Economia Aplicada, v. 24, n. 4, 2020, pp. 533-554

PREDIÇÃO DE SINISTROS AGRÍCOLAS: UMA ABORDAGEM COMPARATIVA UTILIZANDO APRENDIZAGEM DE MÁQUINA

Arthur Lula Mota [*]
Daniel Lima Miquelluti [†]
Vitor Augusto Ozaki [‡]

Resumo

O seguro agrícola tem ganho maior atenção no Brasil desde o início da década passada, com a implementação do Programa de Subvenção ao Prêmio do Seguro Rural. O presente estudo testou o desempenho de algoritmos de *Machine Learning* para as seguradoras anteciparem a ocorrência de sinistro, elaborando previsões por meio de dados de apólices e bases de dados climáticas entre os anos de 2006 e 2017. Foram testados os algoritmos *Random Forest*, *Support Vector Machine e k-Nearest Neighbours*, sendo que o segundo método mostrou melhor performance preditiva de sinistros quando avaliada pelas métricas Acurácia, Precisão, Taxas de Verdadeiro Positivo e Negativo e Correlação de Matthews. No entanto, todos os métodos apresentaram baixa capacidade preditiva para a ocorrência de sinistros.

Palavras-chave: seguro agrícola, sinistro, previsão, machine learning.

Abstract

Crop insurance has gained greater attention in Brazil since the beginning of the past decade, with the implementation of the Rural Insurance Premium Subvention Program. The present study tested the performance of Machine Learning algorithms for insurers to forecast the occurrence of a claim, using data from policies and climate databases between the years of 2006 and 2017. The Random Forest, Support Vector Machine and k-Nearest Neighbors algorithms were tested, and the second method showed a better predictive performance of claims when evaluated by the metrics Accuracy, Precision, Positive and Negative True Rates and Matthews Correlation. However, all methods presented a low predictive capacity for the occurrence of claims.

Keywords: crop insurance, insurance claim, forecast, machine learning.
JEL classification: G22, Q13, Q18.
DOI: http://dx.doi.org/10.11606/1980-5330/ea161194

[*] Exame Research. E-mail: arthurmota@usp.br
[†] LES - ESALQ/USP. E-mail: danielmiq@usp.br
[‡] LES - ESALQ/USP. E-mail: vitorozaki@usp.br

Recebido em 19 de agosto de 2019 . Aceito em 28 de outubro de 2020.

1 Introdução

A atividade das seguradoras é de grande importância para a sociedade, garantindo a estabilidade de mercados, desempenhando funções importantes na proteção de empresas e pessoas para uma grande gama de riscos. Em grandes linhas, o negócio da seguradora é transferir para si o eventual custo dos riscos que os tomadores de seguros não podem ou não gostariam de arcar. A apólice protege o segurado quando há a ocorrência de um sinistro, que nada mais é do que "a manifestação concreta do risco previsto no contrato de seguro e que ocasiona prejuízo ou responsabilidade" (Oliveira 2005). Na ótica da seguradora, a possibilidade de prever sinistros é fundamental para estabelecer um contrato com esperança não negativa de lucro, além de auxiliar na estimação das provisões técnicas que as empresas são obrigadas a alocar.

Uma das atuações do principal órgão regulador, a Superintendência de Seguros Privados (SUSEP), é feita na regulação preventiva por meio das provisões e reservas técnicas. Elas são compostas por um montante de recursos que as seguradoras precisam manter como forma de garantia para eventos e obrigações futuras, entrando como passivo das seguradoras em linha com o estabelecido pelo Conselho Nacional de Seguros Privados (Rodrigues & Martins 2009). A subestimação da probabilidade de um evento de sinistro pode impactar diretamente a saúde financeira de uma empresa de seguros, dado que o nível de provisões pode não ser suficiente para a cobertura do evento.

A possibilidade de melhoria na predição de ocorrência de sinistros é alvo de estudos por parte das seguradoras e o presente trabalho tenta contribuir nesse sentido. Buscou-se a aplicação de métodos envolvendo algoritmos de aprendizagem de máquinas para a predição, dado que uma taxa maior de acerto na predição dos sinistros auxilia em um planejamento financeiro e no estabelecimento de novos contratos de seguro. As aplicações desses modelos no âmbito nacional e internacional são variadas, sendo que as referências não se esgotam nos exemplos acima. Cabe destacar algumas abordagens poucos utilizadas na previsão das ocorrências de sinistro e que mostraram boa performance em Pijl (2017): algoritmos do Random Forest (RF) e Support Vector Machine (SVM), que foram utilizados para estimar a ocorrência de sinistros no ramo automotivo, com 10 mil observações e 14 variáveis para predição, mostrando performance superior a 95% de acurácia e ressaltando que o SVM foi o modelo com erro mínimo na estimação no volume de sinistros.

As seguradoras trabalham com técnicas estatísticas e matemáticas para estimar as condições futuras e os custos resultante de uma operação, tendo a ciência atuarial se dedicado nesse sentido desde o seu início. A informação antecipada por meio da previsão impacta toda a programação da seguradora. Um caso especial e importante para o planejamento financeiro das seguradoras reside nas provisões técnicas, um montante de recursos que as seguradoras precisam manter como forma de garantia para eventos e obrigações futuras, entrando como passivo das seguradoras em linha com o estabelecido pelo Conselho Nacional de Seguros Privados (Rodrigues & Martins 2009).

As provisões garantem que a companhia de seguros terá a capacidade de reunir as estimativas necessárias, com o objetivo de honrar os compromissos que assumiu perante os assegurados. O grande cuidado que o mercado de seguro tem é de evitar os erros elevados nas estimativas desse volume e que podem resultar em alguns problemas, tais como: em caso de provisão excessiva, a seguradora pode ter sua rentabilidade afetada, tendo que sustentar

custos que eram desnecessários; em caso de provisão insuficiente, há o risco de um cenário de insolvência. Cabe destacar que a maior parte da alocação dessas provisões segue as metodologias propostas pela SUSEP, mas ainda há provisões que carecem de metodologias e que as próprias seguradoras precisam desenvolver e encaminhar como notas técnicas para o órgão fiscalizador (Brasil 2017).

Embora não atuariais, os presentes métodos apresentados no trabalho podem auxiliar tanto na estimação das Outras Provisões Técnicas (OPT), que é uma complementação das Provisão de Prêmios Não Ganhos (PPNG), que representa o valor esperado a pagar relativo a despesas e sinistros a ocorrer. Além desta, pode auxiliar no cálculo da Provisão de Sinistros Ocorridos e Não Avisados (IBNR), que não tem metodologia determinada pela SUSEP, podendo antecipar a ocorrência dos sinistros e preparar a seguradora, juntamente com modelos auxiliares que buscam identificar o período do ano que o sinistro acontece, destacando as técnicas usuais como o chain ladder proposto por Harnek (1966) e a sua versão estocástica de Mack et al. (1994), que se baseia no algoritmo chamado triângulo de run-off. No entanto, para que seu uso incorra em sucesso, deve-se considerar a incorporação de outras variáveis e metodologias à análise, pois, conforme demonstrado neste trabalho, a utilização desses métodos por si só, não garante os resultados esperados.

Outra possível utilidade para os resultados dos métodos apresentados é no auxílio do cálculo de prêmio de seguro, visto que a SUSEP não define a metodologia. Além de despesas administrativas, impostos e demais custos, o prêmio também deve cobrir as despesas com sinistros, que podem ser estimadas a partir das predições dos métodos abordados assim como feito em Pijl (2017), utilizando a experiência passada e calculando a indenização esperada.

O diferencial do estudo é tratar especificamente dos sinistros do seguro agrícola, ramo que está crescendo no Brasil e que saiu de R$ 1,46 bilhões de importância segurada em 2006 para próximo de R$ 10,59 bilhões em 2018, tendo atingindo um pico de R$ 17,52 bilhões em 2014. A atividade agrícola incorre em riscos particulares, em que a probabilidade de ocorrência de sinistro não é independente entre os segurados, dado que o evento gerador reflete mudanças em variáveis climáticas (chuva, geada, temperatura etc.), tendo impactos em larga escala espacial. Estudos sobre esse segmento ainda são escassos quando comparado a outros ramos do mercado de seguro nacional[1], sobretudo quando se busca métodos para a predição de ocorrências de sinistros agrícolas. O trabalho buscou verificar se os métodos Random Forest, Support Vector Machine e k-NN apresentam boa capacidade de previsão da ocorrência de sinistros, dado que essas metodologias estão sendo aplicadas a outras áreas do conhecimento e outros ramos de seguro (Castro & Braga 2011, Pijl 2017).

O trabalho se inicia com uma breve revisão da literatura. Na sequência é dado um panorama geral sobre os eventos de sinistro no mercado agrícola brasileiro, seguido da apresentação da base de dados e da metodologia empregada. Por fim, o artigo é finalizado com a discussão dos resultados encontrados, avaliando o desempenho das predições e a conclusão.

[1] Por exemplo, o número de artigos com o termo "seguro agrícola" produzidos em âmbito nacional entre 2006 e 2018 é de apenas 82, ao passo que aqueles com termo "seguro de automóvel" para igual período chega a 215, conforme os dados do Portal de Busca Integrada (PBI) da Universidade de São Paulo.

2 Revisão de literatura

A modelagem envolvendo sinistros no mercado de seguro está concentrada no ramo do seguro automotivo, que apresenta o maior volume de estudos, tanto pela importância econômica desse mercado, quanto pela maior disponibilidade de dados. Nesse contexto, diversas abordagens foram testadas, por exemplo, o estudo de Ye et al. (2018) Ye et al. (2018) trabalhou na projeção dos sinistros do segmento automotivo, mostrando que há um ganho de performance preditiva ao combinar linearmente o resultado de diversos modelos como regressão linear, regressão quantílica e regressões adaptativas. Já Yang et al. (2018) utilizou modelos para a distribuição Poisson-Tweedie por meio de algoritmo *gradient boosting*, baseado em árvore, para prever o tamanho dos sinistros automotivos, mostrando superioridade aos métodos existentes no sentido de gerar previsões mais precisas e ajudando a resolver o problema de seleção adversa. Por sua vez, Baumgartner et al. (2015) buscou estimar a perda total baseado em um modelo bivariado que trata do tamanho do sinistro e sua quantidade (número de chamados). Os autores utilizaram um Modelo Linear Generalizado Misto (GLMM) bivariado de forma a capturar relações de dependência entre as variáveis estudadas para o mercado de seguro automotivo alemão, mostrando um ganho de performance preditiva na estimação da perda total e mitigação do risco por parte da seguradora.

Na literatura nacional também há predominância no ramo automotivo, com Zaniboni & Montini (2015) focando na estimação da probabilidade de ocorrência do sinistro e do número de sinistros, utilizando as distribuições Poisson Inflada de Zeros (ZIP) e Binomial Negativa Inflada de Zeros (ZINB), motivados pela alta frequência de 0 na distribuição de probabilidade de sinistros. As variáveis utilizadas se aproximam dos estudos internacionais com destaque para características do motorista e do carro, estado de residência, número de expostos e importância segurada. Freitas (2010), por sua vez, utilizou o modelo Logit para estimar a probabilidade da ocorrência de um sinistro nos estados brasileiros no mercado automotivo, encontrando que estados com riscos diferentes apresentavam a mesma probabilidade de sinistros para níveis de prêmios diferentes, o que poderia sugerir uma precificação errada por parte das seguradoras.

No caso do sinistro agrícola, Sousa (2010) utilizou modelos lineares generalizados (GLM) para modelar os dados de sinistros agrícolas do mercado brasileiro, mais especificamente no Rio Grande do Sul, encontrando evidências de que a variação acumulada e a temperatura média dos 15 municípios pesquisados no ano de 2004 não exerceram influência no montante de sinistros registrados, considerando um modelo com distribuição Tweedie, ao passo que o estudo encontra evidências de que há influência sobre o número de sinistros ao considerar um modelo que trabalha com a distribuição Binomial Negativa.

É importante notar que trabalhos envolvendo a predição com dados binários por meio de aprendizagem de máquinas podem sofrer interferência em seu desempenho caso a base de dados seja desbalanceada, isto é, uma participação muito elevada de apenas umas das categorias (frequência muito maior de 0 do que 1, por exemplo). Como ressalta Menardi & Torelli (2014), se as classes não forem perfeitamente separáveis ou se o problema tiver alta complexidade, o não tratamento do desequilíbrio leva a consequências pesadas, tanto na estimação do modelo quanto na avaliação da sua precisão.

O modelo tende a apresentar um elevado viés para a categoria conside-

rada majoritária, dado que busca minimizar o erro e pode considerar o foco na variável majoritária como um caminho mais "fácil" na busca por essa minimização, interferindo negativamente em situações em que o custo de prever erroneamente a classe minoritária é muito alto em comparação com o custo da majoritária.

Um exemplo pode ser dado no caso de diagnósticos de câncer ou não câncer, e até na própria estimativa de sinistro ou não sinistro (Abd Elrahman & Abraham 2013), em que o erro da predição apresenta um impacto negativo elevado. Weiss et al. (2007) discutiu o problema, tratando uma situação de modelo de árvore de decisão usando dois métodos: a sobreamostragem (oversampling), que replica as observações de forma aleatória da classe minoritária, ou a subamostragem (undersampling), que reduz o volume da majoritária, de forma a deixar as duas classes praticamente balanceadas. Os autores encontraram evidências de que as duas práticas são igualmente boas para melhorar a performance do modelo, driblando o problema da base desbalanceada e o seu viés.

Outros estudos também relatam que os dois métodos se revezam em qualidade do aumento de performance, dependendo dos dados e de cada modelo, sendo que a maior desvantagem da subamostragem é excluir dados potencialmente úteis para o processo de treinamento, ao passo que a sobreamostragem aumenta o tamanho do conjunto e o seu tempo de treinamento (Bekkar & Alitouche 2013).

Para tratar especialmente do problema do sinistro do segmento de seguro agrícola, foi necessário verificar as possíveis variáveis que impactam na produção. Pensando em efeitos do clima, Berlato et al. (2005) mostram para o caso do milho que há uma forte tendência do El Niño em favorecer a cultura na região Sul, com ganhos de produtividade, ao passo que o evento La Niña gera efeito contrário. Os impactos na região Sul são de grande interesse, dado que o mercado de seguro agrícola é concentrado naquela região, além de ter elevada representatividade no conjunto de dados do presente estudo.

Há mais leituras mostrando impactos desses fenômenos em outras culturas, tais como Cirino et al. (2015) que mostra os efeitos desses fenômenos na região Nordeste e Sul do país, com efeitos negativos do El Niño no primeiro e La Niña no segundo. Já Cunha et al. (1999) avaliou estados do Sul, Sudeste e Centro-Oeste, encontraram-se impactos dos dois fenômenos na produtividade agrícola da região, revezando entre efeitos positivos e negativos. Ainda assim, cabe apontar que poucos estudos exploraram as interações entre o seguro agrícola e as previsões baseadas nos eventos El Niño e La Niña, sendo sempre de fontes internacionais (Liu et al. 2008, Cabrera et al. 2006).

É importante ressaltar que a região do cultivo impacta no sucesso ou não da colheita, dado que os riscos são variáveis ao longo do país, de um município para o outro, conforme mostra o Zoneamento Agrícola de Risco Climático (ZARC) do Ministério da Agricultura, Pecuária e Abastecimento (MAPA) e que usa a metodologia da EMBRAPA (Steinmetz & SILVA 2017). Segundo Cunha & Assad (2001), dentre os resultados do Programa de Zoneamento Agrícola do MAPA, que traz orientações dos períodos de semeadura por município, cultura e tipo de solo, há a possibilidade de redução das taxas de sinistralidade.

Outra variável de interesse para a predição de sinistro são as próprias seguradoras. De fato, Borde et al. (1994) e a sua revisão de literatura mostram que há grande divergência entre níveis de risco no mercado das seguradoras,

tanto pela composição da carteira, quanto pelas variáveis financeiras. Dessa forma, seguradoras podem se expor a risco de diferentes formas em cada ramo, com o objetivo de diversificar seu portfólio. Os autores ainda sugerem que o aumento no volume de prêmio representa um aumento na exposição, reflexo de um nível mais alto de risco. Além disso, o prêmio é uma sinalização importante de avaliação de risco a priori por parte das seguradoras, sendo considerado atuarialmente justo no momento em que a probabilidade de um sinistro ocorrer se igualar ao prêmio por unidade de compensação, ou então quando tivermos um prêmio equivalente ao estimado para as indenizações (Ozaki 2008).

3 Metodologia

3.1 Dados

Os dados para a elaboração do estudo foram coletados do MAPA, em sua seção de Relatórios Estatísticos do Seguro Rural, que contém informações de indenizações do período de 2006 a 2017, ou seja, estamos tratando apenas da parcela do mercado de seguro agrícola brasileiro que recebe incentivo do Programa de Subvenção ao Prêmio do Seguro Rural (PSR).

Dados abertos envolvendo sinistros são escassos e difícil de se obter sem alguma parceria com seguradoras, portanto as informações do MAPA são a principal fonte do presente estudo. Foram coletadas 70.358 observações de 8 variáveis envolvendo contratos de seguro, a saber: ano do contrato, a cultura segurada, a seguradora (por nome), se ocorreu sinistro naquele contrato e o tipo, a importância segurada, o volume de prêmios, o município e o estado.

Ao todo foram usadas informações de 57 culturas, 12 seguradoras, 23 estados e 1.416 municípios. Além dessas informações, foram criadas mais outras 6 variáveis binárias para os fenômenos El Niño e La Niña, construindo 3 variáveis para cada uma de forma a separar os eventos fracos, moderados e fortes por meio dos dados da Golden Gate Weather Services (Null 2015), que estabelecem a ocorrência dos eventos e suas intensidades para cada ano.

A variável de interesse no estudo foi a ocorrência de sinistro, inicialmente representada em uma variável qualitativa, que informava se o contrato havia registrado sinistro e o evento gerador (geada, chuva etc.). Dessa forma, a variável foi transformada em binária: quando não houve sinistro, foi dado valor 0, e valor 1 no caso contrário. Por sua vez, as variáveis qualitativas (nomes) município, estado e seguradora foram transformadas em 3 variáveis numéricas, representadas por um vetor de números inteiro em que o número de cada vetor representa um estado, um município ou uma seguradora. Em resumo, os dados estão dispostos conforme a Tabela 1:

Avaliando os 70.538 contratos de seguro, cerca de 23% apresentou algum sinistro no período. A Tabela 2 mostra quais foram os eventos que mais geraram sinistros nas mais de 20 mil situações avaliadas no estudo, com destaque para o elevado número de ocorrências por conta da seca e a média de indenizações paga por ocorrência, superando o granizo, que embora tenha uma maior frequência, pagou volume menor de indenizações.

Há uma grande concentração na distribuição das causas de sinistros por alguns motivos. Dentre eles, devemos destacar a própria concentração de mercado. Conforme mostra Brasil (2018), a região Sul representou em média 70% do total de produtores rurais beneficiados pelo programa de subsídio gover-

Tabela 1: Relação das variáveis utilizadas

Variável	Tipo
Sinistro	Binária (0 e 1)
Cultura	Numérica (um número para cada cultura)
Seguradora	Numérica (um número para cada seguradora)
Município	Numérica (um número para cada município)
Importância segurada	Em R$
Prêmio	Em R$
El Niño Forte	Binária (0 e 1)
El Niño Fraco	Binária (0 e 1)
El Niño Moderado	Binária (0 e 1)
La Niña Forte	Binária (0 e 1)
La Niña Fraco	Binária (0 e 1)
La Niña Moderado	Binária (0 e 1)
Estado	Numérica (cada número um estado)
Ano	Cada ano

Fonte: Elaboração própria.

Tabela 2: Maiores causas de sinistro do seguro agrícola brasileiro entre 2006 e 2017

Evento de sinistro	Número de ocorrências	Indenizações pagas (B)	B/A
Seca	6.567	R$1.577.132.742	R$240.160,31
Granizo	7.570	R$1.062.030.350	R$140.294,63
Geada	3.939	R$473.453.860	R$120.196,46
Chuva excessiva	2.907	R$338.116.095	R$116.311,01
Inundação/Tromba d'água	359	R$38.830.705	R$108.163,52
Variação excessiva de temperatura	69	R$5.535.640	R$80.226,67
Incêndio	149	R$8.580.452	R$57.586,93
Ventos fortes/frio	631	R$25.123.579	R$39.815,50
Queda de parreiral	76	R$857.152	R$11.278,31
Demais causas	502	R$4.351.146	R$8.667,62

Elaboração própria com dados do MAPA.

namental, além de 50% do total de área coberta e 58,3% do total de recursos disponibilizados para subsídio.

Dentre as culturas, temos que a soja (42,5%), milho (22,3%), maçã (9,0%) e trigo (7,5%) são as mais representativas dentre aquelas seguradas. A concentração das operações nessas regiões explica em parte a maior exposição ao risco de intempéries climática Brasil (2018).

Avaliando os dados em que o estudo trabalhou, a distribuição vai em linha com a concentração de mercado, visto que a maior parte dos municípios que mostraram eventos de sinistros são da região Sul do país, com destaque para Rio Grande do Sul e Paraná, que sofrem com esses tipos específicos de problemas, sobretudo em culturas como soja, trigo, milho, uva e maçã, que representa mais de 80% do total de eventos de sinistros no período avaliado.

3.2 *Random Forest* (RF)

O algoritmo de aprendizagem de máquinas *Random Forest* (RF) foi desenvolvido por Breiman (2001), tratando de problemas de classificação e regressão por métodos de aprendizagem em árvore a partir do bootstrap dos dados de treinamento, ampliando a aplicação do algoritmo conhecido como CART (*Classification e Regression Trees*) proposto por Breiman et al. (1984). No geral, os métodos em árvore de decisão apresentam baixo viés, mas elevada variância, o que acaba produzindo um sobreajuste (*overfitting*)[2]. A utilização do RF trata o problema da variância ao trabalhar com um número n de árvores de forma que ao produzir n observações independentes $\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_n$, uma para cada nó final das árvores, é possível reduzir a variância por meio da média, dado que a variância individual de cada árvore é maior do que a variância da média delas (ou $\sigma_n^2 > \sigma^2/_n$).

3.3 *Support Vector Machine* (SVM)

O chamado *Support Vector Machine* é um método que pode ser usado para predição, buscando em um dado espaço de entrada encontrar um vetor de observações de um conjunto de treinamento, permitindo a construção de um hiperplano que possibilite a separação do conjunto de pontos nas categorias que se deseja prever (duas, no caso do atual estudo). O método foi sugerido inicialmente por Vapnik (2006) e depois Cortes & Vapnik (1995) fizeram modificações para um método mais geral que abrange o caso em que não há formas de separação linear.

3.4 *k-Nearest Neighbours*

O último algoritmo que será utilizado, e que não consta em Pijl (2017), é o que usa informações dos dados de k-vizinhos feita por um classificador baseado em memória chamado *k-Nearest Neighbours* (k-NN), sendo parte da família dos algoritmos de *Machine Learning* da chamada aprendizagem preguiçosa (*lazy learning*). O método é considerado não paramétrico, a despeito da necessidade de estabelecer um k a priori para estabelecer o número de vizinhos que o algoritmo vai avaliar. Conforme Cunningham & Delany (2007) mostram, a classificação por k-NN pode ser dividida em duas partes, sendo a primeira aquela que determina os chamados vizinhos mais próximos e a segunda é a determinação da classe usando tais vizinhos. Torgo (2016) ressalta que o método é fortemente dependente da noção de similaridade entre as observações, estabelecida com a ajuda de uma métrica que avalia distâncias no espaço de entrada, sendo a distância euclidiana a mais usual de todas.

3.5 Algoritmo SMOTE

O algoritmo SMOTE (*Synthetic Minority Oversampling Technique*) executa uma abordagem de sobreamostragem para reequilibrar o conjunto de treinamento original. Em vez de aplicar uma replicação simples das instâncias de classe minoritária, a ideia principal do SMOTE é apresentar exemplos sintéticos. Esses novos dados são criados por interpolação entre várias instâncias de classe

[2] O processo ocorre quando o método apresenta uma grande performance no conjunto de dados no qual foi treinado, mas o desempenho piora significativamente quando se passa para o conjunto de dados para teste.

minoritária que estão dentro de uma vizinhança definida. Por esse motivo, diz-se que o procedimento está focado no "espaço de *features*" e não no "espaço de dados", ou seja, o algoritmo é baseado nos valores das *features* e sua relação, em vez de considerar os pontos de dados como um todo (Chawla et al. 2002).

3.6 Validação Cruzada

O processo de seleção dos hiperparâmetros de cada algoritmo será feito a partir de um método conhecido como validação cruzada (*cross-validation*), em que se separa de forma aleatória o conjunto de treinamento em vários outros subconjuntos menores, com o objetivo de atingir a configuração de algoritmo que apresente à maior acurácia durante as predições desses subconjuntos. O método usado no presente trabalho, descrito em Hastie et al. (2009), é conhecido como *k-fold cross-validation*, que consiste em dividir o conjunto de dados em k partes de igual tamanho, de tal forma que o algoritmo separa a $k-ésima$ parte e estabelece o treinamento nas k-1 partes restantes. Após treino, o conjunto separado é utilizado para a validação, estimando o erro de predição. Cada $k-ésima$ parte será usada como conjunto para a validação com o objetivo de chegar em uma configuração de modelo que consiga o máximo possível de generalização na predição da base de dados. Não há uma regra estabelecida para a seleção de k, mas Hastie et al. (2009) ressalta que o k é escolhido de forma que cada subconjunto seja grande o suficiente para ser estatisticamente representativo do conjunto de dados e que $k = 10$ é a escolha mais usual dentre os pesquisadores, também sendo escolhido para esse trabalho.

3.7 Medidas para avaliar o desempenho

A avaliação de performance não se limita apenas ao cálculo da acurácia, que pode mascarar um desempenho aquém do esperado ao prever corretamente apenas uma das classes da variável resposta, enquanto apresenta uma baixa capacidade preditiva na categoria complementar. Dessa forma, outras métricas foram utilizadas para analisar o desempenho dos algoritmos: Matriz de Confusão, Sensibilidade, Sensitividade, Precisão e Coeficiente de Correlação de Matthews (MCC).

A partir da base inicial, que compreende os anos de 2006 a 2017, foi separado o ano de 2017 para a avaliação da performance dos modelos ajustados, compreendendo 7764 observações com um percentual de 81,5%/18,5% para as classes 0 e 1, respectivamente.

Matriz de Confusão

Para avaliar os problemas envolvendo variáveis binárias as métricas mais comuns de análise de performance são derivadas da chamada matriz de confusão, que pode ser ilustrada conforme a Figura 1:

Dessa forma, em uma dada amostra de tamanho N com 2 categorias, sendo p observações de 0 e $1-p$ observações de 1, o melhor desempenho é atingindo quando $VN = p$ e $VP = 1-p$, ou seja, a soma da diagonal principal da matriz é N. O desempenho pode ser observado no que é chamado de medida de acurácia:

Figura 1: Matriz de confusão

		Efetivo	
		0	**1**
Predito	**0**	Verdadeiro Negativo (VN)	Falso Negativo (VN)
	1	Falso Positivo (FP)	Verdadeiro Positivo (VP)

Fonte: Elaboração do autor com base em Abd Elrahman & Abraham (2013).

$$Acurácia: \frac{VN+VP}{VN+FN+FP+VP} \tag{1}$$

em que quanto mais próximo de 1, mais próximo $VN+VP$ estará de N, representando 100% de acerto na predição, ao passo que 0 indicaria 0% de acerto. É difícil atingir 100% de acerto, portanto, é necessário olhar o desempenho individual de cada categoria. Além de olhar o valor absoluto em cada entrada da matriz, é possível calcular as Taxas de Verdadeiro Positivo (TVP) e Taxas de Verdadeiro Negativo (TVN), conforme as expressões 2 e 3.

$$TVP = \frac{VP}{VP+FN} \tag{2}$$

$$TVN = \frac{VN}{VN+FP} \tag{3}$$

Limiar, Curva ROC e AUC

O vetor predito, fruto das regressões RF e SVM, é composto por valores contínuos entre 0 e 1, ao passo que a variável original que se pretende prever é composta pelos valores discretos 0 ou 1. Assim, é preciso alterar os dados preditos de forma a torná-los binários, como mostra Fawcett (2006), caso contrário não será possível estabelecer a Matriz de Confusão ou outro critério de performance relevante para o método. A transformação é feita de forma que:

$$\hat{y}_n^* = \begin{cases} 0, & \hat{y}_n < \tau \\ 1, & \hat{y}_n \geq \tau \end{cases} \tag{4}$$

em que $\hat{y}_n$ são as observações preditas pelos modelos, $\hat{y}_n^*$ são as observações transformadas na variável binária, 0 ou 1, dependendo de τ, conforme Pijl (2017). O desafio agora é encontrar tal limiar τ (conhecido como *threshold*[3]) que guiará essa divisão, sendo ele definido a partir da chamada curva ROC. A curva ROC (Receiver Operating Characteristics) é uma técnica de visualização gráfica utilizada para aumentar o acerto e reduzir o volume de "alarmes falsos" (Fawcett 2006). Com o vetor de $\tau \in [0,1]$ é possível calcular diversas matrizes de confusão, extraindo o TVP, conhecido na métrica da ROC como Sensitividade, e o TVN, chamado de Especificidade. De fato, o τ que se deseja buscar é aquele que resulta na maior Sensitividade e Especificidade. Além disso, a elaboração da curva ROC permite calcular outra medida, a área embaixo da ROC, ou AUC (*Area under the ROC Curve*). A avaliação da AUC se dá entre 0,5 e 1 e indica quanto o modelo é capaz de distinguir entre classes, ou seja, quanto maior a AUC, maior o acerto do modelo (Fawcett 2006).

[3]Uma revisão teórica e discussão da literatura envolvendo a determinação da *threshold* para avaliação da predição está disponível em Liu et al. (2005) e Pei et al. (2013).

Conforme Elrahman e Abraham (2013) mostram, a área pode ser calculada por:

$$AUC = \frac{TVP + TVN}{2}. \tag{5}$$

Precisão e coeficiente de correlação de Matthews

É recomendado avaliar ainda o desempenho em duas métricas: a precisão e o coeficiente de correlação de Matthews (MCC). Como mostram Abd Elrahman & Abraham (2013) e Boughorbel et al. (2017):

$$\text{Precisão Categoria Negativa} = \frac{VN}{VN + FN} \tag{6}$$

$$\text{Precisão Categoria Positiva} = \frac{VP}{VP + FP} \tag{7}$$

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \tag{8}$$

A precisão foca na proporção de acertos na categoria positiva (ou 1, no caso do exemplo binário) em relação aos totais de acertos. Ou seja, além disso, também há a estimativa para a precisão da categoria negativa (0, no caso). No caso em que desejamos estudar, ter uma elevada precisão na categoria positiva significa acertar a ocorrência de sinistro, por isso a importância de ter uma elevada precisão. Já o coeficiente de correlação de Matthews é outra medida de performance, levando em conta as taxas de acerto tanto da categoria positiva, quanto da negativa, mostrando uma espécie de correlação entre o valor real e o predito.

Os métodos foram aplicados utilizando-se o software R (Team et al. 2013), sendo que foram utilizados os pacotes *DMwR* (Torgo & Torgo 2013) para a aplicação do método SMOTE e os pacotes *caret* (Kuhn et al. 2008), *ranger* (Wright & Ziegler 2015) e *kernlab* (Karatzoglou et al. 2004) para os demais métodos.

4 Resultados e Discussão

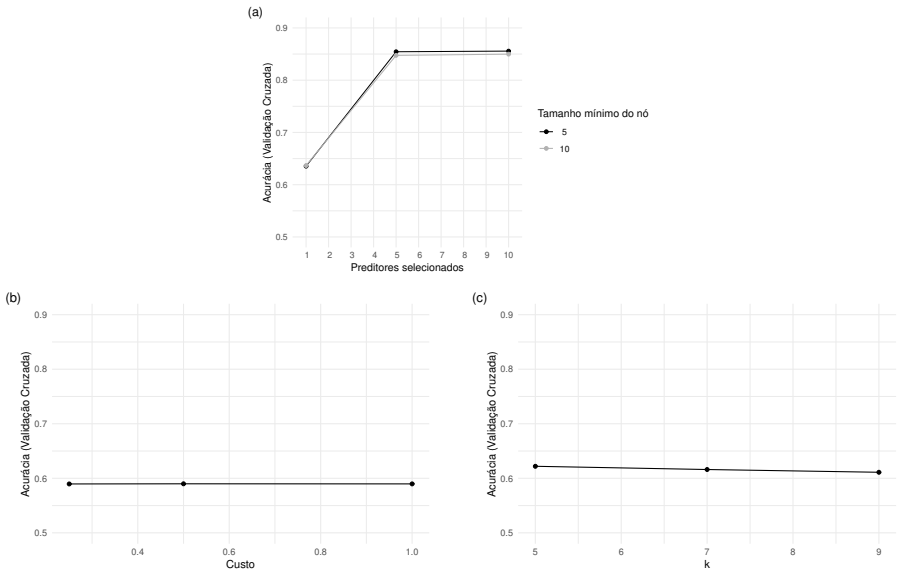
Esta seção mostra os resultados separadamente para cada método, com exceção das medidas de acurácia de predição do conjunto de teste, que são apresentadas conjuntamente ao final de modo a facilitar a comparação entre os métodos.

O rebalanceamento da base de treino pelo algoritmo SMOTE resultou em uma base de 104.356 observações das 13 variáveis preditivas, sendo que a variável binária que reporta a ocorrência de sinistro, que antes era composta por uma relação 76,3%/23,7% de 0 e 1, agora está balanceado em 57,1%/42,9%. Por sua vez, o conjunto de teste representa cerca de 7.764 observações de tais variáveis, mas segue desbalanceado de forma a representar o cenário real com uma relação 81,5%/18,5%.

A escolha dos parâmetros foi feita por meio do processo de validação cruzada, com 10 *folds*, sendo apresentados os resultados na Figura 2.

O algoritmo *Random Forest* foi treinado com um número fixo de 500 árvores e utilizando o índice de Gini como regra de divisão. Os parâmetros "tamanho mínimo do nó" e "número de preditores" foram definidos por meio

Figura 2: Erro Quadrático Médio (EQM) de acordo com os parâmetros escolhidos, por método. (a) Random Forest; (b) Support Vector Machine; (c) k-Nearest Neighbours



Fonte: Elaboração própria.

da validação cruzada, chegando-se aos valores de 5 e 10, respectivamente. Entretanto, como se observa na figura 2a, a acurácia se mostrou similar em comportamento e valor para as diversas combinações de parâmetros testadas.

Por sua vez, para a aplicação do *Support Vector Machine*, utilizou-se a função de base radial[4]. O parâmetro sigma foi definido como $3,49e^{-10}$ pelo método heurístico descrito em Caputo et al (2002) e o parâmetro custo foi definido por meio da validação cruzada, chegando-se a um valor de 0,5 (Figura 2b).

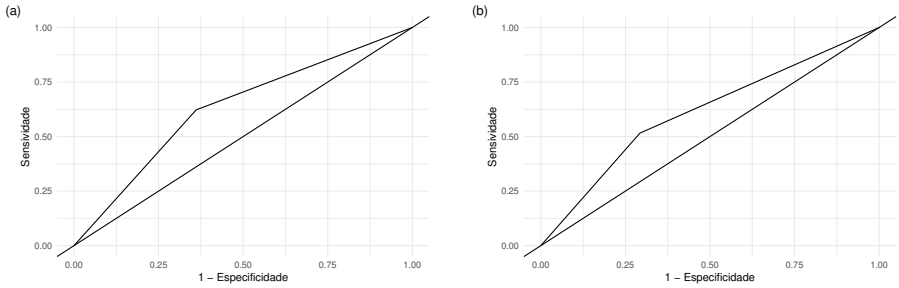
Para o k-NN, optou-se por testar a acurácia com um k variando de 5 a 9 vizinhos, conforme a Figura 1c. O EQM é reduzido ao selecionar $k = 5$, embora com uma performance menor de 5% superior em relação à um k igual a 9, que obteve o pior desempenho (Figura 2c).

Após selecionar as melhores configurações de cada método foram realizadas as predições para o conjunto de teste, que envolvem os dados de 2017. Para transformar os vetores preditos das ocorrências de sinistros em variáveis binárias, no caso do RF e SVM, a estimativa do *threshold* se deu pela curva ROC, buscando a melhor relação entre a sensitividade e a especificidade. A Figura 3 mostra como foi o resultado considerando os dois modelos para um conjunto de *threshold*.

O algoritmo RF mostrou um desempenho preditivo superior ao SVM, sobretudo no ponto escolhido como τ em ambos os casos, podendo destacar que ambos estão acima da linha de 45º, que sinaliza o limite mínimo da curva AUC (0,5) (Figura 4). A superioridade é confirmada ao se calcular a AUC, que é de

[4]Conforme mostra Hastie et al. (2009), a função radial pode ser definida como $f(x) = \sum_{j=1}^{M} K_{\lambda_j}(\xi_j, x)\beta_j = \sum_{j=1}^{M} D\left(\frac{\|x-\xi_j\|}{\lambda_j}\right)\beta_j$, em que ξ_j é um parâmetro de localização, λ_j é parâmetro de escala e D pode ser uma função gaussiana.

Figura 3: Curvas ROC para os modelos Random Forest e SVM



Fonte: Elaboração própria.

0,63 no caso da RF e de 0,61 no caso da SVM. No caso do método k-NN escolhido, a resposta já é binária e, portanto, não necessitou de uma configuração de um *threshold*, podendo estabelecer uma avaliação do método logo após a classificação do conjunto de teste.

É possível ver nas matrizes de confusão da Figura 4 a performance preditiva de cada método, ressaltando qual mostrou melhor desempenho tanto para o não sinistro (0) quanto para a ocorrência de sinistro (1). De forma geral, os algoritmos RF e SVM apresentam uma melhor capacidade preditiva, principalmente por apresentar um número menor em sua diagonal secundária. No entanto, nenhum dos algoritmos apresentou desempenho adequado na previsão da ocorrência de sinistro, ou seja, o verdadeiro positivo (prever 1 quando de fato era 1). A importância desse resultado consiste em errar justamente em situações em que de fato o sinistro aconteceu, não conseguindo antecipar o seu evento.

Figura 4: Matriz de confusão por método, RF, SVM e k-NN

RF

		Efetivo	
		0	**1**
Predito	**0**	4041	544
	1	2283	896

SVM

		Efetivo	
		0	**1**
Predito	**0**	4471	697
	1	1853	743

Fonte: Elaboração própria.

O custo financeiro em prever que não aconteceria um sinistro em um contrato que sofreu sinistro (prever 0 quando era 1) é maior do que no caso em que se prevê sinistro para um contrato que de fato não ocorreu sinistro (prever 1 quando era 0), visto que no primeiro caso acontecerá uma saída de caixa (indenização) não esperada e no segundo caso haverá apenas uma alocação de recursos para uma reserva técnica, portanto, sem perdas para a seguradora. A seguradora pode não estar com liquidez para a primeira situação, incorrendo em custos de empréstimos e penalizações pelo órgão regulador, ao passo que no segundo caso ela terá apenas separado o dinheiro, podendo ser investido nos ativos descritos em Central (2015).

Após a projeção de cada método, cabe a comparação das principais medidas de desempenho elencadas na seção anterior, para então ter a certeza do melhor modelo. A Tabela 3 mostra desempenho o preditivo similar do *Random Forest* e SVM quando se analisa as demais de comparação, isto é, Acu-

rácia, TVP, TVN, Precisão e MCC. Observa-se a proximidade do desempenho dos métodos RF e SVM, sendo o primeiro superior na previsão de não sinistros e o último levemente superior na previsão dos sinistros. Vale ressaltar que mesmo treinando o algoritmo com uma base balanceada, o desempenho foi pior em prever 1 do que a categoria 0 em todos os métodos testados, ressaltando a dificuldade em se antever eventos de sinistro.

Tabela 3: Principais medidas de avaliação de predição por método.

	Acurácia	TVN	TVP	Precisão (0)	Precisão (1)	MCC
RF	0,6359	0,6390	**0,6222**	**0,8813**	0,2818	**0,2065**
SVM	**0,6716**	**0,7070**	0,5160	0,8651	**0,2862**	0,1837
k-NN	0,6073	0,6395	0,4687	0,8409	0,2284	0,0866

Elaboração própria.

Abrindo os resultados, é possível observar a diferença por estado na Tabela 4, também revelando a alternância de melhor capacidade preditiva entre os métodos RF e SVM. Considerando apenas aqueles estados com mais de 500 contratos para o ano de 2017 na base de dados estudada, os algoritmos RF e SVM mostraram um índice de acurácia médio de 0,65 (ou 65% de acerto) superior a acurácia do k-NN (0,60) para a média simples dos mesmos contratos.

Tabela 4: Acurácia da predição dos sinistros ocorrido em 2017, por estado e por método

Estado	Número de observações	RF	SVM	k-NN
AL	2	**1,00**	**1,00**	**1,00**
BA	45	0,69	**0,80**	0,71
DF	8	**0,62**	**0,62**	0,37
ES	9	**0,67**	0,44	0,44
GO	537	0,62	**0,77**	0,66
MA	2	0,00	0,00	0,00
MG	557	0,56	**0,65**	0,60
MS	325	0,60	**0,81**	0,72
MT	209	0,64	**0,77**	0,70
PE	1	**1,00**	0,00	0,00
PR	3852	**0,66**	**0,66**	0,59
RO	1	0,00	0,00	0,00
RS	723	**0,61**	**0,61**	0,55
SC	270	0,57	0,56	**0,63**
SE	2	**1,00**	0,50	0,50
SP	1220	0,62	**0,67**	0,62
TO	1	**1,00**	**1,00**	0,00

Fonte: Elaboração própria.

Destacando apenas a região Sul, que apresenta 4.845 das 7.764 observações desse conjunto de teste, a média de acurácia foi de 0,61 para o RF, de 0,61 para o SVM e de 0,59 para o k-NN. Na região Sudeste (correspondendo a 23% da base de teste), o indicador de acurácia médio ficou em 0,62, 0,59 e 0,55, respectivamente. Naqueles estados em que há somente uma observação de sinistro para o período abordado há uma dificuldade adicional para a predição, visto que no próprio conjunto de treinamento o número de observações para

esses estados também era reduzido, destacando os estados da região Nordeste e Nordeste do país.

Considerando-se as principais culturas em termos de apólices e sinistros, sejam elas a soja, milho 2ª safra e trigo, representando 73,3% do número de sinistros para o ano de 2017, temos que o algoritmo SVM apresentou desempenho médio superior aos demais. Enquanto o algoritmo SVM apresentou Acurácia de 0,75, 0,68 e 0,54 para as três culturas, respectivamente, o algoritmo RF registrou previsão inferior, com Acurácia de 0,68, 0,52 e 0,52, sendo que o mesmo aconteceu no caso do k-NN, com 0,65, 0,59 e 0,54 (Tabela 5).

Por fim, cabe destacar o desempenho preditivo por seguradora. Em 2017, 10 das 12 seguradoras que aturam no mercado de seguro agrícola naquele ano apresentaram algum tipo de sinistro. O algoritmo SVM mostrou acurácia predominantemente superior aos demais, seguido pelo RF e o k-NN. Destaca-se, no entanto, a utilidade relativa dessas metodologias na predição dos eventos em que ocorreu sinistros em si, dado que conforme já discutido, a precisão para a classificação de eventos de sinistro foi baixa para todos os algoritmos (Tabela 6).

5 Conclusões

Este estudo teve o objetivo avançar na predição de ocorrências de sinistros no mercado de seguro agrícola, sendo essa uma variável de relativa importância para a estabilidade das operações das seguradoras, em geral.

Em ramos tradicionais do mercado segurador, por exemplo, ramo de automóveis, tais questionamentos já estão relativamente pacificados. Ao longo do tempo, diferentes abordagens foram testadas utilizando amostras suficientemente grandes para garantir a precisão dos resultados. Em ramos mais recentes, por exemplo, o ramo rural, que possui idiossincrasias que o distanciam das abordagens tradicionais, diversas questões ainda precisam ser respondidas.

Uma das principais e foco deste estudo, refere-se à ausência de metodologias capazes de antecipar e predizer com relativa precisão os sinistros agrícolas. Para as companhias seguradoras isso implica em alocar um volume incerto de provisão. Se superestimar o volume necessário, a seguradora pode ter sua rentabilidade afetada, tendo que sustentar custos desnecessários. Caso contrário, há risco de incapacidade de pagamento das indenizações aos segurados com severas penalizações pela SUSEP ou até mesmo risco de insolvência.

Cabe destacar que a maior parte da alocação dessas provisões segue as metodologias propostas pela SUSEP, mas ainda há provisões que carecem de metodologias e que as próprias seguradoras precisam desenvolver e encaminhar como notas técnicas para o órgão fiscalizador (Brasil 2017).

Nesse contexto, aplicam-se neste estudo os métodos de aprendizagem de máquinas *Random Forest* (RF), *Support Vector Machine* (SVM) e *k-Nearest Neighbors* (k-NN) foram escolhidos para elaborar a predição da ocorrência ou não do sinistro agrícola na parte do mercado que participa do Programa de Subvenção ao Prêmio do Seguro Rural (PSR) do ano de 2017, após o treino do algoritmo com os dados do período entre 2006 e 2016.

Os resultados mostraram que os métodos Random Forest e SVM registraram melhor desempenho, com o primeiro exibindo uma performance levemente superior na predição do não sinistro. No entanto, ambos apresentaram

Tabela 5: Acurácia da predição dos sinistros ocorrido em 2017, por cultura e por método

Cultura	Observações	RF	SVM	k-NN
Abacate	1	0,00	0,00	**1,00**
Abóbora	3	0,33	0,33	**0,67**
Abobrinha	1	**1,00**	**1,00**	**1,00**
Alface	1	**1,00**	0,00	**1,00**
Algodão	15	0,60	**1,00**	0,80
Alho	33	0,58	**0,79**	0,67
Ameixa	77	0,57	**0,61**	0,53
Arroz	54	0,61	**0,76**	0,70
Atemoia	2	**1,00**	0,50	**1,00**
Aveia	9	**1,00**	0,56	0,56
Banana	4	**0,50**	0,25	0,00
Batata	20	**0,55**	0,50	0,45
Berinjela	3	**0,67**	0,00	0,33
Beterraba	1	0,00	0,00	0,00
Café	218	**0,82**	0,68	0,63
Cana-de-açúcar	155	**0,96**	0,74	0,59
Canola	7	0,57	0,57	**0,71**
Caqui	47	0,53	**0,64**	0,51
Cebola	55	0,49	**0,55**	0,53
Cenoura	7	**0,71**	**0,71**	0,43
Cevada	32	**0,66**	0,50	0,41
Chuchu	1	0,00	**1,00**	**1,00**
Couve-flor	1	**1,00**	**1,00**	**1,00**
Feijão	186	**0,66**	0,57	0,59
Goiaba	10	0,40	**0,60**	0,30
Kiwi	8	**0,63**	0,38	0,50
Laranja	7	0,57	0,29	**0,71**
Maçã	122	0,47	0,61	**0,64**
Mandioca	8	**1,00**	0,50	0,63
Melancia	12	**0,58**	**0,58**	0,50
Melão	1	0,00	0,00	0,00
Milho 1ª safra	392	**0,73**	0,53	0,50
Milho 2ª safra	1577	0,52	**0,68**	0,59
Nectarina	19	**0,79**	0,53	0,53
Pepino	1	**1,00**	**1,00**	0,00
Pêra	22	0,32	0,45	**0,64**
Pêssego	67	0,58	0,60	**0,61**
Pimentão	18	**0,72**	0,61	**0,72**
Repolho	1	**1,00**	0,00	0,00
Soja	3296	0,68	**0,75**	0,65
Sorgo	19	**0,74**	0,47	0,47
Tangerina	18	0,28	0,72	**0,83**
Tomate	116	0,64	**0,71**	0,62
Trigo	818	**0,55**	0,54	0,54
Triticale	4	0,25	0,00	**0,50**
Uva	295	**0,64**	0,56	0,59

Fonte: Elaboração própria.

Tabela 6: Acurácia da predição dos sinistros ocorrido em 2017, por seguradora e por método

Seguradora	Número de observações	RF	SVM	k-NN
Aliança do Brasil	1600	**0,70**	**0,70**	0,61
Allianz	886	0,67	**0,71**	0,61
Essor	1146	0,56	**0,58**	0,56
Excelsior	191	**0,65**	0,61	0,56
Fairfax	493	0,64	**0,71**	0,61
Mapfre	1073	0,59	**0,65**	0,62
Porto Seguro	209	0,62	**0,63**	**0,63**
Sancor	959	0,62	**0,66**	0,59
Swiss Re	1068	0,66	**0,73**	0,65
Tokio Marine	139	**0,55**	**0,55**	0,54

Fonte: Elaboração própria.

uma performance insuficiente na predição do sinistro. Tais resultados revelam achados importantes no processo investigatório. As metodologias empregadas neste estudo e que foram aplicadas em outros ramos, com resultados comparativamente melhores, mostraram relativa incapacidade em prever os eventos de sinistro.

Em futuros estudos, novas abordagens poderão ser utilizadas, bem como novas variáveis que influenciam os sinistros. Como alternativa, poder-se-ia utilizar no lugar de uma variável binária para a ocorrência de sinistros, uma variável numérica indicando valores monetários para cada tipo de sinistro, e não apenas sua ocorrência ou não.

Ademais, a impossibilidade de acessar dados históricos de produtividade em nível de propriedade rural pode ter prejudicado a capacidade de predição dos modelos. O problema nesse caso refere-se à granularidade dos dados. Por exemplo, no caso do ramo de automóveis, os modelos utilizam dados individualizados, por segurado, e não dados municipais como aqueles utilizados neste estudo.

O fato é que captar a variabilidade produtiva, em nível de propriedade rural, utilizando dados em escala municipal acaba por criar dificuldades adicionais no processo de modelagem. Em outras palavras, a atenuação da variabilidade produtiva causada pela agregação dos dados pode ter afetado a performance preditiva dos modelos, principalmente no caso dos eventos de sinistro. Dessa forma, entende-se que a principal limitação do estudo se baseia na estrutura dos dados. Um maior detalhamento das características do produtor e do local de cultivo, por exemplo, solo e nível tecnológico, poderia aumentar a capacidade preditiva dos métodos utilizados.

Os métodos *Random Forest* e SVM registram o melhor desempenho, com o primeiro exibindo uma performance levemente superior na predição do não sinistro, ambos apresentando, no entanto, uma performance insuficiente na predição do sinistro. Este resultado indica a incapacidade das variáveis utilizadas em prever os eventos de sinistro. Como aprimoração deste estudo, seria possível no lugar de uma variável binária para ocorrência de eventos, criar uma variável numérica dando valor para cada tipo de sinistro, e não apenas sua ocorrência ou não. Ademais, a ausência de dados históricos de produtividade a nível de fazenda também prejudica a capacidade de predição, dado que não se pode mensurar adequadamente a variabilidade de produção a ní-

vel municipal e intra-municipal. Para que isso seja possível, é necessária uma maior variedade e maior precisão de dados nas bases de dados disponíveis ao público. Desta forma, a principal limitação do artigo está concentrada na questão dos dados, pois foi preciso trabalhar apenas com dados de bases abertas, muito mais limitados e com um volume de informações mais particulares a cada contrato de seguro muito inferior ao que as próprias seguradoras possuem. Informações mais precisas das características do produtor e do local de cultivo, como solo e tecnologia empregada, poderiam aumentar a potência preditiva dos métodos utilizados. Outro ponto para futuros estudos seria o estudo dos contratos de seguro agrícola que sofrem com mais de um sinistro durante a sua vigência e como isso poderia ser predito, sendo necessária também uma base de dados com estas informações para cada contrato.

Referências Bibliográficas

Abd Elrahman, S. M. & Abraham, A. (2013), 'A review of class imbalance problem', *Journal of Network and Innovative Computing* **1**(2013), 332–340.

Baumgartner, C., Gruber, L. F. & Czado, C. (2015), 'Bayesian total loss estimation using shared random effects', *Insurance: Mathematics and Economics* **62**, 194–201.

Bekkar, M. & Alitouche, T. A. (2013), 'Imbalanced data learning approaches review', *International Journal of Data Mining & Knowledge Management Process* **3**(4), 15.

Berlato, M. A., Farenzena, H. & Fontana, D. C. (2005), 'Associação entre El Niño, Oscilação Sul e a produtividade do milho no estado do Rio Grande do Sul', *Pesquisa Agropecuária Brasileira* **40**(5), 423–432.

Borde, S. F., Chambliss, K. & Madura, J. (1994), 'Explaining variation in risk across insurance companies', *Journal of Financial Services Research* **8**(3), 177–191.

Boughorbel, S., Jarray, F. & El-Anbari, M. (2017), 'Optimal classifier for imbalanced data using Matthews correlation coefficient metric', *PloS one* **12**(6), e0177678.

Brasil (2017), Provisões técnicas: orientações da SUSEP ao mercado de seguros, previdência complementar aberta, capitalização e resseguro local, Technical report, Superintendencia De Seguros Privados (SUSEP).
URL: *http://www.agricultura.gov.br/assuntos/riscosseguro/seguro-rural/relatorios-estatisticos*

Brasil (2018), Dados de indenizações – 2006 a 2017, Technical report, Ministerio Da Agricultura, Pecuária E Abastecimento (MAPA).
URL: *http://www.agricultura.gov.br/assuntos/riscosseguro/seguro-rural/relatorios-estatisticos*

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), 'Classification and regression trees. Belmont, CA: Wadsworth', *International Group* **432**, 151–166.

Cabrera, V. E., Fraisse, C. W., Letson, D., Podestá, G. & Novak, J. (2006), 'Impact of climate information on reducing farm risk by optimizing crop insurance strategy', *Transactions of the ASABE* **49**(4), 1223–1233.

Castro, C. L. & Braga, A. P. (2011), 'Aprendizado supervisionado com conjuntos de dados desbalanceados', *Sba: Controle & Automação Sociedade Brasileira de Automatica* **22**(5), 441–466.

Central, B. (2015), 'Resolução no 4.444, de 13 de novembro de 2015'.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'Smote: synthetic minority over-sampling technique', *Journal of artificial intelligence research* **16**, 321–357.

Cirino, P. H., Féres, J. G., Braga, M. J. & Reis, E. (2015), 'Assessing the impacts of ENSO-related weather effects on t he Brazilian agriculture', *Procedia Economics and Finance* **24**, 146–155.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.

Cunha, G. R. & Assad, E. D. (2001), 'Uma visão geral do número especial da RBA sobre zoneamento agrícola no Brasil', *Revista Brasileira de Agrometeorologia* **9**(3), 377–385.

Cunha, G. R., Dalmago, G. & Estefanel, V. (1999), 'ENSO influences on wheat crop in Brazil', *Revista Brasileira de Agrometeorologia* **7**(1), 127–138.

Fawcett, T. (2006), 'An introduction to ROC analysis', *Pattern recognition letters* **27**(8), 861–874.

Freitas, M. A. L. (2010), 'Modelo logístico aplicado ao mercado de seguros de auto no Brasil: cálculo da probabilidade de sinistros', *Indicadores Econômicos FEE* **37**(3).

Harnek, R. F. (1966), Formula loss reserves, Technical report, Insurance Accounting and Statistical Association.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. (2004), 'kernlab-an s4 package for kernel methods in R', *Journal of statistical software* **11**(9), 1–20.

Kuhn, M. et al. (2008), 'Building predictive models in R using the caret package', *Journal of statistical software* **28**(5), 1–26.

Liu, C., Berry, P. M., Dawson, T. P. & Pearson, R. G. (2005), 'Selecting thresholds of occurrence in the prediction of species distributions', *Ecography* **28**(3), 385–393.

Liu, J., Men, C., Cabrera, V. E., Uryasev, S. & Fraisse, C. W. (2008), 'Optimizing crop insurance under climate variability', *Journal of Applied Meteorology and Climatology* **47**(10), 2572–2580.

Mack, T. et al. (1994), 'Which stochastic model is underlying the chain ladder method', *Insurance: mathematics and economics* **15**(2-3), 133–138.

Menardi, G. & Torelli, N. (2014), 'Training and assessing classification rules with imbalanced data', *Data Mining and Knowledge Discovery* **28**(1), 92–122.

Null, J. (2015), 'El Niño and La Niña years and intensities', *Golden Gate Weather Services (5 Sep 2013)* .

Oliveira, N. V. (2005), Mercados de Seguros: Solvência, Riscos e Eficácia Regulatória, PhD thesis, Erasmus University Rotterdam.

Ozaki, V. A. (2008), 'Em busca de um novo paradigma para o seguro rural no Brasil', *Revista de Economia e Sociologia Rural* **46**(1), 97–119.

Pei, Y., Kim, T.-K. & Zha, H. (2013), Unsupervised random forest manifold alignment for lipreading, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 129–136.

Pijl, T. (2017), A framework to forecast insurance claims, PhD thesis, Erasmus University Rotterdam.

Rodrigues, A. & Martins, E. (2009), 'Gerenciamento da informação contábil através das provisões técnicas constituídas por sociedades seguradoras', *Revista Universo Contábil* **6**(1), 46–66.

Sousa, K. M. M. (2010), Modelos lineares generalizados e modelos de dispersão aplicados à modelagem de sinistros agrícolas, PhD thesis, Universidade de São Paulo.

Steinmetz, S. & Silva, S. (2017), 'Início dos estudos sobre zoneamento agrí-cola de risco climático (ZARC) no Brasil', *Santo Antônio de Goiás: Embrapa Arroz e Feijão* .

Team, R. C. et al. (2013), 'R: A language and environment for statistical computing'.

Torgo, L. (2016), *Data mining with R: learning with case studies*, CRC press.

Torgo, L. & Torgo, M. L. (2013), 'Package 'dmwr'', *Comprehensive R Archive Network* .

Vapnik, V. (2006), *Estimation of dependences based on empirical data*, Springer Science & Business Media.

Weiss, G. M., McCarthy, K. & Zabar, B. (2007), 'Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?', *Dmin* **7**(35-41), 24.

Wright, M. N. & Ziegler, A. (2015), 'Ranger: a fast implementation of random forests for high dimensional data in C++ and R', *Journal of Statistical Software*.

Yang, Y., Qian, W. & Zou, H. (2018), 'Insurance premium prediction via gradient tree-boosted tweedie compound Poisson models', *Journal of Business & Economic Statistics* **36**(3), 456–470.

Ye, C., Zhang, L., Han, M., Yu, Y., Zhao, B. & Yang, Y. (2018), 'Combining predictions of auto insurance claims', *ArXiv* .

Zaniboni, N. & Montini, A. (2015), 'Modelos de Poisson inflada de zeros e binomial negativa inflada de zeros na previsão de sinistro de automóveis', *E&G Economia e Gestão* **15**(41).