Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Previsão de prêmio e a ocorrência de sinistros no mercado de seguro agrícola brasileiro

Arthur Augusto Lula Mota

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Economia Aplicada

Piracicaba
2019

Arthur Augusto Lula Mota
Bacharel em Ciências Econômicas

Previsão de prêmio e a ocorrência de sinistros no mercado de seguro agrícola brasileiro

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **VITOR AUGUSTO OZAKI**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Economia Aplicada

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Mota, Arthur Augusto Lula

Previsão de prêmio e a ocorrência de sinistros no mercado de seguro agrícola brasileiro / Arthur Augusto Lula Mota. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2019.

90 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Seguro agrícola 2. Sinistro 3. Prêmio 4. Previsão 5. Séries de tempo I. Título

AGRADECIMENTOS

Ao Prof. Dr. Vitor Augusto Ozaki, professor e orientador, pela sua experiência e guia durante o processo desse trabalho. Esta foi uma das raras oportunidades de vida em que se pode trabalhar com alguém que é referência em sua área.

Aos meus pais, Magda e Silas, e meu irmão Pedro por todo o apoio desde a decisão de fazer o mestrado até a fase final. Assim como em vários outros momentos de minha vida, vocês foram fundamentais em mais essa etapa.

A minha namorada Carol, pela compreensão e apoio em todos esses meses em que ficamos distante, oferecendo amor e parceria que espero poder retribuir por muitos anos. Agradeço também a Pandora e Sofia pela companhia diária durante essa jornada.

Obrigado aos colegas de mestrado, em especial Fábio, Guilherme, Augusto, Pedro, Yuri e Lucas, que me ajudaram a superar as dificuldades, tanto durante os estudos quanto no dia a dia de Piracicaba.

Aos meus vários colegas de profissão que incentivaram a fazer o mestrado, decisão que terá impacto pelo resto da minha vida.

À CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, e ao CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico, que me concederam bolsa durante a realização deste mestrado.

Por fim, deixo minha gratidão para todos aqueles que participaram direta ou indiretamente para a concretização deste trabalho.

EPÍGRAFE

“O imprevisto é uma espécie de deus avulso, ao qual se deve reservar algumas ações que podem ter voto decisivo na assembleia dos acontecimentos.”.

Machado de Assis

SUMÁRIO

RESUMO 6
ABSTRACT 7
LISTA DE FIGURAS 8
LISTA DE TABELAS 9
1. INTRODUÇÃO 11
REFERÊNCIAS 14
2. APLICAÇÃO DE MÉTODOS DE SÉRIES DE TEMPO E APRENDIZAGEM DE MÁQUINA PARA PROJEÇÃO DE PRÊMIO DO SEGURO AGRÍCOLA NAS REGIÕES DO BRASIL 17
RESUMO 17
ABSTRACT 17
2.1. INTRODUÇÃO 18
2.2. REFERENCIAL TEÓRICO 19
2.3. MÉTODOS 26
2.4. RESULTADOS E DISCUSSÃO 35
2.5. CONCLUSÃO 42
REFERÊNCIAS 43
3. PREDIÇÃO DE SINISTROS AGRÍCOLAS: UMA ABORDAGEM COMPARATIVA UTILIZANDO INTELIGÊNCIA ARTIFICIAL 49
RESUMO 49
ABSTRACT 49
3.1. INTRODUÇÃO 49
3.2. REVISÃO DE LITERATURA 51
3.3. MÉTODOS 55
3.4. RESULTADOS E DISCUSSÃO 66
3.5. CONCLUSÕES 76
REFERÊNCIAS 76
4. CONCLUSÃO 81
APÊNDICES 83

RESUMO

Previsão de prêmio e sinistro do mercado de seguro agrícola brasileiro

Um dos maiores desafios do produtor agrícola é lidar com uma diversidade de riscos, sendo eles exógenos ou endógenos a sua atividade. O que estão em alcance do produtor são tratados com tecnologia, mas há ainda aqueles riscos exógenos, sobretudo ligado a acontecimentos climáticos. O seguro agrícola serve para abordar tais riscos, por meio da transferência dos custos para um agente econômico, a seguradora. O prêmio e a ocorrência de sinistros são variáveis de máxima importância para os participantes desse mercado, sendo de interesse das seguradoras a sua previsão para a possibilidade de um melhor planejamento financeiro. O presente trabalho teve por objetivo avançar nessas duas frentes: na primeira frente, o prêmio do seguro agrícola regional foi alvo de modelagem para sua predição. As séries de tempo do volume de prêmio do seguro agrícola regional foram projetadas com modelos SARIMA, *Neural Network Autoregressive* (NNAR), *Multiple Aggregation Prediction Algorithm* (MAPA), *Trigonometric Box-Cox Transformation, ARMA Errors, Trend And Seasonal Components* (TBATS) e o *Extreme Learning Machine* (ELM), além de uma variável mensal derivada do volume de subvenção anual ao prêmio do seguro agrícola. O modelo SARIMA mostrou melhor desempenho de previsão na região Sul, Centro-Oeste e Nordeste ao passo que o ELM registrou melhor previsão para a região Sudeste. Por sua vez, na segunda frente foi tratado o problema da ocorrência de sinistro, que é o evento de maior saída de dinheiro das seguradoras. Utilizou-se os algoritmos *Random Forest* (RF), Support Vector Machine (SVM) e *k-Nearest Neighbors* (k-NN) para o treinamento e teste de uma base de dados contendo mais de 70 mil contratos de seguro agrícola. Com o objetivo de projeção da ocorrência de sinistro no contrato, o método RF mostrou a melhor performance em termos de redução do Erro Quadrático Médio (MSE), tanto durante o treinamento quanto no conjunto de teste, sendo que o último contém exclusivamente os contratos do ano de 2017.

Palavras-chave: Seguro agrícola, Sinistro, Prêmio, Previsão, Séries de tempo

ABSTRACT

Prediction of premium and claim of Brazilian agricultural insurance market

One of the biggest challenges of the agricultural producer is to deal with a diversity of risks, exogenous or endogenous to their activity. What is in reach of the producer is treated with technology, but there are still those exogenous risks, especially linked to climatic events. Agricultural insurance serves to address such risks by transferring costs to an economic agent, the insurer. The premium and the occurrence of claims are variables of the utmost importance for the participants of this market, and it is in the interests of the insurers that they anticipate the possibility of better financial planning. The objective of this study was to advance on these two fronts: in the first front, the regional agricultural insurance premium was the target of modeling for its prediction. The regional agricultural insurance premium volume time series were predicted with SARIMA, Neural Network Autoregressive (NNAR) models, Multiple Aggregation Prediction Algorithm (MAPA), Trigonometric Box-Cox Transformation, ARMA Errors, Trend And Seasonal Components (TBATS) and Extreme Learning Machine (ELM), plus a monthly variable derived from the annual subsidy volume of the agricultural insurance premium. The SARIMA model showed better prediction performance in the South, Midwest and Northeast regions, while the ELM recorded a better forecast for the Southeast region. In turn, the second front dealt with the problem of the occurrence of claims, which is the event of greater outflow of money from insurers. Using Random Forest (RF), Support Vector Machine (SVM) and K-Nearest Neighbors (k-NN) for training and database testing containing more than 70,000 contracts agricultural insurance. The Random Forest (RF), Support Vector Machine (SVM) and k-Nearest Neighbors (k-NN) algorithms were used to train and test a database containing more than 70,000 agricultural insurance contracts. With the objective of projecting the occurrence of claims in the contract, the RF method showed the best performance in terms of reduction of the Mean Square Error (MSE), both during training and in the test set, and the latter contains exclusively the contracts of the 2017.

Keywords: Crop insurance, Claims, Premium, Prediction, Time series

LISTA DE FIGURAS

Figura 1. Relação entre volume de subvenção anual do PSR, em R$ milhões a preços correntes, e número de apólices do seguro agrícola, em milhares 23

Figura 2. Distribuição de apólices do seguro agrícola no Brasil em 2006 e 2018 24

Figura 3. Fluxo de algoritmo MAPA 29

Figura 4. Estrutura do algoritmo SLFN 31

Figura 5. Representação da validação cruzada e séries de tempo 33

Figura 6. (a) Série de subvenção anual do seguro agrícola, por região, em R$ milhões, deflacionado pelo IGP-M com ano base em 2018; (b) Série mensal derivada da subvenção do seguro agrícola anual, por região, em R$ milhões, deflacionado pelo IGP-M com mês base em janeiro de 2018, utilizando método proposto em Sax e Steiner (2013). 37

Figura 7. (a) Ajuste do modelo SARIMA(0,0,0)(1,1,0) à série de prêmio mensal do seguro agrícola da região Sul; (b) Ajuste do algoritmo ELM à série de prêmio mensal do seguro agrícola da região Sudeste; (c) Ajuste do modelo SARIMA(1,0,0)(1,1,2) à série de prêmio mensal do seguro agrícola da região Centro-Oeste; (d) Ajuste do modelo SARIMA(13,0,0)(2,1,0) à série de prêmio mensal do seguro agrícola da região Nordeste. ... 40

Figura 8. Representação do modelo *Random Forest*. 57

Figura 9. Representação do modelo SVM não linear. 60

Figura 10. Representação do k-*fold cross validation*..... 62

Figura 11. Matriz de confusão. 63

Figura 12. Evolução do Erro Quadrático Médio (MSE) por *fold*. (a) *Random Forest*; (b) *Support Vector Machine*; (c) *k-Nearest Neighbours*..... 67

Figura 13. (a) Relações entre o número de árvores e MSE; (b) Relações entre as variáveis por nó das árvores e MSE..... 68

Figura 14. Evolução do Erro Quadrático Médio (MSE) por número de vizinhos (k) selecionados para o k-NN. 69

Figura 15. Curvas ROC para os modelos estimados..... 70

Figura 16. Matriz de confusão por método, RF, SVM e k-NN..... 71

LISTA DE TABELAS

Tabela 1. Volume de prêmio recebidos pelas seguradoras que atuaram no ramo agrícola em 2018 25

Tabela 2. Resultados das medidas de acurácia das projeções por região, por modelo univariado, calculado a partir da validação cruzada 35

Tabela 3. Resultados das medidas de acurácia das projeções por região e por modelo com covariável mensal estimada para o subsídio 38

Tabela 4. Estimação dos parâmetros dos modelos SARIMA para o volume de prêmio do seguro agrícola mensal por região do Brasil entre 2006 e 2017 (erro padrão) 39

Tabela 5. Comparação entre valor original e predito das séries de prêmio por região do país para os meses de 2018, em R$ milhões, de acordo com cada modelo selecionado 41

Tabela 6. Causas de sinistro do seguro agrícola brasileiro de 2006 a 2017 55

Tabela 7. Erros quadráticos médios do SVM por tipo de função *kernel* 68

Tabela 8. Principais medidas de avaliação de predição por método 71

Tabela 9. Acurácia da predição dos sinistros ocorrido em 2017, por estado e por método 72

Tabela 10. Acurácia da predição dos sinistros ocorrido em 2017, por cultura e por método .. 73

Tabela 11. Acurácia da predição dos sinistros ocorrido em 2017, por seguradora e por método 74

1. INTRODUÇÃO

A atividade rural é fonte de renda para milhões de famílias no Brasil, representando parte importante do cotidiano do país, visto que pelo menos 4.500 dentre os 5.665 municípios brasileiros são considerados "municípios rurais", com alguma relação direta com a agricultura (BUAINAIN *et al*. 2011). O agronegócio brasileiro é um dos segmentos mais importantes para a economia brasileira, gerando emprego, renda e produtos para a nossa sociedade. A soma de bens e serviços produzidos no agronegócio atingiu 21,1% do PIB nacional em 2018 (CNA/Cepea, 2019), ressaltando a alta representatividade de um setor que enfrenta riscos peculiares.

De fato, a importância do agronegócio para o país se fica mais claro quando avaliado seu desempenho comercial. O comércio exterior desse segmento é a principal fonte de estabilização das contas externas brasileiras, com destaque para o saldo comercial, visto que o setor sempre superou o saldo comercial total do país. Entre 2011 e 2018 as exportações do agronegócio representaram em média 42,4% do total de exportações nacionais, conforme MAPA (2019), ressaltando sua importância para a solvência do Balanço de Pagamentos e para a acomodação de variações cambiais em momentos de estresse financeiro no país e no mundo. Mais do que isso, a relevância dessa atividade não se limita a geração de empregos e resultados comerciais, mas serve como um importante "colchão anticíclico", evitando o aprofundamento de cenários recessivos (BUAINAIN et al. 2011). O exemplo mais recente ocorreu em 2017, primeiro ano após uma profunda recessão que o Brasil passou entre 2014 e 2016, visto que o PIB da agropecuária avançou 12,5%, ao passo que a Indústria (-0,5%) e Serviços (0,5%) não apresentaram o mesmo desempenho (IBGE, 2019), o que encerrou o período de crise econômica.

No que tange especificamente a atividade agrícola, Duarte (2018) ressalta que há uma grande quantidade de riscos envolvendo o setor, com destaque para eventos climáticos, incêndios, pragas, utilização incorreta de tecnologia e também os riscos envolvendo a parte financeira, como a oscilação dos preços. Além disto, os grandes problemas enfrentados pelo segmento no que tange à precariedade da infraestrutura para transporte no país, a burocracia e outros componentes que oneram os custos de produção. Esses fatores impulsionaram a criação de mecanismos e decisões para a mitigação do risco, podendo citar exemplos como a migração de culturas de uma região para outra, sistemas de irrigação (reduzir impactos de secas), melhoria das sementes e até mudanças na infraestrutura na área de produção (BUAINAIN et al. 2011). Dentre outros efeitos positivos resultantes da contratação do seguro,

há a possibilidade de maior área cultivada e uso de tecnologia, estimulados pelo potencial aumento de acesso ao crédito que o produto proporciona. Podemos afirmar que o seguro:

> "Não apenas contribui para a obtenção de taxas de juros mais baixas pelo produtor rural, já que a redução do risco agrícola se traduz em menor risco de crédito, como também contribui para o desenvolvimento dos mercados financeiros, seguradores e de capitais. Em consequência, minimiza a pressão por crédito subvencionado e por socorro financeiro governamental *ex-post*, reduzindo a recorrente pressão por renegociações de dívidas rurais" (MAPA, 2018).

A gestão desse risco é necessária para reduzir a exposição da produção e a renda esperada do produtor aos custos desse risco, tendo o seguro agrícola importante papel nesse sentido. De fato, ao não conseguir tratar e mitigar todos os riscos, a transferência das suas consequências pode ser feita pelos contratos de seguro, possibilitando ao produtor assegurar parte de sua renda mesmo quando há a ocorrência de um evento de sinistro.

Como a agropecuária enfrenta riscos específicos e muitas vezes de difícil antecipação, além de afetar uma grande quantidade de unidades produtivas, as condições normais de segurabilidade ficam comprometidas, necessitando da atuação de políticas públicas para amenizar o problema, situação registrada pela literatura e pela experiência internacional[1]. De forma geral, podemos dizer que as condições para que um risco possa ser "segurável", ou seja, que o seguro tenha efetividade, são conhecidas como condições de segurabilidade. Essas condições envolvem: i) as condições de um sinistro, ou o evento de perda, devem ser bem definidos, acidentais e não intencionais, exigindo uma avaliação técnica e auditoria que valide esses pontos; ii) os objetos segurados devem ser homogêneas e independentes, permitindo uma previsão das perdas com um nível aceitável economicamente; iii) as perdas esperadas devem ser mensuráveis, bem como sua frequência média; iv) o prêmio deve ser economicamente viável; v) não ocorrer perdas catastróficas e sem riscos sistêmicos (OZAKI, 2005; MIRANDA; GLAUBER, 1997).

Dadas essas condições, o grande desafio do mercado de seguros - e o seguro agrícola é apenas um caso particular desse mercado, sofrendo dos mesmos problemas - é lidar com a gestão do risco, sobretudo analisando a frequência e as probabilidades dos sinistros, tratando o problema de forma matemática e estatística (OLIVEIRA, 2005). Além disto, a determinação do prêmio também é fundamental para a maximização de lucro da firma, ressaltando também a importância de sua previsibilidade para a gestão financeira das seguradoras. O cálculo

[1] O mercado de seguro dos Estados Unidos e da Espanha podem se destacar dentre aqueles com atuação eficaz do Estado.

errado do prêmio ou uma previsão equivocada pode resultar em perdas financeiras para as empresas, o que não é incomum conforme destaca a literatura: mercados de seguro agrícola nos Estados Unidos, Japão, Canadá, Índia e também no Brasil já mostraram situações em que o volume de prêmio arrecadados pelas seguradoras não cobriam o montante de indenizações pagas, resultados de taxas de prêmios mal calculadas (OZAKI, 2005).

A política implementada na primeira década desse século pelo governo brasileiro, o Programa de Subvenção ao Prêmio do Seguro Rural (PSR)[2], iniciou uma nova fase deste segmento no país, possibilitando novas formas de análise deste mercado e contribuindo com dados estatísticos antes não disponíveis. A assimétrica distribuição regional do seguro agrícola, com maior concentração na região Sul, pode ser um problema na hora de diversificação do risco nas carteiras das seguradoras e a predição da demanda regional, bem como a predição da ocorrência de sinistro, podem apresentar contribuições nesse sentido, antecipando movimentos e eventos desse mercado.

Desta forma, o objetivo principal desta dissertação é contribuir com a previsão de duas variáveis de grande importância no mercado de seguro agrícola: o prêmio e o sinistro. Eles serão apresentados na forma de capítulos e se dividem em: i) elaborar projeções do prêmio do seguro para as regiões Nordeste, Centro-Oeste, Sudeste e Sul do país por meio de métodos que abordem séries de tempo, com objetivo de selecionar aquele com menor erro ao comparar modelos com e sem influência de uma variável estimada a partir do volume de subvenção anual, aumentando a chance de prever corretamente uma importante variável do mercado; e ii) testar métodos para a predição de eventos de sinistros do seguro agrícola baseado em algumas variáveis disponíveis antes do evento.

O trabalho utiliza três bases principais, com destaque para os dados mensais de prêmio do seguro agrícola disponíveis no Sistema de Estatísticas da SUSEP (SES) para o período entre 2003 e 2018, por Unidade Federativa, que foram agregados de forma regional, além das informações do Atlas do Seguro Rural do Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Por sua vez, para os sinistros foram utilizadas as informações disponíveis na base de dados de indenizações entre 2006 a 2017 do MAPA, que informa as ocorrências de sinistros nesse período, mostrando região, causa, cultura, indenização, valor assegurado, etc.

A maior motivação do trabalho é testar o desempenho de métodos que auxiliem no aumento de informação no mercado de seguro agrícola ao possibilitar a projeção de variáveis

[2] Ver Apêndice.

que impactam o resultado financeiro das seguradoras, visto que o mercado de seguro agrícola tem apresentado crescimento[3] desde o início do programa de subvenção no final de 2005, sendo que há uma quantidade pequena de trabalhos tratando da predição de variáveis específicas do seguro agrícola, como no caso do prêmio e da ocorrência de sinistro.

O atual capítulo é uma seção introdutória para as duas pesquisas realizadas, esclarecendo o objetivo do trabalho. O segundo capítulo já inicia o ensaio das predições do prêmio do seguro agrícola por região, utilizando os os modelos ARIMA Sazonal, *Neural Network Autoregressive* (NNAR), *Multiple Aggregation Prediction Algorithm* (MAPA), *Trigonometric Box-Cox Transformation, ARMA Errors, Trend And Seasonal Components* (TBATS) e o *Extreme Learning Machine* (ELM), de forma a aprimorar os passos iniciais dados por Adami e Ozaki (2012), que encontraram o ARIMA sazonal como melhor método para predição do prêmio nacional. O terceiro capítulo utiliza os algoritmos *Random Forest* (RF), *Support Vector Machine* (SVM) e *k-Nearest Neighbors* (k-NN) para tratar da predição da ocorrência de eventos de sinistros, adotando metodologias de *Machine Learning* e as principais medidas de avaliação de acurácia.

Referências

BUAINAIN, A; VIEIRA, P; CURY, J. **Gestão do Risco e Seguro na Agricultura Brasileira**. Funenseg, 2011.

DUARTE, G. V. **Precificação do seguro agrícola: novas abordagens utilizando distribuições de probabilidades alternativas e o uso de cópulas bidimensionais e tridimensionais**. Dissertação de mestrado disponível em: http://www.teses.usp.br/teses/disponiveis/11/11134/tde-17012019152335/pt-br.php. Acesso em: 20 jan. 2019.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Contas Nacionais Trimestrais**. Disponível em: https://ww2.ibge.gov.br/home/estatistica/indicadores/pib/defaultcnt.shtm. Acesso em: 25 fev. 2019.

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO. **Seguro Rural - Relatório Anual - Resultado Geral 2017**. 2018. Disponível em: http://www.agricultura.gov.br/assuntos/riscos-seguro/seguro-rural/relatorios-estatisticos. Acesso em: 7 out. 2018.

[3] Cabe destacar que o número de apólices com subvenção federal entre 2006 e 2018 chegou a 800.298, conforme dados do Atlas do Seguro Rural.

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO. **Estatísticas de Comercio Exterior do Agronegócio Brasileiro**. 2019. Disponível em: http://sistemasweb.agricultura.gov.br/pages/AGROSTAT.html. Acesso em: 7 fev. 2019.

OLIVEIRA; N. V. **Mercados de Seguros: Solvência, Riscos e Eficácia Regulatória**. Tese de doutorado disponível em: http : //www.ens.edu.br/publicacoes − detalhes/104 Acesso em: 12 fev. 2019.

2. APLICAÇÃO DE MÉTODOS DE SÉRIES DE TEMPO E APRENDIZAGEM DE MÁQUINA PARA PROJEÇÃO DE PRÊMIO DO SEGURO AGRÍCOLA NAS REGIÕES DO BRASIL

Resumo

O setor agrícola enfrenta problemas de risco muito particulares, com destaque para aqueles que tangem as incertezas climáticas. Como forma de compartilhar a consequência do risco das atividades agrícolas, o mercado de seguro agrícola foi desenvolvido com participação do governo, mas ainda enfrenta problemas de estimação e predição de variáveis importantes para o mercado. As regiões do Brasil foram objeto deste estudo, que comparou o desempenho entre métodos para projeção de uma das variáveis mais importantes no mercado de seguro: o prêmio. Foram utilizados os modelos SARIMA e os algoritmos NNAR, TBATS, MAPA e ELM, com e sem uma variável mensal derivada da subvenção federal anual. Os modelos foram aplicados aos dados mensais do volume prêmio do seguro agrícola das regiões Sul, Sudeste, Centro-Oeste e Nordeste do Brasil nos anos de 2006 até 2018. O modelo SARIMA univariado mostrou melhor previsão e menor erro tanto na região Centro-Oeste quanto Nordeste, ao passo que o SARIMA com a covariável registrou melhor resultado na região Sul e o ELM com covariável para a região Sudeste.

Palavras-chave: Seguro agrícola, Prêmio, Previsão, Subvenção

Abstract

The agricultural sector faces very particular risk problems, especially climate uncertainties. As a way of sharing the consequence of the risk of agricultural activities, the agricultural insurance market was developed with government participation, but still faces problems of estimation and prediction of important variables for such a market. The regions of Brazil were the targets of this study that compared the performance among methods for forecasting one of the most important variables in the insurance market: the insurance premium. SARIMA models and the NNAR, TBATS, MAPA and ELM algorithms were used, with and without a monthly variable derived from the annual federal subsidy. The methodologies were applied to the monthly on data of the premium volume of the agricultural insurance of the South, Southeast, Central-West and Northeast regions of Brazil from 2006 to 2018. It was observed that the univariate SARIMA model showed better result both in the Central-West and Northeast, while the SARIMA with the covariate recorded a better result, both in the South and the ELM with covariate for the Southeast.

Keywords: Agricultural insurance, Premium, Forecast, Subvention

JEL codes: G22, Q13, Q18

2.1. Introdução

Motivado pelo elevado encadeamento setorial, envolvendo não apenas setores ligados diretamente a produção agrícola, mas segmentos da agroindústria e de serviços, o governo trabalha com políticas agrícolas para a garantia de parte da renda do produtor, que enfrenta uma ampla gama de riscos. Dentre as políticas, devemos destacar o seguro agrícola, cujo mercado está crescendo consideravelmente desde a criação do Programa de Subvenção ao Prêmio do Seguro Rural (PSR) em 2005. Esta é uma das maneiras encontradas para proteger tanto o produtor, que pode perder toda sua fonte de renda por conta de um evento aleatório no clima, bem como para os demais elos da cadeia do agronegócio, como por exemplo os agentes que fornecem o crédito para a produção.

De forma geral, a dinâmica do mercado de seguro agrícola com o PSR começa com os produtores contratando apólices de seguro e solicitando subvenção federal para abater parte do prêmio por meio das seguradoras, que enviam as apólices para o Ministério da Agricultura, Pecuária e Abastecimento (MAPA), e este por sua vez avaliará questões legais e burocráticas, bem como a disponibilidade de recursos.

Como o volume de prêmios pagos pelos assegurados tem influência do volume de subvenção disponibilizado pelo governo e reconhecendo a importância do PSR para a massificação do seguro rural no país, o presente estudo tem por objetivo a comparação de métodos para projeção dos prêmios de seguro agrícola, que poderão ajudar nas futuras elaborações dos Planos Trienais do Seguro Rural, utilizando informações de volume de subvenção disponibilizados pelo governo. A possibilidade de reduzir critérios subjetivos para a análise do mercado envolvendo o volume de prêmios, variável que reflete a demanda de mercado (OZAKI, 2005; SHOWERS; SHOTICK, 1994), aumenta a informação disponível tanto para fins fiscais, quanto para as seguradoras, que tem o prêmio como principal fonte de receita.

Foram avaliadas as regiões Nordeste, Centro-Oeste, Sudeste e Sul, ficando a região Norte de fora por conta da baixa disponibilidade de dados, embora a sua ausência não prejudique totalmente as análises que envolvem o mercado, visto a baixa participação tanto na produção agrícola brasileira como no mercado de seguro agrícola. Utilizou-se métodos que trabalham com série de tempo para a previsão do prêmio do seguro para cada uma das regiões. Além disto, o trabalho contou com uma variável mensal derivada do volume de subvenção regional, elaborada a partir dos dados anuais de subvenção publicados pelo MAPA

e utilizando o método iniciada em Denton (1971) e adaptado por Sax e Steiner (2013) para transformar séries anuais em uma aproximação mensal.

Inicialmente, esse artigo apresenta uma breve discussão sobre o mercado de seguro agrícola no Brasil. Após ressaltar algumas estatísticas, discute-se os métodos utilizados, bem como as bases de dados que foram utilizados. Por fim, há a discussão dos resultados e a seleção dos modelos com menor erro de projeção, bem como a conclusão do trabalho.

2.2. Referencial teórico

O seguro é tema de estudo pelos economistas pelo menos desde metade do século passado, em linha com a evolução da compreensão dos impactos que a assimetria de informação tem nos resultados do mercado competitivo. O sistema de preços pode ser limitado no que tange a otimização dada a nossa incapacidade de fazer distinções (do tipo de consumidor, por exemplo) necessárias para atingir pontos ótimos numa situação sob incerteza, como é o caso dos mercados em que os seguros atuam. Há fatores que não são de mercado, como os princípios morais e de confiança, e que de certa forma são essenciais para atingir a eficiência (ARROW, 1968).

Nesse contexto, Jensen e Meckling (1976) desenvolveram um modelo, que ficou conhecido posteriormente como modelo agente-principal, que trata do problema da informação assimétrica no mercado, avaliando impactos e consequências de uma relação entre agentes em que uma parte (o agente) tem mais informação que a outra (principal). No geral, o agente presta algum serviço para o principal, e como ambos buscam maximizar a própria utilidade, é possível que o agente não apresente escolhas que coincidem com a maximização da utilidade do principal. O principal tem ferramentas para criar incentivos que devem atingir o agente, computando custos adicionais nas suas ações de modo a atingir um dado objetivo. O custo que o principal incorre para garantir que o agente tomará as decisões ótimas não é nulo, resultando em dispêndios positivos de monitoramento, e ainda assim haverá divergências no processo de maximização de utilidade (JENSEN; MECKLING, 1976). Essa situação pode ocorrer no mercado de seguro no momento em que o segurado começa adotar comportamentos mais propensos ao risco após o contrato, comportamento este que não aconteceria caso não tivesse fechado o acordo, situação conhecida como risco moral. No mercado automotivo, por exemplo, um mecanismo que as seguradoras encontraram para reduzir a chance de sinistros é a adoção da franquia, que compartilha parte dos custos de um evento danoso com o agente.

Por sua vez, Rothschild e Stiglitz (1976) discutiram os impactos da informação imperfeita, trabalhando num modelo em que o prêmio do seguro é uma das variáveis de decisão importantes tanto para o consumidor quanto para a firma (ou produtor e seguradora, no contexto do seguro agrícola), com impactos diretos na utilidade e no lucro esperado. A seleção adversa é um dos complicadores da falta de informação e possivelmente uma avaliação incorreta do prêmio justo, que reflete uma situação em que o prêmio é igual a indenização esperada (OZAKI, 2008).

No modelo teórico, se assumirmos que há dois estados da natureza para um indivíduo representativo, sendo eles o de acidente e o sem acidente, com a probabilidade do primeiro estado ocorrer sendo dada por p, e seu complementar dada por $1-p$, podemos formular uma função de utilidade esperada[4] para esse indivíduo. Considerando que ao ocorrer o estado acidente, há uma perda d, mas há um mercado que cobre um montante a_2 via seguro pelo prêmio a_1, podemos mostrar que a renda (W) dos indivíduos pode assumir dois valores caso contrate um seguro:

$$W_1 = W - a_1 \tag{2.1}$$

$$W_2 = W - d + a_2 - a_1 \tag{2.2}$$

sendo W_1 a renda num estado sem acidente, e W_2 no caso em que há o acidente. Desta forma, se tomarmos $U(\cdot)$ como a função de utilidade da renda, teremos que o valor esperado assume tal função:

$$V_1(W_1, W_2, p) = (1-p)U(W_1) + pU(W_2) = V_1(W, a_1, a_2, d, p) \tag{2.3}$$

Desta forma, o indivíduo vai fazer sua escolha baseado na maximização desta função, lembrando que deverá comparar com o valor esperado de um cenário onde ele não contrata o seguro, ou seja, que valerá tal desigualdade

$$V_1(W_1, W_2, p) \geq V_2(W, d, p) = pU(W-d) + (1-p)U(W) \tag{2.4}$$

sendo o valor a_1 (o prêmio) uma variável importante na determinação da utilidade esperada e na tomada de decisão do indivíduo.

[4] Para o conceito de função de utilidade esperada, ver Cusinato (2003).

Por outro lado, o prêmio também tem importância no lucro esperado das seguradoras (π), e se considerarmos tais firmas como neutras ao risco, num modelo simples elas vendem o produto aos indivíduos buscando maximizar tal função de lucro esperado dada por

$$\pi_1(p, a_1, a_2) = (1 - p)a_1 - pa_2 = a_1 - p(a_1 + a_2) \tag{2.5}$$

na situação em que inexistem os demais custos, com o de transação. Se todas as expressões forem válidas em um dado mercado, é garantido que um contrato de seguro que for demandado pelo indivíduo e que maximize $V_1(\cdot)$ e $\pi_1(\cdot)$ será ofertado (ROTHSCHILD, STIGLITZ, 1976).

A situação muda se considerarmos que o indivíduo não é mais representativo e há mais de um tipo de indivíduo: o de baixo e o de alto risco. A existência de indivíduos de maior risco gera uma externalidade negativa, tornando a situação dos indivíduos de risco baixo pior do que em um cenário onde não há aqueles de alto risco.

Se fosse possível a identificação de cada indivíduo por parte das seguradoras, teríamos um equilíbrio separador[5], a questão é que a assimetria de informação nesse mercado dificulta a tal identificação. Ahsan *et al*. (1982) tentam estabelecer um mecanismo de seguro obrigatório para forçar a sinalização do risco de cada contratante, pois os indivíduos de baixo risco contratariam apenas o volume obrigatório, e o de maior risco possivelmente buscariam volumes adicionais. O problema é, conforme apontam Rothschild e Stiglitz (1976), que esse não é um meio rentável, dado que a sinalização só viria após a realização do contrato, ou seja, é possível que (2.5) não tenha sido maximizada. Os autores mostram que não há equilíbrio em um mercado competitivo de seguros, principalmente pelo fato de que o custo de agregar os indivíduos de menor risco é baixo, e o custo de separar os indivíduos de maior risco é alto.

Esta é a situação tratada por Pauly (1974) em que a seguradora não consegue definir precisamente características fundamentais do segurado e que são utilizadas para o cálculo da probabilidade dos estados futuros do objeto a ser segurado. Essa ausência de informação completa pode resultar em estados nos quais o segurador capta apenas consumidores de alto risco, o que pode tornar a esperança do lucro da firma negativo e insustentável.

---

[5] Considere que p_a é a probabilidade de acidente dos indivíduos de alto risco e p_b daqueles de baixo risco, sendo θ a fração desses últimos na população total. Caso o mercado consiga diferenciar os indivíduos e venda diferentes contratos para diferentes indivíduos, teremos um equilíbrio separador, do contrário, será o mesmo contrato para todos por meio da probabilidade média desse mercado de seguros dada por $\bar{p} = \theta p_b + (1 - \theta)p_a$, resultando no equilíbrio agregador.

No caso do seguro agrícola, Ahsan *et al*. (1982) ressaltam que cada produtor tem mais informações sobre seus próprios riscos do que os demais agentes do mercado de seguro, o que pode resultar no problema de seleção adversa, ou seja, apenas os produtores em regiões e de atividades de maior risco vão adquirir seguro. Apenas seguradoras que abrangem uma vasta variabilidade de áreas com climas, culturas e perfis diferentes, podem reduzir os riscos na atuação deste mercado. Num cenário extremo, onde a informação assimétrica é tão forte de forma que não é possível identificar diferença entre os produtos, o mercado de seguro agrícola, mesmo com programas do governo, se torna inviável (MIRANDA; GLAUBER, 1997). O risco moral, no contexto desse seguro, ocorre quando um produtor registra um menor cuidado com a manutenção de uma elevada produtividade ou proteção de riscos climáticos após a adesão do seguro.

2.2.1. O seguro agrícola no Brasil

Embora exista o mercado de seguro agrícola, é necessário ressaltar que ele se difere fundamentalmente dos demais mercados de seguro, conforme apontam Buainain *et al*. (2014). Enquanto as condições tradicionais de segurabilidade envolvem riscos mensuráveis e é necessário que as unidades expostas sejam independentes entre si, no caso do seguro agrícola tais condições são prejudicadas pela não independência dos casos de sinistro, visto a proximidade entre os segurados e dos produtos. Por exemplo, quando o rendimento de uma dada cultura, que é concentrada em uma região, for prejudicado pela falta ou excesso de chuva em seu período de plantio, possivelmente todos os produtores serão atingidos e, se tiverem seguro, resultam num volume concentrado de sinistro para as seguradoras arcarem.

Nesse caso, para o desenvolvimento deste mercado é importante a participação do Estado, inicialmente com subvenções, para compartilhar o custo com o risco e derrubar custo para todos os agentes que participam do mercado (produtores e seguradoras), aumentando a oferta e demanda pelo seguro agrícola e criando condições para a evolução desse segmento (GUIMARÃES; NOGUEIRA, 2009). No Brasil, o mercado de seguro agrícola começou a ser desenvolvido desde a década de 50 e 60 com a criação da Companhia Nacional de Seguro Agrícola e do Fundo de Estabilidade do Seguro Rural (FESR). Entretanto, o produto ganhou efetivamente maior adesão a partir do final de 2005, com o Programa de Subvenção ao Prêmio do Seguro Rural (PSR). O PSR busca expandir o seguro agrícola e o inserir como amenizador das consequências dos riscos climáticos das atividades agrícolas, sobretudo na estabilidade da renda e diminuir o preço de aquisição da apólice pelo prêmio (MAPA, 2017).

O número de apólices vem subindo, acompanhando a evolução no volume de subvenção, exibindo uma correlação positiva[6] entre o período de 2006 e 2018, conforme a Figura 1. Quanto maior for o volume de recurso destinado ao PSR, maior será o número de produtores que será contemplado pelo programa, ao passo que a redução do volume de subvenção impacta no desenvolvimento do seguro agrícola (ADAMI; OZAKI, 2012).

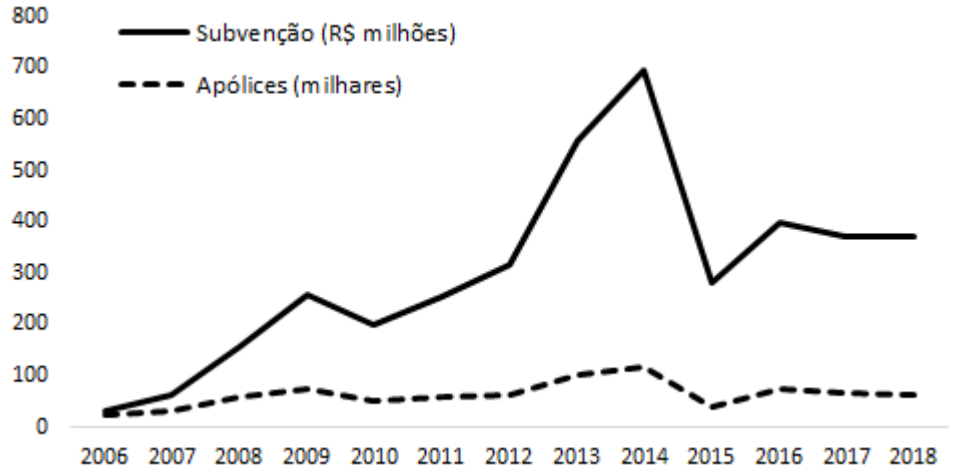

Figura 1. Evolução doe volume de subvenção anual do PSR, em R$ milhões a preços correntes, e número de apólices do seguro agrícola, em milhares

Fonte: Elaboração própria com dados do MAPA

Ainda em relação ao número de apólices, é possível observar uma significativa evolução entre os anos de 2006 e 2018. De fato, o número de apólices passou de 21.783 para 63.554, aumentando sua presença em outros lugares além da região Sudeste e Sul do país (Figura 2).

Utilizando os dados do Censo do Seguro Rural, Macedo *et al*. (2013) mostraram a evolução do PSR entre o período 2006 e 2010, ilustrando a participação regional nesse mercado e ressaltando a concentração do programa na região Sul e Sudeste do país, o que pode ser verificado também na Figura 2. Os autores ressaltam que os recursos disponibilizados para o PSR no período estudado foram insuficientes para a evolução adequada do mercado, sobretudo por contingenciamentos no MAPA em 2009 e 2010.

[6] A correlação significativa dos dados apólices e volume de subvenção disponíveis no MAPA foi obtida a partir do coeficiente de correlação de Pearson, proposto em Pearson (1896). O coeficiente estimado foi de 0,93, rejeitando a hipótese nula que o coeficiente é igual a zero (p-valor = 4,014e-06), com um intervalo de confiança de 0,778459 e 0,979335.

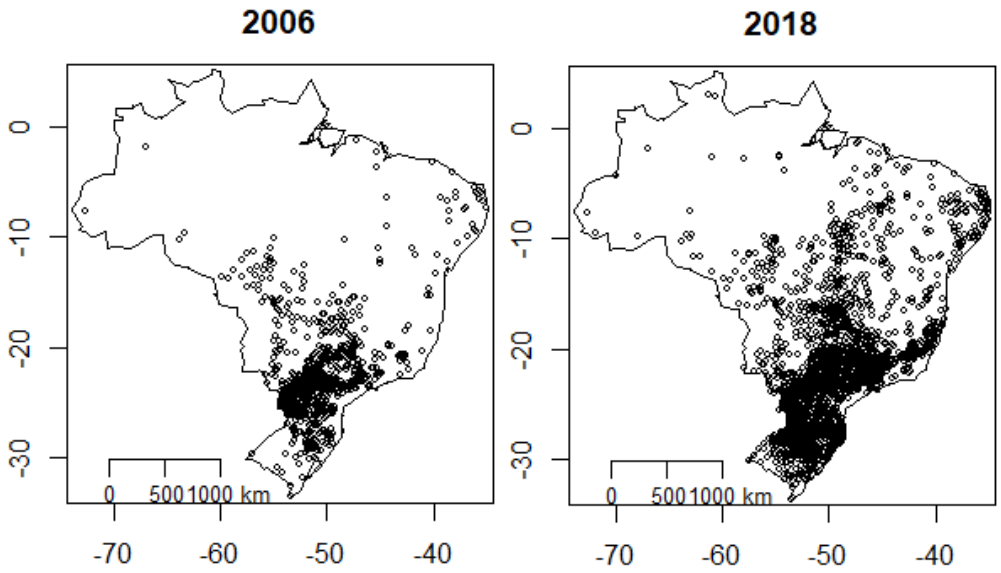

Figura 2. Distribuição de apólices do seguro agrícola no Brasil em 2006 e 2018
Fonte: Elaboração própria com dados do MAPA

Além das diferenças regionais, há grande diferença na importância do seguro agrícola nas carteiras das empresas que atuam nesse mercado. Das variáveis que impactam a performance financeira da seguradora, o volume de prêmio tem grande importância na otimização do lucro dessas empresas, sendo que no contexto do mercado brasileiro, o volume de prêmio recebido por todas as seguradoras do mercado, independente do ramo de atuação, chegou a R$ 112,62 bilhões em 2018, representando grande fonte de faturamento nas demonstrações contábeis divulgadas no Sistema de Estatísticas da SUSEP. Neste mesmo ano o total de prêmio pagos referente apenas ao ramo agrícola, envolvendo as 11 empresas que atuaram nesse segmento, foi de R$ 1,05 bilhão, representando 4% do total de prêmios recebidos por essas mesmas 11 empresas naquele período.

Tabela 1. Volume de prêmio recebidos pelas seguradoras que atuaram no ramo agrícola em 2018

Empresa	Agrícola (A)	Total (B)	A/B
Aliança do Brasil Seguros S.A.	R$220.606.796	R$629.429.632	35,0%
Allianz Seguros S.A.	R$75.357.636	R$3.062.175.482	2,5%
Companhia Excelsior de Seguros	R$25.477.351	R$172.128.205	14,8%
Essor Seguros S.A.	R$180.729.066	R$391.074.354	46,2%
Fairfax Brasil Seguros Corporativos S.A	R$48.174.573	R$490.836.461	9,8%
Mapfre Seguros Gerais S.A.	R$111.923.073	R$6.253.104.315	1,8%
Markel Seguradora do Brasil S.A.	R$68.595.265	R$68.815.720	99,7%
Porto Seguro Companhia de Seguros Gerais	R$30.998.843	R$8.954.513.664	0,3%
Sancor Seguros do Brasil S.A.	R$153.540.695	R$318.812.646	48,2%
Swiss Cre Corp. Solutions Br. Seg. S.A.	R$128.153.851	R$760.106.825	16,9%
Tokio Marine Seguradora S.A.	R$8.390.771	R$5.117.050.810	0,2%
Total	R$1.051.947.920	R$26.218.048.114	4,0%

Nota: O volume agrícola refere-se aos ramos 1101, 1102, 1161 da SUSEP, e o volume total comporta todos os ramos de seguro do mercado brasileiro.

Fonte: Elaboração do autor com dados da SUSEP.

O mercado de seguro agrícola ainda é pequeno quando comparado com o tamanho do mercado do seguro brasileiro. Além disso, a literatura que envolve os dados do PSR e do prêmio de seguro agrícola ainda é escassa[7]. Podemos ressaltar que Ozaki (2005) apresentou métodos estatísticos com abordagem paramétrica e não-paramétrica envolvendo a produtividade para precificação de contratos de seguro agrícola. Até então, o mercado brasileiro não contava com PSR, sendo que os programas vigentes, como o Programa de Garantia da Atividade Agropecuária (Proagro), apresentavam grandes índices de sinistralidade, ou seja, a relação entre indenizações pagas e os prêmios arrecadados era alta. A determinação incorreta da taxa de prêmio resultava em problemas como risco moral e seleção adversa, sendo que o estudo buscou apresentar três métodos para melhorar o cálculo das taxas de prêmio. A incorporação de componentes espaciais em metodologias de determinação de prêmio melhora as estimativas das distribuições de probabilidade. Já Ozaki (2013) utilizou o modelo ARIMA na determinação do volume ideal de subvenção para cobrir o volume de crédito rural, de forma a garantir a estabilidade do sistema e a massificação do seguro rural. Para o volume do prêmio do seguro agrícola, Adami e Ozaki (2012) estudaram métodos para

[7] 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.

a projeção do volume nacional com metodologias já conhecidas nos estudos de séries de tempo, como os modelos SARIMA.

A projeção do volume de prêmio nos mercados regionais pode ser importante para o planejamento e atuação do mercado por parte dessas seguradoras, sobretudo no planejamento de suas provisões técnicas. Além disto, o maior volume de informação relativa ao contexto regional contribui para a diversificação de risco das carteiras das seguradoras, principalmente considerando que o volume de prêmio é variável relevante para decisão para determinar o portfólio das seguradoras (LI, 1995, apud THOMPSON *et al*., 1972). Por sua vez, o presente estudo optou pela abrangência regional e a adição de métodos que usam redes neurais para previsão. A comparação entre previsão com modelos de redes neurais e modelos SARIMA já foi feita em pesquisas envolvendo o agronegócio, embora não no seguro agrícola (OLIVEIRA *et al*., 2002; SOARES *et al.*, 2015; FLORES; WERNER, 2007), sendo que o bom desempenho dos modelos em termos de previsão levanta a hipótese de que a utilização em outras séries do segmento agrícola podem superar os resultados dos demais modelos de previsão de séries de tempo.

2.3. Métodos

O critério de seleção de modelos foi a Raiz do Erro Quadrático Médio (RMSE, na sigla em inglês) como método de acurácia (HYNDMAN; KOEHLER, 2006). Dado uma série Z_t e sua previsão $\bar{Z}_t$, pode-se definir a métrica conforme a expressão:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Z_t - \bar{Z}_t)^2} \tag{2.6}$$

Desta forma, quanto menor for tal métrica, melhor será o valor do ajuste da previsão em relação a série original, servindo como um parâmetro para tomar decisão entre modelos de métodos distintos (ALEXANDER, 2006). O trabalho utilizou o software R, usando os pacotes *forecast, nnfor, MAPA, tempdisagg, deflateBR* e *lmtest.*

2.3.1. SARIMA

Os modelos SARIMA (ou ARIMA sazonal) são uma extensão dos modelos ARIMA, descrita no Apêndice, em que uma série de tempo Z_t é gerada por um processo SARIMA(p,q,d)(P,D,Q) se vale

$$(B)\Phi(B^s)(1-B)^d(1-B^s)^D Z_t = \theta(B)\Theta(B^s)a_t \tag{2.7}$$

sendo que

1. $\phi(B) = 1 - \phi_1 B^1 - \phi_2 B^2 - \cdots - \phi_p B^p$ é o operador autoregressivo AR estacionário de ordem *p*;
2. $\theta(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - \cdots - \theta_q B^q$ é o operador de médias móveis MA invertível de ordem *q*;
3. $(1-B)^d$ é o processo de integração de ordem *d*;
4. $\Phi(B) = 1 - \Phi_1 B^{1s} - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$ é o operador autoregressivo sazonal estacionário de ordem *P*;
5. $\Theta(B) = 1 - \Theta_1 B^{1s} - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$ é o operador de médias móveis sazonal de ordem *Q*;
6. $(1-B^s)^D$ é o operador de diferença sazonal;
7. a_t é um ruído branco,

sendo *B* o operador de defasagem, *s* é a dimensão do período sazonal.

No caso em que há série de tempo X_t exógena como covariável, o modelo SARIMA é conhecido como SARIMAX (ou ARIMAX sazonal) e como descrito em Yang *et al*. (1995) e Huang *et al*. (2005) a expressão (2.7) se torna:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Z_t = \psi(B)X_t + \theta(B)\Theta(B^s)a_t \tag{2.8}$$

em que $\psi(B)$ é o operador de defasagem e os coeficientes da covariável, valendo $\psi(B) = 1 - \psi_1 B^1 - \psi_2 B^2 - \cdots - \psi_m B^m$.

Inicialmente foram testadas a estacionariedade das séries pelo teste de raízes unitárias Dickey-Fuller Aumentado (ADF) proposto em Dickey e Fuller (1981), depois foram identificados os modelos pelas chamadas função de autocorrelação (FAC) e função de autocorrelação parcial (FACP)[8] e selecionado aquele com menor AIC (*Akaike Information Criterion*[9]). Após sua estimação, é necessário fazer alguns testes com os resíduos: o teste de

[8] A FAC é dada por $\rho_{t,s} = \frac{Cov(Z_t,Z_s)}{\sqrt{(Var(Z_t)Var(Z_s))}}$ e a FACP dada por $\phi_{k,k}$, podendo ser encontrada ao satisfazer os sistemas das equações de Yule-Walker ($\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \cdots + \phi_{kk}\rho_{j-p}$) pela regra de Cramer.

[9] Trata-se de um critério para comparação e seleção de modelos, proposto em Akaike (1973).

Box e Pierce (1970) ou de Ljung e Box (1978), para verificar se não há autocorrelação dos resíduos, e o teste de Shapiro e Wilk (1965) para verificar se seguem uma distribuição normal.

2.3.2. NNAR

Os modelos *Neural Network Autoregression* (NNAR) são modelos híbridos de redes neurais do tipo *feedforward* e dos modelos clássicos autorregressivos (AR), constituída de uma possível parte linear e uma não linear. A literatura apresenta redes neurais como uma rede de neurônios organizados em camadas: as entradas (*inputs*) formam a camada inferior e as previsões (ou *outputs*) formam a camada superior, sendo intermediadas por uma camada oculta, que contém os neurônios, que são valores transformados do *input* inicial por algum peso e que serão novamente transformados por funções chamadas funções de ativação. O componente linear é formado pelo intercepto e pela ponderação de i defasagens da variável objetivo, assim como nos modelos AR. A parte não linear, que contém os neurônios ocultos, transforma as variáveis imputadas, ponderadas por um parâmetro γ_{ij} e um viés γ_{0j}, através de uma função não linear $\psi(.)$. Dietz (2010) e Zhang (2003) mostram que a função completa pode ser escrita como

$$\mathrm{Z_t} = \mathrm{a_0} + \sum_{i=1}^{p} a_i Z_{t-i} + \sum_{j=1}^{n} \beta_j \psi\left(\gamma_{0j} + \sum_{i=1}^{p} \gamma_{ij} Z_{t-i}\right) + \varepsilon_t \qquad (2.9)$$

sendo β_j a ponderação das funções não lineares e ε_t o componente de erro, assumido homocedástico e normalmente distribuído. Formalmente, como mostra Hyndman (2014) podemos chamar o modelo de NNAR(p,P,k), onde p são defasagens da série de tempo usada na entrada, como nos modelos autorregressivos, P as defasagens sazonais e k o número de neurônios na camada oculta.

2.3.3. Algoritmo MAPA

O modelo conhecido como *Multiple Aggregation Prediction Algorithm* (MAPA) foi proposto por Kourentzes *et al.*(2013) e envolve a agregação da série de tempo Z_t em alguns componentes de diferentes frequências k (por exemplo, uma série mensal pode ser agregada em série bimestral, em que $k = 2$, trimestral, em que $k = 3$ e etc.), que são projetados individualmente por métodos de suavização exponencial da família dos modelos *Exponential Smoothing* (ETS), descrita em Hyndman *et al.*(2008) e no Apêndice deste trabalho. O ETS

capta os componentes estruturais de uma série de tempo, como tendências (T_t), sazonalidades (S_t), nível (L_t) e ruídos (ε_t), elaborando a projeção dos três primeiros.

Assim, as novas séries Z_t^k são projetadas h períodos à frente por meio de seus componentes estruturais e novamente combinadas para resultar nas previsões da série original. Athanasopoulos *et al.* (2011) iniciaram a modelagem empírica e a avaliação do impacto no desempenho dos métodos de suavização exponencial, com a agregação de dados mensais e trimestrais para a frequência anual e realizando projeções dessas séries.

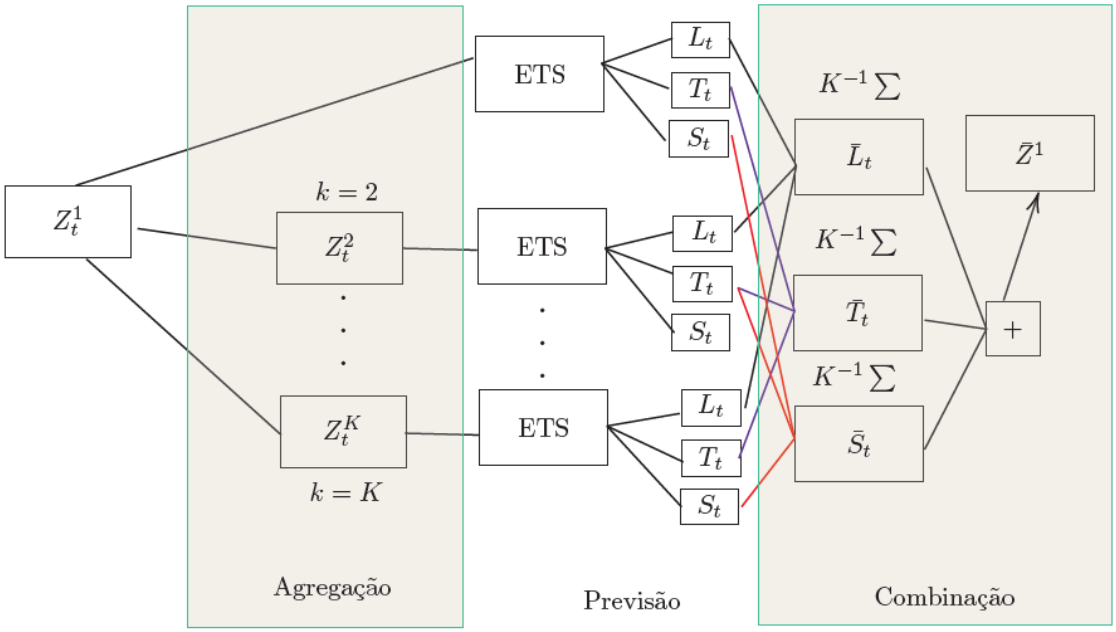

Figura 3. Fluxo de algoritmo MAPA
Fonte: Elaboração própria baseado em Kourentzes *et al.*(2013).

A combinação ponderada, como por exemplo a média, é a mais comum na literatura, podendo alcançar até uma menor variância dos erros de previsão (HIBON; EVGEBIOU, 2005). A projeção final deverá ser aquela em que o modelo ETS minimizou o AIC e com uma agregação que resulte num baixo RMSE.

2.3.4. Algoritmo TBATS

O algoritmo *Trigonometric Box-Cox Transformation, ARMA Errors, Trend And Seasonal Components* (TBATS) proposto em De Livera *et al.* (2010) é um avanço na proposta de Taylor (2003) para a versão linear do modelo Holt-Winters, envolvendo

modificações de modelos de espaço de estados, incorporando transformações de Box-Cox, representações de Fourier com coeficientes variáveis e correção de erros do tipo ARMA, além de componentes trigonométricos que acomodam séries de tempo com sazonalidades mais complexas, bem como tratar de possíveis não linearidades, podendo ser descrito para uma série Z_t como

$$\mathrm{Z}_{\mathrm{t}}^{\omega} = \begin{cases} \dfrac{\mathrm{Z}_{\mathrm{t}}^{\omega} - 1}{\omega}, \text{ se } \omega \neq 0 \\ \log \mathrm{Z}_{\mathrm{t}}, \textit{do contrário} \end{cases} \tag{2.10}$$

$$\mathrm{Z}_{\mathrm{t}}^{\omega} = \mathrm{L}_{\mathrm{t}-1} + \phi \mathrm{T}_{\mathrm{t}-1} + \sum_{i=1}^{T} S_{t-1}^{i} + \mathrm{d}_{\mathrm{t}}; \tag{2.11}$$

$$\mathrm{L}_{\mathrm{t}} = \mathrm{L}_{\mathrm{t}-1} + \varphi \mathrm{T}_{\mathrm{t}-1} + \alpha \mathrm{d}_{\mathrm{t}} \tag{2.12}$$

$$\mathrm{T}_{\mathrm{t}} = (1 - \varphi)\mathrm{T} + \varphi \mathrm{T}_{\mathrm{t}-1} + \beta \mathrm{d}_{\mathrm{t}} \tag{2.13}$$

$$\mathrm{S}_{\mathrm{t}}^{\mathrm{i}} = \sum_{j=1}^{k_i} \mathrm{S}_{\mathrm{j,t}}^{\mathrm{i}} \tag{2.14}$$

$$\mathrm{S}_{\mathrm{j,t}}^{\mathrm{i}} = \mathrm{S}_{\mathrm{j,t-1}}^{\mathrm{i}} \cos\lambda_{\mathrm{j}}^{\mathrm{i}} + \mathrm{S}_{\mathrm{j,t-1}}^{*\mathrm{i}} \mathrm{sen}\lambda_{\mathrm{j}}^{\mathrm{i}} + \gamma_1^{\mathrm{i}} \mathrm{d}_{\mathrm{t}} \tag{2.15}$$

$$\mathrm{S}_{\mathrm{j,t}}^{*\mathrm{i}} = -\mathrm{S}_{\mathrm{j,t-1}}^{\mathrm{i}} \mathrm{sen}\lambda_{\mathrm{j}}^{\mathrm{i}} + \mathrm{S}_{\mathrm{j,t-1}}^{*\mathrm{i}} \cos\lambda_{\mathrm{j}}^{\mathrm{i}} + \gamma_2^{\mathrm{i}} \mathrm{d}_{\mathrm{t}} \tag{2.16}$$

$$\mathrm{d}_{\mathrm{t}} = \sum_{i=1}^{p} \phi_{\mathrm{i}} \mathrm{d}_{\mathrm{t}-1} + \sum_{i=1}^{q} \theta_{\mathrm{i}} \varepsilon_{\mathrm{t}-1} + \varepsilon_{\mathrm{t}} \tag{2.17}$$

onde ω é o parâmetro da transformação Box-Cox, L_t é o nível da série, T é a tendência de longo prazo, T_t a tendência de curto prazo, s_t^i é o i-ésimo componente sazonal, m_1,..,m_t são os períodos sazonais, d_t é o processo $ARMA(p,q)$, ε_t é um ruído branco e α, β e γ_i são parâmetros de suavização. Nas expressões que tratam da sazonalidade, $\lambda_j^i = \frac{2\pi_j}{m_i}$, γ_1^i e γ_2^i são os parâmetros de suavização, $s_{j,t-1}^i$ são os i-ésimos componentes sazonais estocásticos e $s_{j,t}^{*i}$ é sua denotação em termos de taxa de crescimento. Por fim, o termo k_i no somatório é dado por $k_i = \frac{m_i}{2}$ para valores pares de m_i e $k_i = \frac{(m_i - 1)}{2}$ para valores ímpares.

2.3.5. Algoritmo ELM

O algoritmo *Extreme Learning Machine* (ELM) foi proposto por Huang *et al.* (2006) usa como base o algoritmo de redes neurais do tipo *feedforward* com camada oculta conhecido como *Single Layer Feedforward Network*, ou SLFN (HUANG; BABRI, 1998), ilustrado na Figura 4.

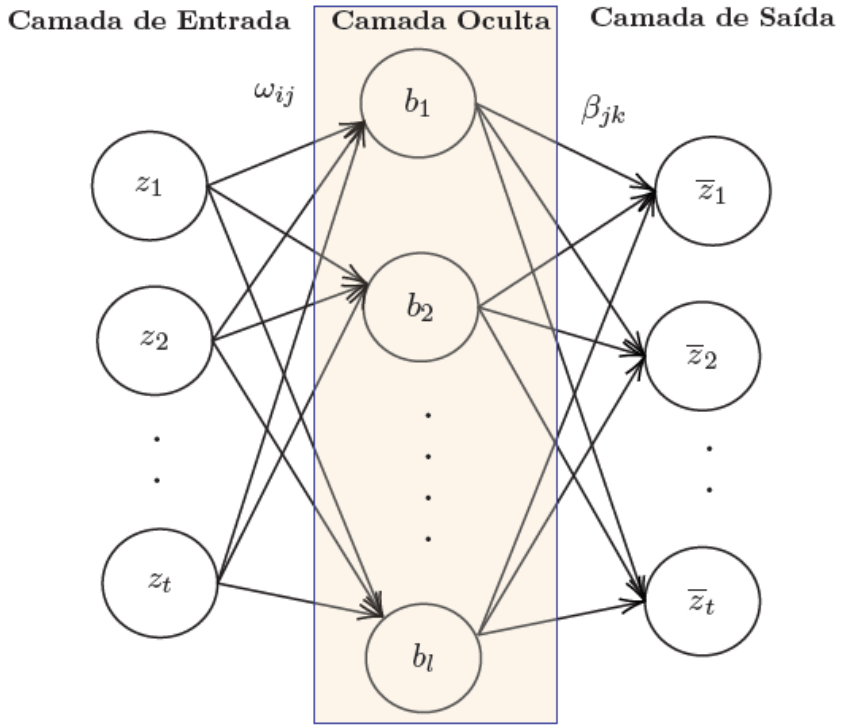

Figura 4. Estrutura do algoritmo SLFN
Fonte: Elaboração própria baseado em Li *et al.*(2018).

O ELM tem uma dinâmica de camadas semelhante ao NNAR, contendo t neurônios na camada de entrada e de saída e i na camada oculta, mas não é tão dependente dos pesos e vieses da sua camada intermediária, e sim da camada final. Considere a série Z_t e uma função de ativação $G(\omega_i, \mathrm{b}_i, Z_t)$ como a saída do i-ésimo neurônio oculto, sendo $b_i \in R$ o chamado viés e $\omega_i = (\omega_{i1}, \dots, \omega_{ij})$ o vetor de pesos que conecta o i-ésimo neurônio da camada oculta com o t-ésimo neurônio da camada de entrada. Assim, podemos considerar uma função de saída é dada por

$$\overline{z_t} = \sum_{i=1}^{l} \beta_i \, G(\omega_i, b_i, z_t) \tag{2.18}$$

em que $\beta_i = (\beta_{i1}, \dots, \beta_{ik})$ é um vetor de pesos que conecta o i-ésimo neurônio da camada oculta com o k-ésimo neurônio da camada de sáida. Considere Z_t, com $t = 1,2, \dots, m$, um vetor de entrada composto pelos dados de treinamento, $\overline{Z_t}$ o correspondente vetor de saída, além de levar em conta $w_i \in R^m$, vetor de pesos, e $b_i \in R$, sendo ambos aleatórios. A expressão pode ser escrita na forma de matrizes tal que vale $H\beta = \overline{Z_t}$, como mostra Singh e Balasundaram (2007), sendo necessário para a solução desse grande conjunto de equações.

O treino do algoritmo ELM, com base no SLFN, corresponde a encontrar a norma mínima da solução dos mínimos quadrados da matriz de equações $H\beta = \overline{Z_t}$ dado por $\min_{\beta} |H\beta - \overline{Z_t}|$, começando com valores aleatórios dos pesos e dos vieses na camada oculta. É calculado H^+, sendo essa a matriz inversa de Moore-Penrose da matriz H (descrita no Apêndice), usada para calcular a solução de sistema de equações lineares que não possuem uma solução única, ou então a norma mínima de tal sistema. Depois, encontra-se a matriz de pesos $\overline{\beta}$ dada por $\overline{\beta} = H^+\overline{Z_t}$ e, por fim, para cada entrada z_t, a saída $\overline{z_t}$ será dada pela função $\overline{z_t} = \sum_{i=1}^{l} \overline{\beta_i} G(\omega_i' z_t + b_i)$.

2.3.6. *Naive* (modelo base)

Seguindo Hyndman (2014) e Pappenberger *et al.* (2015), optou-se por comparar os métodos descritos acima com um modelo mais básico possível, de fácil reprodução e sem complexidade computacional, que servirá de base de comparação (*baseline*). O objetivo deste procedimento é que, além de escolher o algoritmo que produz o menor erro de predição dentre os métodos apresentados, a abordagem selecionada deve também superar o desempenho de um algoritmo simples. O *Naive* Sazonal com deslocamento descrito em Hyndman (2014) foi o método considerado como *baseline*. A projeção de h períodos à frente de Z_t é dada por

$$\bar{Z}_{t+h} = Z_{t+h-m(k+1)} + \frac{h}{t-1}\sum_{t=2}^{n}(Z_t - Z_{t-1}) = Z_{t+h-m(k+1)} + h\left(\frac{Z_t - Z_1}{t-1}\right) \tag{2.19}$$

em que m é o período sazonal (12, no caso de séries mensais) e k é o inteiro resultante de $\frac{h-1}{m}$. Em suma, o algoritmo assume que a projeção h períodos à frente é determinada por um valor

passado já conhecido (a primeira parte do termo do lado direito da expressão acima) acrescido de uma constante de deslocamento.

2.3.7. Validação cruzada (VC) e índice de habilidade (*Skill*)

O cálculo do RMSE foi avaliado no contexto de um processo conhecido como validação cruzada, que permite a confirmação da capacidade de generalização do algoritmo utilizado (BERGMEIR *et al*., 2017). A validação cruzada para séries de tempo proposta em Hyndman (2014) diferencia-se do processo de validação cruzada utilizado em outros tipos de conjuntos de dados, como aqueles em corte transversal. Na abordagem proposta, o conjunto de dados de treino é dividido em subconjuntos diferentes, em que existe um primeiro conjunto com um número mínimo *k* de observações da série e os próximos conjuntos dispõem de uma observação a mais que o anterior. O processo consiste em selecionar $\mathrm{k} + \mathrm{i}$ observações da série para treino e deste conjunto utilizar $k + i + h$ para o teste do modelo, elaborando a projeção das últimas *h* observações e computando o erro de projeção.

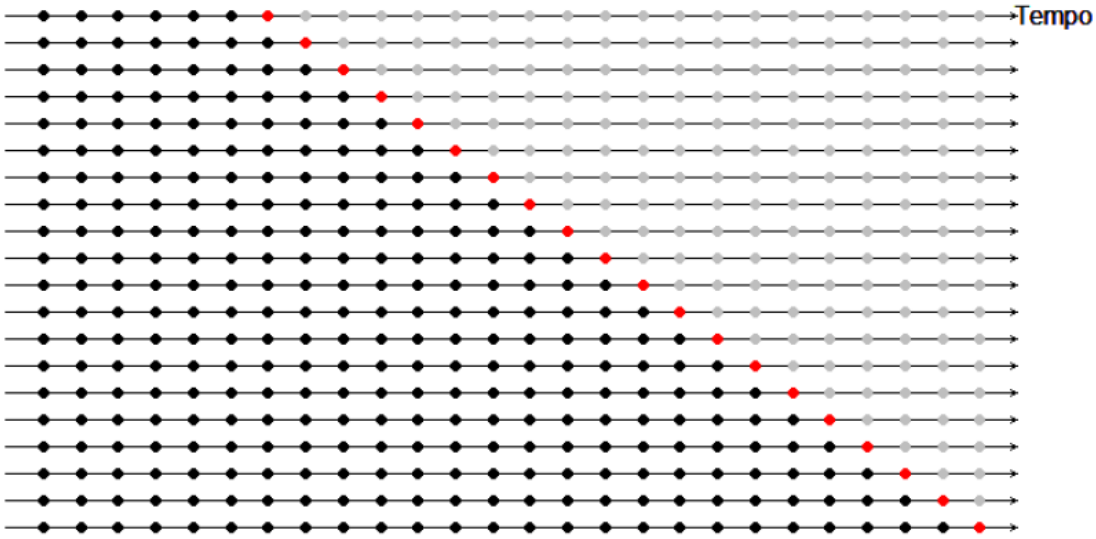

Figura 5. Representação da validação cruzada e séries de tempo

Nota: Observações em preto = conjunto de treino; observações em vermelho: projeções dos respectivos conjuntos de treino; observações em cinza: conjunto da série que não foi utilizada nem para treino e não foi projetada.

Fonte: Elaboração própria baseado em Hyndman (2014).

Ao todo, se o conjunto tiver n observações, então a validação cruzada terá $n - h - k$ cortes temporais testados e um valor de RMSE para cada teste, que são compilados em uma média final. Para facilitar o confrontamento dos métodos para cada região com o *baseline*, adotou-se o índice de proficiência (ou *Skill*) proposto em Pappenberger *et al.* (2015), que vale:

$$\text{Skill} = \frac{\mathrm{f}\left(\overline{\mathrm{Z}}_{\mathrm{t}}^{\mathrm{i}}\right)}{\mathrm{f}\left(\overline{\mathrm{Z}}_{\mathrm{t}}^{\mathrm{b}}\right)} \tag{2.20}$$

em que $\overline{Z}_t^i$ são as projeções do algoritmo i, $\overline{Z_t^b}$ são as projeções do *baseline* e $f(\cdot)$ é uma função de performance, que no presente trabalho será representada pelo RMSE.

2.3.8. Base de Dados

Para a elaboração dos modelos de previsão utilizou-se as séries mensais de prêmio de seguro direto disponíveis por unidade federativa no Sistema de Estatística da SUSEP (SES), que foram deflacionados pelo Índice Geral de Preços do Mercado (IGP-M) do IBRE/FGV para que o modelo não tenha que projetar a inflação dos períodos, ou seja, trabalhou-se com as séries em valores reais. A frequência dos dados é mensal e compreende o período de janeiro de 2003 a dezembro de 2018, e estão a preços de janeiro de 2018. Além disto, foram utilizados as séries anuais do volume de subvenção ao seguro agrícola regional disponibilizado pelo governo, tendo o MAPA como fonte.

Com o objetivo de testar previsões do volume de prêmios pagos anualmente, retirou-se as últimas 12 observações de cada série de tempo e após a seleção do modelo pela validação cruzada foi projetado o prêmio do período retirado, que comportam os meses entre janeiro e dezembro de 2018, seguindo o procedimento utilizado em Thielbar e Dickey (2011).

2.3.9. Desagregando a série de subvenção

O equilíbrio do volume total de prêmios arrecadados e as indenizações pagas ocorreram após o apoio governamental por meio da subvenção (OZAKI, 2005), podendo ser variável relevante para a projeção do volume de prêmio de cada ano. A grande limitação é que existe apenas dados anuais do volume de subvenção concedidos. Com o objetivo de aproveitar essa informação, foi usado um método de desagregação temporal para replicar um comportamento mensal do volume anual de subvenção, ainda que não reflita o comportamento real mês a mês da série, mas captando o nível que cada ano apresenta. O

objetivo da desagregação temporal é encontrar uma série desconhecida $\widehat{Z_t}$ de frequência h de tal forma que sua soma resulte em uma série conhecida Z_t de frequência l, em que $l < h$. O método apresentado em Denton (1971) utiliza uma variável P_t de frequência h e do mesmo tamanho que $\widehat{Z_t}$ como apoio para a desagregação. A configuração do método utilizado pelo trabalho é aquele descrito em Sax e Steiner (2013) em que P_t é composto apenas por 1, de forma que $\widehat{Z_t}$ é definido conforme

$$\widehat{\mathrm{Z}_\mathrm{t}} = \mathrm{P}_\mathrm{t} + \mathrm{DU}_\mathrm{t} \tag{2.21}$$

em que

$$\mathrm{U}_\mathrm{t} = \mathrm{Z}_\mathrm{t} - \mathrm{CP}_\mathrm{t}, \tag{2.22}$$

tal que D é uma matriz chamada distribuição $n \times m$, sendo n o comprimento de $\widehat{Z_t}$ e m o de Z_t, U_t uma matriz $m \times 1$, e C é uma matriz de conversão conforme especificado em Sax e Steiner (2013). A determinação de D segue o proposto em Denton (1971) em que vale $D = \Sigma C'(C\Sigma C')^{-1}$, sendo $\Sigma = (\Delta'\Delta)^{-1}$, onde Δ é uma matriz $n \times n$ com 1 na diagonal principal, -1 na diagonal imediatamente abaixo da principal e 0 nas demais entradas.

2.4. Resultados e discussão

Iniciando pela avaliação dos modelos para cada região no caso univariado, a Tabela 2 retrata o resultado daqueles que exibiram as melhores medidas de acurácia, compreendendo o RMSE e ajuste do modelo dos prêmios de seguro agrícola do período de janeiro de 2006 a dezembro de 2017 pela validação cruzada, além das estimativas do *Skill* a partir dos mesmos RMSE.

Tabela 2. Resultados das medidas de acurácia das projeções por região, por modelo univariado, calculado a partir da validação cruzada

	Sul		Sudeste		Centro-Oeste		Nordeste	
	RMSE	Skill	RMSE	Skill	RMSE	Skill	RMSE	Skill
SARIMA	19,19	0,93	10,76	0,80	15,10	0,93	1,65	0,47
NNAR	25,76	1,25	14,71	1,09	16,83	1,03	3,53	1,01
ELM	17,79	0,86	12,58	0,93	15,17	0,93	3,34	0,95
TBATS	35,21	1,71	15,35	1,14	23,95	1,47	6,66	1,90
MAPA	24,91	1,21	17,61	1,30	21,24	1,30	4,67	1,33
Naive	20,58	1,00	13,50	1,00	16,31	1,00	3,51	1,00

Fonte: Resultado da pesquisa

O destaque é que nenhuma configuração dos algoritmos TBATS e MAPA superaram o desempenho do *Naive* em termos de RMSE para as regiões, refutando a hipótese de que esses métodos poderiam contribuir com a projeção das séries de prêmio do seguro agrícola regional. A média dos RMSE resultante do processo de validação cruzada desses algoritmos equivalem a pelo menos 115% do resultado do *Naive* sazonal, no caso de Sudeste, chegando até 190% no caso do Nordeste, considerando as melhores configurações para tais métodos.

Considerando a região Sul, o modelo SARIMA mostrou bom resultado durante a VC, superando o *Naive*, mas ficando abaixo do algoritmo ELM que apresentou um RMSE equivalente a 86% daquele exibido pelo *Naive* durante o VC. No caso do Sudeste, o SARIMA e o ELM também superam o *Naive*, mas o primeiro método mostrou a melhor acurácia em suas projeções, registrando apenas 80% do RMSE exibido pelo *Naive* sazonal. Resultado similar foi encontrado nos casos de Centro-Oeste e Nordeste, com bom desempenho do SARIMA e do ELM, mas com o primeiro mostrando melhor performance tanto em relação aos demais, quanto em relação ao *Skill*.

A diferença de nível de RMSE entre as regiões ocorre pela heterogeneidade de nível das próprias séries, sendo que a região Sul apresenta o maior volume de prêmio ao longo do ano e o Nordeste a menor, influenciando a escala desta métrica de acurácia.

Por sua vez, no caso em que os métodos foram estimados com covariáveis, foi necessário estimar a variável derivada da subvenção mensal por região pelo método descrito em 2.3.9. A Figura 6 ilustra os resultados da aplicação do método proposto em Sax e Steiner (2013), mostrando como a evolução da variável de subvenção anual fica suavizada na frequência mensal, mas conserva sua trajetória.

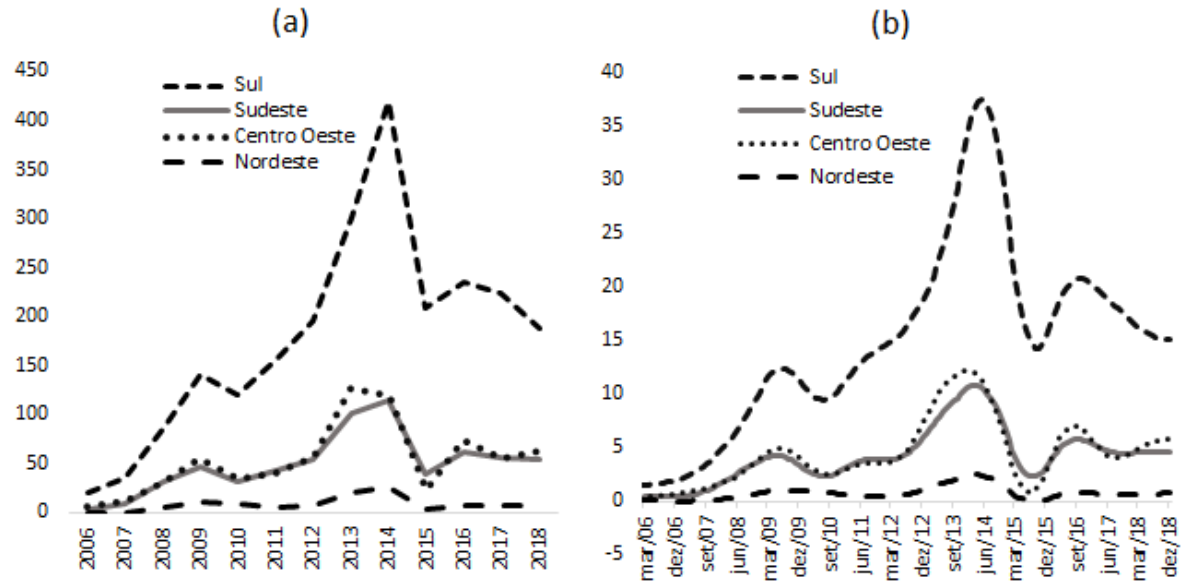

Figura 6. (a) Série de subvenção anual do seguro agrícola, por região, em R$ milhões, deflacionado pelo IGP-M com ano base em 2018; (b) Série mensal derivada da subvenção do seguro agrícola anual, por região, em R$ milhões, deflacionado pelo IGP-M com mês base em janeiro de 2018, utilizando método proposto em Sax e Steiner (2013).

Fonte: Resultado da pesquisa.

Os algoritmos TBATS e MAPA não permitem a projeção com covariável, mas como não apresentaram baixo RMSE mesmo no caso univariado, não há prejuízo na avaliação ao não os considerar. A Tabela 3 exibe os resultados da validação cruzada dos métodos no caso em que a variável estimada da subvenção foi utilizada, e tanto o ELM quanto o SARIMA superaram o desempenho *Naive* (que também foi estimado com a covariável) em todas as regiões.

Na região Sul, diferentemente do caso univariado, o SARIMA com covariável mostrou o menor RMSE e *Skill*, ao passo que o melhor método no caso do Sudeste mudou para o ELM, com ambos *Skill* próximo aos 0.70. O Nordeste também teve o SARIMA como melhor método, mas os seus resultados não superaram aqueles do caso univariado, embora com um resultado muito próximo. Apenas o Centro-Oeste mostrou o NNAR como melhor método de predição durante a validação cruzada no caso em que foi utilizado a *proxy* de subvenção, com 15,12 de RMSE e superando sutilmente o caso do NNAR univariado (16,83 de RMSE), com um *Skill* menor visto que o *baseline* não tem o seu desempenho alterado no caso em que é considerado a covariável.

Tabela 3. Resultados das medidas de acurácia das projeções por região e por modelo com covariável mensal estimada para o subsídio

	Sul		**Sudeste**		**Centro-Oeste**		**Nordeste**	
	RMSE	**Skill**	**RMSE**	**Skill**	**RMSE**	**Skill**	**RMSE**	**Skill**
SARIMA	13,90	0,68	9,90	0,73	19,13	1,17	1,74	0,49
NNAR	24,57	1,19	13,12	0,97	15,12	0,93	3,32	0,94
ELM	15,69	0,76	9,79	0,73	15,29	0,94	3,34	0,95
Naive	20,58	1,00	13,50	1,00	16,31	1,00	3,51	1,00

Fonte: Resultado da pesquisa

Desta forma podemos selecionar a partir do RMSE os 4 métodos que mostraram melhor performance preditiva durante o VC, considerando casos univariado e com covariáveis: no caso da região Sul temos o SARIMA(0,0,0)(1,1,0) e a covariável com o menor RMSE; na região Sudeste o ELM com covariável também se mostrou o método com melhor acurácia; no Centro-Oeste o SARIMA(1,0,0)(1,1,2)[10] univariado foi o método com menor RMSE; e por fim, no Nordeste também tivemos SARIMA(13,0,0)(2,1,0)[11] univariado como o método selecionado. Os resultados dos coeficientes e testes estão dispostos na Tabela 4 abaixo. No caso dos modelos SARIMA estimados, todos os coeficientes dos componentes AR, MA e MA sazonal mostraram-se estatisticamente significantes, os testes Ljung-Box e Box-Pierce não rejeitaram a hipótese nula de erros independentes e identicamente distribuídos, o teste Shapiro-Wilk rejeitou a hipótese de não normalidade dos resíduos.

[10] Considerando apenas as defasagens 1 do AR, 1 do AR sazonal e 2 do MA sazonal.
[11] Considerando apenas as defasagens 1, 11 e 13 do AR, além do 1 e 2 do AR sazonal.

Tabela 4. Estimação dos parâmetros dos modelos SARIMA para o volume de prêmio do seguro agrícola mensal por região do Brasil entre 2006 e 2017 (erro padrão)

	Sul	Nordeste	Centro-Oeste
	SARIMA (0,0,0)(1,1,0)	SARIMA (13,0,0)(2,1,0)	SARIMA (1,0,0)(1,1,2)
Subvenção	0,576*** (0,201)	-	-
AR(1)	-	0,472*** (0,078)	0,496*** (0,077)
AR(11)	-	0,177** (0,090)	-
AR(13)	-	-0,075* (0,076)	-
SAR(1)	-	-0,612*** (0,095)	-0,567*** (0,098)
SAR(2)	-0,179** (0,087)	-0,321*** (0,091)	-
SMA(2)	-	-	-0,339*** (0,117)
Observações	132	132	132
Teste de Box-Pierce (autocorrelação dos resíduos)	0,1817	0,8662	0,1587
Teste Ljung-Box (lag 1,3,6,9)	0,14; 0,11; 0,16; 0,22	0,75; 0,54; 0,70; 0,79	0,54; 0,54; 0,45; 0,14
Teste Shapiro-Wilk (Normalidade)	0,0001275	$1,036 \times 10^{-9}$	$2,658 \times 10^{-5}$

Nota: (*) denota a rejeição de H0 à 10% de significância, (**) denota a rejeição de H0 à 5% de significância e (***) denota à rejeição de H0 à 1% de significância.
Fonte: Elaboração própria a partir dos dados da pesquisa

Por sua vez, a Figura 7 mostra que os modelos estimados apresentaram bom ajuste à séries de prêmios. Percebe-se, porém, que o modelo subestimou os picos nos meses de outubro de 2008, maio de 2014 e maio de 2016 no caso da região Sul; agosto de 2010 e quase todo o ano de 2013 para a região Sul; dezembro de 2013, agosto de 2014 e janeiro e outubro de 2017 na região Centro-Oeste; setembro de 2009, setembro de 2010, novembro de 2011 e agosto de 2015 na região Nordeste. Ainda assim cabe notar que os resultados também

seguiram a tendência de elevação do volume de prêmios ao longo do tempo, conforme encontrado em Adami e Ozaki (2012).

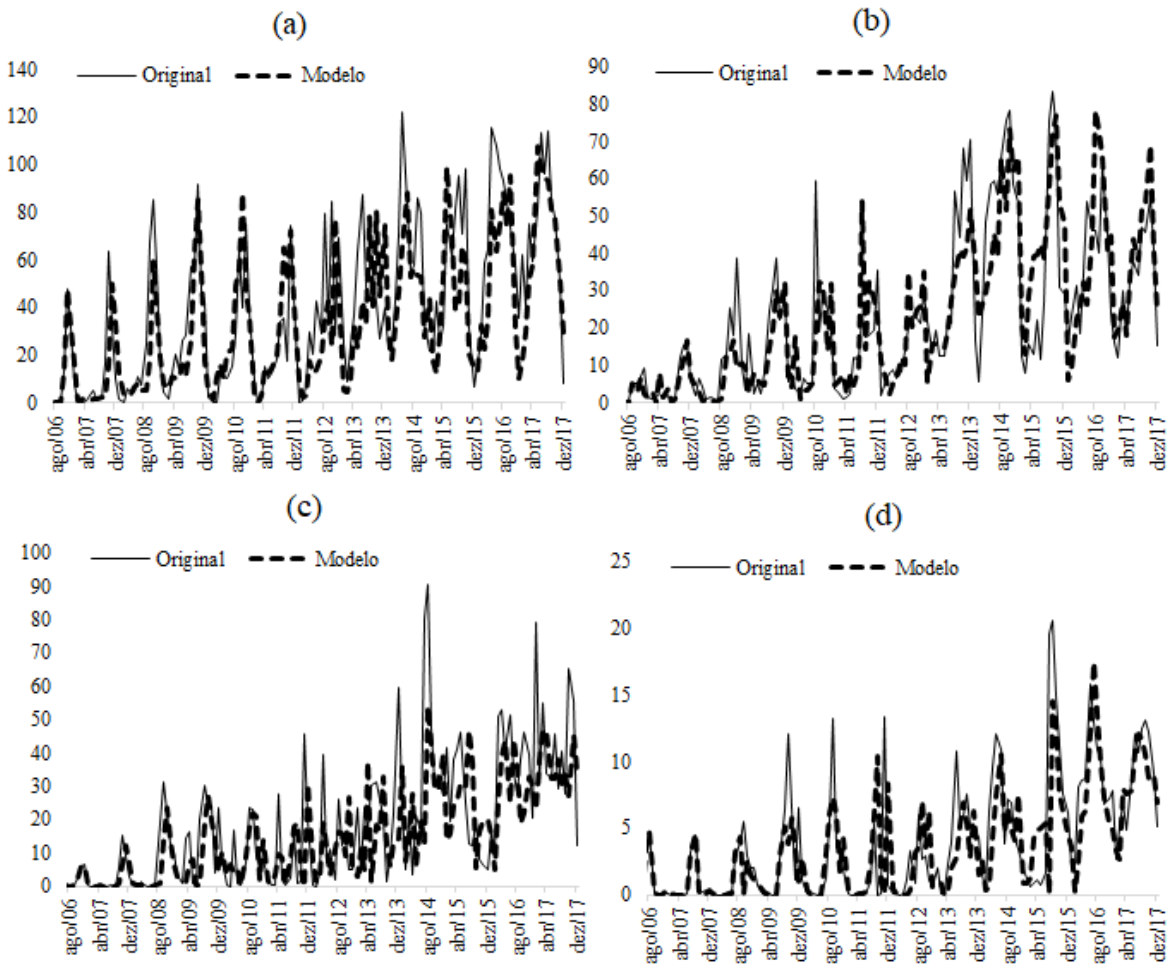

Figura 7. (a) Ajuste do modelo SARIMA(0,0,0)(1,1,0) à série de prêmio mensal do seguro agrícola da região Sul; (b) Ajuste do algoritmo ELM à série de prêmio mensal do seguro agrícola da região Sudeste; (c) Ajuste do modelo SARIMA(1,0,0)(1,1,2) à série de prêmio mensal do seguro agrícola da região Centro-Oeste; (d) Ajuste do modelo SARIMA(13,0,0)(2,1,0) à série de prêmio mensal do seguro agrícola da região Nordeste.

Fonte: Resultado da pesquisa.

2.4.1. Utilização prática dos métodos escolhidos

O resultado encontrado vai em direção ao registrado em Adami e Ozaki (2012), que mostra o método SARIMA como melhor para elaborar a predição da série de volume de prêmio nacional, sendo o caso para 3 das 4 séries abordadas. A diferença no presente trabalho é que dois métodos selecionados contam com a covariável estimada, além do fato do método ELM acompanhar de forma muito próxima a performance do SARIMA da respectiva série

projetada, sendo que apenas no caso do Nordeste que o RMSE ficou mais distante (próximo do dobro).

Após a seleção, foi elaborado as predições dos volumes de prêmio de janeiro de 2018 a dezembro de 2018 com o melhor método de cada região, conforme a Tabela 5.

Tabela 5. Comparação entre valor original e predito das séries de prêmio por região do país para os meses de 2018, em R$ milhões, de acordo com cada modelo selecionado

	Sul		Sudeste		Centro-Oeste		Nordeste	
	Projeção	Efetivo	Projeção	Efetivo	Projeção	Efetivo	Projeção	Efetivo
jan/18	52,3	37,0	15,4	24,0	43,6	51,9	2,3	5,3
fev/18	36,2	32,2	15,1	13,6	17,1	19,2	3,2	2,9
mar/18	70,4	70,5	32,2	32,9	43,3	43,4	4,9	3,9
abr/18	58,2	88,1	18,7	37,9	35,1	46,4	4,6	6,9
mai/18	85,3	84,7	33,2	43,5	33,1	42,5	6,6	7,8
jun/18	110,1	112,7	42,1	35,6	38,1	51,3	9,7	14,2
jul/18	94,0	114,0	38,3	37,9	30,2	38,3	14,0	13,2
ago/18	107,8	118,6	47,7	48,9	35,0	46,0	15,1	15,2
set/18	81,8	72,9	43,6	55,4	28,0	51,9	13,9	13,2
out/18	74,6	54,7	56,3	62,5	43,4	50,4	11,1	10,5
nov/18	49,1	32,1	41,8	18,1	38,6	48,8	8,0	7,6
dez/18	11,0	4,8	17,6	20,0	16,5	14,3	6,0	6,1

Fonte: Resultado da pesquisa

A predição dos meses de 2018 da região Sul resultaram em um RMSE de 14,34, um pouco superior ao resultado durante a validação cruzada (13,90), com subestimação em abril, julho e agosto. Por sua vez, o volume de prêmio projetado para a região Sudeste resultou em um RMSE de 10,56, também superando um pouco aquele visto durante o VC, de 9,79, subestimando os resultados efetivos verificados em abril, maio e setembro. Com RMSE de 10,72 e VC de 15,10, a projeção do prêmio para a região Centro-Oeste foi subestimada em boa parte do ano de 2018, acertando o movimento, mas errando o nível. Por fim, a predição da região Nordeste para esse período gerou um RMSE de 1,79, muito próximo dos 1,65 obtidos durante o VC, com maior desvio verificado em junho.

Com os métodos apresentados, as seguradoras podem estimar o volume de prêmio esperado para cada região do país e, ao conhecer uma estimativa mais robusta para a demanda daquela região, construir um plano estratégico de negócios mais assertivo. No caso em que é necessário a variável de subvenção, cabe destacar que não só há as informações disponíveis no Plano Trienal do Seguro Rural, como também o volume divulgado no começo de todo ano no Diário Oficial da União ou no Lei Orçamentária Anual (LOA) aprovada, podendo as

previsões serem atualizadas ao longo do ano tão logo as novas informações mensais fiquem disponíveis ou eventuais mudanças no volume de subvenção que o governo divulgar. A volatilidade do volume disponibilizado para subvenção, bem como a sua mudança ao longo do ano podem ser limitantes para o desempenho dos métodos que dependem dessa covariável. É possível que a performance preditiva melhore com o acúmulo de informações dos próximos anos e se o programa de subvenção mostrar uma maior estabilidade no volume de recursos disponibilizado anualmente. Os próximos estudos também podem avançar em formas de incorporar informações envolvendo variáveis do mercado de crédito regional como forma de auxiliar na projeção, ainda que a obtenção de dados futuros para essas variáveis também sejam particularmente difíceis de se encontrar[12].

Em suma, destaca-se que os métodos elencados são os mais indicados para as previsões dos volumes das prêmios regiões, visto que mostram o menor RMSE e, portanto, previsões mais acuradas. As decisões econômicas-financeiras como de gestão de risco com base nesses métodos serão, em teoria, as melhores possíveis dentro do conjunto de métodos estudados no trabalho. Seguradoras que trabalham no Brasil todo podem se beneficiar com projeções e mais informações na hora do planejamento anual, sobretudo na hora de diversificar os riscos, ao saber o comportamento esperado para cada região.

2.5. Conclusão

O mercado de seguro rural, sobretudo agrícola, tem se desenvolvido desde 2006, com o suporte do programa de subvenção. Ainda assim, o ritmo e a abrangência do programa ainda são pequenos, perto da dimensão do agronegócio brasileiro, reforçando a necessidade de mais estudos na área.

Nesse sentido, o presente estudo abordou modelos de projeção para o prêmio do seguro agrícola por região, com o objetivo de aumentar o volume de informação para o mercado, sobretudo para as seguradoras. Os resultados mostram que, conforme encontrado em Adami e Ozaki (2012) para o caso nacional, o método SARIMA com e sem covariáveis mostrou-se o mais eficaz na elaboração de projeções de tais séries, em termos de maior acurácia. Com a hipótese de melhoria das predições, foram derivadas variáveis mensais do volume de subvenção disponibilizada pelo Ministério da Agricultura, Pecuária e Abastecimento (MAPA), utilizando o método proposto em Sax e Steiner (2013) a partir da

[12] Uma possível opção é trabalhar com médias móveis de *h* períodos das variáveis de crédito, sendo *h* o período que se deseja prever.

variável de frequência anual, mas apenas em dois casos essas variáveis se mostram eficazes na redução do erro de predição.

Selecionado a partir do processo de validação cruzada, a região Sul do país mostrou no SARIMA(0,0,0)(1,1,0) o menor RMSE dentre os métodos testados, além de utilizar a variável estimada para a subvenção mensal. A região Sul teve o *Extreme Learning Machine* (ELM) como melhor método de predição, também com a sua respectiva variável de subvenção. Por sua vez, para o caso de Centro-Oeste e Nordeste o modelo univariado mostrou a melhor acurácia, mostrando a configuração do SARIMA(1,0,0)(1,1,2) e SARIMA(13,0,0)(2,1,0) como as melhores opções, respectivamente.

Por fim, os métodos permitem a utilização por parte das seguradoras na hora de preparar o planejamento anual de suas operações no ramo agrícola, sobretudo com as informações do programa de subvenção divulgadas no início de cada ano, permitindo uma informação adicional da demanda de cada região do país.

Referências

ADAMI, A.; OZAKI, V. **Modelagem estatística dos prêmios do seguro rural**. Revista de Política Agrícola, ano 21, n. 1, p. 60-75, 2012.

AHSAN, S.; ALI, A.; KURIAN, N. **Toward a Theory of Agricultural Insurancel**. American Journal of Agricultural Economics, Volume 64, p. 510–529, 1982.

AKAIKE, H. **Information theory and the maximum likelihood principle**. 2nd International

Symposium on Information Theory, 1973

ALEXANDER, G. D. **Simple kNN-Method for Times Series Prediction**. Neural-Forecasting-Competition.Com, p. 2, 2007.

ARROW, K. **The Economics of Moral Hazard: Further Comment**. The American Economic Review, Vol. 58, No. 3, p. 537-539, 1968.

BERGMEIR, C.; HYNDMAN, R.; KOO, B. **A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction.** Computational Statistics E Data Analysis, Vol. 120, p. 70-83, 2018.

BUAINAIN, A; VIEIRA, P; CURY, J. **Gestão do Risco e Seguro na Agricultura Brasileira**.

Funenseg, 2014., p. 16, 2014.

BOX, G.; JENKINGS, G. **Time Series Analysis: forecasting and control**. Revised edition, San Francisco: Holden-Day. 1976.

BOX, G.; PIERCE, D. **Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models**. Journal of the American Statistical Association, Vol. 65, p. 1509-1526, 1970

DE JONG, P. **The diffuse Kalman filter**. The Annals of Statistics 19, P. 1073–1083, 1991.

DE LIVERA, A. M.; HYNDMAN, R. J.; SNYDER, R. **Forecasting time series with complex seasonal patterns using exponential smoothing**. Journal of the American Statistical Association, Vol. 106, n. 496, p. 1513–1527, 2011.

DENTON, F. **Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization**. Journal of the American Statistical Association, Vol. 66, n. 333, 1971

DICKEY, D.; FULLER, W. **Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root**. Econometrica, Vol. 49, n. 4, p. 1057-1072, 1981.

DIETZ, D. **Autoregressive Neural Network Processes: Univariate, Multivariate and Cointegrated Models with Application to the German Automobile Industry**. Tese de doutorado disponível em: https://opus4.kobv.de/opus4−uni−passau/files/142/DietzSebastian.pdf. Acesso em: 12 nov. 2018.

CUSINATO, R. T. **Teoria de decisão sob incerteza e a hipótese da utilidade esperada: conceitos analíticos e paradoxos**. Dissertação disponível em: https://lume.ufrgs.br/handle/10183/1961 Acesso em: 12 jan. 2019.

FLORES, J. H.; WERNER, L. **Aplicação de redes neurais artificiais à previsão de vendas de máquinas agrícolas: um estudo de caso junto a AGCO do Brasil Ltda**. Trabalho de conclusão de curso, disponível em: https://lume.ufrgs.br/handle/10183/135441. Acesso em: 02 jan. 2019.

GARDNER, E.; MCKENZIE, E. **Forecasting Trends in Time Series**. Management Science, p.1237–1246, 1985.

GEMAN, S.; DOURSAT, R.; BIENENSTOCK, E. **Neural Networks and the Bias/Variance Dilemma. Neural Computation**, v. 4, p. 1–58, 1992.

GHEYAS, I. A.; SMITH, L. S. A Neural Network Approach to Time Series Forecasting. **Proceedings of the World Congress on Engineering,** v. 2, n. 1, p. 1–5, 2009.

GORR, W. L. **Research prospective on neural network forecasting**. International Journal of Forecasting, p. 1-4, 1994.

GUIMARAES, M. F.; NOGUEIRA, J. M. **A experiência norte-americana com o seguro agrícola: lições ao Brasil?** Revista de Economia e Sociologia Rural, vol.47 n.1, 2009.

HIBON, M.; EVGENIOU, T. **To combine or not to combine: selecting among forecasts and their combinations**. International Journal of Forecasting, Vol. 21, p. 15 – 24, 2005.

HUANG, G.; BABRI, H. A. **Upper bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions**. IEEE Transactions on Neural Networks, vol. 9, p. 224-229, 1998.

HUANG, C.; HUANG, C.; WANG, M. **A Particle Swarm Optimization to Identifying the ARMAX Model for Short-Term Load Forecasting**. IEEE Transactions on Power System, vol. 20, n. 2, 2005.

HYNDMAN, R. J.; KOEHLER, A. B. **Another look at measures of forecast accuracy**.International Journal of Forecasting, v. 22, n. 4, p. 679–688, 2006.

HYNDMAN, R. J.; KOEHLER, A. B.; SNYDER, R. D. **Forecasting with Exponential Smoothing: The State Space Approach**. Springer, 2008.

HYNDMAN, R. J.; KHANDAKAR, Y. **Automatic time series forecasting: The forecast package for R**. Journal of Statistical Software, v. 27, n. 3, 2008.

JENSEN, M. C.; MECKLING, W. H. **Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure**. Journal of Financial Economics, Vol. 3, N. 4, 1976

KITAGAWA, G. **State space modeling of time series:** an informational approach. Springer, 1994.

LI, G.; MA, X.; YANG, H. **Hybrid Model for Monthly Precipitation Time Series Forecasting Based on Variational Mode Decomposition with Extreme Learning Machine**. Information, Vol. 9, No. 177, 2018.

LJUNG, G.; BOX, G. **On a Measure of Lack of Fit in Time Series Models**. Biometrika, Vo. 65, n. 2, p. 297-303, 1978.

MACEDO, L. O.; PACHECO, A. B.; SANTO, E. **A evolução do Programa de Subvenção do Prêmio do Seguro Rural: uma avaliação do período 2006-10**. Revistas Eletrônicas FEE, 40, n. 4, p. 99-112, 2013.

MEDEIROS, E. A. **Avaliação da implementação do programa de subvenção do prêmio do Seguro Rural**. Revista de Economia e Sociologia Rural, vol.51 n.2, 2013.

MINISTERIO DA AGRICULTURA, PECUARIA E ABASTECIMENTO. **Seguro Rural. Relatório das indenizações pagas - 2006 a 2015**. 2018. Disponível em: http://www.agricultura.gov.br/assuntos/riscos-seguro/seguro-rural/relatorios-estatisticos. Acesso em: 10 nov. 2018.

MINISTERIO DA AGRICULTURA, PECUARIA E ABASTECIMENTO. **Relatório Seguro Rural. Resultado Geral 2018**. 2019. Disponível em: http://www.agricultura.gov.br/assuntos/riscosseguro/seguro-rural/relatorios-estatisticos. Acesso em: 5 jan. 2019.

MIRANDA, M; GLAUBER, J. **Systemic Risk, Reinsurance, and the Failure of Crop Insurance Markets**. American Journal of Agricultural Economics, Vol. 79, p. 206–215, 1997.

OLIVEIRA, A. C.; SOUZA, A. A.; LACERDA, W. S.; GONC¸ALVES, L.R. **Aplicação de redes neurais artificiais na previsão da produção de álcool**. Revista Ciência e Agrotecnologia, Vol.34, n.2, 2010.

OZAKI, V. **Métodos atuariais aplicados à determinação da taxa de prêmio de contratos de seguro agrícola:** um estudo de caso. Tese de doutorado disponível em: http://www.teses.usp.br/teses/disponiveis/11/11132/tde-11072005-160540/pt-br.php

OZAKI, V. **Em busca de um novo paradigma para o seguro rural no Brasil**. Revista de Economia e Sociologia Rural, vol.46 n.1, 2008.

OZAKI, V.; SHIROTA, R. **A experiência do seguro agrícola nos EUA: evolução e performance**. Revista Brasileira de Risco e Seguro, Vol.1, n.2, p 68-87, 2006.

OZAKI, V; **Em busca de um novo paradigma para o seguro rural no Brasil**. Revista de Economia e Sociologia Rural, vol.46 n.1, 2008.

PAPPENBERGER, F.; RAMOS, M. H.; CLOKE, H. L.; WETTERHALL, F.; ALFIERI, L.; BOGNER, K.; MUELLER, A. **How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction**. Journal of Hydrology, Vol. 522, p. 697–713, 2015.

PEARSON, K. **Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia**. Philosophical Transactions of the Royal Society of London, p. 253–318, 1896.

ROTHSCHILD, M.; STIGLITZ, J. **Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information**. The Quarterly Journal of Economics, Vol. 90, No. 4, pp. 629-649, 1976.

SANTOS, Gesmar; SILVA, Fabiano. **Dez anos do Programa de Subvenção ao Prêmio de Seguro Agrícola: Proposta de índice técnico para análise do gasto público e ampliação do seguro**. IPEA, Texto para discussão 2290, 2017.

SAX, C.; STEINER, P. **Temporal Disaggregation of Time Series**. The R Journal, Vol. 5/2, 2013.

SHOWERS, V.; SHOTICK, A. **The Effects of Household Characteristics on Demand for Insurance: A Tobit Analysis**. The Journal of Risk and Insurance, Vol. 61, No. 3, p. 492-502, 1994.

SKEES, J.; HAZELL, P.; MIRANDA, M. **New approaches to crop yield insurance in developing countries**. EPTD Discussion Paper, n. 55, 1999

SINGH, R.; BALASUNDARAM, S. **Application of extreme learning machine method for time series analysis**. International Journal of Computer and Information Engineering, p.256–262, 2007.

SOARES, F.C.; ROBAIANA, A.D.; PEITER, M.X.; RUSSI, J.L. **Predição da produtividade da cultura do milho utilizando rede neural artificial**. Ciência Rural, V.45, n.11, p.1987-1993, 2015.

SOUZA, L.R.; SMITH, J. **Effects of temporal aggregation on estimates and forecasts of fractionally integrated processes: a Monte-Carlo study**. International Journal of Forecasting, Vol. 20, p.487 – 502, 2004.

SUPERINTENDENCIA DE SEGUROS PRIVADOS. **Dados de prêmio do seguro agrícola por Unidade Federativa**. Disponível em: http://www2.susep.gov.br/menuestatistica/SES/premiosdesegurosporuf.aspx?id=4. Acesso em: 10 nov. 2018.

SUPERINTENDENCIA DE SEGUROS PRIVADOS. **7º Relatório de análise e acompanhamento dos mercados supervisionados**. Disponível em http://www.susep.gov.br/menuestatistica/SES/RelatAcompMercado2019.pdf/view. Acesso em: 15 mar. 2019.

TAYLOR, J. **Short-term electricity demand forecasting using double seasonal exponential smoothing.** Journal of the Operational Research Society 54, p.799–805, 2003.

TEALAB, A.; HEFNY, H.; BADR, A. **Forecasting of nonlinear time series using ANN**. Future Computing and Informatics Journal, v. 2, n. 1, p. 39–47, 2017.

TERASVIRTA, T.; LIN, C.; GRANGER, C. **Power of the Neural Network Linearity Test**. Journal of Time Series Analysis, v. 14, n. 2, p. 209–220, 1993.

THIELBAR, M.; DICKEY, D. **Neural Networks for Time Series Forecasting**. Future Computing and Informatics Journal, v. 2, n. 1, p. 39–47, 2017.

WHITE, H. **An additional hidden unit test for neglected nonlinearity in multilayer feedforward network**. International Joint Conference on Neural Networks, p. 451–455 vol.2, 1989.

YANG, H.; HUANG, C.; HUANG. C. **Identification of ARMAX Model for Short Term Load Forecasting**: an evolutionary programming approach. IEEE Transactions on Power Systems, Vol. 11, 1996.

ZHANG, P. **Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model**. Neurocomputing 50, P. 159-175. 2003.

3. PREDIÇÃO DE SINISTROS AGRÍCOLAS: UMA ABORDAGEM COMPARATIVA UTILIZANDO APRENDIZADO DE MÁQUINA

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 (PSR). O presente estudo testou o desempenho de métodos para as seguradoras anteciparem a possibilidade de ocorrência de sinistro, elaborando previsões dos dados envolvendo sinistro no mercado agrícola entre 2006 e 2017. Foram testados os algoritmos *Random Forest*, *Support Vector Machine* e *k-Nearest Neighbours*, sendo que o primeiro método mostrou a melhor performance preditiva de sinistros quando avaliado por métricas como Acurácia, Precisão, Taxas de Verdadeiro Positivo e Negativo, além da Correlação de Matthews.

Palavras-chave: Seguro agrícola, Sinistro, Previsão, *Machine Learning*

Abstract

The agricultural insurance has gained more attention in Brazil since the beginning of the last decade, with the implementation of the Crop Insurance Program. The present study tested the performance of methods for insurers to anticipate the occurrence or non-occurrence of claims by predicting the data involving claims in the agricultural market between 2006 and 2017. We tested the Random Forest, Support Vector and k-Nearest Neighbours. The first method showed the best predictive performance of claims when evaluated by metrics such as Accuracy, Precision, True Positive and Negative Rates, and Matthews Correlation Coefficient.

Keywords: Agricultural insurance, Insurance claim, Forecast, Machine learning

JEL codes: G22, Q13, Q18

3.1. 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 órgão regulador do mercado de seguros, a Superintendência de Seguros Privados (SUSEP), é feita na regulação preventiva por meio das provisões e reservas técnicas. Estas são compostas pelo montante de recursos que as seguradoras precisam manter como garantia para eventos e obrigações futuras, entrando como passivo das seguradoras em linha com o estabelecido pelo Conselho Nacional de Seguros Privados (RODRIGUES, MARTINS, 2010). A subestimação da possibilidade 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 ou ser excessiva.

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 novos métodos de predição para aumentar a informação no mercado de seguro agrícola, dado que uma taxa maior de acerto na predição dos sinistros auxilia em um planejamento financeiro e no estabelecimento de novos contratos de seguro.

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, segundo os dados do MAPA (2018). 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, sobretudo quando se busca métodos para a predição de ocorrências de sinistros agrícolas. O trabalho buscou uma modelagem que aponte se haverá sinistro ou não, de forma binária, utilizando algoritmos já empregados em outras áreas do conhecimento, como diagnóstico médico e detecção de fraudes (CASTRO; BRAGA, 2011).

Esse artigo inicia com uma breve revisão da literatura, que aborda métodos para tratar do problema da predição dos sinistros, mas que em sua maioria trata exclusivamente do ramo automotivo. A intenção foi buscar indicativos dos métodos que os pesquisadores utilizaram para trabalhar com a variável sinistro. Na sequência é dado um panorama geral sobre os eventos de sinistro no mercado agrícola brasileiro, seguido da apresentação do método empregado, bem como os dados. Por fim, o artigo é finalizado com a discussão dos resultados encontrados, avaliando o desempenho das predições e a conclusão. Os modelos do

trabalho foram baseados na experiência de Pijl (2017) para o mercado de seguro automotivo, com a adição do algoritmo k-NN.

3.2. 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, sendo relevante citar a abordagem e a experiência de alguns deles. Ye *et al*. (2018) trabalharam 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*. (2016) utilizaram 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 (números de chamados). Os autores utilizaram um Modelos Lineares Generalizados Mistos (GLMM) bivariados para as variáveis de forma a capturar suas relações de dependência para o mercado de seguro automotivo alemão, mostrando ganho de performance preditiva para estimar a perda total e mitigar o risco por parte da seguradora.

Na literatura nacional, também há predominância no ramo automotivo, com Zaniboni e 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 condutor e do carro, estado de residência, número de expostos e importância segurada. Freitas (2009), por sua vez, utilizou modelo Logit para estimar a probabilidade da ocorrência de um sinistro nos estados brasileiros também no mercado automotivo, encontrando que estados com riscos diferentes apresentam 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 que há influência sobre o número de sinistros ao considerar um modelo que trabalha com a Binomial Negativa.

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 pouco 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.

É importante notar que trabalhos envolvendo predição com dados binários 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 e Torelli (2010), 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 considerada 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 (ELRAHMAN, ABRAHAM, 2013), em que o erro da predição apresenta um impacto negativo elevado. Weiss *et al*. (2007) discutiram 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 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). Por sua vez, o presente estudo abordou o problema dos dados desbalanceados a partir do algoritmo proposto por Mernardi e Torelli (2010), denominado ROSE (acrônimo de *Random Over Sampling Examples*).

Para tratar especialmente do problema do sinistro do segmento de seguro agrícola, foram necessárias 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 mostram 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) avaliaram estados do Sul, Sudeste e Centro-Oeste, encontrando 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 previsões baseadas em eventos *El Niño* e *La Niña*, sendo sempre de fontes internacionais (LIU *et al.*, 2007; 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 e 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 divergência entre níveis de risco no mercado das seguradoras, tanto pela composição da carteira, quanto pelas variáveis financeiras. Desta 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). Há um conjunto de outras variáveis que impactam na produção agrícola, sobretudo ligadas ao clima, mas essas são de difícil previsão com larga antecedência e os modelos testados têm por objetivo a predição o mais antecipadamente possível e com variáveis mais fáceis de estimar.

O mercado de seguro agrícola no Brasil vem crescendo desde o início do Programa de Subvenção ao Prêmio do Seguro Rural (PSR), no fim de 2005. O estudo aborda o período de 2006 até 2017, intervalo em que estão disponíveis os dados do MAPA para parte dos sinistros no mercado de seguro agrícola que recebeu subvenção do PSR. O Atlas do Seguro Rural, elaborado pelo Ministério da Agricultura, Pecuária e do Abastecimento (MAPA), mostra que nesse período foram pagos cerca de R$ 3,53 bilhões em indenizações referentes a chamados de sinistros, de um total de R$ 103,86 bilhões de importância acumulada ao longo de todos esses anos, o que rendeu um prêmio acumulado de R$ 6,75 bilhões para esse mercado. Por outro lado, se considerarmos os dados da SUSEP, que envolvem todo o mercado de seguro agrícola (não apenas a parcela subvencionada), o total de prêmio acumulado de 2006 a 2017 chega a R$ 10,09 bilhões[13]. Avaliando os 70.538 contratos de seguro, cerca de 23% apresentou algum sinistro no período. A Tabela 6 abaixo 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.

[13] A estimativa foi feita com base nos ramos 1101, 1102, 1161 do Sistema de Estatísticas da SUSEP (SES).

Tabela 6. Causas de sinistro do seguro agrícola brasileiro de 2006 a 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

Fonte: Elaboração própria com dados do MAPA (2018).

Há grande concentração na distribuição das causas de sinistros por alguns motivos, dentre eles devemos destacar a própria concentração de mercado. Conforme mostra MAPA (2018), a região Sul representou em média 70% do total de produtores rurais beneficiados pelo programa de subsídio governamental, 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 (MAPA, 2018). Avaliando os dados utilizado nesse estudo, 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.

3.3. Métodos

O trabalho utilizou o software R, usando os pacotes *randomForest*, *pROC*, *ROSE*, *e1071* e *class*.

3.3.1. Random Forest

O método *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 de decisão em árvore apresentam baixo viés, mas elevada variância, o que acaba produzindo um sobreajuste (*overfitting*[14]). 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 $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_n$, um 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 $\sigma_n^2 > \sigma^2/_n$).

Para o treinamento, o algoritmo RF cria um conjunto B de árvores pelo método CART, cada uma treinada em uma amostra n, sendo que essa amostra foi formada por *bootstrap* dos dados de um conjunto N separado para treinamento, com reposição. O conjunto de treinamento N e esse subconjunto n possuem o mesmo tamanho. O processo se difere do método conhecido como *Bagging* proposto em Breiman (1994) ao usar apenas um número m das p variáveis preditoras em cada divisão da árvore, em que $m < p$ e $m \approx \frac{p}{3}$ (JAMES *et al.*, 2013), escolhidas de forma aleatórias, com reposição, criando modelos independentes e árvores descorrelacionadas.

De forma geral, o algoritmo pode ser descrito como: Dado um conjunto de teste $X = x_1, \dots, x_n$ que se deseja usar para prever $Y = y_1, \dots, y_n$ são retirada as amostras aleatórias n cerca de B vezes por *bagging* com reposição, formando os conjuntos X_b e Y_b, em que $b = 1, \dots B$, sendo B o número de árvores do processo. Os erros são calculados a cada etapa recursiva, num método chamado de *out-of-bag error*, conforme Hastie *et al.* (2009), elaborando a projeção e medindo a acurácia do conjunto de dados de treino que não foram amostrados naquela oportunidade. A avaliação da acurácia do método é feita pelo Erro Quadrático Médio (MSE, na sigla em inglês) definido, no caso da predição, como

$$\mathrm{MSE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2. \tag{3.1}$$

sendo $\hat{y}_i$ a projeção da observação y_i. O algoritmo treina regressões em árvore T_b, escolhendo a melhor variável das m variáveis e dividindo um nó em outros dois, repetindo recursivamente em cada nível das árvores até chegar a ponto em que o CART detecta que novas divisões não geram nenhum ganho adicional ou então que o algoritmo atinja o limite de

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

alguma regras de parada pré-definidas[15], produzindo $T_b(x)$ que somado as demais árvores formam uma floresta ${T_b}^B$ com B árvores. A predição de novos pontos é obtida por meio de $\hat{f}^B = \frac{1}{B}\sum_{b=1}^{B} T_b x$, ou seja, por meio da média dos resultados finais das árvores.

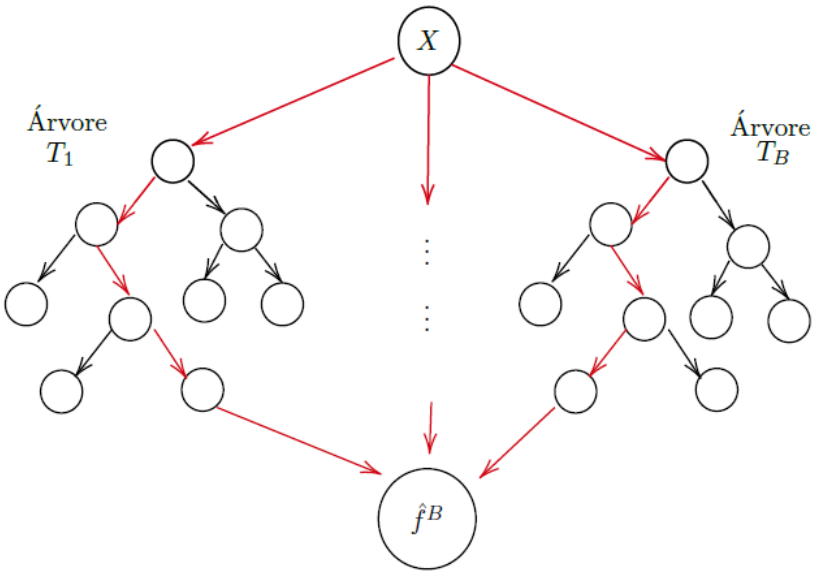

Figura 8. Representação do modelo *Random Forest*.
Fonte: Elaboração própria.

3.3.2. Support Vector Machine (SVM)

O Support Vector Machine é um método que pode ser usado para predição, buscando num 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 (1979) e depois Cortes e Vapnik (1995) fizeram modificações para um caso mais geral, que abrange a situação em que não há formas de separação linear.

Conforme Gunn (1998) e Smola e Scholkopf (2004), dado um conjunto de treinamento $X = x_1, \dots, x_n$ para prever $Y = y_1, \dots, y_n$, sendo $(X, Y) \subset \mathrm{S} \times \mathbb{R}$ e S um espaço de entrada, deseja-se encontrar $f(x)$ que resulte no menor desvio ε em relação a um conjunto de variável resposta de interesse. Ao considerar inicialmente a função linear $\mathrm{f(x)} = \langle \mathrm{w}, \mathrm{x} \rangle + \mathrm{b}$

[15] No geral, há uma escolha que envolve custo-benefício entre tamanho da árvore e qualidade do ajuste, e as regras de penalidades para "podar" a árvore buscam maximizar essa relação, conforme mostra Hastie *et al.* (2009).

com $w \in S$ e $b \in R$, temos que o problema se torna uma otimização, sendo necessário a minimização da norma

$$\begin{gathered} min\ {}^{1}/_{2}\,\|w\|^2 \\ \text{sujeito a } y_n - \langle w, x_n \rangle - b \leq \varepsilon \\ \langle w, x_n \rangle + b - y_n \leq \varepsilon. \end{gathered} \tag{3.2}$$

Em um grande número de situações essa otimização não é possível (não há convergência) dado a disposição dos dados, levando Cortes e Vapnik (1995) a introduzirem termos ξ_n e ξ_n^* para afrouxar a restrição, resultando em um problema que pode ser caracterizado de forma primal como

$$\begin{gathered} \min \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}(\xi_n + \xi_n^*) \\ \text{sujeito a } y_n - \langle w, x_n \rangle - b \leq \varepsilon + \xi_n \\ \langle w, x_n \rangle + b - y_n \leq \varepsilon + \xi_n^* \\ \xi_n, \xi_n^* \geq 0 \end{gathered} \tag{3.3}$$

em que C é um parâmetro não negativo que permite esse relaxamento no erro, controlando uma penalidade imposta às observações fora da margem ε. É importante ressaltar que um C pequeno resulta em margens estreitas e pouca violação, mas possivelmente gerando *overfitting* pelo baixo viés e alta variância, ao passo que um C elevado eleva o viés do método. Já o problema na forma dual permitirá a introdução de funções não lineares posteriormente. Considere a construção do lagrangeano a partir da função primal, levando ao problema dual dado por

$$\begin{aligned} L \coloneqq \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{n=1}^{N}(\xi_n + \xi_n^*) - \sum_{n=1}^{N}(\eta_n\xi_n + \eta_n^*\xi_n^*) \\ - \sum_{n=1}^{N}\alpha_n(\varepsilon + \xi_n - y_n + \langle w, x_n \rangle + b) \\ - \sum_{n=1}^{N}\alpha_n^*(\varepsilon + \xi_n^* + y_n - \langle w, x_n \rangle - b) \end{aligned} \tag{3.4}$$

em que η_n, η_n^*, α_n, α_n^* são os multiplicadores de Lagrange não negativos. Após as derivações e substituições[16], temos que o processo dual pode ser escrito como

[16] A derivação está disponível em Smola e Scholkopf (2004).

$$\min \frac{1}{2}\sum_{n,i=1}^{N}(\alpha_n-\alpha_n^*)(\alpha_i-\alpha_i^*)\langle x_n,x_i\rangle-\varepsilon\sum_{n=1}^{N}(\alpha_n-\alpha_n^*)$$
$$+\sum_{n=1}^{N}y_n(\alpha_n-\alpha_n^*) \tag{3.5}$$
$$\text{sujeito a } \sum_{n=1}^{N}(\alpha_n-\alpha_n^*)=0$$
$$\alpha_n,\alpha_n^*\in[0,C]$$

o que nos permite obter

$$\mathrm{w}=\sum_{\mathrm{n}=1}^{\mathrm{N}}(\alpha_\mathrm{n}-\alpha_\mathrm{n}^*)\mathrm{x_n} \tag{3.6}$$

$$\mathrm{f(x)}=\sum_{\mathrm{n}=1}^{\mathrm{N}}(\alpha_\mathrm{n}-\alpha_\mathrm{n}^*)\langle \mathrm{x_n},\mathrm{x}\rangle+\mathrm{b} \tag{3.7}$$

sendo essa a função desejada para a regressão SVM no caso linear. A maior parte dos problemas não são facilmente enquadrados nos casos lineares, mas a formulação anterior permite uma extensão para casos não lineares. A diferença reside em substituir o produto escalar dado por $\langle x_n,x_i\rangle$ por uma função *kernel*, $K(x_n,x_i)=\langle\varphi(x_n),\varphi(x_i)\rangle$, que não é linear, em que $\varphi(x)$ é uma transformação que mapeia x para a um espaço de dimensão maior. O teorema de Cover nos garante que é possível transformar um espaço de entradas com padrão não linear em um novo espaço em que os padrões são lineares.

O *kernel* permite utilizar funções mais flexíveis, como as polinomiais, gausianas e radiais, ajudando a estabelecer a similaridade entre duas observações[17]. No geral, os algoritmos já programados encontram a melhor função para o problema de forma autônoma, e o problema dual para o caso não linear será dado pelo lagrangeano:

$$\min \frac{1}{2}\sum_{n,i=1}^{N}(\alpha_n-\alpha_n^*)(\alpha_i-\alpha_i^*)\varphi(x_n,x_i)-\varepsilon\sum_{n=1}^{N}(\alpha_n-\alpha_n^*)$$
$$+\sum_{n=1}^{N}y_n(\alpha_n-\alpha_n^*) \tag{3.8}$$
$$\text{sujeito a } \sum_{n=1}^{N}(\alpha_n-\alpha_n^*)=0$$
$$\alpha_n,\alpha_n^*\in[0,C]$$

resultando em

$$\mathrm{w}=\sum_{\mathrm{n}=1}^{\mathrm{N}}(\alpha_\mathrm{n}-\alpha_\mathrm{n}^*)\varphi(x_n) \tag{3.9}$$

$$f(x)=\sum_{\mathrm{n}=1}^{\mathrm{N}}(\alpha_n-\alpha_n^*)K(x_n,x_i)+b \tag{3.10}$$

[17] Essa relação tem como base o teorema de Mercer.

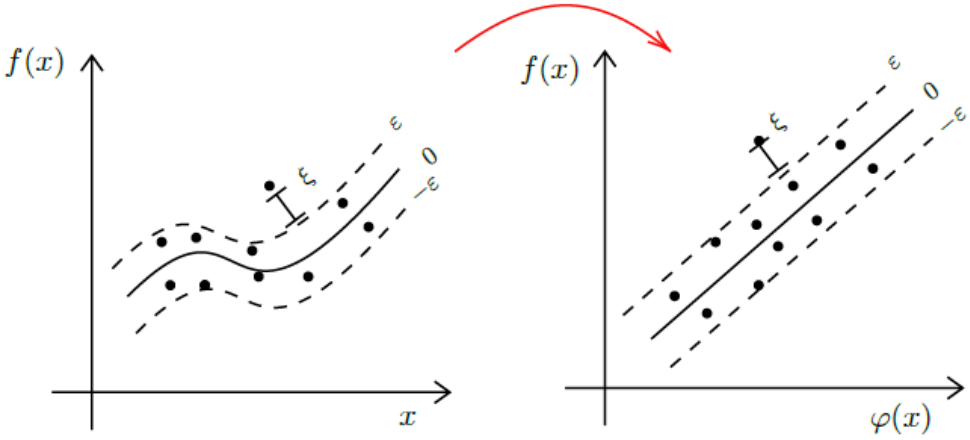

Figura 9. Representação do modelo SVM não linear.
Fonte: Elaboração própria.

Em suma, o método trabalha com uma fronteira de classes não lineares no espaço de entrada de dimensão m utilizando uma função *kernel* $\varphi: \mathbb{R}^m \rightarrow \mathbb{R}^q$ e mapeando o problema num espaço maior de dimensão q em que é possível tratá-lo de forma linear.

3.3.3. k-Nearest Neighbours

O último método 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), que faz parte da família dos algoritmos 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 irá avaliar. Conforme Cunningham e 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 (2011) 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.

Como existem outras métricas de distâncias, o método pode ser apresentado de uma forma mais generalizada: considere o conjunto de treino X que reúne o conjunto de vetores de observações x_i de um conjunto de preditores P tal que as variáveis de treinamento estão

ligados à variável resposta $y_i \in Y$. Deseja-se utilizar para a classificação a observação $\hat{x}_i$ que pertence ao conjunto de teste $\hat{X}$ de tal forma que para cada $x_i \in X$ podemos calcular a distância entre $\hat{x}_i$ e $\mathrm{x_i}$ por

$$\mathrm{d}(\hat{\mathrm{x}}_\mathrm{i}, \mathrm{x}_\mathrm{i}) = \sum_{\mathrm{p} \in \mathrm{P}} \omega_\mathrm{p} \delta\left(\hat{\mathrm{x}}_\mathrm{ip}, \mathrm{x}_\mathrm{i,p}\right) \tag{3.11}$$

tal que $\delta(\cdot)$ é a função que é usado como métrica de distância e ω_p é alguma métrica de peso que aumenta a importância dos vizinhos mais próximos. O processo é similar a uma "eleição" para qual será a categorização da observação, por meio de uma votação ponderada da distância, onde os k vizinhos votam na classe e tais votos são ponderados pelo inverso de sua distância tal que vale

$$Voto(y_i) = \sum_{p=1}^{P} \frac{1}{d(\hat{x}_i, x_c)^n} 1(\hat{y}_i, y_c). \tag{3.12}$$

em que x_c são as observações vizinhas, n pode ser usado para reduzir ainda mais a influência de vizinhos mais distantes e $1(\hat{y}_i, y_c)$ é uma função que fará a divisão das classes em 1 ou 0, o que completa a estimação. Cabe notar que a escolha do parâmetro k é dependente dos dados em questão, mas que em geral um k maior suprime os efeitos de ruído (ou *outliers*), mas torna os limites de classificação mais difíceis de se identificar.

3.3.4. Algoritmo ROSE

Quando a variável objetivo tem uma frequência desbalanceada, o treinamento é feito de forma desequilibrada, pois como o objetivo do algoritmo é minimizar o erro total (e não por classe), tenderá a focar sua precisão na classe majoritária, dando pouco peso a classe minoritária (BEKKAR, ALITOUCHE, 2013). O algoritmo ROSE (*Random Over-Sampling Examples*), desenvolvido em Menardi e Torelli (2010), simula novos exemplos da classe minoritária a partir de uma estimativa das densidades condicionais das duas classes e o método *kernel*, em que essa geração de dados novos reduz risco de *overfitting*, mantendo a capacidade de generalização do modelo que pode ser perdida pelo *oversampling*, buscando alterar a distribuição das classes para obter uma amostra mais equilibrada. Com o objetivo de contornar o problema de dados desbalanceados, o algoritmo fornece novos exemplos na vizinhança das classes estudadas, de acordo com uma abordagem de *bootstrap* suavizado, descrito em termos computacionais em Lunardon *et al.* (2014). Após o balanceamento do

conjunto de treino, os dados são utilizados no modelo que será testado pelo conjunto de teste que não teve sua distribuição alterada (MERNARDI, TORELLI, 2010).

3.3.5. Validação Cruzada

O processo de treinamento de cada algoritmo será feita 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 objetivo de atingir a configuração de algoritmo que apresente o menor MSE 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 numa configuração de modelo que consiga o máximo possível de generalização na predição da base de dados. O MSE final é calculado pela média dos MSE dos subconjuntos que foram utilizados para a validação em cada iteração. 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.

1º Conjunto	2º Conjunto	3º Conjunto	···	k-ésimo Conjunto
Teste	Treino	Treino	···	Treino
Treino	**Teste**	Treino	···	Treino
		⋮		
Treino	Treino	Treino	···	**Teste**

Figura 10. Representação do k-*fold cross validation*.
Fonte: Elaboração própria.

3.3.6. Medidas para avaliar o desempenho

A avaliação de performance desse método não se limita apenas aos cálculos do MSE, que pode mascarar um desempenho muito bom em prever apenas uma parte da variável binária, e uma execução pobre na categoria complementar. Desta forma, outras métricas foram desenvolvidas para analisar o desempenho do algoritmo e do resultado da predição, a saber: Matriz de Confusão, Acurácia, Sensibilidade, Sensitividade, Precisão e Coeficiente de Correlação de Matthews (MCC).

3.3.6.1. Matriz de Confusão

Para avaliar o problema envolvendo variável binária, em especial aquelas que abrangem 0 e 1, as métricas mais comuns de análise de performance são derivadas da chamada matriz de confusão, que pode ser ilustrado conforme a Figura 11:

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

Figura 11. Matriz de confusão.
Fonte: Elaboração do autor com base em Elrahman e Abraham (2013).

Desta forma, numa 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, ou seja:

$$Acurácia = \frac{VN + VP}{VN + FN + FP + VP} \tag{3.13}$$

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 o 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), que fornece esse valor individual, conforme as expressões (3.14) e (3.15).

$$TVP = \frac{VP}{VP + FN} \tag{3.14}$$

$$TVN = \frac{VN}{VN + FP} \tag{3.15}$$

3.3.6.2. Limiar, Curva ROC e AUC

O vetor predito, fruto das regressões RF e SVM, é composto por números diversos entre 0 e 1, ao passo que a variável original que se pretende prever é composta por 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:

$$\hat{y}_{\mathrm{n}}^{*} = \begin{cases} 0, & \hat{y}_{\mathrm{n}} < \tau \\ 1, & \hat{y}_{\mathrm{n}} \geq \tau \end{cases} \tag{3.16}$$

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, como usou Pijl (2017). O desafio agora é encontrar tal limiar τ (conhecido como *threshold*[18]) que guiará essa divisão, sendo ele definido a partir da chamada curva ROC. A curva ROC (*Receiver Operating Characteristics*) é uma técnica por meio de visualização gráfica, auxiliando na decisão 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 dela o TVP, conhecido na métrica da ROC como Sensitividade, e o TVN, chamada de Especificidade. De fato, o τ que se deseja buscar é aquele que resulta tanto na maior Sensitividade e Especificidade. Além disto, 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). Conforme Elrahman e Abraham (2013) mostram, a área pode ser calculada por:

$$AUC = \frac{TVP + TVN}{2}. \tag{3.17}$$

[18] 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).

3.3.6.3. 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 Elrahman e Abraham (2013) e Boughorbe *et al*. (2017):

$$\text{Precisão Categoria Negativa} = \frac{VN}{VN + FN} \tag{3.18}$$

$$\text{Precisão Categoria Positiva} = \frac{VP}{VP + FP} \tag{3.19}$$

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \tag{3.20}$$

A precisão foca na proporção de acerto na categoria positiva (ou 1, no caso do exemplo binário) em relação aos totais de acertos, ou seja, além disso há também há 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.

3.3.7. 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 por meio do Programa de Subvenção ao Prêmio do Seguro Rural (PSR). Dados abertos envolvendo sinistros são escassos e difíceis 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 através dos dados da *Golden Gate Weather*

Services, 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.). Desta 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 transformados em 3 variáveis numéricas, representadas por um vetor de números inteiro em que o número de cada vetor representa um estado, ou um município ou uma seguradora. Em resumo, os dados estão dispostos conforme o Quadro 1 abaixo:

Quadro 1. Relação das variáveis utilizadas

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.

Os dados foram separados num conjunto de treino, correspondendo aos períodos de 2006 a 2016, e o conjunto de teste, referente aos dados de 2017. O rebalanceamento da base de treino pelo algoritmo ROSE resultou em uma base de 70.000 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 47,9%/52,1%. 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%.

3.4. Resultados e Discussão

Esta seção mostra os resultados de cada método, mas para facilitar a comparação as principais medidas de acurácia de predição do conjunto de teste são apresentadas

conjuntamente no final. O treinamento foi feito por meio do processo de validação cruzada, com $k = 10$ e a Figura 12 ilustra a evolução do MSE de cada *fold* separado do conjunto de treino. A média do MSE para o caso do RF foi de 0,07112378, com desvio padrão de $6{,}967696 \times 10^{-5}$, ao passo que o SVM mostrou MSE médio de 0,2476398 (e $5{,}709922 \times 10^{-3}$ desvião padrão) com o *kernel* radial, que apresentou menor MSE dentre as funções *kernel* conforme será mostrado a seguir. Por sua vez, o k-NN mostrou MSE médio de 0,30886141 (e $2{,}313447 \times 10^{-3}$ de desvio padrão) quando selecionado $k = 1$.

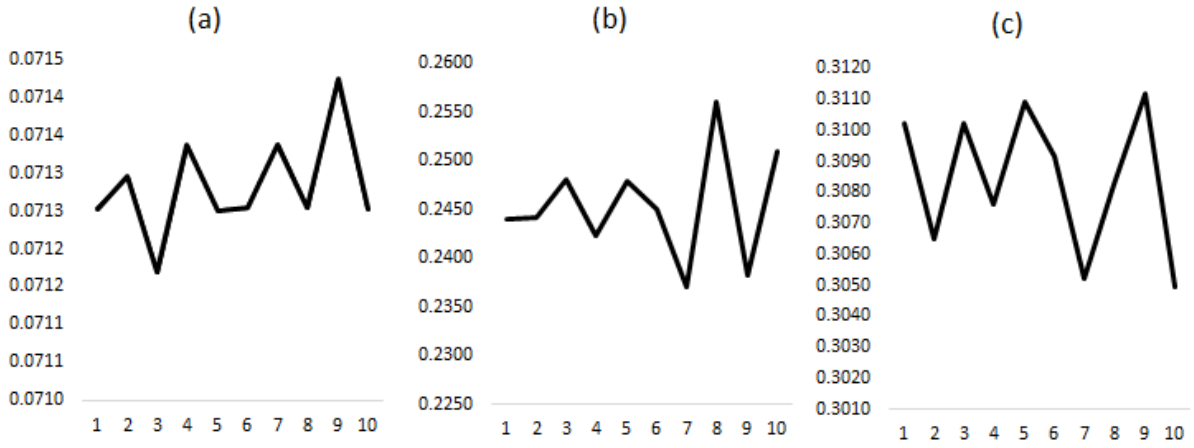

Figura 12. Evolução do Erro Quadrático Médio (MSE) por *fold*. (a) *Random Forest*; (b) *Support Vector Machine*; (c) *k-Nearest Neighbours*.

Fonte: Resultado da pesquisa.

O algoritmo *Random Forest* foi treinado a partir de 200 árvores, entretanto o MSE não mostrou redução significativa após chegar próximo da quinquagésima árvore, revelando que o número ideal para se extrair a média das previsões poderia ser menor. Conforme a Figura 13, o MSE registra redução mínima após esse volume de árvores, visto que o erro permanece muito próximo do 0,07112378. Além disso, foi necessário confirmar o número de variáveis para a divisão da árvore em cada nó, que tem por padrão nesse caso ficar próximo da vizinhança de $m = \frac{p}{3}$. Como o problema trata de $p = 13$, o resultado ficou próximo dado que $m = 5$ apresenta o menor MSE.

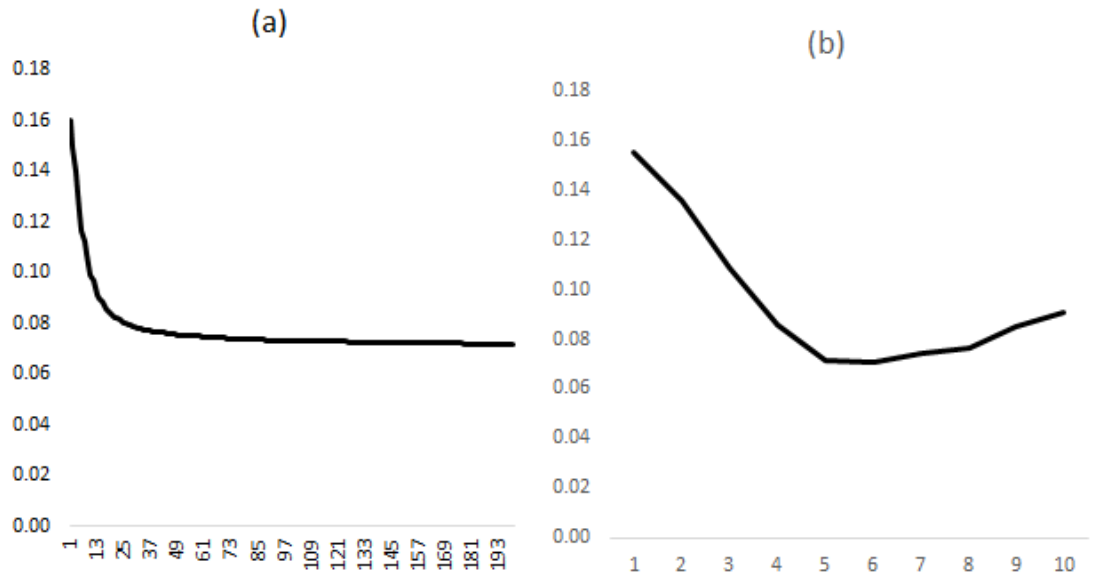

Figura 13. (a) Relações entre o número de árvores e MSE; (b) Relações entre as variáveis por nó das árvores e MSE.

Fonte: Resultado da pesquisa.

Por sua vez, para a aplicação do *Support Vector Machine*, a função radial[19] mostrou quase a metade do MSE registrado pelas funções linear e sigmoide, superando também o resultado da *kernel* polinomial, como mostra a Tabela 7:

Tabela 7. Erros quadráticos médios do SVM por tipo de função *kernel*

Função φ(.)	MSE
Radial	0,2476398
Linear	0,4142696
Polinomial	0,2820777
Sigmoide	0,4629098

Fonte: Elaboração própria.

Para o k-NN optou-se por testar os resultados de 1 a 100 vizinhos, conforme a Figura 14. O MSE é reduzido ao selecionar $k = 1$, ou seja, apenas um único vizinho mais próximo, o que deve ocorrer principalmente pela quantidade de pontos (vetores) e que devem estar excessivamente próximos um dos outros e adoção de apenas 1 vizinho próximo transforma o método na sua configuração mais simples, em que não é preciso a decisão por votos ponderadas para a classificação. O MSE cresce rapidamente a partir da adição de novos vizinhos na análise, se estabilizando próximo de 0,45 a partir do 14ª vizinho.

[19] 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.

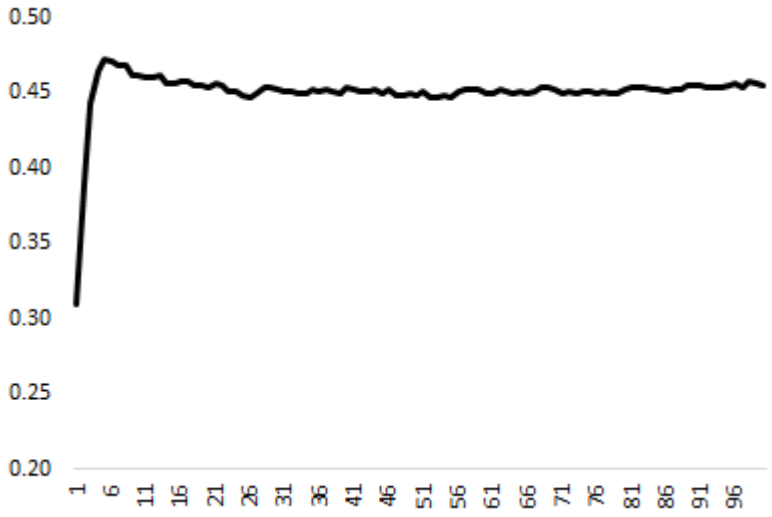

Figura 14. Evolução do Erro Quadrático Médio (MSE) por número de vizinhos (k) selecionados para o k-NN.

Fonte: Resultado da pesquisa.

Após selecionar as melhores configurações de métodos, as predições para o conjunto de teste, que envolvem os dados de 2017, foram realizadas. 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 especificidade. A Figura 15 mostra como foi o resultado considerando os dois modelos para um conjunto de *threshold*.

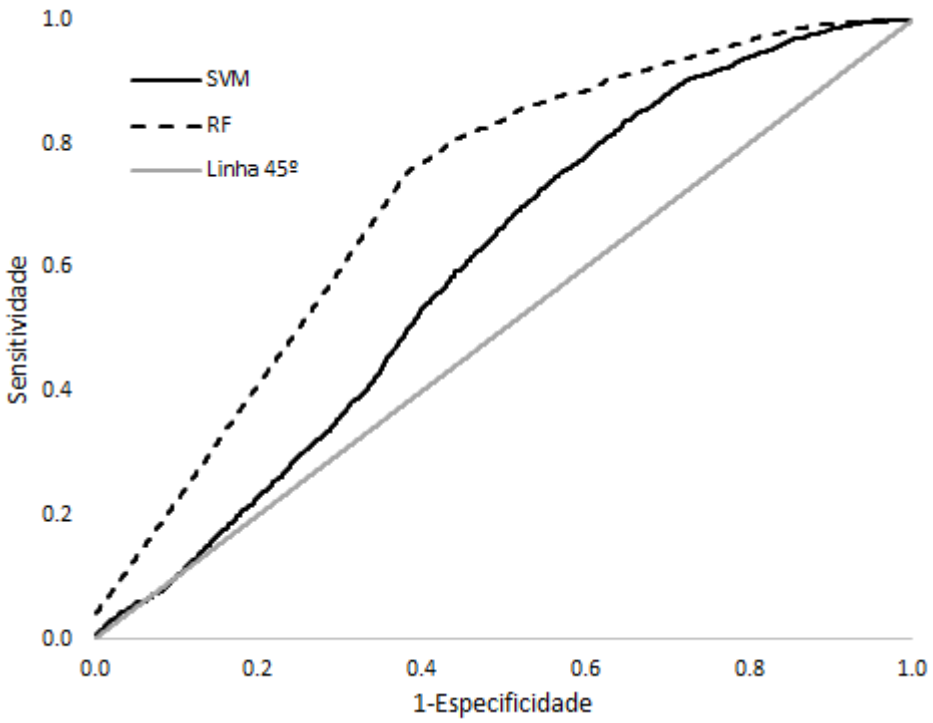

Figura 15. Curvas ROC para os modelos estimados.
Fonte: Resultado da pesquisa.

A figura já ilustra que o algoritmo RF mostrou um desempenho preditivo significativamente 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). A superioridade é confirmada ao se calcular a AUC, que é de 0,7543307 no caso da RF e de 0,6627510 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 16 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, o algoritmo RF dá sinais de uma melhor capacidade preditiva, principalmente por apresentar um número menor em sua diagonal secundária, tendo apresentado o melhor desempenho para prever a ocorrência de sinistro, ou seja, o verdadeiro positivo (prever 1 quando de fato era 1). A importância desse resultado consiste em errar menos em situações em que de fato o sinistro aconteceu, antecipando o seu evento.

RF

		Efetivo	
		0	**1**
Predito	**0**	5557	259
	1	767	1181

SVM

		Efetivo	
		0	**1**
Predito	**0**	4626	843
	1	1698	597

k-NN

		Efetivo	
		0	**1**
Predito	**0**	4900	974
	1	1424	466

Figura 16. Resultado de predição compilado na matriz de confusão por método
Fonte: Resultado da pesquisa.

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 teve (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, apresentando apenas uma perda por um possível custo de oportunidade para as seguradoras. 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 Banco 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 8 mostra melhor desempenho preditivo do *Random Forest* comparado aos demais métodos, em todas as métricas de comparação, isto é, em termos de Acurácia, TVP, TVN, Precisão e MCC. A melhor precisão do RF inclusive na categoria 1, ou positiva, é o que chama mais atenção quando comparado com as demais, mostrando o seu diferencial para a predição de ocorrências de sinistros no mercado de seguro agrícola, o que é reafirmado pela métrica de correlação MCC. 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 8. Principais medidas de avaliação de predição por método

	Acurácia	TVN	TVP	Precisão (0)	Precisão (1)	MCC
RF	0,9014684	0,9199873	0,8201388	0,9573802	0,7000592	0,6975587
SVM	0,6727202	0,7314991	0,4145833	0,8458585	0,2601307	0,1244313
k-NN	0,6911386	0,7748260	0,3236111	0,8341845	0,2465608	0,0891535

Fonte: Resultado da pesquisa.

Abrindo os resultados, é possível observar a diferença por estado na Tabela 9, também revelando maior performance de acurácia para o método RF. Considerando apenas aqueles estados com mais de 500 contratos para o ano de 2017 na base de dados estudada, o algoritmo de RF mostrou um índice de acurácia médio de 0,86 (ou 86% de acerto) consideravelmente superior a acurácia do k-NN (0,71) e SVM (0,53) para a média simples dos mesmos contratos.

Tabela 9. 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	0,50	1,00
BA	45	0,84	0,64	0,78
DF	8	0,75	0,50	0,75
ES	9	1,00	0,78	0,89
GO	537	0,82	0,69	0,75
MA	2	0,50	0,70	1,00
MG	557	0,85	0,00	0,79
MS	325	0,85	0,73	0,74
MT	209	0,87	0,71	0,76
PE	1	1,00	1,00	0,00
PR	3852	0,89	0,69	0,67
RO	1	1,00	0,00	0,00
RS	723	0,87	0,61	0,60
SC	270	0,72	0,51	0,67
SE	2	1,00	1,00	1,00
SP	1220	0,88	0,67	0,72
TO	1	1,00	1,00	1,00

Fonte: Resultado da pesquisa.

Destacando apenas a região Sul, que apresenta 4.845 das 7.764 observações desse conjunto de teste, a média de acurácia foi de 0,83 para o RF, 0,60 para o SVM e 0,65 para o k-NN. Na região Sudeste (correspondendo a 23% da base de teste), o indicador de acurácia médio ficou em 0,91, 0,48 e 0,80, 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.

Por sua vez, olhando pelas culturas envolvidas na base que foi predita, temos que a média da Acurácia para o RF foi de 0,85, ao passo que o SVM (0,65) e k-NN (0,69) também registraram performance inferior. O grande destaque são os desempenhos de predição no caso da soja, milho 2ª safra e do trigo, que foram as culturas que representam 73,3% do número de

sinistros na base trabalhada para o ano de 2017, em linha com a ocorrência de maior sinistros na região Sul do país, que apresenta grandes áreas plantadas das três culturas. Enquanto o algoritmo RF apresentou Acurácia de 0,91, 0,91 e 0,85 para as três culturas, respectivamente, o algoritmo SVM registrou previsão inferior, com Acurácia de 0,72, 0,57 e 0,58, sendo que o mesmo aconteceu no caso do k-NN, com 0,72, 0,61 e 0,55.

Tabela 10. Acurácia da predição dos sinistros ocorrido em 2017, por cultura e por método

Cultura	Observações	RF	SVM	k-NN
Abacate	1	1,00	1,00	1,00
Abóbora	3	0,67	0,00	1,00
Abobrinha	1	1,00	1,00	1,00
Alface	1	1,00	1,00	1,00
Algodão	15	1,00	1,00	0,87
Alho	33	0,79	0,76	0,55
Ameixa	77	0,58	0,64	0,65
Arroz	54	0,78	0,83	0,78
Atemoia	2	1,00	1,00	0,00
Aveia	9	1,00	1,00	0,78
Banana	4	0,50	0,50	0,75
Batata	20	0,95	1,00	0,75
Berinjela	3	0,67	0,67	1,00
Beterraba	1	1,00	1,00	1,00
Café	218	0,91	0,95	0,86
Cana-de-açúcar	155	0,94	0,97	0,85
Canola	7	1,00	1,00	0,71
Caqui	47	0,64	0,57	0,66
Cebola	55	0,71	0,67	0,62
Cenoura	7	1,00	1,00	0,71
Cevada	32	0,75	0,81	0,59
Chuchu	1	0,00	1,00	0,00
Couve-flor	1	1,00	1,00	1,00
Feijão	186	0,89	0,86	0,75
Goiaba	10	0,80	0,70	0,20
Kiwi	8	0,88	0,50	0,50
Laranja	7	0,86	0,43	0,71
Maçã	122	0,76	0,45	0,59
Mandioca	8	1,00	0,75	0,75
Melancia	12	0,92	0,50	0,58
Melão	1	1,00	0,00	0,00
Milho 1ª safra	392	0,96	0,43	0,77
Milho 2ª safra	1577	0,91	0,57	0,61
Nectarina	19	0,79	0,16	0,84
Pepino	1	1,00	0,00	1,00

Pêra	22	0,77	0,18	0,64
Pêssego	67	0,76	0,49	0,63
Pimentão	18	0,83	0,39	0,61
Repolho	1	1,00	0,00	1,00
Soja	3296	0,91	0,72	0,72
Sorgo	19	0,79	0,16	0,79
Tangerina	18	0,94	0,50	0,78
Tomate	116	0,83	0,67	0,73
Trigo	818	0,85	0,58	0,55
Triticale	4	1,00	1,00	0,50
Uva	295	0,81	0,65	0,59

Fonte: Resultado da pesquisa,

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 RF mostrou Acurácia média de 0,84 na predição dos sinistros, seguido por 0,69 do k-NN e 0,65 do SVM, destacando a utilidade do primeiro algoritmo para o caso da Aliança do Brasil, seguradora de maior representatividade no mercado de seguro agrícola brasileira, com 0,91 de Acurácia, ou seja, acertando cerca de 91% dos 1.600 sinistros ocorrido em 2017. O mesmo caso pode ser visto em grandes seguradoras do mercado agrícola como Essor, Mapfre e Swiss Re.

Tabela 11. 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	0,91	0,68	0,75
Allianz	886	0,79	0,66	0,70
Essor	1146	0,89	0,60	0,62
Excelsior	191	0,79	0,58	0,73
Fairfax	493	0,81	0,65	0,67
Mapfre	1073	0,91	0,73	0,71
Porto Seguro	209	0,81	0,60	0,68
Sancor	959	0,85	0,65	0,63
Swiss Re	1069	0,90	0,77	0,70
Tokio Marine	139	0,73	0,56	0,71

Fonte: Resultado da pesquisa

3.4.1. Utilização dos resultados encontrados

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, 2010).

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 os métodos propostos 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 (SUSEP, 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 método de estimação determinado 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 (1994), que se baseia no algoritmo chamado triângulo de *run-off*.

Outra possível utilidade para os resultados dos métodos apresentados é no auxílio do cálculo de prêmio de seguro, visto que a SUSEP não define o método de cálculo. 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.

3.5. Conclusões

Este estudo teve o objetivo de avançar na predição de ocorrências de sinistros no mercado de seguro agrícola, sendo essa uma variável de máxima importância para a estabilidade das seguradoras e do mercado de seguro em geral. Os métodos *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) no ano de 2017, após o treino do algoritmo com os dados do período entre 2006 e 2016.

O método *Random Forest* registrou o melhor desempenho para esse caso particular de dados, exibindo uma boa performance na predição tanto do não sinistro, como do sinistro, com destaque para esse último evento, cujo Erro Quadrático Médio (MSE) foi inferior aos demais métodos. Por sua vez, as métricas de análise conhecidas como Acurácia, Taxa de Verdadeiro Positivo, Taxa de Verdadeiro Negativo, Precisão e Coeficiente de Correlação de Matthews (MCC) também foram as melhores para o RF.

Como aprimoramento 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. Para que isso seja possível, é necessária uma variedade maior e mais precisa de dados contando com informação do contrato de seguro. 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 formas de estudar aqueles contratos de seguro agrícola que sofrem com mais de um sinistro durante a sua vigência e como isso poderia ser predito, sendo necessário também uma base de dados com mais informações para cada contrato.

Referências

BANCO CENTRAL, **Resolução no 4.444, de 13 de novembro de 2015**. 2015

BAUMGARTNER, C.; GRUBER, L. F.; CZADO, C. **Bayesian total loss estimation using shared random effects**. Insurance: Mathematics and Economics, p. 194-201, 2015.

BEKKAR, M.; ALITOUCHE, T. A. **Imbalanced data learning approaches review**. International Journal of Data Mining & Knowledge Management Process, Vol. 3, n. 4, 2013.

BERLATO, M. A.; FARENZENA, H.; FONTANA, D. C. **Associação entre El Niño Oscilação Sul e a produtividade do milho no Estado do Rio Grande do Sul**. Pesquisa Agropecuária Brasileira, Vol. 40, n. 5, p. 423-432, 2005.

BOUGHORBEL, S.; JARRAY, F.; EL-ANBARI, M. **Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric.** PLoS ONE, 12(6): e0177678, 2017. Disponível em: https://www.ncbi.nlm.nih.gov/pubmed/28574989. Acesso em: 10 fev. 2019.

BREIMAN, L. Random Forest Machine Learning, Vol. 45, p. 5–32, 2001.

BREIMAN, L.; FRIEDMAN, J.; STONE, C.; OLSHEN, R. A. **Classification and Regression Trees**. Wadsworth Publishing, 1983

CABRERA, V. E.; FRAISSE, C.; LETSON, D.; PODESTA, G.; NOVAK, J. **Impact of climate information in reducing farm risk by selecting crop insurance programs.** Transactions of the ASABE, Vol. 49, p. 1223–1233, 2006.

CASTRO, C.; BRAGA, A. **Aprendizado supervisionado com conjunto de dados desbalanceados**. Revista Controle & Automação, Vol.22, n. 5, 2011.

CIRINO, P. H.; FERES, J. G.; BRAGA, M. J.; REIS, E. **Assessing the Impacts of ENSO-related weather effects on the Brazilian Agriculture**. Procedia Economics and Finance, Vol. 24, p. 146–155, 2015.

CORTES, C. VAPNIK, V.; **Support-vector networks**. Machine Learning, Vol. 20, n. 3, p 273-298, 1995.

CUMMINS, J. D.; GRIEPENTROG, G. L. **Forecasting automobile insurance paid claim costs using econometric and arima models**. International Journal of Forecasting, Vol. 1, p. 203-215, 1985.

CUMMINS, J. D.; DERRIG, R. A. **Fuzzy Trends in Property-Liability Insurance Claim Costs**. The Journal of Risk and Insurance, Vol. 60, n. 3, p. 429-465, 1993.

CUNHA, G. R.; ASSAD, E. **Uma visão geral do número especial da RBA sobre zoneamento agrícola no Brasil**. Revista Brasileira de Agrometeorologia, Vol. 9, n. 3, p. 377-385, 2001.

CUNHA, G. R.; DALMAGO, G. A.; ESTEFANEL, V. **Influências do fenômeno ENSO sobre a cultura de trigo no Brasil**. Revista Brasileira de Agrometeorologia, Vol. 7, n. 1, p. 127-138, 1999.

CUNNINGHAM, P.; DELANY, S. J. **k-Nearest neighbour classifiers**. Technical Report UCD-CSI-2007-4, 2007.

ELRAHMAN, S. M.; ABRAHAM, A. **A Review of Class Imbalance Problem**. Journal of Network and Innovative Computing, Vol. 1, p. 332-340, 2013.

EL-BASSIOUNI, M. Y.; EL-HABASHI, M. H. **A Multiple Resampling Method for Learning from Imbalanced Data Sets**. Computational Intelligence, Vol. 20, n. 1; p. 18-36, 2004.

ESTABROOKS, A. **Optimizing Crop Insurance under Climate Variability**. Journal of Applied Meteorology and Climatology, Vol. 47, 2007.

FAWCETT, T. **An introduction to ROC analysis**. Pattern Recognition Letters, Vol. 27, p. 861–874, 2006.

FREITAS, M. A. **Modelo logístico aplicado ao mercado de seguros de auto no Brasil: cálculo da probabilidade de sinistros**. Revista Indicadores Econômicos FEE, Vol. 37, n. 3, 2009.

GOLDEN GATE WEATHER SERVICES **El Niño and La Ninã Years and Intensities**. Disponível em: https://ggweather.com/enso/oni.htm. Acesso em: 5 fev. 2019.

GUNN, S. R. **Support Vector Machines for Classification and Regression**. Technical Report, University of Southampton, Faculty of Engineering, Science and Mathematics; School of Electronics and Computer Science, 1998.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer, 2009.

HERNEK, R. F. **Formula loss reserves**. Insurance Accounting and Statistical Association, 1966.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: with Applications in R**. Springer, 2013.

JAPKOWCIZ, N.; STEPHEN, S. **The class imbalance problem: A systematic study**. Intelligent Data Analysis, Vol.6, p. 203-231, 2002.

LIAW, A.; WIERNER, M. **Classification and Regression by randomForest**. R News, Vol. 2, p. 18-22, 2002.

LIU, J.; MEN, C.; CABRERA, V.; URYASEV, S.; FRAISSE, C. W.; **Optimizing Crop Insurance under Climate Variability**. Journal of Applied Meteorology and Climatology, Vol. 47, 2007.

LIU, C.; BERRY, P.; DAWSON, T.; PEARSON, R. **Selecting thresholds of occurrence in the prediction of species distributions**. Ecography, Vol. 28, N. p. 385-393, 2005.

LUNARDON, N.; MENARDI, G.; TORELLI, N. **ROSE: A Package for Binary Imbalanced Learning**. The R Journal, Vol. 6, 2004.

MACK, T. **Which stochastic model is underlying the chain ladder method?**. Insurance: Mathematics and Economics, Vol. 15, n. 2, p. 133-138, 1994.

MINISTERIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO. **Dados de indenizações – 2006 a 2017**. 2018. Disponível em: http://www.agricultura.gov.br/assuntos/riscosseguro/seguro-rural/relatorios-estatisticos. Acesso em: 5 jan. 2019.

MINISTERIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO. **Relatório Seguro Rural - Resultado Geral 2018**. 2019. Disponível em: http://www.agricultura.gov.br/assuntos/riscosseguro/seguro-rural/relatorios-estatisticos. Acesso em: 5 jan. 2019.

MENARDI, G.; TORELLI, N. **Training and assessing classification rules with unbalanced data**. Università di Trieste, Working Paper Series, n. 2, 2010.

OLIVEIRA, N. V. **Mercados de Seguros: Solvência, Riscos e Eficácia Regulatória**. Tese de doutorado disponível em: http://www.ens.edu.br/publicacoes-detalhes/104. Acesso em: 12 jan.2019.

OZAKI, V. A. **Em busca de um novo paradigma para o seguro rural no Brasil**. Revista de Economia e Sociologia Rural, Vol .46, n. 1, p. 97-119, 2008.

PEI, Y.; KIM, T.; ZHA, H. **Unsupervised Random Forest Manifold Alignment for Lipreading**. 2013 IEEE International Conference on Computer Vision, Sydney, p. 129-136, 2013.

PIJL, T. **A Framework to Forecast Insurance Claims**. Dissertação de mestrado disponível em: https://thesis.eur.nl/pub/39731/Pijl.pdf. Acesso em: 20 jan. 2019.

RODRIGUES, A.; MARTINS, E. **Gerenciamento da informação contábil através das provisões técnicas constituídas por sociedades seguradoras**. Revista Universo Contábil, Vol. 6, n. 1, p. 46-66, 2010.

SAMSON, D.; THOMAS, H. **Linear Models as Aids in Insurance Decision Making: The Estimation of Automobile Insurance Claims**. Journal of Business Research, Vol. 15, p. 247-256, 1987.

SMOLA, A.; SCHOLKOPF, B. **A tutorial on support vector regression**. Statistics and Computing, Vol. 14, p. 199–222, 2004.

SOUSA, K. M. **Modelos lineares generalizados e modelos de dispersão aplicados à modelagem de sinistros agrícolas**. Dissertação de mestrado disponível em: http://www.teses.usp.br/teses/disponiveis/11/11134/tde-16032010-163501/pt-br.php; Acesso em: 5 jan. 2019.

STEINMETZ, S.; SILVA, S. C. **Início dos Estudos sobre Zoneamento Agrícola de Risco Climático (ZARC) no Brasil**. Santo Antônio de Goiás: Embrapa Arroz e Feijão, 2017.
SUPERINTENDENCIA DE SEGUROS PRIVADOS. **Provisões Técnicas: Orientações da SUSEP ao Mercado de Seguros, Previdência Complementar Aberta, Capitalização e Resseguro Local**. Disponível em http://susep.gov.br/setores−susep/cgsoa/copra/arquivos−copra/orientacoes/Provisoes%20Tecnicas%20−%20Versao%20−%20201701−NOVO-NOVO.pdf. Acesso em: 15 abr. 2019. 2017

TORGO, L. Data Mining with R - **Learning with Cases Studies.** Chapman and Hall/CRC, 2011.

VAPNIK, V. **Estimation of Dependences Based on empirical Data**. Springer, 1979.

WEISS, G.; MCCARTHY, K.; ZABAR, B, **Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?** International Conference on Data Mining, Las Vegas, p. 35–41, 2007.

YANG, Y.; QIAN, W.; ZOU, H. **Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models**. Journal of Business & Economic Statistics, p. 1–15, 2017.

YE, C.; ZHANG, L.; HAN, M.; YU, Y.; ZHAO, B.; YANG, Y. **Combining Predictions of Auto Insurance Claims**. arXiv:1808.08982, Disponível em https://arxiv.org/abs/1808.08982. Acesso em: 10 jan. 2019.

ZANIBONI, N.; MONTINI, A. **Modelos de poisson inflada de zeros e binomial negativa inflada de zeros na previsão de sinistro de automóveis**. E&G Economia e Gestão, Vol. 15, n. 41, 2015.

4. CONCLUSÃO

O mercado de seguro agrícola é de importância para a proteção da renda dos produtos desse segmento e a previsão de variáveis importantes é fundamental para aumentar a informação e possibilitar o planejamento de todos os agentes que participam desse mercado. Desta forma, o trabalho atual buscou estabelecer métodos para a previsão de duas variáveis importantes nesse segmento: o prêmio e o sinistro.

Pela necessidade de mais estudos envolvendo o seguro agrícola com foco regional buscou-se modelos para o prêmio das principais regiões do país, utilizando metodologias de séries de tempo como o SARIMA, além de alternativas mais recentes como o TBATS, NNAR, ELM e o MAPA. De fato, não se encontrou um método melhor para cada série de prêmio das regiões estudadas, sendo que a região Sudeste mostrou melhor performance com o ELM, a Nordeste, Sul e Centro-Oeste com o SARIMA, tendo essa última região mostrado a pior performance entre as abordadas. O desempenho dos métodos em termos de predição melhorou quando se considerou uma variável mensal derivada da subvenção do prêmio do seguro agrícola, melhorando a performance em duas das quatro regiões abordadas.

Por sua vez, também buscou-se modelar a ocorrência de sinistros ao longo de todo país, utilizando uma base de dados que compreende o período de 2006 a 2017, com mais de 70 mil observações de contratos. Utilizou-se métodos como *Random Forest*, *Support Vector Machine* e *k-Nearest Neighbor* para treinar um conjunto de observações de 13 variáveis. A variável objetivo que seria prevista tem uma composição binária, de modo que 0 indica a não ocorrência de sinistro e 1 a ocorrência, sendo necessário balancear o conjunto de dados por um outro algoritmo (ROSE) para que não tenhamos um método enviesado, visto que a frequência de sinistro é menor. Após avaliar a performance dos métodos a partir de um conjunto de treino separado, correspondendo aos dados do ano de 2017, e por métricas estabelecidas na literatura, conclui-se que o algoritmo *Random Forest* exibe o menor erro dentre os três métodos utilizados para a predição de sinistro na base estudada.

APÊNDICES

APÊNDICE A. Programa de Subvenção ao Prêmio do Seguro Rural

O PSR, criado em 2003 e com sua operação iniciada no fim de 20055 é um programa que busca oferecer apoio aos produtores rurais que desejam proteger suas lavouras contra riscos climáticos adversos. A partir da disponibilização de subvenção econômica ao prêmio do seguro rural, o Governo Federal apoia financeiramente àqueles produtores que contratarem o seguro, arcando com parcela dos custos de aquisição deste produto. O percentual de subvenção disponibilizado é de 30% a 35%, conforme o MAPA. O PSR comporta as modalidades de seguro agrícolas, pecuária, florestas e aquícola.

APÊNDICE B. Informações adicionais dos métodos

A.1 Base para o modelo SARIMA

Os modelos ARMA (p,q), que agregam componentes autoregressivo (AR) e de médias móveis (MA), foram inicialmente propostos por Box e Jenkis (1976) e servem de base para uma ampla gama de trabalhos que buscam estabelecer modelos para previsões. Formalmente, se considerarmos que um processo estocástico, ou então a série de tempo $\{Z_t, t \in Z\}$ temos que o ARMA(p,q) pode ser apresentado na forma

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \qquad \text{(A.1)}$$

com a_t sendo um ruído branco[20], independente de todas as defasagens da série de tempo. Tais modelos trabalham apenas com séries ditas fracamente estacionárias[21], ou seja, vale que

1. $E(Z_t)^2 < \infty$;
2. $E(Z_t) = \mu$ para todo $t \in \mathbb{Z}$;
3. $\mathrm{E}(\mathrm{Z_t} - \mu)(\mathrm{Z_{t-j}} - \mu) = \gamma_j$ sendo γ_j a autocovariância.

[20] O ruído branco tem as seguintes características: $E(a_t) =, E(a_t^2) = \sigma_a^2, E(a_t a_s) =, \forall t, s \neq$.

[21] Uma das formas de avaliar se o processo é estacionário é o teste Dickey-Fuller Aumentado (ADF), verificando a presença de raiz unitária. O teste tem como hipótese nula $H_0: \delta = 0$, sendo δ o coeficiente de presença de raiz unitária da regressão geral dada por $\Delta y_t = a_0 + a_1 t + \delta y_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta y_{t-i} + \varepsilon_t$, onde a_0 é o intercepto, $a_1 t$ o *drift*, e m é o número de defasagens. O teste é feito após o cálculo da estatística $DF = \frac{\delta}{SE(\bar{\delta})}$, em que $SE(\bar{\delta})$ é a estimativa do desvio padrão do erro de $\bar{\delta}$

Desta foram, a série é fracamente estacionária se possui média, variância e autocovariância que permanecem constantes ao longo do tempo, ou seja, a série tenderá a retornar para o seu valor médio.

Teremos uma solução estacionária para Z_t se as raízes da equação característica do componente AR(p), dada por $\phi(x) = 1 - \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} = 0$, forem em módulo maiores do que um (estão fora do círculo unitário). Por sua vez, o componente MA(q) será invertível se a sua equação característica $\theta(x) = 1 - \theta_1 x_{t-1} - \theta_2 x_{t-2} - \cdots - \theta_q x_{t-q} = 0$ também apresentar raízes maiores que a unidade em valor absoluto.

Os modelos ARIMA(p,d,q), por sua vez, são casos gerais dos modelos ARMA(p,q) e podem trabalhar com séries não estacionárias, como são os casos de diversas séries econômicas, e que usualmente se tornam estacionárias após a aplicação de uma quantidade d de diferenças. Se isto ocorrer, a série Z_t será reconhecida como integrada de ordem d, ou $I(d)$, explicando a diferença entre os modelos ARMA e ARIMA.

Se estabelecermos B como um operador de defasagem de tal sorte que $BZ_t = Z_{t-1}$ ou, no caso geral, $B^j Z_t = Z_{t-j}$, é possível ver que $(1 - B)Z = Z_t - Z_{t-1}$ e também que $(1 - B)^2 Z_t = (1 - 2B + B^2)Z_t = Z_t - 2Z_{t-1} + Z_{t-2}$ e assim por diante. Desta forma, o processo ARIMA(p,d,q) também poderá ser escrito na forma

$$\phi(B)(1 - B)^d Z_t = \theta(B) a_t \tag{A.2}$$

em que

1. $\phi(B) = 1 - \phi_1 B^1 - \phi_2 B^2 - \cdots - \phi_p B^p$ é o operador autoregressivo (AR) estacionário de ordem p;
2. $\theta(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - \cdots - \theta_q B^q$ é o operador de médias móveis (MA) invertível de ordem q;
3. $(1 - B)^d$ é o processo de integração de ordem d;
4. a_t é um ruído branco.

e que equivale ao modelo ARMA dado por $\phi(B)W_t = \theta(B)a_t$, notando que W_t é uma transformação de d diferenças da série Z_t, ou seja, Z_t é integral de W_t.

A.2 Demais informações do modelo NNAR

As redes são treinadas recursivamente, ponderando as funções com algoritmos de aprendizagem de forma a minimizar os erros mensurados por funções como a própria *RMSE*.

Além disso, é possível a previsão pela simulação de amostras futuras ao gerar aleatoriamente valores de ε_t a partir de uma distribuição normal ou então pelos valores históricos da série. Ao simular repetidamente, o algoritmo adquire o conhecimento da distribuição dos valores futuros baseados na rede neural. Esses algoritmos costumam apresentar bom desempenho nos conjuntos de treinamento, entretanto podem falhar na predição de valores extremos (tipo *outliers*), como boa parte dos modelos de séries de tempo.

Por fim, Geman et al.(1992) ressaltaram que os modelos de redes neurais devem ser parcimoniosos, observando o *trade-off* entre viés-variância. De fato, a escolha deve balancear a ocorrência de *overfitting* - quando se escolhe um modelo complexo e que tem boa performance no conjunto de treino e performance ruim no conjunto de teste, dado sua variância elevada - ou então a ocorrência do *underfitting* – quando o modelo tem um desempenho ruim no conjunto de treino pela sua simplicidade, mas apresenta um erro semelhante num conjunto de teste, refletindo alto viés e baixa variância.

A.3 Adicionais do Algoritmo MAPA e modelo ETS

Formalmente, o modelo MAPA trata de uma série de tempo Z_t de uma dada frequência, sendo z_t cada observação, e que pode ser desagregada em séries de tempo de frequência k. Para cada série agregada Z_t^k é estimado um modelo ETS com menor AIC possível, possibilitando o vetor de estado Y_i^k de cada série armazenar os componentes de nível, tendência e sazonalidade, insumo necessário para as previsões dos dados. Além da série original Z_t (ou Z_t^1), quando temos $k > 1$ o tamanho da previsão h é ajustado por $h^k = h/k$, ou seja, numa previsão mensal de $h = 12$ (12 meses) numa série agregada trimestralmente, teremos que $h^3 = 12/3 = 4$.

Após a estimação do modelo e da projeção, os componentes são combinados, sendo necessário transformá-los na frequência da série original Z_t^1 através de

$$z_t = \sum_{j=1}^{k} \omega_j z_i^k, \quad t = 1,2,\dots,n; \quad i = t/_k \tag{A.3}$$

em que z_i são as observações da série de frequência k e $\omega = 1/_k$, ou por pesos que não são iguais, conforme o método proposto em Spithourakis et al.(2012) e que também somam um.

Após a equivalência de frequências, é necessário a combinação de todas os componentes de cada série, ou seja, $\overline{Z_t^l}$, $\overline{T_t}$, $\overline{S_t}$ conforme as expressões:

$$\bar{Z}^{lk}_{i+h^{[k]}} = K^{-1}\sum_{k=1}^{K}\bar{Z}^{lk}_{i+h^{[k]}} \tag{A.4}$$

$$\bar{S}^{k}_{i+h^{[k]}} = K^{-1}\sum_{k=1}^{K}\bar{S}^{k}_{i+h^{[k]}} \tag{A.5}$$

$$\bar{T}^{k}_{i+h^{[k]}} = K^{-1}\sum_{k=1}^{K}\bar{T}^{k}_{i+h^{[k]}} \tag{A.6}$$

Por fim, para chegar à projeção da série original, as séries são combinadas de forma aditiva (componentes multiplicativos devem ser transformados em aditivos) tal que vale

$$\bar{Z}^{k}_{i+h^{[k]}} = \bar{Z}^{lk}_{i+h^{[k]}} + \bar{T}^{k}_{i+h^{[k]}} + \bar{S}^{k}_{i+h^{[k]}} \tag{A.7}$$

considerando que haja sazonalidade, caso contrário, o último elemento será zerado, sendo $\bar{Z}^{k}_{i+h^{[k]}}$ a projeção final desejada.

$$z_i^k = \frac{1}{k}\sum_{t=1+(i-1)k}^{ik} z_t \tag{A.8}$$

em que $i = 1,2,3,\dots,n/k,\ t,k \in N, z_i^k$ são as observações da nova série Z_t^k. É importante notar que a série Z_t^k terá n/k observações, ou seja, uma série com 99 observações mensais e que será agregada para frequência trimestral terá $n/k = 99/3 = 33$ observações, sendo cada z_i^k a média aritmética de três observações.

A.3.1 ETS

Os modelos ETS utilizam o arcabouço dos modelos de estado espaço, permitindo avaliar um conjunto grande de modelos de séries temporais por meio de sistemas de equações, que envolvem as chamadas equações de medições e a equações de estado, ou transição, e que de forma recursiva se tornam um algorítmo capaz de previsões e estimação de modelos. Uma configuração padrão e simples para o modelo pode ser adaptado de Espinoza (2004) e Kitagawa (1994) na seguinte forma:

$$Z_t = F_t Y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim RB(0, H_t), i.i.e. \tag{A.9}$$

$$Y_t = C_t Y_{t-1} + \eta_t, \quad P_t\eta_t \sim RB(0, Q_t), i.i.e. \tag{A.10}$$

em que as expressões são as equações de medições e de estado, respectivamente. A despeito da notação, cabe notar que Z_t, a série de tempo, é um vetor de ordem $p \times 1$, Y_t é o vetor de estado que tem ordem $m \times 1$, F_t, C_t, P_t, H_t, Q_t são as chamadas matrizes do sistema. Os parâmetros são estimados por meio da máxima verossimilhança, sendo usualmente utilizado o filtro de Kalman num processo recursivo que possibilidade inferências sobre o vetor de estado e previsões da série de tempo.

Ao utilizar a abordagem de estado espaço, a família ETS apresenta 12 tipos de combinação de suavização exponencial, que envolvem diferentes tipos de sazonalidades (S_t), tendências (T_t) e ruídos brancos (ε_t) conforme abaixo:

1. Modelos Aditivos: Consideram que a série temporal é soma de componentes, na forma $Z_t = T_t + S_t + \varepsilon_t$;
2. Modelos Multiplicativos: Consideram a série de tempo como produto de seus componentes tais que $Z_t = T_t S_t \varepsilon_t$;
3. Modelos Não Aditivos e Multiplicativos (ou *None*): A série não segue os padrões anteriores, sem apresentar componentes de tendência e/ou sazonalidade, apenas o nível;
4. Modelos Amortecidos: Se a série de tempo apresentar tendência constante (crescente ou decrescente), há interferência no desempenho das projeções. Gardner e McKenzie (1985) introduziram um parâmetro φ que amortece a tendência.

Utilizando a metodologia do espaço estado, Ord et al.(1997) mostram que a estrutura da suavização exponencial é composta por um vetor de estado Y_t e um espaço de equações dada por

$$Z_t = h(Y_{t-1}) + k(Y_{t-1})\varepsilon_t \tag{A.11}$$

$$Y_t = f(Y_{t-1}) + g(Y_{t-1})\varepsilon_t \tag{A.12}$$

em que $Y_t = \left(L_t, T_t, S_t, S_{t-1}, \dots, S_{t-(m-1)}\right)$, e se chamarmos $u_t = h(Y_{t-1})$ e $e_t = k(Y_{t-1})\varepsilon_t$ temos então que (A.11) se torna $Z_t = u_t + e_t$. Assim, o material apresentado acima pode ser combinado em 12 modelos conforme a Tabela A.1.

Tabela A.1. Família de modelos ETS

Tendência	Sazonalidade		
	None	Aditivo	Multiplicativo
None	$u_t = L_{t-1}$ $L_t = L_{t-1} + \alpha\varepsilon_t$	$u_t = L_{t-1} + S_{t-m}$ $L_t = L_{t-1} + \alpha\varepsilon_t$ $S_t = S_{t-m} + \gamma\varepsilon_t$	$u_t = L_{t-1}S_{t-m}$ $L_t = L_{t-1} + \alpha\varepsilon_t/S_{t-m}$ $S_t = S_{t-m} + \gamma\varepsilon_t/L_{t-1}$
Aditivo	$u_t = L_{t-1} + T_{t-1}$ $L_t = L_{t-1} + T_{t-1} + \alpha\varepsilon_t$ $T_t = T_{t-1} + \alpha\beta\varepsilon_t$	$u_t = L_{t-1} + T_{t-1} + S_{t-m}$ $L_t = L_{t-1} + T_{t-1}\alpha\varepsilon_t$ $T_t = T_{t-1} + \alpha\beta\varepsilon_t$ $S_t = S_{t-m} + \gamma\varepsilon_t$	$u_t = (L_{t-1} + T_{t-1})S_{t-m}$ $L_t = L_{t-1} + T_{t-1}\alpha\varepsilon_t/S_{t-m}$ $T_t = T_{t-1} + \alpha\beta\varepsilon_t/S_{t-m}$ $S_t = S_{t-m} + \gamma\varepsilon_t/(L_{t-1} + T_{t-1})$
Multiplicativo	$u_t = L_{t-1}T_{t-1}$ $L_t = L_{t-1}T_{t-1} + \alpha\varepsilon_t$ $T_t = T_{t-1} + \alpha\beta\varepsilon_t/L_{t-1}$	$u_t = L_{t-1}T_{t-1} + S_{t-m}$ $L_t = L_{t-1}T_{t-1} + \alpha\varepsilon_t$ $T_t = T_{t-1} + \alpha\beta\varepsilon_t$ $S_t = S_{t-m} + \gamma\varepsilon_t$	$u_t = L_{t-1}T_{t-1}S_{t-m}$ $L_t = L_{t-1}T_{t-1}\alpha\varepsilon_t/S_{t-m}$ $T_t = T_{t-1} + \alpha\beta\varepsilon_t/(S_{t-m}L_{t-1})$ $S_t = S_{t-m} + \gamma\varepsilon_t/(L_{t-1} + T_{t-1})$
Amortecido	$u_t = L_{t-1} + T_{t-1}$ $L_t = L_{t-1} + T_{t-1} + \alpha\varepsilon_t$ $T_t = \varphi T_{t-1} + \alpha\beta\varepsilon_t/L_{t-1}$	$u_t = L_{t-1} + T_{t-1} + S_{t-m}$ $L_t = L_{t-1} + T_{t-1} + \alpha\varepsilon_t$ $T_t = \varphi T_{t-1} + \alpha\beta\varepsilon_t$ $S_t = S_{t-m} + \gamma\varepsilon_t$	$u_t = (L_{t-1} + T_{t-1})S_{t-m}$ $L_t = L_{t-1} + T_{t-1}\alpha\varepsilon_t/S_{t-m}$ $T_t = \varphi T_{t-1} + \alpha\beta\varepsilon_t/S_{t-m}$ $S_t = S_{t-m} + \gamma\varepsilon_t/(L_{t-1} + T_{t-1})$

Fonte: Elaboração própria com dados de Hyndman *et al.* (2008)

Desta forma, é possível estimar os modelos de acordo com a função de verossimilhança condicional, que deve ser minimizada, dada por

$$L^*(\Omega, Y_0) = n\log\left(\sum_{t=1}^{n} e_t^2 / k^2(Y_{t-1})\right) + 2\sum_{t=1}^{n} \log|\mathrm{k}(Y_{t-1})| \tag{A.13}$$

sendo que $\Omega = (\alpha, \beta, \varphi, \gamma)$ são os parâmetros, em que $\alpha, \beta, \varphi, \gamma \in (0,1)$, e $Y_0 = (Z_t, T_t, S_t, S_{t-1})$ é o vetor de estado inicial. Assim, o modelo é selecionado pelo menor AIC, que pode ser definido por $L^*(\widehat{\Omega}, \widehat{Y_o}) + 2p$, sendo p o número de parâmetros.

A.4 Adicionais do TBATS

Em relação ao componente ARMA(p,q) do modelo, o procedimento de correção de erros é dado por duas etapas: Primeiramente é estimado um modelo TBATS sem o componente ARMA, ou seja, com p e q iguais a zero, e depois encontrando o melhor modelo ARMA(p,q) para os termos de erros de tal modelo, pelos métodos já descritos[22]. Assim, a segunda etapa é encontrar é a elaboração de um novo TBATS, mas com o componente ARMA encontrado no passo anterior, ou seja, incluindo a correção de erros.

[22] Não há descrição de modelo desta primeira etapa em De Livera et all. (2010), sendo necessário descobrir pela programação do pacote *forecast* do software R que os autores elaboraram para utilizar a metodologia.

A projeção depende da seleção adequada de k_i usado para o componente sazonal, e utiliza-se o método proposto em Hyndman *et al* (2008), em que após a extração da tendência da série de tempo é utilizado uma regressão linear dada por $\sum_{i=1}^{T}\sum_{j=1}^{k_i} a_j^i cos(\lambda_j^i t) + b_j^i sen(\lambda_j^i t)$ e iniciando com k_1 e adicionando demais componentes que devem ser significativos e que minimizem o AIC.

A.5 Adicionais do ELM

O ELM consegue aperfeiçoar a velocidade de aprendizado e o desempenho das redes, superando problemas como a localização do mínimo local, *overfitting* e a taxa de aprendizagem, situações comuns em algoritmos que utilizam os gradientes (LI *et al*., 2018). A matriz $H\beta = \overline{Z}_t$ pode ser escrita como:

$$H_{m\times l} = \begin{bmatrix} G(\omega_1, b_1, z_1) & \cdots & G(\omega_l, b_l, z_1) \\ \vdots & \ddots & \vdots \\ G(\omega_1, b_1, z_m) & \cdots & G(\omega_l, b_l, z_m) \end{bmatrix},\ \beta_{l\times j} = \begin{bmatrix} \beta_1' \\ \vdots \\ \beta_j' \end{bmatrix}, Z_{m\times j} = \begin{bmatrix} \overline{z}_1' \\ \vdots \\ \overline{z}_j' \end{bmatrix} \tag{A.14}$$

em que a i-ésima coluna de *H* será a saída do i-ésimo neurônio das entradas de $z_1, \dots, z_m$. Vale pontuar que a série de tempo Z_t é um caso particular em que o vetor tem dimensão $Z_{t\times 1}$. Os vetores de peso e o viés podem ser obtidos pelos mínimos quadrados dado por

$$min|H\beta - \overline{Z}_t| \tag{A.15}$$

sendo |.| a norma euclidiana e que tem como solução $\overline{\beta} = H^+\overline{Z}_t$, sendo H^+ a chamada matriz inversa de Moore-Penrose para a matriz *H*, que será apresentada a seguir.

A.5.1 Matriz inversa de Moore-Penrose

De forma geral, a inversa de Moore-Penrose, também conhecida como pseudoinversa, é uma ferramenta para calcular a melhor solução, em termos de ajuste, para um sistema de equações lineares que não possuem uma solução única, ou então a norma mínima de tal sistema. O ajuste é feito pelo método dos mínimos quadrados.

Dado um sistema qualquer de equações lineares $Ax = y$, em que A pode ser singular e não quadrada, Huang *et al*.(2006) mostra que *G* será uma matriz de ordem $n \times m$ será uma inversa de Moore-Penrose de *A* (que tem ordem $m \times n$) se valer

$$AGA = A, \quad GAG = G, \quad (AG)^T = AG, \quad (GA)^T = GA \tag{A.16}$$

e para facilitar a notação, considere A^+ como essa tal matriz inversa. No caso do sistema de equações lineares, $\hat{x}$ será a única menor solução do mínimos quadrados caso $|A\hat{x} - y| = min|Ax - y|$ e logo se Gy é a norma mínima de mínimos quadrados que soluciona o sistema $Ax = y$, então $G = A^+$. Há diversos métodos para o cálculo da matriz A^+, destacando-se o método iterativo e decomposição de valores singulares (ou SVD).