Mg2*); Alumínio trocável (*Al3_Altrocavel*); Al com extrator Acetato de Cálcio (*H_Al3*); Soma de bases trocáveis (*SB*); Capacidade de troca catiônica efetiva (*t*); Capacidade de troca catiônica a pH 7,0 (*T*); Índice de Saturação por Bases (*V*); Índice de Saturação por Alumínio (*m*); Fósforo remanescente (*P_rem*); Nitrogênio (*N*); Carbono (*C*); Razão Carbono Nitrogênio (*CN*); Isótopos estáveis de Carbono (*d13C*).

2.2. Análise de Regressão Linear

No presente estudo, um modelo de Regressão Linear Múltipla foi ajustado para predição do δ^{15}N do solo para o bioma Cerrado. Modelos de regressão são capazes de descrever relações lineares entre variáveis e estimar mudanças médias em uma variável dependente (y) a partir da variação observada em uma ou mais variáveis independentes (x). No caso de duas ou mais variáveis independentes, o modelo recebe o nome de Regressão Múltipla e é dado por

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \qquad (2)$$

onde y é a variável dependente independentes e x_1, x_2, ..., x_p são as varávies preditoras. β_0, também conhecido como o intercepto do modelo, representa o valor médio de y quanto todos os preditoreas são iguais a 0; β_i é o coeficiente de regressão para a i-ésima variável independente, sendo $i: 1, 2, \ldots, p$; ϵ é o erro do modelo (MONTGOMERY *et al*., 2012).

Para ajustar um modelo de Regressão Linear Múltipla deve-se assumir alguns pressupostos:

- Relação linear entre a variável dependente e as variáveis independentes;
- Ausência de multicolinearidade, ou seja, as variáveis independentes x_1, x_2, ..., x_p não devem ser correlacionadas;
- Homocedasticidade, ou seja, o resíduo ϵ da regressão deve ser uma variável aleatória de média $\mu = 0$ e variância σ^2 constante;
- Normalidade do erro, ou seja $\epsilon \sim N(0, \sigma^2)$;
- Independência do erro.

A análise de regressão é realizada através do método dos Mínimos Quadrados Ordinários, que visa determinar uma reta de regressão que passe o mais próximo possível de todos os valores da variável dependente. Ou seja, esse método encontra a equação ótima com o menor ϵ possível (MONTGOMERY *et al*., 2012).

2.2.1. Suposição de correlação linear entre variável dependente e variáveis independentes

Na primeira etapa do trabalho foram estudadas as correlações lineares entre a variável resposta e cada uma das variáveis explicativas. Foram realizados testes de normalidade de Shapiro-Wilk para a variável dependente e todas as variáveis independentes (SHAPIRO e

WILK, 1965). Vale destacar que o teste de normalidade feito para as variáveis preditoras não constitui um pressuposto do modelo de regressão, e foi realizado apenas para definir o teste de significância a ser realizado para as correlações.

Para identificar as variáveis independentes significativamente correlacionadas com a variável dependente, foram realizadas análises de correlação de Pearson, com variáveis que seguem uma distribuição normal; e de Spearman para aquelas que não apresentam distribuição normal. O coeficiente de correlação de Pearson (r) (Eq. 3) indica o grau de correlação linear entre duas variáveis quantitativas com distribuição normal e é calculado por meio da padronização da covariância entre as duas variáveis por seus respectivos desvios.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (3)$$

onde x_i e y_i são os valores das observações individuais das variáveis x e y; $\bar{x}$ e $\bar{y}$ são as médias das variáveis x e y, respectivamente.

Para as variáveis independentes que não apresentaram distribuição normal foi aplicada a análise de correlação não paramétrico de Spearman, que avalia a relação entre duas variáveis transformadas em ordinais. A análise de correlação de Spearman atribui um ranking a cada uma das observações em cada variável. Os rankings são determinados pela classificação das observações em ordem crescente, atribuindo a cada observação o número correspondente ao seu posto na sequência classificada. Posteriormente, o coeficiente de correlação de Spearman (r_s) é calculado pela Eq. 4.

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)} \quad (4)$$

onde, $\sum d^2$ é a soma dos quadrados das diferenças entre os rankings das duas variáveis e n é o número de pares de observações.

Foi utilizada a estatística de teste *t* para determinar a significância das análises de correlação linear. Foram incluídas no modelo de regressão linear múltipla todas as variáveis independentes que apresentaram correlação linear significativa com o δ^{15}N do solo, ao nível de 5% de significância.

2.2.2. Suposição de ausência de multicolinearidade entre variáveis independentes

Modelos de regressão múltipla com multicolinearidade entre variáveis independentes diminuem a confiabilidade dos estimadores. Para garantir a ausência de multicolinearidade no modelo, as 46 variáveis independentes foram divididas em subconjuntos de variáveis (variáveis climáticas, biofísicas, geográficas e edáficas). Cada subconjunto foi submetido a uma análise de correlação linear par a par e dispostos em matrizes de correlação. Variáveis preditoras com correlação $r \geq |0{,}70|$ foram excluídas de cada subconjunto de variáveis. Para cada para de variáveis com correlação maior que o limite estabelecido, a variável mantida foi escolhida considerando sua influência potencial na variável dependente de acordo com a literatura especializada. Em seguida, os subconjuntos sem variáveis correlacionadas foram agrupados novamente em um conjunto de dados com todas as variáveis remanescentes. Esse novo conjunto de dados passou por uma nova matriz de correlação e busca por variáveis correlacionadas. Nesta etapa, variáveis preditoras que apresentaram $r \geq |0{,}70|$ com variáveis de outros subconjuntos também foram excluídas.

2.2.3. Ajuste dos modelos de regressão linear múltipla

O ajuste do modelo final preditivo para o δ^{15}N do solo passou por etapas que envolveram ajustar modelos de regressão anteriores. As variáveis que apresentaram uma correlação com a variável dependente considerando um nível de significância de 5% foram incluídas no primeiro modelo de regressão linear múltipla (*Modelo 1*). Portanto, o *Modelo 1* não considerou a suposição de ausência de multicolinearidade. Este primeiro modelo serviu apenas para comparação com os demais modelos e para observar a capacidade preditiva das variáveis independentes para o δ^{15}N do solo, desconsiderando a interpretabilidade dos estimadores do modelo.

No *Modelo 2* de regressão linear múltipla ajustado foram inseridas todas as variáveis independentes não correlacionadas entre si que apresentaram correlação linear significativa com o δ^{15}N do solo, ao nível de 5% de significância, respeitando assim as duas primeiras suposições. Para reforçar a suposição de ausência de multicolinearidade, foi calculado o Fator de Inflação da Variância (VIF) do *Modelo 2*. O VIF detecta a multicolinearidade entre as variáveis preditoras de um modelo. O VIF representa o incremento da variância devido à presença de multicolinearidade (Montgomery *et al.* 2012) e é calculado por

$$VIF_k = \frac{1}{1 - R_k^2} \qquad k = 1, 2, \ldots, p \tag{5}$$

onde p é a quantidade de variáveis preditoras; R_k^2 é o coeficiente de determinação da regressão entre a variável independente X_k e todas as outras variáveis independentes. Foram desconsideradas as preditoras com *VIF > 10*.

Para a definição do *Modelo final*, o *Modelo 2* passou por um processo de seleção de variáveis por meio do método *Stepwise* AIC (YAMASHITA *et al.*, 2007). Este método gera modelos adicionando uma variável preditora significativa x_p em um modelo e compara o valor do critério de todos os modelos anteriores. Então, adiciona a variável x_p ao modelo se ela fornecer o melhor valor do critério AIC. O AIC é dado por

$$AIC = 2k - 2\,ln(L) \tag{6}$$

em que AIC é o Akaike Information Criterion; k é o número de parâmetros do modelo, incluindo o intersepto; *ln* é a função logaritmo natural; *L* é a verossimilhança do modelo. Essa escolha pode funcionar por método de seleção (*forward*) ou por método de remoção (*backward*) de variáveis (YAMASHITA *et al.*, 2007). O objetivo é encontrar um modelo com o menor valor de AIC possível, o que indica um melhor ajuste aos dados com um número mínimo de variáveis.

2.2.4. Suposições envolvendo os erros do modelo de Regressão Linear Múltipla

Os resíduos do *Modelo final* foram submetidos à análise gráfica para a avaliação dos pressupostos da regressão linear múltipla. Além disso, os resíduos foram testados para homocedasticidade, normalidade, independência dos erros e autocorrelação espacial, uma vez que os dados apresentam localização espacial conhecida.

Para testar a suposição de normalidade dos resíduos, foi aplicado o método de Shapiro-Wilk, que testa a hipótese nula que uma amostra aleatória segue uma distribuição normal através da estatística W apresentada na Eq. 7.

$$W = \frac{\left(\sum_{i=1}^{n} a x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{7}$$

onde W é a estatística de teste, n é o tamanho da amostra, $x_{(i)}$ é o i-ésimo menor valor observado, $\bar{x}$ é a média da amostra e a são os coeficientes calculados a partir das médias, covariâncias e variâncias teóricas da distribuição normal padrão. O valor da estatística W é comparado com valores críticos tabelados para determinar se a amostra segue uma distribuição normal.

A distribuição da soma dos quadrados e das estatísticas dos testes de hipóteses da regressão linear múltipla dependem da suposição de homocedasticidade. Isso porque quando os resíduos são heterocedásticos, os estimadores de mínimos quadrados ordinários (MQO) dão mais peso para as amostras com maior variância. O teste de homocedasticidade testa a hipótese nula de que a variância dos erros é constante. Aqui, foi aplicado o teste de Breusch-Pagan (BP), que usa a estatística LM calculando sua significância pela distribuição qui-quadrado.

$$LM = n \, . R^2_{\hat{\epsilon}^2} \sim \chi^2 \tag{8}$$

onde n é o número de observações e $R^2_{\hat{\epsilon}^2}$ é o coeficiente de determinação obtido a partir de uma regressão auxiliar dos resíduos ao quadrado ($\hat{\epsilon}^2$) em relação às variáveis independentes. A estatística *LM* é então comparada ao valor crítico da distribuição qui-quadrado (χ^2) para determinar se há evidências de heterocedasticidade.

Por se tratar de dados com distribuição espacial conhecida, a independência dos resíduos do modelo foi testada a partir da autocorrelação espacial, usando o Índice Global de Moran. O índice de Moran calculado para os resíduos do *Modelo final* é dado por

$$I = \frac{n}{\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\epsilon_i - \bar{\epsilon})\left(\epsilon_j - \bar{\epsilon}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \tag{9}$$

em que ϵ_i é o valor do resíduo do modelo na localidade da amostra *i* e ϵ_j é o valor do resíduo do modelo na localidade da amostra *j*; o peso w_{ij} é dado pela distância entre os pares de amostras (*i, j*).

2.3. Avaliação da performance do modelo

Considera-se que a Soma dos quadrados totais (SQT) = Soma dos quadrados do modelo (SQM) + Soma dos quadrados dos resíduos (SQR), ou seja

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2 \quad (10)$$

Onde y_i é o valor observado da variável dependente para cada observação *i*, $\bar{y}$ é a média da amostra, $\hat{y}$ é o valor predito pelo modelo e *n* é o número de observações.

Assim, é possível calcular o Coeficiente de Determinação (R^2), que mostra o efeito das variáveis independentes (*x*) na explicação da variabilidade da variável dependente (*y*):

$$R^2 = \frac{SQM}{SQT} = \frac{SQT - SQR}{SQT} = 1 - \frac{SQR}{SQT} \quad (11)$$

A partir do R^2 é calculado o coeficiente de determinação ajustado (R^2_{adj}), que minimiza o efeito da quantidade de covariáveis inseridas no modelo. O R^2_{adj} é dado por

$$R^2_{adj} = 1 - \left[\left(\frac{n-1}{n-p-1}\right)\right](1 - R^2) \quad (12)$$

em que p é o número de variáveis independentes e n é o número de observações da amostra.

Para testar o poder de predição do modelo foi realizada uma validação cruzada *Leave-One-Out* (LOOCV). Esse método de validação cruzada ajusta o modelo de regressão usando todas as observações, exceto a observação *i*. Porteriormente, realiza a predição para a observação *i* usando o modelo ajustado e calcula o resíduo para a observação *i* como a diferença do valor predito $\hat{y}_i$ pelo valor observado y_i, repetindo esse procedimento até contemplar todas as observações do banco de dados. Por fim, é calculado o erro médio quadrático (MSE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (13)$$

em que *n* é o número total de observações no conjunto de dados; y_i é o valor real da variável de resposta para a observação *i*; $\hat{y}_i$ é o valor previsto da variável de resposta para a observação *i* usando o modelo ajustado sem esta observação. Outra medida de erro para descrever a performance dos modelos foi a *Root Mean Squared Error* (*RMSE*), que é calculada a partir da raiz quadrada do *MSE*. É uma métrica comum para medir o quão próximo as previsões do modelo estão em relação aos valores reais. O *RMSE* é expresso na mesma unidade da variável

resposta e fornece uma medida absoluta do desvio do modelo. Quanto menor o valor do *RMSE*, melhor o ajuste do modelo aos dados, indicando que as previsões estão mais próximas dos valores reais.

Todas as análises estatísticas foram realizadas no programa R v. 4.2.2 (R Core Team 2022). O gráfico das suposições do modelo foi elaborado por meio da função *check_model* do Pacote *performance* (LÜDECKE *et al.,* 2021). O *script R* está disponível no APÊNDICE D.

3. RESULTADOS

3.1. Análise descritiva da variável resposta

Os valores observados de δ^{15}N do solo das amostras se aproximam de uma distribuição normal, conforme apontado pelo teste de Shapiro-Wilk (W = 0,985; *p-valor* = 0,35) (Figura 2).

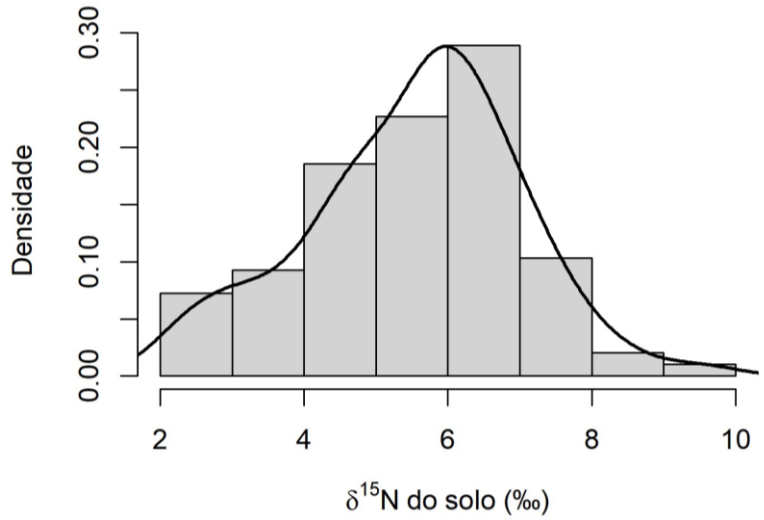

Figura 2. Distribuição de densidade dos valores observados de δ^{15}N do solo das amostras.

Os valores medidos de δ^{15}N do solo apresentaram uma amplitude de 7,48 ‰. O valor mínimo medido foi 2,09 ‰ e o máximo foi de 9,57 ‰, com uma média de 5,53 ‰, muito próximo da mediana 5,71 ‰. As medidas resumo do δ^{15}N do solo estão dispostas na Tabela 1.

Tabela 1. Medidas resumo da variável dependente.

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	Desvio Padrão
2,09	4,60	5,71	5,53	6,46	9,57	1,49

3.2. Seleção de variáveis preditoras

3.2.1. Variáveis preditoras correlacionadas com o δ^{15}N do solo

Os testes de Correlação de Pearson realizados entre a variável resposta e cada uma das cinco variáveis preditoras que apresentaram distribuição normal indicaram apenas três variáveis com relação significativa: *bio3* (r = -0,40*; p-valor* < 0,05), *ndvi_sep* (r = -0,22*; p-valor* < 0,05) e *soilwater1500* (r = 0,20*; p-valor* < 0,05).

Entre as 41 variáveis preditoras que não apresentaram distribuição normal, 28 apresentaram correlação de Spearman significativa com o δ^{15}N do solo (Figura 3). A variável *bio4* é a preditora com maior força de correlação com δ^{15}N do solo do Cerrado (r_s = 0,63; *p-valor* < 0,05). A latitude também apresenta uma correlação significativa com o δ^{15}N do solo (r_s = -0,49; *p-valor* < 0,05). A variável edáfica com maior correlação com δ^{15}N do solo é o teor de carbono do solo (r_s = 0,63; *p-valor* < 0,05). A produção primária líquida (*npp*) foi a variável biofísica com maior coeficiente de correlação de Spearman com o δ^{15}N (r_s = 0,34 *p-valor* < 0,05).

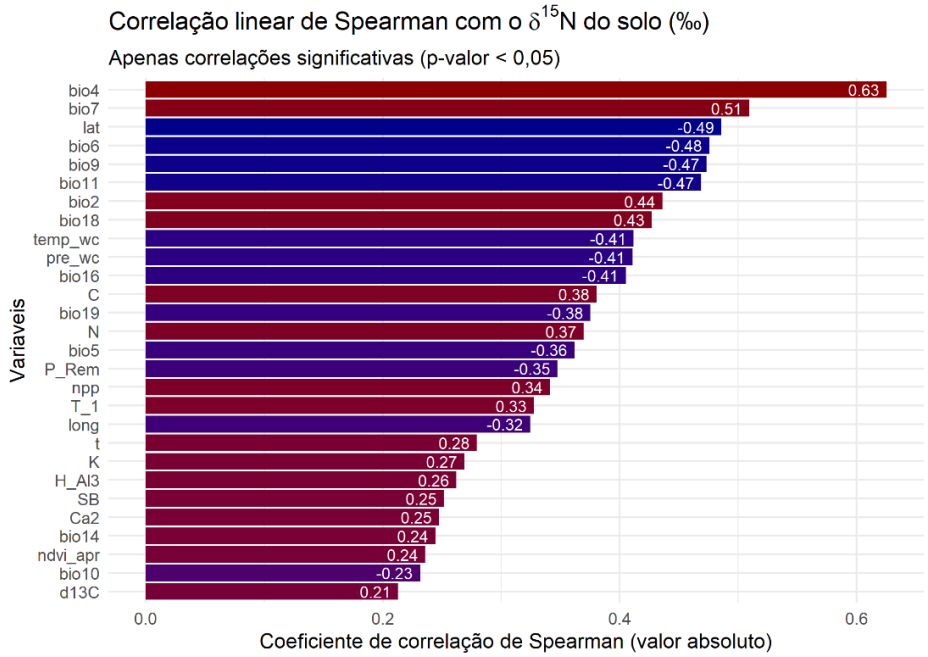

Figura 3. Correlação de Spearman entre as vaiáveis preditoras e o δ^{15}N do solo. O gráfico mostra apenas as correlações significativas (*p-valor* < 0,05).

3.2.2. Detecção e eliminação de multicolinearidade

As matrizes de correlação para cada grupo de variáveis e a matriz de correlação das variáveis selecionadas podem ser acessadas no APÊNDICE B. Os grupos de variáveis com maior presença de correlação forte foram os das variáveis climáticas e edáficas. A maioria das variáveis climáticas são derivadas da temperatura média ou da precipitação anual. Por isso apresentam muitas variáveis correlacionadas (Figura A1). Das 15 variáveis climáticas, nove foram excluídas por apresentarem $r \geq |0{,}70|$. O subconjunto de variáveis climáticas selecionadas nesta etapa está representado na matriz de correlação da Figura 4.

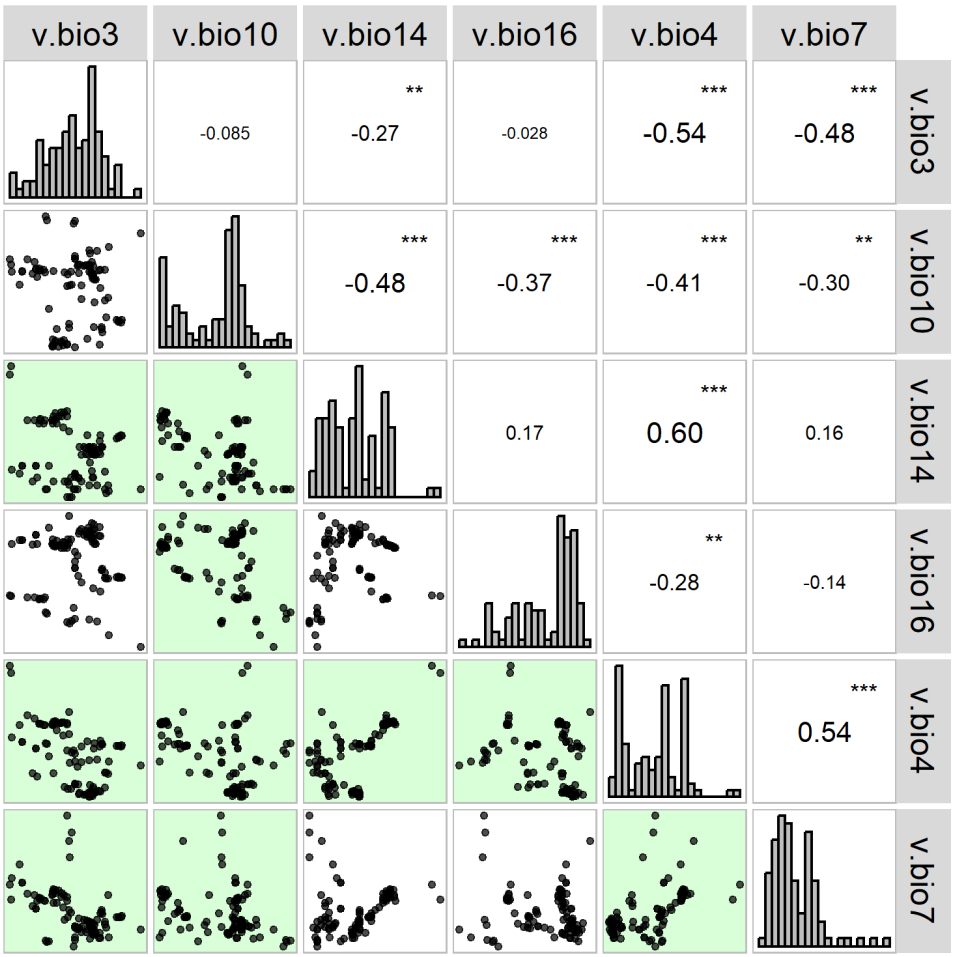

Figura 4. Matriz de correlação com as variáveis climáticas selecionadas.

As variáveis edáficas apresentaram alta colinearidade pois muitas delas dependem dos mesmos processos que ocorrem no solo. Por exemplo, teores de carbono e nitrogênio dependem dos teores de matéria orgânica do solo e da dinâmica de decomposição da matéria orgânica, por isso espera-se que apresentem forte correlação. Das 12 variáveis edáficas, sete foram excluídas

por apresentarem $r \geq |0{,}70|$. O subconjunto de variáveis edáficas selecionadas nesta etapa está representado na matriz de correlação da Figura 5.

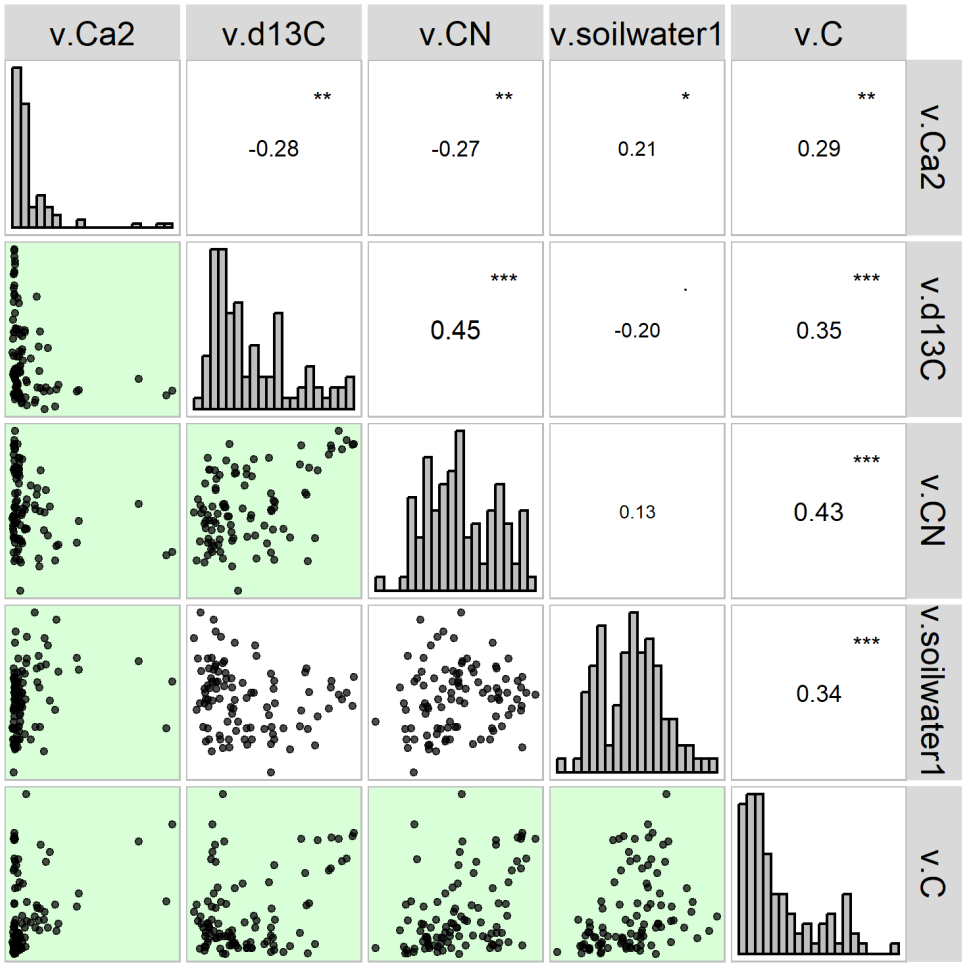

Figura 5. Matriz de correlação com as variáveis edáficas selecionadas.

O conjunto de dados contendo as variáveis selecionadas em cada grupo também passou por uma matriz de correlação para identificação de multicolinearidade. No total, foram 16 variáveis preditoras testadas. Destas, três foram eliminadas por apresentarem $r \geq |0{,}70|$ com outras variáveis. Foram eliminadas a longitude (*long*) e a latitude (*lat*), que apresentaram alta correlação com as variáveis climáticas; e o *npp*. A matriz de correlação com as variáveis selecionadas pode ser acessada na Figura A3, no APÊNDICE B.

3.3. Modelo de regressão linear múltipla

O *Modelo 1* incluiu todas as preditoras que apresentaram correlação significativa com δ^{15}N do solo, ainda desconsiderando a suposição de ausência de multicolinearidade do modelo de regressão linear múltipla. Esse primeiro modelo serviu apenas para comparação com os

modelos ajustados posteriormente. O *Modelo 1* teve o R^2 de 0,81, com um R^2_{adj} de 0,66 (*p-valor < 0,001*). Os coeficientes da regressão, a tabela da análise de variância e os parâmetros do *Modelo 1* podem ser encontradas no APÊNDICE C (Tabela A2).

O *Modelo 2* foi ajustado após exclusão das variáveis preditoras que apresentaram multicolinearidade. O segundo modelo com as 13 variáveis teve um R^2 de 0,61 e R^2 ajustado de 0,55. Os coeficientes da regressão, a tabela da análise de variânvia e os parâmetros do *Modelo 2* também podem ser encontrados no APÊNDICE C (Tabela A2). O *Modelo final* foi gerado a partir do método *Stepwise* AIC aplicado ao *Modelo 2*. O método selecionou as seguintes variáveis: *bio3, bio16, Bio4, bio7, d13C e Ca2*. A variável *bio3* apresentou um nível de significância elevado (*p-valor* > 0,14) após seleção de variáveis por *Stepwise* AIC e foi excluída do modelo final.

A variável *bio16* representa a precipitação do trimestre mais úmido e é dado em milímetros (mm) de chuva. A média desta variável foi 714,4 mm, com um desvio padrão de 100,7 mm. As variáveis *bio4* representa a sazonalidade da temperatura do local da amostra, dada em °C*100 (ver APÊNDICE A). O valor médio foi de 115,1, com um desvio padrão de 44,51. A variável *bio7* é a medida da amplitude de temperatura para cada local amostrado. A menor amplitude de temperatura anual entre os pontos de coleta foi 16,1 °C. A maior amplitude de temperatura anual foi de 22,47 °C. A variável *Ca2* é o valor da concentração do íon Ca_2^+ medido nas amostras de solo. O valor mínimo encontrado foi 0, com uma média de 0,71 cmol/dm^3. Com um valor máximo de 7,65 cmol/dm^3 e uma média muito maior que a mediana, é evidente que essa concentração apresenta naturalmente alguns valores extremos (*outliers*) devido à natureza da variável. Portanto, foi transformada em logaritmo natural no modelo. O *d13C*, que representa o δ^{13}C do solo, apresentou uma média de -23,7 ‰, com valore mínimo de -28,1 ‰ e máximo de -14,6 ‰. As medidas resumo das variáveis preditoras usadas no *Modelo final* são apresentadas na Tabela 2.

Tabela 2. Medidas resumo das variáveis preditoras usadas no *Modelo final.*

Variável	Mínimo	1º Quartil	Mediana	Média	3ºQuartil	Máximo	Desvio Padrão
bio16	435,10	656,00	763,90	714,40	793,00	860,10	100,69
bio4	54,95	69,12	123,71	115,09	148,18	251,40	44,51
bio7	16,10	16,98	17,51	17,76	18,43	22,47	1,14
Ca2	0,00	0,10	0,28	0,71	0,66	7,65	1,31
d13C	-28,11	-26,55	-24,95	-23,73	-21,60	-14,60	3,57

O modelo final apresentou R^2 de 59,31% e R^2_{adj} de 57,07%. As variáveis *bio4, bio7, Ca2* e *d13C* foram fatores associados positivamente com o δ^{15}N do solo. A variável *bio7*, por exemplo, apresentou um coeficiente de 0,26 no modelo final (Tabela 3). Isso indica que a cada 1 °C de aumento da amplitude anual da temperatura, espera-se um aumento médio de 0,26 ‰ no δ^{15}N do solo, considerando constante as demais variáveis independentes. Apenas a variável *bio16* apresentou uma associação negativa. O efeito parcial de *bio16* no modelo indica que, mantendo as demais variáveis constantes, um aumento de 1 mm de precipitação no trimestre mais úmido diminui em média 0,004 ‰ no δ^{15}N do solo. Todas as variáveis do modelo final apresentaram um nível de significância *p-valor* < 0,05 e o VIF < 2 (Tabela 3).

Tabela 3. Coeficientes da regressão, tabela de análise de variância e os parâmetros do modelo final para o δ^{15}N do solo do Cerrado.

X	Coef.	Sd. Error	t	Pr(>|t|)	g.l	Sum of Sq.	RSS	AIC	F	Pr(>F)	VIF
(Intercept)	5,810	2,11	2,75	0,01			87,03	1,47			
bio16	-0,004	0,00	-3,45	0,00	1	11,40	98,42	11,41	11,92	0,00	1,19
bio4	0,010	0,00	3,60	0,00	1	12,42	99,45	12,42	12,99	0,00	1,64
bio7	0,259	0,10	2,49	0,01	1	5,95	92,97	5,89	6,22	0,01	1,42
log1p(Ca2)	1,108	0,25	4,50	0,00	1	19,33	106,36	18,93	20,22	0,00	1,25
d13C	0,162	0,03	5,05	0,00	1	24,38	111,41	23,43	25,49	0,00	1,33

3.4. Análise dos resíduos do modelo final

A avaliação gráfica do modelo final indica que as suposições da regressão linear múltipla foram alcançadas (Figura 6). Nota-se que a distribuição dos valores preditos pelo modelo se aproxima da distribuição dos dados observados. O modelo parece ser homocedástico e o gráfico *qqplot* indica que os resíduos seguem uma distribuição normal.

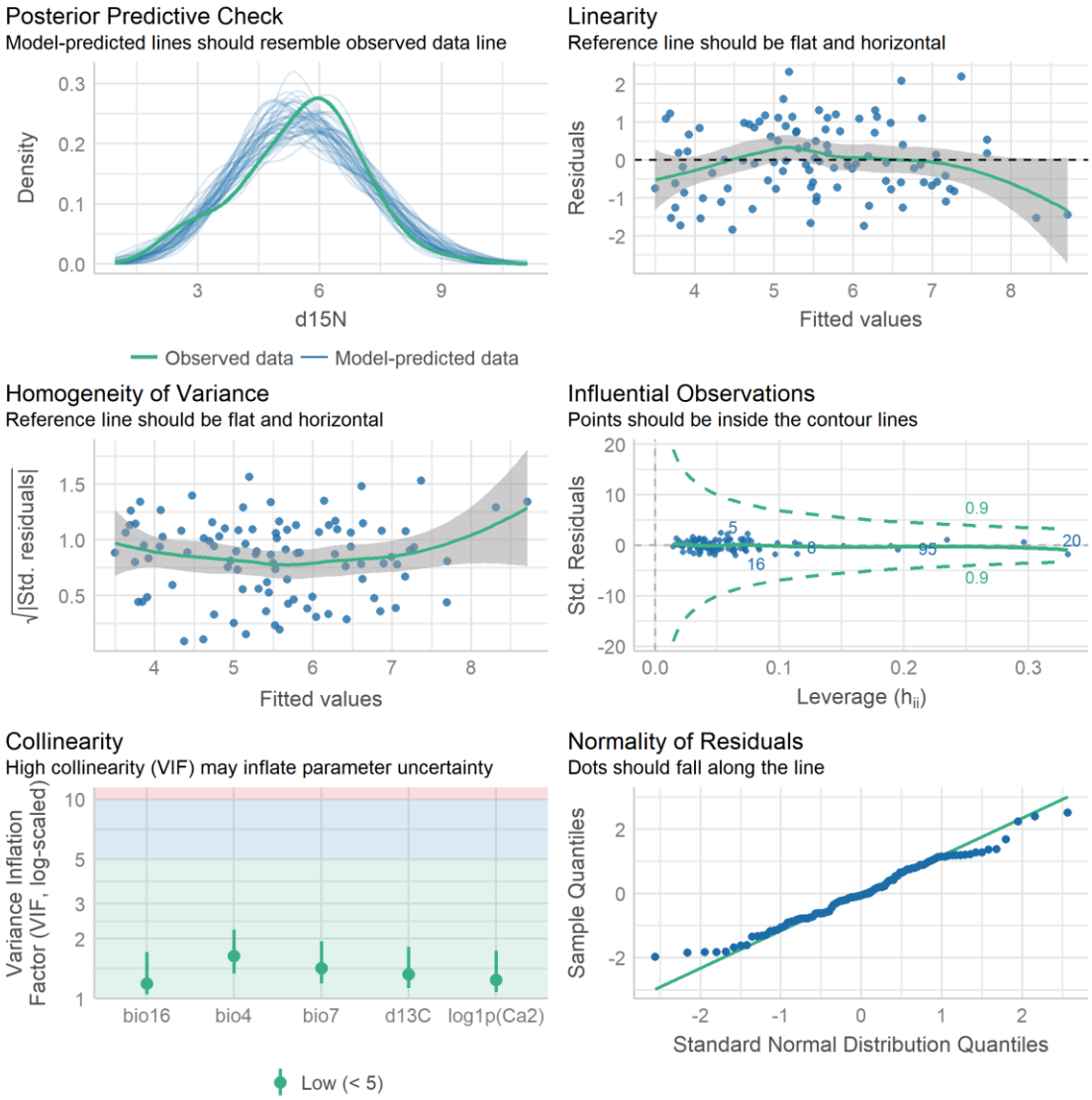

Figura 6. Avaliação visual dos pressupostos do modelo de regressão linear múltipla aplicado na predição de δ^{15}N do solo no Cerrado.

A Tabela 4 apresenta os valores resumo dos resíduos, que apresentaram média zero, muito próxima da mediana (-0,06). O resíduo máximo gerado pelo modelo final foi 2,35 e o mínimo -1,84 (Tabela 4). O teste de Shapiro-Wilk não apresentou evidência estatística para rejeitar a hipótese de normalidade (W = 0,9816; *p-valor* = 0,19). Portanto, o modelo atende a suposição de normalidade dos resíduos.

Tabela 4. Medidas resumo dos resíduos do modelo de regressão linear múltipla.

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
-1.84	-0,73	-0,06	0,00	0,75	2.32

O teste de Breusch-Pagan indicou que não há evidência estatística para rejeitar a hipótese nula de homoscedasticidade (*BP* = 3,49; *gl* = 6; *p-valor* = 0,62). Portanto, o modelo respeitou a suposição de homocedasticidade. O Índice Global de Moran indicou ausência de autocorrelação espacial entre os resíduos (*Morans'I* = 0,02; *p-valor* = 0,40). A Figura 7 mostra o Índice Local de Moran por classes de distância e um mapa mostrando o padrão espacial dos resíduos. Nota-se a ausência de um padrão espacial dos resíduos.

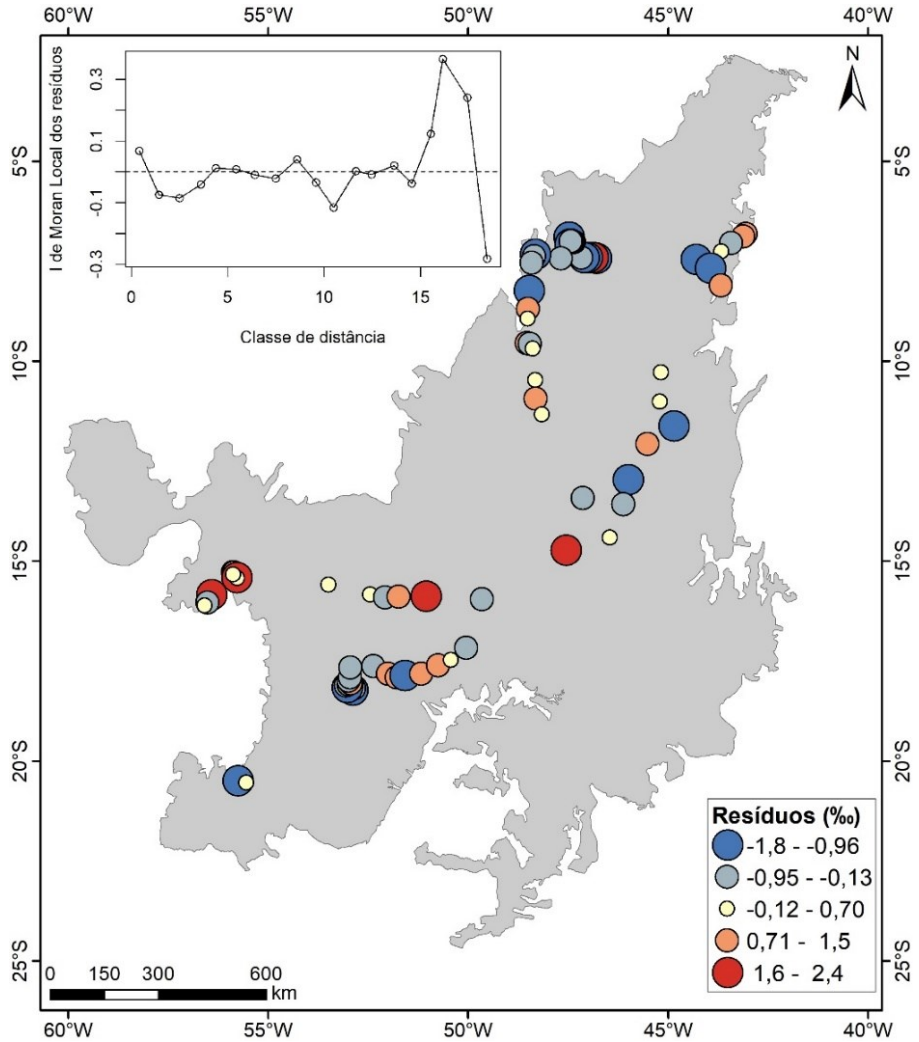

Figura 7. Mapa da distribuição espacial dos resíduos e correlograma.

3.5. Avaliação da performance preditiva do modelo final por Validação Cruzada Leave-One-Out

A performance preditiva do modelo linear múltiplo final após validação cruzada LOO obteve um R^2 menor do que os modelos gerados nas etapas anteriores do *Stepwise* AIC. Também apresentou um maior RMSE (Figura 8). O *Modelo 1* teve um R^2 preditivo de 0,81 e um RMSE de 0,63. O *Modelo 2* gerou o R^2 preditivo de 0,61, com RMSE de 0,92. O modelo final teve R^2 da validação cruzada de 0,59, com RMSE de 0,94 (Tabela 5).

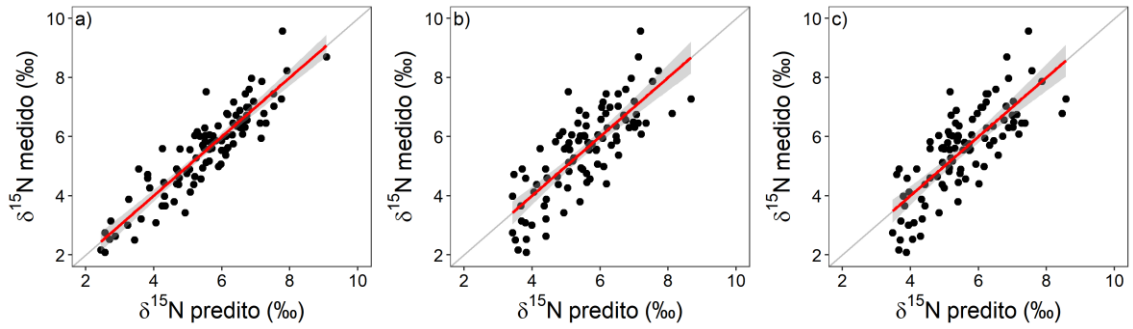

Figura 8. Gráfico de dispersão mostrando os valores observados de $\delta^{15}N$ e os valores preditos após a LOOCV: a) performance preditiva do Modelo 1; b) performance preditiva do Modelo 2, após exclusão de variáveis pelo VIF; c) validação cruzada do Modelo final, com variáveis selecionadas pelo método sepwise AIC. A linha cinza representa a linha 1:1.

A avaliação dos parâmetros do modelo (Tabela 5) mostrou uma performance de predição inferior do *modelo final* em relação aos modelos 1 e 2 quando considerados o R^2 e o *RMSE*. Porém, o *modelo final* obteve um o R^2_{adj} maior que o *Modelo* 2 a menor diferença entre o R^2 e o R^2_{adj}, indicando uma menor influência da quantidade de variáveis na performance do modelo. Mesmo com uma quantidade muito inferior de variáveis e seguindo todos os pressupostos da regressão linear múltipla, o *modelo final* apresentou ainda o valor de AIC um muito próximo que o *Modelo 1*.

Tabela 5. Performance dos modelos de regressão gerados no processo e do modelo final.

Modelo	***AIC***	***R^2***	***R^2_{adj}***	***RMSE***	***F***	***p-valor***
Modelo 1	277,56	0,81	0,66	0,63	5,48	0,00
Modelo 2	297,19	0,61	0,55	0,92	10,19	0,00
Modelo final	278,74	0,59	0,57	0,94	26,53	0,00

4. DISCUSSÃO

4.1. Desempenho dos modelos

A comparação da performance mostra que o *Modelo 1* teve um desempenho superior que os *Modelos 2* e *final* (Figura 8; Tabela 5). Entretanto, o *Modelo 1* foi usado apenas como parte do processo exploratório dos dados. Ele poderia ser usado para a predição dos valores da variável dependente, mas os coeficientes ajustados no *Modelo 1* não são confiáveis. A Figura 8a mostra que o *Modelo 1* foi capaz de realizar predição de valores máximos e mínimos do banco de dados com mais precisão em comparação com os *modelos 2* e *final* (Figura 8bc). Esse fato pode indicar que os critérios de seleção de variáveis usados aqui pode ter excluído uma ou mais variáveis independentes capazes de a explicar os valores extremos da variável dependente. Entretanto, essa performance também pode caracterizar *overfitting* por violar o princípio da parcimônia incluindo mais variáveis que o necessário ou aumentando a complexibilidade e flexibilidade para gerar um melhor desempenho (HAWKINS, 2004). Isso ocorre porque um modelo com mais variáveis tem maior flexibilidade para capturar padrões complexos no banco de dados (RENCHER e PUN, 1980; HAWKINS, 2004). Porém, modelos com muitas variáveis correm o risco de se ajustar excessivamente aos dados, capturando não apenas os padrões reais, mas também o ruído ou flutuações aleatórias.

Modelos que apresentam esse tipo de *overfitting* podem ser sensíveis a variações aleatórias nos dados de entrada. Se essas variações forem específicas para o conjunto de dados de entrada, isso pode tornar o modelo inadequado para a predição de novos dados, ou seja, o modelo pode não generalizar bem para novas amostras e atualizações do banco de dados (HAWKINS, 2004). Modelos múltiplos com excesso de variáveis também podem aumentar a complexidade da interpretação e diminuir a confiabilidade dos coeficientes da regressão (RENCHER e PUN, 1980). Mesmo que tenham um bom desempenho em predições, esses modelos não permitem a compreensão dos fatores que realmente estão impulsionando as previsões.

Para evitar *overfitting* e permitir a quantificação do efeito de cada variável preditora no modelo final, foi necessário um processo de seleção de variáveis preditoras que excluiu as que apresentaram colinearidade. Posteriormente, o método *stepwise AIC* usado para ajustar o Modelo final buscando selecionar um subconjunto das variáveis independentes com um equilíbrio entre o ajuste do modelo e a complexidade do modelo. O método *stepwise AIC* também pode ajudar a explicar o menor R^2 do modelo final. Durante o processo de escolha das variáveis, o método pode excluir algumas variáveis que têm um pequeno impacto na variável

dependente, resultando em um valor de R^2 menor do que um modelo com todas as variáveis incluídas (YAMASHITA *et al.*, 2007).

4.2. Explicações sobre a variabilidade do $\delta^{15}N$ do solo do Cerrado a partir do modelo final

O modelo final ajustado aos dados utilizou seis variáveis preditoras selecionadas entre as 48 presentes no banco de dados. As variáveis selecionadas foram *bio16, bio4, bio7, d13C e Ca2*. O modelo selecionado apresentou um coeficiente de determinação de 0,59 e um RMSE de 0,94. Ou seja, as seis variáveis preditoras conseguiram explivar aproximadamente 59% da variância do $\delta^{15}N$ do solo nas observações. Embora pareça uma baixa performance, esses valores podem ser considerados elevados considerando a complexidade da variável predita e a baixa quantidade de amostras para representar o fenômeno. Estudos prévios de predição de isótopos estáveis de nitrogênio já se contentaram com performances semelhantes ou inferiores (SENA-SOUZA *et al.*, 2020; CRAINE *et al.*, 2015; AMUNDSON *et al.*, 2003).

Uma das vantagens do modelo ajustado neste estudo em relação à trabalhos anteriores foi a inclusão de variáveis edáficas provenientes de dados primários. Isso porque modelos preditivos de isótopos estáveis geralmente buscam aplicar a predição na elaboração de um mapa e usam modelos espaciais de dados secundários (NEVES *et al.*, 2021; SENA-SOUZA *et al.*, 2019; 2020). A inclusão de variáveis preditoras medidas nas mesmas amostras de solo podem ajudar a compreender os mecanismos do fracionamento isotópico do $\delta^{15}N$ do solo no Cerrado (BUSTAMANTE *et al.*, 2004).

O método *stepwise AIC* incluiu duas variáveis edáficas no *Modelo final*. O *d13C* foi uma delas e corresponde ao $\delta^{13}C$ do solo. Essa varável tem a mesma notação da variável dependente, e mostra a razão isotópica para os isótopos estáveis de carbono (ver Eq. 1). O $\delta^{13}C$ do solo teve um coeficiente positivo no modelo e apresentou uma correlação positiva com o $\delta^{15}N$ do solo (Tabela 3). Essa relação entre o $\delta^{13}C$ e o $\delta^{15}N$ do solo já foi observada em outros ambientes (PERI *et al.*, 2012). Ambas as variáveis são indicativas a eficiência na ciclagem do N e dinâmica da Matéria Orgânica do Solo (MOS) e perdem isótopos mais leves ao longo do processo de decomposição da matéria orgânica (PERI *et al.*, 2012). Além disso, ambientes com presença de algumas gramíneas têm o valor de $\delta^{13}C$ mais elevado, e nesses ambientes pode haver mais perda de N por vias gasosas, elevando também o valor de $\delta^{15}N$ (SENA-SOUZA *et al.*, 2023).

Outra variável edáfica selecionada para o *Modelo final* foi o *Ca2*, que representa a concentração dos íons de Cálcio (Ca^{2+} cmol/dm^3) no solo. O coeficiente positivo da

concentração Ca^{2+} no solo no *Modelo final* é coerente com a dinâmica do nitrogênio no solo (Tabela 3). Embora existam poucos estudos dessa natureza em ambientes cársticos (formado em rochas calcárias, ricas em Ca), sabe-se que Ca^{2+} pode ter um impacto indireto no $\delta^{15}N$ por meio de sua influência no pH do solo e nos processos de nitrificação e desnitrificação. O Ca^{2+} no solo pode competir com íons hidrogênio (H^+) e amônio (NH_4^+) nas trocas catiônicas, levando a uma diminuição da acidez e aumento do pH do solo (PERAKIS *et al.*, 2013). Em solos com pH mais alcalino, a nitrificação é favorecida, resultando na conversão do amônio em nitrato (NO_3^-). Durante a nitrificação, ocorre um maior fracionamento isotópico que tende a acumular o ^{15}N (HOULTON *et al.*, 2006). Portanto, um aumento na concentração de Ca^{2+} que leva ao aumento do pH do solo pode resultar em uma proporção mais alta de ^{15}N em relação ao ^{14}N, aumentando o $\delta^{15}N$ do solo. Por outro lado, em solos com baixa concentração de Ca^{2+} e pH mais ácido, a nitrificação pode ser inibida, favorecendo a retenção de amônio (NH_4^+) no solo, que tem um menor valor de $\delta^{15}N$ (HÖGBERG, 1997).

As demais variáveis preditoras que compõem o *Modelo final* são variáveis climáticas extraídas de dados secundários da base de dados WorldClim (HIJMANS *et al.*, 2005). Três das quatro variáveis climáticas representam medidas de variação da temperatura do ambiente: Sazonalidade de temperatura (*bio4*); e Amplitude anual da temperatura. Isotermalidade é dada em °C e indica a variação viária de temperatura. Essa variável apresentou um efeito negativo no modelo. A sazonalidade da temperatura é o desvio padrão da temperatura do ano múltiplicado por 100. A Amplitude anual de temperatura é dada em °C. As duas últimas tiveram um coeficiente positivo no modelo.

Os efeitos das variáveis derivadas da temperatura são coerentes com o que se espera, uma vez que a sazonalidade climática do Cerrado é um dos princiais fatores que afetam o $\delta^{15}N$ do solo. Sabe-se que há uma correlação forte entre o $\delta^{15}N$ do solo e o $\delta^{15}N$ das plantas em cada área (NARDOTO *et al.*, 2014). Existe uma grande variação de valores de $\delta^{15}N$ foliar entre as plantas do Cerrado devido à sazonalidade, forma de vida e a habilidade da planta em fixar ou não o N atmosférico. Além disso, pode haver diferença entre indivíduos de uma mesma espécie vivendo em fitofisionomias diferentes (BUSTAMANTE *et al.*, 2004). As principais fontes de N para as plantas do Cerrado apresentam diferentes composições isotópicas (COPLEN *et al.*, 2002; ROBINSON, 2001).

A variável *bio16* também foi selecionada para ajuste do modelo final. A variável representa a precipitação pluviométrica do trimestre mais úmido, uma das medidas da sazonalidade climática derivada da precipitação anual. O resultado é semelhante a tendências encontradas em escala global. Craine *et al.* (2015) mostraram que em média, ambientes secos

tendem a ter mais perda de N do solo por vias gasosas, deixando o solo com maiores valores de δ^{15}N do que em ambientes mais chuvosos. Padrões de maiores valores de δ^{15}N do solo em ambientes de savana com menor disponibilidade de água também foram encontrados na África (WANG *et al.*, 2013). Esse padrão indica um ciclo mais aberto nesses ambientes, com maiores perdas relativas de N para a atmosfera em comparação aos ambientes com maior disponibilidade de água (WANG *et al.*, 2013).

5. CONCLUSÃO

O modelo de regressão linear múltipla ajustado para os dados de δ^{15}N do solo do Cerrado confirmou algumas tendências encontradas realizados em diversos biomas. As principais evidências apontavam para uma influência negativa da precipitação nos valores de δ^{15}N, confirmada pela entrada da variável *bio16*. Áreas com menor precipitação estão associadas a uma maior amplitude e maior sazonalidade da temperatura, que tiveram influência positiva no modelo. Apesar da confirmação dos padrões, algumas variáveis edáficas provenientes de dados primários ajudaram a ajustar o modelo e foram significativas, reforçando a necessidade de incluir esse tipo e dados na explicação de valores isotópicos do solo. A relação do δ^{15}N com o δ^{13}C do solo já era conhecida e foi confirmada pelo modelo ajustado no presente estudo. Um possível avanço do modelo apresentado é a concentração de Ca^{2+} como uma variável explicativa significativa. O efeito positivo do Ca^{2+} no δ^{15}N do solo foi explicado por sua influência direta na acidez do solo e no aumento da capacidade de troca de cátions. Entretanto, ainda há um campo científico para ser explorado considerando a relação direta entre Ca^{2+} no δ^{15}N do solo. O ajuste de modelos explicativos do δ^{15}N do solo em ambientes cársticos podem ajudar a esclarecer o padrão encontrado e identificar mecanismos diretos de influência do Ca^{2+}.

Foi possível ajustar um modelo de regressão linear múltipla aos dados disponíveis. O *Modelo final* ajustado seguiu todas as suposições da regressão linear múltipla. O coeficiente de determinação de 60% após a validação cruzada mostrou um modelo com bom desempenho preditivo além da interpretabilidade dos coeficientes. Apesar do bom desempenho, os critérios de seleção de variáveis podem ter excluído variáveis de grande relevância para explicar a variável dependente. O alto poder preditivo do modelo com todas as variáveis demonstra que o conjunto de variáveis pode conter mais explicação para a variância do δ^{15}N do solo, embora esse desempenho também pode indicar *overfitting*. Neste sentido, futuros trabalhos podem explorar outros métodos de seleção de variáveis, bem como métodos que integram grupos de variáveis, como análise de componentes principais, antes de ajustar o modelo de regressão linear múltipla.

A vantagem do modelo final ajustado é a confiabilidade na interpretação dos coeficientes. Porém, outros tipos de modelagem podem ser testados neste conjunto de dados para explorar outras relações lineares e não lineares. Por exemplo, o uso de outros modelos de aprendizagem de máquina, como Floresta Aleatória (*Random forest*), podem ser aplicados na tentativa de melhorar a predição ou encontrar relações não lineares no conjunto de dados.

REFERÊNCIAS BIBLIOGRÁFICAS

AMUNDSON, R. *et al.* Global patterns of the isotopic composition of soil and plant nitrogen. **Global Biogeochemical Cycles**, v. 17, n. 1, p. 1031, 2003.

AUSTIN, A T. *et al.* Latin America's Nitrogen Challenge. **Science**, v. 340, n. 6129, p. 149, 2013.

BOWEN, G. J. Isoscapes: Spatial Pattern in Isotopic Biogeochemistry. **Annual Review of Earth and Planetary Sciences**, p. 161–187, 2010.

BUSTAMANTE, M. M. C. *et al.* ^{15}N natural abundance in woody plants and soils of central Brazilian Savannas (Cerrado). **Ecological Applications**, v. 14, n. sp4, p. 200–213, 2004.

CAMACHO, F. *et al*. GEOV1: LAI, FAPAR essential climate variables and FCOVER global time series capitalizing over existing products. Part 2: validation and intercomparison with reference products. **Remote Sensing of Environment**. v. 137, p. 310–329, 2013

COMPTON, J. E. *et al.* Ecosystem services altered by human changes in the nitrogen cycle: a new perspective for US decision making. **Ecology Letters**, v. 14, n. 8, p. 804–815, 2011.

COPLEN, T. B. *et al.* Isotope-abundance variations of selected elements (IUPAC Technical Report). **Pure and Applied Chemistry**, v. 74, n. 10, p. 1987–2017, 2002.

CRAINE, J. M. *et al.* Convergence of soil nitrogen isotopes across global climate gradients. **Scientific Reports**, v. 5, p. 8280, 2015.

GARCIA-PEREZ, B.; HOBSON, K. A. A multi-isotope (d2H, d13C, d15N) approach to establishing migratory connectivity of Barn Swallow (*Hirundo rustica*). **Ecosphere**, v. 5, n. February, p. 1–12, 2014.

HAWKINS, D. M. The Problem of Overfitting. **Journal of Chemical Information and Computer Sciences**, v. 44, n. 1, p. 1–12, 2004.

HIJMANS, R. J. *et al.* Very high-resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, v. 25, n. 15, p. 1965–1978, 2005.

HOBSON, K. A. *et al.* A multi-isotope (d13C, d15N, d2H) feather isoscape to assign Afrotropical migrant birds to origins. **Ecosphere**, v. 3, n. May, p. 1–20, 2012.

HÖGBERG, P. ^{15}N natural abundance in soil-plant systems. **New Phytologist**, v. 137, n. 2, p. 179–203, 1997.

HOULTON, B. Z.; MARKLEIN, A. R.; BAI, E. Representation of nitrogen in climate change forecasts. **Nature Publishing Group**, v. 5, n. 5, p. 398–401, 2015.

HOULTON, B. Z.; SIGMAN, D. M.; HEDIN, L. O. Isotopic evidence for large gaseous nitrogen losses from tropical rainforests. **Proceedings of the National Academy of Sciences**, v. 103, n. 23, p. 8745–8750, 2006.

LÜDECKE D., BEN-SHACHAR M., PATIL I., WAGGONER P., MAKOWSKI D. O. performance: An R Package for Assessment, Comparison and Testing of Statistical Models. **Journal of Open Source Software**, v. 6, n. 60, p. 3139, 2021.

MALLETTE, J. R. *et al.* Geographically sourcing cocaine's origin – delineation of the nineteen major coca growing regions in South America. **Scientific Reports**, v. 6, n. March, p. 23520, 2016.

MONTGOMERY, D. C.; PENCK, E. A.; VINING, G. **Introduction to Linear Regression Analysis**. 5. ed. New York: [s.n.], 2012.

NARDOTO, G. B. *et al.* Basin-wide variations in Amazon Forest nitrogen-cycling characteristics as inferred from plant and soil ^{15}N:^{14}N measurements. **Plant Ecology & Diversity**, v. 7, n. 1–2, p. 173–187, 2014.

NEVES, G. *et al.* Spatial distribution of soil $\delta^{13}C$ in the central Brazilian savanna**. Journal of Environmental Management**, v. 300, n. February, p. 113758, 2021.

PERAKIS, S. S. *et al.* Forest calcium depletion and biotic retention along a soil nitrogen gradient. **Ecological Applications**, v. 23, n. 8, p. 1947–1961, 2013.

PERI, P. L. *et al.* Carbon ($\delta^{13}C$) and nitrogen ($\delta^{15}N$) stable isotope composition in plant and soil in Southern Patagonia's native forests. **Global Change Biology**, v. 18, n. 1, p. 311–321, 2012.

RENCHER, A. C.; PUN, F. C. Inflation of r^2 in best subset regression. **Technometrics**, v. 22, n. 1, p. 49–53, 1980.

ROBINSON, D. $\delta^{15}N$ as an integrator of the nitrogen. **Trends in Ecology & Evolution**, v. 16, n. 3, p. 153–162, 2001.

SALEMI, L. F. *et al.* Past and present land use influences on tropical riparian zones: an isotopic assessment with implications for riparian forest width determination. **Biota Neotropica,** v. 16, n. 2, p. 1–8, 2016.

SENA-SOUZA, J. P. *et al.* Reconstructing continental-scale variation in soil $\delta^{15}N$: a machine learning approach in South America. **Ecosphere**, v. 11, n. 8, 2020.

SENA-SOUZA, J. P. *et al.* Mapping the effects of Melinis minutiflora invasion on soil nitrogen dynamics in the Brazilian savanna: A dual-isotope approach. **Pedobiologia**, v. 96, 2023.

SENA-SOUZA, J. P.; COSTA, F. J. V.; NARDOTO, G. B. Background and the use of isoscapes in the Brazilian context: essential tool for isotope data interpretation and natural resource management. **Revista Ambiente e Agua**, v. 14, n. 2, 2019.

SHAPIRO, A. S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, n. 3, p. 591–611, 1965.

VITOUSEK, P. M. *et al.* Human alteration of the global nitrogen cycle: Sources and consequences. **Ecological Applications**, v. 7, n. 3, p. 737–750, 1997.

WANG, L. *et al.* Ecosystem-scale spatial heterogeneity of stable isotopes of soil nitrogen in African savannas. **Landscape Ecology**, v. 28, n. 4, p. 685–698, 2013.

WEINTRAUB, S. R. *et al.* Topographic controls on soil nitrogen availability in a lowland tropical forest. **Ecology**, v. 96, n. 6, p. 1561–1574, 2015.

YAMASHITA, T.; YAMASHITA, K.; KAMIMURA, R. A stepwise AIC method for variable selection in linear regression. **Communications in Statistics - Theory and Methods**, v. 36, n. 13, p. 2395–2403, 2007.

APÊNDICE A – Lista de variáveis independentes utilizadas

Tabela A1. Significado e fonte das variáveis independentes usadas neste estudo.

Nome no código	Significado	Fonte
temp_wc	Temperatura média anual (°C)	WorldClim 2.0 (Hijmans et al., 2005)
bio2	Média mensal da variação diária de temperatura (max temp - min temp)	WorldClim 2.0 (Hijmans et al., 2005)
bio3	Isotermalidade	WorldClim 2.0 (Hijmans et al., 2005)
bio4	Sazonalidade de temperatura (BIO2/BIO7) * 100	WorldClim 2.0 (Hijmans et al., 2005)
bio5	Temperatura máxima do mês mais quente (°C)	WorldClim 2.0 (Hijmans et al., 2005)
bio6	Temperatura mínima do mês mais frio (°C)	WorldClim 2.0 (Hijmans et al., 2005)
bio7	Amplitude anual de temperatura (BIO5-BIO6)	WorldClim 2.0 (Hijmans et al., 2005)
bio8	Temperatura média do trimestre mais úmido (°C)	WorldClim 2.0 (Hijmans et al., 2005)
bio9	Temperatura média do trimestre mais seco (°C)	WorldClim 2.0 (Hijmans et al., 2005)
bio10	Temperatura média do trimestre mais quente (°C)	WorldClim 2.0 (Hijmans et al., 2005)
bio11	Temperatura média do trimestre mais frio (°C)	WorldClim 2.0 (Hijmans et al., 2005)
pre_wc	Precipitação anual (mm)	WorldClim 2.0 (Hijmans et al., 2005)
bio13	Precipitação do mês mais chuvoso (mm)	WorldClim 2.0 (Hijmans et al., 2005)
bio14	Precipitação do mês mais seco (mm)	WorldClim 2.0 (Hijmans et al., 2005)
bio15	Sazonalidade de precipitação (Coeficiente de Variação da precipitação anual)	WorldClim 2.0 (Hijmans et al., 2005)
bio16	Precipitação do trimestre mais úmido (mm)	WorldClim 2.0 (Hijmans et al., 2005)
bio17	Precipitação do trimestre mais seco (mm)	WorldClim 2.0 (Hijmans et al., 2005)
bio18	Precipitação do trimestre quente (mm)	WorldClim 2.0 (Hijmans et al., 2005)
bio19	Precipitação do trimestre mais frio (mm)	WorldClim 2.0 (Hijmans et al., 2005)
ndvi_apr	Média de abril (1999-2017) do índice de vegetação de diferença normalizada	Copernicus Global Land Service dataset (CGLS) (Camacho et al., 2013)
ndvi_sep	Média de setembro (1999-2017) do índice de vegetação de diferença normalizada	Copernicus Global Land Service dataset (CGLS) (Camacho et al., 2013)
npp	Produção primária líquida	MODIS17
gpp	Produção primária bruta	MODIS17
fapar	Fração de radiação absorvida por atividade fotossintética	Copernicus Global Land Service dataset (CGLS) (Camacho et al., 2013)
soilwater1500	Conteúdo de água no solo a 1500 kpa	Copernicus Global Land Service dataset (CGLS) (Camacho et al., 2013)
lat	Latitude	Este estudo
long	Longitude	Este estudo
alt	Altitude	SRTM (http://earthexplorer.usgs.gov)
pH_H2O	pH em água	Este estudo
P	Fóstoro (mg/dm3)	Este estudo
K	Potássio (mg/dm3)	Este estudo
Ca2	Cálcio (cmol/dm3)	Este estudo
Mg2	Magnésio (cmol/dm3)	Este estudo
Al3_Altrocavel	Alumínio trocável (cmol/dm3)	Este estudo

H_Al3	H + Al - Extrator Acetato de Cálcio 0,5mol/L - pH 7,0	Este estudo
SB	Soma de bases trocáveis (cmol/dm3)	Este estudo
t	Capacidade de troca catiônica efetiva (cmol/dm3)	Este estudo
T	Capacidade de troca catiônica a pH 7,0 (cmol/dm3)	Este estudo
V	Índice de Saturação por Bases (%)	Este estudo
m	Índice de Saturação por Alumínio (%)	Este estudo
P_rem	Fósforo remanescente	Este estudo
N	Nitrogênio (%)	Este estudo
C	Carbono (%)	Este estudo
CN	Razão Carbono Nitrogênio	Este estudo
d13C	Isótopos estáveis de Carbono	Este estudo

APÊNDICE B – Matrizes de correlação entre variáveis independentes

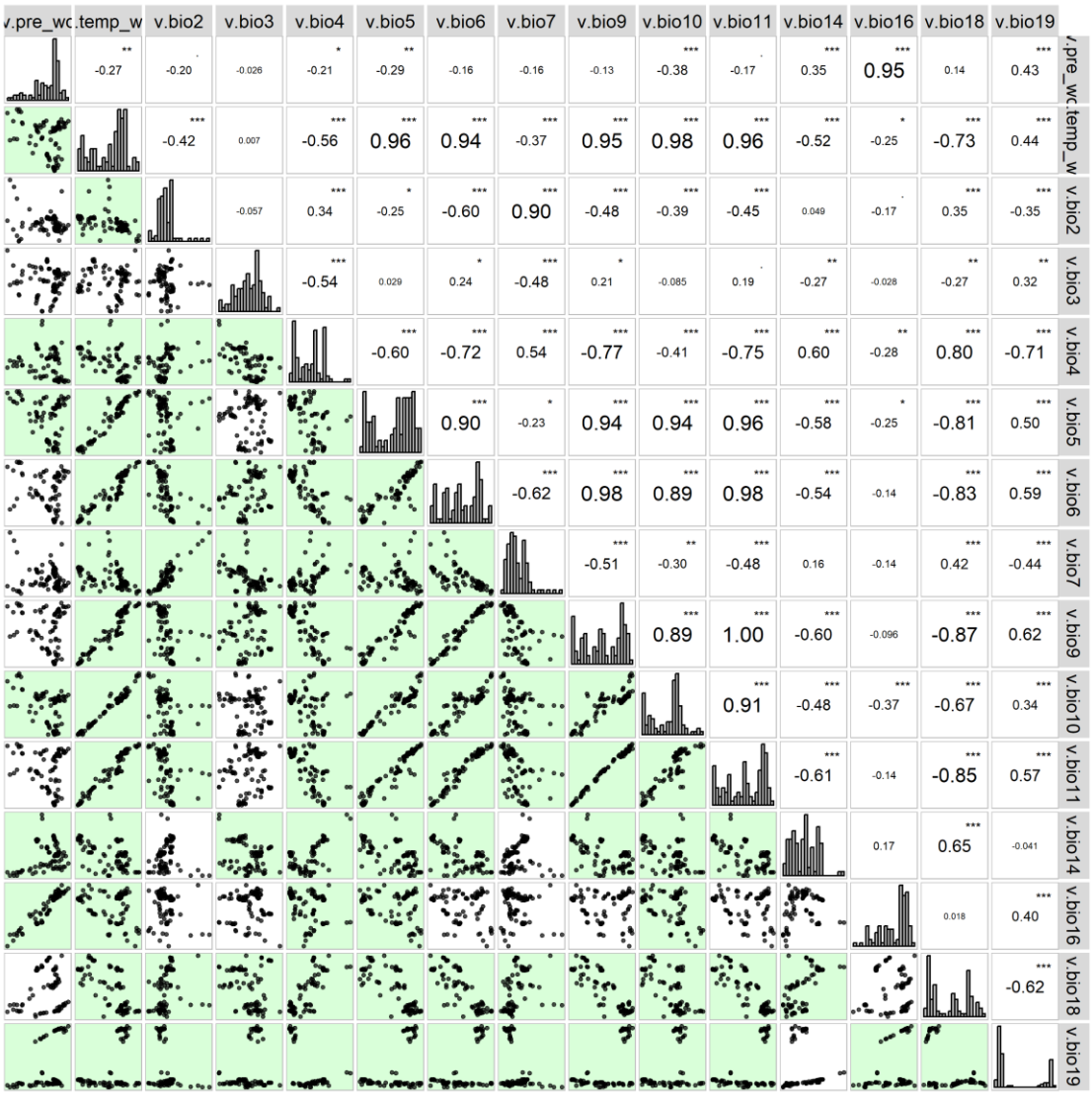

Figura A9. Matriz de correlação entre variáveis climáticas.

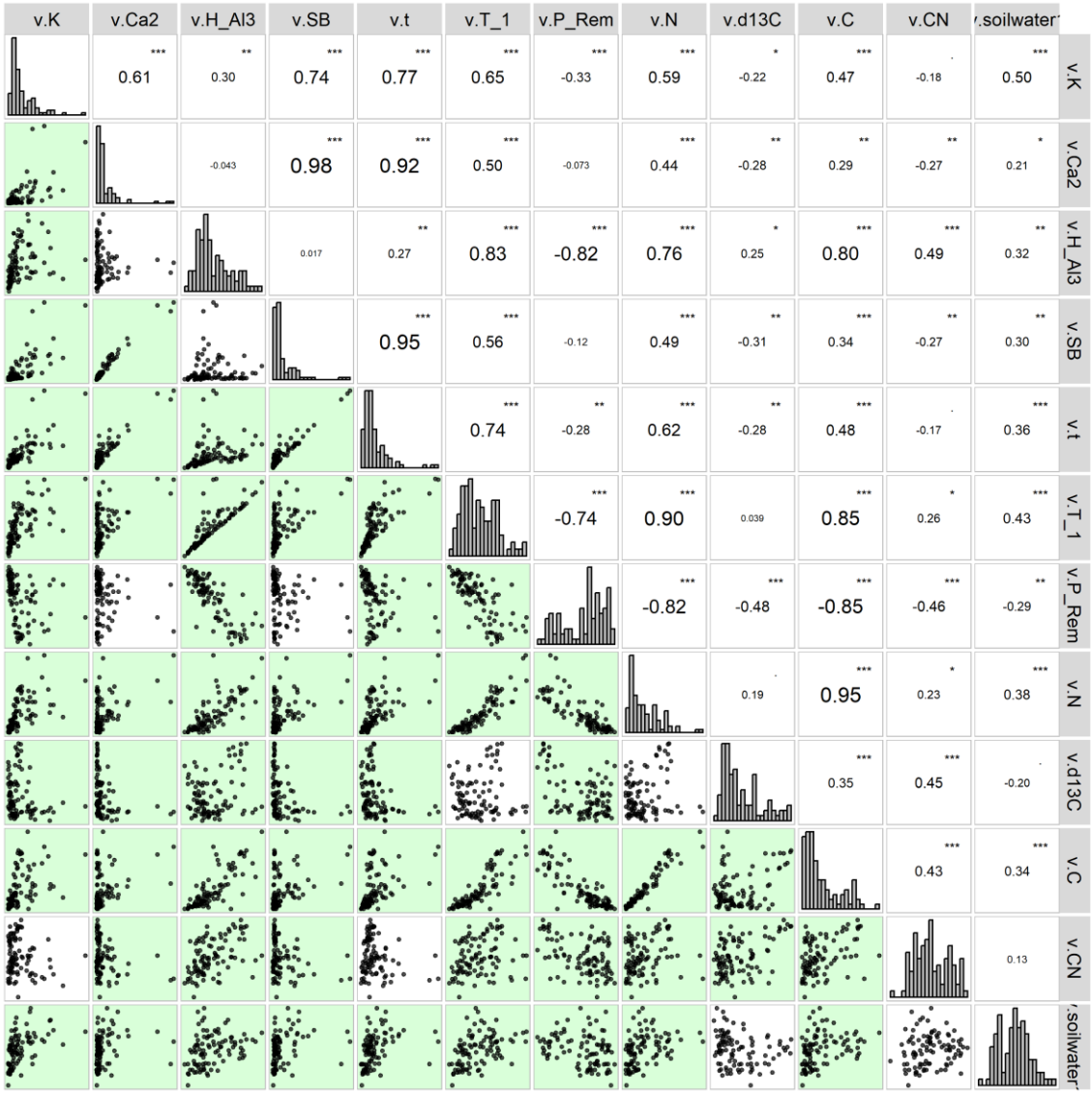

Figura A2. Matriz de correlação entre variáveis edáficas.

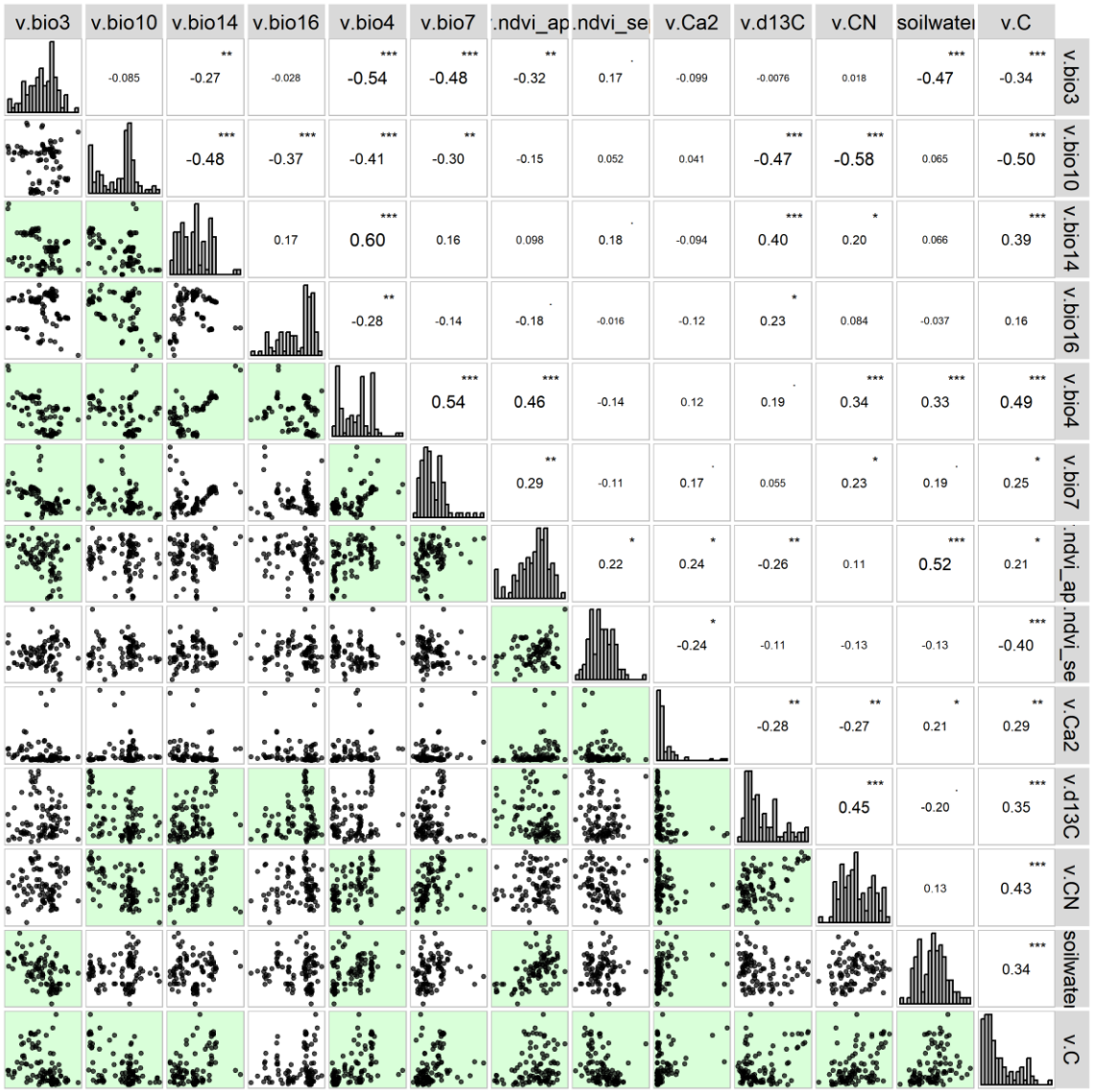

Figura A3. Matriz de correlação entre variáveis selecionadas pela ausência de multicolinearidade.

APÊNDICE C – Tabelas de análise de variância, Coeficientes da regressão, e os parâmetros de modelos gerados no processo metodológico

Tabela A2. Tabela de coeficientes e estatísticas do modelo de regressão múltipla gerado com todas as variáveis que apresentaram correlação linear de Pearson ou de Spearman significativa (p-valor < 0,05) com a variável dependente.

	Estimate	Std.Error	t	Pr(>|t|)	Df	Sum of Sq	RSS	AIC	F	Pr(>F)
(Intercept)	159.50	120.20	1.33	0.19			39.27	0.29		
lat	-0.355	0.38	-0.95	0.35	1	0.66	39.93	-0.09	0.89	0.35
long	-0.059	0.35	-0.17	0.86	1	0.02	39.29	-1.65	0.03	0.86
pH_H2O	-1.150	0.84	-1.37	0.18	1	1.39	40.66	1.67	1.88	0.18
P	-0.019	0.14	-0.14	0.89	1	0.01	39.29	-1.67	0.02	0.89
K	-0.043	0.11	-0.38	0.71	1	0.11	39.38	-1.44	0.14	0.71
Ca2	-16.760	44.31	-0.38	0.71	1	0.11	39.38	-1.45	0.14	0.71
Mg2	-16.940	44.43	-0.38	0.70	1	0.11	39.38	-1.44	0.15	0.70
Al3_Altroc	0.415	0.78	0.53	0.60	1	0.21	39.48	-1.20	0.28	0.60
H_Al3	-0.461	0.26	-1.79	0.08	1	2.36	41.63	3.95	3.19	0.08
SB	16.580	44.32	0.37	0.71	1	0.10	39.38	-1.45	0.14	0.71
V	0.023	0.04	0.53	0.60	1	0.21	39.48	-1.20	0.28	0.60
m	-0.014	0.02	-0.64	0.53	1	0.30	39.57	-0.97	0.41	0.53
P_Rem	-0.067	0.03	-2.37	**0.02**	1	4.18	43.45	8.10	5.64	0.02
N	5.011	16.36	0.31	0.76	1	0.07	39.34	-1.54	0.09	0.76
d13C	0.198	0.06	3.59	**0.00**	1	9.55	48.82	19.40	12.89	0.00
C	0.239	0.92	0.26	0.80	1	0.05	39.32	-1.58	0.07	0.80
CN	-0.145	0.14	-1.05	0.30	1	0.81	40.08	0.28	1.10	0.30
pre_wc	-0.011	0.01	-1.34	0.19	1	1.33	40.61	1.53	1.80	0.19
temp_wc	3.981	5.37	0.74	0.46	1	0.41	39.68	-0.71	0.55	0.46
bio2	9.662	10.09	0.96	0.34	1	0.68	39.95	-0.04	0.92	0.34
bio3	-1.682	1.76	-0.96	0.34	1	0.68	39.95	-0.05	0.91	0.34
bio4	-0.022	0.10	-0.23	0.82	1	0.04	39.31	-1.61	0.05	0.82
bio5	-6.833	7.04	-0.97	0.34	1	0.70	39.97	0.00	0.94	0.34
bio6	5.211	6.75	0.77	0.44	1	0.44	39.71	-0.62	0.60	0.44
bio8	-4.136	4.19	-0.99	0.33	1	0.72	40.00	0.06	0.98	0.33
bio9	-3.732	4.25	-0.88	0.38	1	0.57	39.84	-0.31	0.77	0.38
bio10	2.298	4.29	0.54	0.59	1	0.21	39.49	-1.18	0.29	0.59
bio11	2.688	3.51	0.77	0.45	1	0.43	39.71	-0.64	0.59	0.45
bio13	-0.054	0.02	-2.68	0.01	1	5.32	44.59	10.62	7.18	0.01
bio14	0.085	0.15	0.57	0.57	1	0.24	39.52	-1.11	0.33	0.57

bio15	-0.227	0.22	-1.04	0.30	1	0.80	40.07	0.24	1.07	0.30
bio16	0.035	0.02	1.73	0.09	1	2.23	41.50	3.64	3.00	0.09
bio17	-0.052	0.06	-0.82	0.42	1	0.50	39.77	-0.48	0.67	0.42
bio18	0.001	0.00	0.46	0.65	1	0.15	39.43	-1.33	0.21	0.65
bio19	0.002	0.00	1.46	0.15	1	1.58	40.85	2.12	2.13	0.15
gpp	0.000	0.00	0.04	0.96	1	0.00	39.27	-1.70	0.00	0.96
npp	0.000	0.00	-0.50	0.62	1	0.19	39.46	-1.24	0.25	0.62
soilwater1	-0.062	0.07	-0.89	0.38	1	0.59	39.86	-0.26	0.80	0.38
soilwater3	-0.023	0.05	-0.45	0.65	1	0.15	39.42	-1.34	0.20	0.65
ndvi_apr	0.366	3.73	0.10	0.92	1	0.01	39.28	-1.69	0.01	0.92
ndvi_sep	-2.930	2.85	-1.03	0.31	1	0.78	40.05	0.20	1.05	0.31
fapar	0.011	0.01	0.93	0.36	1	0.64	39.91	-0.14	0.86	0.36
alt	-0.003	0.01	-0.40	0.69	1	0.12	39.39	-1.42	0.16	0.69

Tabela A3. Tabela de coeficientes e estatísticas do *Modelo 2,* ajustado após exclusão de variáveis preditoras com multicolinearidade. Este modelo passou pelo *stepwise AIC* para o ajuste do *Modelo final*.

	Estimate	**Std. Error**	**t**	**Pr(>|t|)**	**Df**	**Sum of Sq**	**RSS**	**AIC**	**F value**	**Pr(>F)**
(Intercept)	17.944	12.65	1.42	0.16			82.40	12.17		
bio3	-0.120	0.11	-1.14	0.26	1	1.29	83.69	11.68	1.30	0.26
bio10	-0.059	0.14	-0.41	0.68	1	0.17	82.56	10.37	0.17	0.68
bio14	-0.021	0.03	-0.77	0.44	1	0.59	82.99	10.86	0.60	0.44
bio16	-0.004	0.00	-2.40	0.02	1	5.72	88.12	16.69	5.77	0.02
bio4	0.009	0.01	1.40	0.17	1	1.94	84.33	12.43	1.95	0.17
bio7	0.171	0.13	1.33	0.19	1	1.75	84.15	12.21	1.76	0.19
ndvi_apr	1.858	2.34	0.79	0.43	1	0.63	83.02	10.91	0.63	0.43
ndvi_sep	-0.766	2.14	-0.36	0.72	1	0.13	82.52	10.32	0.13	0.72
d13C	0.183	0.04	4.54	0.00	1	20.45	102.85	31.68	20.60	0.00
CN	-0.017	0.07	-0.26	0.80	1	0.07	82.46	10.25	0.07	0.80
soilwater1	-0.007	0.05	-0.13	0.90	1	0.02	82.41	10.19	0.02	0.90
C	-0.020	0.21	-0.10	0.92	1	0.01	82.41	10.18	0.01	0.92
log1p(Ca2)	0.979	0.32	3.07	0.00	1	9.35	91.74	20.60	9.42	0.00

APÊNCIDE D – Script R

```
##### Script para TCC de Especializacao em Estatistica
##### Joao Paulo Sena Souza

## Definindo o diretorio

setwd("C:\\...\\")

## Carregando os pacotes

library(corrplot); library(caret); library(lares);
library(leaps); library(raster); library(MASS); library(sp); library(gstat)
library(ncf); library(gridExtra); library(ggpmisc); library(dplyr); library(car)
library(metan); library(readr); library(GWmodel); library(FactoMineR); library(lmtest);
library(sp); library(terra); library(spdep); library(performance); library(rgdal)

##### Abrindo o banco de dados #####

dados <- read.csv("dados_d15Nsolo_cerrado_TCC_joaopaulo.csv", header = T,
                  sep = ";")


##### Analise exploratoria da variavel dependente d15N do solo #####

# Estatistica descritiva da variavel dependente
summary(dados$d15N)
sd(dados$d15N)
boxplot(dados$d15N)

png("hist.png", width = 5, height = 4, units = 'in', res = 300)
hist(dados$d15N, breaks = 10, freq = F, xlab = expression
(delta^{15}*N~do~solo~"(\u2030)"),
     ylab = 'Densidade', main = ' ')
lines(density(dados$d15N, na.rm = T, from = 1, to = 12),col = "black", lwd = 2)
dev.off()

shapiro.test(dados$d15N)


##### Teste de normalidade das covariaveis e separação dos conjuntos de dados#####
# Criando vetores para armazenar as variaveis normais e nao normais
variaveis_normais <- c()
variaveis_nao_normais <- c()
```

```
43
44  # Testando a normalidade de cada variavel usando o teste Shapiro-Wilk
45  for (col in 2:48) {
46    p_valor <- shapiro.test(dados[, col])$p.value
47
48    # Definindo um limite de significancia de 0,05
49    if (p_valor >= 0.05) {
50      variaveis_normais <- c(variaveis_normais, col)
51    } else {
52      variaveis_nao_normais <- c(variaveis_nao_normais, col)
53    }
54  }
55
56  # Dividindo o conjunto de dados com base na normalidade das variaveis
57  dados_normais <- dados[, variaveis_normais]
58  dados_nao_normais <- dados[, variaveis_nao_normais]
59
60  ncol(dados_nao_normais)
61  ##### Testando a relacao linear da variavel dependente com as variaveis preditivas #####
62  ##### Escolha das variaveis nao normais
63  dados_nao_normais <- data.frame(dados$d15N, dados_nao_normais)
64
65  corr.var.nnorm <- corr_var(dados_nao_normais, dados.d15N, method = "spearman", plot = F,
66  top = 46, max_pvalue = 0.05)
67  corr.var.nnorm$variables
68
69  ## Plotando as variaveis nao normais correlacionadas
70
71  png("correlations_nnorm_d15n.jpg", width = 7, height = 5, units = 'in', res = 300)
72  ggplot(corr.var.nnorm, aes(x=abs(corr), y=reorder(variables, abs(corr)),fill=corr))+
73    geom_bar(stat="identity", position="dodge")+
74    geom_text(aes(label=format(round(corr, 2), nsmall = 1)),
75              position=position_dodge(width=0.9), hjust=1.2,  colour = "white", size=3) +
76    scale_fill_gradient(low = "blue4",
77                        high = "red4") +
78    labs(x = "Coeficiente de correlação de Spearman (valor absoluto)", y = "Variaveis",
79         title = expression (Correlação~de~Spearman~com~o~delta^{15}*N~do~solo~"(\u2030)"),
80         subtitle = "Apenas correlações significativas (p-valor < 0,05)")+
81    theme_minimal()+
82    guides(fill = FALSE)
83  dev.off()
84
85  ##### Escolha das variaveis normais
86
```

```
corr.var.norm <- corr_var(dados_normais, d15N, method = "pearson", plot = F, top = 6,
max_pvalue = 0.05)
corr.var.norm$variables

## Plotando as variaveis nao normais correlacionadas

png("correlations_norm_d15n.jpg", width = 7, height = 5, units = 'in', res = 300)
ggplot(corr.var.norm, aes(x=abs(corr), y=reorder(variables, abs(corr)),fill=corr))+
  geom_bar(stat="identity", position="dodge")+
  geom_text(aes(label=format(round(corr, 2), nsmall = 1)),
            position=position_dodge(width=0.9), hjust=1.2,  colour = "white", size=3) +
  scale_fill_gradient(low = "blue4",
                      high = "red4") +
  labs(x = "Coeficiente de correlação de Pearson (valor absoluto)", y = "Variaveis",
       title = expression (Correlação~de~Pearson~com~o~delta^{15}*N~do~solo~"(\u2030)"),
       subtitle = "Apenas correlações significativas (p-valor < 0,05)")+
  theme_minimal()+
  guides(fill = FALSE)
dev.off()


##### Criando o data.frame apenas com as covariaveis correlacionadas e a variavel
dependente ####
v <- data.frame(dados$d15N, dados[,c(corr.var.nnorm$variables)],
dados[,c(corr.var.norm$variables)])

colnames(v) <- c("d15N", c(corr.var.nnorm$variables), c(corr.var.norm$variables))

##### SELEÇÂO DE VARIÁVEIS ######
### Separando grupos de variáveis - Climáticas
var_clim<-data.frame(v$pre_wc, v$temp_wc, v$bio2, v$bio3, v$bio4, v$bio5, v$bio6,
                     v$bio7, v$bio9, v$bio10, v$bio11, v$bio14, v$bio16, v$bio18, v$bio19)

par(mfrow=c(1,1))
corr_plot(var_clim)
climCor <- findCorrelation(cor(var_clim), cutoff = .70)
varclim_cor <- subset(var_clim, select=c(-climCor))
corr_plot(varclim_cor)
varclim_cor <- data.frame(varclim_cor[,-c(1,6)], v$bio4, v$bio7) ## incluindo variáveis com
forte correlação com a variável alvo, respeitando a multicolinearidade
corr_plot(varclim_cor)

### Separando grupos de variáveis - Vegetação
var_veg<-data.frame(v$ndvi_apr,v$ndvi_sep,v$npp)
```

```
131
132 par(mfrow=c(1,1))
133 corr_plot(var_veg)
134
135
136 ### Separando grupos de variáveis - Solo
137 var_soil<-data.frame(v$K, v$Ca2, v$H_Al3, v$SB, v$t,
138                      v$T_1, v$P_Rem, v$N, v$d13C, v$C, v$CN,v$soilwater1)
139 par(mfrow=c(1,1))
140 corr_plot(var_soil)
141 soilCor <- findCorrelation(cor(var_soil), cutoff = .70)
142 varsoil_cor <- subset(var_soil, select=c(-soilCor))
143 corr_plot(varsoil_cor)
144 varsoil_cor <- subset(var_soil, select=c(-soilCor))
145 varsoil_cor<- data.frame(varsoil_cor[,-2],v$C)
146 corr_plot(varsoil_cor)
147
148
149 ### Juntando todas as variáveis sem multicolinearidade intragrupos para testar
150 multicolinearidade intergrupos
151
152 todas_var_cor<-data.frame(v$long,v$lat,varclim_cor,var_veg,varsoil_cor)
153
154 png("matrizcor_16.jpg", width = 9, height = 9, units = 'in', res = 300)
155 corr_plot(todas_var_cor)
156 dev.off()
157
158 allCor <- findCorrelation(cor(todas_var_cor), cutoff = .70)
159 var.final <- subset(todas_var_cor, select=c(-allCor))
160
161 png("matriz_cor_final.jpg", width = 9, height = 9, units = 'in', res = 300)
162 corr_plot(var.final)
163 dev.off()
164
165 df.final<-data.frame(v$d15N,var.final)
166 names(df.final)
167 colnames(df.final)<-c("d15N","bio3","bio10","bio14","bio16","bio4","bio7",
168                       "ndvi_apr","ndvi_sep","Ca2","d13C","CN","soilwater1",
169                       "C")
170
171
172 ##### AJUSTANDO MODELOS DE REGRESSÃO LINEAR MÚLTIPLA #####
173
174 ##### Rodando o primeiro modelo de regressão linear múltipla com todas as variáveis #####
```

```
### variáveis correlacionadas com a variável dependente
# Gerando modelo1 sem variáveis que geraram NA nos coeficientes.
modelo1 <- lm(d15N ~ .-t-T_1-bio7, data = dados [,-1])
summary(modelo1)

drop1(modelo1, test = "F")
AIC(modelo1)
sqrt(sum((dados$d15N - modelo1$fitted.values)^2)/nrow(dados))

shapiro.test(modelo1$residuals)
dwtest(modelo1)
bptest(modelo1)

##### Rodando Modelo 2, com variaveis excluídas pela multicolinearidade #####
## A variável Ca2 foi transformada nesta etapa para amenizar a influência de outliers
modelo2<-lm(formula = d15N ~ .-Ca2+log1p(Ca2), data = df.final)

summary(modelo2)
vif.modelo2<-as.data.frame(vif(modelo2))
drop1(modelo2, test = "F")
AIC(modelo2)
sqrt(sum((df.final$d15N - modelo2$fitted.values)^2)/nrow(df.final))

shapiro.test(modelo2$residuals)
dwtest(modelo2)
bptest(modelo2)

##### Obtendo o melhor modelo aplicando o metodo Stepwise AIC #####
stepAIC(modelo2)
?stepAIC
modelo.final <- lm(formula = d15N ~ bio3 + bio16 + bio4 + bio7 + log1p(Ca2) +
                     d13C, data = df.final)
summary(modelo.final)

modelo.final <- lm(formula = d15N ~ bio16 + bio4 + bio7 + log1p(Ca2) +
                     d13C, data = df.final) # Excluindo bio3
summary(modelo.final)

vif.mfinal<-as.data.frame(vif(modelo.final))
drop1(modelo.final, test = "F")
AIC(modelo.final)
plot(modelo.final)

dados[20,]
```

```
sqrt(sum((df.final$d15N - modelo.final$fitted.values)^2)/nrow(df.final))

##### Análise descritiva das variaveis do modelo final #####

var_mod_final <- df.final[,c("d15N", "bio3", "bio16", "bio4", "bio7", "Ca2",
                             "d13C")]

summary(var_mod_final)
sapply(var_mod_final, sd)

##### Análise dos resíduos do modelo final #####

png("lm_performance.jpg", width = 8, height = 8, units = 'in', res = 300)
performance::check_model(modelo.final, alpha = 0.5)
dev.off()


summary(modelo.final$residuals)
hist(modelo.final$residuals)
shapiro.test(modelo.final$residuals)
dwtest(modelo.final)
bptest(modelo.final)



##### Testando autocorrelação espacial dos resíduos #####

##### Testando autocorrelação espacial pelo Índice de Moran Global
resid.lm <- data.frame(dados[,4],dados[,3], modelo.final$residuals)
colnames(resid.lm) <- c("long", "lat", "residuals")

coordinates(resid.lm) <- c("long", "lat")


nb <- knn2nb(knearneigh(coordinates(resid.lm)), row.names = NULL)
w <- nb2listw(nb)
moran.test(resid.lm$residuals, listw=w, randomisation = F)

### Índice de Moran local por classe de distância
ncf.cor <- correlog(dados[,4],dados[,3], modelo.final$residuals,
                    increment = 1, resamp = 0, quiet = T)

png("correlog_residlmfinal.jpg", width = 5, height = 4, units = 'in', res = 300)
plot(ncf.cor, xlab = 'Classe de distância',
```

```
     ylab = "I de Moran Local dos resíduos" )
abline(h = 0, lty = 2)
dev.off()

resid <- vect(resid.lm)
writeVector(resid, "resid_lmfinal.shp")


##### Testando o poder preditivo dos modelos LOOCV #####

## Modelo1

y <- dados$d15N
SQT <- sum((y - mean(y))^2)

yhat <- rep(NA, nrow(dados))

for(cont in 1:nrow(dados)){
  modelo <- lm(d15N ~ .-t-T_1-bio7, data = dados [,-1])
  yhat[cont] <- predict(modelo, newdata = dados[cont,])
}

SQres <- sum((y - yhat)^2)

RMSE <- sqrt(sum(((y - yhat)^2)/nrow(dados)))

(R2 <- 1 - SQres/SQT)

res <- y - yhat
sd(res)
summary(lm(y ~yhat))

##### Gráfico de validação após LOOCV do lm.1
valid.lm <- data.frame(y, yhat)

(p.lm1<-ggplot(valid.lm,aes(x = yhat, y = y)) + geom_point(alpha = 1,size = 2) +
    geom_abline(intercept = 0, slope = 1,colour = "grey") +
    stat_smooth(method = lm, colour = "red", se = T) +
    scale_x_continuous(limits=c(2, 10)) +
    scale_y_continuous(limits=c(2, 10)) +
    xlab(expression (delta^{15}*N~predito~"(\u2030)")) +
    ylab(expression (delta^{15}*N~medido~"(\u2030)")) +
    ggtitle("a)")+
    theme_bw()+
```

```
307       theme(panel.grid = element_blank(),axis.title = element_text(size = 16),
308             axis.text = element_text(size = 12, colour = "black"),
309             plot.title=element_text( hjust=.01, vjust=-7)))
310
311
312
313
314   ##### Testando o poder preditivo do lm por LOOCV do modelo 2
315   y <- dados$d15N
316   SQT <- sum((y - mean(y))^2)
317
318   yhat <- rep(NA, nrow(dados))
319
320   for(cont in 1:nrow(dados)){
321     modelo <- lm(formula = d15N ~ .-Ca2+log1p(Ca2), data = df.final)
322     yhat[cont] <- predict(modelo, newdata = dados[cont,])
323   }
324
325   SQres <- sum((y - yhat)^2)
326
327   (RMSE <- sqrt(sum(((y - yhat)^2)/nrow(dados))))
328
329   (R2 <- 1 - SQres/SQT)
330
331   res <- y - yhat
332   sd(res)
333   summary(lm(y ~yhat))
334
335   AIC(lm(y ~yhat))
336   ##### Gráfico de validação após LOOCV do modelo 2
337   valid.lm <- data.frame(y, yhat)
338
339   (p.lm2<- ggplot(valid.lm,aes(x = yhat, y = y)) + geom_point(alpha = 1,size = 2) +
340       geom_abline(intercept = 0, slope = 1,colour = "grey") +
341       stat_smooth(method = lm, colour = "red", se = T) +
342       scale_x_continuous(limits=c(2, 10)) +
343       scale_y_continuous(limits=c(2, 10)) +
344       xlab(expression (delta^{15}*N~predito~"(\u2030)")) +
345       ylab(expression (delta^{15}*N~medido~"(\u2030)")) +
346       ggtitle("b)")+
347       theme_bw()+
348       theme(panel.grid = element_blank(),axis.title = element_text(size = 16),
349             axis.text = element_text(size = 12, colour = "black"),
350             plot.title=element_text( hjust=.01, vjust=-7)))
```

```
351
352
353  ##### Testando o poder preditivo do lm por LOOCV do modelo final
354  y <- dados$d15N
355  SQT <- sum((y - mean(y))^2)
356
357  yhat <- rep(NA, nrow(dados))
358
359  for(cont in 1:nrow(dados)){
360    modelo <- lm(formula = d15N ~ bio16 + bio4 + bio7 + log1p(Ca2) +
361                   d13C, data = df.final)
362    yhat[cont] <- predict(modelo, newdata = dados[cont,])
363  }
364
365  SQres <- sum((y - yhat)^2)
366
367  RMSE <- sqrt(sum(((y - yhat)^2)/nrow(dados)))
368
369  (R2 <- 1 - SQres/SQT)
370  res <- y - yhat
371  sd(res)
372  summary(lm(y ~yhat))
373
374  AIC(lm(y ~yhat))
375  ##### Gráfico de validação após LOOCV do modelo final
376  valid.lm <- data.frame(y, yhat)
377
378  p.lm3 <- ggplot(valid.lm,aes(x = yhat, y = y)) + geom_point(alpha = 1,size = 2) +
379    geom_abline(intercept = 0, slope = 1,colour = "grey") +
380    stat_smooth(method = lm, colour = "red", se = T) +
381    scale_x_continuous(limits=c(2, 10)) +
382    scale_y_continuous(limits=c(2, 10)) +
383    xlab(expression (delta^{15}*N~predito~"(\u2030)")) +
384    ylab(expression (delta^{15}*N~medido~"(\u2030)")) +
385    ggtitle("c)")+
386    theme_bw()+
387    theme(panel.grid = element_blank(),axis.title = element_text(size = 16),
388          axis.text = element_text(size = 12, colour = "black"),
389          plot.title=element_text( hjust=.01, vjust=-7))
390
391  png("ValidacaoLOOCV.jpg", width = 11, height = 3.5, units = 'in', res = 300)
392  grid.arrange(p.lm1,p.lm2,p.lm3,nrow = 1)
393  dev.off()
```