

Modelagem Atuarial de Perdas em Apólices de Seguro de Soja no Brasil

Helton Graziadei

Departamento de Estatística, Universidade Federal de São Carlos (UFSCar).
helton@ufscar.br
<https://orcid.org/0000-0001-7381-9015>

Eduardo F.L. de Melo

Escola de Matemática Aplicada, Fundação Getúlio Vargas (EMAp/FGV). Universidade Estadual do Rio de Janeiro (UERJ). Superintendência de Seguros Privados (SUSEP).
eduardofraga@ime.uerj.br

Rodrigo S. Targino

Escola de Matemática Aplicada, Fundação Getúlio Vargas (EMAp/FGV). Observatório do Seguro Rural (OSR), Fundação Getúlio Vargas.
rodrigo.targino@fgv.br

Resumo

No Brasil, a soja se destaca como a cultura mais importante em um setor essencial para o país. A previsão para a safra de 2023/2024 é de uma produção superior a 160 milhões de toneladas, equivalente à produção combinada dos países da OCDE. Contudo, sua produção enfrenta vulnerabilidades climáticas que impactam tanto a produtividade quanto os preços do seguro rural. Regiões Sul e Centro-Oeste têm sofrido com secas ou excesso de chuvas, afetando gravemente as colheitas e aumentando a sinistralidade das seguradoras. Neste estudo, utilizamos modelos de aprendizagem atuarial para prever perdas nas lavouras de soja, considerando variáveis climáticas como precipitação e temperatura. Foram comparados modelos lineares generalizados, modelos aditivos generalizados e floresta aleatória, aplicados tanto para a ocorrência de sinistros quanto para a severidade. O desempenho preditivo dos modelos foi avaliado e, com base no modelo mais adequado, extrapolamos as perdas para medir a quebra de produtividade em propriedades com e sem seguro, por município e em nível nacional. Os resultados indicam caminhos práticos para a precificação, gestão de riscos, regulação e formulação de políticas públicas agrícolas.

Palavras-Chave

Modelos de aprendizagem atuarial; seguro agrícola; produção de soja; florestas aleatórias; perdas agrícolas.

Sumário

1. Introdução. 2. Dados. 2.1 Dados de Seguro Rural. 2.2 Dados climáticos e do solo. 3. Modelagem. 3.1 Modelo Atuarial. 3.2 Avaliação de Desempenho dos Modelos. 3.3 Interpretabilidade. 4. Cálculo do Prêmio e Extrapolação das Perdas para Propriedades Não Seguradas. 4.1 Cálculo do Prêmio. 4.2 Extrapolação de Perdas para Propriedades Não Seguradas. 5. Considerações finais. 6. Agradecimentos. 7. Referências Bibliográficas. Anexo A. Anexo B.

Abstract

Actuarial Modeling of Losses in Brazilian Soybean Insurance Policies

Helton Graziadei

Assistant Professor, Department of Statistics, Federal University of São Carlos.

helton@ufscar.br

<https://orcid.org/0000-0001-7381-9015>

Eduardo F.L. de Melo

School of Applied Mathematics, Getulio Vargas Foundation, Brazil. Rio de Janeiro State University. The Superintendency of Private Insurance (SUSEP).

eduardofraga@ime.uerj.br

Rodrigo S. Targino

School of Applied Mathematics, Getulio Vargas Foundation, Brazil. Rural Insurance Observatory, Getulio Vargas Foundation.

rodrigo.targino@fgv.br

Summary

Soybean stands out as the most important crop in Brazil, a key sector for the country's economy. The forecast for the 2023/2024 harvest is over 160 million tons, equivalent to the combined production of OECD countries. However, soybean production faces climatic vulnerabilities that affect both productivity and rural insurance prices. The South and Midwest regions have suffered from droughts or excessive rainfall, severely impacting yields and increasing insurers' loss ratios. In this study, we apply actuarial learning models to predict soybean crop losses, incorporating climatic variables such as precipitation and temperature. We compare generalized linear models, generalized additive models, and random forests, applied to both claim occurrence and claim severity. Predictive performance was assessed, and based on the best-performing model, we extrapolated losses to measure productivity shortfalls in insured and uninsured farms, at both municipal and national levels. The findings are relevant for pricing, risk management, regulation, and the design of agricultural public policies.

key-words

Actuarial learning models; crop insurance; soybean production; random forests; agricultural losses.

Contents

1. Introduction. 2. Data. 2.1 Rural Insurance Data. 2.2 Climate and Soil Data. 3. Methodology. 3.1 Actuarial Model. 3.2 Model Performance Evaluation. 3.3 Interpretability. 4. Premium Calculation and Loss Extrapolation to Non-Insured Areas. 4.1 Premium Calculation. 4.2 Loss Extrapolation to Non-Insured Properties. 5. Concluding Remarks. 6. Acknowledgments. 7. Bibliographic References. Appendix A. Appendix B.

1. Introdução

O setor agrícola é de suma importância para o Brasil, contribuindo com cerca de 25% do PIB nacional em 2022¹. Dessa forma, qualquer queda na produção rural impacta diretamente a renda de um setor econômico crucial. A produção da soja projetada do Brasil reflete a produção total de todas as nações da OCDE, atingindo 160 milhões de toneladas em 2023/2024². No contexto da guerra no Leste Europeu, aumento de preços e condições climáticas extremas, a produção agrícola no Brasil torna-se ainda mais relevante para o país e para o mundo.

Dada sua importância, uma cobertura de seguro adequada é essencial para essa atividade econômica. O seguro rural desempenha um papel crítico na mitigação de perdas para os agricultores, tornando-se um componente expressivo da agricultura e da economia do país. O risco inerente ao seguro rural está relacionado às perdas decorrentes de eventos climáticos e pragas, ambos capazes de afetar a produtividade das lavouras.

A Organização das Nações Unidas para a Alimentação e a Agricultura (FAO) enfatiza a importância desse seguro para aumentar a produção de alimentos (FAO (2020) e FAO (2021)), posicionando-o como uma ferramenta eficiente para aumentar a resiliência dos produtores rurais contra eventos de grande escala que impactam a população rural.

Nas regiões mais produtivas do Brasil em termos agrícolas, abrangendo tanto o Sul quanto o Centro-Oeste, recentes secas e chuvas excessivas causaram um aumento substancial nas taxas de sinistralidade da carteira de seguro rural. Segundo a Superintendência de Seguros Privados (SUSEP)³, houve um aumento na sinistralidade do seguro rural. No primeiro semestre de 2022, essa sinistralidade atingiu aproximadamente 281,6%, um aumento acentuado em relação aos 89,7% no mesmo período em 2021. Essa elevação destaca o impacto significativo das condições climáticas adversas na agricultura.

¹ CPEA. PIB do agronegócio brasileiro. **Cepea.org.br**, [S.l.], 17 jun. 2025. Disponível em: <https://www.cepea.esalq.usp.br/br/pib-do-agronegocio-brasileiro.aspx>. Acesso em: 19/08/2025.

² CONAB. Ajustes na área de milho e soja resultam em uma produção de 295,45 milhões de toneladas na safra 2023/2024. **Gov.br**, [S.l.], 19 jul. 2024. Disponível em: <https://www.gov.br/conab/pt-br/assuntos/noticias/ajustes-na-area-de-milho-e-soja-resultam-em-uma-producao-de-295-45-milhoes-de-toneladas-na-safra-2023-2024-1>. Acesso em: 19/08/2025.

³ SUSEP. SES. Sistema de Estatísticas da SUSEP. **SUSEP**, [S.l.], [S.d.]. Disponível em: <https://www2.susep.gov.br/menuestatistica/ses/resp-premiosesinistros.aspx>. Acesso em: 19/08/2025.

Uma consequência importante desse cenário catastrófico é que o país pode enfrentar, ou já enfrenta, um problema de escassez na oferta de seguros. As resseguradoras, que historicamente assumem a maior parte dos riscos rurais segurados, podem se retrair acentuadamente na oferta de limites de contratação, essenciais para que os produtos sejam oferecidos aos segurados⁴. Isso constitui um cenário desfavorável, considerando a relevância que o agronegócio tem nas economias nacional e internacional. Fica clara a importância de estudos debruçados sobre este setor, inclusive para fins de regulação.

Neste contexto, modelos mais sofisticados para estimar corretamente os riscos e geri-los, com ferramentas avançadas para quantificar provisões técnicas e para precificação, são fundamentais tanto para seguradoras quanto para a promoção de políticas públicas. Uma modelagem de risco mais precisa e adequada traz mais segurança para o lado da oferta, ou seja, para (res)seguradoras, reduzindo a incerteza no processo de estimativa e também sendo capaz de auxiliar na diminuição das margens de risco (incerteza) nos prêmios, aumentando assim a penetração e a cobertura de seguros, o que é de grande interesse para o Brasil.

No campo do seguro rural, uma parte significativa da literatura tem se dedicado ao desenvolvimento de metodologias para prever efetivamente sinistros de seguro rural. Entre eles, o estudo conduzido por Mota, Miquelluti e Ozaki (2020) focou principalmente em prever a ocorrência de sinistros usando modelos de aprendizagem de máquina. No entanto, ainda há uma lacuna na previsão da magnitude financeira desses sinistros.

Nosso trabalho estende essas metodologias para estimar não apenas a probabilidade, mas também o potencial valor das perdas dos sinistros. Para isso, utilizamos bancos de dados de várias fontes, incluindo o Ministério da Agricultura (MAPA), a Unidade de Pesquisa Climática da Universidade de East Anglia (fornecendo dados meteorológicos, como precipitação e temperatura, coletados de estações meteorológicas em todo o país, por localização – latitude e longitude), a Agência Nacional de Águas (umidade do solo) e dados da Columbia Climate School (índices de ocorrência de El Niño e La Niña).

⁴ WALENDORFF, Rafael. Perdas elevadas reduzem apetite de seguradoras. **Valor**, São Paulo, 16 jun. 2023. Disponível em: <https://valor.globo.com/agronegocios/noticia/2023/06/16/perdas-elevadas-reduzem-apetite-de-seguradoras.ghml>. Acesso em: 19/08/2025.

Os resultados obtidos neste artigo são relevantes para a formulação de políticas públicas, incluindo subsídios do governo federal. Apresentamos também resultados sobre a precificação em municípios, seguindo uma metodologia similar descrita em Liu e Ramsey (2023). Além disso, ao incorporar dados de produção municipal total, que incluem áreas seguradas e não seguradas, estimamos também perdas totais potenciais nos municípios.

2. Dados

Nesta seção fornecemos uma visão geral dos conjuntos de dados utilizados no processo de modelagem atuarial.

2.1 Dados de Seguro Rural

Obtivemos os dados das apólices de seguro rural no site do MAPA⁵. As apólices estão incluídas no Programa de Subvenção ao Seguro Rural (PSR). O objetivo principal do PSR é fornecer assistência financeira para a compra de apólices de seguro, cobrindo uma parte dos prêmios cobrados dos produtores rurais. Restringimos nossa análise a contratos iniciados entre março de 2016 e fevereiro de 2022 e selecionamos apenas aqueles baseados em produtividade. Além disso, focamos em apólices de cultivo de soja, que representam 48% de todas as apólices no período analisado. Uma descrição completa das principais variáveis contidas nesses dados é dada na Tabela 1.

Definimos o ano da safra como abrangendo todas as apólices com datas de início de vigência variando de março de um ano a fevereiro do ano subsequente. Essa delimitação está alinhada com os cronogramas de plantio de soja usuais, que tipicamente se iniciam em outubro e culminam em janeiro⁶. As apólices com data de início entre março e outubro correspondem a aproximadamente 95% das apólices de soja.

Em nosso banco de dados, a seca é responsável por 73.0% dos sinistros relacionados à soja. A segunda principal causa é o excesso de chuva, representando 15.0% dos sinistros.

⁵ MAPA. Sistema de Subvenção Econômica ao Prêmio do Seguro Rural (SISSER). **Portal de Dados Abertos do Ministério da Agricultura e Pecuária**, [S.l.], [S.d.]. Disponível em: <https://dados.agricultura.gov.br/dataset/sisser3>. Acesso em: 19/08/2025.

⁶ Calendário de Plantio e Colheita de Grãos no Brasil 2020. Companhia Nacional de Abastecimento (CONAB). Disponível em: <https://www.gov.br/conab/pt-br/aceso-a-informacao/institucional/publicacoes/arquivos-de-paginas/calendriozalteradozmarz2021.pdf>.

2.2 Dados climáticos e do solo

Diversos estudos destacam a importância dos dados históricos de clima, especialmente precipitação e temperatura, na determinação da produtividade para o seguro rural e, como consequência, influenciando as perdas nas lavouras. Por exemplo, Liu e Ramsey (2023) utilizaram metodologias bayesianas para prever perdas com base dados climáticos. Da mesma forma, Zhu, Porth e Tan (2019) destacaram o papel dos dados climáticos na melhoria da precisão das previsões de produtividade das culturas. Além disso, Berlato e Fontana (2001) investigaram os efeitos do fenômeno El Niño-Oscilação Sul (ENSO) na precipitação e os relaciona aos requisitos específicos das culturas. Os dados de precipitação e temperatura que utilizaremos são provenientes do conjunto de dados de *Gridded Time Series da Climate Research Unit* (CRU-TS – consultem-se Harris *et al.* (2020)) da Universidade de East Anglia. Esse conjunto de dados em particular é renomado por seu amplo uso em diversos domínios e aplicações, especialmente em modelos que dependem fortemente de dados locais de clima como discutido em Harris *et al.* (2020).

Tabela 1 – Variáveis no conjunto de dados de seguro rural obtidos do Ministério da Agricultura do Brasil

Variável	Descrição
Latitude	Latitude da fazenda
Longitude	Longitude da fazenda
Município	Município onde a fazenda está localizada
DataContrato	Data de início do contrato
Estado	UF onde a fazenda está localizada
Área	Área segurada (ha)
Produtividade	Produtividade segurada (kg/ha)
ValorSinistro	Valor do sinistro (BRL)
Prêmio	Prêmio cobrado (BRL)
Subvenção	Valor da subvenção (BRL)

Os mapas representados na Figura 3 mostram os padrões de precipitação em dezembro em diferentes municípios brasileiros para os anos agrícolas de 2016/2017 a 2021/2022. Notadamente, a safra de 2021/2022 registrou precipitações acumuladas mais baixas no sul do Brasil. Por outro lado, a Figura 4 mostra as temperaturas médias durante o mesmo período. Especialmente em dezembro (e janeiro, não mostrado aqui) da safra de 2021/2022, o sul do Brasil experimentou algumas das temperaturas mais altas em comparação aos anos agrícolas anteriores.

Para relacionar essas variáveis com a produtividade da cultura da soja, é necessário recorrer a estudos de agronomia sobre a planta. O ciclo fenológico da soja pode ser segmentado em quatro fases: estabelecimento (≈ 12 dias); fase vegetativa (≈ 40 dias); fase reprodutiva (≈ 40 dias); e maturação (≈ 28 dias). Zanatta (2019) observou que as lavouras de soja são particularmente vulneráveis à escassez de água durante a fase reprodutiva.

Como delineado na Seção 2.1, o plantio de soja no Brasil geralmente começa entre outubro e dezembro. Além disso, o Zoneamento Agrícola de Risco Climático (ZARC)⁷ recomenda esse mesmo período – de outubro a dezembro – para a maioria dos municípios como a janela ótima para plantar sementes de soja. Considerando que o período da soja mais sensível à água inicia aproximadamente 52 dias após a semeadura, utilizamos os dados mensais acumulados de precipitação de novembro a fevereiro como um indicador representativo de quando as lavouras de soja são mais suscetíveis à escassez de água. Alguns estudos têm consistentemente destacado uma relação entre a variabilidade climática do fenômeno El Niño–Oscilação Sul (ENSO) e anomalias na precipitação (Berlato e Fontana (2001). Ainda: Diaz, Studzinski e Mechoso (1998)). Especificamente, para as regiões Sul, Sudeste e Centro-Oeste sul do Brasil, El Niño tende a gerar precipitação que ultrapassa a média climatológica, enquanto La Niña geralmente resulta em chuvas abaixo dessa média (Berlato e Fontana (2001)). Os efeitos mais pronunciados tanto do El Niño quanto da La Niña na precipitação são observados na primavera e início do verão – coincidentemente, este é também o período em que a soja tem as maiores necessidades hídricas.

⁷ MAPA. Painel de indicação de riscos. **Gov.br.**, [S.l.], [S.d.]. Disponível em: <https://mapa-indicadores.agricultura.gov.br/publico/extensions/Zarc/Zarc.html>. Acesso em: 19/08/2025

Neste contexto, consideramos dois índices: o Índice Oceânico Niño (ONI) e o Índice de Oscilação Sul (SOI). O ONI é definido como a média móvel de três meses de anomalias de temperatura – desvios nas águas superficiais do Pacífico Tropical Centro-Leste, próximo à Linha Internacional de Data. Em resumo, quando o ONI excede um limite de 0.5 ou o ultrapassa, sugere a ocorrência de um evento El Niño. Por outro lado, um valor abaixo de -0.5 indica a ocorrência de La Niña (ver Banholzer e Donner (2014)). Em contraste, o SOI é calculado a partir das diferenças padronizadas na pressão do nível do mar (SLP) observadas entre Tahiti e Darwin, Austrália. Um valor de SOI abaixo de -1.5 é indicativo de condições de El Niño, enquanto um valor acima de 1.5 aponta para La Niña. As séries temporais para SOI e ONI são fornecidas na Figura 6. Em nosso trabalho, incorporamos o número de vezes em que SOI e ONI excedem 1.5 (ou caem abaixo de -1.5) dentro de um ano da safra específico como covariáveis, para representar a intensidade e continuidade de El Niño/La Niña.

Também levamos em consideração a capacidade de água disponível (AWC, ou “*Available Water Capacity*”), uma medida da capacidade do solo de reter água para as plantas. Esse conjunto de dados, mantido pela Agência Nacional de Águas (ANA) e Universidade Federal do Paraná⁸, dados de AWC para municípios brasileiros em 2021. Os dados de AWC são visualmente representados na Figura 5.

3. Modelagem

Nesta seção, descrevemos o processo de modelagem para prever perdas de seguro em safras de soja. São definidas a variável-resposta que representa as perdas municipais e as covariáveis que compõem o modelo preditivo. Em seguida, apresentamos várias escolhas para os modelos de dois estágios, consistindo em modelos lineares generalizados, modelos aditivos generalizados e florestas aleatórias.

3.1 Modelo Atuarial

Para definir o modelo, denota-se:

$$L_{i,j} = \frac{\sum_{k \in i} Y_{k,j} \frac{S_j}{S_k}}{\sum_{k \in i} A_{k,j} P_{k,j}}$$

onde $L_{i,j}$ é a perda no município i na safra j , $A_{k,j}$ é a área segurada para a fazenda k na safra j , S_k representa o preço da soja em BRL na data de assinatura do contrato, enquanto S_j é o preço da soja em BRL em 30 de dezembro de 2022. Já $Y_{k,j}$ representa o valor do sinistro para a fazenda k na safra j e $P_{k,j}$ a produção de soja segurada.

⁸ Consultar dados planilhados em: https://metadados.snirh.gov.br/geonetwork/srv/api/records/28fe4baa-66f3-4f6b-b0d2-890abf5910c4/attachments/CAD_AWC_media_municipal.xlsx

O conjunto de variáveis que consideramos para a modelagem é: (i) Ano da safra, de 1 a 6, (ii) Latitude do município, (iii) Longitude do município, (iv) AWC, (v) Prec. em novembro, (vi) Prec. em dezembro, (vii) Prec. em janeiro, (viii) Prec. em fevereiro, (ix) Temp. média em novembro, (x) Temp. média em dezembro, (xi) Temp. média em janeiro, (xii) Temp. média (Celsius) em fevereiro, (xiii) Número de vezes $SOI < -1.5$, (xiv) Número de vezes $SOI > 1.5$, (xv) Indicador – pelo menos um sinistro e (xvi) Perda relativa: razão entre total dos sinistros deflacionados e a produção.

Considerando que Z_{ij} é um indicador de pelo menos um sinistro no município i no ano da safra j , assumimos o seguinte processo gerador dos dados: $D = \{(X_{ij}, Z_{ij}, L_{ij}) : i = 1, \dots, n_j, j = 1, \dots, J\}$, onde n_j é o número de municípios com uma área segurada positiva no ano da safra j e J é o número total de anos da safra.

Além disso, $X_{ij} \in \mathbb{R}^d$ é um vetor de preditores no município i no ano da safra j , $Z_{ij} \in \{0, 1\}$ é igual a 1 no caso em que o município tem pelo menos um sinistro no ano da safra j e 0, caso contrário, e L_{ij} é a perda relativa definida em 3.1.

Nosso objetivo principal é prever as perdas esperadas L_{ij} com base nas covariáveis X_{ij} . Isso é estimado pela expectativa condicional $E(L_{ij} | X_{ij})$ que pode ser fatorada, pela propriedade de torre, como:

$$E[L_{ij} | X_{ij}] = E[L_{ij} | X_{ij}, Z_{ij} = 1] \cdot P(Z_{ij} = 1 | X_{ij}), \quad (1)$$

sob a suposição de que $E(L_{ij} | X_{ij}, Z_{ij} = 0) = 0$. Tal suposição simplesmente afirma que se não houver sinistros no município i no ano da safra j , então a perda relativa é igual a zero.

Portanto, a partir da Equação 1, começamos ajustando um modelo de classificação para Z_{ij} dado X_{ij} , usando todo o conjunto de dados. Em seguida, ajustamos um modelo de regressão para L_{ij} restrito a instâncias em que $Z_{ij} = 1$. Assim, consideramos a seguinte hierarquia de modelo de duas etapas:

$$\begin{aligned} Z_{ij} | \mathbf{X}_{ij} &\sim F_{ij} \\ Y_{ij} | Z_{ij}, \mathbf{X}_{ij} &\sim G_{ij} \end{aligned}$$

Os modelos utilizados nas etapas de frequência e severidade foram: Modelos Lineares Generalizados (GLM), Modelos Aditivos Generalizados (GAM) e Florestas Aleatórias (*floresta aleatória*). A seguir, apresentamos uma breve descrição de cada um desses modelos:

- **Modelos Lineares Generalizados (GLM):** Este modelo é uma extensão dos modelos de regressão linear que permite à resposta ser distribuída de acordo com diferentes distribuições da família exponencial (por exemplo, normal, Poisson, binomial). Ele utiliza uma função de ligação para conectar a esperança da resposta ao preditor linear. Para uma descrição detalhada dos Modelos Lineares Generalizados e suas aplicações, ver Cordeiro, Demétrio e Moral (2024). Para o modelo de frequências, que indica ocorrência de sinistro no município, foi utilizada a regressão logística, enquanto que o modelo de severidade ajustado foi uma regressão Gama.
- **Modelos Aditivos Generalizados (GAM):** Este modelo estende os GLMs ao permitir que a relação entre a variável-resposta e os preditores seja aditiva e não linear. Isso é alcançado através do uso de funções suaves, como *splines*, para modelar cada preditor individualmente.
- **Florestas Aleatórias (Random Forest):** Este é um método de aprendizado de máquina que opera construindo múltiplas árvores de decisão durante o treinamento e combinando as predições através da classe mais frequente (classificação) ou da média das predições (regressão) das árvores individuais, ajustadas utilizando-se reamostragens dos dados observados. As florestas aleatórias são conhecidas por sua capacidade de lidar com conjuntos de dados de alta dimensionalidade e complexidade, permitindo capturar efeitos não lineares e interações entre variáveis.

Para mais detalhes sobre os modelos GAM e florestas aleatórias, o leitor interessado pode consultar James *et al.* (2023), bem como Hastie, Tibshirani e Friedman (2009).

3.2 Avaliação de Desempenho dos Modelos

Para avaliar o desempenho preditivo desses algoritmos, empregamos um procedimento iterativo, descrito da seguinte forma: em cada iteração, o conjunto de treinamento inclui apólices da safra 1 até a safra $i - 1$, enquanto o conjunto de teste inclui políticas do ano da safra i . Esse procedimento é realizado para $i = 2, \dots, 6$, onde $i = 2$ é o ano da safra 2017/2018, e $i = 6$ é o ano da safra 2021/2022. Em seguida, medimos a área sob a curva ROC (Hastie, Tibshirani e Friedman (2009)) para o modelo de classificação de sinistros e o erro absoluto médio para o modelo de perda.

A Tabela 3 exibe a média das áreas sob a curva ROC para o modelo de classificação, bem como essa métrica apenas para a última iteração do procedimento (safra 2021/2022 como conjunto de teste). Para ambas as medidas de desempenho preditivo, a floresta aleatória demonstrou desempenho superior à regressão logística. Além disso, a Tabela 4 exibe a média dos erros absolutos médios para o modelo de perda, considerando os vários conjuntos de teste. Também fornece, na terceira coluna, o erro absoluto médio da última iteração. De acordo com os resultados do modelo de classificação, a floresta aleatória também superou o desempenho da regressão Gamma. Portanto, utilizamos um primeiro modelo de floresta aleatória para calcular as probabilidades de sinistros, seguido pela floresta aleatória para o montante de perda.

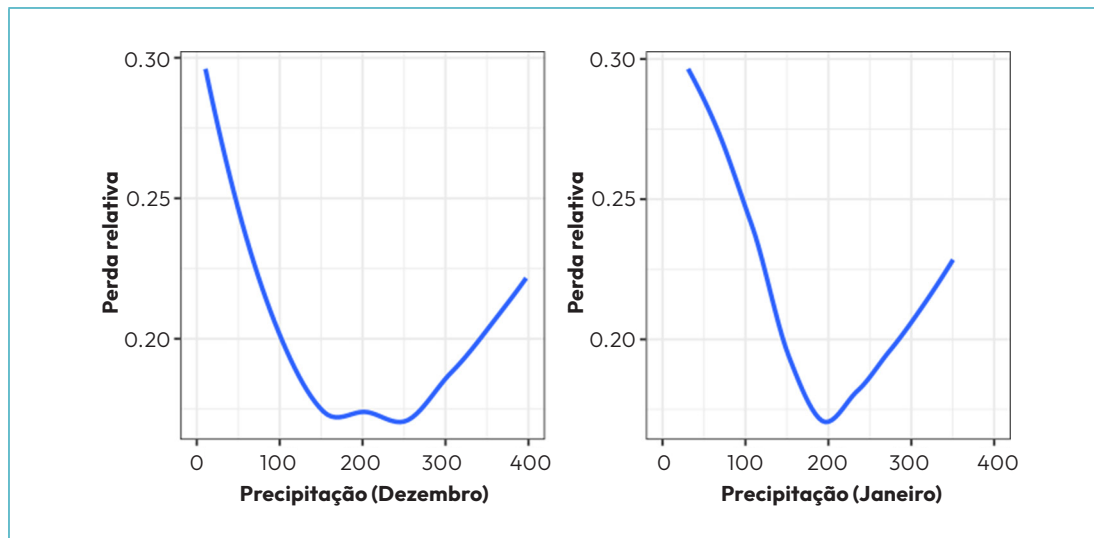
3.3 Interpretabilidade

A Figura 2 apresenta o gráfico de importância das variáveis gerado pelo modelo de floresta aleatória, destacando a precipitação em dezembro como o principal fator na previsão de perdas nas lavouras de soja. Esse resultado é compatível com o ciclo da cultura, já que dezembro corresponde, na maioria das regiões, ao período de maturação da soja – etapa particularmente sensível à escassez hídrica. A latitude surge como a segunda variável mais relevante, o que também está de acordo com as expectativas, dado que os municípios da região Sul, mais suscetíveis a eventos climáticos extremos, tendem a apresentar maiores riscos.

Para compreender com maior precisão a contribuição marginal de cada variável meteorológica no modelo de perdas preditivas, utilizamos gráficos de dependência parcial (PDP), conforme descrito por Friedman (2001) e Zhao e Hastie (2021) – ver Figura 2. Foram analisados os PDPs para as precipitações acumuladas em dezembro e janeiro, identificadas como as variáveis meteorológicas mais relevantes no modelo. Observa-se que, à medida que o volume de chuva em dezembro aumenta, a perda relativa projetada tende a diminuir. No entanto, esse efeito é revertido quando a precipitação ultrapassa aproximadamente 250 mm, indicando que volumes excessivos de chuva passam a contribuir para o aumento das perdas.

Esse padrão sugere que níveis moderados de precipitação em dezembro podem atenuar as perdas, possivelmente em razão de benefícios como umidade do solo ideal, menor necessidade de irrigação e condições agrônômicas favoráveis ao desenvolvimento da cultura. No entanto, quando a precipitação ultrapassa 250 mm mensais, observa-se um aumento nas perdas projetadas. Vale destacar que a variável de perda contempla sinistros decorrentes tanto de seca quanto de excesso de chuva.

Figura 1 – Gráficos de dependência parcial mostrando as previsões médias das perdas relativas com base nos valores totais de precipitação de dezembro e janeiro (com curvas suavizadas em azul)



4. Cálculo do Prêmio e Extrapolação das Perdas para Propriedades Não Seguradas

4.1 Cálculo do Prêmio

Na análise realizada, inicialmente assumimos, de forma hipotética, que os valores de precipitação e temperatura de um determinado ano-safra são conhecidos. No entanto, essa suposição não é realista, vez que tais informações meteorológicas não estão disponíveis no início do ano-safra j . Reconhecendo essa limitação, utilizamos a propriedade da torre (lei da esperança total) para incorporar adequadamente a incerteza associada às covariáveis climáticas ainda não observadas.

Em termos práticos, essa abordagem equivale a considerar a média ponderada de diferentes cenários climáticos, observando a probabilidade de ocorrência de cada um. Para isso, definimos o vetor de covariáveis $X_{ij} = (W_{ij}, X^*)$, em que W_{ij} representa as covariáveis dinâmicas e climáticas específicas do município i na safra j , e X^* corresponde ao vetor de covariáveis estáticas cujos valores são constantes ao longo das safras. Então, aproximamos o valor esperado da perda relativa no município i no ano da safra j por meio da seguinte esperança condicional:

$$E(L_{iJ} | X_i^*) = E\left[E(L_{iJ} | X_i^*, W_{iJ})\right]$$

que pode ser aproximada pela média empírica:

$$E(L_{iJ} | X_i^*) \approx \frac{1}{J-1} \sum_{j=1}^{J-1} \hat{E}[L_{ij} | X_i^*, W_{ij}, Z_{ij} = 1] \cdot \hat{P}(Z_{ij} = 1 | X_i^*, W_{ij})$$

Essa aproximação nos permite estimar a indenização esperada para um município específico no início de uma safra, mesmo na ausência de informações climáticas futuras, o que viabiliza a definição do prêmio de forma prospectiva.

Na etapa seguinte da análise, avaliamos a taxa de sinistralidade com base tanto nos prêmios efetivamente cobrados quanto naqueles estimados pelo modelo. Seja Y_{ij} a perda total observada no município i durante a safra j , então, as taxas de sinistralidade são definidas da seguinte forma:

$$LR_j^{(charged)} = \frac{\sum_{i=1}^{n_j} Y_{ij}}{\sum_{i=1}^{n_j} \pi_{ij}},$$

para $j = 1, \dots, J$, onde $\pi^{(charged)}$ representa o prêmio total cobrado no município i no ano da safra j .

Além disso,

$$LR_j^{(estimated)} = \frac{\sum_{i=1}^{n_j} Y_{ij}}{\sum_{i=1}^{n_j} \hat{\pi}_{ij}},$$

é a *taxa de sinistralidade* para o prêmio estimado com base no modelo proposto. A taxa de sinistralidade para os prêmios cobrados foi igual a $LR^{(charged)} = 2,59$, enquanto a sinistralidade correspondente ao modelo proposto resultou em uma taxa de sinistralidade de $LR^{(estimated)} = 1,07$. Isso evidencia que os prêmios estimados pelo modelo apresentaram uma correspondência mais favorável com a indenização observada em comparação com os prêmios cobrados.

4.2 Extrapolação de Perdas para Propriedades Não Seguradas

Nesta seção, propomos um método para estimar as perdas em toda a área plantada de soja de cada município. Para isso, assumimos que a perda relativa observada na área subsidiada (e segurada) é representativa da perda relativa em toda a região plantada, abrangendo tanto propriedades seguradas quanto não seguradas. Com base nessa suposição, perda total é calculada multiplicando-se a perda relativa prevista pela produção total de soja (em quilogramas) obtida do banco de dados de Produção Agrícola Municipal do Instituto Brasileiro de Geografia e Estatística⁹ (IBGE). Dessa forma, obtemos estimativas das perdas totais de soja em cada município e, por agregação, para o país como um todo. Assim, a perda total de soja atribuída a eventos de seca ou chuva excessiva no município i durante a safra j , denotada por τ_{ij} , é definida como:

$$\hat{\tau}_{ij} = \hat{E}[L_{ij} \mid X_{ij}, Z_{ij} = 1] \cdot \hat{P}(Z_{ij} = 1 \mid X_{ij}) \cdot prod_{ij}$$

onde $prod_{ij}$ é a produção total de soja no município i na safra j .

A Tabela 2 exibe apenas os resultados extrapolados para o Brasil, assim como as perdas observadas advindas do seguro rural (subsidiado). A maior perda observada para áreas seguradas e subsidiadas atingiu R\$ 3,94 bilhões no ano da safra 2021/2022. Isso coincide com o ano que registrou uma perda total extrapolada de R\$ 46,78 bilhões para todo o país.

Tabela 2 – Perda total prevista (em BRL) de culturas de soja no Brasil para cada ano da safra, considerando áreas seguradas e não seguradas, além de perdas observadas (em BLR) de soja apenas para áreas com seguro rural subsidiado (baseado em produtividade)

Safra	Perda Total Extrapolada	Perda Observada
2017/2018	R\$ 9,51 bi	R\$ 0,12 bi
2018/2019	R\$ 17,64 bi	R\$ 0,46 bi
2019/2020	R\$ 22,52 bi	R\$ 0,86 bi
2020/2021	R\$ 8,55 bi	R\$ 0,30 bi
2021/2022	R\$ 46,76 bi	R\$ 3,94 bi

⁹ IBGE. Produção Agrícola Municipal. **Sistema IBGE de Recuperação Automática**, [S.l.], [S.d.]. Disponível em: <https://sidra.ibge.gov.br/tabela/5457>. Acesso em: 19/08/2025.

5. Considerações finais

Neste trabalho, propusemos modelos de predição de perdas em culturas de soja a partir de dados de apólices de seguro subsidiadas pelo Programa de Subvenção ao Seguro Rural (PSR) do Governo Federal Brasileiro. Ao combinar dados das apólices subvencionadas com dados climáticos de precipitação e temperatura do CRU-TS, capacidade de umidade do solo e índices El Niño e La Niña, construímos modelos preditivos baseados em Modelos Lineares Generalizados (GLMs), Modelos Aditivos Generalizados (GAMs) e florestas aleatórias para prever perdas agregadas em nível municipal no Brasil. O modelo de duas etapas (frequência-severidade) baseado em floresta aleatória apresentou capacidade preditiva superior tanto na classificação da ocorrência de sinistros quanto na estimativa da magnitude das perdas.

Além disso, devido à ausência de dados climáticos no início da safra, utilizamos dados de safras anteriores para estimar perdas enquanto marginalizamos variáveis climáticas. Propusemos também um método de extrapolação para áreas não seguradas, assumindo homogeneidade de risco, com base na produção agrícola municipal total. Apesar dos resultados promissores, reconhecemos que o modelo apresenta limitações, sobretudo no que se refere à capacidade de capturar a variabilidade espacial e temporal das condições climáticas futuras. Assim, futuros estudos devem explorar a incorporação de modelos climáticos mais sofisticados, capazes de melhorar a predição de variáveis meteorológicas e, conseqüentemente, refinar a estimativa de perdas. Além disso, recomenda-se o desenvolvimento de modelos em nível de propriedade, com dados individualizados, a fim de capturar heterogeneidades dentro dos municípios.

6. Agradecimentos

Os autores agradecem o apoio financeiro da Fundación MAPFRE, que viabilizou a realização desta pesquisa. Rodrigo S. Targino agradece o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico [Bolsa 200293/2022-2] e da Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro [Auxílios E-26/201.350, E-26/211.426 e E-26/211.578].

7. Referências Bibliográficas

- BANHOLZER, S.; DONNER, S. The influence of different El Niño types on global average temperature. **Geophysical Research Letters**, [S.l.], v.41, n.6, p.2093-2099, 2014.
- BERLATO, M. A.; FONTANA, D. C. Impacts of El Niño and La Niña on agricultural production in southern Brazil and the use of climate forecasts in agriculture. **Applications of climate forecasting for better decision-making processes in agriculture**. Passo Fundo: Embrapa Trigo, 2001. p. 217-241.
- CONAB. Ajustes na área de milho e soja resultam em uma produção de 295,45 milhões de toneladas na safra 2023/2024. **Gov.br.**, [S.l.], 19 jul. 2024. Disponível em: <https://www.gov.br/conab/pt-br/assuntos/noticias/ajustes-na-area-de-milho-e-soja-resultam-em-uma-producao-de-295-45-milhoes-de-toneladas-na-safra-2023-2024-1>. Acesso em: 19/08/2025.
- CORDEIRO, G. M.; DEMÉTRIO, C. G. B.; MORAL, R. A. **Modelos lineares generalizados e aplicações**. São Paulo: Blucher, 2024.
- CPEA. PIB do agronegócio brasileiro. **Cepea.org.br**, [S.l.], 17 jun. 2025. Disponível em: <https://www.cepea.esalq.usp.br/br/pib-do-agronegocio-brasileiro.aspx>. Acesso em: 19/08/2025.
- DIAZ, A. F.; STUDZINSKI, C. D.; MECHOSO, C. R. Relationships between precipitation anomalies in Uruguay and southern Brazil and sea surface temperature in the Pacific and Atlantic oceans. **Journal of Climate**, [S.l.], v.11, n.2, p.251-271, 1998.
- FAO. **Protecting livelihoods**: Linking disaster risk insurance and social protection in agriculture. **FAO.Org.**, [S.l.], 2020. [fonte? Não achei no repositório] <https://openknowledge.fao.org/bitstreams/70f8c4fa-32b1-480f-a0f7-50af8946e27e/download>. Acesso em: 19/08/2025.
- FAO. **Protecting livelihoods**. Linking agricultural insurance and social protection. **FAO.Org.**, [S.l.], 2021. Disponível em: <https://openknowledge.fao.org/server/api/core/bitstreams/70f8c4fa-32b1-480f-a0f7-50af8946e27e/content>. Acesso em: 19/08/2025.
- FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of Statistics**, [S.l.], v. 29, n. 5, p. 1189-1232, 2001.
- HARRIS, I.; OSBORN, T. J.; JONES, P.; LISTER, D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. **Scientific Data**, [S.l.], v.7, n.1, p.109, 2020.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. 2ed. Nova York: Springer, 2009.
- IBGE. Produção Agrícola Municipal. **Sistema IBGE de Recuperação Automática**, [S.l.], [S.d.]. Disponível em: <https://sidra.ibge.gov.br/tabela/5457>. Acesso em: 19/08/2025.
- JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R.; TAYLOR, J. **An introduction to statistical learning**: with applications in python. Nova York: Springer Nature, 2023.
- LIU, Y.; RAMSEY, A. F. Incorporating historical weather information in crop insurance rating. **American Journal of Agricultural Economics**, [S.l.], v.105, n.2, p. 546-575, 2023.

MAPA. Painel de indicação de riscos. **Gov.br.**, [S.l.], [S.d.]. Disponível em: <https://mapa-indicadores.agricultura.gov.br/publico/extensions/Zarc/Zarc.html>. Acesso em:19/08/2025.

MAPA. Sistema de Subvenção Econômica ao Prêmio do Seguro Rural (SISSER). **Portal de Dados Abertos do Ministério da Agricultura e Pecuária**, [S.l.], [S.d.]. Disponível em: <https://dados.agricultura.gov.br/dataset/sisser3>. Acesso em:19/08/2025.

MOTA, A. L.; MIQUELLUTI, D. L.; OZAKI, V. A. Predição de sinistros agrícolas: uma abordagem comparativa utilizando aprendizagem de máquina. **Economia Aplicada**, [S.l.], v.24, n.4, p.533-554, 2020. SUSEP. SES. Sistema de Estatísticas da SUSEP. **SUSEP**, [S.l.], [S.d.]. Disponível em: https://www2.susep.gov.br/menuestatistica/ses/resp_premiossinistros.aspx. Acesso em:19/08/2025.

WALENDORFF, Rafael. Aumento de indenizações leva empresas a deixar de ofertar seguro rural. **Globo Rural**, [S.l.], 16 jun. 2023. Disponível em: <https://globorural.globo.com/negocios/noticia/2023/06/aumento-de-indenizacoes-leva-empresas-a-deixar-de-ofertar-seguro-rural.ghtml>. Acesso em:19/08/2025.

WALENDORFF, Rafael. Perdas elevadas reduzem apetite de seguradoras. **Valor**, São Paulo, 16 jun.2023. Disponível em: <https://valor.globo.com/agronegocios/noticia/2023/06/16/perdas-elevadas-reduzem-apetite-de-seguradoras.ghtml>. Acesso em: 19/08/2025.

ZANATTA, M. A. **Variabilidade espaço-temporal da precipitação e efeito sobre a produtividade da soja**. 2019. Tese (Doutorado) – Universidade de São Paulo, São Paulo, 2019.

ZHAO, Q.; HASTIE, T. Causal interpretations of black-box models. **Journal of Business & Economic Statistics**, [S.l.], v.39, n.1, p.272-281, 2021.

ZHU, W.; PORTH, L.; TAN, K. S. A credibility-based yield forecasting model for crop reinsurance pricing and weather risk management. **Agricultural Finance Review**, [S.l.], v.79, n.1, p.2-26, 2019.

Anexo A – Avaliação do desempenho do modelo

Apresentamos os resultados de desempenho preditivo para os modelos de classificação e regressão descritos na Seção 3.1.

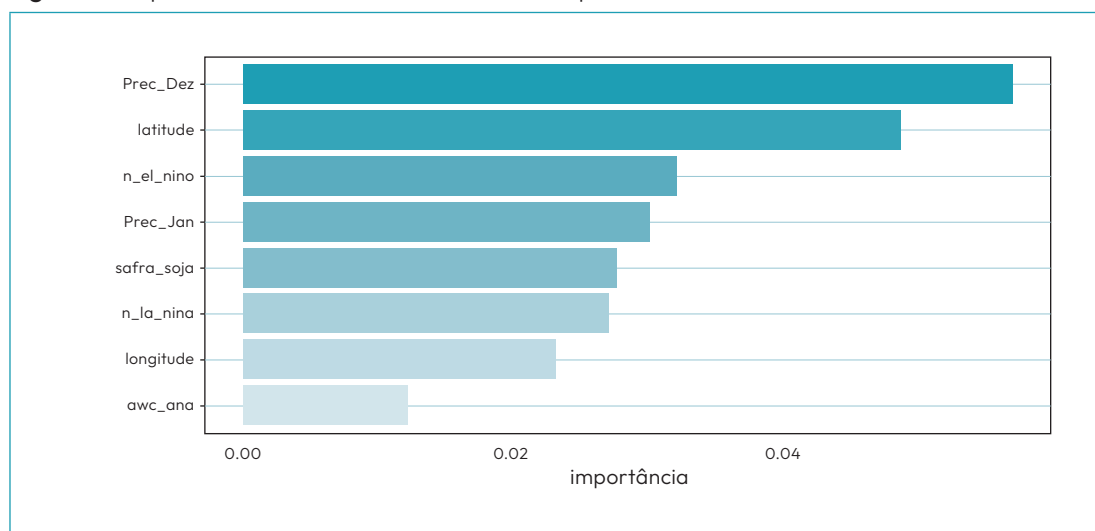
Tabela 3 – Área sob as curvas ROC (AUC) para os modelos de classificação. A segunda coluna apresenta a área média sob as curvas calculada em uma série de conjuntos de teste

Modelo	AUC
Regressão Logística	0.738
GAM	0.753
<i>Floresta Aleatória</i>	0.802

Tabela 4 – Erros médios absolutos dos modelos de perda. A segunda coluna apresenta o erro médio absoluto calculado em uma série de conjuntos de teste

Modelo	MAE
Regressão Gamma	0.644
GAM	0.468
<i>Floresta Aleatória</i>	0.408

Figura 2 – Importância das variáveis nos modelos de perda



Anexo B – Mapas de variáveis meteorológicas

Figura 3 – Distribuição total de precipitação no Brasil para dezembro

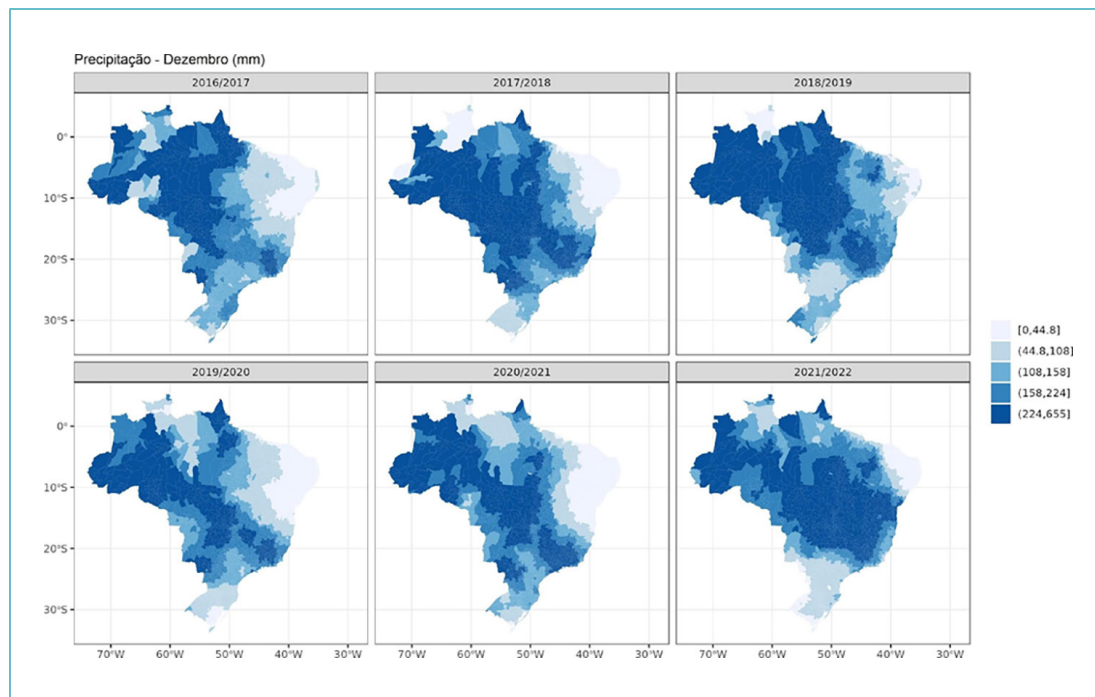


Figura 4 – Distribuição de temperatura média no Brasil para dezembro

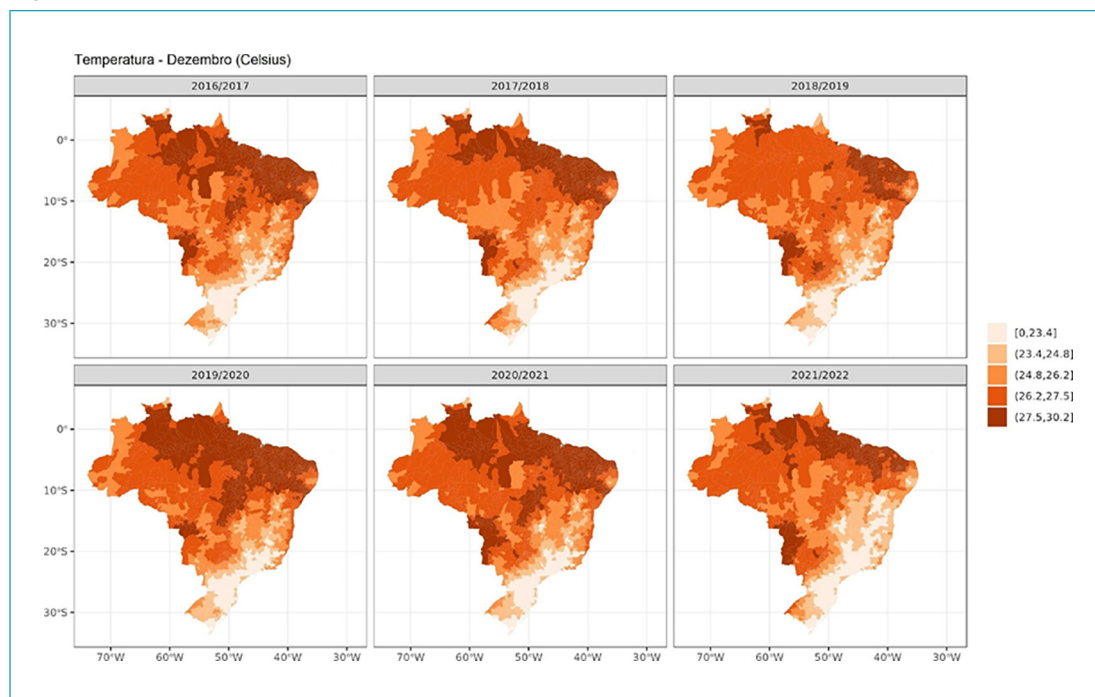


Figura 5 – Capacidade de água disponível (AWC) para municípios no Brasil

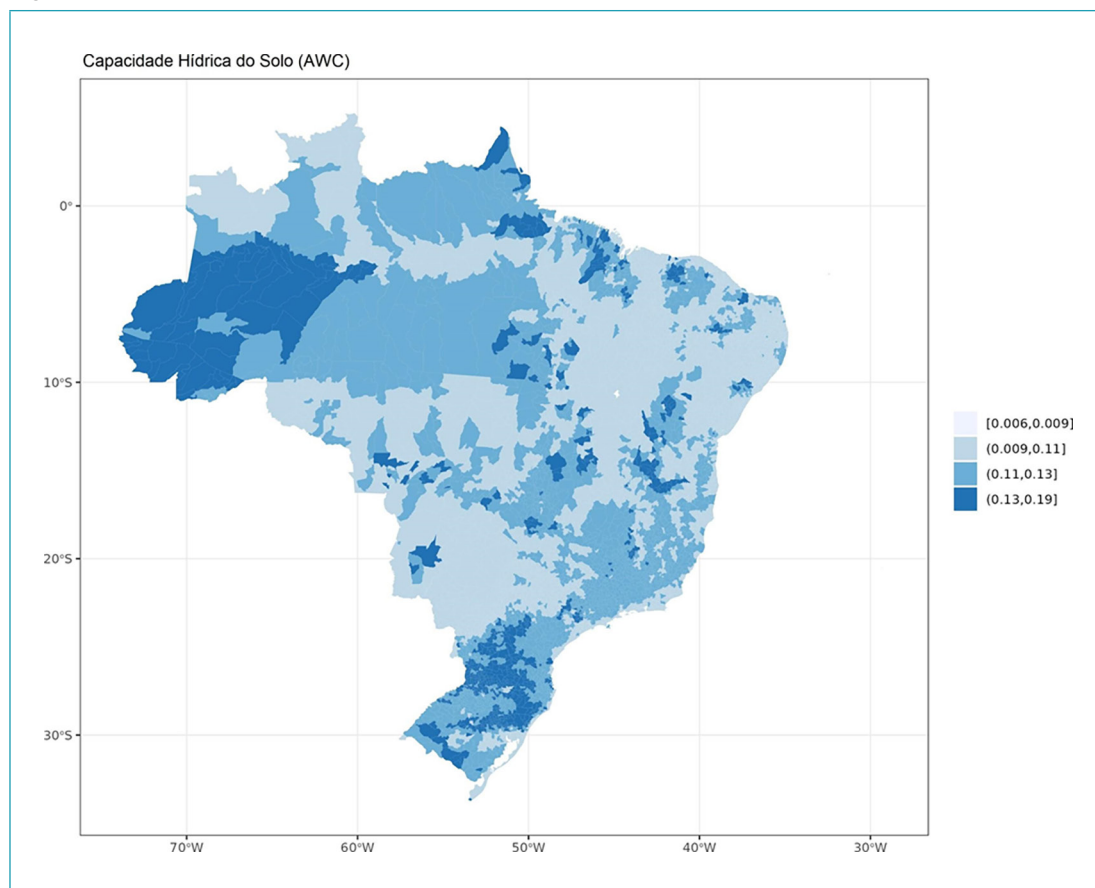


Figura 6 – Séries temporais para o Índice Oceânico Niño (ONI) e Índice de Oscilação do Sul (SOI), abrangendo de janeiro de 2015 a abril de 2023. As linhas vermelhas marcam o limiar para eventos El Niño, enquanto as linhas azuis denotam o limiar para La Niña

