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

Programa de fiscalização reduzida em embalagens de madeira para acondicionamento de importados: impacto do programa no comportamento das empresas e comparação entre modelos preditivos para o manejo do risco

Fábio Souza Aranha

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

Piracicaba
2021

Fábio Souza Aranha
Bacharel em Ciências Econômicas

Programa de fiscalização reduzida em embalagens de madeira para acondicionamento de importados: impacto do programa no comportamento das empresas e comparação entre modelos preditivos para o manejo do risco

versão revisada de acordo com a resolução CoPGr 6018 de 2011.

Orientadora:
Profa. Dra. **SÍLVIA HELENA GALVÃO DE MIRANDA**

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

Piracicaba
2021

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

Aranha, Fábio Souza

Programa de fiscalização reduzida em embalagens de madeira para acondicionamento de importados: impacto do programa no comportamento das empresas e comparação entre modelos preditivos para o manejo de risco / Fábio Souza Aranha. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2021.

93 p.

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

1. Inspeção fitossanitária 2. Gerenciamento de risco 3. Comércio exterior 4. Inspeção reduzida I. Título

AGRADECIMENTOS

Agradeço, em primeiro lugar, a Deus.

Agradeço aos meus familiares, cujo apoio fora imprescindível durante o mestrado. Em especial, aos meus país, pelo apoio e compreensão nos momentos difíceis. Não tenho palavras para expressar o orgulho que tenho ao chamá-los de “meus pais”. Também agradeço ao meu irmão, Lucas, pelo companheirismo e pelo bom humor ao longo de todos esses anos.

Agradeço a todos os professores e funcionários do departamento de Economia da Escola Superior de Agricultura “Luiz de Queiroz”.

Ainda agradeço a minha orientadora Prof.ª Drª. Sílvia Helena Galvão de Miranda pela paciência, compreensão, orientação e apoio nos momentos de dificuldade.

Agradeço a todos os colaboradores e gestores da Aeroportos Brasil Viracopos S.A., pela hospitalidade, colaboração e por fornecer os registros de movimentação de cargas acondicionadas em MEMR e de sua não conformidade, importadas via Aeroporto Internacional de Viracopos.

Agradeço a todos os servidores do Vigiagro/Ministério da Agricultura, Pecuária e do Abastecimento (MAPA) do Aeroporto Internacional de Viracopos. Em especial, ao André Marcondes, pela hospitalidade, ensinamentos e colaboração para a execução desta pesquisa.

Agradeço ao todos os membros da república “Casa Nova”, pela recepção e por sempre estarem de portas abertas quando necessitei.

Agradeço a toda turma de mestrado em economia aplicada de 2018.

A Regiane, Ana, Bruna e Marcelo, pelas palavras de apoio compartilhadas em momentos de dificuldade e pela companhia ao longo de todo o mestrado.

Ao Yuri e Lucas, pelo companheirismo e ocasionais noites de jogatina.

Ao Arthur Lula, pela amizade e conselhos ao longo de todo o mestrado.

Por fim, agradeço, especialmente, ao Augusto e Guilherme. Ao Augusto, por toda a amizade ao longo desses anos e conselhos. Ao Guilherme, por toda a amizade, conselhos e companheirismo nos momentos difíceis. Sem vocês, meus dias na ESALQ teriam sido menos felizes.

Por fim, agradeço a CAPES, pelo apoio financeiro ao longo de todo o mestrado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

SUMÀRIO

RESUMO 6
ABSTRACT 7
1. INTRODUÇÃO 9
Referências 15
2. COMPARAÇÃO ENTRE MÉTODOS PREDITIVOS PARA AS NÃO CONFORMIDADES DAS EMBALAGENS DE MADEIRA. 19
RESUMO 19
2.1. Introdução 19
2.2. Revisão de Literatura 20
2.2.1. Fatores Associados ao Risco Fitossanitário 20
2.2.2. Gerenciamento do Risco 22
2.3. Metodologia 24
2.3.1. Eventos Raros e Dados Desbalanceados 24
2.3.2. Synthetic Minority Oversampling Technique 26
2.3.3. Oversample e Undersample 27
2.3.4. Regressão Logística 28
2.3.4.1. Correção de King e Zeng (2001) 28
2.3.5. Random Forest 29
2.3.6. Support Vector Machines (SVM) e Alternating Genetic Algorithm SVM (ALGA - SVM) 30
2.3.6.1. Support Vector Machines 30
2.3.6.2. SVM e funções de Kernel 32
2.3.6.3. Complexidade e Tamanho da Amostra no SVM 34
2.3.6.4. ALGA – SVM 35
2.3.7. Critérios de Avaliação e Seleção dos modelos 38
2.3.8. Validação Cruzada 41
2.3.9. Fonte de Dados e Variáveis 42
2.4. Resultados e Discussão 44
2.4.1. Resumo Estatístico 44
2.4.2. Validação Cruzada 53
2.4.3. ALGA-SVM 56
2.4.4. Árvores no Random Forest 57

2.4.5. Regressão Logística 58
2.4.6. Random Forest 60
2.4.7. Resultados e Comparação entre os modelos no período de outubro 2018 a março 2019 61
2.5. Considerações Finais 63
Referências 65
3. FISCALIZAÇÃO REDUZIDA DE EMBALANGES DE MADEIRA: UMA ANÁLISE SOBRE A ALTERAÇÃO NO COMPORTAMENTO DAS EMPRESAS IMPORTADORAS. 69
RESUMO 69
3.1. Introdução 69
3.2. Metodologia 71
3.2.1. Diferenças-em-Diferenças 73
3.2.1.1. Pressuposições do método Diferenças-em-Diferenças 75
3.2.2. Variáveis de Origem 76
3.2.2.1. Análise Fatorial de Dados Mistos 77
3.2.2.2. Índice Anual de Importações 77
3.2.3. Base de dados 78
3.3. Resultados 78
3.3.1. Índice Anual de Caracterização da Procedência das Importações 78
3.3.2. Diferenças-em-Diferenças 81
3.4. Considerações Finais 87
Referências 88
4. CONSIDERAÇÕES FINAIS 91
APÊNDICES 93

RESUMO

Programa de fiscalização reduzida em embalagens de madeira para acondicionamento de importados: impacto do programa no comportamento das empresas e comparação entre modelos preditivos para o manejo do risco

O aumento das transações comerciais via comércio exterior eleva o risco de transmissão de doenças e pragas entre os países. Entretanto, a fiscalização confere custos operacionais aos órgãos de fiscalização, bem como custos indiretos aos importadores e exportadores. Logo, ferramentas associadas ao gerenciamento do risco podem reduzir os custos associados a fiscalização mantendo um nível de segurança fitossanitária desejável, bem como aumentar a competitividade dos produtos nacionais no mercado internacional. Um dos modos de se introduzir o gerenciamento de risco na fiscalização ocorre com a adoção do sistema de checagem reduzida. Assim, em dezembro de 2016, entrou em vigor o projeto piloto para a criação de um sistema de fiscalização reduzida das importações acondicionadas em Material de Embalagens de Madeira Regulamentado (MEMR) no Aeroporto Internacional de Viracopos. O projeto prevê que, entre as 50 empresas com maior volume de importação operando nesse aeroporto, estão aptas a ter suas importações fiscalizadas por amostragem as empresas que apresentarem uma taxa de não conformidade das embalagens de madeira inferior a 3%. Com a instituição do sistema de checagem reduzida, algumas questões são instauradas quanto a sua efetividade e operacionalidade, dentre elas: como manter a eficiência da fiscalização das embalagens de madeira enquanto se reduz o volume total fiscalizado e qual o impacto da introdução do sistema de fiscalização reduzida sobre o comportamento das empresas importadoras. Para manter a efetividade da fiscalização, um dos problemas centrais do sistema é a escolha das cargas a serem fiscalizadas, de modo que se propõe a adoção de modelos estatísticos para a previsão das cargas com embalagens de madeira que apresentam não conformidades. Para tanto, comparam-se os métodos: Regressão Logística, *Random Forest* e SVM. Para verificar o impacto da introdução do sistema de checagem reduzida, foram estimados modelos com dados em painel, com os parâmetros do método de diferenças-em-diferenças. Os resultados apontam que os modelos da família *Random Forest* detiveram o melhor desempenho na predição dos MEMR com não conformidades e a adoção desse método para a seleção das cargas a serem fiscalizadas levaria, no período de outubro de 2018 a março de 2019, à redução de 77.616 MEMR fiscalizados. Os resultados dos modelos em dados em painel com os parâmetros de diferenças-em-diferenças apontaram que a adoção do sistema de checagem reduzida não impactou o número de MEMR importados com irregularidades, após sua implementação. Não se verificou, tampouco, impacto na probabilidade de não atender ao critério da taxa de não conformidade inferior a 3%. Conclui-se que, apesar do número significativo de informações analisadas, o modelo pode não estar captando, ainda, as consequências dessa mudança na estratégia da política de inspeção, dado o curto horizonte de análise.

Palavras-chave: Inspeção fitossanitária, Gerenciamento de risco, Comércio exterior, Inspeção reduzida

ABSTRACT

Reduced inspection program in wood package material: impacts of the program on firms behavior and comparison between predict models to risk management

The growth of commercial transactions in global trade raises the risk of pest transmission between the countries. However, the phytosanitary inspection implies operational costs to the inspection agencies, and indirect costs to importers and exporters. Therefore, tools associated with risk management can reduce the costs associated with phytosanitary inspection, maintaining a desirable level of phytosanitary safety and increasing the competitiveness of domestic products in the international market. One of the ways to introduce risk management in phytosanitary inspection occurs with the adoption of the reduced inspection system. Thus, in December 2016, were introduced a pilot project for a reduced inspection system of wood packing material on imports on Viracopos International Airport. With the proposed system, the 50 companies with highest import volume, operating at this airport, are able to have their imports inspected by sampling, if they present a non-conformity rate of wood packages below 3%. With the institution of the reduced inspection system, some questions are raised about its effectiveness and operationality, among them: how to maintain the efficiency of the wood packaging material inspection while reducing the total volume inspections and what is the impact of the introduction of the reduced inspection system on the importing companies behavior. To maintain the inspection effectiveness, one of the main questions for the reduced inspection system is the choice of the wood packaging material to be inspected. So it is proposed the adoption of statistical models to forecast the non-conformity in wood packaging material. So the performance of some statistical models forecast on the non-conformity are compared: Logistic Regression, Random Forest and SVM. To verify the impact of the introduction of the reduced inspection system, models with panel data were estimated, with the parameters of the differences-in-differences method. The results indicate that the models of the Random Forest family had the best performance in the prediction of non-conformity in wood packaging material and the adoption of this method would lead in a reduction of 77,616 wood packaging material inspected, on october 2018 to march 2019. The results of the models in panel data with the difference-in-differences parameters pointed out that the adoption of the reduced inspection system did not change the number of wood packaging material with non-conformity by the imported by the importers. There was also no impact on the probability of not meeting the criterion of non-conformity rate below 3%. It is concluded that, despite the significant number of information analyzed, the model may not yet be capturing the consequences of this change in inspection policy strategy, given the short analysis horizon.

Keywords: Phytosanitary inspection, Risk management, World trade, Reduced inspection

1. INTRODUÇÃO

Com o crescimento da economia mundial e, por consequência, o aumento do comércio entre as nações, elevou-se o risco de espraiamento de doenças e pragas via comércio exterior. Dentre os materiais transportados internacionalmente associados a transmissão de enfermidades, destacam-se as embalagens e suportes de madeira (MONTEFERRANTE *et. al*, 2018; HAACK; PETRICE, 2009). Estima-se que 70% dos produtos transacionados entre as nações estão acondicionados a Material de Embalagem de Madeira Regulamentado (MEMR) (USDA, 2003).

Praga, de acordo com a Convenção Internacional de Proteção dos Vegetais (CIPV) a partir da Norma Internacional de Medidas Fitossanitárias n° 5 (NIMF-5), é definida como: "qualquer espécie, animal, agente patogênico, planta, nocivos a plantas ou produtos vegetais" (FAO, 2009). Ademais, define-se praga quarentenária como: pragas não presentes, ou não amplamente distribuídas e controladas, com potencial em afetar o nível de produção de uma determinada região (GAZZONI, 2018).

A introdução de pragas em uma determinada região pode trazer prejuízos biológicos e econômicos a região afetada. As pragas invasoras podem afetar as práticas de manejo florestal, bem como podem afetar a fauna nativa, desmatamento e produtividade agropecuária. Ademais, é comum que as pragas introduzidas não possuam predadores naturais na nova região, desregulando o ecossistema em que foram introduzidas (WALLNER, 1996). Estima-se que a introdução de pragas via comércio gera um prejuízo anual de US$ 314 bilhões, considerando a perda de produção e custos com o controle das pragas (PIMENTEL, 2002). Entretanto, se essa projeção for atualizada com base no crescimento da produção agrícola mundial, os prejuízos situam-se entre US$ 454 bilhões e US$ 500 bilhões (GAZZONI, 2018).

Ao longo dos anos, diversos países relataram os impactos que a introdução de fitossanidades via movimentação de mercadorias e pessoas detiveram sobre a produção local e ecossistemas nativos. Estima-se que 19 das 70 principais pragas florestais dos Estados Unidos foram introduzidas via movimentação de mercadorias (PIMENTEL, 1986). Segundo Haack (2006), ao menos 25 espécies de coleóptera estabelecidas no ecossistema dos Estados Unidos provem de outras nações e adentraram ao país via comércio exterior. Ademais, ainda foram relatadas a introdução de *Anoplophora glabripennis* (besouro asiático) em diversas nações europeias e da américa do norte (HERARD *et al*., 2006), bem como a introdução da vespa da madeira, *Sirex noctilio*, nos países do hemisfério sul (HURLEY *et al.*, 2007). Desse modo, a

introdução de fitossanidades via comércio é um problema que afeta todas as nações e causa prejuízos diretos a produção local.

No Brasil, a introdução de pragas via comércio afetou diversas espécies e culturas agrícolas. Dentre as espécies de madeira, o *Eucalyptus* foi afetado pela introdução de diversas fitossanidades ao longo da metade do século XX e início do século XXI (IEDE, 2005). Dentre as pragas que afetaram o *Eucalyptus* cabe citar: A broca *Phoracantha semipunctata* na década de 1950 (BERTI FILHO *et al.* 1995), *Phoracantha recurva* em 2001 (WILCKEN *et al.*, 2002) e *Glycaspis brimblecombei* em 2003 (WILCKEN *et al.*, 2003). Outra espécie de madeira afetada por introdução de pragas via comércio foi o *Pinus*. Durante a década de 90, o Pinus foi afetado pela introdução de algumas espécies de pulgões (IEDE *et al.*, 1998; LAZZARI; ZONTA-DE-CARVALHO, 2000) enquanto nos anos 2000 registrou-se a *Pissodes castaneus* e *Pineus boerneri* (IEDE *et al.*, 2004). Além das espécies já introduzidas nos biomas brasileiros, outras espécies de pragas podem afetar a produção nacional de *Eucalyptus* e *Pinus*. Segundo Iede (2005), outras 18 espécies de pragas são potencialmente quarentenárias para o *Pinus* e 2 para o *Eucalyptus*. A introdução de algumas dessas pragas pode gerar diversos impactos econômicos ao setor extrativo, dada a importância da produção de *Eucalyptus* e *Pinus* no setor (IEDE, 2005).

Nesse cenário de transferências de fitossanidades via comércio, diversos procedimentos regulatórios e medidas de fiscalização foram aplicados em escala global, a fim de se reduzir o risco de associado à transmissão de pragas em MEMR. Embora a fiscalização de mercadorias requeira um montante de recursos alocados em aspectos operacionais da inspeção, os custos associados aos aspectos técnicos são inferiores aos custos de controle da praga, uma vez que a fitossanidade se instalou na região em questão (LOVETT *et al.*, 2016).

Desse modo, CIPV publicou, em 2002, a Norma Internacional de Medidas Fitossanitárias n° 15 (NIMF-15). A NIMF-15 estabeleceu diretrizes para o tratamento das embalagens de madeira, bem como as regras para o reaproveitamento e reparo do MEMR, sob a perspectiva do tratamento químico desempenhado no MEMR e ainda estabeleceu as regras para a identificação que o MEMR foi tratado (IPPC, 2002). Dois tratamentos foram prescritos pela NIMF-15, sendo eles: Tratamento térmico da embalagem de madeira ou fumigação. O tratamento térmico pode ser realizado de duas formas: O MEMR deve ser aquecido a 56 graus Célsius por 30 minutos ou via aquecimento dielétrico por 1 minuto a 60 graus Celsius. Já a fumigação pode ser realizada com brometo de metila ou fluoreto de sulfuril. A NMIF-15 também estipula o padrão informações referentes a realização do tratamento na madeira estão dispostas no MEMR. As embalagens devem ser devidamente carimbadas pós tratamento. O

carimbo deve conter: Símbolo do IPPC, Código ISO de dois dígitos do país em que o tratamento foi realizado, Código para a identificação do órgão que realizou o tratamento e código do tratamento, em que: HT refere-se ao tratamento de Calor, DH refere-se ao tratamento de calor via aquecimento dielétrico, MB refere-se à fumigação com brometo de metila e SF refere-se à fumigação com fluoreto de sulfuril. A NMIF-15 ainda prescreve os procedimentos a serem realizados caso o MEMR importado não detenha tratamento ou não possua o carimbo. Os procedimentos incluem: Incineração do MEMR e devolução ao país exportador (IPPC, 2002).

A NIMF-15, em termo gerais, foi uma das maiores ações em termos de proteção fitossanitária de embalagens de madeira para o transporte internacional de cargas (IEDE, 2012). Estima-se que o impacto da normativa nos custos econômicos foram inferiores aos seus benefícios, tendo em vista os prejuízos na produção devido às pragas (STRUTT *et al.*, 2012).

Em termos de fiscalização das embalagens de madeira, não há um padrão único adotado por todas as nações (ROCHA, 2013). Segundo Everett (2000), cerca de 18% dos 300 mil contêineres marítimos com destino a Nova Zelândia são fiscalizados. Nos Estados Unidos apenas 2% das cargas que adentraram ao país via portos marítimos e aeroportos são fiscalizadas (WORK *et al.*, 2005).

No Brasil, O Ministério da Agricultura, Pecuária e Abastecimento (MAPA) executa a fiscalização das embalagens de madeira através do Sistema de Vigilância Agropecuária Internacional (Vigiagro). De acordo com Rocha (2013), o primeiro dispositivo legal a instaurar a fiscalização de embalagens e suportes de madeira foi a Portaria Interministerial MMA/MF n° 499, de 03 de novembro de 1999. Essa estabelecia o tratamento obrigatório das embalagens advindas dos seguintes países: China, Coréia do Norte, Coreia do Sul e Estados Unidos da América. Posteriormente, o Ministério da Agricultura, Pecuária e Abastecimento introduziu a Instrução Normativa SDA/MAPA n° 04, de 6 de janeiro de 2004 (BRASIL, 2004). Em termos gerais, a Instrução Normativa n° 4 internalizava os preceitos da Norma Internacional de Medidas Fitossanitárias (NIMF) n° 15 e procedimentos operacionais para a inspeção fitossanitária.

Entretanto, embora a fiscalização dos MEMR seja necessária, a inspeção das embalagens de madeira tornou-se um desafio para os órgãos reguladores. O aumento contínuo do fluxo de mercadorias internacionalmente comercializadas exige à expansão do volume de importações fiscalizadas, fato o qual esbarra nas restrições orçamentárias dos órgãos responsáveis pelas inspeções (ROCHA, 2013; PONTES, 2013; SURKOV, 2007). Ademais, do ponto de vista dos importadores, os procedimentos de inspeção nas embalagens de madeira, embora necessários, representam um custo a mais no processo produtivo. Segundo Arvis *et. al*

(2012), a retenção das cargas importadas e exportadas nas unidades de fiscalização das fronteiras representam custos às empresas, bem como colaboram para a perda de competitividade dos produtos nacionais.

Nesse sentido, a Instrução Normativa n° 32 de 23 de setembro de 2015 e a Instrução Normativa n° 39, de 27 de novembro de 2017 acenam para instrumentos voltados a análise e gerenciamento do risco na fiscalização das embalagens de madeira que acondicionam as importações, de forma a reduzir os custos ligados ao fluxo de mercadorias (MONTEFERRANTE *et. al*, 2018).

Em consonância com as diretrizes estipuladas pela Instrução Normativa n° 39 e os entraves da fiscalização, estabeleceu-se o projeto piloto para a introdução de um sistema de fiscalização reduzida no Aeroporto Internacional de Viracopos. O projeto tem como objetivo a criação de uma lista de empresas cujas importações, acondicionadas por MEMR, serão fiscalizadas via amostragem. Antes da introdução do projeto, todas as importações acondicionadas em MEMR eram fiscalizadas. A fiscalização dos MEMR segue os padrões internacionais expostos na NIMF-15. Assim, o MEMR fiscalizado é condenado caso apresente: presença de pragas, resquícios de casca na embalagem de madeira, galeria ativa, ausência de carimbo que comprove o tratamento do MEMR como prescrito na NIMF-15 ou marca da inspeção não conforme com os parâmetros descritos na NIMF-15.

A escolha das empresas integrantes do sistema de fiscalização reduzida do Aeroporto Internacional de Viracopos obedece a dois critérios: Pertencer ao conjunto de empresas alto volume de importação acondicionada a MEMR e a taxa anual de não conformidades das embalagens de madeira não deve exceder 3%. Ademais, a fim de facilitar o controle sobre as empresas candidatas ao canal, o primeiro critério foi traduzindo em uma lista de 50 empresas (“Top 50”). A lista “Top 50” foi construída com base nas empresas com 2000 ou mais MEMR internalizados no aeroporto, no período que percorre 2014, janeiro a agosto de 2015 e janeiro a novembro de 2016 (ADAMI; MIRANDA, 2019). Desse modo, apenas as empresas pertencentes a lista “Top 50” são elegíveis ao sistema de fiscalização reduzida. Dentre as “Top 50”, apenas as empresas que detiveram da taxa de não conformidade dos MEMR inferior a 3% foram inseridas no sistema de checagem reduzida. Ademais, estabeleceu-se que o mesmo procedimento de avaliação ocorrera anualmente, promovendo a exclusão ou entrada de empresas conforme os critérios de volume de cargas com MEMR importado e taxa de não conformidade das embalagens de madeira estabelecidos.

Com a introdução do sistema de fiscalização reduzida, esperava-se a redução com custos de inspeção, bem como redução do tempo alocado no processo de importação. De 2016 à março

de 2019, 18468 empresas importaram via Aeroporto Internacional de Viracopos. Do total de empresas, cerca de 58,35% de todo os Pallet de Carga Granel[1] (PCG) importados com MEMR provem das empresas da lista “Top 50” e 55,98% do total de PCG. Assim, mais da metade de todo o volume de PCG importado via o Aeroporto Internacional de Viracopos provem desse grupo de empresas. Quanto a não conformidade dos MEMR, cerca de 2,55% dos PCG com MEMR das empresas da lista “Top 50” detiveram não conformidades, as demais empresas importadoras detiveram, como um todo, 4,39%. Logo, embora o alto volume de PCG, o grupo de empresas escolhidas para o sistema de fiscalização reduzida detém menor nível de não conformidade do que os demais importadores do aeroporto. Em termos dos motivos para a condenação, em ambos os grupos de empresas, a ausência ou não conformidade do carimbo foi o grande motivo para a condenação do MEMR, representando 98,89% das condenações do grupo “Top 50” e 99,16% das condenações das demais empresas. Por fim, cabe destacar que nenhuma condenação por presença de pragas foi registrada nos MEMR das “Top 50”.

Assim, a metodologia do projeto estabeleceu que, ao fim de cada ano, as empresas pertencentes ao grupo das 50 maiores importadoras são reavaliadas quanto às taxas de não conformidade encontradas nas cargas e, a partir desses resultados, inseridas ou retiradas do sistema de fiscalização reduzida. O teste de efetividade do sistema de fiscalização reduzida da madeira teve início em meados de dezembro de 2016. Desde então o projeto está em operação.

Logo, o projeto de fiscalização reduzida no Aeroporto Internacional de Viracopos baseia-se no histórico e volume de importação para definir quais mercadorias serão inspecionadas. Entretanto, há outros modos de operacionalizar um sistema de fiscalização reduzida. Uma forma de reduzir o volume de mercadorias fiscalizadas é com o uso de modelos estatísticos voltados a predição da não conformidade do MEMR. Assim, pode-se definir quais cargas serão inspecionadas a partir de um conjunto específicos de informações dos MEMR. Logo, o uso dos modelos estatísticos permite expandir a sistema de fiscalização reduzida para um conjunto maior de mercadorias, uma vez que não está restrito ao conjunto de empresas com histórico de importação.

O uso de modelos estatísticos no sistema de fiscalização reduzida já foi explorado por outros autores como: Rocha (2013), Pontes (2013) e Surkov (2007). Estima-se que caso o modelo desenvolvido por Rocha (2013) fosse empregado no porto de Santos, em 2013, R$ 1,84

[1] Neste trabalho, toda a análise de não conformidade em MEMR decorre sobre as unidades de PCG, dado que a inspeção das importações, bem como o registro do seu resultado são realizados para cada unidade de PCG.

milhão em custos seriam economizados pelos importadores, bem como 1.841,25 de horas de serviços públicos.

Assim, este trabalho procurou explorar o uso de modelos de previsão para a não conformidade dos MEMR. Para tanto, comparou-se o desempenho de algumas técnicas estatísticas comumente empregadas em problemas semelhantes à classificação da não conformidade das embalagens de madeira, sendo essas: *Random Forest*, SVM e Regressão Logística. Ademais, os modelos empregados neste trabalho realizam a projeção a partir de um conjunto de características associadas a variável objetivo. É a partir desse conjunto de características, traduzidas com variáveis explicativas, que são realizadas as previsões das não conformidades dos MEMR. Logo, a escolha das variáveis explicativas influência diretamente a performance dos métodos estatísticos aplicados, bem como a sua aplicação prática.

Para garantir a aplicabilidade prática dos modelos e a replicação desse estudo em outras unidades de fiscalização, optamos por usar regressores com cuja fonte é de fácil acesso e não restrita. Ademais, o conjunto de variáveis explicativas foi selecionado com base nos fatores destacados na literatura como determinantes para a seleção das cargas a serem inspecionadas, sendo esses: Variáveis associadas a origem do MEMR (ROCHA, 2013; EYRE; HAACK, 2017; HAACK, 2006; EYRE *et al.*, 2018; CATON; DOBBS; BRODEL, 2006; MONTEFERRANTE *et. al*, 2018; SURKOV, 2007), tipo de produto acondicionado na embalagem de madeira (EYRE; HAACK, 2017; ROCHA, 2013; HAACK, 2006; EYRE *et. al.*, 2018; CATON; THOMAS; BRODEL, 2006) e frequência de cargas com não conformidade longo do ano. (CATON; THOMAS; BRODEL, 2006; SZYNISZEWSKA *et al.*, 2016; SURKOV, 2007).

Além de reduzir os custos aos importadores e aos órgãos de fiscalização, o sistema de fiscalização reduzida deve oferecer um estímulo as empresas para importar com MEMR de melhor qualidade.

Uma vez que as mercadorias dos participantes da lista de fiscalização reduzida são fiscalizadas por amostragem, as mercadorias não retidas para a fiscalização não estão sujeitas aos processos de inspeção. Logo, as mercadorias não selecionadas para a inspeção detêm menos etapas no processo de importação, e por consequência, detém menor custo ao importador. Sendo assim, é possível que a introdução dos sistemas de fiscalização reduzida ofereça um estímulo as empresas da lista “Top 50” a melhorar o perfil do MEMR que acopla suas importações. Para verificar se esse estímulo foi suficiente para alterar o perfil do MEMR importado, estimamos modelos de diferenças-em-diferenças.

Desse modo, os objetivos explorados nesse trabalho traduzem-se nos seguintes objetivos específicos:

a) Comparação entre os métodos comumente aplicados em problemas de classificação na predição da não conformidade dos MEMR importados;
b) Identificação dos fatores comumente associados a não conformidade, bem como tradução desses fatores em variáveis explicativas nos modelos de predição, enquanto respeita-se os critérios de acessibilidade às variáveis;
c) Identificação dos impactos que o sistema de fiscalização reduzida deteve sobre a taxa de não conformidade e volume de cargas com MEMR não conforme das empresas a lista "Top 50", através dos modelos de diferenças-em-diferenças;

Os objetivos específicos foram traduzidos em dois artigos. O primeiro abrange a comparação dos modelos predição e o primeiro e segundo objetivos. Já o segundo artigo abrange os impactos do sistema de fiscalização reduzida e o terceiro objetivo.

Referências

ADAMI, A. C. O.; MIRANDA, S. H. G. Análise das informações das Inspeções de embalagens e suportes de madeira em 2018, Reavaliação dos critérios para o Canal Verde e Análise dos Benefícios da implementação do canal. Abr./2019.

ARVIS, J.-F. *et al.* **Connecting to Compete 2012:** Trade Logistics in the Global Economy – the Logistics Performance Index and its Indicators. Washington: World Bank. 2012.

BERTI FILHO, E.; CERIGNONI, J. A.; SOUZA JUNIOR.C. N. Primeiro registro de *Phoracantha semipunctata* (Fabricius 1775) (Coleoptera: Cerambycidae) no Estado de São Paulo. **Revista de Agricultura**, v. 70, n. 1, p. 1-16, 1995.

BRASIL. Instrução Normativa DAS/MAPA n° 4, de 06 de janeiro de 2004. **Estabelece, em caráter emergencial, até que se complete o processo de ajustamento da Legislação Fitossanitária Brasileira, a Norma Internacional e cumprimento dos prazos de notificação aos organismos internacionais, os procedimentos de inspeção e fiscalização de embalagens e suportes de madeira utilizados no transporte de mercadorias no comércio internacional.** Diário Oficial da União, Brasília, Distrito Federal, 2004.

BRASIL. Instrução Normativa SDA/MAPA nº 32, de 23 de setembro de 2015. **Estabelece os procedimentos de fiscalização e certificação fitossanitária de embalagens, suportes ou peças de madeira, em bruto, que serão utilizadas como material para confecção de embalagens e suportes, destinados ao acondicionamento de mercadorias importadas ou exportadas pelo Brasil.** Diário Oficial da República Federativa do Brasil, Brasília, Distrito Federal, 2015.

BRASIL. Instrução Normativa SDA/MAPA nº 4, de 6 de janeiro de 2004. **Aprova os procedimentos de inspeção e fiscalização de embalagens e suportes de madeira utilizados no transporte de mercadorias no comércio internacional.** Diário Oficial da República Federativa do Brasil, Brasília, Distrito Federal, 2004.

CATON; B. P.; DOBBS, T. T.; BRODEL, C. F. Arrivals of hichhking insect pests on international cargo aircraft at Miami International Airport. **Biological Invasions**, v. 8, p. 765-785, 2006.

ERYE, D. *et al.* Variation in Inspection Efficacy by Member States of Wood Packaging Material Entering the European Union. **Journal of Economic Entomology**, v. 111, n. 2, p. 707-715. 2018.

EVERETT, R. A. Patterns and pathways of biological invasions. **Trends in Ecology & Evolution**, v. 15, n. 5, p. 177–178, 2000.

EYRE, D.; HAACK, R. A. Invasive Cerambycid Pests and Biosecurity Measures. *In:* WANG, Q. (Org.). **Cerambycidae of the World**: biology and pest management. Boca Raton, FL: CRC, 2017. Cap. 13, p. 563-618.

FAO. **Normas Internacionais para Medidas Fitossanitárias nº 15. Regulamentação de Material de Embalagem de Madeira no Comércio Internacional.** Roma, Itália. Food and Agriculture Organizationo of the United Nations, 2009.

GAZZONI, D. L. Introdução *In:* Fidelis, E. G. *et al.* **PRIORIZAÇÃO DE PRAGAS QUARENTENÁRIAS AUSENTES NO BRASIL**. Roraima: Empraba Roraima, 2018. p. 15-19.

HAACK, R. A. Exotic bark- and wood-boring Coleoptera in United States: recent establishments and interception, **Can. J. for Res.** v. 36, p. 269-288, 2006.

HAACK, R. A.; PETRICE, T. R. Bark- and Wood-Borer Colonization of Logs and Lumber After Heat Treatment to ISPM 15 Specifications: The Role of Residuals Bark. **Journal of Economic Entomology**, v. 102, n. 3, p 1075-1084, 2009.

HÉRARD, F *et al.* Anoplophora species in Europe: infestations and management processes. **EPPO Bulletin**, v. 36, p. 470-474, 2006.

HURLEY, B. P.; SLIPPERS, B.; WINGFIELD, M. J. A. A comparison of control results for the alien invasive woodwasp, *Sirex noctilio*, in the southern hemisphere. **Agricultural and Forest Entomology**, v. 9, p. 159–171, 2007.

IEDE, E. T. A CIPV e as normas internacionais de medidas fitossanitárias NIMF 15: a evolução dos tratamentos quarentenários de madeira. In: SEMINÁRIO INTERNACIONAL SOBRE PRAGAS QUARENTENÁRIAS FLORESTAIS, Colombo, 2012, **Anais...**

IEDE, E. T. *et al.* Ocorrência de *Cinara pinivora* (Homoptera: Aphididae) em Pinus elliottii e P. taeda em Cambará do Sul (RS) e Lages (SC). In: CONGRESSO BRASILEIRO DE ZOOLOGIA, 12., 1998, Recife. **Resumos**. [S.l.]: Sociedade Brasileira de Zoologia. 1998, p. 141.

IEDE, E. T. **Importância das Pragas Quarentenárias Florestais no Comércio Internacional:** Estratégias e Alternativas para o Brasil. Colombo: Embraba Florestas, 2005. (Embrapa Florestas. Boletim de pesquisa e desenvolvimento, 22).

IPPC. International Plant Protection Convention. International Standards for Phytosanitary Measures: Guidelines for Regulating Wood Packaging Material in International Trade, Publ. No. 15. **Food and Agriculture Organization of the United Nations**, Roma, Italy. 2002.

LAZZARI, S. M. N.; ZONTA-DE CARVALHO, R. C. Aphids (Homoptera: Aphididae: Lachninae: Cinarini) on *Pinus* spp. and *Cupressus* sp. in Southern Brazil. In: INTERNATIONAL CONGRESS OF ENTOMOLOGY, 21., 2000, Foz do Iguassu. **Abstracts**. Londrina: Embrapa Soja, 2000. v. 1, p. 493. (Embrapa Soja. Documentos, 143).

LOVETT, G. M. *et al.* Nonnative forest insects and pathogens in the United States: Impacts and policy options. **Ecological Applications**, vol. 26, n. 5, pp. 1437-1455, 2016.

MONTEFERRANTE, E. C. *et al.* Interceptações de pragas em embalagens de madeira no Porto de Santos. **Revista de Política Agrícola**, v. 27, n. 1, p. 132-141, 2018.

PIMENTEL, D. Biological invasions of plants and animals in agriculture and forestry. *In:* MOONEY, H. A.; DRAKE, J. A**. Ecology of biological invasions of North America and Hawaii**. New York: Springer-Verlag. p. 149-162. 1986

PIMENTEL, D. Biological invasions: economic and environmental costs of alien plant, animal, and microbe species. **CRC Press**, 2002.

PONTES, F. S. S. **Técnicas de mineração de dados aplicada na inspeção de importações para verificação dos status fitossanitário de embalagens de madeira em contêineres**. 2013, 67 f. Dissertação (Mestrado em Defesa Sanitária Vegetal) – Universidade Federal de Viçosa, Viçosa. 2013.

ROCHA, B. G. D. **Identificação de perfis de risco associados a envios importados com probabilidade de abrigar pragas quarentenárias.** 2013. 73 f. Dissertação (Mestrado em Defesa Sanitária Vegetal) – Universidade Federal de Viçosa, Viçosa. 2013.

STRUTT. A *et al.* Evaluating the impacts of an international phytosanitary standard for wood packaging material: Global and United States trade implications. **Forest Policy and Economics**, v. 27, p. 54-64, 2012.

SURKOV, I. V.**Optimising import phytosanitary inspection.** 2007. 196 f. Tese - Wageningen University, Wageningen. 2007.

SZYNISZEWSKA, A. M.; LEPPLA, N. C.; HUANG, Z.; TATEM, A. J. Analysis of Seasonal Risk for Importation of the Mediterranean Fruit Fly,Ceratitis capitata(Diptera: Tephritidae), via Air Passenger Traffic Arriving in Florida and California. **Journal of Economic Entomology**, v. 0 n. 0, p. 1-12, 2016.

USDA. **Regulatory impact analysis of the proposed rule to adopt the international standard on wood packing material in international trade.** 2003

WALLNER, W. E. Invasion of the tree snatchers: the Asian gypsy moth's entrance into North America shows that exotic-pest invasions can happen in unexpected ways. **American Nurseryman**, v. 183, n. 6, p. 28-31, 1996.

WILCKEN, C. F. *et al.* Ocorrência de *Phoracantha mrecurva* (Coleoptera: Cerambycidae) em eucalipto no Brasil. In: CONGRESSO BRASILEIRO DE ENTOMOLOGIA, 19., 2002, Manaus. **A entomologia no século 21 e o manejo da biodiversidade**: resumos. [S.l.]: Sociedade Entomológica do Brasil, Manaus: INPA: Fundação Universidade do Amazonas. p. 153, 2002.

WING, C.; SIMON, K.; BELLO-GOMEZ, R. A. Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. **Annual Review of Public Health**, v. 39, p. 453-469, 2018.

WORK, T. T. *et al.* Arrival rate of nonindigenous insect species into the United States through foreign trade. **Biological Invasions**, v. 7, n. 2, p. 323–332, 2005.

2. COMPARAÇÃO ENTRE MÉTODOS PREDITIVOS PARA AS NÃO CONFORMIDADES DAS EMBALAGENS DE MADEIRA.

RESUMO

O aumento das transações comerciais via comércio exterior eleva o risco de transmissão de doenças e pragas entre as nações. Assim, cabe ao Estado, via órgãos de fiscalização, impedir a entrada de mercadorias com fitossanidades nas fronteiras nacionais. Entretanto, a fiscalização confere custos operacionais, bem como custos indiretos aos importadores e exportadores. Logo, ferramentas associadas ao gerenciamento do risco podem reduzir os custos associados a fiscalização, bem como aumentar a competitividade dos produtos nacionais no mercado internacional. Um dos modos de se introduzir o gerenciamento de risco na fiscalização ocorre com a adoção do sistema de checagem reduzida. Nesse caso, apenas parte das importações são fiscalizadas. Assim, um dos problemas centrais do sistema é a escolha das importações a serem fiscalizadas. A fim de soluciona tal problemática, esse artigo propõe a adoção de métodos estatísticos na previsão das importações com embalagens de madeira que apresentam não conformidades. Para tanto, comparou-se os métodos: Regressão Logística, *Random Forest* e SVM.

2.1. Introdução

O aumento do comércio entre as nações elevou o risco de espraiamento de doenças e pragas via comércio exterior. Um dos principais veículos no transporte das enfermidades é o Material de Embalagem de Madeira Regulamentado (MEMR). Desse modo, cabe as autoridades locais de cada nação estabelecer normas e procedimentos a fim de fiscalizar os MEMR transacionados via comercio exterior. Entretanto, os procedimentos de fiscalização do MEMR conferem custos aos importadores e órgãos reguladores, bem como a perda de competitividade dos produtos no mercado global. Desse modo, a preocupação do Estado brasileiro com a competitividade dos produtos nacionais e os custos crescentes com a fiscalização de mercadorias culminaram em mudanças das diretrizes dos procedimentos fitossanitários relacionados ao comércio exterior, a partir da NIMF-39. A NIMF-39 prevê que a fiscalização do trânsito de mercadorias exportadas e importadas ocorra via instrumentos voltados ao gerenciamento de risco (BRASIL, 2017). Nesse cenário, trabalhos como Rocha (2013), Pontes (2013) e Surkov (2007), que propuseram o uso de instrumentos estatísticos na fiscalização de cargas importadas, são imprescindíveis para a adoção das novas diretrizes de fiscalização.

Tanto Rocha (2013), quanto Surkov (2007) basearam-se no uso da tradicional regressão logística como método para reduzir o volume de MEMR a serem fiscalizados. Em ambos os

casos, a regressão logística é utilizada para classificar quais cargas devem ser fiscalizadas, com base em características de origem, tipo da carga e importador. Já Pontes (2013), baseou-se no uso do algoritmo de arvores de decisão, mas com o mesmo objetivo de Rocha (2013) e Surkov (2007). Embora a regressão logística e o algoritmo de árvores de decisão sejam comumente empregados no tratamento de eventos dicotômicos, e, portanto, cabe sua aplicação na escolha de cargas a serem inspecionadas, outros métodos estatísticos demonstraram performances superiores no seu uso para o gerenciamento do risco (GHATASHEH, 2014). Dentre os métodos, cabe citar: *Random Forest*, *Support Vector Machine* (SVM) e Redes Neurais. Sendo assim, este artigo propõe a comparação entre as técnicas estatísticas para a predição das cargas de madeira não conformes. As técnicas analisadas compreendem as famílias de modelos: *Random Forest*, Regressão Logística e SVM. Para tanto, utilizam-se dados provenientes dos registros da fiscalização das MEMR do Aeroporto Internacional de Viracopos, realizada pelo Sistema de Vigilância Agropecuária Internacional (Vigiagro), da Secretaria de Defesa Agropecuária (SDA) - MAPA, no período de 2016 a março de 2019.

Além desta introdução, este artigo contém mais cinco seções. A segunda seção apresenta a revisão sobre os fatores associados ao risco fitossanitário, bem como as ferramentas de gerenciamento do risco, e os procedimentos de inspeção fitossanitária realizados no Brasil. A seção três apresenta os métodos utilizados, critérios de comparação de desempenho dos modelos e suas variáveis independentes. A seção quatro apresenta os Resultados e Discussão. A seção cinco detém as Referências Bibliográficas.

2.2. Revisão de Literatura

2.2.1. Fatores Associados ao Risco Fitossanitário

Diversos autores analisaram os perfis das cargas de produtos transacionados, a fim de se compreender os fatores associados ao risco fitossanitário no comércio internacional. Em linhas gerais, há um consenso de que as origens das importações apresentam diferentes taxas de não conformidade, sendo um importante fator para a análise do risco fitossanitário (ROCHA, 2013; EYRE; HAACK, 2017; HAACK, 2006; EYRE *et al.*, 2018; CATON; DOBBS; BRODEL, 2006; MONTEFERRANTE *et. al*, 2018; SURKOV, 2007).

Nos Estados Unidos, entre 1985 e 2008, os seguintes países de origem registraram o maior número de cargas com problemas fitossanitários: China, Itália, México, Turquia e Espanha (HAACK, 2006; EYRE; HAACK, 2017; HAACK *et al.*, 2014). Para a União

Europeia, a China também figura como uma das origens com maior incidência de cargas não conformes. Em 2013, estabeleceu-se que toda importação de produtos minerais ou derivados de minerais, importados da China, deveriam ser submetidas a fiscalização, em todos os países da União Europeia (ERYE *et al.*, 2018).

Já no Porto de Santos, entre 2015 a 2017, o continente asiático figurou como a principal origem de cargas interceptadas com a presença de pragas. Nesse período, 57,36% das importações com pragas quarentenárias tiveram a Índia como origem. Cingapura, China, Vietnã e Indonésia completam as cinco maiores procedências das pragas interceptadas naquele Porto no período (MONTEFERRANTE *et. al*, 2018).

Outros fatores, como o tipo de produto acondicionado na embalagem de madeira, também demonstraram ser importantes para a presença de pragas. Os MEMR usados para acondicionar ou apoiar produtos grandes e pesados costumam ser feitos a partir de madeiras de qualidade inferior, e também com qualidade inferior dos tratamentos a que são submetidas, em comparação às embalagens de madeira que acondicionam os produtos destinados ao consumidor final (EYRE; HAACK, 2017).

Entre 1985 e 2008, nos Estados Unidos, os seguintes produtos registraram as maiores taxas de incidência de não conformidades em suas embalagens de madeira: Telhas, Maquinário, Mármore, Aço e Ferragens (HAACK, 2006; EYRE; HAACK, 2017). No aspecto geral, Eyre *et. al.* (2018) relatam que os produtos derivados do setor extrativo mineral reportam taxas de não conformidade superioras aos demais. Já Caton, Thomas e Brodel (2006), em análise das importações no Aeroporto Internacional de Miami, em 1998 e 1999, argumentam que há diferenças significativas de taxas de não conformidade em cargas contendo plantas e roupas.

Além da natureza dos produtos acondicionados, as empresas importadoras também se configuram como variáveis associadas ao risco fitossanitário (ROCHA, 2013). Além da origem, tipos de carga, identidade da empresa importadora, a frequência de cargas com não conformidade aparenta ser sazonal ao longo do ano. Importações cujo país de origem está em período do ano com maior umidade do ar parecem estar relacionadas com uma maior incidência de pragas (CATON; THOMAS; BRODEL, 2006).

Szyniszewska *et al.* (2016) demonstram que o risco de se introduzir a *Ceratitis capitata (mosca das frutas)* no estado da Califórnia é maior entre janeiro e agosto, em comparação ao período de setembro a dezembro. Essa sazonalidade ainda se relaciona à origem da carga. Esses autores argumentam que o risco de introdução da praga na Florida é maior durante os meses de janeiro a agosto em voos originados dos seguintes países: Brasil, Colômbia, Panamá, Venezuela, Argentina e Equador. No caso do estado da Califórnia, o maior nível de risco é

identificado durante os meses de maio a agosto, com origem nos seguintes países: Brasil, Panamá, Colômbia e Itália.

Surkov (2007), em seu trabalho sobre as importações holandesas de plantas ornamentais, para o período de 1998 a 2001, advoga que há um aumento na probabilidade de se rejeitar a importação de uma carga, devido a presença de pragas, no outono e primavera europeias.

2.2.2. Gerenciamento do Risco

Embora necessária, a fiscalização fitossanitária do fluxo de mercadorias comercializadas internacionalmente gera custos aos importadores e aos órgãos fiscalizadores (FURLAN; PINTO, 2015). Nesse cenário, ferramentas capazes de manejar o risco, com base nas características das cargas, podem reduzir os custos associados a fiscalização (ROCHA, 2013; MONTEFERRANTE *et. al*, 2018).

Como fatores associados a escolha das cargas a serem inspecionadas, a Norma Internacional de Medidas Fitossanitárias n° 23 (NIMF-23) (IPPC, 2006) propõe que a seleção dos produtos de origem vegetal ocorra com base nos seguintes fatores: Medidas de mitigação das pragas adotadas por parte do país exportador; Tipo de *commodity* e o seu uso; Local de produção; Tamanho da carga; Frequência, Volume e Duração do transporte; Experiências previas com o exportador; Qualidade da embalagem; Recursos Técnicos e Financeiros do exportador e dos procedimentos de inspeção; Histórico de não conformidade do exportador e dificuldade de se identificar irregularidade em uma carga.

Dados os ganhos com a seleção de cargas a serem amostradas, autores como Rocha (2013) e Surkov (2007) propuseram o uso de métodos estatísticos para identificar os principais fatores associados ao risco fitossanitário nas importações e para a escolha das cargas a serem fiscalizadas. Rocha (2013), desenvolveu seu trabalho para as embalagens de madeira dos produtos importados via Porto de Santos, no período de 2008 a 2012. O autor sugere que a seleção das cargas a serem inspecionadas ocorra via predição de um modelo de regressão logística, cuja variável dependente é a não conformidade das importações. Como preditores foram utilizados: A origem das cargas, tipo de *commodity* e empresa importadora. A partir da validação externa, realizada com dados do ano de 2013, o autor advoga que 1.841,25 horas de serviços públicos seriam economizadas com a adoção do modelo, e aproximadamente R$1.840.000,00 do orçamento voltado aos órgãos de fiscalização.

Surkov (2007) utilizou uma regressão logística multinomial, com o intuito de predizer e identificar as variáveis associadas ao risco fitossanitário e demais não conformidades, nas importações de plantas ornamentais na Holanda. Com dados do período de 1998 a 2001, o risco fitossanitário foi associado a: continente de origem e nível de renda do exportador; Participação de plantas, flores e produtos agrícolas na pauta de exportação da origem da carga; Tipo de *commodity* importada; Número de espécies, Volume de cargas e Certificações do importador; Quantidade de cargas no transporte; e Estações do ano no hemisfério Norte. O autor ainda utilizou dummies para identificar cada um dos anos do período analisado - 1998, 1999, 2000 e 2001. No contexto geral, a exceção do nível de renda da nação de origem, todos fatores mostraram-se associados ao risco de pragas. Entretanto, somente o nível de renda foi significativo na predição das demais não conformidades.

Tanto Rocha (2013), como Surkov (2007), utilizaram e regressões logística para predizer o risco fitossanitário, sendo que tal prática é comum na análise de outras atividades econômicas. Dentre essas, destaca-se a análise de risco de inadimplência, comumente associada a oferta de crédito. Nessa, além da regressão logística, a análise de discriminante é usada para a predição do risco, tanto para empresas, quanto para consumidores (CROOK; EDELMAN; THOMAS, 2007; HAND; NIALL; 2000). Entretanto, além das técnicas convencionais, outros métodos podem ser empregados para a análise de risco de inadimplência (GHATASHEH, 2014). Logo, diversos trabalhos focaram na comparação de outras técnicas de predição à tradicional regressão logística e análise de discriminante. Fantazzini e Figini (2009) compararam a regressão logística e o método *Random Survival Forest* na predição do risco de inadimplência de pequenas e médias empresas. Os resultados apontaram que na predição de dados fora da amostra utilizada na construção dos modelos, a regressão logística deteve melhores resultados.

Já Malekipirbazari e Aksakalli (2015), com o intuito de prever o risco nas transações de empréstimo eletrônico *peer-to-peer*, utilizaram os métodos: *Random Forest*, Vizinhos Principais, Regressão Logística, *Support Vector Machine* e um modelo matemático. Os autores observam que o *Random Fore*st foi superior aos demais métodos estatísticos e aos usuais scores de crédito na predição dos “bons pagadores”, entretanto, não obteve uma boa performance para o conjunto de consumidores de alto risco.

Kruppa *et al.* (2013) investigaram o risco de inadimplência em contratos de financiamento de domicílios. Para tanto, os autores utilizaram os seguintes métodos: Regressão Logística, *Random Forest*, Vizinhos Principais e Vizinhos Principais com *bagging*. Os

resultados apontaram que o *Random Forest* foi superior aos demais métodos na predição da inadimplência.

2.3. Metodologia

2.3.1. Eventos Raros e Dados Desbalanceados

No contexto geral, desproporção entre as categorias, em uma variável categórica de interesse, impõe obstáculos ao uso de técnicas estatísticas para sua predição. De acordo com Paal (2014), os conjuntos de dados com variáveis categóricas cuja frequência de uma categoria é muito inferior as demais, são denominados: Dados com Eventos Raros ou Dados Desbalanceados. A desproporção entre as categorias costuma afetar a predição por técnicas de classificação (AMIN *et al*, 2016), bem como produz parâmetros e predições de probabilidade viesados, no caso da regressão logística (KING; ZENG, 2001).

A proporção entre os MEMR com alguma irregularidade no Aeroporto Internacional de Viracopos fora de 3,4%, entre janeiro de 2016 a março de 2019. A tabela 2.1 ilustra a evolução da taxa de não conformidade anual do Material de Embalagem de Madeira Regulamentado (MEMR), inspecionado no Aeroporto Internacional de Viracopos nesse período.

Tabela 2.1: Proporção anual de MEMR condenadas e não condenadas, de janeiro de 2016 a março de 2019. Aeroporto Internacional de Viracopos – Campinas/SP

	2016	2017	2018	2019	Total
MEMR não condenadas	149.874	133.492	139.288	27.495	450.149
MEMR condenadas	4.055	5.040	5.436	788	15.319
Proporção entre MEMR condenadas e não condenadas (em %)	0,027	0,038	0,039	0,029	0,034

Fonte: Elaboração própria a partir de dados fornecidos pela ABV.

De acordo com a Tabela 2.1, a proporção entre MEMR condenados e não condenados manteve-se abaixo 0,04 em todos os anos. Logo, a não conformidade das embalagens de madeira caracteriza-se como um “Evento Raro”. Sendo assim, procedimentos corretivos foram adotados antes da estimação dos modelos propostos. Paal (2014) denota que a desproporção entre categorias pode ser tratada adotando os seguintes métodos: Replicação/criação de observações da categoria menos frequente (*Overampling*) ou retirada das observações da

categoria mais frequente (*Undersampling*); Métodos que introduzam penalizações na estimação dos coeficientes dos métodos; Introdução de modificações nas funções de ligação tradicionais (para o caso dos modelos Logit e Probit).

Entretanto, de acordo com Amin *et al.* (2016), a exclusão de observações da categoria mais frequente pode resultar na exclusão de informações importantes para a modelagem, afetando a performance dos modelos estimados. Já os métodos de aumento do número de observações da categoria menos frequente não induzem a alterações no desempenho dos classificadores, embora possam afetar o tempo de cálculo de tais métodos. Nesse quesito, Weiss (2004) destaca que, para o caso da Regressão Logística, a replicação de observações da categoria menos frequente altera apenas o intercepto da regressão, enquanto os coeficientes angulares mantêm-se os mesmos. Já King e Zeng (2001), baseando-se em métodos de penalização aos estimadores, propõem o ajuste dos coeficientes via correção, a partir do viés dos parâmetros. Os autores ainda propõem um fator de ajuste para as probabilidades estimadas.

Desse modo, foram adotados três procedimentos corretivos: A correção dos estimadores e probabilidades estimadas propostas por King e Zeng (2001) para a regressão logística. Ademais, utilizamos duas técnicas comumente aplicadas na literatura para a correção de amostras em desbalanço, sendo essas: *Oversampling*, *Synthetic Minority Oversampling Technique* (SMOTE). O *Oversample* foi utilizado para a correção da amostra empregada na Regressão Logística, assim como em: Mohammed, Jumanah e Abdullah (2020) e Alrasheed e Che (2018). Ademais, o *Oversample* também foi aplicado para a correção das amostras do *Random Forest*, assim como em: Mohammed, Jumanah e Abdullah (2020), Sawangarreerak e Thanathamathee (2020) e Qu *et al.* (2020). Assim como o Oversample, o SMOTE foi aplicado na Regressão Logística (MOHAMMED; JUMANAH; ABDULLAH, 2020; ALRASHEED; CHE, 2018) e *Random Forest* (MOHAMMED; JUMANAH; ABDULLAH, 2020; ALRASHEED; CHE, 2018; LE *et al.*, 2018; QU *et al.*, 2020; SAWANGARREERAK; THANATHAMATHEE, 2020; ABDOH; RIZKA; MAGHRABY, 2018; BHAGAT; PATIL, 2015; TAN *et al.*, 2019). Para o caso do SVM, a própria aplicação do algoritmo ALGA-SVM pressupõe *K* observações de cada categoria na composição da amostra, e assim, não há problema de desbalanço entre as categorias com essa aplicação.

2.3.2. *Synthetic Minority Oversampling Technique*

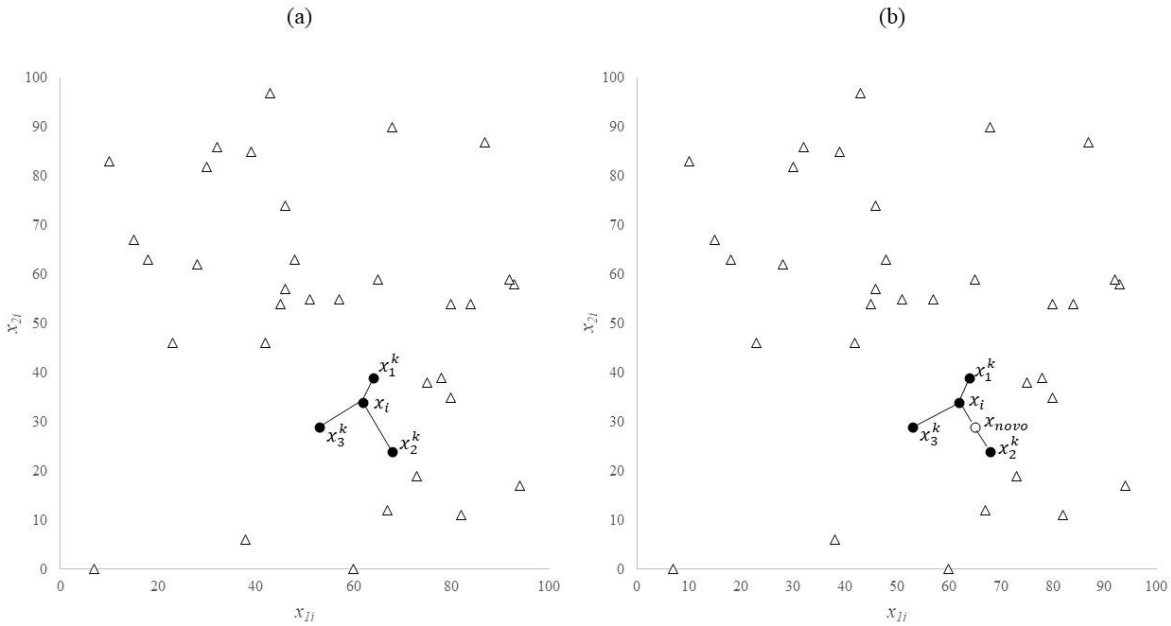
O SMOTE é uma técnica empregada para gerar observações sintéticas a partir das observações presentes na amostra. A técnica baseia-se nos k vizinhos mais próximos de uma determinada observação i, sendo os k vizinhos e a observação pertencentes a mesma classe, para gerar as observações sintéticas. Logo, para os pares $(x_i, y_i) \in Y_M$, $i = 1, 2, \dots, N$ em que Y_M é o conjunto de observações da categoria minoritária, temos que para cada $x_i \in Y_M$ existe um conjunto $Y_i^k \subset Y_M$ cujas observações $x_j^k \in Y_i^k$, $j = 1, 2, \dots, k$ detém as menores distâncias para a observação x_i. Desse modo, as novas observações são construídas a partir da escolha aleatória de uma observação x_j^k a partir da seguinte expressão:

$$x_{novo} = x_i + \theta\left(x_j^k - x_i\right). \tag{2.1}$$

Em que: x_{novo} denota a nova observação; x_i denota as variáveis independentes do indivíduo i; θ denota uma constante aleatória em que $\theta \in [0,1]$; x_j^k denota as variáveis independentes do indivíduo j que está entre as k observações com menor distância para i.

A Figura 2.1 ilustra o processo de construção de novas observações.

Figura 2.1: Aplicação do SMOTE para 2 variáveis independentes e 3 vizinhos, desde a identificação dos vizinhos mais próximos de x_i (a) à construção da observação x_{novo} (b).



Fonte: Modificado de Le *et al.* (2018).

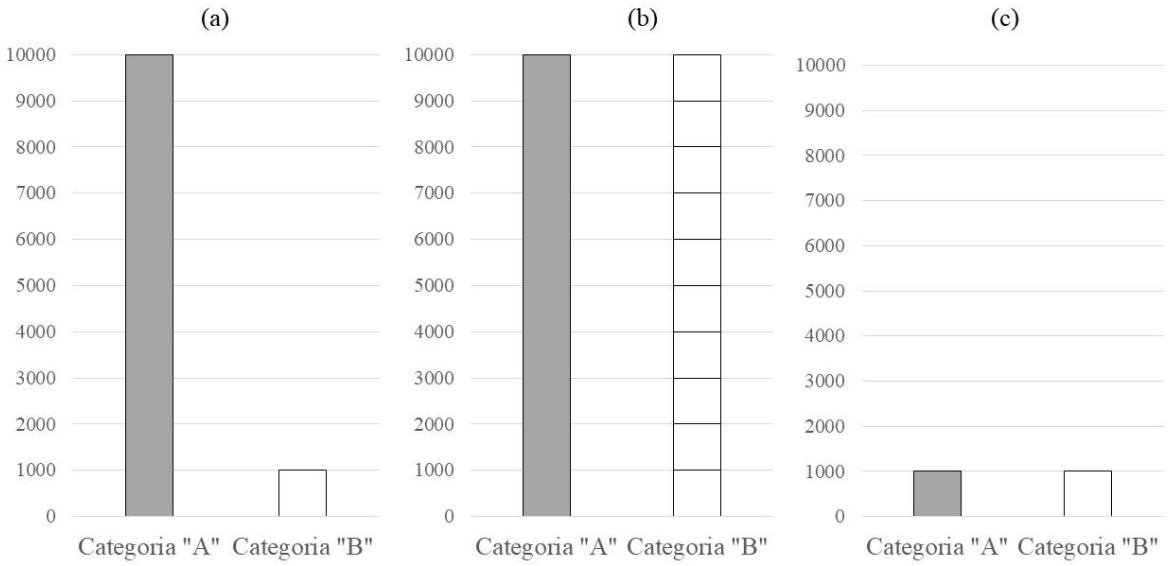
A Figura 2.1 demonstra a criação de uma nova observação a partir de x_i com $k = 3$. Nesse caso, x_2^k foi escolhido aleatoriamente e, a partir de (2.1), gerou-se x_{novo}. O processo exposto na Figura 2.1 é executado até que a proporção entre as duas categorias de y sejam equivalentes.

2.3.3. *Oversample* e *Undersample*

O *Oversample* remete ao processo de amostragem aleatória da categoria minoritária, de modo a gerar novas observações ou a replicação de observações pré-existentes, ampliando o número de observações da categoria minoritária. Assim como o *Oversample*, o *Undersample* baseia-se em métodos voltados a amostragem aleatória, mas nesse caso pretende-se reduzir o tamanho da categoria majoritária. Para tanto, são utilizados métodos que, de alguma forma, excluem observações da categoria principal (MOHAMMED; JUMANAH; ABDULLAH, 2020).

A Figura 2.2 ilustra o intuito principal de ambos os métodos.

Figura 2.2: Exemplo de amostra com desbalanço entre as categorias “A” e “B” (a), correção do desbalanço com *Oversample* da categoria minoritária (b) e correção do desbalanço com *Undersample* da categoria majoritária (c).



Fonte: Modificado Mohammed, Jumanah e Abdullah (2020).

Como ilustrado na Figura 2.2, a amostra original (a) é desbalanceada, de modo que a proporção entre a categoria minoritária e majoritária é de 1:10. Em (b), foi aplicado alguma técnica de amostragem na categoria minoritária, com o intuído de balancear a amostra a partir da introdução de novas observações (*Oversample*). Com a adição de 9000 observações a categoria minoritária, a proporção entre as categorias passou para 1:1, eliminando os problemas de desbalanço entre as categorias. Já em (c), assim como em (b), a proporção entre as categorias situa-se em 1:1. O *Undersample* realizado em (c) reduziu a desproporção entre as categorias a partir da eliminação de observações da categoria majoritária.

Neste trabalho, adotamos a amostragem aleatória com repetição da categoria minoritária como *Oversample*, bem como tal método foi aplicado para corrigir o desbalanço entre os MEMR com não conformidades e sem não conformidades.

2.3.4. Regressão Logística

Na regressão logística, a variável dependente é binária e denota duas categorias mutualmente exclusivas, de modo que: $Y \epsilon \{0,1\}$. Assume-se que, para $i = 1, ..., n$ observações, $Y_i \sim$ Bernoulli $(Y_i \mid \pi_i)$. Nesse caso, o termo π_i equivale à:

$$\pi_i = \frac{1}{1 + e^{-X_i\beta}}. \tag{2.2}$$

Em que: β representa o vetor de coeficientes, de ordem $k \times 1$; X_i representa o vetor de variáveis independentes.

Logo, a partir da função de densidade de probabilidade da destruição de Bernoulli, define-se que: $\Pr(Y_i \mid \pi_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$. Assim, pressupondo independência das observações, a regressão logística pode ser estimada via máxima verossimilhança.

Ademais, a expressão (2.2) denota Pr $(Yi = 1 \mid X)$. Logo, utiliza-se (2.2) para computar as probabilidades estimadas[2].

2.3.4.1. Correção de King e Zeng (2001)

De acordo com King e Zeng (2001), a predição de probabilidades em (2.2) é afetada por dois problemas distintos, nos casos de amostras finitas com dados não balanceados: O viés dos

[2] Para mais detalhes sobre a estimação da regressão logística, ver Greene (2001).

parâmetros estimados e o modo como a probabilidade é calculada. Para a correção dos parâmetros, King e Zeng (2001) estimam o viés através da seguinte expressão[3]:

$$Vies(\hat{\beta}) = (\boldsymbol{X'WX})^{-1}(\boldsymbol{X'W\zeta}). \tag{2.3}$$

Em que: $\zeta_i = 0.5Q_{ii}[(1+c_1)\hat{\pi}_i - c_1]$; Q_{ii} são os elementos da diagonal da matriz $\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X'XWX})^{-1}\boldsymbol{X'}$; $\boldsymbol{W}$ = diagonal$\{\hat{\pi}_i(1-\hat{\pi}_i)w_i\}$; a variável c_1 é representada por: $c_1 = \tau / \gamma$; τ representa a proporção populacional entre as categorias; γ representa a proporção amostral entre as categorias.

Assim, obtêm-se os parâmetros corrigidos por meio de: $\tilde{\beta} = \hat{\beta} - Vi\acute{e}s(\hat{\beta})$. Da mesma forma, a variância dos parâmetros é dada por: $V(\tilde{\beta}) = [n/(n+k)^2]V(\hat{\beta})$. Uma vez que $[n/(n+k)^2] < 1$, tem-se: $V(\tilde{\beta}) < V(\hat{\beta})$. Portanto, a correção do viés também reduz a variância do estimador (KING; ZENG, 2001).

Para o cálculo das probabilidades, utiliza-se a expressão (2.4):

$$P(Y_i = 1) \approx \tilde{\pi}_i + (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)x_0 V(\tilde{\beta})x_0'. \tag{2.4}$$

Desse modo, o cálculo ajustado das probabilidades é composto da estimativa usual da probabilidade de ocorrência do evento *Y=1*, junto a um fator de correção.

2.3.5. *Random Forest*

Árvores de decisão são um método popular empregado em problemas de classificação e análise de regressão. A ideia central do método consiste em dividir um espaço de *d* dimensões em *M* regiões, de modo que o indivíduo *i* é classificado de acordo com os demais membros da região *m.* Para a sua estimação, adota-se um algoritmo recursivo de múltiplas e sucessivas divisões dos espaço entre 2 regiões, ou seja, a partir de um regressor X_1, a primeira divisão do espaço ocorre de modo a estabelecer o limite t_1 em que $X_1 > t_1 \in R_1$ (região 1) e $X_1 < t_1 \in R_2$ (região 2). Após a primeira divisão, o algoritmo aplica o mesmo princípio nas regiões resultantes, para os demais regressores, de modo a gerar no máximo 4 regiões. As sucessivas aplicações de divisão do espaço entre regiões podem ser interpretadas como uma árvore de decisão, conferindo assim o nome ao método. Nesse caso, cada divisão é denominada "nó".

[3] Detalhes sobre a aplicação da correção disponíveis em King e Zeng (2001).

Embora as árvores de decisão despontem como uma alternativa com baixo viés aos problemas de classificação e análise de regressão, sua alta variância produz estimativas com largos intervalos de confiança (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Assim, para melhora a performance das árvores de decisão Breiman (2001) propõe um método que estimação utilizando *bagging*, conhecido como *Random Forest*. De acordo Hastie, Tibshirani e Friedman (2008), o *Random Forest* é aplicado seguindo o seguinte algoritmo:

1 - Construa conjuntos amostrais Z_i com reposição do conjunto de dados original;

2- Para cada Z_i aplique o aplique o algoritmo de árvores de decisão, seguindo a seguinte regra para cada "nó":

2.1 – Do grupo de variáveis dependentes *p*, escolha aleatoriamente *m* candidatas para a divisão no "nó";

2.2- Escolha a melhor variável dentre as *m* candidatas;

2.3- Realize a divisão com base na variável escolhida;

3- Realize a previsão com base na média dos valores computados pelas árvores, para o caso da regressão e opção com mais votos, para o caso da classificação.

Logo, o *Random Forest* é composto *I* arvores derivadas dos conjuntos amostrais Z_i.

2.3.6. *Support Vector Machines* (SVM) e *Alternating Genetic Algorithm* SVM (ALGA - SVM)

2.3.6.1. Support Vector Machines

O *Support Vector Machine* (SVM) é um método comumente empregado em problemas de classificação. Diferentemente da regressão logística e *Random Forest*, o SVM enquadra-se nos problemas focados em obter o hiperplano que separa as categorias analisadas.

Assim, para uma determinada variável categórica $Y \in \{-1,1\}$ com duas categorias, pode-se utilizar um conjunto de características (X_i) para predizer os valores futuros de Y, assim, a partir dos pares (X_i, Y_i), $i = 1, 2, \dots, N$ onde $X_i \in R^n$, define-se o problema de estimação do hiperplano a partir de:

$$\{x: f(x) = x^T\beta + \beta_0 = 0\}. \tag{2.5}$$

Em que: $f(x)$ denota a função do hiperplano; x^T denota o conjunto de variáveis independentes do indivíduo; β denota os parâmetros do hiperplano.

Associado ao classificador:

$$Class(x) = Sing[f(x)]. \tag{2.6}$$

Em que: $Class(x)$ denota a função de classificação; $Sing(x)$ denota a função que atribui a classificação a partir do sinal de $f(x)$; $f(x)$ denota a função do hiperplano.

Nesse quesito, Hastie, Tibshirani e Friedman (2008) denotam que o problema exposto em (2.5) não possui solução única e ótima, sem a adição de algumas restrições ao problema de estimação. Nesse caso, como define Vapnik (1966), o hiperplano ótimo é o hiperplano capaz de separar as categorias além de produzir a maior distância entre as observações das categorias.

Desse modo, o problema de estimação pode ser traduzido no seguinte ´problema de minimização:

$$\begin{gathered} \min_{\beta,\beta_0} \|\beta\| \\ sujeito\ a\ y_i(x_i^T\beta + \beta_0) \geq 1, i = 1, \dots, N. \end{gathered} \tag{2.7}$$

Em que: β denota os parâmetros da função do hiperplano; y_i denota a variável dependente do indivíduo i; x_i^t denota o conjunto de variáveis independentes do indivíduo i.

O problema exposto em (2.7) exige que exista separabilidade nos dados em relação a variável objetivo, ou seja, o hiperplano estimado deve dividir totalmente ambas as categorias, sem que nenhuma observação esteja na área contrária à sua classificação. Entretanto, em diversos problemas, não há separabilidade nos dados. Assim, para esses casos é necessário que se adote algumas medidas para seguir com a estimação do hiperplano. Uma das medidas adotadas é permitir que existam observações na zona errada da margem do hiperplano e atribuíam-se restrições ao erro de classificação dessas observações. Assim, o problema de minimização pode ser rescrito como:

$$\begin{gathered} \min_{\beta,\beta_0} \|\beta\| \\ sujeito\ a: \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \delta_i, i = 1, \dots, N \\ \delta_i \geq 0, \sum \delta_i \leq C \end{cases}. \end{gathered} \tag{2.8}$$

Em que: β denota os parâmetros da função do hiperplano; y_i denota a variável dependente do indivíduo i; x_i^t denota o conjunto de variáveis independentes do indivíduo i; δ_i denota a distância entre as observações no lado oposto ao grupo que pertencem e a margem do lado correto; C denota uma constante exógena a estimação.

Logo, o problema exposto em (2.8) permite que se determine o nível de erro na estimação do SVM, uma vez que o valor de C é determinado fora do problema de otimização. Ademais, a equação (2.8) exibe um problema de otimização com restrições. Assim, para solucionar o problema, recorre-se aos multiplicadores de Lagrange. Para tanto, é conveniente

em termos computacionais rescrever o problema como (HASTIE; TIBSHIRANI; FRIEDMAN, 2008):

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\delta_i$$
$$sujeito\ a: \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \delta_i, i = 1, \dots, N \\ \delta_i \geq 0 \end{cases}. \tag{2.9}$$

Em que: β denota os parâmetros da função do hiperplano; y_i denota a variável dependente do indivíduo i; x_i^t denota o conjunto de variáveis independentes do indivíduo i; δ_i denota a distância entre a observação i e a margem da área em que a observação pertence (δ_i = 0 caso a observação esteja no lado correto do hiperplano); C denota uma constante exógena a estimação.

Assim, o Lagrangeano de (2.9) é definido por:

$$L_P = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\delta_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \delta_i)] - \sum_{i=1}^{N}\mu_i\delta_i. \tag{2.10}$$

Em que: β denota os parâmetros da função do hiperplano; C denota uma constante exógena a estimação; δ_i denota a distância entre a observação i e a margem da área em que a observação pertence (δ_i = 0 caso a observação esteja no lado correto do hiperplano); α_i denota o multiplicador de Lagrange associado a primeira restrição; y_i denota a variável dependente do indivíduo i; x_i^t denota o conjunto de variáveis independentes do indivíduo i; μ_i denota o multiplicador de Lagrange associado a segunda restrição.

A partir da substituição das condições de primeira ordem[4] em (2.10), temos:

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\,\alpha_{i'}y_iy_{i'}x_i^Tx_{i'}. \tag{2.11}$$

Nesse caso, (2.11) é um problema de otimização em uma função concava e sua otimização a partir do algoritmo apropriado levará a solução do SVM.

2.3.6.2. *SVM* e funções de *Kernel*

Em (2.11), o processo descrito gerará a estimação de hiperplanos lineares. Entretanto, Hastie, Tibshirani e Friedman (2008) descrevem que é possível flexibilizar tal resultado, a partir do alargamento do espaço de característica original, através de expansões de base. Ou seja,

[4] Condições de primeira ordem refere-se a tomar a primeiras derivadas e igualá-las a 0.

pretende-se levar o problema de classificação para uma dimensão mais alta. Essa manobra permite a flexibilização da linearidade do hiperplano, e, por consequência, o melhor desempenho do SVM em problemas com baixa separabilidade das categorias. Assim, a partir da seleção da função de base adotada $h_m(x)$ em que $m = 1, \dots, M$, temos que as variáveis independentes podem ser expressas como: $h(x_i) = (h_1(x_1), h_2(x_2), \dots, h_M(x_i))$, $i = 1, \dots, N$. Com isso, a estimação do hiperplano pode ser expressa como:

$$\{x: f(x) = h(x)^T\beta + \beta_0 = 0\}. \tag{2.12}$$

Em que: $f(x)$ denota a função do hiperplano; $h(x^T)$; $h(.)$ denota a função para expansão da base; x denota o conjunto de variáveis independentes do indivíduo; β denota os parâmetros do hiperplano.

Associado ao classificador:

$$G(x) = Sing[f(x)]. \tag{2.13}$$

Em que: $G(x)$ denota a função de classificação; $Sing(x)$ denota a função que atribui a classificação a partir do sinal de $f(x)$; $f(x)$ denota a função do hiperplano.

Para esse caso, (2.11) pode ser expresso como[5]:

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} \alpha_i\, \alpha_{i'} y_i y_{i'} \langle h(x_i), h(x_{i'})\rangle. \tag{2.14}$$

Assim, a expressão (2.14) envolve o produto interno entre as funções de base. Assim, para realizar a estimação de (2.14), é necessário que se especifique a função de *Kernel* equivalente ao produto interno entre as funções de base. Como expresso em (2.15):

$$K(x, x') = \langle h(x), h(x')\rangle. \tag{2.15}$$

Em que: $K(x, x')$ denota a função de *Kernel*; $h(x)$ denota a função para expansão da base; x denota o conjunto de variáveis independentes do indivíduo.

Dentre as opções de função de *Kernel* comumente utilizadas na literatura, utilizamos a função de Base Radial, que pode ser expressa como:

$$K(x, x') = e^{\left(-\gamma\|x - x'\|^2\right)}. \tag{2.16}$$

[5] Assim como em , Hastie, Tibshirani e Friedman (2008), (2.14) foi expresso na versão com produto interno entre as variáveis independentes na função de base adotada.

Em que: $K(x, x')$ denota a função de *Kernel*; γ denota uma constante exógena; x denota o conjunto de variáveis independentes do indivíduo.

2.3.6.3. Complexidade e Tamanho da Amostra no SVM

De acordo com Nalepa e Kawulok (2018), o SVM possui funções de complexidade em relação ao tempo de estimativa e memória requerida de, respectivamente, $O(t^3)$ e $O(t^2)$, em que t representa o número de observações na amostra. Desse modo, tanto o tempo de estimação, quanto o espaço a memória computacional elevam-se exponencialmente, conforme o tamanho amostral, impossibilitando o uso do SVM em aplicações em grandes conjuntos de dados.

Dentre os métodos propostos para corrigir tal problemática, Nalepa e Kawulok (2018) os classificam em duas abordagens: Técnicas voltadas a elevar a velocidade de estimação do SVM e técnicas que reduzem o tamanho do conjunto de treino. Segundo os autores, as técnicas voltadas a redução do conjunto de treino podem abrangem: Análise das características do conjunto de teste, agrupamentos com base na vizinhança, algoritmos evolucionários, e aprendizagem ativa e amostragem aleatória do conjunto de teste. Quanto aos métodos voltados a elevar a velocidade de estimação, esses baseiam-se em técnicas voltadas a reduzir a complexidade ou aprimorar alguns aspectos do processo de estimação. Entretanto, tais métodos ainda sofrem com o elevado uso de capacidade computacional (NALEPA; KAWULOK, 2018).

Embora as técnicas voltadas a reduzir o tamanho conjunto de dados usados para o treino do SVM sejam efetivas na redução da complexidade do algoritmo, a exclusão de algumas observações pode afetar a efetividade do SVM. Dado que ao retirar algumas observações do conjunto de dados, pode-se alterar o hiperplano ótimo (KAWULOK; NALEPA; DUDZIK, 2017). A alteração decorre da relação do método para com as observações. No SVM, apenas algumas observações são importantes para a determinação do hiperplano, de modo que a exclusão de observações com menor importância não deverá afetar a efetividade da técnica (KAWULOK; NALEPA, 2012).

Por outro lado, Kawulok, Nalepa e Dudzik (2017) argumentam que o hiperplano ótimo, e sua estimação, altera-se conforme alteram-se os hiperparâmetros do SVM. Ou seja, para cada possível valor dos hiperparâmetros de uma determinada função de kernel empregada em um SVM, há um hiperplano ótimo. Logo, para cada possível valor dos hiperparâmetros, há um conjunto de observações ideal para a estimação do hiperplano. Assim, os autores sugerem que a busca pelo hiperparâmetro ótimo, bem como o conjunto de observações a ser usado na

estimação ocorra simultaneamente. Para tanto, Kawulok, Nalepa e Dudzik (2017) sugerem o uso de um algoritmo genético, denominado *Alternating Genetic Algorithm* (ALGA).

2.3.6.4. ALGA – SVM

O algoritmo ALGA, de autoria de Kawulok, Nalepa e Dudzik (2017), tem como principal objetivo a seleção simultânea do conjunto de treino e hiperparâmetros de um modelo SVM. Para tanto, o algoritmo baseia-se na aplicação de um algoritmo genético em um processo interativo entre a seleção das observações mais relevantes a estimação do hiperplano e seleção dos melhores hiperparâmetros. Assim, a otimização do conjunto de treino e hiperparâmetros ocorre da seguinte forma: A partir de um conjunto de hiperparâmetros, selecionados aleatoriamente, testa-se o desempenho de diversas sub amostras do conjunto inicial de dados. A partir dos resultados do teste, as piores sub amostras passam por transformações (combinações e mutações) enquanto as com melhor desempenho são mantidas, e todas são testadas novamente. Esse processo perdura até que uma sub amostra perdure por duas rodadas de teste. Em seguida, inverte-se o processo e a sub amostra com o melhor desempenho é utilizada para testar diversos valores do hiperparâmetro. Desse processo, seleciona-se o melhor hiperparâmetro e as sub amostras são submetidas novamente a otimização. Esse processo de otimização ocorre até que se encontre os melhores hiperparâmetros e sub amostra.

Inicialmente, para operacionalizar o ALGA, define-se R sub amostras e U valores candidatos a hiperparâmetro. Assim, temos p_i i= 1, 2, 3, ..., R sub amostras e q_i i= 1, 2, 3, ..., U hiperparâmetros, em que $p_i \subset \boldsymbol{T'}$ e $q_i \subset M'$. As sub amostras p_i são formadas por K observações de cada categoria. Desse modo, para o caso de um modelo de classificação entre duas categorias dicotômicas, cada sub amostra p_i possuirá $2K$ observações, escolhidas aleatoriamente[6]. Já os q_i hiperparâmetros são formados por conjuntos de possíveis valores para os *inputs* da função de *Kernel*. Ademais, os testes dentro do algoritmo genético são realizados em um conjunto de dados selecionado previamente a aplicação do algoritmo, de modo que as sub amostras que compõem o conjunto de teste não compartilham observações com o conjunto de dados voltados para treino.

O Pseudocódigo para a execução do algoritmo é demonstrado em Algoritmo 1.

[6] Uma vez que cada sub amostra *p*i possui *K* observações de cada categoria, não há desbalanceamento entre as frequências das categorias. Assim, o algoritmo ALGA-SVM já corrige os problemas de desbalanceamento na composição das sub amostras de teste.

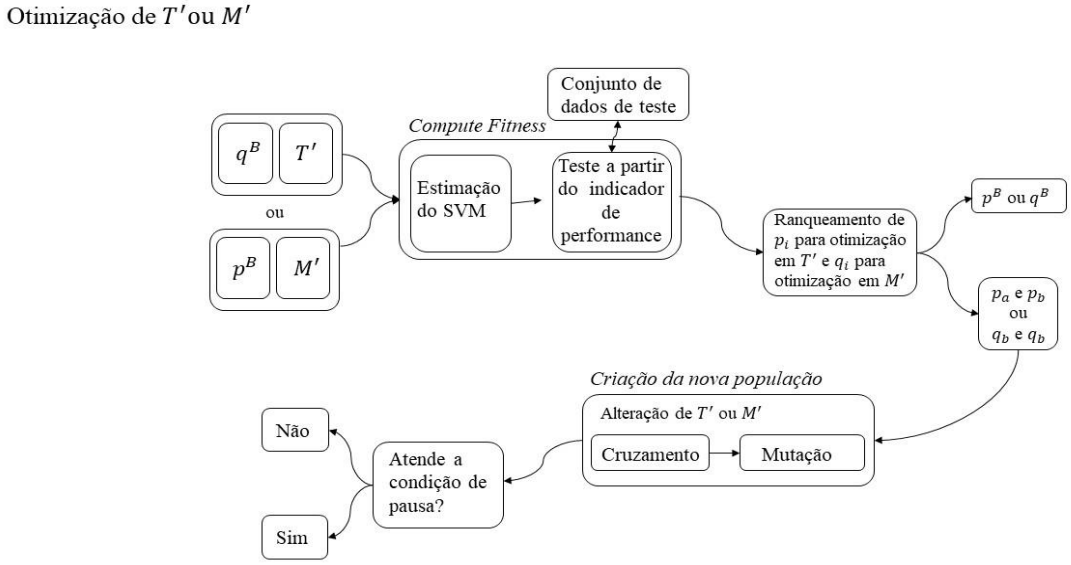
Algoritmo 1. *Alternating Genetic Algorithm* (ALGA)

1: A partir de uma população ***T'*** ($\{p_i\}$) de tamanho N;
2: A partir de uma população ***M'*** ($\{q_i\}$) de tamanho M;
3: $q_{init} \leftarrow$ GETRANDON ($\{q_i\}$);
4: **for all** ($\{p_i\}$) do
5: $\eta_i \leftarrow$ COMPUTEFITNESS(p_i, q_{init});
6: **end for**
7: Selecione p^B – individuo com o melhor indicador de performance em $\{p_i\}$;
8: **repeat**
9: **repeat**
10: **for all** ($\{q_i\}$) **do**
11: $\eta_i \leftarrow$ COMPUTEFITNESS(p^B, q_i);
12: **end for**
13: Selecione q^B – individuo com o melhor indicador de performance em $\{q_i\}$;
14: Crie uma nova população $\{q_i\}$;
15: **until** LOCALSTOP;
16: **if** ¬ GLOBALSTOP **then**
17: **repeat**
18: **for all** ($\{p_i\}$) **do**
19: $\eta_i \leftarrow$ COMPUTEFITNESS(p_i, q^B);
20: **end for**
21: Selecione p^B – individuo com o melhor indicador de performance em $\{p_i\}$;
22: Crie uma nova população $\{p_i\}$;
23: **until** LOCALSTOP;
24: **end if**
25: **until** GLOBALSTOP;
26: **return** (p^B, q^B);

Fonte: Modificado de Kawulok, Nalepa e Dudzik (2017).

Quanto as etapas de criação de novos conjuntos de dados e hiperparâmetros, essas ocorrem junto ao passo 11 e 19 do pseudocódigo. Ademais, há três processos distintos aplicados em meio a essas etapas: Mutação, Cruzamento e Seleção. Esses processos ocorrem em meio a aplicação do algoritmo, tanto nas etapas de otimização do hiperparametro e sub amostra, assim como o ilustrado na Figura 2.3

Figura 2.3: Processo de Otimização de *T'* Ou *M'*.



Fonte: Elaboração própria com base em em Kawulok, Nalepa e Dudzik (2017).

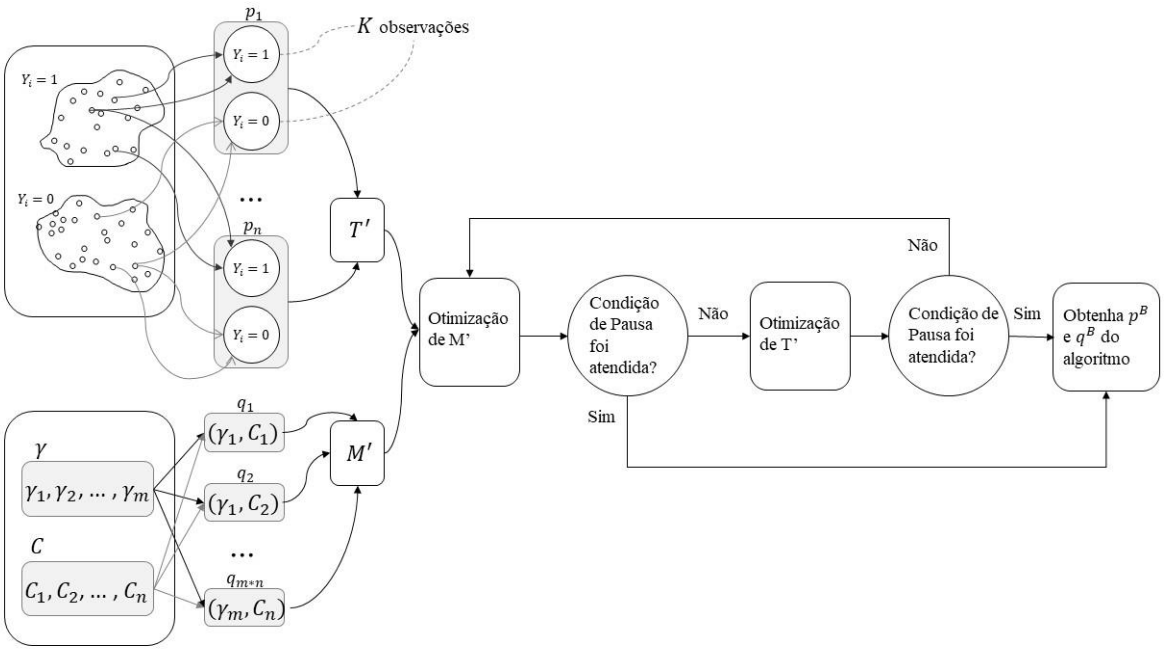
A partir da Figura 2.3, observa-se que o processo de *Compute Fitness* abrange a estimação do SVM, bem como o teste de performance das estimações com os candidatos da para amostra e hiperparâmetro. Após as estimações, ocorre o ranqueamento dos candidatos, bem como a divisão entre os grupos A e B para o cruzamento.

O cruzamento é descrito como o processo entre a junção de dois candidatos. Para a aplicação do ALGA SVM, Kawulok, Nalepa e Dudzik (2017) definem que o cruzamento ocorrerá entre os indivíduos com melhor e pior desempenho na predição do conjunto de testes, a partir do resultado do indicador de performance selecionado. Assim, os autores definem a divisão dos indivíduos em dois grupos (Grupo A e Grupo B), sendo que o grupo é composto A é composto pelos indivíduos com os melhores indicadores de seleção, enquanto o grupo *B* é composto pelos indivíduos com os piores. Desse modo, os indivíduos do grupo A e grupo B cruzam-se com uma determinada probabilidade (P_c^M para o grupo de candidatos a hiperparametros). Para as sub amostras, a reprodução ocorre via seleção aleatória das observações de ambos os parentes, de modo a respeitar *K* observações por categoria. Para o caso dos hiperparametros, os descendentes são gerados a partir da seguinte expressão: $x_{a+b} = x_a + \alpha_m.(x_a - x_b)$. Em que o parâmetro xa e xb são parâmetros de interesse de indivíduos do grupo *A* e *B*, respectivamente, enquanto α_m é uma variável escolhida aleatoriamente [-0.5,1.5] e definisse-se o resultado dessa expressão como não negativo ($x_{a+b} > 0$). Assim como no cruzamento, a mutação ocorre a partir de uma determinada probabilidade (P_m^M para o grupo de candidatos a hiperparametros) e

consiste em selecionar aleatoriamente os valores da variável de interesse, a partir do intervalo $x \in [x - \delta_m . x,\ x + \delta_m . x]$. Em que δ_m denota um parâmetro utilizado para definir o tamanho do intervalo da variável de interesse na mutação.

Desse modo, o algoritmo ALGA-SVM pode ser resumido no exposto na Figura 2.4.

Figura 2.4: Composição das sub-amostras e conjuntos de hiperparâmetros na aplicação do ALGA-SVM com função de *Kernel* RBF e operacionalização do algoritmo.



Fonte: Elaboração própria com base em em Kawulok, Nalepa e Dudzik (2017).

De acordo com a Figura 2.4, o ALGA-SVM em um algoritmo voltado a otimização simultânea de dois aspectos do SVM: A amostra e os hiperparâmetros. Após a formação de *T'* e *M'*, o algoritmo entra em um processo de otimização simultânea, mas particionada, de ambos os objetos.

2.3.7. Critérios de Avaliação e Seleção dos modelos

Para o caso de classificação, os métodos devem ser capazes de discriminar corretamente os indivíduos (KRUPPA *et al.*, 2013). Da mesma forma, os métodos voltados a estimação das probabilidades devem ser capazes de prever corretamente a ocorrência dos eventos. Na literatura empírica, um dos modos utilizados para a classificação das predições dos modelos

são as estatísticas e indicadores derivados da matriz de confusão (NEHRECECKA, 2018). A Tabela 2.2 apresenta o esquema básico da matriz de confusão para 2 eventos:

Tabela 2.2: Matriz de Confusão para 2 eventos.

		Evento Predito	
		$\hat{Y} = 1$	$\hat{Y} = 0$
Evento Observado	Y=1	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Y=0	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Elaboração própria.

Assim, a matriz de confusão classifica as observações em quatro categorias. As categorias Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN) denotam as observações que o método previu corretamente. Já as categorias Falso Negativo (FN) e Falso Positivo (FP) denotam os erros de classificação do método. No contexto geral, os erros de classificação possuem implicações diferentes, conforme o objeto de estudo.

Para o caso do presente artigo, a ocorrência de Falso Positivo implica em embalagens de madeira, sem não conformidades, que serão fiscalizadas. O resultado é um custo desnecessário ao órgão de fiscalização e um custo indireto para o importador, dado que a fiscalização faz com que o tempo médio de permanência das cargas importadas no aeroporto aumente. Ainda, assim, estes custos são mais baixos em comparação a ocorrência de Falsos Negativo. O Falso Negativo implica em embalagens com não conformidade que não serão fiscalizadas e, por consequência, no aumento da probabilidade de introdução de pragas no ecossistema de uma determinada região. Logo, a ocorrência de Falsos Negativos produz, em geral, custos bem maiores para a sociedade do que a ocorrência de Falsos Positivos, no problema de fiscalização de importações.

Ademais, a classificação disposta na matriz de confusão serve de base para o cálculo de indicadores e estatísticas que mensuram a performance dos modelos. Dentre os possíveis indicadores, os seguintes foram usados para a comparação entre os métodos: Acurácia, Sensibilidade, Especificidade, Taxa de Falso Negativo, Taxa de Falso Positivo e *F-Score*. Os indicadores podem ser calculados pelas seguintes expressões:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}; \tag{2.17}$$

$$Sensibilidade = \frac{VP}{VP + FN}; \tag{2.18}$$

$$Especificidade = \frac{VN}{VN + FP}; \tag{2.19}$$

$$Taxa\ de\ Falso\ Negativo = \frac{FN}{VP + FN}; \tag{2.20}$$

$$Taxa\ de\ Falso\ Positivo = \frac{FP}{FP + VN}; \tag{2.21}$$

$$F - Score = \frac{VP}{VP + \frac{1}{2}(FP + FN)}. \tag{2.22}$$

Embora a matriz de confusão apresente um panorama geral entre as proporções de erros e acertos na predição, métodos como a Regressão Logística não projetam a ocorrência do evento, mas sim a sua probabilidade de ocorrência. Nesse caso, é comum especificar um limiar de probabilidades previstas, pelo qual valores acima do limiar denotam a ocorrência do evento, e valores abaixo à não ocorrência do evento. Tal procedimento ainda pode ser estendido para a proporção de votos em uma função de voto. Neste trabalho os limiares de cada método são definidos pelo desempenho da acurácia em conjunto da taxa de falso negativo, dado a importância desses dois indicadores para o problema a ser resolvido.

Ademais, há diversas formas adotadas na literatura para definir o limiar ótimo. Dentre elas, destaca-se o uso da curva "Característica de Operação do Receptor" (ROC), dada sua capacidade de escolha do limiar e avaliação do desempenho dos métodos empregados.

Em termos gerais, a curva ROC mede a relação entre a Sensibilidade e 1-Espeficidade (proporção de Falsos Positivos) para cada ponto de quebra (limiar) das probabilidades. Assim, quanto mais próximo do canto esquerdo superior do plano cartesiano à curva ROC, de um determinado método, está, maior é a Sensibilidade junto a baixos valores da proporção de Falsos Positivos. Sendo assim, um dos critérios para a seleção de limiares é a distância dos pontos da curva para com o canto superior esquerdo, de modo que se escolhe o limiar com a menor medida de distância[7]. Ademais, a curva ROC ainda é base para outra medida empregada na literatura empírica na avaliação da performance de métodos: A Área Abaixo de Curva ROC (AUC). A AUC é calculada a partir da área de baixo da curva ROC, assim valores da AUC superiores a 0,5 indicam boa performance do método.

Além da tradicional divisão do conjunto de dados entre Conjunto de Teste e Conjunto de Treino, este trabalho segue Rocha (2013) e simula o desempenho dos modelos na fiscalização desde o início de um exercício. Assim, pretende-se demonstrar o real impacto da

[7] Comumente emprega-se a distância euclidiana entre os pontos da curva ROC e o ponto (Sensibilidade, 1-Especificidade) = (1,0)

aplicação dos modelos, caso fossem utilizados no sistema de seleção de MEMR a serem fiscalizados. Portanto, a amostra de treino é composta dos registros dos MEMR referentes aos anos de 2017 a setembro de 2018. A simulação dos resultados com a adoção dos modelos no sistema foi realizada para os três últimos meses de 2018 (Outubro, Novembro e Dezembro) e três primeiros meses de 2019 (Janeiro, Fevereiro e Março).

2.3.8. Validação Cruzada

De acordo com Hastie, Tibshirani e Friedman (2008), a validação cruzada é um método comumente empregado para se estimar o erro de previsão, dado as propriedades dessa técnica. Assim, costuma-se empregá-la na escolha dos hiperparâmetros, como o valor da constante C no caso do SVM ou o número de variáveis por sorteio, no caso do *Random Forest*.

Em termos gerais, a validação cruzada é empregada da seguinte forma: Divide-se a amostra em k partes. A partir da divisão em k partes, k-1 são empregadas na estimação do modelo, enquanto a parte não empregada na estimação é utilizada para medir a performance do modelo, a partir dos indicadores de performance selecionados. Após esse procedimento, retira-se uma das k-1 partições da amostra utilizada para a estimação e a partição utilizada para teste é introduzida no conjunto de dados, enquanto a partição retirada é alocada para o treino. Esse processo segue até que as k partições sejam empregadas para o teste de performance do modelo.

A Figura 2.5 ilustra a validação cruzada com a divisão da amostra em 5 partes.

Figura 2.5: Validação cruzada com divisão da amostra em 5 partes.

	(1)	(2)	(3)	(4)	(5)
Divisão 1	Teste	Treino	Treino	Treino	Treino
Divisão 2	Treino	Teste	Treino	Treino	Treino
Divisão 3	Treino	Treino	Teste	Treino	Treino
Divisão 4	Treino	Treino	Treino	Teste	Treino
Divisão 5	Treino	Treino	Treino	Treino	Teste

Fonte: Modificado de Malekipirbazari e Aksakalli (2015).

Assim como exposto na Figura 2.5, a validação cruzada com a divisão da amostra em 5 partes prove a estimação de 5 modelos com as *k*-1 partes da amostra empregadas não empregadas para teste da divisão de amostra. Ao término das 5 estimações e 5 testes, os resultados são agrupados ou calcula-se a média do indicador, de acordo com as particularidades do indicador de performance escolhido.

2.3.9. Fonte de Dados e Variáveis

Como já citado anteriormente, a variável dependente das metodologias aplicadas para o presente problema foi a presença de não conformidades (Y_i=1) e a ausência de não conformidades (Y_i=0). Já as variáveis independentes foram escolhidas e construídas com base na NIMF-23 e disponibilidade de acesso. A Tabela 2.3 apresenta os preditores selecionados.

Tabela 2.3: Variáveis independentes dos modelos estimados para previsão da não conformidade em PCG com MEMR a serem inspecionados

Grupo	Subgrupo	Nome da Variável/ Característica ou Tipo ou Natureza ou classificação
Variáveis do Importador		Setor do Importador (Categórica)[8]
		Anos de Experiência do Importador (Contínua)
Variáveis de Origem		Classificação do Aeroporto de Origem da Importação (Categórica)
		Continente de Origem da Importação (Categórica)
		Classificação do Nível de Renda da Origem da Importação (Categórica)[9]
		Adoção dos Procedimentos do NIMF-15 pelo exportador (Categórica)
		Mês do Ano em que ocorreu a Importação (Categórica)
Carga Transportada		Quantidade de PCG's presentes no mesmo voo (Contínua)
Histórico de Não Conformidades	Importador	Taxa de Não Conformidades do Importador, do Ano Anterior (Contínua)
		Quantidade de PCG importados pelo Importador no Ano Anterior (Contínua)
		Taxa de Não Conformidades do Importador, no mesmo mês do Ano Anterior (Contínua)
	Aeroporto de Origem	Taxa de Não Conformidades do Aeroporto de Origem, do Ano Anterior (Contínua)
		Quantidade de PCG importados do Aeroporto de Origem no ano anterior (Contínua)
		Taxa de Não Conformidades do Aeroporto de Origem, no mesmo mês do Ano Anterior (Contínua)

Fonte: Elaboração própria.

Logo, os métodos empregados contêm variáveis relacionadas às: características do importador, características de origem, sazonalidade e histórico de não conformidades.

O conjunto de dados é composto pelos registros de todas as importações que ocorreram no Aeroporto Internacional de Viracopos, de Campinas/SP, entre 2016 e 2019. Assim como em Nehrececka (2018), a base de dados utilizada na estimação dos métodos foi dividida entre: Conjunto de "Treino" e Conjunto de "Teste". O conjunto de treino abrange os anos de 2016 a setembro de 2018. Já o conjunto de teste abrange o período de outubro de 2018 a março de 2019. Ademais, a fim de melhorar o desempenho dos estimadores e selecionar corretamente os hiperparametros dos modelos, optamos por adotar a validação cruzada nos processos de seleção dos hiperparametros.

Quanto a fonte dos dados: Os dados referentes às não conformidades nas embalagens de madeira, volume de cargas importadas e volume de cargas acometidas a inspeção foram cedidos via acordo de parceria entre o Aeroporto Internacional de Viracopos e Serviço de

8 Os setores foram agrupados de acordo com o tipo de importação, a partir da classificação CNAE. A agregação está disponível no Apêndice A.

9 Classificação de acordo com World Bank (2018).

Vigilância Agropecuária Internacional (Vigiagro) do Ministério da Agricultura, Pecuária e de Abastecimento e o Cepea-Fealq.

Os dados referentes aos países participantes do NIMF-15 foram obtidos da *Food and Agriculture Organization* (FAO); enquanto a classificação das nações quanto ao nível de Renda está disponível no Banco Mundial (WORLD BANK, 2018). Ademais, os dados referentes a classificação dos aeroportos foram obtidos do projeto *OurAirports*. Por fim, os dados referentes ao setor e anos de experiência do importador foram obtidos a partir da receita federal. Os anos de experiência foram calculados com base na diferença do ano de importação e a data de criação da empresa, enquanto os setores foram obtidos e agrupados a partir do CNAE declarado pela empresa[10].

Para a composição da base de dados, foram excluídos os registros dos PCG cuja origem não é compatível com a codificação *International Air Transport Association* (IATA) para aeroportos. Da mesma forma foram excluídos os registros de importação que não apresentaram informações compatíveis com o CNPJ para as empresas.

Ademais, todas as variáveis empregadas na composição da base de dados são facilmente acessáveis nos respectivos *websites* das organizações referidas. Um dos objetivos iniciais da modelagem construída versa sobre a empregabilidade das metodologias aplicadas em um sistema de seleção das cargas com MEMR a serem inspecionadas.

2.4. Resultados e Discussão

Esta seção apresenta os resultados das estimações dos modelos estatísticos propostos na seção anterior. As estimativas, bem como todo o tratamento de dados, foram realizados no *software R versão 3.6.3*.

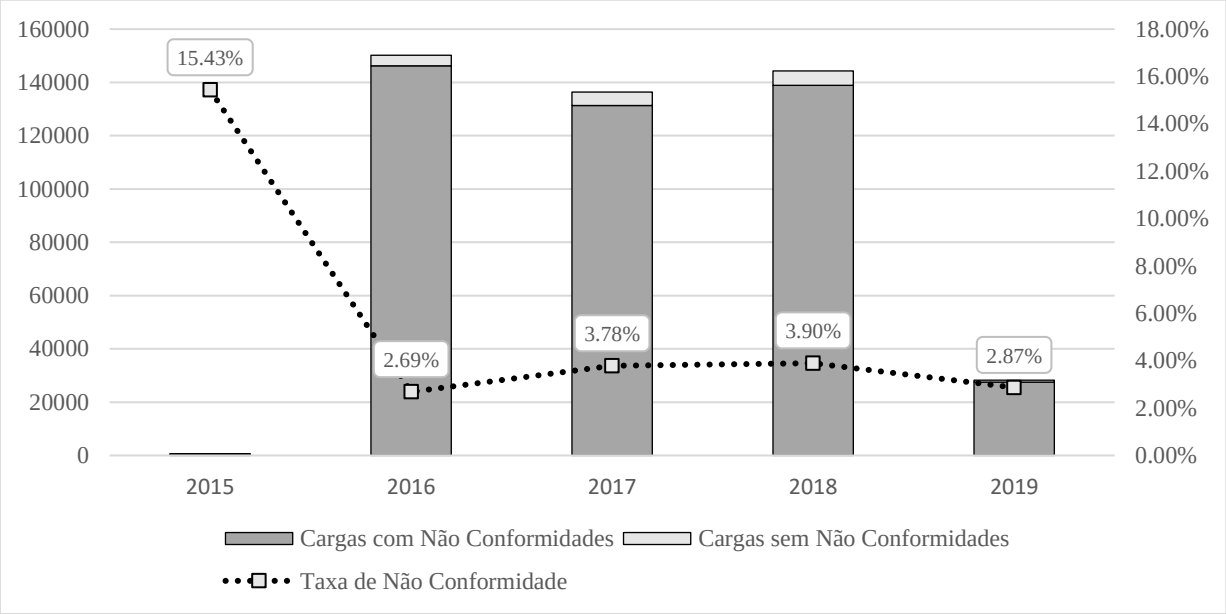
2.4.1. Resumo Estatístico

Essa subseção tem como objetivo demonstrar o comportamento das taxas de não conformidade dos MEMR conforme: variáveis de origem da carga, características do importador, sazonalidade e histórico de transação de MEMR. No período de dezembro de 2015 a março de 2019[11].

[10] A agregação está disponível no Apêndice.
[11] Os dados contidos nessa seção estão de acordo com os processos de limpeza nos dados operacionalizados para a execução dos modelos.

A Figura 2.6 demonstra o volume de PCG importados na última quinzena de 2015, 2016, 2017, 2018 e de janeiro a março de 2019.

Figura 2.6: Número de PCG com MEMR importados com não conformidades (NC), sem NC e Taxa de NC por ano, via Aeroporto de Viracopos – Campinas/SPde dezembro de 2015 a março de 2019

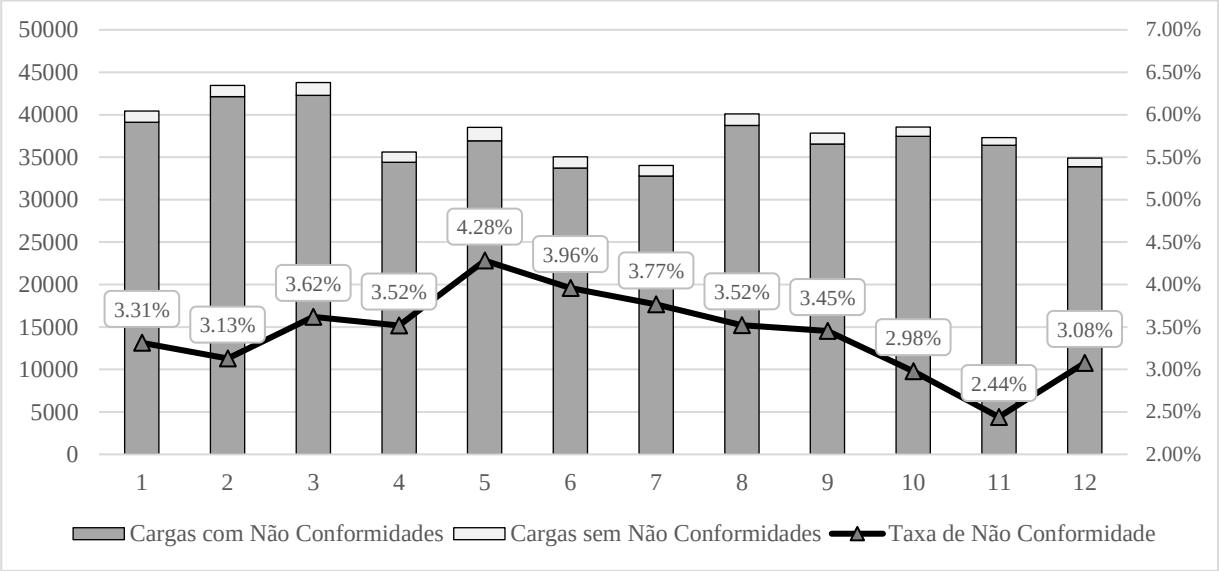


Fonte: Elaboração própria com base em dados fornecidos pela ABV

Entre 2016 e 2017, as taxas de não conformidade das embalagens de madeira situaram-se próximas a 3%, limite estabelecido pelo Vigiagro para a inclusão das empresas importadoras ao programa de fiscalização reduzida. Ademais, o volume de MEMR movimentados anualmente no Aeroporto Internacional de Viracopos, durante 2016 a 2018, manteve-se em próximo a 140 mil unidades.

A sazonalidade em importações pode ser verificada quando se avaliam os volumes de cargas importados, inclusive, por natureza de produtos, por origem de aquisição, e em função de variáveis não somente relacionadas à oferta, mas também à demanda. Logo, a inclusão desta variável no presente estudo visa a identificar se a sazonalidade em termos de tipos de cargas e origens pode também se refletir em variações nas taxas de não conformidade ao longo dos distintos períodos do ano. A Figura 2.7 apresenta a distribuição do volume de PCG em MEMR ao longo dos meses durante o período de dezembro de 2015 a março de 2019.

Figura 2.7: Número de PCG com MEMR importados com NC, sem NC e Taxa de NC por mês, , via Aeroporto de Viracopos – Campinas/SP, de dezembro de 2015 a março de 2019.



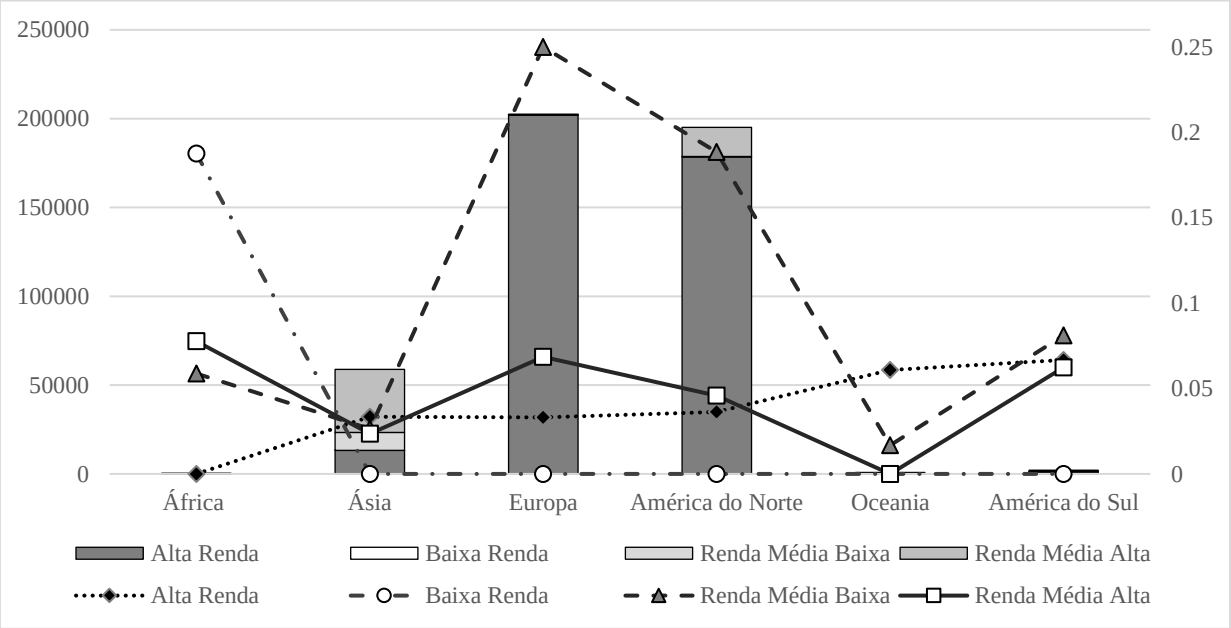
Fonte: Elaboração própria, com base em dados fornecidos pela ABV.

A Figura 2.7 relata o volume de PCG importados ao longo dos meses, entre o período de dezembro de 2015 a março de 2019. Nota-se que nesse período, o mês de maio apresentou a maior taxa de não conformidade, seguido pelo mês de junho e julho, que coincide com o período de redução na média de MEMR importados. Já outubro, novembro e dezembro apresentaram o menor índice de MEMR com registro de não conformidades.

Logo, assim como o relatado em Caton, Thomas e Brodel (2006); Szyniszewska *et al.* (2016) e Surkov (2007), o volume de MEMR não conformes não é uniforme ao longo do ano. Ademais, aparentemente os maiores índices de não conformidade ocorrem ao final do 1° semestre e início do 2° semestre do ano, enquanto os menores indicies ocorrem ao final do 2° semestre.

Outro fator relevante a análise das características discriminantes da não conformidade em embalagens é a origem das cargas. Assim, a Figura 2.8 e Figura 2.9 demonstram a distribuição do volume de cargas por nível de renda e classificação do aeroporto de origem.

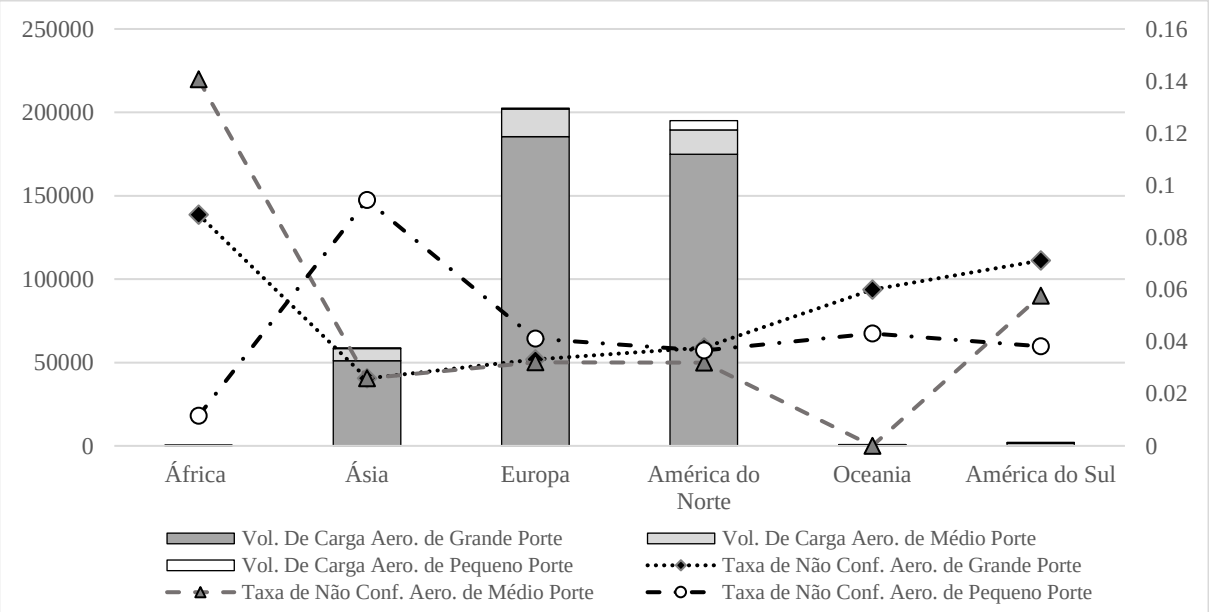
Figura 2.8 Número de PCG com MEMR importados e Taxa de NC conforme a classificação do país de origem por renda, de dezembro de 2015 a março de 2019.



Fonte: Elaboração própria, com base em dados fornecidos pela ABV.

A partir da Figura 2.8, nota-se que, no período analisado, grande parte dos PCG acondicionadas em MEMR importadas via Aeroporto Internacional de Viracopos tiveram como origem os países do continente Europeu e da América do Norte. Em ambas as regiões, grande parte das importações são provenientes de países com alto nível de renda. Entretanto, nota-se que, em ambas as regiões, os maiores níveis de não conformidade nas MEMR foram registradas nas cargas originadas em países de renda "Média-Baixa", enquanto o índice de não conformidade das MEMR em cargas vindas de países de alta renda situa-se abaixo de 5%. Já no caso da Ásia, terceira maior fonte dos PCG importados, grande parte das importações provieram de países com renda "Média-Alta", enquanto os índices de não conformidade situam-se em torno de 3%.

Figura 2.9 Número de PCG com MEMR importados via Aeroporto de Viracopos e Taxa de NC conforme a classificação do aeroporto de origem, de dezembro de 2015 a março de 2019.



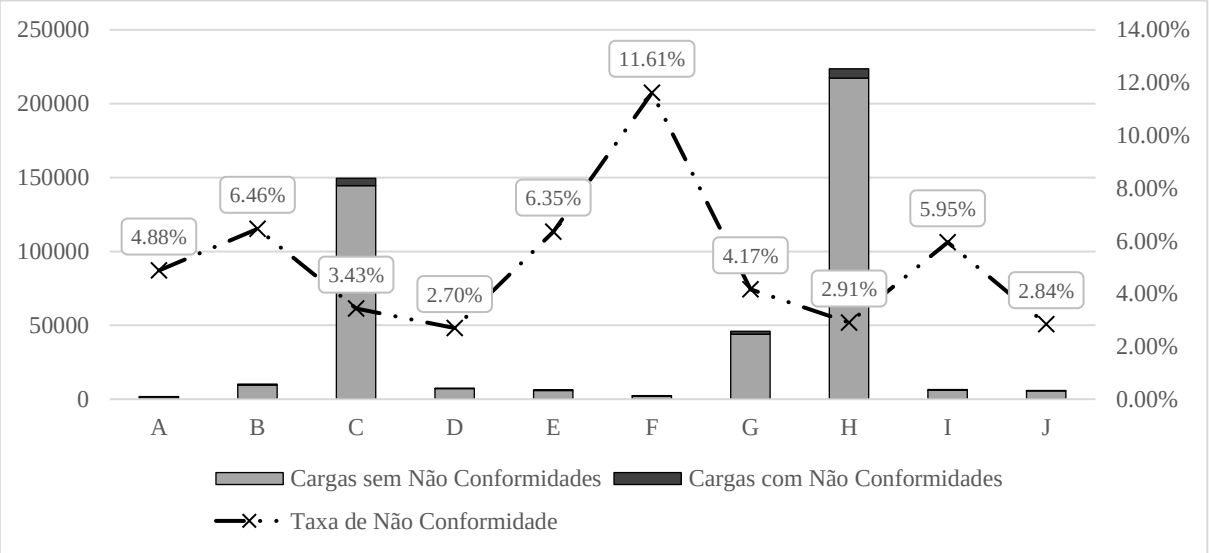
Fonte: Elaboração própria.XXXXX

De acordo com a Figura 2.9, grande parte do volume de PCG acondicionadas em MEMR movimentados no Aeroporto Internacional de Viracopos provêm de aeroportos de grande porte. Nesse caso, nota-se que nos três maiores continentes de origem (Europa, América do Norte e Ásia) as taxas de não conformidade dos aeroportos de grande parte situam-se abaixo de 4%. Apenas nos casos dos aeroportos de grande porte da Oceania e América do Sul são registradas taxas de não conformidade acima de 6%, embora ambos os continentes possuam um baixo movimento de cargas em MEMR destinados ao Aeroporto Internacional de Viracopos, em comparação ao movimento da Europa, América do Norte e Ásia.

Logo, assim como o descrito por: Rocha (2013); Eyre e Haack (2017); Haack (2006); Eyre *et al.* (2018); Caton e Dobbs, Brodel (2006); Monteferrante *et. al* (2018); Surkov (2007), nota-se que as características de origem de carga exibem taxas de não conformidades diferentes e, portanto, são fatores relevantes para a determinação do nível de risco fitossanitário esperado de uma determinada MEMR.

Além das características de origem, outros fatores relevantes são as características do importador. A Figura 2.10 e 2.11 apresentam a distribuição dos MEMR quanto aos setores dos importadores e à experiência do importador.

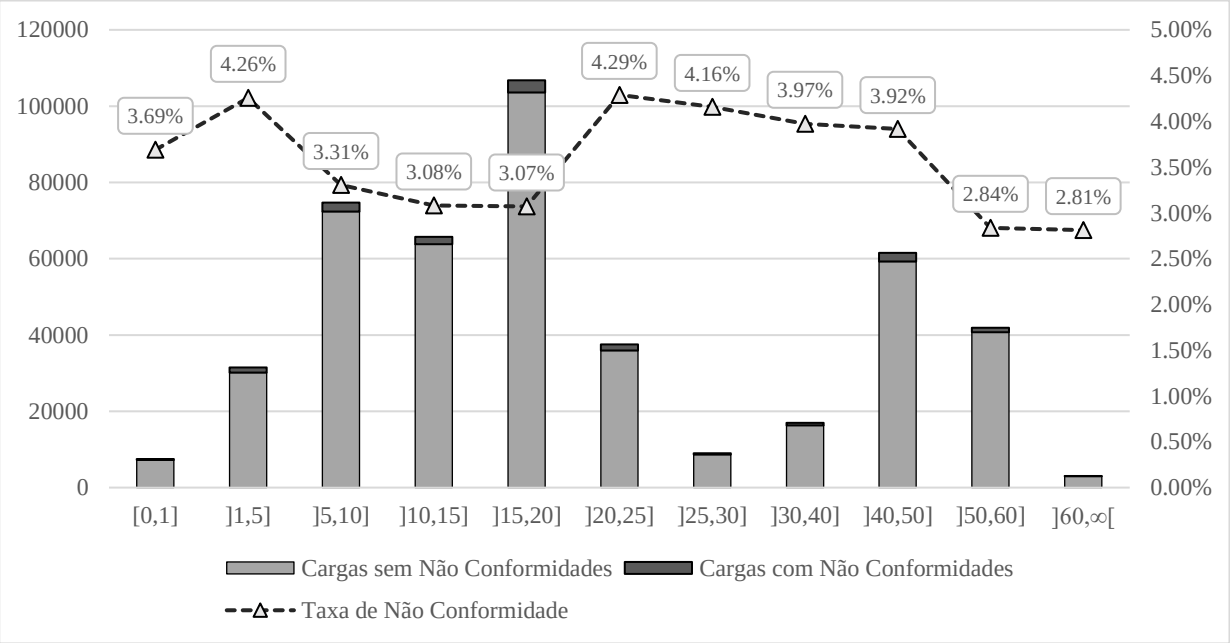
Figura 2.10: Número de PCG com MEMR importados com NC, sem NC e Taxa de NC por setor do importador, de dezembro de 2015 a março de 2019.



Fonte: Elaboração própria.

De acordo com a Figura 2.10, os setores C (Máquinas. Equipamentos, Computadores e Moveis), H (Automóveis, Equipamentos Automotores, Máquinas e Equipamentos) e G (Químicos, Petróleo, Plásticos e Farmacêuticos) respondem pelas maiores movimentações de cargas importadas em MEMR via Aeroporto Internacional de Viracopos, no período considerado. Dentre esses setores, apenas o setor H apresenta taxa de não conformidade inferior a 3%. Dentre os demais setores, apenas o setor J (Varejo, Atacado de Alimentos, Farmacêuticos, Livros e Perfumaria) apresentou taxa de não conformidade abaixo de 3%, enquanto o setor F (Transporte; Construção; Água, Esgoto e Descontaminação) deteve a maior taxa de não conformidade, em 11,61%.

Figura 2.11: Número de PCG com MEMR importadas com NC, sem NC e Taxa de NC por anos de experiência do importador, de dezembro de 2015 a março de 2019.

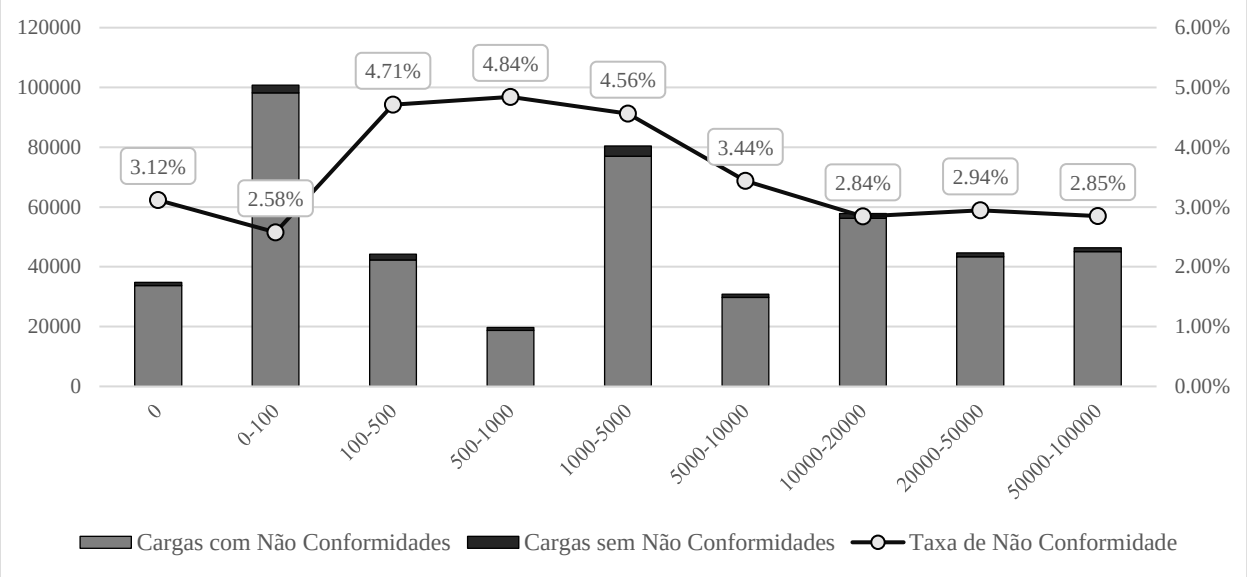


Fonte: Elaboração própria.

Quanto à relação entre experiência do importador e taxa de não conformidade, nota-se que grande parte das cargas em MEMR foram movimentadas por importadores com 5 a 20 anos ou com 40 a 60 anos de experiência e apenas as empresas com mais de 50 anos de experiência detiveram taxas de não conformidade médias abaixo de 3%. Entretanto, os importadores com 5 a 20 anos de experiência detiveram taxas de não conformidade em torno dos 3%, limite mínimo estabelecido para a adesão ao programa de fiscalização reduzida. Ou seja, para que a inspeção dos MEMR passe a ser amostral e não mais censitária, como ocorre para as demais empresas fora do programa piloto.

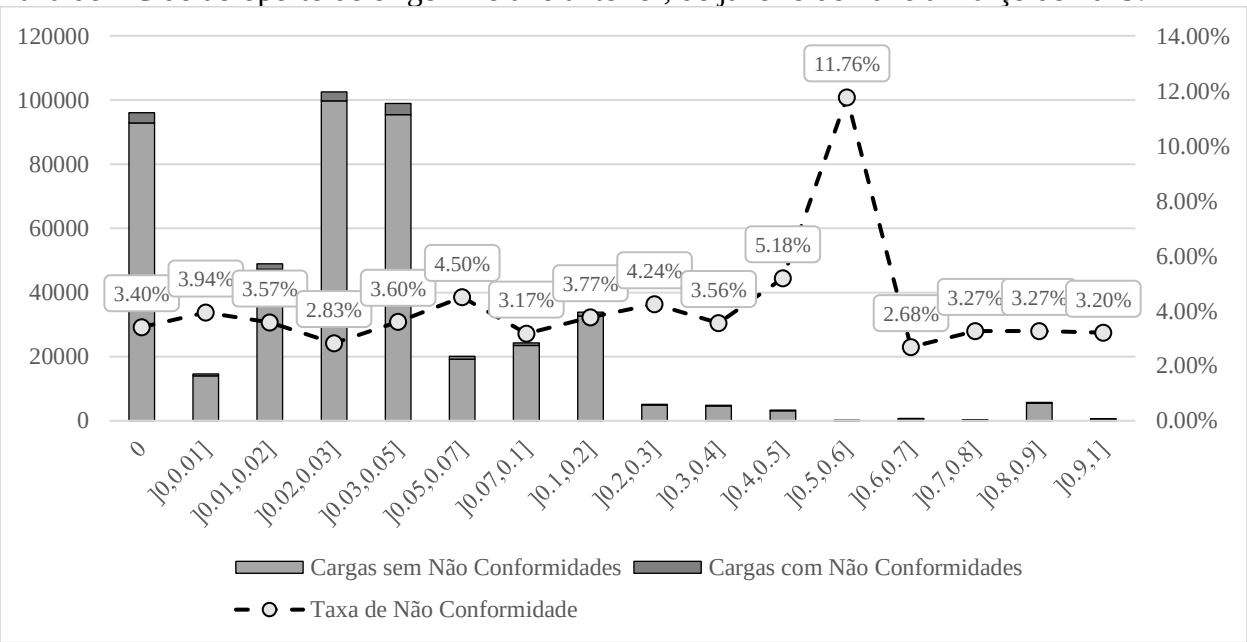
Por fim, outro fator relevante para a identificação e predição do risco fitossanitário via importação de MEMR é o histórico de importação da empresa. Assim, a Figura 2.12 e a Figura 2.13 apresentam as taxas de não conformidade conforme o volume de MEMR movimentado no ano anterior pelos agentes importadores e as taxas de não conformidade reportadas no ano anterior, respectivamente. Da mesma forma, as Figuras 2.14 e 2.15 apresentam a taxa de não conformidade conforme volume de importação usando MEMR do importador no ano anterior e a taxa de não conformidade do ano anterior.

Figura 2.12: Número de PCG com MEMR importado com NC, sem NC e Taxa de NC por volume importado pelo aeroporto de origem no ano anterior, de janeiro de 2016 a março de 2019.



Fonte: Elaboração própria.

Figura 2.13: Número de PCG com MEMR importado com NC, sem NC e Taxa de NC por Taxa de NC do aeroporto de origem no ano anterior, de janeiro de 2016 a março de 2019.

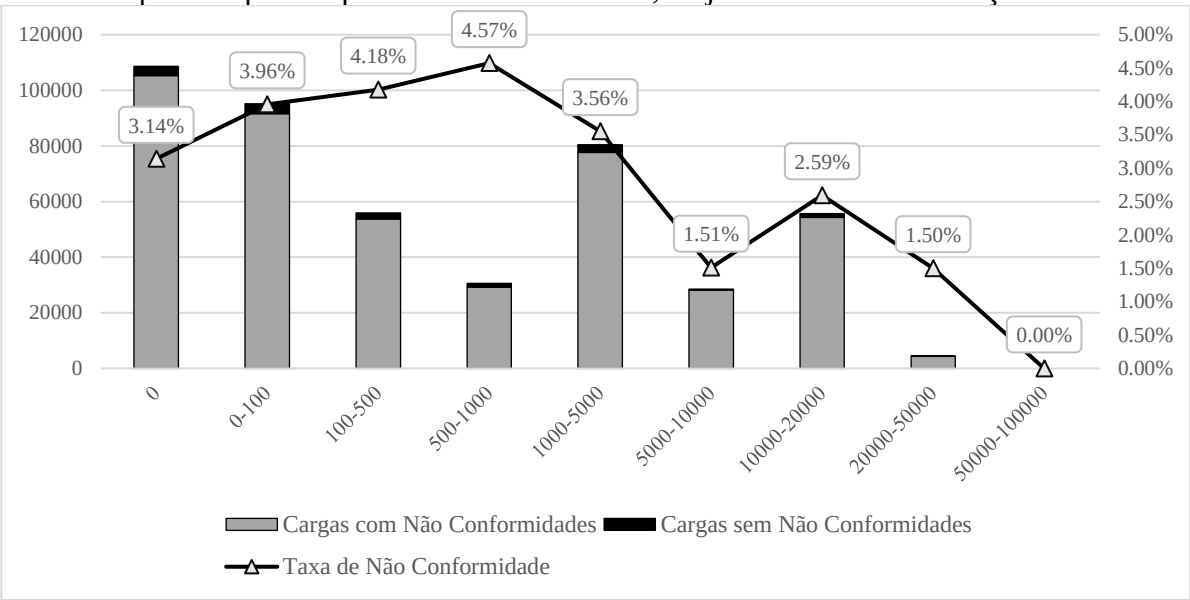


Fonte: Elaboração própria.

De acordo com a Figura 2.12, a maior parte do volume de PCG com MEMR importados se origina de aeroportos que destinaram, no máximo 100 cargas em MEMR para o Aeroporto Internacional de Viracopos no ano anterior. Da mesma forma apenas os aeroportos de origem com volume até 100 unidades de MEMR e superior a 10000 detiveram taxa de não

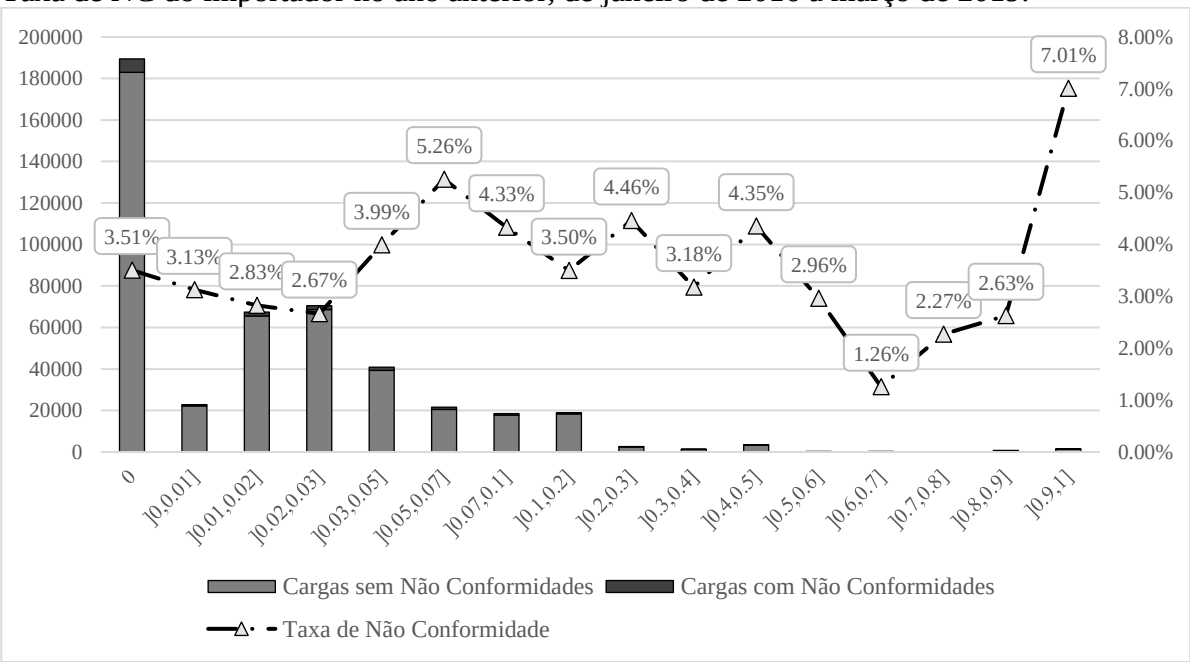
conformidade inferior a 3%. Já para o caso de histórico da taxa de não conformidade, nota-se que apenas o grupo de aeroportos com taxa de não conformidade entre 2% e 3% no ano anterior detiveram taxa de não conformidade menor do que 3%.

Figura 2.14: Número de PCG com MEMR importado com NC, sem NC e Taxa de NC por volume importado pelo importador no ano anterior, de janeiro de 2016 a março de 2019.



Fonte: Elaboração própria.

Figura 2.15: Número de PCG com MEMR importado com NC, sem NC e Taxa de NC por Taxa de NC do importador no ano anterior, de janeiro de 2016 a março de 2019.



Fonte: Elaboração própria.

Quanto às variáveis do histórico do importador, nota-se, a partir dos resultados expostos na Figura 2.15, que os maiores importadores são do grupo de empresas que não importaram ou importaram até 100 unidades no ano anterior, enquanto o grupo de empresas com no mínimo 5000 cargas importadas no ano anterior detiveram taxa não conformidade abaixo dos 3%. Da mesma forma, de acordo com os resultados expostos na Figura 2.15, os importadores com taxas de não conformidade no ano anterior de 1% a 2% e 2% a 3% mantiveram-se em patamares inferiores a 3%. Quanto ao grupo de importadores com 0% de taxa de não conformidade no ano anterior, provavelmente esse resultado advenha da não importação no ano anterior, enquanto o grupo de empresas com taxa de não conformidade no ano anterior de 0% a 1% detém um volume de importação baixo, em comparação aos demais, o que justifica a taxa de não conformidade de 3,13% no ano corrente.

Assim, nota-se que as empresas com histórico de grandes movimentações de PCG acondicionadas em MEMR apresentam menores taxas médias de não conformidade, bem como as empresas que, historicamente, apresentam taxas de não conformidade médias inferiores a 3% tendem a manter-se com baixas taxas de não conformidade nos períodos seguintes.

Logo, os resultados das Figuras 2.14 a Figura 2.15 corroboram com aqueles encontrados na literatura consultada, bem como denotam que a taxa de não conformidade varia conforme características de origem da carga acondicionada em MEMR e do importador, cujas necessidades condicionam o tipo e qualidade da embalagem da carga importada.

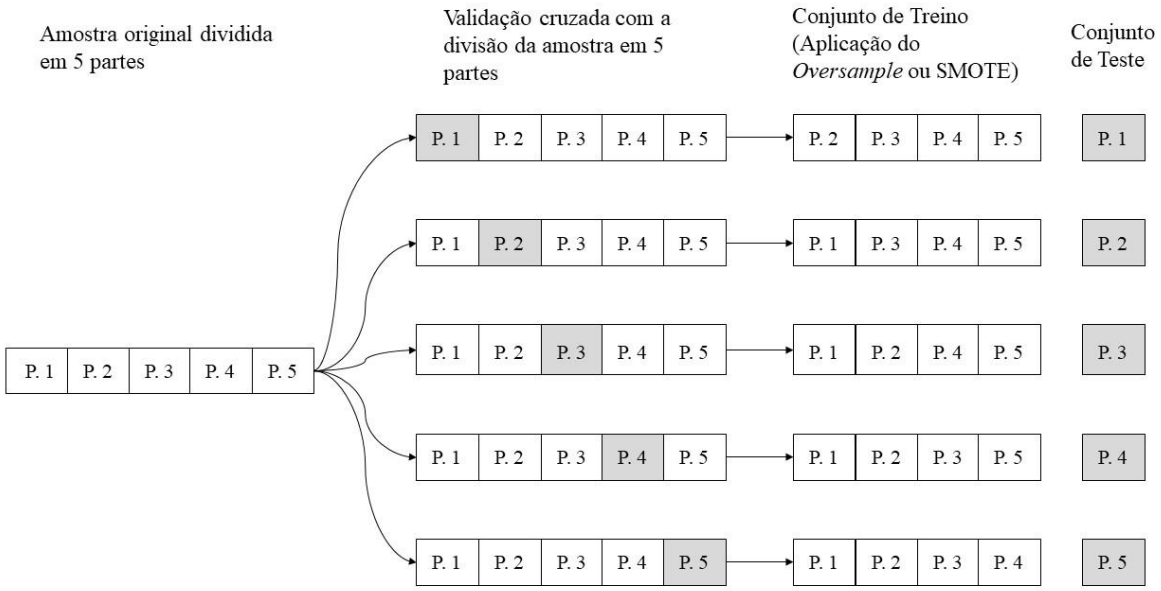
2.4.2. Validação Cruzada

A tabela 4 apresenta os resultados da validação cruzada para os modelos: Regressão Logistica com correção de King e Zeng (2001), Regressão Logística estimada com amostra corrigida via *Oversample*, Regressão Logística estimada com amostra corrigida via SMOTE, *Random Forest* estimado com amostra corrigida via *Oversample*, *Random Forest* estimado com amostra corrigida via SMOTE, e SVM estimado via algoritmo ALGA-SVM. Para aplicar o algoritmo SMOTE na amostra, 5 vizinhos mais próximos das observações minoritárias foram utilizados na composição das novas observações. Quanto a validação cruzada, a amostra foi dividida em 5 partes. Por fim, aplicação do SMOTE e *Oversample* na validação cruzada ocorreu após cada separação do conjunto de testes e de treino na validação cruzada, assim como o

recomendado por Blagus e Lusa (2015)[12]. Para tanto, a amostra inicial foi dividida em 5 partes e, após a divisão, replicada 5 vezes. Após a replicação, operou-se a aplicação do SMOTE ou *Oversample* no conjunto de treino, respeitando a preposição de que cada uma das 5 partes resultantes da divisão da amostra foi utilizada para o teste de desempenho.

Os procedimentos de aplicação da validação cruzada com SMOTE ou *Oversample* foram ilustrados na Figura 2.16.

Figura 2.16: Validação Cruzada com SMOTE e *Oversample*.



Fonte: Elaboração própria.

Como ilustrado na Figura 2.16, o procedimento adotado para a aplicação da validação cruzada com SMOTE ou *Oversample* preservaram as características iniciais da validação cruzada, bem como atenderam a recomendação de Blagus e Lusa (2015). Desse modo, todas as 5 partes resultantes da divisão inicial da amostra serviram como conjunto de dados para teste. A partir desses resultados, foi calculado a Acurácia e Taxa de Falso Negativo, com base nos conjuntos de dados de teste da validação cruzada, de todos os pontos de corte da probabilidade[13]

[12] Blagus e Lusa (2015) advogam que os métodos de correção de amostra não devem ser aplicados em conjunto de dados de teste. Assim, a aplicação dos métodos de correção deve ocorrer após a divisão da amostra entre conjunto de treino e conjunto de teste, apenas no conjunto de teste.

[13] Ponto de corte refere-se ao limiar probabilidade em que uma observação é classificada como $\hat{y}_i = 1$, caso a probabilidade estimada para a observação seja superior ao limiar, ou $\hat{y}_i = 0$ caso a probabilidade estimada seja inferior ao limiar.

(cut-off) da Regressão Logistica, com a correção do estimador e correções para a amostra. No caso do *Random Forest*, foi calculado os pontos de corte[14] para as funções de voto, de 1 à 10 variáveis disponíveis por sorteio, com a correção e amostra via SMOTE e *Oversample*. Ademais, calculamos a distância euclidiana do ponto com 100% de Acurácia e 0% de Taxa de Falso Negativo, para todos os pontos de corte das estimações da Regressão Logística e *Random Forest*. Nesse caso, atribuímos peso 0.75 para a acurácia e 0.25 para a taxa de falsos negativos no cálculo da distância euclidiana[15].

A Tabela 2.4 exibe os resultados para a validação cruzada da Regressão Logística e *Random Forest*.

Tabela 2.4: Acurácia, Taxa de falso negativo, Distância Euclidiana e ponto de corte da validação cruzada dos modelos Logit e *Random Forest*

Variáveis por Sorteio	*Cutoff*	Acurácia	Taxa de Falso Negativo	Distância Euclidiana	*Cutoff*	Acurácia	Taxa de Falso Negativo	Distância Euclidiana
Regressão Logística Correção King								
-	0.045	0.747	0.591	36.776	-	-	-	-
Oversample					SMOTE			
Regressão Logística								
-	0.486	0.57	0.361	39.719	0.161	0.961	0.993	49.789
Random Forest								
1	0.558	0.782	0.437	28.858	0.161	0.796	0.696	39.008
2	0.334	0.789	0.425	28.063	0.161	0.906	0.693	35.608
3	0.366	0.830	0.484	28.361	0.161	0.903	0.641	33.166
4	0.302	0.822	0.470	28.061	0.161	0.900	0.627	32.512
5	0.286	0.830	0.474	27.919	0.161	0.900	0.622	32.283
6	0.226	0.820	0.457	27.662	0.161	0.901	0.618	32.088
7	0.202	0.822	0.452	27.356	0.161	0.901	0.616	31.959
8	0.162	0.814	0.436	27.099	0.161	0.902	0.619	32.082
9	0.170	0.828	0.420	27.039	0.161	0.902	0.616	31.948*
10	0.162	0.832	0.410	26.998*	0.161	0.902	0.617	31.973

Fonte: Elaboração própria.

De acordo com a tabela 2.4, dez variáveis por sorteio minimizam a distância euclidiana[16] para o modelo *Random Forest* com correção de amostra via *Oversample*. Para o caso do modelo *Random Forest* com correção de amostra via SMOTE, nove variáveis por sorteio minimizam a distância euclidiana. Assim, adotamos em ambos os modelos *Random*

[14] Assim como o ponto de corte da probabilidade, o ponto de corte refere-se ao limiar de votos que uma observação necessita ter para ser classificada como $\widehat{y_i} = 1$. Nesse caso, se para uma determinada observação projetada com o *Random Forest* deteve o número de votos superior ao estipulado pelo ponto de corte, essa é classificada como $\widehat{y_i} = 1$, caso contrário é atribuído a essa observação $\widehat{y_i} = 0$.

[15] Nesse caso, a distância euclidiana pode ser representada por: $d(Acurácia_i, Taxa\ de\ Falso\ Negativo_i) = (0.75(Acurácia_i - 1)^2 + 0.25(Taxa\ de\ Falso\ Negativo_i - 0)^2)^{\frac{1}{2}}$.

[16] Além da distância euclidiana, adotamos um segundo critério para a seleção do ponto de corte: O ponto de corte deve ser superior a 0,161.

Forest o número de variáveis por sorteio que minimizam a distância euclidiana, bem como os *cutoff* na classificação das projeções da Regressão Logística.

2.4.3. ALGA-SVM

Para a execução do algoritmo ALGA-SVM, alguns parâmetros e valores iniciais foram atribuídos ao algoritmo. Essa subseção detalha os valores atribuídos e regras de execução adotadas.

Inicialmente, o conjunto de dados fora divido entre cinco partes. Quatro quintos foram destinados a construção da amostra de treino enquanto o último quinto foi reservado a servir como conjunto de teste.

Para a estimação do ALGA-SVM definimos como número de sub samples e hiperparâmetros candidatos, em ambos os casos, 20 unidades (R=20 e U=20). Os valores iniciais de cada candidato a hiperparâmetros são fornecidos pela combinação entre possíveis valores do parâmetro de custo e parâmetro da função de *Kernel*, em que os valores iniciais da função de custos são: 0,1; 1; 10; 100. Para o parâmetro da função de kernel, adotamos: 0,01; 0,1; 1, 10, 100. Para o caso das sub amostras, essas são formadas por 5000 observações amostradas de cada categoria (presença e ausência de não conformidades). Logo, cada subamostra contém 10000 observações. Em ambas as categorias a amostragem foi feita com repetição.

Como função de Kernel, seguimos Kawulok, Nalepa e Dudzik (2017) em sua aplicação inicial do algoritmo ALGA, de modo que utilizamos a tradicional função RBF. Assim, os pares de hiperparâmetros a serem otimizados são: γ (parâmetro da função do RBF) e *C* (parâmetro de custo do problema de minimização do hiperplano).

Para a verificar a performance dos SVM estimados com as sub amostras em *T'* e hiperparâmetros em *M'*, elegemos com critério de performance a menor distância Euclidiana entre 100% de acurácia e 0% de taxa de falsos negativos[17].

---

[17] Nesse caso, denota-se a distância euclidiana como: $d(acc_{p_iq_i}, fnr_{p_iq_i}) = \left((acc_{p_iq_i} - 1)^2 + (fnr_{p_iq_i} - 0)^2\right)^{\frac{1}{2}}$. Em que: $acc_{p_iq_i}$ denota a acurácia da sub amostra *i* (p_i) com o hiperparametro *i* (q_i); $fnr_{p_iq_i}$ denota a taxa de falsos negativos da sub amostra *i* (p_i) com o hiperparametro *i* (q_i); $d(.)$ denota a função da distância euclidiana.

Assim, espera-se que o par p_b e q_b selecionado detenha a maior acurácia, e, ao mesmo tempo, a menor taxa de falsos negativos.

Quanto ao processo de reprodução nas sub amostras, optamos por adotar as seguintes regras: Após a estimação e ranqueamento, com base no critério de performance, os indivíduos (p_i) são alocados entre os grupos *A* e *B*, de modo que p_{ia} em que ia = 1, 2, ..., *R*/2 e p_{ib} em que i b= 1, 2, ..., *R*/2. Os indivíduos do grupo *A* são automaticamente alocados no conjunto *T'*, assim como *R*/2 indivíduos resultantes do cruzamento entre o grupo *A* e *B*. O cruzamento dos indivíduos do grupo *A* e *B* foi realizado a partir da escolha aleatória entre os indivíduos de ambos os grupos. Assim, o indivíduo p_{a+b}, fruto do cruzamento, é dado por a escolha aleatória de indivíduos de p_a e p_b, respeitando a regra de *K* observações em cada categoria. O cruzamento ocorre até que se gere *R*/2 indivíduos. Desse modo, o conjunto *T'* resultante de tal etapa manterá *R* indivíduos para a próxima fase do processo de otimização.

Já o processo de reprodução dos hiperparâmetros ocorre da seguinte forma: Após a estimação, ranqueamento e divisão entre grupo dos candidatos a hiperparâmetro (mesmo processo que o adotado para os sub-samples), os indivíduos do grupo *A* são alocados em *M'*. O restante dos indivíduos de *M'* provém da reprodução entre os grupos *A* e *B*, cuja probabilidade de reprodução (P_c^M) é de: 0,7. Assim, a partir de todas as combinações entre o grupo *A* e *B* que resultaram na reprodução entre os indivíduos, escolhemos aleatoriamente as U/2 combinações que resultaram em reprodução.

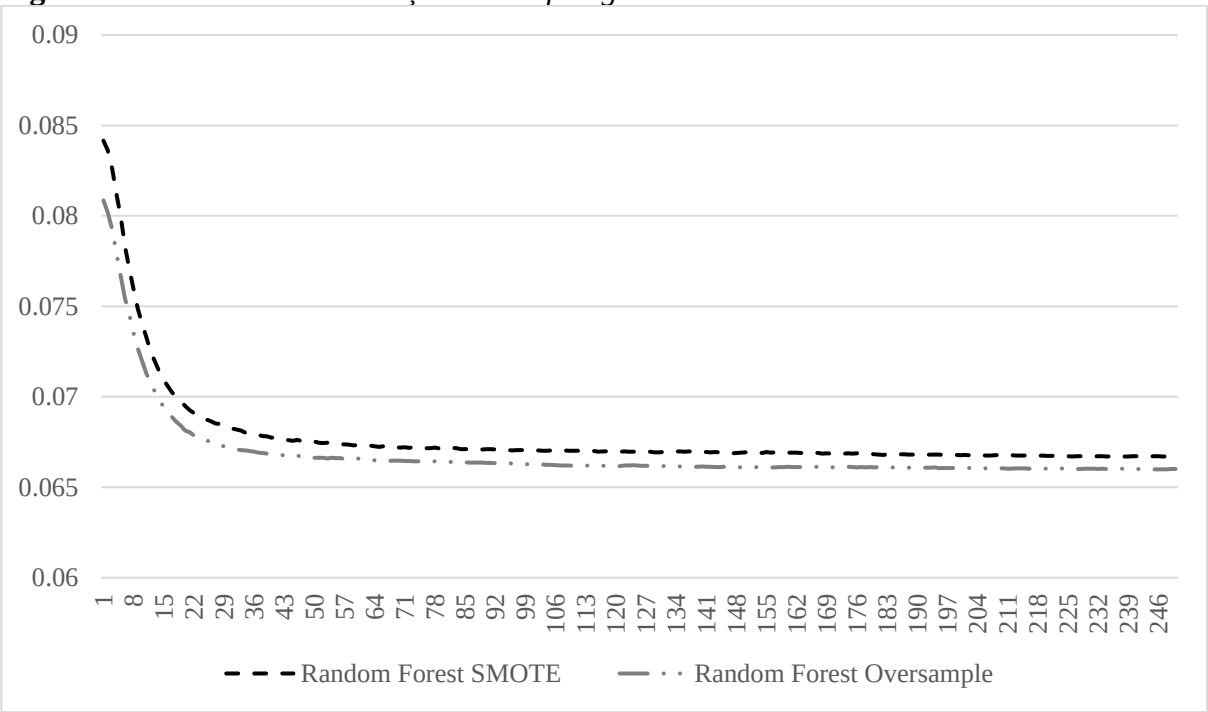
Para o processo de mutação, assim como em Kawulok, Nalepa e Dudzik (2017), adotamos a probabilidade de mutação (P_m^M) como 0,2 e δ_m como 0,1.

Por fim, estabelecemos como condições de parada do algoritmo ALGA a equivalência entre os indivíduos selecionados como ótimo nas etapas *n e n + 1*. Ou seja, para cada etapa de otimização (refinamento de *T'* e *M'*), o algoritmo executa cada etapa até que o mesmo individuo seja escolhido duas vezes consecutivas. Assim, o *looping* inicial em *M'* é finalizado assim que o mesmo par de hiperparâmetros é encontrado, assim como o *looping* em *T'* é finalizado.

2.4.4. Árvores no *Random Forest*

Para a escolha do tamanho da floresta, optou-se pelo número de árvores que produzem o menor erro de classificação. A Figura 11 apresenta o comportamento a média do erro de classificação *"out-of-bag*" de acordo com o número de árvores adotadas para a classificação.

Figura 2.17: Erro de Classificação "*Out-of-Bag*" até 250 árvores.



Fonte: Elaboração própria.

De acordo com a Figura 2.17, em ambos os casos, o percentual de erros de classificação decresce com o aumento do tamanho da floresta, mas não apresenta redução expressiva a partir da quinquagésima árvore. Entretanto, optamos por estimar a versão final de ambos os modelos com 100 árvores.

2.4.5. Regressão Logística

A Tabela 2.5 apresenta os coeficientes da regressão logística corrigidos pelo método de King e Zeng (2001), com a amostra com *Oversample* e amostra corrigida via SMOTE.

Tabela 2.5: Coeficiente da Regressão Logística com correção de King e Zeng (2001), *Oversample* e SMOTE

Variável	Coeficiente	Erro Padrão	*P-Valor*	Coeficiente	Erro Padrão	*P-Valor*	Coeficiente	Erro Padrão	*P-Valor*
	SMOTE			*Oversample*			*King e Zeng (2001)*		
Intercepto	2.571	0.028	0.000***	-0.228	0.025	0.000***	-3.337	0.088	0.000***
Variáveis do Importador									
Setor A: Agropecuária e Setor Extrativo	-0.470	0.087	0.000***	0.188	0.049	0.000***	0.193	0.162	0.234 ns
Setor B: Industria Ali., têxtil e diversos	0.355	0.029	0.000***	0.743	0.018	0.000***	0.693	0.056	0.000***
Setor C: Máquinas, Equip., Computadores e Moveis	0.001	0.011	0.961 ns	0.207	0.007	0.000***	0.187	0.026	0.000***
Setor D: Industria e comércio de Madeira	-0.500	0.046	0.000***	-0.206	0.028	0.000***	-0.202	0.105	0.055*
Setor E: Industria e comércio de metal	0.338	0.033	0.000***	0.359	0.023	0.000***	0.341	0.075	0.000***
Setor F: Transporte; Cons. e Água, Esgoto.	0.204	0.062	0.001***	0.446	0.038	0.000***	0.630	0.111	0.000***
Setor G: Químicos; Petróleo e Plástico, farmacêuticos	0.230	0.016	0.000***	0.218	0.011	0.000***	0.203	0.037	0.000***
Setor I: Pesquisa e Serviços	0.518	0.034	0.000***	0.274	0.024	0.000***	0.315	0.080	0.000***
Setor J: Varejo e Atacado de Alimentos, Farmacêuticos, Livros e Perfumaria	0.810	0.032	0.000***	-0.217	0.027	0.000***	-0.252	0.104	0.016**
Experiência	0.000	0.000	0.491 ns	0.000	0.000	0.623 ns	0.000	0.001	0.965 ns
Variáveis de Origem									
Aeroporto: Médio	1.746	0.012	0.000***	-0.131	0.011	0.000***	-0.120	0.041	0.004***
Aeroporto: Pequeno	1.648	0.024	0.000***	-0.003	0.022	0.910 ns	-0.120	0.079	0.131 ns
África	1.966	0.103	0.000***	0.995	0.321	0.002***	0.797	0.296	0.007***
Europa	-1.342	0.014	0.000***	-0.205	0.028	0.000***	0.153	0.055	0.005***
América do Norte	-1.232	0.014	0.000***	0.000	0.013	0.992 ns	0.168	0.051	0.001***
Oceania	1.808	0.061	0.000***	1.066	0.108	0.000***	0.192	0.222	0.387 ns
América do Sul	-0.964	0.053	0.000***	0.151	0.015	0.000***	0.347	0.132	0.008***
Renda Baixa	0.974	0.452	0.031**	0.188	0.014	0.000***	1.134	0.706	0.108 ns
Renda Média Baixa	2.840	0.019	0.000***	0.055	0.068	0.421 ns	-0.237	0.103	0.022**
Renda Média Alta	-0.162	0.015	0.000***	0.327	0.042	0.000***	0.050	0.044	0.254 ns
NIMF-15	-2.469	0.015	0.000***	-0.323	0.023	0.000***	-0.333	0.080	0.000***
Fevereiro	0.457	0.023	0.000***	-0.039	0.016	0.017**	-0.043	0.063	0.499 ns
Março	0.201	0.024	0.000***	0.175	0.015	0.000***	0.194	0.059	0.001***
Abril	0.265	0.024	0.000***	0.439	0.015	0.000***	0.393	0.057	0.000***
Maio	0.377	0.023	0.000***	0.759	0.015	0.000***	0.705	0.054	0.000***
Junho	0.812	0.022	0.000***	0.768	0.015	0.000***	0.724	0.055	0.000***
Julho	0.221	0.025	0.000***	0.683	0.015	0.000***	0.620	0.057	0.000***
Agosto	0.565	0.023	0.000***	0.717	0.015	0.000***	0.655	0.055	0.000***
Setembro	0.213	0.024	0.000***	0.609	0.015	0.000***	0.560	0.056	0.000***
Outubro	-0.164	0.031	0.000***	0.439	0.018	0.000***	0.349	0.068	0.000***
Novembro	0.264	0.029	0.000***	0.126	0.020	0.000***	0.077	0.078	0.325 ns
Dezembro	-0.577	0.034	0.000***	0.144	0.019	0.000***	0.109	0.073	0.136 ns
Carga Transportada									
PCG no Voo	-0.009	0.000	0.000***	-0.003	0.000	0.000***	-0.003	0.000	0.000***
Histórico de Não Conformidades									
Taxa de NC do Importador no Ano Anterior	0.859	0.096	0.000***	2.103	0.066	0.000***	1.403	0.148	0.000***
PCG com MEMR do Importador no Ano Anterior	0.000	0.000	0.000***	0.000	0.000	0.000***	0.000	0.000	0.000***
Taxa de NC do Importador no Mês da Importação no Ano Anterior	-1.079	0.077	0.000***	0.201	0.041	0.000***	0.188	0.122	0.122 ns
Taxa de NC do Aeroporto de Origem no Ano Anterior	3.078	0.148	0.000***	5.211	0.117	0.000***	2.852	0.198	0.000***
PCG com MEMR do Aeroporto de Origem no Ano Anterior	0.000	0.000	0.000***	0.000	0.000	0.000***	0.000	0.000	0.000***
Taxa de NC do Aeroporto de Origem no Mês da Importação no Ano Anterior	-2.653	0.095	0.000***	-0.089	0.058	0.125 ns	0.181	0.166	0.273 ns

Fonte: Elaboração própria. (ns) denota não significância (*) denota significância à 10%, (**) denota significância à 5% e (***) denota significância à 1%.

A partir dos resultados expostos na Tabela 2.5, destaca-se que grande parte dos coeficientes das variáveis independentes foram estatisticamente significativos, nos modelos estimados com métodos de correção diferentes. Assim, embora a análise acerca dos determinantes da não conformidade em embalagens de madeira esteja além do escopo desse trabalho, a breve análise dos parâmetros estimados aponta fatores relevantes ao comportamento da determinação da não conformidade nas embalagens.

Primeiramente, a análise dos parâmetros aponta para a existência de diferenças entre a probabilidade de importações com embalagens com não conformidades conforme o setor do importador. Quanto ao conjunto de variáveis de origem, nota-se que há diferenças na probabilidade esperada conforme o nível de renda do país de origem, bem como o tamanho do aeroporto de origem, continente de origem e adoção da NIMF-15 no país de origem da carga. Ademais, os resultados da regressão logística com tipos de correções amostrais diferentes apontaram para o impacto da sazonalidade sobre a probabilidade de um MEMR apresentar não conformidades. Nota-se nas três equações há diferenças entre a probabilidade esperada conforme o mês em que ocorreu a importação. Por fim, as variáveis de histórico do importador e origem também foram estatisticamente significativas em uma ou mais equações. Desse modo, espera-se que indivíduos que, historicamente, detém maior taxa de não conformidade apresentem maior probabilidade de importar uma carga com embalagem de madeira não conforme, tanto para o caso do aeroporto de origem, quanto para a empresa importadora.

2.4.6. *Random Forest*

A Tabela 2.6 apresenta a média do coeficiente de Gini das variáveis empregadas no *Random Forest* nos modelos estimados com *Oversample* e SMOTE.

Tabela 2.6: Média do coeficiente de Gini

Variável	Média do Índice de Gini (*Oversample*)	Média do Índice de Gini (SMOTE)
Adoção dos Procedimentos do NIMF-15 pelo exportador	47104.02	515.559
Classificação do Nível de Renda do País de Origem	44434.79	1344.73
Continente de Origem da Importação	20316.32	3072.88
Volume de PCG com MEMR do Aeroporto de Origem no Ano Anterior	20045.85	13448.08
Taxa de Não Conformidade do Aeroporto de Origem no Ano Anterior	14481.03	16674.57
Classificação do Aeroporto de Origem	12495.01	1023.01
Taxa de Não Conformidade do Aeroporto de Origem no Mês da Importação no Ano Anterior	8419.721	13392.14
Volume de PCG com MEMR do Importador no Ano Anterior	7295.538	22313.27
Número de PCG no Voo	6486.84	50994.17
Experiência do Importador	4700.817	14123.06
Taxa de Não Conformidade do Importador no Ano Anterior	3989.816	13440.67
Mês da Chegada	3904.22	20830.47
Setor do Importador	2568.205	8540.974
Taxa de Não Conformidade do Importador no Mês da Importação no Ano Anterior	1432.712	7184.294

Fonte: Elaboração própria.

De acordo com a Tabela 2.6, as seis variáveis com maior importância para a predição da não conformidade no modelo feito a partir do *Oversample* são: Número de PCG no Voo, Mês Chegada, Volume de PCG com MEMR do Importador no Ano Anterior, Taxa de Não Conformidade do Aeroporto de Origem no Ano Anterior e Experiência do Importador. Para o caso do modelo com SMOTE, as variáveis mais importantes para a predição são: Adoção dos Procedimentos do NIMF-15 pelo exportador, Classificação do Aeroporto de Origem, Classificação do Nível de Renda do País de Origem, Continente de Origem da Importação, Taxa de Não Conformidade do Importador no Mês da Importação no Ano Anterior.

2.4.7. Resultados e Comparação entre os modelos no período de outubro 2018 a março 2019

A Tabela 2.7 apresenta os indicadores de performance selecionados e aplicados sobre os resultados da previsão de todos os modelos para o conjunto de teste. A regressão logística, estimada junto a correção proposta por King e Zeng (2001), apresentou a maior acurácia, bem como a maior taxa de falsos negativos. Assim, embora o método detenha alta capacidade de detectar as embalagens de madeira com não conformidades, 84% parte das embalagens não

conformes foram classificadas como regulares. Desse modo, tal método deteve o pior desempenho em prever corretamente a irregularidade nas embalagens, fato que, obviamente, inviabiliza a aplicação prática desse modelo.

Tabela 2.7: Comparação entre a performance dos modelos *Random Forest*, Regressão Logística e SVM na predição das MEMR, em outubro de 2018 a março de 2019.

Modelo	Sensibilidade	Especificidade	Acurácia	Taxa de Falso Negativo	Taxa de Falso Positivo	F-Score	AUC
Logit (King)	0.159	0.929	0.905	0.840	0.070	0.528	0.655
Logit (*Oversample*)	0.407	0.793	0.781	0.592	0.206	0.104	0.661
Logit (SMOTE)	0.500	0.618	0.614	0.499	0.381	0.074	0.584
Random Forest (*Oversample*)	0.542	0.798	0.815	0.457	0.201	0.139	0.768
Random Forest (SMOTE)	0.743	0.529	0.535	0.256	0.47	0.091	0.668
ALGA-SVM	0.496	0.741	0.733	0.504	0.258	0.105	-

Fonte: Elaboração própria.

Já as aplicações da regressão logística estimadas com correções na amostra (SMOTE e *Oversample*) apresentaram desempenhos semelhantes, mas com diferenças cruciais para a aplicação. Enquanto o modelo Logit estimado com a amostra corrigida via *Oversample* deteve acurácia superiora 75% e 16,7 p.p. superior a estimação com correção de amostra via SMOTE, o modelo com a amostra corrigida com o SMOTE apresentou uma taxa de falso negativo inferior a 50% e 9,9 p.p. menor que a taxa apresentada com a estimação com correção de amostra via *Oversample*. Logo, enquanto uma técnica proporcionou um nível maior de acurácia na previsão, a outra deteve maior capacidade em acertar corretamente as observações com maior custo de erro de previsão.

Quanto aos modelos *Random Forest*, esses apresentaram o mesmo padrão da regressão logística. Enquanto a correção de amostra via *Oversample* apresentou maior acurácia em comparação a correção via SMOTE, o modelo estimado com a amostra corrigida pelo SMOTE foi capaz de prever com maior exatidão as observações com não conformidades. O modelo SVM estimado via algoritmo genético ALGA errou pouco mais de 50% das embalagens de madeira com não conformidades, embora previu corretamente 73,3% das observações.

No aspecto geral, o *Random Forest* deteve o melhor desempenho entre as metodologias testadas. A aplicação com correção da amostra via *Oversample* apresentou a melhor combinação entre acurácia e erros de previsão das não conformidades. Já a aplicação com o

SMOTE, embora pouco acurada, errou apenas pouco mais de um quarto das embalagens de madeira com não conformidades.

Por fim, a Tabela 2.8 apresenta o resultado caso ambas as versões as versões do Randon Forest fossem aplicadas na previsão das embalagens de MEMR não conformes no Aeroporto Internacional de Viracopos, entre outubro de 2018 e março de 2019.

Tabela 2.8: Número de PCG com MEMR Inspecionadas e Não Inspecionadas entre outubro de 2018 e março de 2019, com a adoção da projeção Modelos *Random Forest* como critério de seleção à inspeção.

	Oversample	SMOTE	Diferença
Inspecionado	20970	47268	26298
Não Inspecionado	77616	51318	26298
Cargas com não conformidades inspecionadas	1685	2309	624
Cargas com não conformidades não inspecionadas	1421	797	624

Fonte: Elaboração própria.

De acordo com a tabela 2.8, caso os resultados do modelo com *Oversample* fossem utilizados como critério para a inspeção de um PCG com MEMR de madeira, 20.970 PCG com MEMR teriam sido inspecionadas entre outubro de 2018 e março de 2019. Caso o modelo *Random Forest* com correção de amostra via SMOTE fosse empregado, 47.268 PCG com MEMR seriam fiscalizadas. A diferença, de 26.298 PCG com MEMR, equivale a 146,1 PCG com MEMR fiscalizados a mais por dia. Entretanto, o *Random Forest* com correção para a amostra via *Oversample* não foi capaz de predizer que 1421 PCG com MEMR não conformes apresentavam a irregularidade enquanto o modelo estimado com SMOTE apresentou apenas 797. Desse modo, embora o modelo com correção via SMOTE apresente menor redução da fiscalização, 51.318 PCG com MEMR deixariam de ser fiscalizadas com o seu uso. O que equivale a 285,1 PCG com MEMR por dia. Redução associada ao menor nível de risco fitossanitário do que o modelo com correção via *Oversample*. Por outro lado, caso o tomador de decisão esteja disposto a aceitar um nível maior de risco, o uso da estimação com *Oversample* resultaria em 21,27% do total de PCG com MEMR fiscalizados, com apenas 624 PCG com MEMR não conformes não fiscalizados a mais do que na estimação com SMOTE.

2.5. Considerações Finais

O presente trabalho deteve como objetivo construir modelos de previsão para MEMR com não conformidades importadas via Aeroporto Internacional de Viracopos, em Campinas –

SP. Para tanto, empregam-se três métodos popularmente utilizados em tarefas de classificação, sendo eles: Regressão Logistica, *Random Forest* e SVM.

Para adequar a problemática exposta pelo desbalanceamento entre os MEMR conformes e não conformes, foram adotados dois métodos de correção de amostra: *Oversample* e SMOTE. Ademais, ainda se utilizou a regressão logística com correção de King e Zeng (2001) para adequar o estimador da regressão logística com a problemática do desbalanceamento entre as categorias de MEMR. Para o caso do SVM, o tamanho do conjunto de dados inviabilizava a estimação por métodos tradicionais, de modo que empregamos o algoritmo ALGA-SVM para realizar a estimação desse método.

De modo geral, os resultados corroboraram o exposto na literatura sobre não conformidades em MEMR. Fatores como o local e características de origem impactam na taxa de não conformidade, bem como o tipo de material importado, expresso pelo setor de atuação do importador. Ademais, assim como identificado na literatura consultada, o volume de PCG com não conformidades no MEMR varia ao longo do ano, fato que remete a sazonalidade da taxa de não conformidade dos MEMR. Por fim, verificou-se que aeroportos de origem e importadores com alto volume de MEMR em períodos anteriores tendem a apresentar taxas de não conformidade inferiores a 3%, bem como os importadores com histórico de baixas taxas de não conformidade nas embalagens de madeira.

Por fim, foram comparados o desempenho seis modelos de previsão na identificação de não conformidade nas embalagens de madeira, sendo esses: Regressão Logistica com correção de King e Zeng (2001), Regressão Logística estimada com amostra corrigida via *Oversample*, Regressão Logística estimada com amostra corrigida via SMOTE, *Random Forest* estimado com amostra corrigida via *Oversample*, *Random Forest* estimado com amostra corrigida via SMOTE, e SVM estimado via algoritmo ALGA-SVM. Dentre os resultados da previsão dos modelos sobre as MEMR importadas no período de outubro de 2018 e março de 2019 destaca-se o desempenho de dois modelos: *Random Forest* estimado com amostra corrigida via *Oversample*, *Random Forest* estimado com amostra corrigida via SMOTE.

O *Random Forest* estimado com amostra corrigida via *Oversample* apresentou o melhor equilíbrio entre acurácia e erros de classificação das MEMR com não conformidades. Já o *Random Forest* estimado com amostra corrigida via SMOTE apresentou baixa acurácia, mas o menor erro de classificação das MEMR com não conformidades. Desse modo, a simulação do uso do *Random Forest* estimado com amostra corrigida via *Oversample*, no período de outubro de 2018 e março de 2019, acarretou a redução de inspeção de 77616 MEMR, 26298 a menos que o modelo *Random Forest* estimado com amostra corrigida via SMOTE. Por outro lado, o

Random Forest estimado com amostra corrigida via SMOTE gerou 797 erros de classificação de MEMR com não conformidades, 624 a menos que o *Random Forest* estimado com amostra corrigida via *Oversample*. Logo, ambas alternativas mostraram-se viáveis em reduzir o número de MEMR fiscalizadas, mantendo um relativo controle do risco fitossanitário.

Referências

ABDOH, S. F.; RIZKA, M. A.; MAGHRABY, F. A. Cervival Cancer Diagnosis Using Random Forest Classifier with SMOTE and Feature Reduction Techniques. **IEEE Access**, v. 6, p. 59475-59485, 2018.

ADAMI, A. C. O.; SUN, L. T.; MIRANDA, S. H. G. Estudo para Implantação de Canais de Inspeção NA Fiscalização de Embalagens, Suportes e Peças de Madeira no Aeroporto Internacional de Viracopos – Campinas/SP: Atualização dos dados com inclusão de agosto de 2017, reavaliação da Lista Piloto de Seleção e Exclusão de empresas do Canal, Levantamento preliminar dos benefícios e custos, Set. 2017.

ALRASHEED, D.; CHE, D. Improving Bankruptcy Prediction Using Oversampling and Feature Selection Techniques. In: Int'l Conf. Artificial Intelligence, 20, 2018, Las Vegas. **Anais...** CSREA, 2018. p. 440-446.

AMIN, A. *et al.* HUSSAIN, A. Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Costumer Churn Prediction Case Study. **IEEE Access**, v.4, p. 7940-7957, 2016.

BLAGUS, R.; LUSA, L. Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. **BMC Bioinformatics**, v.16, 2015.

BRAGAT, R. C.; PATIL, S. S. Enhanced SMOTE Algorithm for Classification of Imbalanced Big-Data using Random Forest. In: IEEE International Advance Computing Conference (IACC), 5, 2015, Banglore. **Anais...** IEEE, 2015. p. 403-408.

BRASIL. Instrução Normativa DAS/MAPA n° 4, de 06 de janeiro de 2004. **Estabelece, em caráter emergencial, até que se complete o processo de ajustamento da Legislação Fitossanitária Brasileira, a Norma Internacional e cumprimento dos prazos de notificação aos organismos internacionais, os procedimentos de inspeção e fiscalização de embalagens e suportes de madeira utilizados no transporte de mercadorias no comércio internacional.** Diário Oficial da União, Brasília, Distrito Federal, 2004.

BRASIL. Instrução Normativa DAS/MAPA nº 39, de 27 de novembro de 2017**. Aprova o funcionamento do Sistema de Vigilância Agropecuária Internacional -Vigiagro**. Diário Oficial da República Federativa do Brasil, Brasília, Distrito Federal, 2017.

BREIMAN, L. Random Forest, **Machine Learning**, v. 45, p. 5-32, 2001

CATON; B. P.; DOBBS, T. T.; BRODEL, C. F. Arrivals of hichhking insect pests on international cargo aircraft at Miami International Airport. **Biological Invasions**, v. 8, p. 765-785, 2006.

CROOK, J. N.; EDELMAN, D. B.; THOMAS, L. C. Recent developments in consumer credit risk assessment. **European Journal of Operational Research**, v. 183, n. 3, p. 1447-1465, 2007.

ERYE, D. *et al.* Variation in Inspection Efficacy by Member States of Wood Packaging Material Entering the European Union. **Journal of Economic Entomology**, v. 111, n. 2, p. 707-715. 2018.

EYRE, D.; HAACK, R. A. Invasive Cerambycid Pests and Biosecurity Measures. In: WANG, Q. (Org.). **Cerambycidae of the World**: biology and pest management. Boca Raton, FL: CRC, 2017. Cap. 13, p. 563-618.

FANTAZZINI, D.; FIGINI, S. Random Survival Forest Models for SME Credit Risk Measurement. **Methodol Compu. Appl. Probab**., v. 11, p. 29-45, 2009.

FURLAN, P. K.; PINTO, M. M. O. Identificação dos procedimentos de fronteira críticos na importação de cargas conteinerizadas: estudo do Porto de Santos. **Production**, v. 25, n. 1, p. 183-189, 2015.

GHATASHEH, N. Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study. **International Journal of Advanced Science and Technology**, v. 72, p. 19-30. 2014.

GREENE, W. H. **Econometric Analisys.** Pearson Education India. p. 1178. 2007.

HAACK, R. A. *et al.* Effectiveness of the International Phytosanitary Standard ISPM n° 15 on Reducing Wood Borer Infestation Rates in Wood Packaging Material Entering the United States. **PLoS ONE**, v. 9, n. 5, p. 1-15, 2014.

HAACK, R. A. Exotic bark- and wood-boring Coleoptera in United States: recent establishments and interception, **Can. J. for Res.** v. 36, p. 269-288, 2006.

HAND, D. J.; NIALL, A. M. Defining attributes for scorecard construction in credit scoring. **Journal of Applied Statistics**, v. 27, n. 5, p. 527-540. 2000.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**: 2. ed. Springer, 2008.

IPPC. International Plant Protection Convention. Guidelines for inspection. Publ. No. 23. **Food and Agriculture Organization of the United Nations**, Roma, Italy. 2006.

IPPC. International Plant Protection Convention. International Standards for Phytosanitary Measures: Guidelines for Regulating Wood Packaging Material in International Trade, Publ. No. 15. **Food and Agriculture Organization of the United Nations**, Roma, Italy. 2002.

KAWULOK, M.; NALEPA, J. Support Vector Machines Training Data Selection Using a Genetic Algorithm. In: GIMEL'FATH, G. et al. (Org.) **Structural, Syntactic, and Statistical Pattern Recognition**, Berlin: Lecture Notes in Computer Science, 2012.

KAWULOK, M.; NALEPA, J.; DUDZIK, W. Na Alternating Genetic Algorithm for Selecting SVM Model and Training Set In: CARRASCO-OCHOA, J. A.; MARTÍNEZ-TRINIDAD, J. F.; OLVERA-LÓPEZ, J. A. **Pattern Recognition**, Huatulco: Lecture Notes in Computer Science, 2017.

KING; G.; ZENG; L. Logistic Regression in Rare Events Data, **Political Analysis**, v.9, n. 2, p. 137-163, 2001.

KRUPPA, J.*et al.* A. Consumer credit risk: Individual probability estimates using machine learning. **Expert Systems with Applications**, v. 40, p. 5125-5131, 2013.

LE, T. *et al.* Oversampling Techniques for Bankruptcy Prediction: Novel Features from a Transaction Dataset. **Symmetry**, v.10, 2018.

MALEKIPIRBAZARI, M.; AKSAKALLI, V. Risk assessment in social lending via random forest. **Expert Systems with Applications**, v. 42, p. 4621-4631, 2015.

MOHAMMED, R.; RAWASHDEH, J.; ABDULLAH, M. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In: International Conference on information and Communication Systems, 11, 2020, Jordan. **Anais...** Jordan: IEEE-Jordan, 2020. p. 243-248.

MONTEFERRANTE, E. C. *et al.* Interceptações de pragas em embalagens de madeira no Porto de Santos. **Revista de Política Agrícola**, v. 27, n. 1, p. 132-141, 2018.

NALEPA, J.; KAWULOK, M. Selecting training sets for support vector machines: a review. **Artif. Intell. Ver**. v. 52, p. 857-900, 2018.

NEHREBECKA, N. Predicting the Default Risk of Companies. Comparison of Credit Scoring Models: Logit vs Support Vector Machines. **ECONOMETRICS EKONOMETRIA**, v. 22, n.2, p. 54-73, 2018.

PAAL, B. **A Comparison of Different Methods for Modelling Rare Events Data.** 2014. 75f. Dissertação (Mestrado em Análise Estatística de Dados) – Universidade de Gante, Gante. 2014.

PONTES, F. S. S. **Técnicas de mineração de dados aplicada na inspeção de importações para verificação dos status fitossanitário de embalagens de madeira em contêineres**. 2013, 67 f. Dissertação (Mestrado em Defesa Sanitária Vegetal) – Universidade Federal de Viçosa, Viçosa. 2013.

QU, Z. et al. Detection of Electricity Theft Behavior Based on Improved Synthetic Minority Oversampling Technique and Random Forest Classifier. **Energies**, v. 13, 2020.

ROCHA, B. G. D. **Identificação de perfis de risco associados a envios importados com probabilidade de abrigar pragas quarentenárias.** 2013. 73 f. Dissertação (Mestrado em Defesa Sanitária Vegetal) – Universidade Federal de Viçosa, Viçosa. 2013.

SAWANGARREERAK, S.; THANATHAMATHEE, P. Random Forest with Sampling Techniques for Handling Imbalance Prediction of University Student Depression. **Information**, v. 11, 2020.

STRUTT. A *et al.* Evaluating the impacts of an international phytosanitary standard for wood packaging material: Global and United States trade implications. **Forest Policy and Economics**, v. 27, p. 54-64, 2012.

SURKOV, I. V.**Optimising import phytosanitary inspection.** 2007. 196 f. Tese - Wageningen University, Wageningen. 2007.

SZYNISZEWSKA, A. M. *et al.* Analysis of Seasonal Risk for Importation of the Mediterranean Fruit Fly,Ceratitis capitata(Diptera: Tephritidae), via Air Passenger Traffic Arriving in Florida and California. **Journal of Economic Entomology**, v. 0 n. 0, p. 1-12, 2016.

TAN, X. *et al.* Wireless Sensor Networks Intrusion Detection Based on SMOTE and Random Forest Algorithm. **Sensors**, v.19, 2019.

USDA. **Regulatory impact analysis of the proposed rule to adopt the international standard on wood packing material in international trade.** 2003

WEISS, G. M.; Mining with rarity: A unifying framework. **ACM SICKDD Explorations Newsletter**, v. 6, n. 1, p. 7-19, 2004.

WORLD BANK. World Bank list of economies. Washington DC: 2018. Disponível em: < https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups>. Acesso em: 05 de Junho de 2020.

3. FISCALIZAÇÃO REDUZIDA DE EMBALANGES DE MADEIRA: UMA ANÁLISE SOBRE A ALTERAÇÃO NO COMPORTAMENTO DAS EMPRESAS IMPORTADORAS.

RESUMO

Em dezembro de 2016, entrou em vigor o projeto piloto para a criação de um sistema de fiscalização reduzida das importações acondicionadas em Material de Embalagens de Madeira Regulamentado no Aeroporto (MEMR). O projeto prevê que as 50 empresas com maior volume de importação estão aptas a ter suas importações fiscalizadas por amostragem, caso apresentem taxa de não conformidade das MEMR inferiores a 3%. Este trabalho investiga se a introdução do projeto causou impacto no volume de cargas importadas com irregularidades, levando-se em consideração o período de janeiro de 2016 a dezembro de 2018. Para tanto, foram estimados modelos com dados em painel, com os parâmetros do método de diferenças-em-diferenças. Os resultados dos modelos apontam que não houve impacto significativo com a introdução do sistema de fiscalização reduzida. Não se verificou, tampouco, impacto na probabilidade de não atender ao critério da taxa de não conformidade inferior a 3%. Conclui-se que, apesar do número significativo de informações analisadas, o modelo pode não estar captando, ainda, as consequências dessa mudança na estratégia da política de inspeção, dado o curto horizonte de análise.

Palavras-Chave: Vigiagro, madeira, importação, Canal Verde, Fiscalização Reduzida.

3.1. Introdução

A entrada de pragas via embalagens de madeira é um problema fitossanitário que detém alto potencial em causar prejuízos econômicos, dependendo da probabilidade de seu estabelecimento no país importador e da gravidade dos potenciais danos da praga. O MAPA, por meio do Vigilância Agropecuária Internacional (Vigiagro), realiza a inspeção das cargas comercializadas. Assim, cabe ao Vigiagro impedir a introdução e disseminação de pragas e agentes etiológicos de doenças que possam gerar danos à agropecuária nacional, de modo a garantir a qualidade dos produtos e insumos agropecuários transacionados no comércio exterior (BRASIL, 2006).

Entretanto, os procedimentos de inspeção das Material de Embalagens de Madeira Regulamentado (MEMR), embora necessários, introduzem custos as empresas importadoras e aos órgãos de fiscalização. Para a iniciativa privada, tais custos compreendem aqueles diretos, como a necessidade de troca de embalagens dos produtos importados quando se verificam a presença de não conformidades com os requisitos estabelecidos pelos órgãos reguladores, e indiretos, como os custos de oportunidade das cargas que permanecem mais tempo na área alfandegada, para os procedimentos de inspeção e liberação das cargas. Para o setor público,

alocam-se AFFAs – Auditores Fiscais Federais da Agricultura, recursos físicos (áreas de fumigação) e outros decorrentes da inspeção.

Para tanto, esta pesquisa utilizou como base os dados obtidos sob amparo do projeto piloto para a introdução de um sistema de fiscalização reduzida de MEMR. O projeto estabeleceu uma lista de empresas, avaliadas anualmente, com alto volume de importação e histórico de baixas taxas de não conformidade de MEMR, adotando-se como linha de corte, o limite de 3% de não conformidades. As não conformidades consideradas foram: Presença de casca no MEMR, Ausência de marca comprobatória que o MEMR passou pelos tratamentos requisitados na NIMF-15, Presença de praga no MEMR, Galeria Ativa e Ausência de carimbo.

Uma vez que o sistema de amostragem reduz os custos de importação diretos e indiretos para as empresas importadoras, espera-se que a introdução do sistema de checagem reduzida estimule essas empresas a trabalharem em prol da redução ou manutenção do número de não conformidades nas MEMR a níveis abaixo do estabelecido como limite para a inclusão no sistema de fiscalização reduzida. Nesse caso, pressupõe-se que a redução das taxas de não conformidade pode ser alcançada com uma seleção mais criteriosa, por parte dos importadores, dos fornecedores das embalagens e suportes ou escolha da origem das mercadorias.

Desse modo, o presente trabalho teve como objetivo avaliar o impacto que a implantação do sistema de checagem reduzida deteve sobre as taxas de não conformidade das importações acondicionadas em MEMR, no Aeroporto Internacional de Viracopos, respondendo a seguinte questão: Ser elegível (pertencer à lista “Top 50”) ao sistema de fiscalização reduzida alterou o volume de embalagens e suportes de madeira não conformes dessas empresas?

Para tanto, utilizou-se a metodologia diferenças-em-diferenças, estimando-se os modelos com dados em painel de efeitos fixos, discriminados a nível das empresas (CNPJ), referentes aos anos de 2016 a 2018. Como variáveis dependentes dos modelos, foram adotadas: O número de PCG com não conformidades e uma *dummy* para o caso das empresas que detêm taxa anual de não conformidade superior a 3%. Assim, os modelos com a primeira variável dependente têm como objetivo mensurar o impacto que pertencer a lista “Top 50” teve sobre a quantidade de MEMR com irregularidades. Entretanto, historicamente, algumas empresas registram taxas de não conformidade inferiores a 3%. Assim, assume-se que a introdução da lista “Top 50” não confere estímulos a esse grupo de empresas para a redução das importações com MEMR não conforme. Também há empresas cujas taxas históricas de não conformidade situam-se pouco acima dos 3%, de modo que para se adequarem ao critério de entrada no sistema de fiscalização reduzida são necessárias pequenas alterações no volume de MEMR não

conforme nas cargas importadas. Logo, em um conjunto de empresas com tais características, o impacto sobre a quantidade de MEMR não conformes não apresentaria grandes alterações. Assim, para adequar-se a essa problemática, foram estimados os modelos com a segunda variável dependente. Desse modo, tais modelos mensuram se houve alterações na probabilidade de as empresas não atenderem ao critério da taxa de não conformidade máxima.

O grupo de modelos com a segunda variável dependente configura-se como o do Modelo de Probabilidade Linear (MPL). Entretanto, Horrace e Oaxaca (2003) descrevem que os estimadores do MPL sob conjuntos de dados com probabilidades estimadas superiores à unidade são viesados e inconsistentes. Desse modo, os autores propõem um estimador alternativo ao tradicional estimador de Mínimos Quadrados Ordinários (MQO), denominado Mínimos Quadrados Sequenciais (MQS)[18]. Desse modo, os modelos com a segunda variável dependente foram estimados pelos dois métodos (MQO e MQS).

Além desta introdução, este artigo contém mais cinco seções. A seção 2 apresenta a Metodologia utilizada; a seção 3 apresenta os Resultados e Discussão; segue-se a seção 4 que compreende as Considerações Finais; e finalmente, a última seção 5 que lista as Referências Bibliográficas citadas.

3.2. Metodologia

A hipótese central deste trabalho prescreve que estar elegível ao sistema de fiscalização reduzida pode ter alterado as estratégias de seleção de fornecedores de MEMR e gestão na logística desses materiais por parte dos importadores que realizam operações via Aeroporto Internacional de Viracopos, em Campinas, estado de São Paulo. Entende-se que a introdução do sistema de amostragem reduzida produz incentivos suficientes à redução ou manutenção das taxas de não conformidade inferior a 3%. Cabe mencionar que, antes de dezembro de 2016, 100% das MEMR movimentadas nesse aeroporto passavam pela inspeção do Vigiagro/MAPA e que as 50 empresas selecionadas para o “Top 50” respondiam em 2016 por 64,5% de todas as cargas de MEMR no Aeroporto Internacional de Viracopos.

[18] O MQS, proposto por Horrace e Oaxaca (2003), é descrito como um estimador sequencial para corrigir as probabilidades estimadas via MQO de um modelo MPL. Desse modo, partindo de uma estimação por MQO, as observações utilizadas para a estimação são empregadas no modelo. Com as probabilidades previstas para as observações, excluem-se da amostra as observações cuja probabilidade é superior a 1. A partir da amostra sem as observações excluídas, aplica-se novamente a estimação por MQO. De posse da estimação, o procedimento de exclusão de observações é executado novamente. Esse processo perdura até que não existam observações com probabilidade estimada superior a 1.

Por se tratar de uma variação na qualidade do MEMR importado advinda da introdução de um choque exógeno sobre um determinado grupo de indivíduos, seu impacto pode ser retratado pelo efeito médio do tratamento sobre os tratados (*ATET*). Nessa situação, a introdução da lista "Top 50" configura o efeito do tratamento. Pode-se expressar este impacto pela expressão abaixo:

$$ATET = E(Y^1 - Y^0 | D = 1). \quad (3.1)$$

Em que: *ATET* (*Average Treatment Effect on the Treated*) descreve o efeito do tratamento; Y^1 descreve a variável de interesse com o tratamento; Y^0descreve a variável de interesse sem o tratamento, D indica se o tratamento ocorreu ($D = 1$) ou não ($D = 0$) e $E(.)$ denota o operador de esperança.

A equação (3.1) ainda pode ser escrita como:

$$ATET = E(Y^1 | D = 1) - E(Y^0 | D = 1). \quad (3.2)$$

As equações (3.1) e (3.2) descrevem que o efeito do tratamento sobre os indivíduos tratados é estimado pelo valor esperado do indivíduo com tratamento, subtraído do valor esperado sem tratamento, dada a ocorrência do tratamento. Assim, observa-se que o segundo termo da equação não pode ser observado, uma vez que este versa sobre o valor da variável na ausência do tratamento, após a mesma ter sido impactada pelo tratamento. Entretanto, o *ATET* ainda pode ser expresso pela seguinte equação, cuja derivação está disponível em Lechner (2010):

$$ATET = [E(Y_1 | D = 1) - E(Y_0 | D = 1)] - [E(Y_1 | D = 0) - E(Y_0 | D = 0)]. \quad (3.3)$$

Em que: Y_1 refere-se ao valor de Y no período *1* e Y_0 refere-se ao valor de Y no período *0*, ou seja, anterior ao período de ocorrência do tratamento.

A expressão (3.3) aponta que o efeito do tratamento sobre os indivíduos tratados é equivalente à diferença entre os períodos 0 e 1 do grupo que recebeu tratamento subtraída da diferença entre os períodos 0 e 1 do grupo que não recebeu o tratamento. Ademais, observa-se que todos os elementos dessa expressão podem ser observados.

Dessa forma, para estimar (3.3), propõe-se usar o método de diferenças-em-diferenças. Tal método é amplamente usado em pesquisas cujo interesse é avaliar o impacto de políticas públicas (LECHNER, 2010; IMBENS; WOOLDRIDGE, 2009; WING; SIMON; BELLO-GOMEZ, 2018).

De acordo com Lechner (2010), o procedimento descrito no método de diferenças-em-diferenças computa a diferenças entre as médias da variável resposta do grupo que foi exposto

ao tratamento e as do grupo de controle (sem exposição ao tratamento), após a ocorrência do tratamento e as subtrai das diferenças na variável de resposta entre os grupos antes do tratamento. Assim, para a estimação do método de diferenças-em-diferenças são necessários 4 conjuntos amostrais: Indivíduos que receberam o tratamento após o tratamento; Indivíduos que receberam o tratamento antes do tratamento, Indivíduos do grupo controle após tratamento e Indivíduos do grupo controle antes do tratamento.

No caso do presente artigo, os dados compreendem todas as empresas que importaram via o Aeroporto Internacional de Viracopos, no período de 2016 a 2018[19], acondicionando suas cargas em MEMR. Uma vez que o sistema de fiscalização da madeira foi introduzido em meados de dezembro de 2016 e o critério para estar apto exige que as empresas candidatas façam parte do grupo de 50 maiores importadores do aeroporto, entende-se que o conjunto de dados dispõe dos quatro conjuntos amostrais exigidos para a estimação dos parâmetros associados ao método de diferenças-em-diferenças.

3.2.1. Diferenças-em-Diferenças

Para mensurar o impacto da implantação da lista "Top 50" sobre a quantidade de MEMR não conformes, a seguinte especificação é estimada[20]:

$$Y_{it} = \alpha_i + \lambda_t + \delta T_{it} + X'\beta + \varepsilon_{it}. \qquad (3.4)$$

Em que: Y_{it} refere-se a variável dependente; α_i refere-se ao efeito fixo para o indivíduo *i*; λ_t refere-se ao efeito fixo para o tempo *t (t = 1,2,3)*; δ indica o efeito do tratamento; T_{it} denota dummy cujo valor é 1 caso o indivíduo *i* pertença ao grupo de tratamento (lista "Top 50") no período *t* após tratamento; X' refere-se ao vetor de variáveis exógenas; β refere-se aos coeficientes das variáveis contidas em X' e ε_{it} refere-se ao resíduo com as propriedades usuais.

A equação (3.4) equivale a um modelo com dados em painel, de efeitos fixos para os indivíduos *i* e tempo *t*. Por definição, as características não variantes no tempo serão representadas pelos efeitos fixos[21]. Uma vez que o objetivo do artigo é mensurar a alteração no

---

[19] Uma vez que a Instrução Normativa n° 32, de 15 de setembro de 2015, alterou os procedimentos e critérios de verificação das cargas, os dados de inspeção dos períodos anteriores e posteriores a alteração não são comparáveis.
20 A especificação da regressão do método de diferenças-em-diferenças para dois períodos inclui variáveis dummy para: Tempo, Grupo de Controle e Grupo de Controle pós Tratamento. A partir de 3 períodos, pode-se generalizar o método em um modelo de dados em painel com efeitos fixos. Para mais detalhes, ver: Imbens e Wooldridge (2009) e Wing, Simon e Bello-gomez (2018).
21 Em um modelo de dados em painel com efeitos fixos, os fatores não variantes no tempo são representados pelos efeitos individuais de cada indivíduo, de forma que não é possível estimar os parâmetros dessas variáveis (GREENE, 2007).

perfil de cargas com não conformidades das empresas, considerou-se como conjunto de indivíduos as empresas importadoras atuantes no Aeroporto Internacional de Viracopos, na frequência anual, nos anos de 2016 a 2018. O Coeficiente δ denota o efeito do tratamento sobre a variável dependente. Nesse caso, denota-se que as empresas da lista "Top 50" foram alocadas no grupo de tratamento, enquanto as demais empresas importadoras pertencem ao grupo de controle.

Tendo em vista a restrição do número de períodos disponíveis dos (somente três anos) pressupõe-se que a empresa está no seu equilíbrio de curto-prazo. Assim, a dotação de capital e tecnológica de cada firma são tratadas como fixas no período. Logo, apenas a quantidade de insumos variáveis se altera ao longo do tempo.

Como variáveis presentes no vetor de variáveis exógenas, temos: Número PCG importados pela empresa *i* no ano *t* e os índices de caracterização das importações. Assim, o número de PCG importados tem como objetivo ser uma proxy do nível de atividade da empresa.

Além da variável responsável por controlar o nível e importação da empresa, controlaram-se as características de origem das importações, uma vez que a origem da carga[22] é um fator determinante para a não conformidade da embalagem de madeira (MONTEFERRANTE *et. al*, 2018). Para controlar as características de origem nos modelos estimados, foram compostos índices de importação, derivados da Análise Fatorial de Dados Mistos. Estes índices compreendem as seguintes variáveis: Tamanho do aeroporto, Volume de PCG exportadas para o Aeroporto Internacional de Viracopos, Taxa de Não conformidade dos PCG com MEMR Exportadas para o Aeroporto Internacional de Viracopos, Adesão do país de origem dos voos com carga à NIMF-15 e classificação do país de origem quanto ao nível de renda.

Como variável dependente da equação (3.4), inicialmente foi adotada a quantidade de PCG acondicionadas em embalagens de madeira nas quais foram constatadas não conformidades, da empresa *i* e no ano *t*. Entretanto, o critério estabelecido para a permanência na lista de empresas no sistema de fiscalização reduzida denota que a taxa de não conformidade anual de cada empresa não ultrapasse 3%. Sendo assim, empresas com taxas históricas de não conformidade abaixo de 3% não são estimuladas a alterar seu perfil de importações com a introdução da lista "Top 50", enquanto as empresas com taxas históricas pouco acima dos 3% requerem pequenas alterações para adequar-se ao critério. Logo, em um conjunto de empresas

[22] Entende-se como origem da carga o país e aeroporto de procedência do voo que adentrou ao aeroporto analisado com a carga. Em diversos casos, importações com conexões em diversos aeroportos são registradas como provenientes do último aeroporto que estiveram antes de adentrar ao aeroporto em questão.

com tais características o impacto sobre a quantidade de MEMR não conformes não apresentaria grandes alterações. Para adequar-se a essa problemática, foram estimados modelos de probabilidade linear[23], cuja variável dependente denota valor 1 para o indivíduo *i* cuja taxa de não conformidade é superior a 3% no ano *t* e 0, caso contrário.

Horrace e Oaxaca (2003) descrevem que os estimadores do modelo de probabilidade linear sob conjuntos de dados com probabilidades estimadas superiores a 1 são viesados e inconsistentes. Assim, os autores propõem um estimador alternativo, denominado Mínimos Quadrados Sequenciais (MQS). Para adequar-se a essa limitação, o modelo de probabilidades linear foi estimado via Mínimos Quadrados Ordinários e Mínimos Quadrados Sequenciais.

Ademais, parte das 50 empresas da amostra importaram apenas em um ou dois períodos. Para manter o conjunto de dados balanceado[24], foram construídos dois conjuntos amostrais: O primeiro com as empresas que importaram, pelo menos uma vez por ano, entre 2016 e 2018 e o segundo com todas as empresas. Assim, foram estimados modelos com os dois grupos amostrais, para ambas as variáveis dependentes e pelo método de Mínimos Quadrados Ordinários e Mínimos Quadrados Sequenciais.

3.2.1.1. Pressuposições do método Diferenças-em-Diferenças

Segundo Lechner (2010), o parâmetro δ denotará o feito causal do tratamento, caso algumas pressuposições sejam atendidas, sendo essas: 1. Estabilidade do Efeito do Tratamento; 2. Exogeneidade das variáveis dependentes; 3. Independência do tratamento na população tratada, no período antes do tratamento; 4. Tendências paralelas entre os grupos de controle e tratamento.

A primeira pressuposição versa sobre a ausência de externalidades, causadas pelo tratamento, entre os membros da população analisada. Assim, pressupõe-se que a introdução da lista “Top 50” não alterou a relação de cada uma das empresas com as demais importadoras via o Aeroporto Internacional de Viracopos, nem foi capaz de induzir à alterações da não conformidade das MEMR movimentadas pelas demais empresas.

---

[23] Os modelos Logit e Probit são usualmente preteridos ao modelo de probabilidade linear, uma vez que não apresentam efeitos marginais constantes. Entretanto, por se tratar de modelos não lineares, a interpretação do coeficiente da variável de tratamento não é trivial (AI; NORTON, 2003; KARACA-MANDIC; NORTON; DOWD, 2012).

[24] Usualmente um conjunto de dados é denotado como balanceado quando todos os indivíduos da amostra possuem valores observados ao longo de todos os períodos.

A pressuposição de exogeneidade das variáveis dependentes indica que os regressores não são influenciados pelo tratamento. Uma vez que decisão de importar envolve a estrutura de custos das empresas, bem como o nível de produção, pertencer a um conjunto de empresas que estão aptas à fiscalização reduzida do MEMR das cargas importadas, por amostragem, não aparenta induzir a alterações no volume importado. Assim como na relação entre o volume de PCG importados e a quantidade de MEMR não conformes, os índices de importação são compostos por variáveis intrínsecas aos aeroportos de origem, logo, não há relação entre a lista "Top 50" e essas variáveis.

Quanto à terceira pressuposição, esta versa sobre a independência da variável dependente com o tratamento avaliado (LECHNER, 2010). Assim como destacado por Adami, Sun e Miranda (2017), a escolha das empresas integrantes da lista "Top 50" ocorreu a partir do volume de importações acondicionadas em MEMR nos anos de 2014 a 2016. Logo, não se espera que o tratamento influencie na variável dependente antes do ano de 2017.

A quarta pressuposição versa sobre a tendência da variável dependente nos grupos: tratamento e controle. Pressupõe-se que a tendência da variável dependente do grupo de tratamento seja paralela à tendência do grupo de controle. Ou seja, assume-se que tanto o grupo de controle quanto o grupo de tratamento deverão apresentar a mesma tendência ao longo do tempo. Dentre os modos de análise do paralelismo da tendência, diversos trabalhos da literatura aplicada optam pela verificação visual (WING; SIMON; BELLO-GOMEZ, 2018). Os resultados para a verificação visual da tendência nas variáveis dependentes de todos os modelos estão disponíveis na seção de resultados.

3.2.2. Variáveis de Origem

A fim de controlar os efeitos da origem das importações, aqui compreendida como origem o aeroporto de saída da aeronave e não a origem das cargas movimentadas ou dos MEMR, foram construídos índices para a caracterização das origens das importações. Os índices são construídos a partir dos escores produzidos via Análise Fatorial de Dados Mistos.

Assim, para compor os escores são utilizadas as seguintes variáveis: Tamanho do aeroporto de origem, Número de PCG importados via Aeroporto Internacional de Viracopos do aeroporto de origem, Taxa de não conformidade dos MEMR do aeroporto de origem, Adesão do país de origem aos critérios de inspeção dos MEMR estabelecidos na NIMF-15 e

Classificação do país de origem quanto ao nível de renda. A fonte destes dados é discriminada adiante.

3.2.2.1. Análise Fatorial de Dados Mistos

De acordo com Kassambara (2017) a Análise Fatorial de Dados Mistos é utilizada em conjuntos de dados que contém variáveis categóricas e contínuas. Assim, ela pode ser compreendida como a junção da Análise de Componentes Principais (PCA), para o caso das variáveis contínuas e da Análise de Correspondência Multiplica (MCA) para o caso das variáveis categóricas.

Segundo Varella (2008) a PCA é realizada a partir dos autovalores e auto vetores da matriz de correlação ou matriz de covariância de um determinado conjunto de dados. Desse modo, a partir da extração dos autovalores das matrizes, são calculados os autovetores normalizados. A combinação linear dos termos dos autovetores normalizados confere os escores de cada um dos componentes principais[25]. O mesmo procedimento, com algumas adaptações, é realizado na MCA para obtenção dos escores[26] (ABDI; VALENTIN, 2007).

3.2.2.2. Índice Anual de Importações

A partir dos escores obtidos na Análise Fatorial de Dados Mistos, calculam-se os índices anuais de importação. Optou-se por calcular *k* índices, a partir do número de escores, e que correspondam, em conjunto, a pelo menos 84% da variância total dos dados. Assim, o índice de importação de cada componente fatorial é dado por:

$$X_{itk} = \sum_{J_{it}=1}^{J_{it}} \phi_{j_{it}} w_{jtk}. \tag{3.5}$$

Em que: X_{itk} refere-se ao índice de importação *k* do indivíduo *i* no tempo *t*; j_{it} refere-se ao aeroporto *j* de origem da carga cuja empresa *i* importou no tempo *t*, assim $j_{it} \in \{1,..,J_{it}\}$; $\phi_{j_{it}}$ refere-se ao peso que o aeroporto de origem *j* ocupa no volume de cargas importado em MEMR da empresa *i* no tempo *t*; w_{jtk} refere-se ao escore do componente *k* do aeroporto *j* no tempo *t*.

[25] Para mais detalhes, ver Varella (2008).
[26] Derivação algébrica disponível em: Abdi e Valentim (2007)

De acordo com a equação (3.5), o índice de importação *k* pode ser entendido como a média ponderada pela pauta de importação das empresas dos escores dos aeroportos. Sendo assim, esses fornecem um indicador da qualidade da origem da carga importada, embora não sejam facilmente interpretáveis. Da mesma forma, os coeficientes estimados para os índices não são trivialmente interpretáveis.

3.2.3. Base de dados

Os dados referentes a não conformidades nas embalagens de madeira, volume de cargas importadas e volume de cargas sujeitas à inspeção foram cedidos pela parceria entre o Aeroporto Internacional de Viracopos, o Serviço de Vigilância Agropecuária Internacional (Vigiagro) e o Cepea-Fealq.

A identificação dos países que participam da CIPV e, portanto, são adotantes da NIMF-15, foi feita com base em informações da *Food and Agriculture Organization* (FAO). Os dados referentes à classificação das nações quanto ao nível de Renda estão disponíveis no Banco Mundial (WORLD BANK, 2018). Por fim, os dados de classificação dos aeroportos foram obtidos do projeto *OurAirports*.

Para a composição da base de dados, foram excluídos os registros de importação cuja origem não é compatível com a codificação *International Air Transport Association* (IATA) para Aeroportos. Da mesma forma foram excluídos os registros de importação que não apresentaram informações compatíveis com o CNPJ para as empresas.

3.3. Resultados

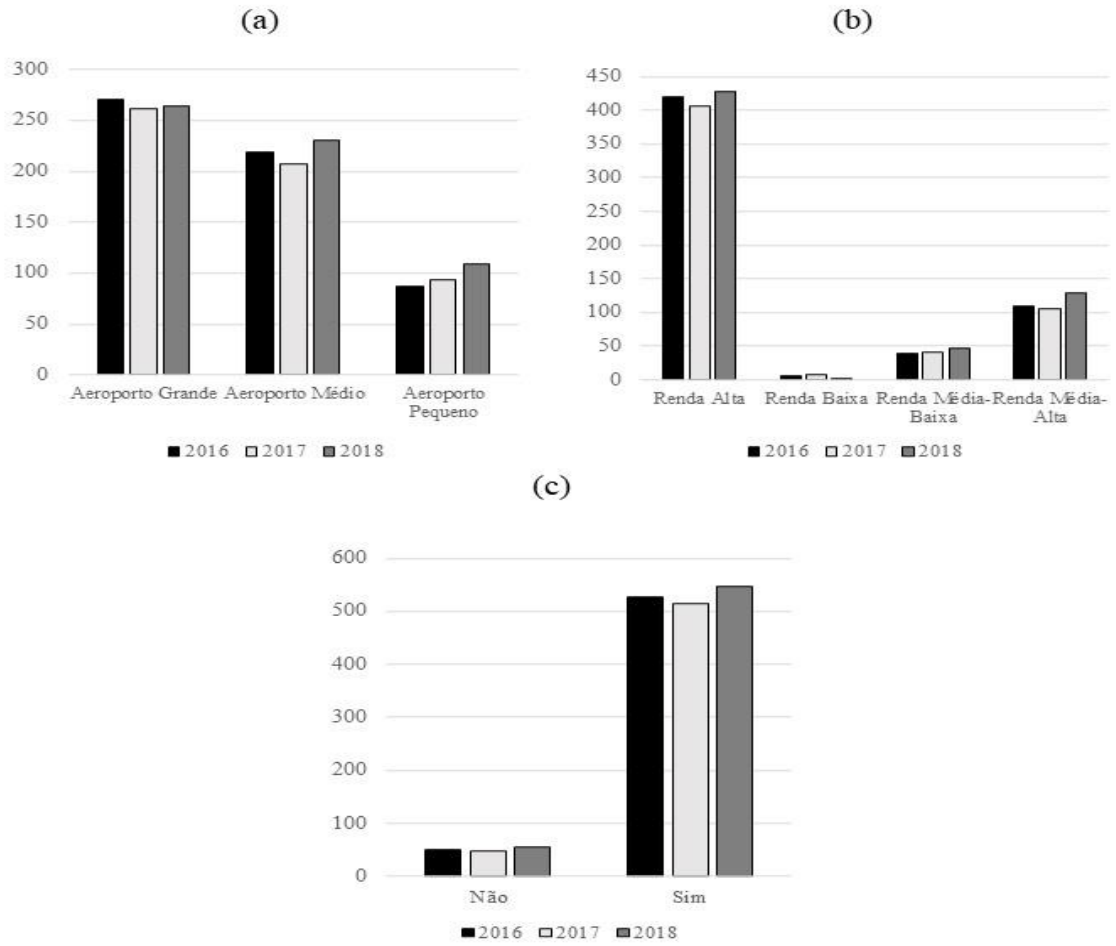
Esta seção apresenta os resultados das estimações dos modelos propostos na seção anterior. As estimativas, bem como todo o tratamento de dados foram realizados no *software R versão 3.6.3*.

3.3.1. Índice Anual de Caracterização da Procedência das Importações

Para controlar os efeitos da origem das cargas importadas nos modelos estimados, foram compostos índices chamados de importação por meio da Análise Fatorial de Dados Mistos. Para compor o índice, consideraram-se as seguintes características: Tamanho do Aeroporto, Número

de PCG totais do aeroporto de origem para o Aeroporto Internacional de Viracopos, Taxa de Não conformidade das cargas exportadas com MEMR para o aeroporto em análise, Adesão do país de origem dos voos com cargas em MEMR à NIMF-15 e classificação do país de origem quanto ao nível de renda. A Figura 3.1 apresenta as frequências absolutas das variáveis categóricas, nos anos: 2016, 2017 e 2018.

Figura 3.1: Frequências Absolutas em número de MEMR movimentado por Classificação dos Aeroportos de Origem (a), Classificação do Nível de Renda do País de Origem (b) e Adesão do país de Origem à NIMF-15 (c) no Aeroporto Internacional de Viracopos, no período de 2016 a 2018.



Fonte: Elaboração Própria.

De acordo com a Figura 3.1, as frequências absolutas em todas as variáveis são estáveis ao longo do período avaliado. Assim, ao longo do período, nota-se que a maior parte das cargas importadas com embalagens de madeira provêm de nações com nível de renda classificado como “Alta” ou “Média-Alta”. Da mesma forma, a maior parte dos países de onde se originam

os voos com as cargas de importados acondicionados em MEMR adotam os requisitos preconizados pela NIMF-15. Quanto ao tamanho dos aeroportos, a maior parte é classificado como “Grande” ou “Médio”.

A Tabela 3.1 apresenta os resultados das estatísticas descritivas para o número de PCG com MEMR importados via Aeroporto Internacional de Viracopos e a taxa de não conformidade dos MEMR.

Tabela 3.1: Estatísticas Descritivas do Número de PCG com MEMR Importados e Taxa de não conformidade dos MEMR, no período de 2016 a 2018.

Ano	Média	Mediana	Desvio Padrão	Mínimo	Máximo
Número de PCG com MEMR Exportadas ao Aeroporto Internacional de Viracopos					
2016	748,90	50,00	3.663,35	1,00	75.657,00
2017	922,80	61,00	4.370,54	1,00	83.072,00
2018	948,00	54,00	4.769,11	1,00	91.933,00
Taxa de Não conformidade das Embalagens de Madeira					
2016	0,0600	0,0000	0,1673	0,000	1,000
2017	0,0846	0,0075	0,1933	0,000	1,000
2018	0,0856	0,0101	0,1938	0,000	1,000

Fonte: Elaboração própria.

De acordo com a Tabela 3.1, tanto no número de PCG com MEMR, quanto nas taxas de não conformidade, não houve diferenças significativas entre os anos de 2017 e 2018. Entretanto, ambos períodos apresentaram valores superiores ao ano de 2016. Embora as médias das taxas de não conformidade sejam superioras a 3% (valor estabelecido como limite para a manutenção de uma Empresa no sistema de fiscalização reduzida) a mediana foi inferior a esse valor nos três anos do período.

A Tabela 3.2 apresenta os resultados dos componentes da Análise Fatorial de Dados Mistos realizada com as variáveis descritas acima.

Tabela 3.2: Autovalores e Porcentagem da Variância dos Componentes do Índice Anual de Importação.

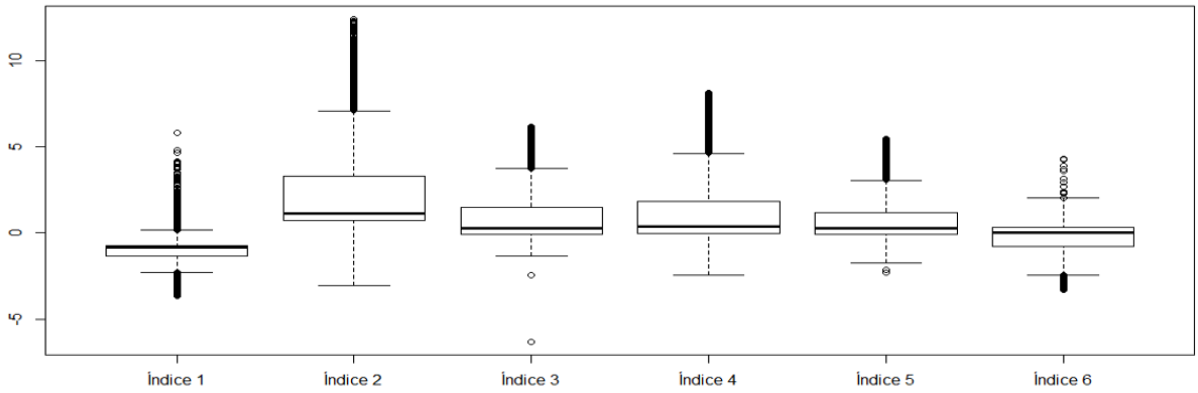
	Autovalor	Porcentagem da Variância representada pelo Autovalor (em %)	Porcentagem Cumulativa da Variância representada pelos Autovalores (em %)
Componente 1	1,600766	20,00958	20,00958
Componente 2	1,201193	15,01491	35,02449
Componente 3	1,064731	13,30913	48,33362
Componente 4	1,021299	12,76623	61,09986
Componente 5	0,941450	11,76812	72,86798
Componente 6	0,922367	11,52959	84,39757
Componente 7	0,783482	9,79352	94,19110
Componente 8	0,464712	5,80809	100,00000

Fonte: Elaboração própria.

A Tabela 3.2 apresenta os Autovalores, Porcentagem da variância de cada autovalor e variância acumulada dos autovalores. A Análise Fatorial de Dados Mistos das variáveis dos aeroportos produziu oito componentes. Optou-se por escolher os seis primeiros para compor as variáveis de origem dos modelos em painel, uma vez que cada um dos componentes, captam no mínimo 10% da variância do grupo de variáveis. Ademais, os seis componentes juntos representam 84,4% da variância total do grupo de variáveis dos aeroportos.

A partir dos seis componentes selecionados, calculou-se, com base na expressão (5), os Índices Anuais de Caracterização da Origem das Importações para cada componente. A Figura 3.2 apresenta a distribuição dos índices.

Figura 3.2: *Boxplot* dos índices de importação construídos.



Fonte: Elaboração própria.

De acordo com a Figura 3.2, a média de todos os índices apresentou valores próximos entre si. Por fim, cabe ressaltar que os índices são gerados a partir da média ponderada dos componentes obtidos na Análise Fatorial de Dados Mistos, de forma que sua interpretação não é trivial. Entretanto, destaca-se que o intuito original dos índices é controlar as características do perfil da origem das importações das empresas, e não inferir sobre seus resultados.

3.3.2. Diferenças-em-Diferenças

A fim de verificar a efetividade dos estímulos advindos do sistema de fiscalização reduzida de MEMR, através da lista “Top 50”, estimaram-se modelos com dados em painel para as empresas, com os parâmetros de interação previstos no método de diferenças-em-diferenças (*diff-in-diff*). Entretanto, a aplicação do método exige que as tendências da variável dependente, dos grupos de controle e tratamento, sejam paralelas. Desse modo, a Tabela 3.3 exibe a média, por empresa, do número de MEMR importados anualmente com não

conformidades, no período de 2016 a 2018. Espera-se que o comportamento das médias em cada grupo seja semelhante, de modo a não violar a pressuposição de paralelismo da tendência.

Tabela 3.3: Médias por empresa do número de PCG importados, acondicionados em MEMR, que registraram não conformidades do Aeroporto Internacional de Viracopos. De 2016 a 2018

	2016	2017	2018
Modelo Empresas Todos os Anos			
Tratamento (“Top 50”)	13,8275	17,3706	19,5431
Crescimento Anual (em %)	-	25,62%	12,50%
Controle	1,0139	1,4949	1,7996
Crescimento Anual (em %)	-	47,44%	20,38%
Modelo Todas as Empresas			
Tratamento (“Top 50”)	10,1988	11,7670	13,0170
Crescimento Anual (em %)	-	15,37%	10,62%
Controle	0,3277	0,4774	0,5705
Crescimento Anual (em %)	-	45,67%	19,48%

Fonte: Elaboração Própria.

De acordo com a Tabela 3.3, em ambos os grupos amostrais (empresas que importaram todos os anos e todas as empresas), a quantidade de embalagens de madeira com não conformidades elevou-se ao longo dos períodos analisados. O crescimento da quantidade de MEMR não conformes em 2017 foi superior ao de 2018, nos grupos de controle e tratamento, em ambos os conjuntos de dados. Assim, a Tabela 3.3 aponta que, entre 2016 a 2018, os grupos de controle e tratamento detiveram a mesma tendência para a quantidade média de MEMR não conformes. Logo, a inspeção visual não apontou para diferenças entre a tendência dos grupos de controle e tratamento, de modo que a pressuposição de tendências paralelas não foi violada, nos dois conjuntos de dados.

A Tabela 3.4 apresenta os coeficientes dos modelos com dados em painel para a amostra com todas as empresas e amostra com as empresas que importaram em todos os anos. Ambos os modelos detêm como variável dependente o número de cargas importadas acondicionadas em MEMR com não conformidades.

Tabela 3.4: Parâmetros dos modelos para a quantidade de PCG com MEMR com não conformidades para o conjunto das empresas que importaram todos os anos e para o conjunto de todas empresas importadoras (Erros-padrão robustos)

Variáveis	Coeficientes	*P-Valor*	Coeficientes	*P-Valor*
	Modelo Empresas Todos os Anos		Modelo Todas as Empresas	
Quantidade de MEMR inspecionadas	0.0147237 (0.0031774)	0.000***	0.0148552 (0.0030901)	0.000***
Quantidade de PCG importados	0.0041804 (0.0023181)	0.071*	0.0042439 (0.0020193)	0.036**
Indice de Importação 1	0.1213021 (0.2011497)	0.547^{ns}	0.015873 (0.0414884)	0.702^{ns}
Indice de Importação 2	-0.4382175 (0.1816595)	0.016**	-0.0845664 (0.0316702)	0.008***
Indice de Importação 3	1.3572288 (0.3352806)	0.000***	0.4161877 (0.0578315)	0.000***
Indice de Importação 4	0.4077386 (0.2645122)	0.123^{ns}	0.1104054 (0.0473495)	0.020**
Indice de Importação 5	0.061505 (0.2456874)	0.802^{ns}	-0.0511351 (0.0512943)	0.319^{ns}
Indice de Importação 6	1.8776457 (0.3809785)	0.000***	0.5971339 (0.0692578)	0.000***
Tratamento (Top 50 x Ano 2017)	2.8829502 (1.8606756)	0.121^{ns}	1.7181241 (1.163127)	0.140^{ns}
R^2	0.42117		0.4153	
Teste F	0.000***		0.000***	
Empresas	1409		5506	

Fonte: Elaboração própria. (^{ns}) denota não significância (*) denota significância à 10%, (**) denota significância à 5% e (***) denota significância à 1%,.

Em ambas as equações estimadas, a Quantidade de PCG importados e Quantidade de MEMR inspecionadas tem impacto positivo sobre a quantidade de cargas em MEMR com não conformidades. Ademais, quanto às variáveis de origem, representadas pelos índices de importação, verifica-se que foram estatisticamente significativas, em ambas as equações: Índice de Importação 2, Índice de Importação 3 e Índice de Importação 6. Ademais, o Índice de Importação 4 foi estatisticamente significativo no modelo com todas as empresas. Assim, embora não sejam diretamente interpretáveis, tais parâmetros denotam que as variáveis de origem são um fator importante na estimação das equações para a quantidade de MEMR não conformes.

Por fim, quanto ao parâmetro que mede o efeito do Tratamento, ou seja, o impacto da empresa pertencer a lista “Top 50”, este não foi significativo para ambos os modelos. Logo, ser elegível ao sistema de fiscalização reduzida (pertencer à lista “Top 50”) não causou impacto sobre o número de cargas de MEMR com não conformidades, considerados os três anos examinados.

Embora os resultados dos modelos para o número de MEMR não conformes demonstrem que a lista “Top 50” não levou a redução do número de PCG com MEMR importadas não conformes, cabe ressaltar que o critério para inclusão de empresas no sistema

de fiscalização reduzida é registrar taxa de não conformidade inferior a 3%. Assim, como destacado anteriormente, empresas com taxas históricas de não conformidade abaixo de 3% não possuem estímulos para alterar seu perfil de importações, enquanto as empresas com taxas históricas pouco acima dos 3% requerem pequenas alterações para adequar-se ao critério.

Para adequar-se a essa problemática, propõe-se a estimação de modelos de probabilidade linear (pelos métodos de Mínimos Quadrados Ordinários e Mínimos Quadrados em Sequência), cuja variável dependente denota valor 1 para os indivíduos com taxas de não conformidade acima de 3% e 0, caso contrário. Assim, o valor do coeficiente da variável tratamento mede o impacto de pertencer a lista "Top 50" sobre a probabilidade de apresentar taxa de não conformidade acima dos 3%.

Assim como no modelo para o número de cargas importadas com não conformidades, é necessário garantir que a variável dependente de ambos os grupos (Controle e Tratamento) detenha a mesma tendência. A Tabela 3.5 apresenta a frequência relativa de empresas com taxas de não conformidades superiores 3%, no período de 2016 a 2018.

Tabela 3.5: Frequência relativa das empresas com taxas de não conformidade superior a 3% ao ano, no período de 2016 a 2018. Modelos de MQO e MQS.

	2016	2017	2018	2016	2017	2018
	MQO			MQS		
Modelo Empresas Todos os Anos						
Tratamento ("Top 50")	0,2241	0,3189	0,4051	0,0823	0,1176	0,2235
Crescimento Anual (em %)	-	42,30%	27,02%	-	42,85%	90,00%
Controle	0,2505	0,3232	0,3758	0,1200	0,1526	0,2052
Crescimento Anual (em %)	-	29,01%	16,26%	-	27,19%	34,48%
Modelo Todas as Empresas						
Tratamento ("Top 50")	0,2045	0,2500	0,3011	0,1411	0,1901	0,2453
Crescimento Anual (em %)	-	22,22%	20,45%	-	34,78%	29,03%
Controle	0,1067	0,1347	0,1559	0,0889	0,1173	0,1390
Crescimento Anual (em %)	-	26,18%	15,73%	-	31,96%	18,49%

Fonte: Elaboração Própria.

De acordo com a Tabela 3.5, nos modelos estimados por MQO e MQS, a frequência de indivíduos com taxa de não conformidade acima dos 3% aumentou ao longo dos anos do período amostrado, nos grupos de controle e tratamento. Nos modelos estimados por MQO, o aumento foi superior em 2017, em comparação a 2018. No caso do conjunto de dados com as empresas que importaram todos os anos no modelo estimado via MQS, a elevação foi superior no ano de 2018, nos grupos de controle e tratamento. Assim, mesmo com o procedimento de correção adotado na estimação via MQS, os grupos de controle e tratamento não apresentaram tendências diferentes no comportamento da variável dependente. Logo, não há indícios de violação da pressuposição de paralelismo da tendência nos modelos de probabilidade linear.

A Tabela 3.6 apresenta os resultados para a estimação do modelo de probabilidade linear pelo método de mínimos quadrados ordinários (MQO).

Tabela 3.6: Parâmetros dos modelos para a probabilidade de não atender ao critério de entrada no sistema de fiscalização reduzida, com o conjunto amostral de todas as empresas e conjunto amostral das empresas que importaram durante todos os anos, estimados via mínimos quadrados ordinários (erros padrões robustos).

Variáveis	Coeficientes	*P-Valor*	Coeficientes	*P-Valor*
	Modelo Empresas Todos os Anos		Modelo Todas as Empresas	
Quantidade de PCG importados	-0.0000064894 (0.0000081924)	0.428^{ns}	0.000032488 (0.00001289)	0.012**
Indice de Importação 1	0.074016 (0.01988)	0.000***	-0.039701 (0.012049)	0.001***
Indice de Importação 2	-0.080561 (0.013727)	0.000***	-0.0025236 (0.0045607)	0.580^{ns}
Indice de Importação 3	0.22722 (0.031888)	0.000***	0.19285 (0.023963)	0.000***
Indice de Importação 4	0.085816 (0.027729)	0.002***	-0.031393 (0.020005)	0.117^{ns}
Indice de Importação 5	-0.01928 (0.030433)	0.526^{ns}	-0.051637 (0.020164)	0.010***
Indice de Importação 6	0.2406 (0.029948)	0.000***	0.19009 (0.016666)	0.000***
Tratamento (Top 50 x Ano 2017)	0.040526 (0.038668)	0.294^{ns}	0.032232 (0.028779)	0.263^{ns}
R^2	0.034588		0.07085	
Teste F	0.000***		0.000***	
Empresas	1409		5506	

Fonte: Elaboração própria. (^{ns}) denota não significância (*) denota significância à 10%, (**) denota significância à 5% e (***) denota significância à 1%,..

De acordo com a Tabela 3.6 os índices de importação 1, 3 e 6 foram significativos em ambas as equações sobre a variável dependente “probabilidade de não atender aos requisitos para entrar no sistema de fiscalização reduzida”. Os índices 2 e 4 foram significativos na equação representativa dessa probabilidade para o conjunto das empresas que importaram todos os anos, enquanto o índice 5 foi significativo no modelo para todas as empresas importadoras. Logo, assim como na equação para o número de PCG com MEMR importados com não conformidades, as variáveis de origem mostraram-se importantes para a estimação das equações da probabilidade de não estar elegível a participar do sistema de fiscalização reduzida.

Quanto ao número de PCG importados, esse detém impacto apenas na equação para todas as empresas. Ademais, nos modelos para ambos os grupos amostrais a variável “Tratamento” não foi significativa. Assim, a introdução da política não deteve impacto sobre a probabilidade de cumprir com o requisito para estar elegível ao grupo de participantes do sistema de fiscalização reduzida.

Para adequar-se à problemática exposta Horrace e Oaxaca (2003), o mesmo modelo de probabilidade foi estimado via Mínimos Quadrados Sequenciais. A estimação por esse método acarretou a exclusão de 374 indivíduos da amostra que contém as empresas que importaram em todos os anos, enquanto 138 empresas foram excluídas do conjunto que contém todas as empresas importadoras via Aeroporto Internacional de Viracopos, que operaram com MEMR. Os coeficientes estimados são reportados na Tabela 3.7.

Tabela 3.7: Parâmetros dos modelos para a probabilidade de não atender ao critério de entrada no sistema de fiscalização reduzida, com o conjunto amostral de todas as empresas e conjunto amostral das empresas que importaram durante todos os anos, estimados via mínimos quadrados sequenciais (erros padrões robustos).

Variáveis	Coeficientes	*P-Valor*	Coeficientes	*P-Valor*
	Modelo Empresas Todos os Anos		Modelo Todas as Empresas	
Quantidade de PCG importados	-0.000004526 (0.000010627)	0.6702^{ns}	0.000045799 (0.000015271)	0.003***
Indice de Importação 1	0.096569 (0.020981)	0.000***	-0.036224 (0.010794)	0.001***
Indice de Importação 2	-0.10007 (0.017544)	0.000***	-0.0091219 (0.00536)	0.089*
Indice de Importação 3	0.24094 (0.034891)	0.000***	0.20889 (0.021328)	0.000***
Indice de Importação 4	0.11555 (0.028964)	0.000***	-0.026784 (0.017599)	0.128^{ns}
Indice de Importação 5	0.016317 (0.030855)	0.597^{ns}	-0.045151 (0.017844)	0.011**
Indice de Importação 6	0.29424 (0.040391)	0.000***	0.21319 (0.018624)	0.000***
Tratamento (Top 50 x Ano 2017)	0.03266 (0.042882)	0.446^{ns}	0.036309 (0.030587)	0.235^{ns}
R^2	0.026519		0.057249	
Teste F	0.000***		0.000***	
Empresas	1035		5368	

Fonte: Elaboração própria. (^{ns}) denota não significância (*) denota significância à 10%, (**) denota significância à 5% e (***) denota significância à 1%,.

Assim como na estimação do modelo de probabilidade linear por Mínimos Quadrados Ordinários, as variáveis de origem foram significativas em ambos os modelos, sendo essas: Índice de Importação 1, Índice de Importação 2, Índice de Importação 3 e Índice de Importação 6. Ademais, o Índice de Importação 4 foi significativo na equação das empresas que importaram em todos os anos, enquanto o Índice de Importação 5 na equação para todas as empresas. Novamente o coeficiente do número de PCG importados foi significativo na equação para todas as empresas. Por fim, as variáveis de tratamento não foram significativas nas equações para os dois grupos amostrais.

Portanto, em termos gerais, as variáveis do grupo responsável por controlar os efeitos de origem foram significativas nos modelos estimados, embora sua interpretação não seja

trivial. Verificou-se ainda, em todos os modelos, que o parâmetro que indica a efetividade da política não foi significativo. Ou seja, as evidências apontam que ser elegível ao sistema de fiscalização reduzida (pertencer a lista “Top 50”) não reduziu o número de PCG com MEMR importadas com não conformidades das empresas. Da mesma forma, não há indícios que a probabilidade de deter taxas de não conformidade superioras a 3% alteraram-se após o início do projeto. Tais resultados indicam que ser elegível ao sistema de fiscalização reduzida não modificou o comportamento das empresas em prol de importações com menor índice de não conformidade, pelo menos no curto período que foi examinado neste estudo.

3.4. Considerações Finais

No início de 2016, foi introduzido o sistema de checagem reduzida das embalagens de madeira no Aeroporto Internacional de Viracopos. O sistema institui que um grupo de empresas com alto volume de importação e taxas de não conformidades em MEMR inferiores a 3% deteriam suas importações fiscalizadas por amostragem. Este trabalho procurou avaliar o impacto da introdução do sistema de checagem reduzida sobre a taxa de não conformidade dos PCG com MEMR. Para tanto, utilizou-se modelos com estimadores de diferenças-em-diferenças, a fim de verificar se ser elegível ao sistema de fiscalização reduzida alterou o volume de MEMR importados com não conformidades e se a fiscalização por amostragem reduziu a probabilidade de uma empresa deter taxas de não conformidade superioras a 3%. Desse modo, foram estimados modelos de diferenças-em-diferenças para a quantidade de MEMR importados por CNPJ e aderência a taxa de não conformidade superior a 3% por CNPJ, no período de 2016 a 2018.

Os resultados para a estimação dos modelos com a quantidade de PCG com MEMR importados como variável dependente apontaram que ser elegível ao sistema de fiscalização reduzida não alterou a quantidade de PCG com MEMR com não conformidades. Da mesma forma, os modelos com a taxa de não conformidade dos PCG com MEMR por CNPJ como variável dependente não demonstraram alterações na taxa de não conformidade para o grupo de empresas elegíveis ao sistema de fiscalização reduzida após a introdução do sistema. Logo, não há evidências que a introdução do sistema de amostragem alterou o comportamento das não conformidades das embalagens de madeira que acondicionam as importações no Aeroporto Internacional de Viracopos.

Por fim, a análise exposta nesse trabalho contempla apenas três anos de importação com MEMR. Nesse período, apenas dois ocorreram após a aplicação da lista “Top 50”. Logo, a

inefetividade da lista “Top 50” pode estar relacionada à fatores como: O breve período de existência do sistema de fiscalização reduzida não fora o suficiente que as empresas da lista “Top 50” remanejassem a importação de insumos em prol de fornecedores com taxas de não conformidades menores, ou exigissem dos seus fornecedores atuais MEMR/processos de importação melhores; O período analisado não foi suficiente para que as empresas contabilizassem os ganhos com a participação no sistema de fiscalização reduzida, de modo que essas não aplicaram medidas em prol da importação com MEMR com menores níveis de não conformidade; A substituição das importações com maior nível de não conformidade exige mudanças em processos estruturais das empresas, de modo que as alterações ocorrem a médio/longo prazo, e não no equilíbrio de curto prazo, como pressuposto nos modelos. Portanto, é factível esperar que as mudanças advindas da introdução da lista “Top 50” ocorram nos próximos anos.

Referências

ABDI, H.; VALENTIN, D. Multiple Correspondence Analysis. IN: SALKIND, N. ENCYCLOPEDIA OF MEASUREMENT AND STATISTICS. Universidade do Texas, Dallas, Texas, 2007.

ADAMI, A. C. O.; MIRANDA, S. H. G. Análise das informações das Inspeções de embalagens e suportes de madeira em 2018, Reavaliação dos critérios para o Canal Verde e Análise dos Benefícios da implementação do canal. Abr./2019.

ADAMI, A. C. O.; SUN, L. T.; MIRANDA, S. H. G. Estudo para Implantação de Canais de Inspeção na Fiscalização de Embalagens, Suportes e Peças de Madeira no Aeroporto Internacional de Viracopos – Campinas/SP: Atualização dos dados com inclusão de agosto de 2017, reavaliação da Lista Piloto de Seleção e Exclusão de empresas do Canal, Levantamento preliminar dos benefícios e custos, Set./2017.

AI, C.; NORTON, E. C. Interaction terms in logit and probit models. **Economics Letters**, v. 80, p. 123-129, 2003.

ARVIS, J. F. *et al.* **Connecting to Compete 2012:** Trade Logistics in the Global Economy – the Logistics Performance Index and its Indicators. Washington: World Bank. 2012.

BRASIL. Instrução Normativa DAS/MAPA nº 39, de 27 de novembro de 2017**. Aprova o funcionamento do Sistema de Vigilância Agropecuária Internacional -Vigiagro**. Diário Oficial da República Federativa do Brasil, Brasília, Distrito Federal, 2017.

BRASIL. Instrução Normativa MAPA Nº. 36, de 10 de novembro de 2006. **Aprova o Manual de Procedimentos da Vigilância Agropecuária Internacional.** Diário Oficial da República Federativa do Brasil, Brasília, Distrito Federal, 2006.

BRASIL. Instrução Normativa SDA/MAPA nº 32, de 23 de setembro de 2015. **Estabelece os procedimentos de fiscalização e certificação fitossanitária de embalagens, suportes ou peças de madeira, em bruto, que serão utilizadas como material para confecção de embalagens e suportes, destinados ao acondicionamento de mercadorias importadas ou exportadas pelo Brasil.** Diário Oficial da República Federativa do Brasil, Brasília, Distrito Federal, 2015.

BRASIL. Instrução Normativa SDA/MAPA nº 4, de 6 de janeiro de 2004. **Aprova os procedimentos de inspeção e fiscalização de embalagens e suportes de madeira utilizados no transporte de mercadorias no comércio internacional.** Diário Oficial da República Federativa do Brasil, Brasília, Distrito Federal, 2004.

BRASIL. Portaria Interministerial MAA/MF n° 499, de 05 de novembro de 1999. **Dispõe sobre a permissão da entrada de madeira, no Brasil, em forma de lenha, após análise de risco de pragas e revoga a Portaria SDA n. 76, de 16 de junho de 1993.** Diário Oficial da República Federativa do Brasil, Brasília, Distrito Federal, 1999.

FAO. **Normas Internacionais para Medidas Fitossanitárias nº 07.Glossário de Termos Fitossanitários.** Roma, Itália. Food and Agriculture Organizationo of the United Nations, 2009.

FAO. **Normas Internacionais para Medidas Fitossanitárias nº 15. Regulamentação de Material de Embalagem de Madeira no Comércio Internacional.** Roma, Itália. Food and Agriculture Organizationo of the United Nations, 2009.

GREENE, W. H. **Econometric Analisys.** Pearson Education India. p. 1178. 2007.

HAACK, R. A.; PETRICE, T. R. Bark- and Wood-Borer Colonization of Logs and Lumber After Heat Treatment to ISPM 15 Specifications: The Role of Residuals Bark. **Journal of Economic Entomology**, v. 102, n. 3, p 1075-1084, 2009.

HORRACE, W. C.; OAXACA, R. L. **New Wine in Old Bottles: A Sequential Estimation Technique for the LPM.** Alemanha: Institute for the Study of Labor, 2003. 30 p. (Discussion Paper, n. 703).

IEDE, E. T. A CIPV e as normas internacionais de medidas fitossanitárias NIMF 15: a evolução dos tratamentos quarentenários de madeira. In: SEMINÁRIO INTERNACIONAL SOBRE PRAGAS QUARENTENÁRIAS FLORESTAIS, Colombo, 2012, **Anais...**

IMBENS, G. W.; WOOLDRIDGE, J. M. Recent developments in the econometrics of program evaluation. **Journal of Economic Literature** v. 47, n.1, p. 5-86, 2009.

KARACA-MANDIC, P.; NORTON, E. C.; DOWD, B. Interaction Terms in Nonlinear Models. **Health Services Research**, v. 47, n. 1, p. 255-274, 2012.

KASSAMBARA, A. **A pratical guide to principle componente methods in R**. STHDA, 2017.

LECHNER, M. The Estimation of Causal Effects by Difference-in-Difference Methods. **Foundations and Trends in Econometrics**, v.4, n.3, p. 165-224, 2010.

MONTEFERRANTE, E. C. *et al.* Interceptações de pragas em embalagens de madeira no Porto de Santos. **Revista de Política Agrícola**, v. 27, n. 1, p. 132-141, 2018.

ROCHA, B. G. D. **Identificação de perfis de risco associados a envios importados com probabilidade de abrigar pragas quarentenárias.** 2013. 73 f. Dissertação (Mestrado em Defesa Sanitária Vegetal) – Universidade Federal de Viçosa, Viçosa. 2013.

ROSA FILHO, O. A. O papel do Vigiagro na prevenção da introdução de pragas quarentenárias no Brasil. In: SEMINÁRIO INTERNACIONAL SOBRE PRAGAS QUARENTENÁRIAS FLORESTAIS, 1, 2012, Curitiba, **Anais...**

STRUTT. A *et al.* Evaluating the impacts of an international phytosanitary standard for wood packaging material: Global and United States trade implications. **Forest Policy and Economics**, v. 27, p. 54-64, 2012.

VARELLA, C. A. A. **Análise de Componentes Principais**. 2008.

WING, C.; SIMON, K.; BELLO-GOMEZ, R. A. Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. **Annual Review of Public Health**, v. 39, p. 453-469, 2018.

4. CONSIDERAÇÕES FINAIS

O presente trabalho deteve como objetivo construir modelos de previsão para MEMR com não conformidades importadas via Aeroporto Internacional de Viracopos e avaliar o impacto da introdução do sistema de checagem reduzida sobre a taxa de não conformidade das embalagens de madeira. Para tanto, o trabalho dividiu-se em duas frentes. A primeira voltada ao uso de métodos de predição comumente utilizadas na literatura em tarefas de classificação, sendo os métodos aplicados: Regressão Logistica, *Random Forest* e SVM. A segunda empregou modelos com estimadores de diferenças-em-diferenças, com o intuito de verificar se ser elegível ao sistema de fiscalização reduzida alterou o volume de MEMR importados com não conformidades e se a fiscalização por amostragem reduzida alterou a probabilidade de as empresas importadoras deterem taxas de não conformidade superioras a 3%.

Assim, foi comparado o desempenho seis modelos de previsão na identificação de PCG com MEMR não conformes, importados no período de outubro de 2018 e março de 2019, sendo esses: Regressão Logistica com correção de King e Zeng (2001), Regressão Logística estimada com amostra corrigida via *Oversample*, Regressão Logística estimada com amostra corrigida via SMOTE, *Random Forest* estimado com amostra corrigida via *Oversample*, *Random Forest* estimado com amostra corrigida via SMOTE, e SVM estimado via algoritmo ALGA-SVM. Dentre os modelos estimados, dois se destacaram quanto a acurácia e capacidade de prever corretamente a não conformidade, sendo esses: *Random Forest* estimado com amostra corrigida via *Oversample* e *Random Forest* estimado com amostra corrigida via SMOTE.

O *Random Forest* estimado com amostra corrigida via *Oversample* apresentou o melhor equilíbrio entre acurácia e erros de classificação das MEMR com não conformidades. Já o *Random Forest* estimado com amostra corrigida via SMOTE apresentou baixa acurácia, mas o menor erro de classificação das MEMR com não conformidades.

A simulação do uso de ambos os modelos *Random Forest*, no período de outubro de 2018 e março de 2019, acarretou a redução de inspeção de 77616 MEMR no modelo estimado com amostra corrigida via *Oversample*, 26298 MEMR a menos que o modelo *Random Forest* estimado com amostra corrigida via SMOTE. Por outro lado, o *Random Forest* estimado com amostra corrigida via SMOTE gerou 797 erros de classificação de MEMR com não conformidades, 624 a menos que o *Random Forest* estimado com amostra corrigida via *Oversample*.

Para o caso dos modelos de diferenças-em-diferenças, foram estimados modelos com a quantidade de MEMR importados por CNPJ e aderência a taxa de não conformidade superior a 3% por CNPJ como variáveis dependentes, para o período de 2016 a 2018.

Os resultados para a estimação dos modelos com a quantidade de MEMR importados como variável dependente apontaram que ser elegível ao sistema de fiscalização reduzida não alterou a quantidade de MEMR com não conformidades. Da mesma forma, os modelos com a taxa de não conformidade dos MEMR por CNPJ como variável dependente não demonstraram alterações na taxa de não conformidade para o grupo de empresas elegíveis ao sistema de fiscalização reduzida após a introdução do sistema. Logo, não há evidências que a introdução do sistema de amostragem alterou o comportamento das não conformidades das embalagens de madeira que acondicionam as importações no Aeroporto Internacional de Viracopos.

Logo, em termos gerais, a introdução do sistema de fiscalização reduzida não produziu redução da quantidade de MEMR com não conformidades. Entretanto, a inefetividade na alteração do volume de MEMR não conforme pode estar associado a fatores como o breve período de análise exposto neste trabalho. Apenas três anos foram analisados, em que em apenas dois foi aplicado o sistema de fiscalização reduzida. Assim, a substituição das importações com maior nível de não conformidade pode exigir mudanças em processos estruturais das empresas, de modo que as alterações no perfil de importação devem ocorrem a médio/longo prazo.

Por fim, cabe ressaltar a escolha dos preditores utilizados nos modelos estatísticos de previsão foi pautada na disponibilidade de acesso a tais informações. Todos os preditores estão disponíveis gratuitamente nos *websites* de órgãos internacionais, instituições governamentais e privadas. Logo, aplicações com melhor desempenho podem ser geradas a partir da inclusão de outros preditores, com menor grau de acessibilidade. Da mesma forma, técnicas que envolvam a combinação de resultados de modelos estatísticos podem gerar previsões mais acuradas, em comparação aos resultados obtidos pelos modelos empregados neste trabalho.

APÊNDICES

Tabela 1:Agregação dos setores por CNAE

Agregação	Divisão	Grupo
A - Agropecuária e Setor Extrativo	1 a 10	Todos
B - Industria Alimentícia, têxtil e diversos	10 a 15, 18 e 32	Todos
C - Máquinas, Equipamentos, Computadores e Moveis	26 a 29 e 31	Todos
D - Industria e comércio de Madeira	16, 17 e 31	Todos
E - Industria e comércio de metal	23 a 25	Todos
F - Transporte; Cons. e Água, Esgoto.	35 a 39 e 49 a 53	Todos
G - Químicos; Petróleo e Plástico, farmacêuticos	19, 20, 22	Todos
H – Automóveis e Equipamentos Automotores	29 e 30	Todos
I - Pesquisa e Serviços	55 a 73 e 77 a 100	Todos
J - Varejo e Atacado de Alimentos, Farmacêuticos, Livros e Perfumaria	46 e 47	Todos

Fonte: Elaboração própria. Os setores que não registraram PCG importados não foram inclusos na classificação