

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ADMINISTRAÇÃO E CIÊNCIAS CONTÁBEIS
MESTRADO PROFISSIONAL EM ADMINISTRAÇÃO PÚBLICA

**PERFIL DO SONEGADOR: UM OLHAR SOBRE CONTRIBUINTES INABILITADOS
DO CADASTRO DO ICMS.**

RICARDO DA SILVEIRA COELHO

JUIZ DE FORA
2018

RICARDO DA SILVEIRA COELHO

**PERFIL DO SONEGADOR: UM OLHAR SOBRE CONTRIBUINTES INABILITADOS
DO CADASTRO DO ICMS.**

Dissertação apresentada como requisito parcial para conclusão do Mestrado Profissional em Administração Pública, da Faculdade de Administração e Ciências Contábeis, Universidade Federal de Juiz de Fora, para obtenção do título de Mestre Administração Pública.

Orientadora:
Prof.^a Dr.^a Heloísa Pinna Bernardo

Coorientador:
Prof. Dr. Lupércio França Bessegato

JUIZ DE FORA
2018

Coelho, Ricardo da Silveira.

PERFIL DO SONEGADOR : UM OLHAR SOBRE
CONTRIBUINTES INABILITADOS DO CADASTRO DO ICMS /

Ricardo da Silveira Coelho. -- 2018.

171 p. : il.

Orientadora: Heloísa Pinna Bernardo

Coorientador: Lupércio França Bessegato

Dissertação (mestrado) - Universidade Federal de Juiz de Fora, Faculdade de
Administração e Ciências Contábeis.

Programa de Pós-Graduação em Administração Pública em Rede Nacional -
PROFIAP, 2018.

1. ICMS. 2. Sonegação. 3. Perfis de conformidade. 4. Regressão Logística. I.
Bernardo, Heloísa Pinna, orient. II. Bessegato, Lupércio França, coorient. III. Título.

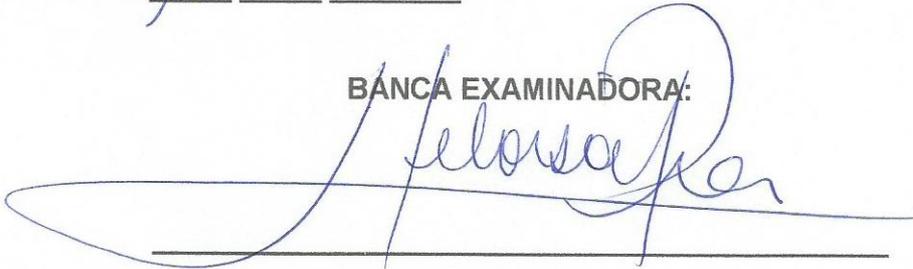
RICARDO DA SILVEIRA COELHO

**PERFIL DO SONEGADOR: UM OLHAR SOBRE CONTRIBUINTES
INABILITADOS DO CADASTRO DO ICMS**

Dissertação de Mestrado submetida à banca examinadora do Curso de Pós-Graduação em Mestrado Profissional em Administração Pública em Rede Nacional da Universidade Federal de Juiz de Fora, como parte dos requisitos necessários para obtenção do Grau de Mestre em Administração Pública.

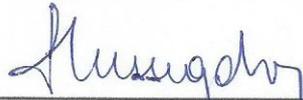
Aprovada em: 11/12/2018

BANCA EXAMINADORA:



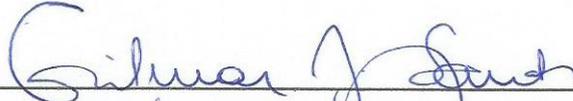
Prof.ª. Dr.ª. Heloísa Pinna Bernardo
Orientadora

Universidade Federal de Juiz de Fora

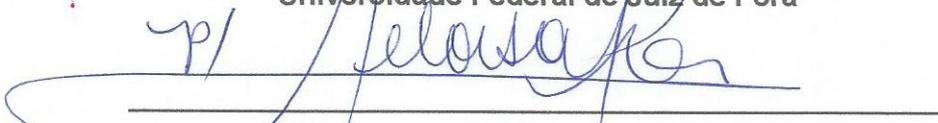


Prof. Dr. Lupércio França Bessegato
Coorientador

Universidade Federal de Juiz de Fora



Prof. Dr. Gilmar José dos Santos
Universidade Federal de Juiz de Fora

71 

Prof. Dr. Luciano Garcia Miguel
Secretaria da Fazenda do Estado de São Paulo

conforme ata

DEDICATÓRIA

À minha esposa Janice e à minha
filha Elisa, pelo amor, carinho e
apoio incondicionais.

AGRADECIMENTOS

À Secretaria da Fazenda do Estado de São Paulo, por tornar esse sonho possível através do Programa de Apoio a Pós-graduação, da Escola Fazendária do Estado de São Paulo – Fazesp.

Aos colegas da Fazesp, em especial à Regina Maria Pokorny, ao Tales Augusto Sad e à Marissol David, por todo profissionalismo e competência na condução do programa.

Ao Mário Takaoka e à Cláudia Formoso, por apoiarem e autorizarem minha inscrição no programa.

Ao Marcelo Fernandez, pelo apoio e incentivo na pesquisa.

Aos meus colegas de trabalho, em especial à Erika Yonehara e ao Fernando Oyama, por suprirem minhas ausências.

Ao Fábio Dickfeldt, ao Carlos Zimmermann e ao Marcos Formoso, pela ajuda com as ferramentas estatísticas e técnicas de Machine Learning.

Ao Alex Otsuki, ao Rodrigo Tomita e ao Marco Haniuda, pelo apoio no acesso aos dados.

Ao Wagner Carvalho, ao Paulo Pacello e à Daniela Nunes, pelo encorajamento e apoio na obtenção de informações.

Aos colegas da DEAT, DETEC e DA, que responderam meus pedidos de informação feitos através do Sistema Integrado de Informações ao Cidadão de SP.

Ao Tatsuo Sasaki, membro da banca, pelas dicas e trocas de ideias.

Ao Luciano Miguel, por aceitar prontamente o convite de integrar a banca e pelas dicas e ensinamentos pessoalmente ou através de sua tese de doutorado, que me serviu de referência.

À Prof.^a Dr.^a. Heloísa Bernardo, minha orientadora, pelo aprendizado e por me fazer crescer como pesquisador.

Ao Prof. Dr. Lupércio Bessegato, meu coorientador, pela clareza com que me ensinou a construir e validar o modelo estatístico.

Ao Prof. Dr. Marcus David, Magnífico Reitor da UFJF, membro da banca de qualificação, pelas dicas preciosas na qualificação.

Ao Prof. Dr. Gilmar José dos Santos, membro da banca, pelos ensinamentos.

Aos meus colegas de turma, pelas trocas de informações e debates enriquecedores.

Aos meus professores do curso, por todos os ensinamentos.

À Camila e à Laura, pelo apoio com os assuntos acadêmicos.

Aos meus pais, irmãos e sobrinhos, por acreditarem e incentivarem meus sonhos.

À minha esposa e à minha filha, meus amores, por suportarem as horas ausentes e ainda assim apoiarem esse projeto.

RESUMO

A sonegação de tributos é um componente de injustiça fiscal na medida em que faz o fardo tributário recair apenas sobre alguns indivíduos, enquanto que outros ganham vantagem competitiva e prejudicam a concorrência. O combate à sonegação, com consequente aumento da conformidade fiscal e diminuição da Brecha Tributária, é objetivo precípua de todas as Administrações Tributárias e passa pela identificação do comportamento e segmentação dos contribuintes de acordo com sua postura de conformidade. Estudos modernos demonstram que a decisão do indivíduo pela sonegação não está sujeita apenas a cálculos econômicos e racionais, pois sofre influência também de fatores Psicológicos, Sociológicos e de Ambiente, e estes devem ser levados em conta na segmentação dos contribuintes nos perfis adequados. A presente pesquisa se propôs a incluir tais fatores na criação de um modelo estatístico capaz de segmentar contribuintes de uma população pouco estudada, qual seja: a de contribuintes do ICMS que sofreram apuração de inatividade. Para tanto, foi feito um estudo empírico que buscou características presentes na amostra que representam os fatores de influência no comportamento tributário e criou-se um modelo estatístico de Regressão Logística, baseado nessas características, capaz de segmentar os contribuintes de acordo com dois perfis: contribuintes sem indícios de irregularidades tributárias além da simples inatividade, e contribuintes com indícios de práticas irregulares, a exemplo da sonegação fiscal. Os resultados da pesquisa demonstraram que as características encontradas foram capazes de segmentar os contribuintes com grau de acerto superior a 84%, e evidências estatísticas demonstraram que o modelo está bem ajustado para ser aplicado e obter resultados parecidos em toda a população do estudo. Tal segmentação pode ser utilizada para que as Autoridades Tributárias tratem os contribuintes da forma apropriada a cada perfil, estimulando a conformidade fiscal e combatendo a sonegação.

Palavras-Chave: ICMS; Sonegação; Perfis de conformidade; Regressão Logística

ABSTRACT

Tax Evasion makes the tax system unfair because it turns the tax burden heavier for some individuals, whereas others get economic advantage over market competitors. Increasing tax compliance by fighting evasion reduces the tax gap, which is the main goal of tax authorities. To accomplish that, one must classify taxpayers according to their tax behaviour. Modern studies demonstrate that tax compliance decisions are not solely based on maximizing economic outcome, but rather influenced by Economic, Psychological, Sociological and Demographic factors. This paper presents a study in which all these factors have been taken in to account to build a mathematical model that is able to classify individuals from a population for which there has been few if any studies conducted so far: taxpayers of ICMS (Brazilian consumption tax) that have presented signs of business inactivity. For such, a sample of this population has been selected and used to identify which variables could best be used in a Logistic Regression Model to categorize individuals in one of two groups: taxpayers with signs of pure business inactivity and taxpayers with signs of evasion previous to business inactivity. The results show that the selected variables could be successfully used to categorize the individuals from the sample with an Accuracy higher than 84%. Moreover, Goodness-of-fit tests show that the model fits the data very well and could be used in the whole population of the study. This would allow Tax Authorities to respond to taxpayers according to their tax behaviour and that would be a possible mechanism to increase tax compliance and reduce tax evasion.

Keywords: ICMS; Tax Evasion; Tax compliance behaviour; Logistic Regression.

LISTA DE ABREVIATURAS E SIGLAS

AUC	Under the ROC Curve
CAF	Coordenadoria da Administração Financeira
CAT	Coordenadoria da Administração Tributária da Secretaria da Fazenda do Estado de São Paulo
CCE	Coordenadoria de Compras Eletrônicas
CGP	Coordenadoria de Gestão de Pessoas
CNAE	Código de Atividade Econômica Principal
CSTC	Coordenadoria de Serviços e Tecnologias Compartilhados
DRT	Delegacia Regional Tributária de jurisdição
HMRC	Her Majesty's Revenue and Customs
ICMS	Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação
IPI	Imposto sobre Produtos Industrializados
ISSQN	Imposto sobre Serviços de Qualquer Natureza
IVA	Imposto sobre Valor Agregado
IVV	Imposto sobre Vendas a Varejo
MEI	Microempreendedor Individual
GOF	Goodness-of-Fit
SN	Simples Nacional
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
ROC	Receiver Operating Characteristic
RPA	Regime Periódico de Apuração
RST	Retail Sales Tax
Sefaz/SP	Secretaria da Fazenda do Estado de São Paulo
VAT	Value Added Tax
VIF	Variance Inflation Factors

LISTA DE FIGURAS

Figura 1: O escopo da pesquisa.	21
Figura 2: Efeitos dos procedimentos de apuração de inatividade.	26
Figura 3: As posturas dos contribuintes.....	30
Figura 4: As posturas dos contribuintes e sua relação com a não conformidade por erro.....	33
Figura 5: Tempo necessário em horas para cumprimento de obrigações tributárias.	34
Figura 6: Tempo de cumprimento em horas na América do Sul.	35
Figura 7: Os componentes da brecha tributária segundo a definição ampla.	36
Figura 8: A brecha tributária no Reino Unido (em bilhões de libras) como percentual da arrecadação potencial.	41
Figura 9: Brecha do IVA na União Europeia em 2014 e 2013.	42
Figura 10: O modelo conceitual da pirâmide de conformidade.	50
Figura 11: Operacionalização da não-cumulatividade pelo método invoice credit.....	55
Figura 12: Causas da brecha administrativa.....	58
Figura 13: Elementos críticos e aporte teórico.	59
Figura 14: População de estudo da pesquisa.	67
Figura 15: Amostra de pesquisa.	68
Figura 16: Composição das variáveis de índices que representam fatores econômico-financeiros.	77
Figura 17: Gráficos de dispersão dos valores das variáveis de índices que representam fatores econômico-financeiros nos grupos de referência.	79
Figura 18: Variáveis que representam fatores psicológicos.....	85
Figura 19: Gráficos de dispersão dos valores das variáveis que representam fatores psicológicos nos grupos de referência.	87
Figura 20: Variáveis que representam fatores de ambiente.	89
Figura 21: Composição das variáveis de índices que representam fatores sociológicos.....	96
Figura 22: Gráficos de dispersão dos valores das variáveis que representam fatores sociológicos nos grupos de referência.	97
Figura 23: Leverage, Standardised Person Residuals e Cook's Distance do modelo.	120
Figura 24: Leverage, Standardised Person Residuals e Cook's Distance sem Influential Values	121

Figura 25: Resultado da verificação de linearidade das variáveis contínuas	123
Figura 26: Resultado da verificação de linearidade das variáveis contínuas com decl_ind2 transformada	124
Figura 27: Resultado da verificação das medidas de ajustes individuais.....	126
Figura 28: Visualização gráfica da Classification Table.....	136
Figura 29: Variáveis e interações do modelo.....	143
Figura 30: Fluxo da solução proposta.....	150

LISTA DE GRÁFICOS

Gráfico 1: Média anual de contribuintes do ICMS no Estado de São Paulo – 2017.....	64
Gráfico 2: Participação na arrecadação por grupo de contribuinte em 2017.....	66
Gráfico 3: Quantitativos da amostra e da população.....	68
Gráfico 4: Distribuição da amostra de acordo com os grupos de referência.....	76
Gráfico 5: Distribuição de valores de cadt_natj.....	90
Gráfico 6: Distribuição de valores de cadt_port.....	91
Gráfico 7: Distribuição de valores de cadt_cnae.....	92
Gráfico 8: Distribuição de valores de cadt_subt.....	93
Gráfico 9: Distribuição de valores de cadt_delg.....	95
Gráfico 10: Influência das variáveis na predição.....	113
Gráfico 11: Direção da influência das variáveis na predição.....	115
Gráfico 12: Valores de Sensitividade e Especificidade por cut-off.....	134
Gráfico 13: ROC Curve do modelo.....	139

LISTA DE QUADROS

Quadro 1: Estrutura parcial da Secretaria da Fazenda do Estado de São Paulo.....	22
Quadro 2: Síntese do aporte teórico pesquisado.....	61
Quadro 3: Fórmula de criação do modelo	111
Quadro 4: O modelo criado	112
Quadro 5: Resultado do outlierTest.....	122
Quadro 6: Resultado do Repeated K-fold Cross Validation.....	137
Quadro 7: Passos para implantação do plano de ação	153

LISTA DE TABELAS

Tabela 1: Contribuintes do RPA inabilitados por inatividade presumida no Estado de São Paulo.	24
Tabela 2: Características dos procedimentos de apuração de inatividade.	24
Tabela 3: Tamanho (percentual PIB) da shadow economy em alguns países – 2000.	40
Tabela 4: Dados coletados e fatores de referência.	70
Tabela 5: Perfil de cumprimento tributário.	71
Tabela 6: Dados de declaração de imposto e de documentos fiscais – Fatores econômico-financeiros	71
Tabela 7: Experiência anterior com o Fisco (percepção de risco) - Fatores psicológicos.	72
Tabela 8: Características demográficas - Fatores de ambiente.	72
Tabela 9: Relação com outros indivíduos - Fatores sociológicos.	73
Tabela 10: Welch Two Sample t-test das variáveis de índices que representam fatores econômico-financeiros contra a variável dilg_irre.	78
Tabela 11: Welch Two Sample t-test das variáveis de índices que representam fatores psicológicos contra a variável dilg_irre.	86
Tabela 12: Lista de naturezas jurídicas.	89
Tabela 13: Lista de portes empresariais.	91
Tabela 14: Lista de códigos de atividade econômica principal.	92
Tabela 15: Lista de delegacias regionais tributárias.	94
Tabela 16: Welch Two Sample t-test das variáveis de índices que representam fatores sociológicos contra a variável dilg_irre.	97
Tabela 17: Síntese das hipóteses de influência no resultado da diligência.	100
Tabela 18: Resultado do Passo 1	105
Tabela 19: Resultado do Passo 2	106
Tabela 20: Resultado do Passo 3	107
Tabela 21: Resultado do Passo 4	107
Tabela 22: Resultado do Passo 5	108
Tabela 23: Resultado do Passo 6	109
Tabela 24: Variáveis e interações selecionadas	110
Tabela 25: Relação entre tolerância e VIF.	117

Tabela 26: Resultado da verificação de Multicolinearidade.....	118
Tabela 27: Medidas após remoção das observações.....	121
Tabela 28: Pearson Chi-square Statistic e Deviance tests	129
Tabela 29: Resultado do Hosmer-Lemeshow Test	130
Tabela 30: Resultado do Stukel's Test	131
Tabela 31: Confusion Matrix com cut-off igual a 0.5.	135
Tabela 32: Pseudo R ² do modelo.....	140

SUMÁRIO

INTRODUÇÃO	19
1 DESCRIÇÃO DO CASO DE ESTUDO.....	21
2 DISCUSSÃO TEÓRICA E PESQUISA EMPÍRICA	28
2.1 Fundamentação teórica	28
2.1.1 <i>Conformidade Tributária</i>	28
2.1.2 <i>Evasão, elisão e elusão fiscal</i>	30
2.1.3 <i>A não conformidade não intencional</i>	32
2.1.4 <i>O fardo de cumprimento tributário</i>	34
2.1.5 <i>A brecha tributária</i>	35
2.1.5.1 <i>Os efeitos da brecha tributária</i>	37
2.1.5.2 <i>O tamanho da brecha tributária</i>	38
2.1.6 <i>O combate às práticas que reduzem a arrecadação</i>	42
2.1.6.1 <i>O combate à elisão e à elusão</i>	43
2.1.6.2 <i>O combate à evasão e o estímulo à diminuição dos erros</i>	44
2.1.7 <i>O comportamento do contribuinte</i>	45
2.1.7.1 <i>A decisão de sonegar: modelo clássico</i>	46
2.1.7.2 <i>A decisão de sonegar: modelos recentes</i>	48
2.1.8 <i>A conformidade nos impostos sobre consumo</i>	52
2.1.9 <i>O aumento da conformidade no âmbito do ICMS</i>	56
2.1.10 <i>Síntese dos elementos críticos abordados</i>	59
2.2 Pesquisa empírica	63
2.2.1 <i>Metodologia</i>	63
2.2.1.1 <i>Classificação da pesquisa</i>	63
2.2.1.2 <i>População</i>	64
2.2.1.3 <i>Amostra</i>	67
2.2.1.4 <i>Coleta dos dados</i>	69
2.2.1.5 <i>Variáveis do estudo</i>	70
2.2.1.6 <i>Limitações da pesquisa</i>	73
2.2.2 <i>Análise dos dados</i>	74
2.2.2.1 <i>Variável resultado da diligência</i>	76
2.2.2.2 <i>Variáveis que representam fatores econômico-financeiros</i>	77
2.2.2.2.1 <i>Variável índice de declarações</i>	80
2.2.2.2.2 <i>Variável índice de crédito e débito das declarações</i>	81
2.2.2.2.3 <i>Variável índice de ausência de lastro de créditos</i>	81

2.2.2.2.4 Variável índice de ausência de lançamento de débitos	82
2.2.2.2.5 Variável índice de lançamento de crédito.....	83
2.2.2.2.6 Variável índice de lançamento de débito.....	84
2.2.2.2.7 Variável índice de ausência de lastro de saídas em relação a entradas	84
2.2.2.3 <i>Variáveis que representam fatores psicológicos</i>	85
2.2.2.3.1 Variável tempo entre inscrição e diligência.....	87
2.2.2.3.2 Variável quantidade de fiscalizações.....	88
2.2.2.3.3 Variável quantidade de autos de infração	88
2.2.2.4 <i>Variáveis que representam fatores de ambiente</i>	88
2.2.2.4.1 Variável natureza jurídica.....	89
2.2.2.4.2 Variável porte empresarial.....	90
2.2.2.4.3 Variável código de atividade econômica principal.....	91
2.2.2.4.4 Variável substituto tributário	93
2.2.2.4.5 Variável delegacia regional tributária.....	94
2.2.2.5 <i>Variáveis que representam fatores sociológicos</i>	96
2.2.2.5.1 Variável índice contador	98
2.2.2.5.2 Variável índice sócio 1	98
2.2.2.5.3 Variável índice sócio 2	99
2.2.2.6 <i>Síntese das hipóteses de influência no evento de interesse</i>	100
2.2.3 <i>Modelo empírico</i>	101
2.2.3.1 <i>Procedimento para seleção de variáveis</i>	103
2.2.3.1.1 Passos	105
2.2.3.2 <i>O modelo estatístico</i>	110
2.2.3.3 <i>Resultados da seleção de variáveis</i>	113
2.2.4 <i>Validação do modelo</i>	116
2.2.4.1 <i>Os pressupostos do modelo Logit</i>	116
2.2.4.1.1 Ausência de Multicolinearidade	117
2.2.4.1.2 Ausência de Influential Values	118
2.2.4.1.3 Relação linear entre as variáveis explicativas contínuas e a variável dependente	122
2.2.4.2 <i>Qualidade do ajuste do modelo</i>	125
2.2.4.2.1 Medidas de ajustes individuais.....	125
2.2.4.2.2 Testes de ajuste geral do modelo	127
2.2.4.2.3 <i>Qualidade do ajuste do modelo: síntese</i>	132
2.2.5 <i>Análise dos resultados do modelo</i>	132
2.2.5.1 <i>Accuracy</i>	133

2.2.5.2 Repeated K-fold Cross Validation	136
2.2.5.3 Area under the ROC Curve	138
2.2.5.4 Pseudo R^2	140
2.2.5.5 Síntese das medidas do modelo	141
2.2.6 Resultados e discussão	141
3 PLANO DE AÇÃO	149
3.1 Solução proposta	150
3.2 Operacionalização da solução	152
CONSIDERAÇÕES FINAIS	154
REFERÊNCIAS	158
APÊNDICES	166

INTRODUÇÃO

Esta dissertação trata da evasão fiscal¹, mais especificamente das características dos sonegadores do Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação – ICMS no Estado de São Paulo que foram submetidos a procedimento de apuração de inatividade. O ICMS é um imposto sobre consumo de competência dos estados brasileiros e do Distrito Federal.

A inscrição no Cadastro é obrigatória para o exercício de atividades sujeitas ao ICMS por contribuintes² deste imposto no Estado de São Paulo e o cumprimento de certos requisitos, também chamados de obrigações tributárias, é mandatório para que a referida inscrição se mantenha eficaz e o contribuinte habilitado a operar (SÃO PAULO, 1993).

Um dos requisitos é a demonstração de atividade regular pelo contribuinte, ou seja, o exercício de forma contínua das atividades para as quais solicitou sua inscrição. Dentre os meios que demonstram o exercício estão a entrega de declarações de apuração do imposto, a emissão de documentos fiscais e o pagamento do imposto apurado (SÃO PAULO, 1995).

O descumprimento dos requisitos legitima a Coordenadoria da Administração Tributária – CAT, unidade ligada à Secretaria da Fazenda do Estado de São Paulo, a apurar a inatividade e inabilitar³ a inscrição do contribuinte (SÃO PAULO, 2006).

A pesquisa em questão foi executada no âmbito dos procedimentos de apuração da inatividade. Conforme será detalhado adiante, um dos procedimentos de apuração de inatividade não é capaz de detectar se houve indícios de práticas sonegatórias conduzidas pelos contribuintes antes da interrupção das operações do estabelecimento empresarial e, portanto, não é possível segmentar tais contribuintes entre aqueles que simplesmente

¹ Evasão será tratada como sinônimo de sonegação neste texto.

² De acordo com Lei paulista nr. 6.374, de 01 de março de 1993, “Contribuinte do imposto é qualquer pessoa, natural ou jurídica, que de modo habitual ou em volume que caracterize intuito comercial, realize operações relativas à circulação de mercadorias ou preste serviços de transporte interestadual ou intermunicipal ou de comunicação” (SÃO PAULO, 1993).

³ O artigo 1º da Portaria CAT 95 de 25 de novembro de 2006, que dispõe sobre a suspensão, cassação e nulidade da eficácia da inscrição no Cadastro de Contribuintes do ICMS, detalha que a inscrição no cadastro pode estar enquadrada como Ativa; Suspensa; Inapta; Baixada; e Nula. Contribuintes em situação Ativa ou Baixada se encontram em situação regular – embora apenas o primeiro esteja apto a operar –, ao passo que os demais se encontram em situação irregular. Para efeitos desse texto, serão considerados contribuintes inabilitados somente aqueles não aptos a operar em função de situação irregular, ou seja, contribuintes com situação Suspensa, Inapta ou Nula.

deixaram de operar e aqueles que previamente ao fim das operações, praticaram ilícitos tributários.

Diante disso e com o intuito de encontrar meios de aprimorar o respectivo procedimento de apuração de inatividade, esta dissertação buscará respostas para o seguinte problema de pesquisa: **quais características permitem segmentar em perfis de conformidade os contribuintes do ICMS do estado de São Paulo que foram submetidos a procedimento de apuração de inatividade?**

Nesse diapasão, o objetivo geral deste estudo é, à luz de teorias que elencam os fatores que determinam o comportamento tributário, identificar características que influenciem ou sejam influenciadas por eles e utilizá-las para segmentar os indivíduos da população de pesquisa entre aqueles que apenas apresentaram indícios de inatividade, e aqueles que, além destes, apresentaram também indícios de práticas sonegatórias.

Serão também buscados os seguintes objetivos específicos:

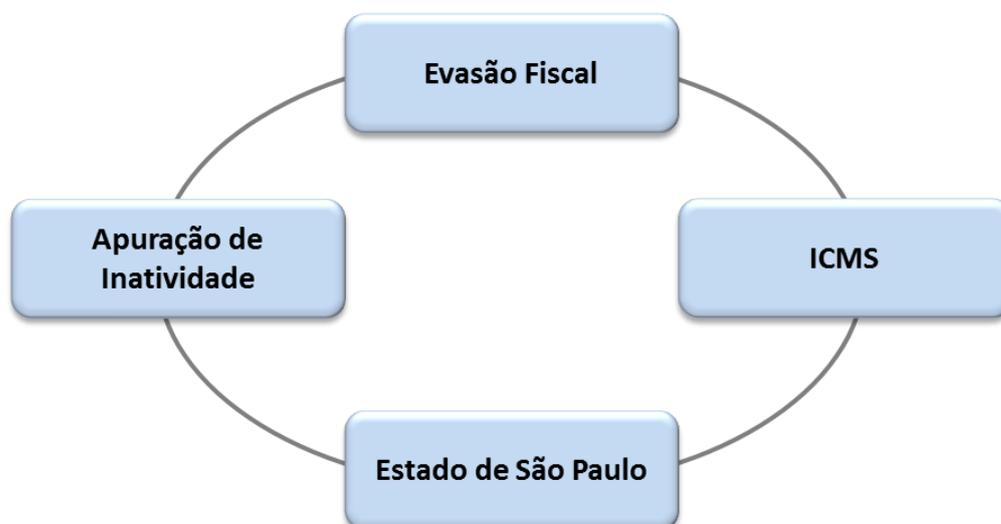
- Determinar que modelo de conformidade proposto na teoria seria apropriado para segmentar corretamente a população estudada.
- Identificar quais fatores são significativos para determinação do perfil do contribuinte de acordo com o modelo escolhido.
- Encontrar, a partir de análise estatística, quais características identificadas na população de estudo poderiam melhor refletir tais fatores.
- Construir um modelo estatístico que contribua para a definição do perfil de conformidade de contribuintes que foram inabilitados do Cadastro do ICMS.

Para tanto, esta dissertação está organizada da seguinte forma: além deste capítulo introdutório em que se apresentou a questão de pesquisa e os objetivos a serem alcançados, o capítulo um detalhará o caso em estudo. O capítulo dois apresentará o Referencial Teórico e a Pesquisa Empírica, enquanto que no capítulo três será proposto um Plano de Ação para que o modelo criado na pesquisa seja implementado na Secretaria da Fazenda do Estado de São Paulo. O último capítulo trará as considerações finais e sugestões para estudos futuros.

1 DESCRIÇÃO DO CASO DE ESTUDO

Esta dissertação trata da **evasão fiscal**, mais especificamente das **características dos sonegadores de ICMS no Estado de São Paulo** que foram submetidos a procedimento de **apuração de inatividade**. A *Figura 1* demonstra de forma esquemática o escopo em que a pesquisa se insere.

Figura 1: O escopo da pesquisa.



Fonte: Elaborado pelo autor da pesquisa.

As normas tributárias em geral, como o Código Tributário Nacional, elencam práticas diversas que são consideradas descumprimento tributário (BRASIL, 1966). Algumas dessas práticas, dentre as quais as atividades fraudulentas, resultam em sonegação de tributos e são, por isso, consideradas irregulares. Há outras, no entanto, que embora não resultem na redução do montante devido de impostos e não sejam necessariamente revestidas de dolo, também são consideradas irregulares.

É o caso da inatividade, entendida como o encerramento das operações da empresa⁴ sem a prévia comunicação ao Fisco⁵. O Superior Tribunal de Justiça – STJ considerou a inatividade passível de punição na Súmula 435, publicada com o seguinte texto:

⁴ O termo tecnicamente mais preciso nesse caso é **estabelecimento empresarial**. De acordo com o artigo 1.142 do Código Civil, “Considera-se estabelecimento todo complexo de bens organizado, para exercício da empresa, por empresário, ou por sociedade empresária”. (BRASIL, 2002). Para fins de facilitação da compreensão, no entanto, será utilizado neste texto o termo empresa.

⁵ Para fins deste texto, o termo Fisco será utilizado como sinônimo de Administração Tributária.

Presume-se dissolvida irregularmente a empresa que deixar de funcionar no seu domicílio fiscal, sem comunicação aos órgãos competentes, legitimando o redirecionamento da execução fiscal para o sócio-gerente (BRASIL, 2010).

No Estado de São Paulo, a apuração da inatividade é responsabilidade da Coordenadoria da Administração Tributária – CAT, que se insere na Secretaria da Fazenda do Estado de São Paulo na forma do organograma apresentado no *Quadro 1*.

Quadro 1: Estrutura parcial da Secretaria da Fazenda do Estado de São Paulo.



Fonte: Adaptado Sefaz/SP (2018).

A CAT foi criada pelo artigo 6º do Decreto 49.899, de 2 de julho de 1968, e reorganizada pelo Decreto 60.812, de 30 de setembro de 2014. Essa coordenação tem como responsabilidade (SEFAZ/SP, 2018):

- Arrecadação dos tributos estaduais;
- Fiscalização do cumprimento das obrigações tributárias.

A apuração de inatividade pode se dar por meio de dois procedimentos, elencados na Portaria CAT 95/2006 (SÃO PAULO, 2006):

- Por **Diligência Fiscal**.
- Pelo Processo de Suspensão e Cassação da Inscrição no Cadastro do ICMS por Inatividade Presumida, doravante denominado neste texto **Cassação por Inatividade Presumida**.

A diligência fiscal é executada por Autoridade Fiscal e pode englobar inspeção *in loco*. Ela é capaz de detectar diversas irregularidades além da inatividade como: simulação (empresas de fachada), ocorrência de práticas sonegatórias, apresentação de documentação falsa, dentre outras.

Para efeitos desta dissertação, o efeito dos resultados das diligências fiscais na situação cadastral das empresas foi agrupado nos seguintes:

- Empresa em situação regular;
- Empresa em situação irregular.

Medeiros (1993) e Kirchler (2007) apontam que, apesar do elevado grau de precisão das diligências fiscais – que contam com sofisticados processos de fiscalização e auditoria –, a natureza detalhada dos procedimentos as torna mais custosas para a Administração Pública, dificultando sua execução em grande volume.

Nesse sentido, a Cassação por Inatividade Presumida preenche em parte a lacuna, pois é capaz de processar dados em massa. Esse procedimento é, por isso, importante instrumento de saneamento e manutenção da qualidade do cadastro, pois procura garantir que somente estabelecimentos em atividade estejam habilitados a operar.

Além do saneamento, um possível efeito ainda mais importante é o combate à evasão fiscal. Esse efeito pode ocorrer porque apenas estabelecimentos com inscrição regular são capazes de emitir documentos fiscais que permitem a geração de créditos de ICMS, conforme será detalhado adiante nesta dissertação. Assim, a suspensão e eventual cassação da inscrição de contribuintes, que porventura estejam envolvidos em práticas sonegatórias, tem a capacidade de impedir diversas dessas práticas.

Ocorre, entretanto, que o combate à sonegação no âmbito da Cassação por Inatividade Presumida é limitado, pois o processo é capaz de detectar apenas a inatividade e não outras irregularidades como o fazem as diligências fiscais. Embora a inabilitação, decorrente do processo, já seja suficiente para impedir certas atividades sonegatórias, ela não é capaz de impedir que novos estabelecimentos dos mesmos titulares sejam inscritos para práticas similares, pois não gera a informação de que além da inatividade da empresa, esta foi utilizada também para práticas de evasão fiscal.

Para demonstrar o poder de processamento em grandes volumes, a **Tabela 1** demonstra o total de contribuintes do chamado Regime Periódico de Apuração – RPA⁶ cassados por inatividade presumida no Estado de São Paulo nos anos de 2015 a 2017 e a relação percentual aproximada frente ao total de contribuintes do respectivo regime.

Tabela 1: Contribuintes do RPA inabilitados por inatividade presumida no Estado de São Paulo.

Ano	Contribuintes RPA inabilitados	% aproximado do total RPA
2015	54.337	15%
2016	64.320	21%
2017	48.900	19%

Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.

Na **Tabela 1** verifica-se o poder de processamento em grandes volumes do processo, pois em todos os anos foi detectado que 15% ou mais do total de estabelecimentos RPA inscritos não estavam operando e estes tiveram sua inscrição inabilitada.

Os números não são capazes, todavia, de apontar em quais casos a inabilitação da inscrição contribuiu apenas para o saneamento do cadastro e em quais casos foi instrumento indireto de combate à evasão fiscal.

A **Tabela 2** resume as características dos dois procedimentos de apuração de inatividade elencados na Portaria CAT 95/2006.

Tabela 2: Características dos procedimentos de apuração de inatividade.

Diligência Fiscal	Cassação por Inatividade Presumida
Executada por Autoridade Fiscal	Executada através de processo computacional
Possui elevado grau de precisão na detecção de diversas irregularidades	Possui precisão apenas na detecção de inatividade
Custosa e de difícil execução em massa	Baixo custo e eficiência na execução em massa

Fonte: Elaborado pelo autor da pesquisa.

Pela **Tabela 2** é possível observar que a Cassação por Inatividade Presumida preenche a lacuna de processamento em massa da Diligência Fiscal, mas é deficiente na detecção de irregularidades diversas da inatividade presumida.

Para ilustrar a questão, considere-se dois contribuintes que possuem inscrição regular no Cadastro do ICMS. Um deles exerce suas atividades empresariais de forma lícita. O outro, no entanto, abriu uma empresa apenas de fachada para gerar créditos de ICMS e reduzir o

⁶ O Regime Periódico de Apuração consiste na sistemática de apropriação de créditos relativos às entradas para compensar com débitos relativos às saídas, considerando sempre um período pré-fixado na legislação (SÃO PAULO, 2000).

montante de impostos devidos por ele e possivelmente por outros associados na prática fraudulenta.

Considere-se, ainda, que antes da detecção da fraude pelo Fisco, o contribuinte envolvido nas práticas sonegatórias as tenha interrompido no âmbito daquela empresa para não gerar suspeitas. Ele poderia ter a intenção de retomá-las posteriormente com a mesma inscrição ou solicitar a inscrição de outro estabelecimento no Cadastro do ICMS.

Voltando ao primeiro contribuinte, considere-se que devido a fatores diversos, como crises econômicas, este tenha encerrado suas atividades, mas não tenha solicitado a baixa cadastral perante o Fisco, incorrendo em prática irregular, porém sem redução de impostos, ou seja, sem sonegação.

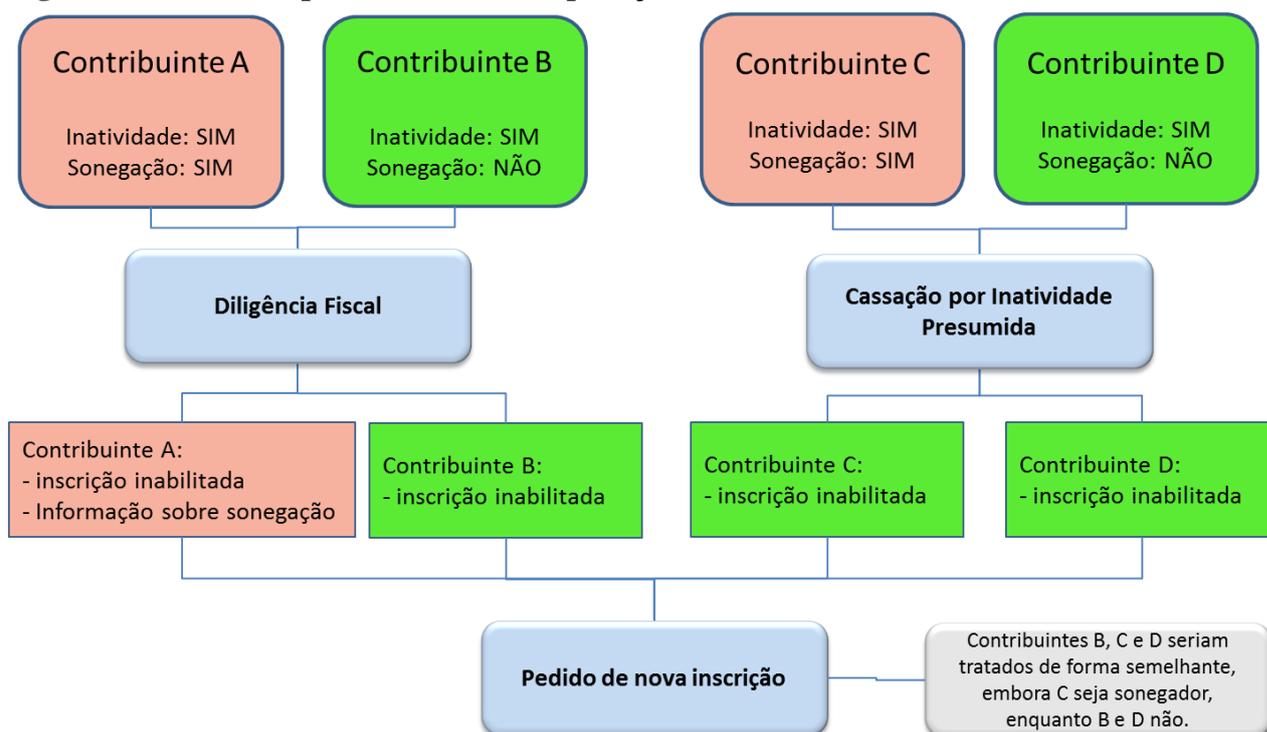
A inatividade de ambos seria detectada na Cassação por Inatividade Presumida e estes teriam suas inscrições inabilitadas. Neste caso, houve tanto o saneamento do cadastro – já que a realidade (inatividade dos contribuintes) foi propriamente identificada – como também o impedimento das práticas sonegatórias no caso do contribuinte mal-intencionado.

Ocorre, entretanto, que a Cassação por Inatividade Presumida gera como resultado apenas a ocorrência ou não da inatividade, e por isso não seria possível diferenciar o comportamento dos contribuintes. Ambos seriam tratados como inabilitados por inatividade, embora um deles, além da inatividade, também tenha se envolvido em práticas sonegatórias. Estas poderiam ter sido detectadas por uma diligência fiscal, mas como argumentado, as diligências são normalmente custosas e de difícil execução em grandes volumes.

Suponha-se, por fim, que após certo tempo ambos tenham solicitado a inscrição de uma nova empresa no cadastro. A informação de que um deles estava previamente envolvido em práticas sonegatórias e o outro não seria de grande valia para as autoridades tributárias analisarem tais pedidos de inscrição.

A **Figura 2** demonstra de forma gráfica os efeitos dos dois procedimentos na detecção de irregularidades além da inatividade. Nela fica claro que a Cassação por Inatividade Presumida se tornaria mais efetiva se fosse capaz de detectar também indícios de evasão fiscal.

Figura 2: Efeitos dos procedimentos de apuração de inatividade.



Fonte: Elaborado pelo autor da pesquisa.

Braithwaite (2003) e Kirchler (2007) reforçam a necessidade de segmentação dos contribuintes de acordo com o perfil de cumprimento tributário. Os autores reconhecem que tal segmentação nem sempre é tarefa fácil, mas argumentam que seus benefícios na diminuição da evasão fiscal e aumento da conformidade tributária são compensadores.

Nessa esteira, diversas iniciativas como as dos Fiscos Australiano, Norte-Americano e Brasileiro foram implementadas para identificar o perfil tributário dos contribuintes e aplicar o devido tratamento a eles (KIRCHLER, 2007; VIOL, 2015).

A **inovação** do presente **estudo** reside na proposta de **aplicação de mecanismos de segmentação em uma população pouco estudada, qual seja: a de contribuintes do ICMS que sofreram apuração de inatividade.**

Não foram encontrados na literatura estudos com propostas de modelos de segmentação de contribuintes que sofreram apuração por inatividade. A segmentação no âmbito dessa população permitirá ampliar o potencial de combate à evasão fiscal da Cassação por Inatividade Presumida, pois fornecerá informações adicionais para análise de pedidos de novas inscrições.

Kirchler (2007) aponta que a maneira como o Fisco percebe os contribuintes determinará sua postura quanto a estes e, por consequência, a postura destes em relação

àquele. Como consequência, a segmentação de contribuintes em perfis é importante para que o tratamento adequado seja dispensado a cada tipo de perfil e o cumprimento tributário seja ampliado. Esse aspecto justifica e reforça a importância do presente estudo, que foi realizado no âmbito do ICMS do estado de São Paulo. Ademais, espera-se que os achados também possam ser aplicados a outros estados e, eventualmente, a outros tributos.

2 DISCUSSÃO TEÓRICA E PESQUISA EMPÍRICA

Na introdução e na descrição do caso de estudo, apresentados anteriormente, foi argumentado que a segmentação dos contribuintes inabilitados do Cadastro do ICMS do Estado de São Paulo seria útil no combate à sonegação.

Para que essa segmentação se torne possível, é necessário identificar quais características permitiriam a segregação dos contribuintes nos perfis de conformidade: *Empresa em situação regular* e *Empresa em situação irregular*.

A busca dessas características será iniciada através da análise das principais teorias concernentes ao tema e será complementada por uma pesquisa empírica que torne possível aplicar os achados na teoria ao caso concreto. É o que se discute nesse capítulo.

2.1 Fundamentação teórica

No presente capítulo, serão estudados os principais conceitos e correntes teóricas que nortearão a pesquisa.

2.1.1 Conformidade Tributária

A *tax compliance*, mais comumente traduzida na literatura especializada como conformidade tributária (ou fiscal), ou ainda, obediência tributária, pode ser definida, de forma simples, como a ausência de sonegação de tributos (WEBER; FOOKEN; HERRMANN, 2014). Em outras palavras, seria a disposição do contribuinte (*taxpayer*) em pagar seus impostos corretamente. A *não conformidade (non-compliance)*, por sua vez, se reflete nas falhas de cumprimento das obrigações tributárias, sejam tais falhas intencionais ou não.

Dos estudos de McBarnet (2001) e Braithwaite (2003b) é possível elencar os seguintes tipos de posturas dos contribuintes em relação ao pagamento de tributos:

- **Conformidade Voluntária (*commitment*):** ocorre quando o contribuinte cumpre suas obrigações corretamente, de forma voluntária e com boa vontade.
- **Conformidade por Rendição (*capitulation*):** ocorre quando o contribuinte se rende às autoridades por reconhecer sua legitimidade e adota uma postura de cooperação e desejo de cumprir suas obrigações, mesmo que nem sempre consiga.
- **Conformidade Criativa (*creative compliance ou game playing*):** aparece quando a conformidade ocorre apenas legalmente, pois do ponto de vista econômico os tributos são elididos⁷ através de brechas na lei tributária.
- **Resistência (*resistance*):** acontece quando o contribuinte adota uma postura de confronto em relação à Administração Tributária e busca a sonegação sempre que tem a percepção de não estar sendo monitorado pelo Fisco.
- **Desengajamento (*disengagement*):** ocorre quando o contribuinte está decidido a não cumprir suas obrigações e adota todos os meios possíveis para evadir tributos.

Braithwaite (2003b) aponta que as posturas não são necessariamente estáticas e podem variar no tempo, com a migração de contribuintes de uma postura para outra, ou ocorrer de forma simultânea com um mesmo contribuinte em relação a diferentes tributos ou entes tributantes.

As três primeiras posturas (conformidade voluntária, conformidade por rendição e conformidade criativa) são chamadas de posturas de conformidade. As duas últimas (resistência e desengajamento), por sua vez, são chamadas de posturas de *não conformidade*.

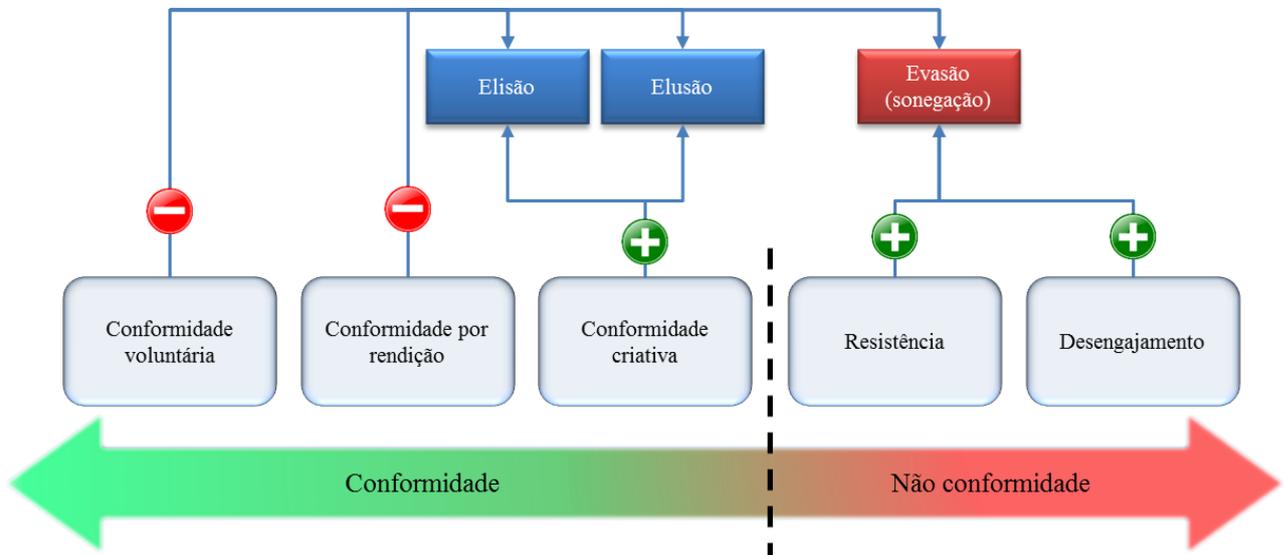
A conformidade voluntária e a conformidade por rendição estão negativamente relacionadas aos conceitos de evasão, elisão e elusão fiscal, discutidos a seguir. Em outras palavras, contribuintes com esses perfis são menos propensos à elisão, elusão e evasão de tributos. As posturas de resistência e desengajamento, por sua vez, estão positivamente relacionadas ao conceito de evasão fiscal. A conformidade criativa, por fim, está positivamente relacionada à elisão e à elusão fiscal (KIRCHLER, 2007).

As práticas de elisão e elusão fiscal, apesar de causarem redução no montante de tributos devidos, são consideradas atitudes de conformidade. As razões para tanto serão detalhadas na próxima seção. Por esse motivo, a conformidade criativa, que está positivamente relacionada a esses conceitos, é considerada uma postura de conformidade (BRAITHWAITE, 2003b).

⁷ Os conceitos de evasão, elisão e elusão serão discutidos no tópico seguinte.

A **Figura 3** relaciona as posturas de contribuintes de tributos com os conceitos de elisão, elusão e evasão fiscal, e de conformidade e *não conformidade*.

Figura 3: As posturas dos contribuintes.



Fonte: Elaborado pelo autor da pesquisa.

Braithwaite (2003a) aponta que o objetivo das Administrações Tributárias deve ser aumentar o número de contribuintes enquadrados nos perfis de conformidade voluntária e conformidade por rendição, nessa ordem.

2.1.2 Evasão, elisão e elusão fiscal

Evasão fiscal ou sonegação (*tax evasion* ou *tax cheating*, em inglês) não se confunde com elisão fiscal (*tax avoidance*). Economicamente, o efeito de ambas é muito parecido e causa impactos diretos na saúde financeira do Estado, pois resulta em diminuição de arrecadação (COBHAM, 2005). Esse efeito é chamado de brecha tributária (*tax gap*) e será visto em detalhes adiante.

Nada obstante à aparente semelhança, juridicamente há uma distinção entre os conceitos. De acordo com a Organização para a Cooperação e Desenvolvimento Econômico – OCDE⁸, evasão fiscal implica em adoção de práticas ilegais para esconder ou ignorar a

⁸ *The Organisation for Economic Co-operation and Development – OECD*, em inglês, inclui as 35 economias mais avançadas do globo (OECD, 2017b).

responsabilidade tributária, ou seja, esconder informações das autoridades fiscais para pagar menos impostos do que é exigido por lei. (OECD, 2017b).

Kirchler (2007) aponta que a evasão se dá por omissão (*omission*) ou comissão (*commission*). Exemplos de atos de omissão seriam: esconder determinado rendimento da Declaração de Imposto de Renda, ou não emitir Documento Fiscal na venda de mercadoria sujeita ao ICMS. Exemplos de atos de comissão, por sua vez, seriam: listar recibos falsos na Declaração de Imposto de Renda para aumentar as deduções, ou emitir Notas Fiscais frias⁹ de venda de mercadoria para aproveitamento de créditos do ICMS. A não emissão de Documento Fiscal e a emissão de Notas Fiscais frias são práticas sonegatórias conhecidas no âmbito do ICMS e serão discutidas ainda neste texto.

Elisão fiscal também é definida como a adoção de práticas para reduzir tributos. Essas práticas, no entanto, apesar de oportunistas, não ferem a legislação. Ao contrário, elas se aproveitam exatamente de brechas legais (*loopholes*) para reduzir o montante dos tributos devidos (OECD, 2017b). De forma simplificada, evasão fiscal é prática ilegal, ao passo que elisão fiscal não configura prática ilegal. (WEBER; FOOKEN; HERRMANN, 2014).

Há ainda outro conceito, encontrado com mais frequência na literatura estrangeira, que seria distinto de evasão e elisão fiscal. É o conceito de elusão¹⁰, ou fuga fiscal. Kirchler (2007) aponta que, sob uma perspectiva econômica, faz sentido avaliar elisão e elusão como sinônimos, visto que ambos causam efeitos financeiros similares aos cofres estatais e não representam, a rigor, uma infração à lei. Etzioni (2010) esclarece, no entanto, que levando em conta aspectos morais, ambos teriam significados distintos.

Foi o que revelou um estudo conduzido por Kirchler, Maciejovsky e Schneider (2003), que buscou mapear as representações sociais dos termos *tax avoidance* (elisão), *tax evasion* (evasão) e *tax flight* (elusão)¹¹. Os autores concluíram que elisão foi considerada um planejamento tributário inteligente, feito por pessoas físicas que não desejam infringir a lei, mas querem pagar o mínimo possível de impostos. Evasão fiscal foi associada a fraude, a práticas criminosas e arriscadas e foi reprovada do ponto de vista moral pelos entrevistados. Elusão, por sua vez, também chamada na literatura brasileira de elisão abusiva (NISHIOKA,

⁹ Para efeitos desse texto, entende-se como Nota Fiscal fria aquela sem lastro material, ou seja, uma Nota Fiscal emitida sem que tenha ocorrido, de fato, uma operação de compra e venda.

¹⁰ Tradução livre do original *Tax Flight ou Tax Circumvention*, em inglês (KIRCHLER, 2007).

¹¹ Esses termos tiveram tradução livre no presente texto e foram identificados, respectivamente, como elisão fiscal, evasão (ou sonegação) fiscal e elusão fiscal. Digno de nota é o fato de que, frequentemente, *tax avoidance* e *tax flight* são ambos traduzidos como elisão fiscal, em uma consideração apenas ao aspecto legal.

2010), foi associada ao planejamento tributário agressivo, como aquele em que são utilizados paraísos fiscais e abuso de formas jurídicas para se evitar impostos. Tais práticas, mesmo que estritamente legais, foram consideradas imorais pela maioria dos entrevistados (KIRCHLER; MACIEJOVSKY; SCHNEIDER 2003).

Um caso emblemático de elusão foi aquele que envolveu a empresa Starbucks (ARMISTEAD, 2012). Em 2012, a empresa reportou 400 milhões de libras em vendas no Reino Unido, porém não recolheu imposto de renda no país. A manobra tributária para reduzir os impostos envolveu pagamento de royalties a coligada holandesa, compra de café de uma coligada na Suíça e pagamentos de juros de empréstimos tomados de partes relacionadas. Embora não tenha sido acusada de violar as leis britânicas, o planejamento tributário agressivo da empresa se tornou alvo de duros protestos da população do Reino Unido, em uma demonstração de reprovação moral à conduta da empresa (EBRAHIMI, 2012).

A elisão e a elusão decorrem do fato de que, apesar das leis tributárias serem geralmente extensas e complexas, ainda assim elas não são capazes de prever todas as situações do mundo real (KIRCHLER, 2007). Segundo estimativas da OCDE, casos de planejamento tributário agressivo podem custar até 240 bilhões de dólares anuais em perdas globais de impostos. (OECD, 2015).

A respeito da utilização dos termos, digno de nota é o fato de que elisão fiscal é muitas vezes utilizada com sentido amplo, abrangendo tanto *tax avoidance* quanto *tax flight*, ou seja, é utilizada apenas considerando-se o aspecto de legalidade, sem levar em conta sua representação social em termos morais. O uso do termo com esse sentido mais amplo pode ser visto em Viol (2015) e Siqueira e Ramos (2005), só para citar alguns.

2.1.3 A não conformidade não intencional

Embora a elisão, a elusão e a evasão sejam práticas deliberadas e intencionais de redução de impostos, apenas a evasão consiste em infração à lei. Por isso, King e Shefrin (2002) e Sandmo (2003) a descrevem como sempre revestida de dolo, de intenção fraudulenta. Os autores definem a evasão como ato deliberado e doloso de *não conformidade*, que resulta no pagamento de menos impostos do que se deveria.

Essa distinção entre ocorrência ou não de dolo é importante porque, além dos conceitos estudados de elisão, elusão e evasão, existe ainda outro relacionado à *não conformidade* fiscal, que também pode resultar em redução do montante de tributos pagos aos

cofres públicos. É o caso da *não conformidade* involuntária, causada pelo descumprimento não intencional e que ocorre, inclusive e principalmente, com contribuintes que adotam posturas de conformidade perante o Fisco.

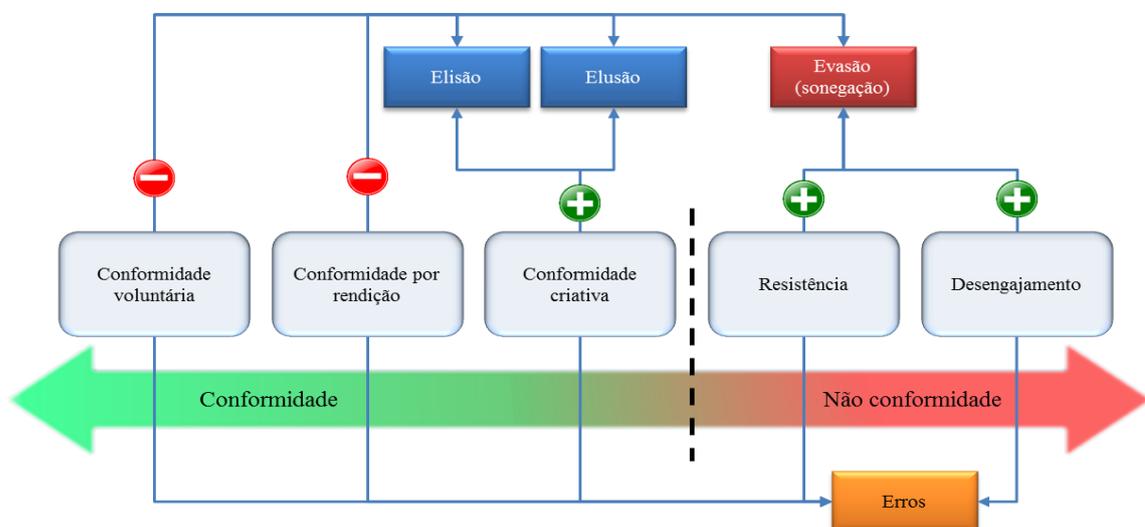
Esse descumprimento não intencional pode ou não causar redução dos montantes devidos aos cofres públicos. O atraso no envio de uma declaração de imposto pode não causar redução nos montantes de tributos devidos e ainda assim ser um descumprimento da norma que estabeleceu o prazo de envio. O envio de uma declaração com valores menores do que os devidos, no entanto, mesmo que de forma involuntária, acarreta redução indevida da responsabilidade tributária (VIOL, 2015).

Para efeitos de simplificação, neste estudo não serão feitas distinções entre *não conformidade* involuntária, que acarreta redução indevida de tributos, e aquela, também involuntária, que não acarreta tais reduções.

As **ocorrências de *não conformidade* involuntária serão chamadas** nesse estudo de **erros**. De acordo com Viol (2015), elas derivam de falhas de cálculo, dificuldade na interpretação da lei tributária, excesso de regulações e formalismos, dentre outras razões. O excesso de regulações e formalismo para cumprimento das obrigações tributárias é chamado de Fardo de Cumprimento Tributário e sua relação com a *não conformidade*, e consequente elevação da brecha tributária, será vista adiante.

A **Figura 4** expande a **Figura 3** para demonstrar que qualquer tipo de contribuinte, independente do perfil de conformidade no qual mais se enquadra, está sujeito à *não conformidade* por erro.

Figura 4: As posturas dos contribuintes e sua relação com a *não conformidade* por erro.



Fonte: Elaborado pelo autor da pesquisa.

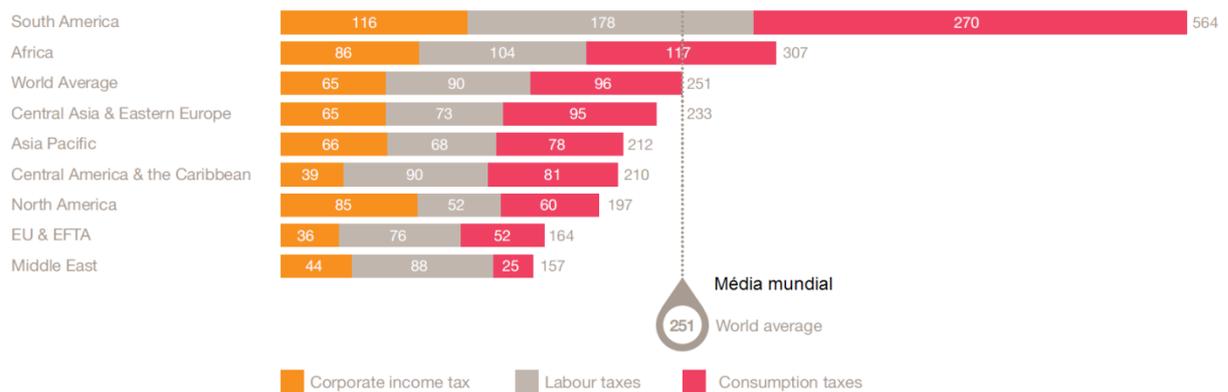
2.1.4 O fardo de cumprimento tributário

Kirchler (2007) argumenta que o excesso de regulações e formalismos, chamado de fardo ou *burden* de cumprimento tributário, impulsiona a ocorrência de erros e estimula a informalidade e a proliferação do número de transações escondidas. Um fardo elevado resulta tanto em ocorrências associadas ao dolo quanto naquelas involuntárias.

O World Bank Group e a Price Waterhouse Coopers defendem que a diminuição do *burden* deve ser objetivo de todos os governos e, por isso, publicam regularmente um relatório que mapeia o fardo de conformidade em diversos países. Esse fardo é medido em quantidade de horas de trabalho necessárias para cumprimento das obrigações tributárias (WORLD BANK GROUP; PWC, 2017).

A **Figura 5**, baseada em relatório de 2017, que se fundamentou em dados de 2015, mostra o tempo de cumprimento por tipo de tributo em diferentes regiões do globo.

Figura 5: Tempo necessário em horas para cumprimento de obrigações tributárias.

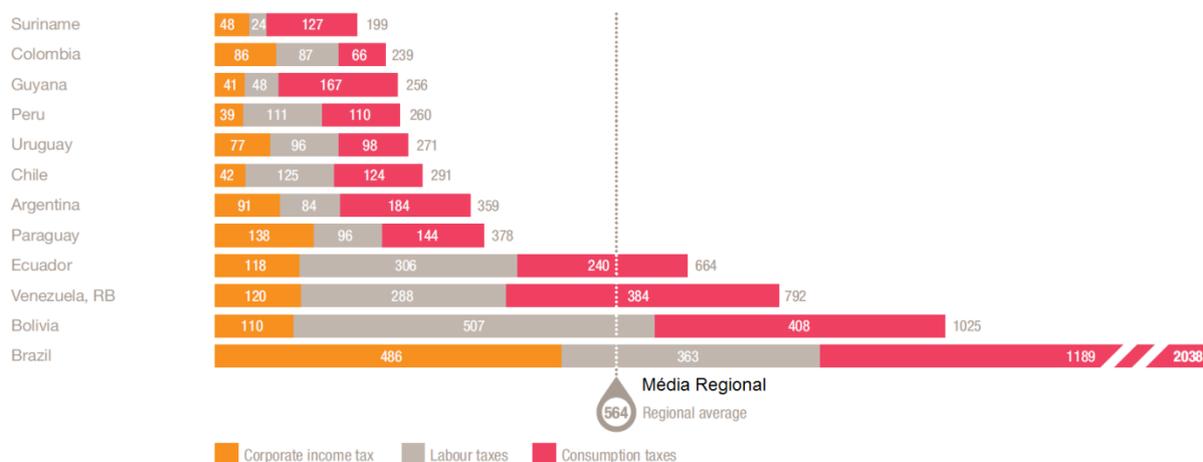


Fonte: Adaptado de World Bank Group e PWC (2017).

De acordo com a **Figura 5**, a América do Sul é a região em que o fardo de cumprimento é maior. Ela demonstra, também, o elevado impacto dos impostos sobre consumo – dentre os quais o ICMS brasileiro – no fardo de cumprimento. Na América do Sul, os impostos sobre consumo representam aproximadamente 48% do tempo total.

No Brasil, esse percentual supera 58% do tempo total, como é possível observar na **Figura 6**.

Figura 6: Tempo de cumprimento em horas na América do Sul.



Fonte: Adaptado de World Bank Group e PWC (2017).

O relatório do World Bank Group e da PWC apontou que, pela primeira vez desde a primeira edição do estudo, há mais de dez anos, o Brasil reduziu 562 horas do tempo total necessário para cumprimento. Essa redução se deveu à implementação de sistemas eletrônicos para preparação, preenchimento e pagamento de impostos e, ainda, redução de algumas obrigações acessórias. Apesar da boa notícia, o total de 2.038 horas é ainda mais de oito vezes a média mundial (WORLD BANK GROUP; PWC, 2017).

Kirchler (2007) lista estudos que sugerem que o excesso de regulamentação e formalismo é a principal causa do elevado número de horas para cumprimento. Esse excesso também é percebido como limitante à atividade empreendedora, estímulo à informalidade, ao uso de transações escondidas e contribui para a elevação da brecha tributária (KIRCHLER, 2007).

2.1.5 A brecha tributária

Existem diferentes definições do termo brecha tributária (*tax gap*). Alguns autores a definem de forma ampla, e nesse caso ela seria a diferença entre a arrecadação efetiva e a legalmente intencionada, ou seja, levaria em conta tanto a letra quanto o espírito da lei (THACKRAY, 2012).

De acordo com essa definição, a brecha tributária seria a diferença entre aquilo intencionado pelo legislador e o efetivado na prática, e compreenderia, além da elisão, elusão, evasão e erros de cumprimento, outras formas de diminuição da carga tributária potencial.

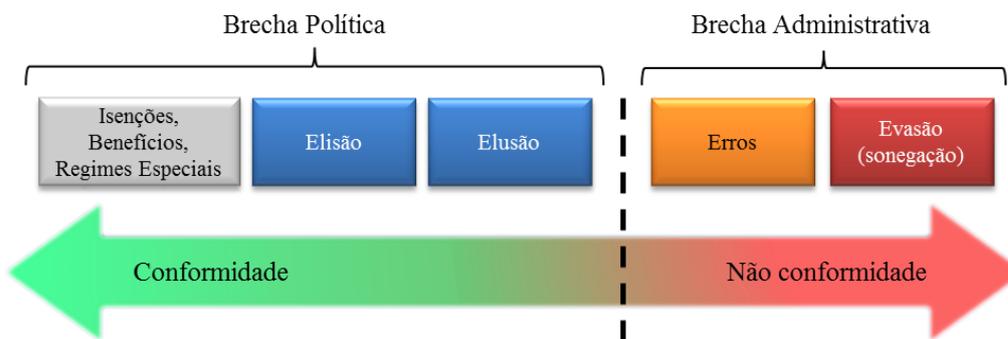
Essas formas se apresentam como isenções, regimes especiais e benefícios fiscais. Por atenderem as formalidades jurídicas impostas, são formalmente classificadas como posturas de conformidade, mesmo que contribuam para o aumento da brecha (VIOL, 2015).

A outra definição de brecha tributária adota um conceito mais estrito e abarca apenas as práticas de *não conformidade* (erros e evasão), ou seja, aquelas que ferem a escrita da lei (ALM, 2012). De acordo com ela, por não se configurarem em desrespeito à escrita da lei, as práticas de isenções, benefícios e regimes especiais, assim como as de elisão e elusão, seriam componentes da chamada brecha política, não abrangidas pelo conceito estrito.

As práticas de evasão e os erros – se estes últimos culminarem em redução indevida de tributos – por sua vez, afrontam diretamente o texto normativo e seriam, portanto, componentes da chamada brecha administrativa (VIOL, 2015).

A **Figura 7** demonstra os conceitos abarcados pelas brechas política e administrativa.

Figura 7: Os componentes da brecha tributária segundo a definição ampla.



Fonte: Elaborado pelo autor da pesquisa.

A distinção entre brecha política e brecha administrativa se faz relevante quando se estudam os responsáveis por combatê-las. Como será discutido no item 2.6 – *O Combate às práticas que reduzem a arrecadação*, a diminuição da brecha administrativa é essencialmente responsabilidade das Administrações Tributárias, ao passo que a brecha política é responsabilidade também de outros atores, como os Poderes Legislativo e Judiciário (VIOL, 2015).

2.1.5.1 Os efeitos da brecha tributária

O aumento da brecha tributária reduz o montante de tributos arrecadados pelo Estado. Os tributos são necessários porque se constituem na principal forma do Estado se financiar para poder cumprir suas funções, dentre as quais as chamadas funções fiscais, que são divididas em função alocativa, função distributiva e função estabilizadora. Essas funções, no entendimento de Musgrave e Musgrave (2004), são definidas conforme a seguir.

A função alocativa diz respeito à alocação de recursos pelo Estado para fornecer bens e serviços públicos. É através dela que é decidido onde serão alocados os recursos arrecadados. Dentre os bens e serviços públicos, pode-se destacar Segurança Pública, Saúde, Educação e Infraestrutura.

A função distributiva, também chamada de redistributiva, é uma forma que o Estado possui de redistribuir a renda na sociedade através de transferências, impostos e subsídios. O objetivo dessa função é realocar a renda da forma que cada sociedade considera mais justa, diminuindo a desigualdade social. Ela ocorre através de diferentes mecanismos. Um deles é a redistribuição direta, em que o Estado tributa mais aqueles com maior capacidade contributiva e menos os indivíduos mais necessitados.

A função estabilizadora, não menos importante, é aquela em que o Estado procura suprir as deficiências do mercado, a fim de amenizar crises econômicas e garantir o nível de emprego e desenvolvimento.

Embora haja um intenso debate atualmente a respeito de como deve ser alocada a carga tributária e sobre qual seria o tamanho ideal do Estado, há certo consenso de que o Estado deve existir e ser financiado pela sociedade para poder cumprir suas funções (MANKIW, 2014; MYLES, 2009; FIELDHOUSE, 2013; PIKETTY, 2014).

Essa necessidade de recursos por parte do Estado faz com que a tributação seja um foco constante de políticas que buscam aumento de receitas. Em momentos de crise e maior escassez de recursos, muitos governos optam por incrementar a arrecadação via “aumentos de alíquotas, instituição de impostos extraordinários, recorrência a bases com efeitos em cascata, ou mesmo atraso na devolução de créditos de IVA¹²” (VIOL, 2015, p. 53).

Essas estratégias compõem a chamada abordagem de bases estreitas e altas alíquotas (*narrow base – high rate approach*), em oposição àquela de bases amplas e baixas alíquotas (*broad base – low rate approach*). O nome “bases estreitas e altas alíquotas” se baseia no fato

¹² O Imposto sobre Valor Agregado – IVA é um tipo de imposto sobre consumo que será detalhado adiante nesse trabalho.

de que aumentos da carga tributária costumam ter alta incidência na parte da população que detém a menor parte da riqueza (OECD, 2010).

Diversos estudos já demonstraram que é mais efetivo diminuir as alíquotas e tornar a incidência mais ampla, de forma a tributar a população de maneira mais justa. O estudo de número 19 da OCDE, por exemplo, aponta que a abordagem *broad base – low rate* é superior porque reduz distorções nas decisões econômicas e melhora o bem-estar social da população. (OECD, 2010).

As práticas de elisão, elusão, evasão fiscal, e até mesmo a *não conformidade involuntária* (erros) – componentes das brechas política e administrativa – vão de encontro à ampliação de bases e diminuição de alíquotas, pois alguns indivíduos, deliberadamente ou não, e de forma dolosa ou não, encontram meios de diminuir sua responsabilidade tributária, ao passo que outros continuam arcando com o fardo (*tax burden*) de financiar o Estado. (VIOL, 2015).

As reduções nos montantes devidos através de tais práticas podem diminuir a receita fiscal e aumentar a brecha tributária a tal ponto que o Estado acabe por lançar mão de aumentos em tributos ou alíquotas, o que na prática se configuraria na adoção da abordagem bases estreitas – altas alíquotas, que penaliza apenas aqueles que já cumprem com suas obrigações tributárias e distorce a competição entre os agentes econômicos. (VIOL, 2015).

2.1.5.2 O tamanho da brecha tributária

Independentemente da adoção de um conceito amplo ou estrito para a brecha tributária, sua mensuração é extremamente difícil. A razão principal é a dificuldade de detecção das práticas que resultam em redução de tributos. Como apontado por Alm (2012), essa dificuldade decorre do fato de que a brecha tributária é majoritariamente causada por práticas que indivíduos e firmas desejam esconder.

Medeiros (1993, p. 7) ressalta que a natureza da evasão implica em ocultação deliberada de informações, o que torna qualquer medida de sonegação uma “aproximação de uma quantidade que nunca é, de fato, conhecida”. Da mesma forma, os regimes especiais, incentivos e benefícios fiscais carecem de transparência e, por isso, estimá-los é uma tarefa normalmente difícil. Detectar erros e práticas de elisão e elusão fiscal também nem sempre é trivial. (AFONSO, 2014).

Apesar da dificuldade inerente à medição da brecha tributária, diversos métodos foram propostos para estimar seus valores. Medeiros (1993) aponta que esses métodos são

geralmente classificados em diretos e indiretos. Os métodos diretos são aqueles que investigam diretamente os agentes econômicos através de amostras estatisticamente significantes ou não. Os métodos indiretos, por sua vez, buscam rastros de evasão, elisão e elusão em variáveis econômicas agregadas.

Dentre os métodos diretos estão as auditorias fiscais. Apesar de contarem com grau razoável de acuracidade, elas são normalmente custosas e a seleção de uma amostra que represente a população apresenta diversos desafios. Além disso, como aponta Medeiros (1993), mesmo uma auditoria profunda e exaustiva não oferece garantia de captar por completo todas as práticas que ensejam redução de tributos.

Outro método direto é a condução de pesquisas com voluntários. Conforme detalha Alm (2012), tais pesquisas buscam medidas morais de comportamento, ou seja, buscam entender como os indivíduos se posicionam em relação a determinados assuntos, a exemplo da evasão. A fragilidade desse método está na possibilidade de distorções dos resultados em virtude de respostas nem sempre verdadeiras. Ainda dentre os métodos diretos, podemos também citar experimentos em laboratórios e estudos de campo, como aqueles detalhados em Torgler (2007).

Os métodos indiretos se baseiam em dados agregados. Dentre eles está um que se fundamenta em indicadores de comparação de variáveis agregadas, como renda declarada e consumo das famílias. (WEBER; FOOKEN; HERRMANN, 2014). Outro método indireto bastante utilizado é o cálculo da sonegação com base na estimativa do tamanho da *shadow economy* (SCHNEIDER, 2005).

A *shadow economy*, traduzida em Tanzi, Schneider e Barbosa Filho (2009) como economia subterrânea, ou ainda economia informal, inclui a produção e os serviços que são ocultados das autoridades públicas com intuítos diversos, dentre os quais a diminuição da incidência tributária.

Shelak (1997) propõe uma definição mais ampla, que englobaria, além dos itens anteriores, aquelas atividades e operações consideradas ilegais em diversos países, a exemplo de tráfico de entorpecentes, contrabando, prostituição e jogos de azar.

Em ambas as definições, a evasão fiscal é componente de destaque e engloba não apenas a sonegação decorrente da completa ocultação de empresas (não inscrição nos cadastros públicos), mas também aquela adotada por empresas devidamente registradas e que ocultam operações e transações de forma parcial.

De acordo com Siqueira e Ramos (2005), a *shadow economy*, pode ser estimada com base na demanda por moeda. Segundo os autores, a estimativa se fundamenta na hipótese de

que a maioria das transações não declaradas é feita em dinheiro e, por isso, seria possível estimar a economia subterrânea como função do aumento dessa demanda.

Alm, Martinez-Vazquez e Schneider (2004), em estudo baseado em dados de 2000, estimaram o tamanho da *shadow economy* no mundo. Na **Tabela 3**, que lista alguns dos países pesquisados, é possível observar que as maiores taxas estimadas de economia subterrânea se encontram nos países em desenvolvimento.

No Brasil, a economia subterrânea foi estimada em 39,8% do Produto Interno Bruto – PIB. Essa medida pode ser um indicativo do tamanho da brecha tributária no país.

Tabela 3: Tamanho (percentual PIB) da shadow economy em alguns países – 2000.

País	Shadow Economy em 2000	País	Shadow Economy em 2000
África do Sul	28,4%	França	15,2%
Alemanha	16,0%	Índia	23,1%
Arábia Saudita	18,4%	Israel	21,9%
Argentina	25,4%	Itália	27,1%
Austrália	14,1%	Japão	11,2%
Áustria	9,8%	México	30,1%
Bélgica	22,2%	Nova Zelândia	12,8%
Bolívia	67,1%	Peru	59,9%
Brasil	39,8%	Portugal	22,7%
Canadá	16,0%	Reino Unido	12,7%
Chile	19,8%	Singapura	13,1%
China	13,1%	Suécia	19,2%
Colômbia	39,1%	Suíça	8,6%
Coreia do Sul	27,5%	Turquia	32,1%
Equador	34,4%	Uruguai	51,1%
Espanha	22,7%	Venezuela	33,6%
Estados Unidos	8,7%		

Fonte: Adaptado de Alm, Martinez-Vazquez e Schneider (2004).

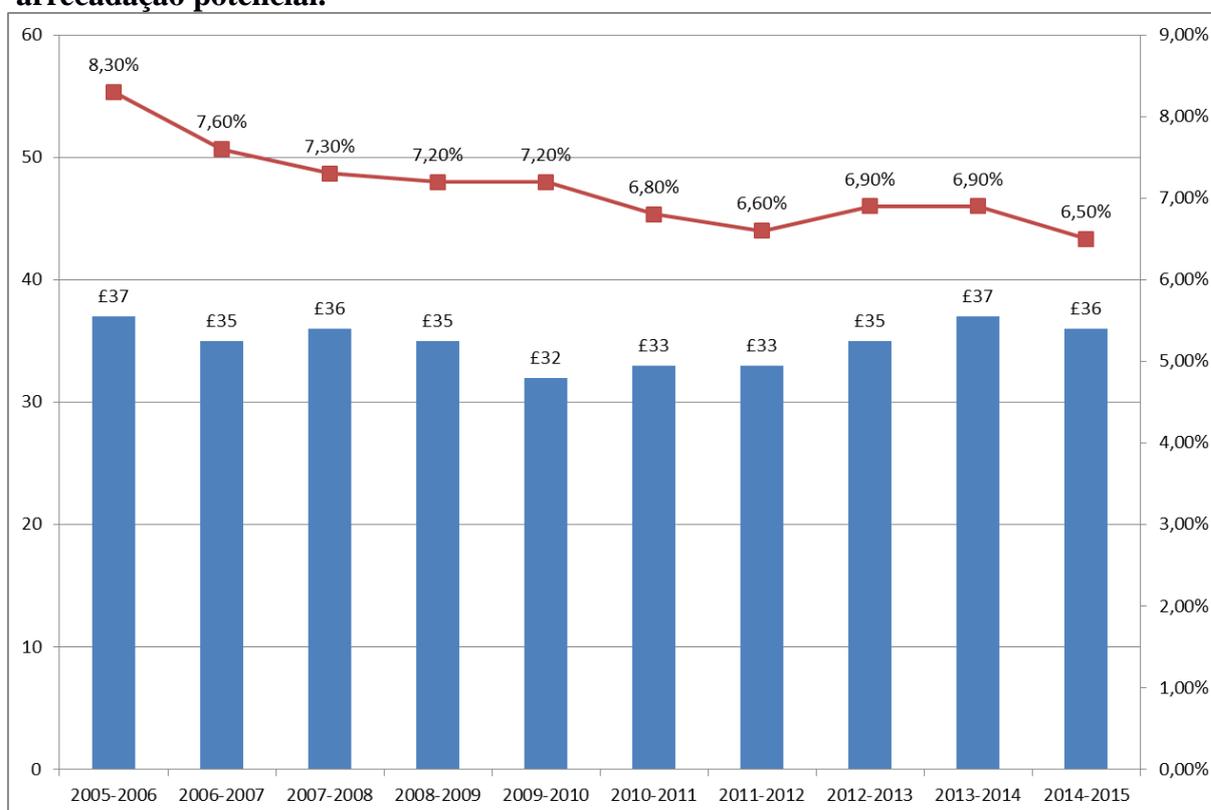
Tanto os métodos diretos como os indiretos são sujeitos a diversas imprecisões e críticas e isso se deve ao fato de que medir a brecha tributária é uma tarefa bastante complexa. (SCHNEIDER, 2005). Nada obstante, a medição é importante para que sejam encontrados meios apropriados de diminuí-la. Por esse motivo, diversos países têm empreendido esforços para estimar o tamanho de seus *tax gaps*.

No Brasil não há uma divulgação oficial dos números pelas Administrações Tributárias. Siqueira e Ramos (2005) citam, no entanto, estudos que estimam que a brecha tributária gira entre 30% e 40%, número este corroborado pelo estudo demonstrado na **Tabela 3**.

Alguns países têm estimado a brecha tributária ao longo do tempo e por tipo de tributo. É o caso do Reino Unido. O Her Majesty's Revenue and Customs – HMRC¹³ tem se esforçado para publicar regularmente a brecha tributária da região. O relatório publicado em outubro de 2016, que se baseou em dados de 2014 e 2015, demonstrou que a brecha vem caindo ano a ano desde 2005, e encontra-se atualmente em patamares próximos a 6,5% do valor potencial de arrecadação, o que corresponderia a 36 bilhões de libras (HMRC, 2016).

A **Figura 8** demonstra a evolução do tax gap no Reino Unido nos últimos anos.

Figura 8: A brecha tributária no Reino Unido (em bilhões de libras) como percentual da arrecadação potencial.



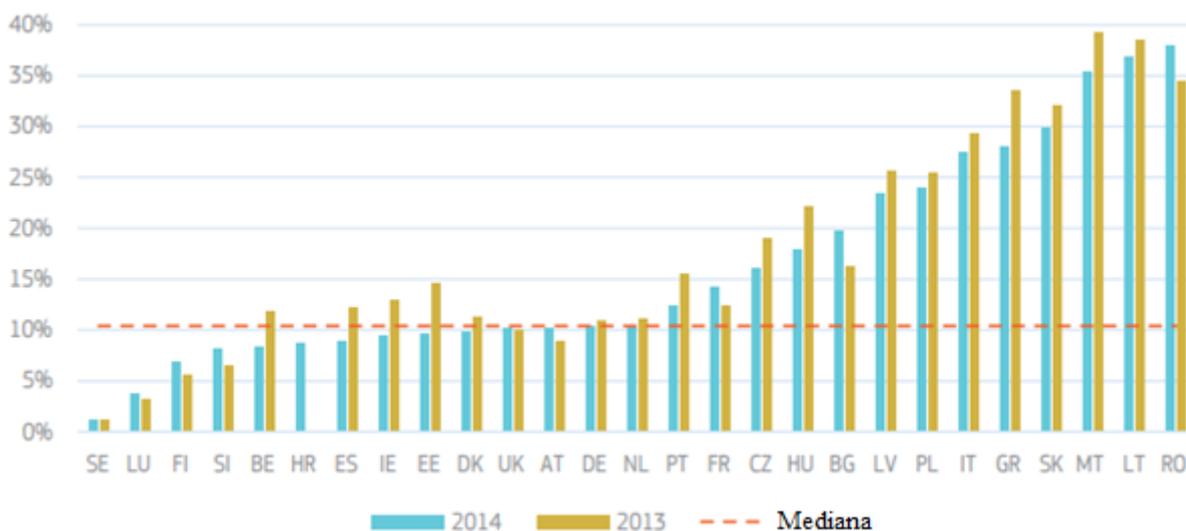
Fonte: Adaptado de HMRC (2016).

O relatório traz, ainda, a brecha por tipo de tributo. Em termos percentuais, o tributo com o maior *gap* é um imposto sobre consumo, o Imposto sobre Valor Agregado – IVA (*Value Added Tax – VAT*, em inglês). A brecha nesse tributo representa 10,3% do total de sua arrecadação potencial (HMRC, 2016).

¹³ Para efeitos de simplificação, propõe-se entender o HMRC como uma espécie de Receita Federal do Reino Unido.

A European Commission divulgou em setembro de 2016 um relatório relativo ao ano de 2014 que apresenta a brecha tributária no âmbito do IVA na União Europeia. De acordo com o relatório, a brecha de arrecadação nesse imposto foi de 10,4%, conforme se vê na **Figura 9**.

Figura 9: Brecha do IVA na União Europeia em 2014 e 2013.



Fonte: Adaptado de European Commission (2016).

2.1.6 O combate às práticas que reduzem a arrecadação

Como demonstrado na **Figura 7**, a brecha tributária possui causas diversas, que vão desde a sonegação até a concessão de isenções fiscais. Coelho et al. (2017) discutem que apesar dos efeitos das isenções, regimes especiais e benefícios fiscais na brecha tributária serem controversos, a argumentação de que essas práticas não a aumentam carece de evidências empíricas. Nada obstante, o aprofundamento quanto a esses componentes da brecha política foge ao escopo deste estudo.

As demais causas da brecha tributária e os mecanismos encontrados na teoria para sua diminuição serão detalhados a seguir.

2.1.6.1 O combate à elisão e à elusão

As práticas de elisão e elusão estão associadas à conformidade criativa, não se tratando, portanto, de *não conformidade* do ponto de vista jurídico. Por isso, tais práticas são consideradas parte da brecha política e não da administrativa. Uma questão que é colocada em função dessa distinção, é de quem seria a responsabilidade de diminuição da brecha política.

Viol (2015) defende que as principais ações de combate à brecha política não caberiam às Administrações Tributárias e sim a outras instâncias estatais, como os Poderes Legislativo e Judiciário. A autora argumenta que "Problemas que envolvem interpretação legal e esquemas de elisão [e elusão] são melhor corrigidos mediante mudanças legais e discussões com agentes externos, como o setor privado e o poder judiciário" (VIOL, 2015, p. 6).

Braithwaite (2005) apresenta uma proposta para combater a elisão e a elusão, que passa por uma reforma da lei tributária, com a integração de regras em princípios. Segundo o autor, as leis tributárias atuais, baseadas puramente em regras, propiciam uma "cultura de gato e rato"¹⁴, bastando uma brecha legal ser fechada para que outra seja encontrada e explorada.

Como solução, é proposta a criação de princípios gerais, que serviriam como guias de comportamentos. A lei tributária listaria regras para transações corriqueiras e os princípios gerais seriam utilizados para interpretá-las.

De forma mais detalhada, a proposta envolveria os seguintes passos (BRAITHWAITE, 2005):

1. Definir princípios gerais e abrangentes, que seriam de observância obrigatória.
2. Incluir entre esses princípios um que determine ser ilegal qualquer esquema que tenha como propósito principal tirar vantagem da lei tributária, mesmo que tal esquema não seja proibido pelas regras tributárias vigentes.
3. Definir um conjunto de regras que abranjam a complexidade da temática tributária.
4. Determinar que quando houver conflito entre uma regra e um princípio, este prevalecerá. Os princípios não seriam meros instrumentos para interpretação das regras. As regras é que seriam instrumentos para garantir a aplicação dos princípios.
5. Definir regras específicas para abranger as transações concretas mais corriqueiras, mas com a clara menção de que não se trata de uma lista exaustiva e sim exemplos de condutas que devem ser seguidas.

¹⁴ Tradução livre do original em inglês: *cat-and-mouse culture*.

6. Frisar, na lei tributária, que a interpretação das regras deve ser feita de forma a privilegiar a aplicação dos princípios gerais. Essa seria uma forma do legislador deixar claro aos contribuintes, às autoridades tributárias e ao judiciário, que os princípios são a base da lei tributária e as situações do dia a dia devem ser sempre interpretadas com eles em mente.
7. No julgamento de um conflito tributário, caso juízes, por eventual força do hábito, se atenham mais a determinada regra 123 do que ao princípio X a que esta se vincula, o legislativo deveria entrar em ação e explicitar na lei tributária que a regra 123 viola o princípio X e, portanto, não deveria mais ser aplicada. Isso eliminaria qualquer precedente que legitimasse o uso da regra 123.
8. Manter uma relação aberta e comunicativa com o judiciário, empresas e a sociedade em geral, a respeito da importância e significado dos princípios estabelecidos.

2.1.6.2 O combate à evasão e o estímulo à diminuição dos erros

Tanto a evasão quanto a *não conformidade* involuntária causada por erros – se esta última culminar em redução indevida de tributos – fazem parte da brecha administrativa. De acordo com Kirchler (2007), o fato da brecha administrativa ser composta por práticas que implicam em violação às leis faz com que as Administrações Tributárias sejam os principais atores de referência no seu combate e diminuição.

Segundo o autor, a evasão deve ser combatida porque (KIRCHLER, 2007):

- Reduz os montantes arrecadados aos cofres públicos.
- Gera impactos negativos na provisão de serviços e bens públicos.
- Limita o potencial de redistribuição de renda que se poderia obter com os tributos, principalmente os progressivos¹⁵.
- Difunde uma sensação de injustiça tributária e social.
- Cria uma cultura de desrespeito à lei.

Os casos de *não conformidade* involuntária, mesmo se causarem alguns dos efeitos da evasão – como “redução dos montantes arrecadados aos cofres públicos” e “geração de

¹⁵ Impostos progressivos são aqueles que levam em consideração as características dos indivíduos no estabelecimento de alíquotas, que são graduadas de acordo com sua capacidade contributiva. É uma forma do Estado tributar mais aqueles que mais podem pagar. Mais detalhes podem ser encontrados em Giambiagi e Além (2011).

impactos negativos na provisão de serviços e bens públicos” – devem ser tratados de forma diferente pelo Fisco. É o que afirmam autores como Braithwaite (2003b), Viol (2015), Kirchler (2007) e Weber, Fooken e Herrmann (2014), só para citar alguns.

Viol (2015, p. 5), por exemplo, afirma que “não se deveriam gastar recursos em auditorias sofisticadas para tratar riscos associados à falta de informação ou erros não intencionais”. Esse entendimento reforça a necessidade de segmentação dos contribuintes em perfis de conformidade.

Weber, Fooken e Herrmann (2014) também entendem que as ferramentas a serem aplicadas devem ser diferentes de acordo com o comportamento tributário do contribuinte. Dentre as ferramentas citadas pelos autores estão as de coerção, para aqueles contribuintes que praticam a evasão, e as de cooperação e facilitação, para diminuir o fardo de cumprimento daqueles que desejam cumprir corretamente suas responsabilidades tributárias, mas nem sempre conseguem.

2.1.7 O comportamento do contribuinte

O objetivo precípua das Administrações Tributárias é o aumento da conformidade fiscal e consequente diminuição da brecha administrativa. Nesse sentido, vale a reprodução da visão do HMRC: “Nós fecharemos a brecha tributária, nossos clientes terão a percepção de que o sistema tributário é simples e isonômico, nós seremos vistos como uma organização eficiente e altamente profissional¹⁶” (TRACKRAY, 2012).

Braithwaite (2003a) e Kirchler (2007) reforçam que para aplicar o tratamento correto aos contribuintes, as Administrações Tributárias devem ser capazes de segmentá-los em perfis de cumprimento, e para isso, devem entender quais fatores afetam o comportamento tributário e quais características estão associadas a esses fatores.

As seções a seguir discutem as principais teorias que buscam apontar os fatores de influência no cumprimento tributário, os chamados modelos de conformidade.

¹⁶ Do original em inglês traduzido de forma livre: *We will close the tax gap, our customers will feel that the tax system is simple for them and even-handed, and we will be seen as a highly professional and efficient organisation* (TRACKRAY, 2012).

2.1.7.1 A decisão de sonegar: modelo clássico

Allingham e Sandmo (1972) foram os pioneiros na proposição de ferramentas econômicas para análise da conformidade tributária. A análise efetuada pelos autores se baseou nos estudos da chamada Economia do Crime – a exemplo daqueles publicados por Becker (1968) e Tulkens e Jacquemin (1971) –, em estudos da Economia da Incerteza – como aqueles publicados por Arrow (1970) e Mossin (1968) –, e nos axiomas de comportamento racional em condições de incerteza – propostos na Teoria da Utilidade de von Neumann, Morgenstern e Khun (1953).

A análise se baseou no imposto de renda, mas as conclusões podem também ser aplicadas a outros tributos, conforme observado por Siqueira e Ramos (2005).

O modelo proposto parte do pressuposto de que a decisão pela sonegação se daria em um cenário de incerteza e seria tomada com base em uma oportunidade percebida e em cálculos racionais e utilitaristas. Apresentada uma oportunidade de escolha, os contribuintes avaliariam o retorno esperado das diversas alternativas e se decidiriam por aquela que apresentasse o maior retorno. A oportunidade percebida seria, portanto, o principal fator de influência na conformidade tributária (ALLINGHAM e SANDMO, 1972).

A evasão fiscal seria, assim, um fenômeno econômico-racional. No momento de decidir por sonegar ou declarar os impostos corretamente, os contribuintes analisariam o risco e o benefício da sonegação, com o objetivo de maximizar o retorno econômico. De um lado, seria analisada a alternativa de pagar os impostos corretamente e de outro, a evasão, que poderia trazer lucro – no caso de não detecção –, ou prejuízo – no caso de detecção pelo Fisco.

A decisão se daria através de cálculos baseados em percepções de probabilidade de detecção, peso da punição, e lucro possível. Se a análise das variáveis apontasse a evasão como melhor escolha utilitarista, o indivíduo, agindo como *homo economicus*, optaria por ela (KIRCHLER, 2007).

O modelo foi baseado em punições pecuniárias apenas, mas os autores não ignoraram que é possível, em alguns países, a aplicação de punições de restrição de liberdade (ALLINGHAM e SANDMO, 1972).

Considerando a motivação de sonegar um fenômeno puramente financeiro e racional, o contribuinte agiria como ser amoral e predisposto a se comportar de maneira desonesta, sendo o cumprimento tributário estimulado multas pesadas e auditorias eficazes. (VIOL, 2015).

Autores como Balestrino e Galmarini corroboram essa visão utilitarista do contribuinte ao afirmar que “no mundo real, as pessoas normalmente se esforçam ao máximo para sonegar e elidir tributos e a maioria dos governos luta uma batalha constante contra essas atividades” (BALESTRINO; GALMARINI, 2003, p. 51).

Outros, no entanto, a exemplo de Andreoni, Erard e Feinstein (1998), consideram o modelo uma descrição pobre do mundo real e apontam que estratégias de estímulo ao cumprimento baseadas apenas em detecção e punibilidade podem até ser um ponto de partida razoável, mas teriam pouca efetividade no longo prazo. Os autores defendem como solução uma abordagem multifacetada, capaz de punir os desvios, mas também de estimular e premiar o bom comportamento.

Outra crítica ao modelo clássico é sua fraca capacidade de explicar os altos níveis de conformidade em muitos países. Simpson (2002) lista alguns estudos que demonstram elevados níveis de economia subterrânea e brecha tributária, mas também muitos outros que demonstram altos níveis de conformidade. Segundo James e Alley (2002), tais níveis de conformidade são maiores do que o modelo clássico poderia explicar.

Os estudos citados por Simpson (2002), que mostram altos níveis de conformidade, refutam a argumentação do modelo clássico ao fundamentar teorias que defendem que apenas uma minoria de contribuintes sonega impostos, enquanto que a maioria cumpre honestamente suas obrigações.

Para explicar a coexistência de estudos que demonstram elevados níveis de economia subterrânea e brecha tributária, e outros que apontam para altos níveis de conformidade, o autor supõe que os poucos contribuintes que sonegam, sonegam grandes montantes, ou seja, o crime corporativo seria intenso.

Por fim, os próprios autores do modelo, Allingham e Sandmo, reconheceram que fatores não econômicos podem exercer influência na decisão de cumprimento, como a preocupação com a reputação própria, por exemplo (ALLINGHAM e SANDMO, 1972). Esses fatores adicionais vão de encontro à tese de que o indivíduo age como *homo economicus* no tocante ao pagamento de tributos.

Nada obstante às críticas, o modelo foi amplamente utilizado e estendido em diversos estudos posteriores, como aqueles elencados por Siqueira e Ramos (2005).

2.1.7.2 A decisão de sonegar: modelos recentes

Os estudos de Webley, Cole e Eidjar (2001) demonstraram que a oportunidade percebida é, de fato, um dos principais elementos na decisão de sonegar, mas não o único.

Nessa esteira, estudos empíricos elencados por Kirchler (2007) demonstraram que a decisão de sonegar não é puramente financeira e utilitarista, mas está, ao contrário, sujeita a diversos outros fatores de natureza não econômica. O autor aponta que esses estudos – efetuados em diversos países, utilizando diferentes metodologias – demonstraram que a grande maioria das pessoas está disposta a pagar seus tributos e não baseia suas decisões tributárias em fatores puramente econômico-financeiros de maximização de resultados.

Long e Swingen (1991) também afirmam que muitos contribuintes não são predispostos à sonegação e não buscam meios de burlar o sistema, mesmo em situações em que o modelo clássico mapearia a evasão como a melhor escolha racional.

Kirchler (2007) argumenta que se o modelo clássico funcionasse para todos os contribuintes, seria razoavelmente fácil para as Autoridades Tributárias aumentarem a conformidade tributária, bastando para isso maior frequência de auditorias e punibilidades mais severas. Se o aumento da frequência das auditorias não fosse possível devido ao alto custo, ainda assim o aumento do peso e rigor das punições já ajudaria a alcançar o efeito de dissuadir os contribuintes da decisão de sonegar¹⁷.

Por esse motivo, se as autoridades tributárias desejam dispensar o tratamento que cada contribuinte merece, deve considerar também os fatores não econômicos que influenciam o comportamento tributário, caso contrário, não será capaz de segmentá-los corretamente (Kirchler, 2007).

Lawler (1998), Alm e Torgler (2006) e Braithwaite (2003b) elencam diversos fatores não econômicos que influenciariam a decisão de cumprimento. Dentre eles estão: valores éticos, padrões morais, crenças religiosas, percepções quanto à justiça tributária, percepção quanto ao poder de controle e detecção do Fisco, grau de conhecimento quanto aos tributos, postura frente ao governo, grau de aversão a risco, características demográficas (idade, gênero, renda, classe social), normas sociais, influência de outras pessoas e grupos de referência, fatores comportamentais e culturais, e fatores psicológicos ligados ao clima de confiança entre contribuintes e autoridades tributárias.

¹⁷ Do original em inglês traduzido de forma livre: If the model worked, it would be easy for finance ministers to increase tax compliance just by increasing the frequency of audits and severity of fines. If audits were too expensive, increasing fines alone would also help (KIRCHLER, 2007, p. 107).

Segundo Kirchler (2007), a adoção do modelo clássico, e consequente desconsideração de fatores não econômicos na determinação do perfil dos contribuintes, pode ter efeito adverso do desejado, pois ao não ser capaz de categorizar os contribuintes corretamente, o Fisco corre o risco de trata-los todos da mesma forma, o que gerará descontentamento naqueles que agem corretamente e sensação de impunidade naqueles mal intencionados (KIRCHLER, 2007).

Se o Fisco adotar uma postura de hostilidade, chamada de “polícia e ladrão” ou “comando e controle”, atitudes de resistência e *não conformidade* tributária seriam estimuladas mesmo naqueles contribuintes que inicialmente eram bem-intencionados. De outro lado, se for adotada pelas autoridades tributárias uma postura “serviço e cliente” com aqueles contribuintes decididos a sonegar, a *não conformidade* recorrente e impune deles pode contaminar outros contribuintes, que veem na impunidade uma injustiça com aqueles que apresentam comportamento honesto¹⁸.

Nesse contexto, Taylor (2003, p. 89) aponta que as autoridades tributárias devem considerar as diferenças entre os contribuintes ao invés de adotar um tratamento padrão a todos eles. A autora argumenta que: “tratar os contribuintes como atores racionais, utilitaristas, que buscam apenas o interesse próprio em todas as situações é limitar nossa habilidade de entender o processo que compreende o comportamento tributário”¹⁹.

Nesse diapasão, é importante reforçar que o objetivo deste texto, ao analisar as diferenças entre o modelo clássico e os modelos mais recentes de conformidade tributária, não é afastar a necessidade de punição a contribuintes mal-intencionados. Ao contrário, nesta dissertação busca-se encontrar formas de segregar contribuintes mal-intencionados da população de estudo, para que seja possível aplicar o rigor da lei sobre eles.

A primeira interpretação dos modelos de conformidade pode induzir o leitor a pensar que a segmentação de contribuintes tem como único fim evitar punições. Essa interpretação decorre do fato de que o viés clássico e predominante das Administrações Tributárias tem sido punitivo e por isso, conclui-se incorretamente que a aplicação das teorias recentes sugere um abrandamento generalizado nas punições.

O autor do texto corrobora, entretanto, o entendimento de estudiosos como Taylor (2003), Braithwaite (2003b), Viol (2015), Kirchler (2007) e Weber, Fooker e Herrmann

¹⁸ Kirchler (2007) chama a postura “polícia e ladrão” e “comando e controle” respectivamente de “*cops and robbers*” e “*command and control*”. A postura “serviço e cliente”, por sua vez, é escrita no original em inglês como “*service and client*”.

¹⁹ Tradução livre do original “...to treat taxpayers as rational, self-interested, utility-maximising actors in all situations is to limit our ability to understand the processes involved in taxpaying behaviour”.

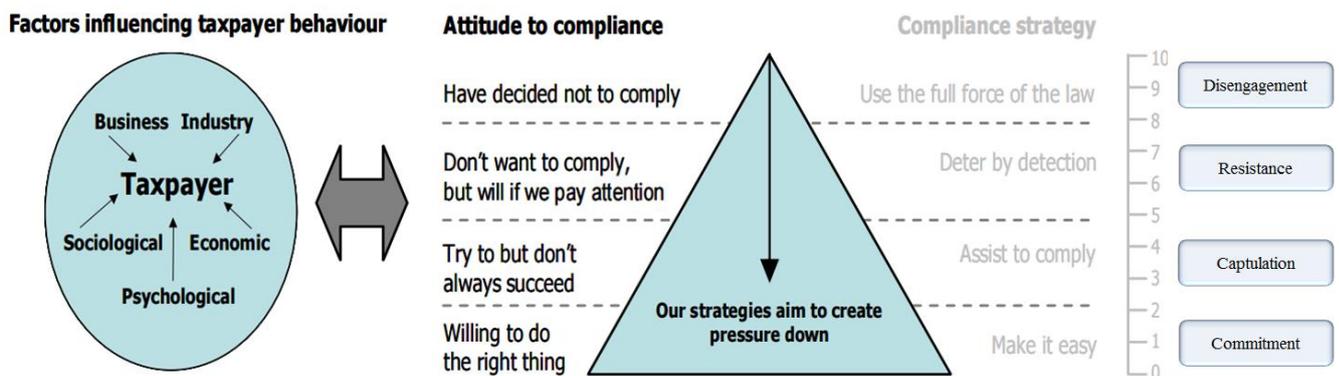
(2014), segundo o qual identificar os fatores que influenciam o comportamento tributário é necessário para que se segmente com mais precisão os indivíduos, de forma que seja possível facilitar o cumprimento daqueles bem-intencionados, e punir com o rigor da lei aqueles que insistem na má conduta.

Para que a segmentação seja corretamente efetuada, Braithwaite (2003b) argumenta que a análise da postura dos contribuintes deve adotar uma concepção multidimensional e não apenas considerar a oportunidade e o cálculo racional como fatores únicos e determinantes de comportamento tributário.

Essa argumentação é corroborada pelo modelo conceitual denominado Pirâmide de Conformidade (*Compliance Pyramid*). Desenvolvida pelo Fisco Australiano com base nas atitudes ou posturas de cumprimento propostas por Braithwaite (2003a), vistas na **Figura 1**, a Pirâmide de Conformidade segmenta os contribuintes de acordo com suas atitudes tributárias, para então apontar as ações que devem ser tomadas pelo Fisco em resposta a cada tipo de comportamento.

A **Figura 10** demonstra os quatro principais perfis de contribuintes e as estratégias apropriadas a cada perfil.

Figura 10: O modelo conceitual da pirâmide de conformidade.



Fonte: Adaptado de OECD (2004) e Braithwaite (2003a).

Das **atitudes** destacadas na pirâmide, é possível extrair dois grupos principais. O grupo de **posturas positivas** compreenderia os contribuintes dispostos a cumprir as obrigações (*Willing to do the right thing*) e os contribuintes que tentam cumpri-las, mas nem sempre conseguem (*Try to but don't always succeed*).

Já o grupo de **posturas negativas** compreende contribuintes que só cumprirão se forem controlados (*Don't want to comply, but will if we pay attention*) e contribuintes decididos a não cumprir suas obrigações (*Have decided not to comply*).

A pirâmide e os estudos que a embasam demonstram que, além da racionalidade econômica (*Economic*), fatores psicológicos (*Psychological*), sociológicos (*Sociological*) e de ambiente (*Business / Industry*) influenciam o comportamento tributário dos contribuintes e, por isso, encontrar características nos contribuintes e em seus comportamentos tributários que possam refletir tais fatores torna-se essencial para segmentá-los em perfis adequados.

Braithwaite (2003a) argumenta que essa segmentação é importante para que o tratamento adequado seja dispensado a cada grupo pelas Administrações Tributárias, pois seria pouco efetivo facilitar (*Make it easy*) ou ajudar no cumprimento (*Assist to comply*) contribuintes decididos a não cumprir. Da mesma maneira, a dissuasão por detecção (*Deter by detection*) e a aplicação de todo o rigor da lei (*Use de full force of the law*) poderiam trazer efeitos negativos se aplicadas a contribuintes dispostos a cumprir suas obrigações, mas que às vezes encontram dificuldades.

Ainda segunda a autora, o topo da pirâmide, sua parte mais estreita, é que comporta os sonegadores (*Disengagement e Resistance*), ao passo que a base, muito mais ampla, abriga os contribuintes que buscam cumprir suas obrigações (*Captulation e Commitment*). Em contraste com o modelo clássico, que atribui aos contribuintes em geral uma propensão à desonestidade, a Pirâmide de Conformidade se baseia em estudos que demonstraram que a maioria dos contribuintes é honesta e predisposta ao cumprimento de suas obrigações tributárias.

Kirchler (2007) aponta o clima de confiança entre contribuintes e Autoridades Tributárias como essencial para elevar a conformidade da base da pirâmide (*Captulation e Commitment*). O autor defende que esse clima deve ser criado por meio de informações claras, transparentes e simples, pois pouca compreensão ou completa falta de compreensão em relação aos tributos e obrigações tributárias²⁰ geram desconfiança e consequente descumprimento.

Foi o que identificaram Eriksen e Fallen (1996) em experimento no qual a educação tributária aprimorou a ética fiscal, aumentou a conformidade e reduziu a propensão à sonegação.

²⁰ Obrigações tributárias não são apenas aquelas de pagar tributos, mas também aquelas exigidas pelas Autoridades Tributárias para que a arrecadação de tais tributos seja possível, como registro nos cadastros, emissão de Notas Fiscais, entrega tempestiva de Declarações, dentre outras.

Em outra vertente, Snow e Warren (2005) demonstraram que o receio quanto a uma possível auditoria e o medo de punições aumentaram o nível de conformidade em contribuintes avessos ao risco (*Resistance*).

Já a parte mais alta da pirâmide (*Disengagement*) não responderia aos estímulos de cumprimento proporcionados pelas Administrações Tributárias, pois comporta indivíduos determinados a sonegar. Para estes, apenas a aplicação de todo o rigor da lei traria algum resultado (OECD, 2004).

Esses estudos reforçam a argumentação de que as diferentes ferramentas à disposição do Fisco – como coerção de um lado e colaboração de outro – podem trazer resultados positivos no aumento da conformidade tributária, desde que aplicadas aos perfis apropriados, e isso enfatiza a necessidade de segmentar adequadamente os contribuintes, como se propõe este estudo.

2.1.8 A conformidade nos impostos sobre consumo

Miguel (2016) aponta a importância dos impostos sobre consumo no Brasil, que respondem por praticamente 50% da arrecadação de todo o país. Dentre eles, o ICMS seria o principal, pois representa 20% da arrecadação tributária total.

Da arrecadação do ICMS, aproximadamente 30% advêm do Estado de São Paulo, como apontam Coelho, Bernardo e Maia (2017).

Nada obstante à sua importância, o referido tributo é repleto de complexidades, pois, segundo Roque Antônio Carrazza (2015), abrange diversas materialidades distintas, ou seja, diversos tributos em um único imposto.

Nesse sentido, o ICMS incide sobre (BRASIL, 1988):

- Operações de circulação de mercadoria.
- Prestação de serviço de transporte interestadual e intermunicipal.
- Prestação de serviço de comunicação.
- Importações.
- Operações relativas a energia elétrica.

O ICMS é um imposto sobre consumo e é similar, em alguns aspectos, ao Imposto sobre Valor Agregado – IVA. Zodrow (1999) aponta que as principais modalidades de

impostos sobre consumo, ou circulação de riqueza, são o Imposto sobre Vendas a Varejo – IVV (*Retail Sales Tax – RST*, em inglês) e o Imposto sobre Valor Agregado – IVA (*Value-Added Tax – VAT*, em inglês).

Considerando apenas a perspectiva econômica, ambos são entendidos como equivalentes. A diferença principal se situa na forma de arrecadação. O IVV é arrecadado apenas na venda ao consumidor final, ao passo que o IVA é coletado ao longo do processo de produção e distribuição (plurifasia ou multifasia). (ZODROW, 1999; MIGUEL, 2016).

O fato do IVV ser devido apenas na venda ao consumidor final torna mais difícil garantir sua arrecadação, pois uma eventual sonegação na ponta reduziria o montante total devido em toda a cadeia. Nesse aspecto, a potencial perda de arrecadação com o IVA é menor, pois o imposto é recolhido em diversas etapas ao longo da cadeia plurifásica (ZODROW, 1999).

Adicionalmente, como o IVA é calculado através de uma sistemática de débitos e créditos, compensando-se o que foi pago na etapa anterior, o próprio adquirente na etapa corrente teria interesse em se certificar de que o alienante emitiu todos os documentos fiscais necessários para que ele possa se creditar do montante pago. Isso torna o IVA uma espécie de imposto auto fiscalizável e também permite que as autoridades fiscais façam referências cruzadas entre os montantes recolhidos e deduzidos nas várias etapas da cadeia (ZODROW, 1999).

Por esses motivos, o IVA é considerado superior ao IVV e já foi adotado em aproximadamente 150 países, enquanto que o último é utilizado essencialmente nos Estados Unidos (CHARLET; OWENS, 2010).

Na maioria dos países existe apenas um IVA e a definição de suas alíquotas é feita pelo governo central (SWISTAK, 2015). No Brasil, no entanto, o legislador constituinte, numa tentativa de dividir a arrecadação, optou por segregar a tributação sobre o consumo em três impostos: o Imposto sobre Produtos Industrializados – IPI, de competência da União, o Imposto sobre Operações relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação – ICMS, de competência estadual, e o Imposto sobre Serviços de Qualquer Natureza – ISSQN (ou Imposto sobre Serviços – ISS), de competência municipal. (MIGUEL, 2016).

Essa pulverização de competências no Brasil em relação à tributação sobre o consumo talvez seja um dos motivos pelos quais o fardo de cumprimento brasileiro, demonstrado na **Figura 6**, seja tão maior que a média mundial, pois resulta na possibilidade de cada ente legislar sobre seus tributos e estabelecer obrigações quanto a eles.

De acordo com Kirchler (2007), o excesso de complexidade e pouca compreensão por parte dos contribuintes em relação aos tributos é um dos fatores que levam à *não conformidade*, tanto dolosa quanto não dolosa.

Apesar de não ser um IVA puro, o ICMS compartilha a principal característica daquele imposto, a não-cumulatividade. O objetivo dela é impedir a incidência em cascata do tributo sobre si mesmo (ZODROW, 1999) e sua importância se deve a uma característica marcante dos impostos sobre consumo: o fato de serem indiretos.

Um tributo indireto é aquele em que há transferência do ônus financeiro, que recai sobre uma pessoa diferente daquela que compõe a relação tributária. Nesse caso, o sujeito passivo do tributo repassa seu ônus financeiro a outrem (CARRAZA, 2015).

No caso do ICMS, o sujeito passivo é aquele que incorre em seu fato gerador. Na venda de uma mercadoria, por exemplo, o sujeito passivo do ICMS é o estabelecimento que efetuou a venda. O ônus tributário, no entanto, não recai sobre ele e sim sobre o consumidor final, que pagou pela mercadoria com o valor do imposto embutido no seu preço.

Alguns doutrinadores, como Coelho (2011), entendem que a transferência do ônus financeiro ocorre em todos os tributos, mas reconhecem que essa característica seja mais marcante nos tributos sobre consumo.

A incidência plurifásica nos IVA teria o potencial de elevar em demasia a carga tributária e afetar negativamente o mercado econômico, pois poderia aumentar de forma considerável o ônus financeiro repassado ao consumidor final.

A não-cumulatividade se apresenta, portanto, como solução à incidência plurifásica e é comumente efetivada através de um método conhecido como *invoice credit* ou *tax on tax*. Esse método faz com que o imposto incida sobre o montante total da operação em cada etapa, mas permite que o valor incidente na operação anterior, destacado na fatura (Nota Fiscal), seja abatido pelo adquirente contribuinte do imposto (EBRIL et al., 2001).

Entender a operacionalização da não-cumulatividade é importante para se entender como ocorrem as práticas de *não conformidade* (evasão e eventuais erros) no ICMS.

A **Figura II** apresenta um exemplo de operacionalização do método *invoice credit*. Para efeitos de simplificação, a característica chamada de cobrança “por dentro²¹” do ICMS foi desconsiderada no exemplo.

²¹ Para mais detalhes sobre essa característica, ver Carraza (2015).

Figura 11: Operacionalização da não-cumulatividade pelo método invoice credit.



Fonte: Elaborado pelo autor da pesquisa.

Na **Figura 11** é possível observar que se a não-cumulatividade fosse ignorada, o valor total do imposto a ser pago seria de R\$ 45,00. Considerando o valor final de venda de R\$ 200,00, a alíquota teria mais que dobrado em relação à de 10% adotada no exemplo.

Pelo exemplo de operacionalização da não-cumulatividade do ICMS demonstrado na **Figura 11**, é possível observar duas formas de redução do imposto, possíveis tanto em casos de evasão quanto em casos de *não conformidade* involuntária. Essas formas seriam operacionalizadas através dos conceitos de omissão e comissão discutidos na seção **Erro! Fonte de referência não encontrada.**

A omissão no exemplo acima poderia ocorrer em qualquer etapa da cadeia através do ato de esconder do Fisco a ocorrência da operação. Se na saída do mercado, por exemplo, o documento fiscal de venda não fosse emitido, haveria *não conformidade*. Essa omissão teria efeitos diretos no aumento da brecha tributária no âmbito do ICMS, em virtude da diminuição da arrecadação devida.

A comissão, por sua vez, poderia ser operacionalizada com a utilização de um documento fiscal frio, que não tivesse lastro material. Suponha-se que uma operação de venda do distribuidor para o mercado não tenha ocorrido, mas mesmo assim este último se utilize de

uma nota fria emitida por uma empresa fictícia com sócios laranjas ou testas de ferro²², e se aproprie dos créditos do imposto supostamente pago.

Essas empresas fictícias teriam que recolher o tributo devido na suposta operação de venda, mas como o intuito, nesse caso, foi exatamente a evasão fiscal, não o fazem. A cobrança pelo Fisco, por sua vez, também não é facilitada, visto que por vezes os sócios sequer existem, ou se existem, talvez nem tenham conhecimento de que uma empresa foi aberta em seu nome (laranjas), ou se tem tal conhecimento (testas de ferro), nem sempre é possível alcançar seu patrimônio pessoal na cobrança (FERRAGUT, 2015).

Os exemplos anteriores apontaram práticas dolosas com intuito de evadir tributos. Os atos de omissão ou comissão, entretanto, podem ser cometidos também em função de erros, como discutido anteriormente e por isso, a Pirâmide de Conformidade sugere que cada tipo de desvio seja abordado de forma diferente pelas Administrações Tributárias.

A omissão involuntária de uma operação de venda, causada por erro, deve ser abordada pelas autoridades fiscais por meios de atitudes que tornem o cumprimento mais fácil, como simplificação das obrigações tributárias, campanhas educativas, dentre outras. A mesma omissão por dolo, no entanto, deve ser abordada com ações punitivas e repressivas que desencorajem o comportamento (BRAITHWAITE, 2003a).

2.1.9 O aumento da conformidade no âmbito do ICMS

A conformidade tributária em relação ao ICMS perpassa as seguintes etapas (VIOL, 2015; SIQUEIRA; RAMOS, 2005):

- Inscrição no cadastro do respectivo ente tributante.
- Entrega das declarações.
- Pagamento das quantias devidas.

A inscrição no cadastro é um dos pré-requisitos para o exercício de atividades econômicas sujeitas ao ICMS. No Estado de São Paulo, esse pré-requisito foi instituído pela Lei nr. 6.374, de 01 de março de 1993 (SÃO PAULO, 1993).

²² Para efeitos deste texto, o termo laranja designa pessoas que, inadvertidamente ou sem conhecimento, têm seus nomes utilizados de maneira fraudulenta, enquanto que o termo testa de ferro descreve indivíduos que agem em conluio com os agentes fraudadores (DOS SANTOS et al., 2017).

As declarações, por sua vez, constituem o cálculo do tributo devido, que é lançado por homologação. Nesse tipo de lançamento, cabe ao contribuinte apurar o montante devido do tributo e antecipar o pagamento que será, posteriormente, homologado ou não pela autoridade administrativa (BRASIL, 1966).

A falta de pagamento das quantias devidas configura a inadimplência para com o poder público e sujeita o devedor à inscrição em dívida ativa, conforme prescreve a Lei 6.830 de 22 de setembro de 1980 (Lei de Execução fiscal).

Viol (2015) aponta que, em relação às etapas acima, o dano causado pela *não conformidade* vai aumentando ao longo da sequência listada. Contribuintes com débitos inscritos em dívida ativa, última etapa, já incorreram em comportamento que lesou o Estado, pois mesmo que o crédito tributário seja recuperado, a máquina estatal foi movida para sua recuperação.

Na etapa imediatamente anterior, embora possa haver a detecção das declarações entregues com valores incorretos e os montantes corretos do tributo serem lançados de ofício pelo Fisco, também foi necessário dispêndio governamental em análises e auditorias.

Viol (2015) defende, portanto, que o mais eficiente seria impedir um comportamento lesivo já no cadastro, ou seja, impedir que contribuintes com atitudes fraudadoras tenham o registro cadastral de suas empresas deferido.

O impedimento da inscrição no cadastro impossibilitaria empresas fictícias de gerarem créditos frios do ICMS. O mesmo impedimento, no entanto, se feito com contribuintes bem-intencionados, seria um desserviço tanto ao contribuinte quanto à sociedade em geral, pois a atividade econômica é fundamental para geração de riqueza, conforme argumenta Mankiw (2014).

Nessa esteira, Braithwaite (2003a) reforça a importância de separar contribuintes com postura de conformidade, que eventualmente cometem erros, daqueles com postura não conforme, dispostos a sonegar. Nessa linha, a Pirâmide de Conformidade apresenta como proposta para tal separação a utilização não apenas de fatores econômico-financeiros, mas também de fatores sociológicos, psicológicos e de ambiente (OECD, 2004).

Kirchler (2007) aponta que nem sempre é possível mapear tais fatores diretamente. Como mapear, por exemplo, valores éticos e padrões morais dos contribuintes, ou seu grau de aversão ao risco? Como solução, o autor argumenta que, embora pareça não ser possível mapear tais fatores diretamente, talvez seja possível identificar variáveis que influenciem ou sejam influenciadas por eles.

O autor cita diversos exemplos de variáveis que podem refletir fatores que influenciam no comportamento tributário. Dentre eles, destaca-se os seguintes.

A **experiência anterior com o Fisco**, estabelecida em auditorias ou autuações sofridas pelo contribuinte, pode ser uma variável que representa **Fatores Psicológicos**. Isso ocorre por que a experiência de uma auditoria prévia e suas consequências podem gerar influências nas percepções de poder de detecção e punibilidade do Fisco e afetar o grau de aversão ao risco do contribuinte.

De forma semelhante, o comportamento de atores que se relacionam com o contribuinte, a exemplo de **contabilistas e sócios**, podem ser variáveis que refletem **Fatores Sociológicos**, pois posturas éticas e morais de outros indivíduos podem se tornar referência para o contribuinte.

Fatores Econômico-financeiros, por sua vez, podem ser refletidos por variáveis que identifiquem padrões em **declarações entregues** ao Fisco ou em **documentos fiscais emitidos**.

Adicionalmente, características demográficas do contribuinte, como **localização e ramo de atividade** podem estar sujeitos a influências de **Ambiente**.

De acordo com a OECD (2004), a adoção do modelo conceitual proposto pela Pirâmide de Conformidade – que considera tanto fatores econômicos como não econômicos na determinação do perfil de conformidade dos contribuintes – é a forma mais efetiva de aumentar a conformidade e diminuir as causas Brecha Administrativa, resumidas na **Figura 12**.

Figura 12: Causas da brecha administrativa



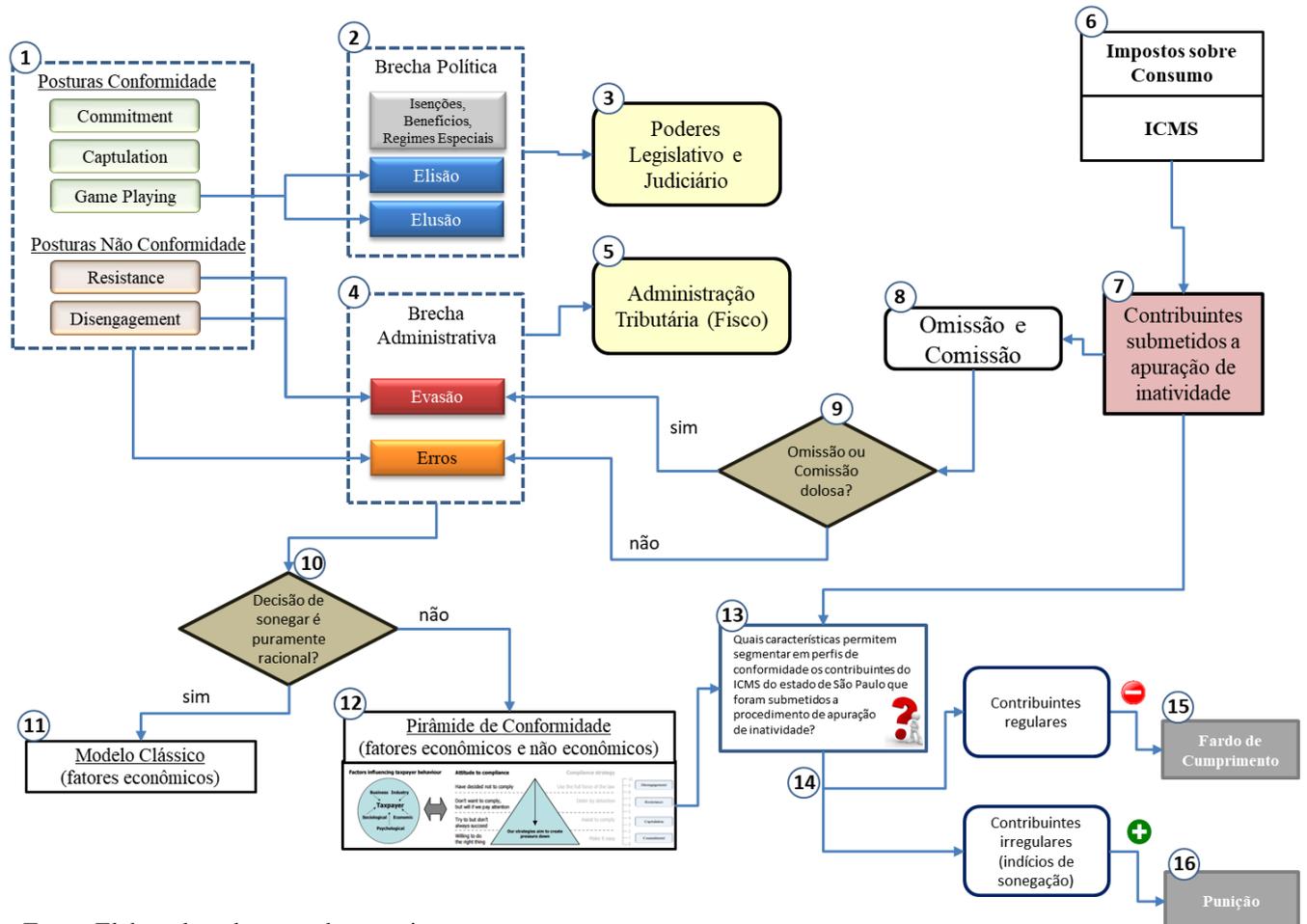
Fonte: Elaborado pelo autor da pesquisa.

Pelos motivos expostos, acredita-se que o modelo conceitual da Pirâmide de Conformidade é o mais adequado para esse estudo, pois aponta fatores além dos econômicos, que serão utilizados para responder à questão de pesquisa e encontrar características que permitam segmentar os contribuintes Cassados por Inatividade Presumida.

2.1.10 Síntese dos elementos críticos abordados

A **Figura 13** resume os principais pontos discutidos e o **Quadro 2** lista os principais autores pesquisados.

Figura 13: Elementos críticos e aporte teórico.



Fonte: Elaborado pelo autor da pesquisa.

Pela **Figura 13** é possível observar que (1) **contribuintes** podem ser **classificados** em perfis, ou **Posturas de Conformidade e de Não Conformidade**. A **conformidade** ocorre quando há o correto **cumprimento das obrigações tributárias** e é retratada pelas **posturas de Commitment e Captulation**. Posturas que levam à busca de **brechas legais** para diminuição da carga tributária são chamadas de **Game Playing** e são **materializadas através da Elisão e Elusão** de tributos. Por não ferirem a escrita da lei, tais posturas também são **consideradas posturas de conformidade a despeito do dano** que podem causar aos **cofres públicos**.

Juntamente com a diminuição de arrecadação decorrente da Elisão e da Elusão, as **concessões** recebidas do poder público (**Isenções, Benefícios e Regimes Especiais**) compõem a chamada **Brecha Política** (2). Por agregar práticas que não ferem a escrita da lei, seu combate é **responsabilidade** precípua de instâncias estatais como os **Poderes Legislativo e Judiciário** (3).

A **não conformidade** é decorrente de posturas de *Resistance e Disengagement* quando ocorre de forma **voluntária** e **dolosa** em atos de **evasão**, e de todas as outras posturas, quando ocorre de forma **involuntária** por **erros** cometidos. Por ferirem a escrita da lei, os **erros** compõem, juntamente com a **evasão**, a **Brecha Administrativa** (4). O combate a esta brecha é **responsabilidade** primária da Administração Tributária (5).

O ICMS é um imposto sobre **consumo** (6) e os seus **contribuintes no Estado de São Paulo** que foram submetidos a procedimento de **apuração de inatividade** (7) podem ter apresentado, durante seu tempo de atividade, atitudes de **omissão ou comissão** (8). Tais atitudes, se cometidas de forma **dolosa** (9), implicam em **evasão fiscal**. Se, ao contrário, forem cometidas de forma **involuntária**, configuram-se como **erros**.

A busca da diminuição da **Brecha Administrativa**, que é **composta** por **evasão e erros**, pode ser baseada em dois **modelos** principais, que se distinguem pelos fatores que entendem como influenciadores na decisão de sonegar (10). O **Modelo Clássico** (11) argumenta que os fatores que influenciam a decisão de sonegação são puramente racionais e, portanto, **todos os contribuintes são potenciais sonegadores** e devem ser tratados com o **rigor da lei**, o que torna sua **segmentação menos relevante**. O **Modelo da Pirâmide de Conformidade** (12) considera que tanto **fatores econômico-financeiros** quanto **não econômico-financeiros** podem influenciar a decisão de sonegar e, por isso, **contribuintes reagirão de forma diferente frente a uma oportunidade de sonegação**, devendo, portanto, serem corretamente **segmentados**.

Este estudo defende que é **necessário separar** os Contribuintes do ICMS que foram submetidos a procedimento de apuração por inatividade entre aqueles **que apresentaram indícios de evasão e aqueles simplesmente deixaram de operar**. Somente assim seria **possível** aplicar o correto tratamento adequado a cada um.

Para que a segmentação seja feita de forma mais precisa, considera-se que o **modelo conceitual mais adequado seria a Pirâmide de Conformidade**, pois considera, além dos **fatores Econômicos**, **fatores Psicológicos**, **Sociológicos** e **de Ambiente** como relevantes na determinação do comportamento tributário. A busca de **características que permitam**

mapear tais fatores é a chave para que seja possível segregar os contribuintes e é o que se propõe nesta dissertação (13).

Encontradas tais características, seria possível **separar os contribuintes nos perfis adequados** (14) e aplicar o tratamento correto a cada um deles (15 e 16).

Quadro 2: Síntese do aporte teórico pesquisado.

Elementos críticos	Aporte teórico relacionado aos elementos críticos
Posturas dos contribuintes em relação aos tributos	McBarnet (2001) aponta como posturas de conformidade: <i>Commiment, Captulation e Creative Compliance (Game Playing)</i> . Braithwaite (2003b) estende a lista incluindo as posturas de <i>não conformidade: Resistance e Disengagement</i> .
<i>Não Conformidade Involuntária</i>	Viol (2015) aponta que nem toda <i>não conformidade</i> é dolosa. Há casos de <i>não conformidade</i> que ocorrem com contribuintes que possuem uma postura de conformidade, mas que podem cometer erros.
Conceitos de Evasão, Elisão e Elusão	Cobham (2005), OECD (2017b) e Weber; Fookan e Herrmann (2014) definem Evasão como prática ilegal e Elisão como prática legal, embora não desejada pelo legislador. King e Shefrin (2002) e Sandmo (2003) apontam que Evasão é sempre ilegal e revestida de dolo, de atitudes fraudulentas. Kirchler, Maciejovsky e Schneider (2003) descrevem Elusão como uma espécie de Elisão, mas mais agressiva e moralmente reprovada. Nishioka (2010) conceitua Elusão como Elisão Abusiva. Viol (2015) e Siqueira e Ramos (2005) utilizam o termo Elisão de forma ampla, indicando tanto elisão quanto elusão. Elisão e Elusão são operacionalizadas com bases em brechas encontradas na lei.
Atos de Omissão e Comissão	Kirchler (2007) aponta que as práticas de evasão são operacionalizadas através de atos de omissão e de comissão.
Fardo de Cumprimento Tributário	Kirchler (2007) define fardo tributário como o excesso de regulações e formalismos e o vê como uma das causas de ocorrência de <i>não conformidade</i> , tanto voluntária quanto involuntária. O World Bank Group e a Price Waterhouse Coopers (2017) publicam relatório regular que mede o tamanho do fardo em diversos países. O número no Brasil é de 2.038 horas anuais para o cumprimento tributário.
Brecha Tributária	Thackray (2012) utiliza a definição ampla de brecha tributária, que seria a diferença entre a arrecadação efetiva e a legalmente intencionada. Para Viol (2015), as causas da brecha são elisão, elusão, evasão, erros de cumprimento e também isenções, regimes especiais e benefícios fiscais. Alm (2012) define a brecha tributária de forma estrita, considerando como seus componentes apenas as reduções de impostos em desacordo com a lei escrita (evasão e erros). Viol (2015) aponta que a conceituação estrita é chamada de brecha administrativa. As demais reduções, decorrentes de situações que não afrontam o texto da lei são chamadas de brecha política. Alm, Martinez-Vazquez e Schneider (2004), Schneider (2005), Siqueira e Ramos (2005) apontam que uma das maneiras de medir a brecha é através da economia subterrânea, que foi estimada em 39,8% do PIB brasileiro no ano 2000. O HMRC (2016) estimou a brecha decorrente de impostos sobre consumo do Reino Unido em 10,3% e a European Commission (2016) a estimou em 10,4% na União Europeia.

Combate à Elisão e Elusão	Viol (2015) aponta que a responsabilidade no combate à elisão e elusão é de atores como os Poderes Legislativo e Judiciário.
	John Braithwaite (2005) propõe reforma na lei tributária com utilização ampla de princípios para impedir tais práticas.
Combate à Evasão e Estímulo à Diminuição de Erros	Para Viol (2015), o combate à brecha administrativa é responsabilidade do Fisco.
	Braithwaite (2003b), Viol (2005), Kirchler (2007) e Weber, Fooken e Herrmann (2014) defendem a segmentação em perfis para que a resposta adequada seja dada a cada um.
Decisão de Sonegar	Modelo Clássico: proposto por Michael G. Allingham e Agnar Sandmo (1972), assume que os fatores de influência na decisão de sonegar são econômicos. Contribuintes são amorais e tendentes, em sua maioria, ao comportamento desonesto.
	Long e Swingen (1991) e Braithwaite (2003b) apontam estudos que indicam que a maioria dos contribuintes é, ao contrário, honesta.
	Andreoni, Erard e Feinstein (1998) consideram o modelo clássico uma descrição pobre do mundo real e de pouca efetividade no longo prazo.
	Simpson (2002) aponta a fraca capacidade do modelo clássico de explicar os altos níveis de conformidade em muitos países.
	Webley, Cole e Eidjar (2001) defendem que fatores econômicos utilizados a partir de uma oportunidade percebida exerceriam apenas parte da influência na decisão de sonegar, que seria baseada também em fatores não econômicos.
	Lawler (1998), Alm e Torgler (2006) e Braithwaite (2003b) elencam fatores não econômicos que influenciariam a decisão de cumprimento.
	Braithwaite (2003a; 2003b) e OECD (2004) propõem a Pirâmide de Conformidade como alternativa ao modelo clássico.
	Pirâmide de Conformidade: OECD (2004) aponta que as atitudes dos contribuintes são influenciadas por fatores econômicos e não econômicos e podem ser agrupadas em diferentes perfis, com uma sugestão de resposta do Fisco a cada perfil.
Conformidade nos Impostos sobre Consumo	Miguel (2016) aponta que o ICMS é estadual e é o principal imposto sobre consumo, respondendo por 20% da arrecadação tributária total do Brasil.
	ICMS é espécie de Imposto sobre Valor Agregado - IVA, que de acordo com Zodrow (1999), Swistak (2015) e Charlet e Owens (2010), é superior ao Imposto sobre Vendas no Varejo - IVV.
	Ebril et al. (2001) apontam que a não-cumulatividade é fundamental nos IVA e que esta é comumente implementada pelo método invoice credit, utilizado no ICMS.
	A não conformidade no ICMS é operacionalizada de acordo com os atos de omissão e comissão definidos por Kirchler (2007).
Aumento da Conformidade no ICMS	Siqueira e Ramos (2005) e Viol (2015) apontam como etapas de conformidade no ICMS a inscrição, a entrega de declarações e o pagamento do imposto. O dano causado pela não conformidade é crescente a cada etapa.
	Ferragut (2015) aponta a dificuldade de recuperar crédito tributário quando atores são empresas fictícias, que conseguiram efetivar sua inscrição no cadastro para poderem operar.
	Viol (2015) defende que é mais eficiente impedir um comportamento lesivo já no cadastro. Contribuintes bem intencionados, no entanto, deveriam, segundo Kirchler (2007), encontrar postura do Fisco que contribua para diminuir seu fardo tributário.
	Braithwaite (2003a e 2003b) reforça a necessidade do Fisco ser capaz de separar um comportamento do outro para que o tratamento adequado possa ser aplicado a cada um.

Braithwaite (2003a e 2003b), OECD (2004) e Kirchler (2007) argumentam que a Pirâmide de Conformidade, além de oferecer o ferramental mais adequado para essa separação, sugere também as respostas apropriadas a cada perfil.

Fonte: Elaborado pelo autor da pesquisa.

2.2 Pesquisa empírica

No presente capítulo será discutida a metodologia utilizada, serão exploradas e caracterizadas as variáveis de estudo, a seleção de variáveis conduzida, o modelo estatístico criado e validado.

2.2.1 Metodologia

Este capítulo apresenta os principais procedimentos que deverão ser observados na condução da pesquisa ora em projeto, bem como as classificações pertinentes ao tipo de estudo a ser conduzido.

2.2.1.1 Classificação da pesquisa

Para identificar as características que permitem segmentar, em perfis de conformidade, os contribuintes do ICMS no Regime Periódico de Apuração – RPA que foram submetidas a procedimento de apuração de inatividade, este estudo foi realizado sob a **abordagem quantitativa**. Essa classificação se baseia na taxonomia proposta por Gerhardt e Silveira (2009) e decorre do fato de que, apesar do fenômeno pesquisado estar relacionado ao comportamento humano, a pesquisa foi baseada em dados numéricos e classificatórios transformados em variáveis das quais se criou um modelo matemático-estatístico.

Em relação ao tipo, trata-se de uma **pesquisa exploratória e descritiva**. Segundo Gil (1991), a pesquisa exploratória proporciona maior familiaridade com o problema, para torná-lo explícito, formular hipóteses, aprimorando ideias e intuições. Quanto ao aspecto descritivo, o autor afirma que vai além da simples identificação da existência de relações entre variáveis, buscando também determinar a natureza dessa relação.

Quanto ao desenho, a pesquisa configura-se como **estudo de caso instrumental**, em que os resultados podem facilitar a compreensão de algo mais amplo, fornecendo *insights* sobre um assunto, ou contestando uma concepção generalizada (STAKE, 1995). De acordo

com Yin (2003, p.19), estudos de casos se apresentam como a estratégia de pesquisa adequada quando o foco se encontra em “[...] fenômenos contemporâneos inseridos em algum contexto da vida real”.

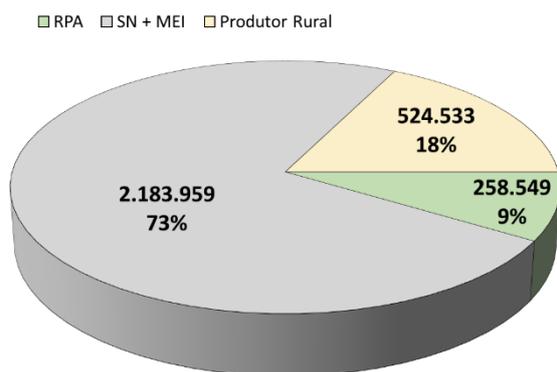
2.2.1.2 População

A **população** da presente pesquisa se constitui de empresas que obedecem a quatro critérios cumulativamente:

1. Sejam contribuintes do ICMS no Estado de São Paulo;
2. Estejam enquadradas no Regime Periódico de Apuração – RPA;
3. Tenham sido submetidas a procedimento de apuração de inatividade e
4. Tenham tido a apuração de inatividade efetuada entre os anos de 2015 e 2017.

A escolha do **Critério 1 – Contribuintes do ICMS do Estado de São Paulo** se deve à proximidade do autor da pesquisa com tais contribuintes, pois este é Agente Fiscal de Rendas da Secretaria da Fazenda do Estado de São Paulo. Os contribuintes enquadrados neste critério abrangem os grupos detalhados no *Gráfico 1*.

Gráfico 1: Média anual de contribuintes do ICMS no Estado de São Paulo – 2017.



Fonte: Adaptado de Sefaz/SP, 2017.

Pelo *Gráfico 1* é possível observar que o universo de empresas contribuintes do ICMS no Estado de São Paulo abrange os **Microempreendedores Individuais – MEI**, os contribuintes do **Simple Nacional – SN**, os **Produtores Rurais** e os contribuintes inscritos no **Regime Periódico de Apuração – RPA**.

Os **MEI** são empreendedores cuja receita bruta não excede o valor anual de R\$ 81.000,00 (oitenta e um mil reais).

Os contribuintes que optaram pelo regime diferenciado de tributação do **SN** são normalmente Microempresas com receita bruta anual igual ou inferior a R\$ 360.000,00 (trezentos e sessenta mil reais) e Empresas de Pequeno Porte cuja receita bruta anual seja superior a R\$ 360.000,00 (trezentos e sessenta mil reais), mas igual ou inferior a R\$ 4.800.000,00 (quatro milhões e oitocentos mil reais). (BRASIL, 2006).

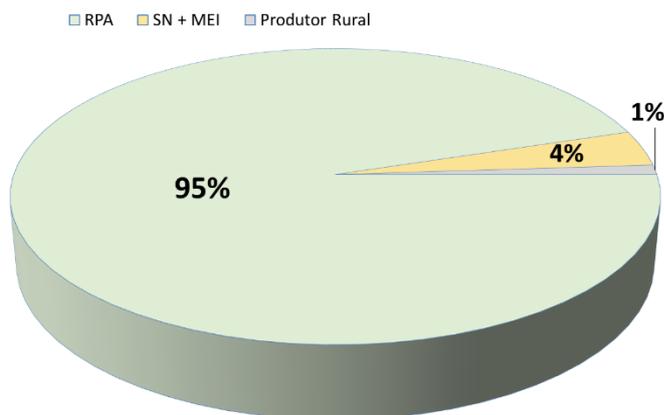
Os **Produtores Rurais** são as pessoas naturais que exercem atividade empresarial rural e que cumprem requisitos como: não serem inscritos no Registro Público de Empresas Mercantis e realizarem profissionalmente atividade agropecuária de extração e exploração vegetal ou animal, de pesca ou de armador de pesca. (SÃO PAULO, 1998).

Dos grupos detalhados no *Gráfico 1*, somente os **enquadrados no Regime Periódico de Apuração – RPA (Critério 2)** foram abrangidos pela pesquisa. Trata-se de pessoas jurídicas (não naturais), cuja receita bruta excede os limites de enquadramento do SN (e por consequência do MEI) ou que, apesar de possuírem receita bruta que os permita serem enquadrados nos respectivos regimes, não solicitaram o enquadramento ou não atenderam aos demais requisitos para tanto. Dentre esses requisitos adicionais, cita-se a ausência de débitos tributários.

O enquadramento nesse regime implica que “o valor do imposto a recolher corresponde à diferença, em cada período de apuração, entre o imposto devido sobre as operações ou prestações tributadas e o cobrado relativamente às anteriores” (SÃO PAULO, 1993). Em outras palavras, significa que o cálculo do imposto é efetuado através da sistemática de débitos e créditos descrita na seção 2.1.8 *A conformidade nos impostos sobre consumo*.

A escolha do **Critério 2** se deveu ao fato de que, embora representem apenas 9% do total de estabelecimentos empresariais no *Gráfico 1*, os contribuintes do RPA respondem por 95% da arrecadação, como demonstrado no *Gráfico 2*.

Gráfico 2: Participação na arrecadação por grupo de contribuinte em 2017.



Fonte: Adaptado de resposta à consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.

Essa alta participação na arrecadação representa grande dano potencial no caso de sonegação, como discutido na seção 2.7.1.1 *A decisão de sonegar: modelo clássico*, o que justifica a inclusão desse critério na presente pesquisa.

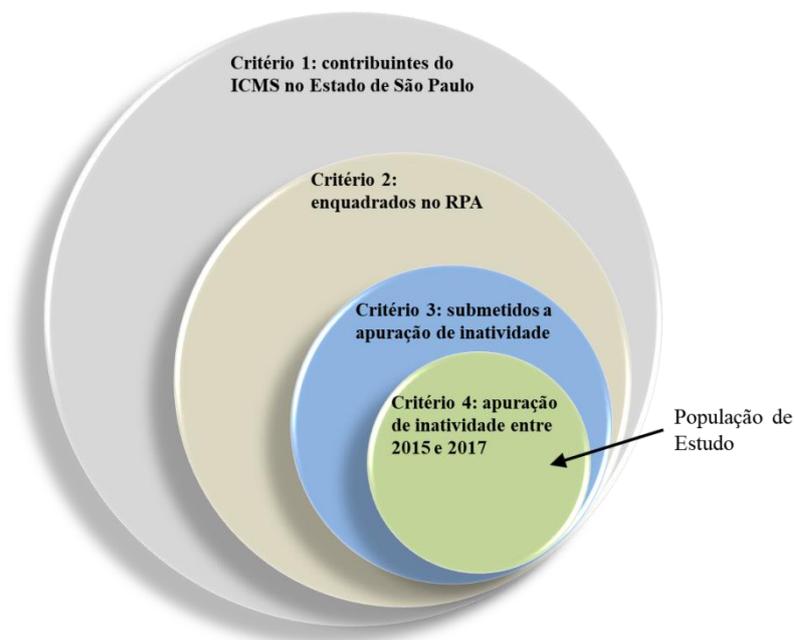
De acordo com o **Critério 3**, nem todos os RPA foram objeto da pesquisa, mas **somente aqueles que foram submetidas a procedimento de apuração de inatividade**. Esse critério foi utilizado pelo fato de não terem sido encontrados outros estudos de determinação do perfil de cumprimento tributário entre esses contribuintes, como se discutiu na seção *1 DESCRIÇÃO DO CASO DE ESTUDO*.

Por fim, o **Critério 4 - apuração de inatividade efetuada entre os anos de 2015 e 2017** foi adotado para favorecer a pesquisa no tocante à precisão dos resultados, pois permitiu selecionar as diligências mais atuais²³ e que refletem práticas sonegatórias e ferramentas de detecção mais recentes.

A **Figura 14** demonstra a população de estudo aplicando-se os quatro critérios de seleção.

²³ O estudo foi realizado no ano de 2018, com dados de 2015 a 2017.

Figura 14: População de estudo da pesquisa.



Fonte: Elaborado pelo autor da pesquisa.

2.2.1.3 Amostra

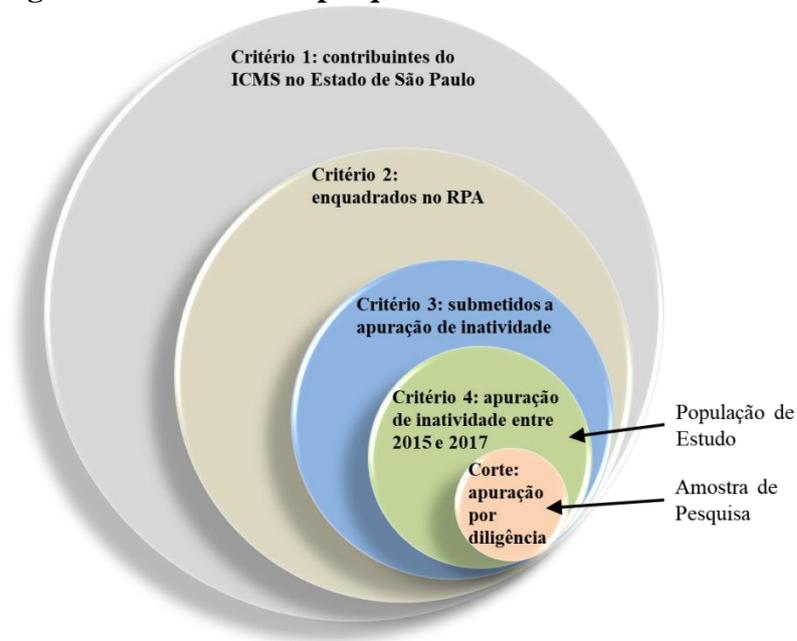
Para seleção da amostra de pesquisa, foi feito um **corte procedimental**. Esse corte se deu pelo fato de que a apuração da inatividade pode se dar por meio de dois procedimentos: **Diligência Fiscal** e **Cassação por Inatividade Presumida**, conforme descrito na seção *1 Descrição do caso de estudo*.

Desses procedimentos, apenas a **Diligência Fiscal** é capaz de indicar se houve ou não indícios de sonegação durante a apuração de inatividade e por isso, foram selecionados os contribuintes submetidos a ela no período indicado para que fosse possível buscar quais características melhor refletem o resultado da diligência e que possam ser utilizadas para prevê-la.

A identificação das características que determinam o perfil sonegador nos contribuintes da amostra teve o intuito de tornar possível a **criação de um modelo estatístico** que possa ser aplicado nos demais indivíduos da população, que não contam com o resultado da diligência, e **ampliar, assim, o poder de combate à evasão fiscal do Processo de Cassação por Inatividade Presumida**.

A **Figura 15** expande a **Figura 14** para apresentar, de forma esquemática, a amostra de pesquisa.

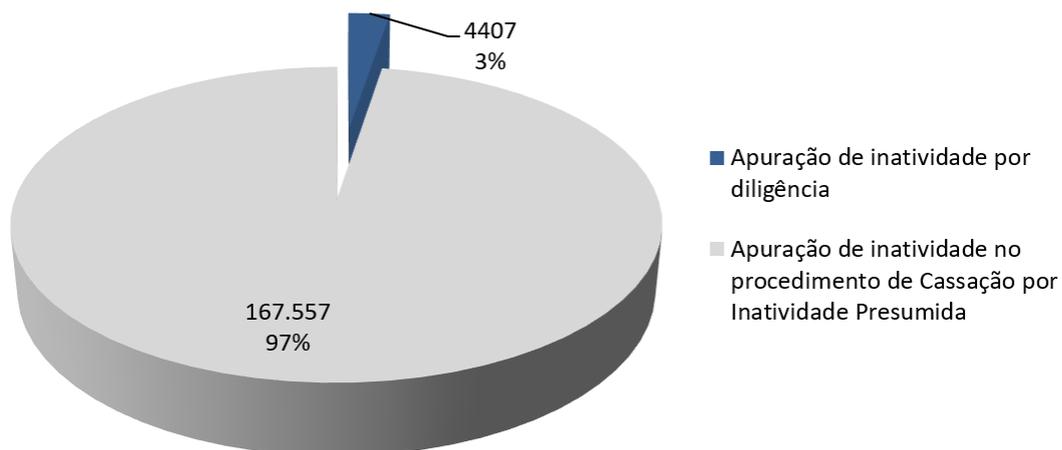
Figura 15: Amostra de pesquisa.



Fonte: Elaborado pelo autor da pesquisa.

O *Gráfico 3* mostra o total de indivíduos da amostra frente ao total da população.

Gráfico 3: Quantitativos da amostra e da população.



Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.

O **total da amostra** é de **4.407 indivíduos**. Além destes, a população possui ainda 167.557 indivíduos (*Tabela 1*), que foram inabilitados no procedimento de Cassação por

Inatividade Presumida. Como é possível observar no **Gráfico 3**, a amostra representa **3% da população**.

Essa diferença na distribuição da população entre os procedimentos de apuração de inatividade demonstra o poder de processamento de grandes volumes de dados da Cassação por Inatividade Presumida e reforça o potencial de contribuição da pesquisa, visto que esta se propõe a criar um modelo estatístico que poderia ser aplicado nos 97% da população que não sofreram diligência fiscal, para que também possam ser corretamente segmentados.

2.2.1.4 Coleta dos dados

Os dados foram obtidos dos bancos de dados da Secretaria da Fazenda do Estado de São Paulo. Foram coletadas informações com potencial de refletir os fatores apontados na teoria como determinantes do comportamento tributário.

De acordo com o modelo conceitual da **Pirâmide de Conformidade**, discutido na seção 2.7.1.2 *A decisão de sonegar: modelos recentes*, os **fatores determinantes** são os seguintes:

- **Econômico-financeiros.**
- **Psicológicos.**
- **Sociológicos.**
- **De ambiente.**

Conforme discutido na seção 2.1.9 *O aumento da conformidade no âmbito do ICMS*, Kirchler (2007) ressalta que nem sempre é possível mapear tais fatores diretamente e por isso, argumenta que devem ser buscadas características que os reflitam. O autor sugere, por exemplo, que fatores econômico-financeiros estão intimamente ligados ao possível benefício da sonegação e probabilidade e peso da punição, da mesma forma que fatores psicológicos refletem o grau de aversão ao risco.

Nesse sentido, espera-se que a análise das movimentações fiscais forneça indícios da percepção do contribuinte em relação ao benefício econômico da sonegação, da mesma forma que a experiência anterior com o Fisco, medida em função de auditorias e autuações sofridas, possa influenciar o grau de aversão ao risco do contribuinte.

Seguindo a mesma linha, Kirchler (2007) aponta que características demográficas podem refletir fatores de ambiente. Dentre as características demográficas estão o local e a área de atuação do contribuinte, assim como o porte empresarial.

Por fim, Kirchler (2007) sugere que fatores sociológicos podem ser refletidos pela rede de interação com indivíduos. Nesse sentido, espera-se que sócios e contadores relacionados à empresa e as relações deles com outras empresas possam dar indícios do comportamento tributário.

Nesse sentido, os dados coletados foram organizados segundo os fatores listados na **Tabela 4**. A partir dos dados, foram construídas as variáveis do estudo, listadas na seção a seguir.

Tabela 4: Dados coletados e fatores de referência.

Dados coletados	Fatores Relacionados
Declarações de imposto	Econômico-financeiros
Documentos fiscais	Econômico-financeiros
Experiência anterior com o Fisco	Psicológicos
Características demográficas	Ambiente
Relação com outros indivíduos	Sociológicos

Fonte: Elaborado pelo autor da pesquisa.

2.2.1.5 Variáveis do estudo

A partir dos fatores elencados na teoria e das características que refletem esses fatores, foram obtidas as **variáveis** descritas nas tabelas a seguir.

A **Tabela 5** descreve o **perfil de cumprimento tributário** do contribuinte, ou seja, o grupo de referência em que foi enquadrado, de acordo com o resultado da diligência.

A **Tabela 6** descreve os **fatores econômico-financeiros** através de variáveis que refletem dados fiscais do contribuinte.

A **Tabela 7** descreve os **fatores psicológicos** ao listar variáveis que refletem a experiência anterior com o Fisco.

A **Tabela 8** descreve os **fatores de ambiente** visto que compreende características demográficas dos contribuintes, como área de atuação e porte empresarial.

Por fim, a **Tabela 9** descreve os **fatores sociológicos** ao abranger variáveis que refletem a relação do contribuinte com outros indivíduos.

Tabela 5: Perfil de cumprimento tributário

Variável	Descrição	Formato
	Situação após a apuração de inatividade por diligência fiscal.	
dilg_irre	<p>0 - Empresa em situação regular; 1 - Empresa em situação irregular.</p> <p>As definições de Regular e Irregular foram determinadas no âmbito da pesquisa e possuem o objetivo de separar ilícitos tributários da simples inatividade. O evento de interesse, quando a variável assume o valor 1, indica a ocorrência de irregularidade tributária além da simples inatividade do contribuinte.</p>	Catégorico

Fonte: Elaborado pelo autor da pesquisa.

Tabela 6: Dados de declaração de imposto e de documentos fiscais – Fatores econômico-financeiros

Variável	Descrição	Detalhamento	Formato
decl_entg	Quantidade de declarações entregues	Quantidade total de declarações apresentadas nos 12 meses anteriores à apuração de inatividade.	Inteiro
decl_movi	Quantidade de declarações com movimento	Do total de declarações apresentadas nos 12 meses anteriores à apuração de inatividade, indica aquelas em que houve operações de entrada e saída de mercadorias.	Inteiro
decl_ind1	Índice de declarações	Índice que relaciona as variáveis decl_entg e decl_movi. A fórmula utilizada no cálculo deste índice foi omitida do texto para preservar o sigilo de algumas informações.	Índice Decimal
decl_cred	Valor total dos créditos nas declarações	Total de créditos de ICMS lançados nas declarações entregues nos 12 meses anteriores à apuração de inatividade (COMPRAS e DEVOLUÇÕES).	Decimal
decl_debi	Valor total dos débitos nas declarações	Total de débitos de ICMS lançados nas declarações entregues nos 12 meses anteriores à apuração de inatividade (VENDAS).	Decimal
decl_ind2	Índice de crédito e débito das declarações	Índice que relaciona as variáveis decl_cred e decl_debi. A fórmula utilizada no cálculo deste índice foi omitida do texto para preservar o sigilo de algumas informações.	Índice Decimal
docf_cred	Valor total dos créditos nos documentos fiscais	Valor total das operações que geram crédito de ICMS lançadas em documentos fiscais nos 12 meses anteriores à apuração de inatividade (COMPRAS e DEVOLUÇÕES).	Decimal
docf_debi	Valor total dos débitos nos documentos fiscais	Valor total das operações que geram débitos de ICMS lançadas em documentos fiscais nos 12 meses anteriores à apuração de inatividade (VENDAS).	Decimal
decl_docf	Índice de ausência de lastro de créditos	Índice que busca detectar atos de COMISSÃO, conforme definidos por Kirchler (2007). Relaciona as variáveis decl_cred e docf_cred. A fórmula utilizada no cálculo deste índice foi omitida do texto para preservar o sigilo de algumas informações.	Índice Decimal
docf_decl	Índice de ausência de lançamento de débitos	Índice que busca detectar atos de OMISSÃO, conforme definidos por Kirchler (2007). Relaciona as variáveis decl_debi e docf_debi. A fórmula utilizada no cálculo deste índice foi omitida do texto para preservar o sigilo de algumas informações.	Índice Decimal
docf_said	Valor total de produtos e serviços em Documentos Fiscais de saída	Indica o valor total dos produtos e serviços lançados em notas fiscais de saída nos 12 meses anteriores à apuração de inatividade (VENDAS).	Decimal

docf_entd	Valor total de produtos e serviços em Documentos Fiscais de entrada	Indica o valor total dos produtos e serviços lançados em notas fiscais de entrada nos 12 meses anteriores à apuração de inatividade (COMPRAS e DEVOLUÇÕES).	Decimal
docf_ind1	Índice de lançamento de crédito	Índice que relaciona as variáveis docf_cred e docf_entd. A fórmula utilizada no cálculo deste índice foi omitida do texto para preservar o sigilo de algumas informações.	Índice Decimal
docf_ind2	Índice de lançamento de débito	Índice que relaciona as variáveis docf_debt e docf_said. A fórmula utilizada no cálculo deste índice foi omitida do texto para preservar o sigilo de algumas informações.	Índice Decimal
docf_ind3	Índice de ausência de lastro de saídas em relação a entradas	Índice que busca detectar atos de COMISSÃO e OMISSÃO, conforme definidos por Kirchler (2007). Relaciona as variáveis docf_entd e docf_said. A fórmula utilizada no cálculo deste índice foi omitida do texto para preservar o sigilo de algumas informações.	Índice Decimal

Fonte: Elaborado pelo autor da pesquisa.

Tabela 7: Experiência anterior com o Fisco (percepção de risco) - Fatores psicológicos

Variável	Descrição	Detalhamento	Formato
dilg_insc	Tempo entre inscrição e diligência	Tempo em meses entre a inscrição no cadastro e o início da apuração de inatividade.	Inteiro
qtde_fisc	Quantidade de fiscalizações	Quantidade de fiscalizações encerradas sofridas nos 12 meses anteriores ao início da apuração de inatividade.	Inteiro
qtde_auto	Quantidade de autos de infração	Quantidade de autos de infração recebidos nos 12 meses anteriores ao início da apuração de inatividade.	Inteiro

Fonte: Elaborado pelo autor da pesquisa.

Tabela 8: Características demográficas - Fatores de ambiente

Variável	Descrição	Detalhamento	Formato
cad_t_natj	Natureza jurídica	Natureza jurídica do estabelecimento.	Fator
cad_t_port	Porte	Porte empresarial.	Fator
cad_t_cnae	CNAE	Código de atividade econômica principal.	Fator
cad_t_subt	ST	Indica se o contribuinte é (1 – Sim) ou não (0 – Não) substituto tributário	Fator
cad_t_delg	DRT	Delegacia Regional Tributária de jurisdição.	Fator

Fonte: Elaborado pelo autor da pesquisa.

Tabela 9: Relação com outros indivíduos - Fatores sociológicos

Variável	Descrição	Detalhamento	Formato
cont_empr	Empresas contador	Total de empresas inscritas no ICMS no estado de São Paulo sob a responsabilidade do contador da empresa que sofreu apuração por inatividade.	Inteiro
cont_empi	Empresas irregulares contador	Total de empresas inscritas no ICMS no estado de São Paulo em situação irregular sob a responsabilidade do contador da empresa que sofreu apuração por inatividade.	Inteiro
cont_idx	Índice contador	Índice que relaciona as variáveis cont_empr e cont_empi. A fórmula utilizada no cálculo deste índice foi omitida do texto para preservar o sigilo de algumas informações.	Índice Decimal
soc1_empr	Empresas sócio 1	Total de empresas inscritas no ICMS no estado de São Paulo do sócio 1 (primeiro sócio com maior participação no capital social da empresa que sofreu apuração por inatividade).	Inteiro
soc1_empi	Empresas irregulares sócio 1	Total de empresas inscritas no ICMS no estado de São Paulo do sócio 1 em situação irregular.	Inteiro
soc1_idx	Índice de irregularidade sócio 1	Índice que relaciona as variáveis soc1_empr e soc1_empi. A fórmula utilizada no cálculo deste índice foi omitida do texto para preservar o sigilo de algumas informações.	Índice Decimal
soc2_empr	Empresas sócio 2	Total de empresas inscritas no ICMS no estado de São Paulo do sócio 2 (segundo sócio com maior participação no capital social da empresa que sofreu apuração por inatividade).	Inteiro
soc2_empi	Empresas irregulares sócio 2	Total de empresas inscritas no ICMS no estado de São Paulo do sócio 2 em situação irregular.	Inteiro
soc2_idx	Índice de irregularidade sócio 2	Índice que relaciona as variáveis soc2_empr e soc2_empi. A fórmula utilizada no cálculo deste índice foi omitida do texto para preservar o sigilo de algumas informações.	Índice Decimal

Fonte: Elaborado pelo autor da pesquisa.

2.2.1.6 Limitações da pesquisa

O aporte teórico apontou os fatores que determinam o cumprimento tributário, bem como sugestões de características que podem refletir tais fatores. O estudo empírico, por sua vez, buscou determinar a relação de dependência entre as características escolhidas e os perfis de cumprimento, apontando o grau de aproximação provável com o respectivo perfil.

Uma das possíveis limitações da pesquisa aqui proposta, entretanto, é que dentre os diversos fatores que podem influenciar o cumprimento tributário, nem todos puderam ser mapeados. Um exemplo, dentre os fatores elencados por Lawler (1998), Alm e Torgler (2006) e Braithwaite (2003b), são as crenças religiosas.

Essa limitação foi, no entanto, mitigada pela técnica estatística adotada, já que esta possui mecanismos para testar e validar a qualidade do ajuste do modelo estatístico, como será descrito adiante.

2.2.2 Análise dos dados

Hosmer e Lemeshow (2000) apontam que para que um modelo estatístico de qualidade possa ser criado, é **fundamental o conhecimento do pesquisador a respeito do fenômeno estudado e dos dados** que o embasam. O **fenômeno** em estudo **foi amplamente detalhado na discussão teórica**. A **análise exploratória dos dados**, por sua vez, será **feita neste tópico**.

Todas as análises estatísticas e visuais desta pesquisa foram executadas utilizando-se o **Software RStudio Versão 1.1.456** com os **pacotes do R Versão x64 3.5.1** (R CORE TEAM, 2018).

A exploração das variáveis foi feita de duas formas:

1. Estatística e
2. Visual.

A **análise estatística** buscou identificar se as médias nos valores de cada variável são estatisticamente diferentes entre os grupos de referência. Para tanto foi aplicado o **Welch Two Sample t-test**.

De acordo com Yuen (1974), esse teste é utilizado para verificar a **hipótese** de que dois **grupos têm médias** iguais no âmbito da variável em análise. A rejeição dessa hipótese, chamada de hipótese nula, indica que as médias são **estatisticamente diferentes em um intervalo de confiança de 95%**.

A autora aponta, ainda, que os *Student's t-tests* assumem que duas populações têm distribuição normal e variâncias equivalentes. O *Welch Two Sample t-test*, por sua vez, embora mantenha o pressuposto de normalidade, é capaz de funcionar mesmo em casos em que não é possível afirmar que as variâncias entre as populações sejam iguais.

Best e Rayner (1987) apontam que esse teste é uma solução aproximada para o *Behrens-Fisher problem* – um dos problemas sem solução na Estatística – exatamente por ser capaz de funcionar mesmo quando não se pode assumir que as variâncias entre as populações sejam iguais.

No caso da presente pesquisa, como os dados da população inteira não estão disponíveis, mas apenas os da amostra, entende-se não ser razoável assumir que as variâncias sejam iguais entre os grupos de referência e, por isso, considerando-se sua maior flexibilidade, optou-se pelo *Welch Two Sample t-test* para a análise estatística das médias.

A **análise visual**, por sua vez, foi feita através de **gráficos de dispersão**. Nela, buscou-se entender como os **dados** no âmbito de cada variável estão **distribuídos entre os grupos de referência para se identificar padrões** nessa distribuição.

Dessa forma, a exploração dos dados objetivou detalhar como as variáveis elencadas nas **Tabelas 6, 7, 8 e 9** se relacionam com o **resultado da diligência fiscal**. Esse resultado encontra-se mapeado na variável **dilg_irre** da **Tabela 5** e permite agrupar os indivíduos da amostra em **0 – Regular e 1 – Irregular**.

Nesse sentido, nas seções a seguir será demonstrada a **relação de cada variável com o resultado da diligência (dilg_irre)** para que se verifique como estão distribuídos os dados nos **grupos de referência (0 – Regular e 1 – Irregular)**.

Durante a análise visual, buscou-se a formulação de hipóteses de influência esperada de cada variável na previsão do resultado da diligência e serviu de auxílio na seleção de variáveis que comporão o modelo, conforme será visto na seção **2.2.3 Modelo empírico**.

Todas as variáveis elencadas nas **Tabelas 5, 6, 7, 8 e 9** foram exploradas, mas devido à sensibilidade de alguns dados, que são protegidos por sigilo fiscal de acordo com o Artigo 198 do Código Tributário Nacional (BRASIL, 1966), somente serão discutidas nas seções a seguir as variáveis que mapeiam dados obtidos pelo autor na condição de cidadão através do Sistema Integrado de Informações ao Cidadão de SP (SIC.SP, 2018) e as variáveis de índices criadas na própria pesquisa. Estas últimas estão identificadas pela palavra Índice no campo Formato das **Tabelas 6 e 9**.

A próxima seção discutirá a variável **dilg_irre (Tabela 5)**, que representa o resultado da diligência nos grupos de referência, e será seguida das seções que analisarão as variáveis das **Tabelas 6, 7, 8 e 9**.

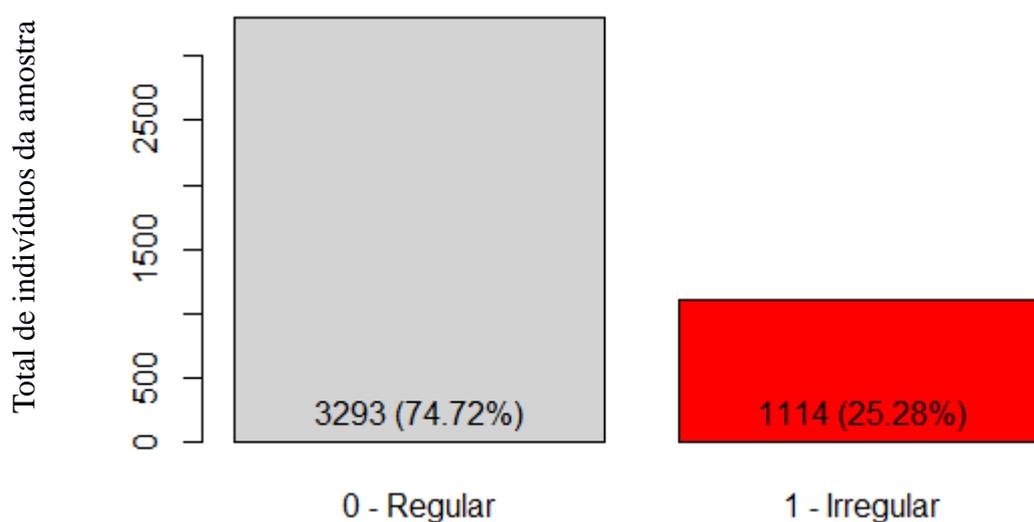
2.2.2.1 Variável resultado da diligência

A variável **dilig_irre**, descrita na **Tabela 5**, representa uma classificação criada pelo autor da pesquisa para o resultado das diligências fiscais. Nessa classificação, procurou-se distinguir contribuintes cuja diligência encontrou apenas indícios de inatividade, daqueles em que foram detectados também outros ilícitos tributários, como sonegação.

Essa **variável binária** assume dois valores: *0 – Regular* e *1 – Irregular*. O **valor 1 - Irregular indica a ocorrência do evento de interesse**, qual seja, a **existência de irregularidades tributárias além da simples inatividade**.

O **Gráfico 4** mostra a distribuição da amostra nos grupos de referência.

Gráfico 4: Distribuição da amostra de acordo com os grupos de referência.



Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.

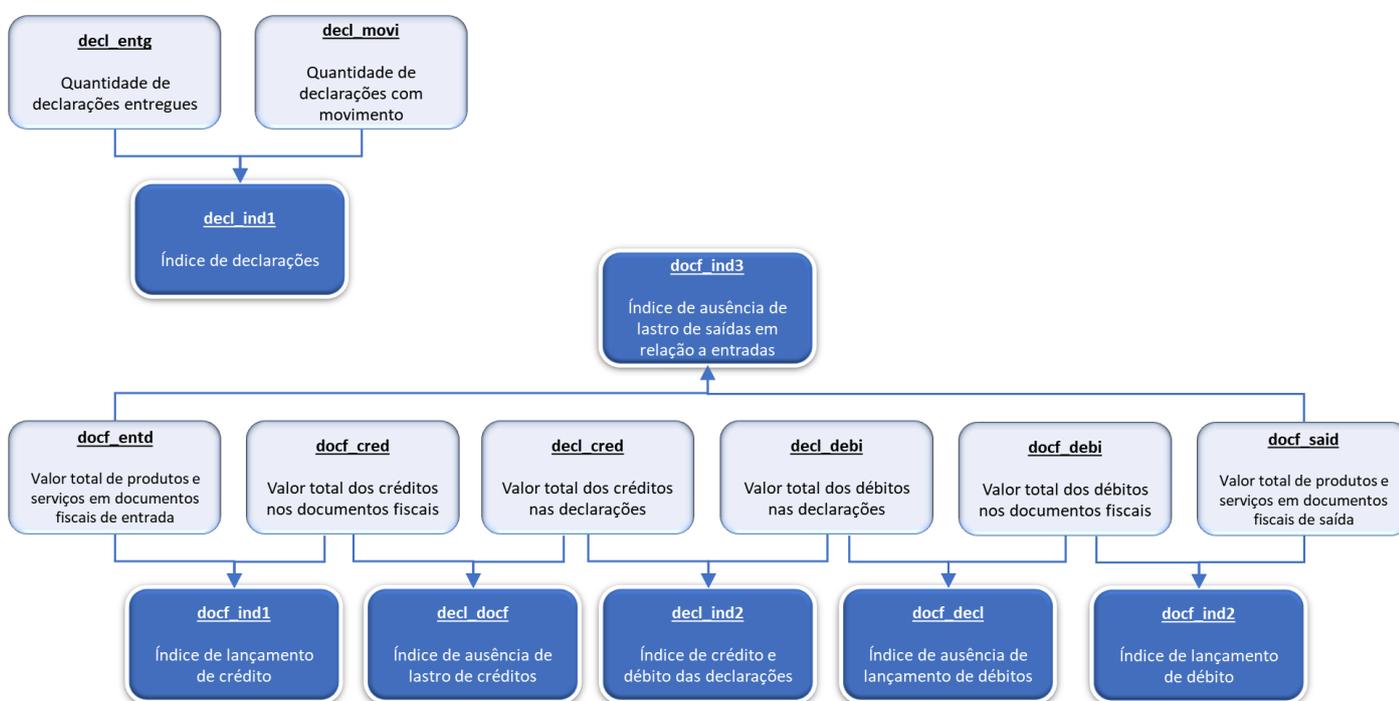
É possível observar que a amostra possui uma distribuição não homogênea em relação aos grupos de referência, pois **74,72%** dos indivíduos estão no grupo *0 – Regular*, ao passo que apenas **25,28%** foram classificados no grupo *1 – Irregular*. Essa distribuição não homogênea impõe uma exigência adicional ao modelo, que será discutida na seção 2.2.5 *Análise dos resultados do modelo*.

2.2.2.2 Variáveis que representam fatores econômico-financeiros

Dentre as variáveis que representam os fatores econômico-financeiros descritas na **Tabela 6**, serão detalhas aquelas criadas na pesquisa, quais sejam: **decl_ind1**, **decl_ind2**, **decl_docf**, **docf_decl**, **docf_ind1**, **docf_ind2**.

As fórmulas utilizadas na criação das variáveis-índices não serão exibidas para preservar o sigilo de algumas informações, mas suas variáveis de origem podem ser vistas na **Figura 16**.

Figura 16: Composição das variáveis de índices que representam fatores econômico-financeiros.



Fonte: Elaborado pelo autor da pesquisa.

Como é possível observar, **há pelo menos um índice relacionado a cada uma das variáveis de origem** (variáveis base). Em função disso, **supõe-se que os índices sejam suficientes para mapear as influências de tais variáveis na previsão do resultado da diligência**. A confirmação ou não dessa suposição será discutida na seção 2.2.3 *Modelo empírico*.

Como mencionado na seção 2.2.2 *Análise dos dados*, a exploração das variáveis foi feita de forma estatística e visual. O resultado da análise estatística (aplicação do *Welch Two Sample t-test*) pode ser visualizado na **Tabela 10**.

Tabela 10: Welch Two Sample t-test das variáveis de índices que representam fatores econômico-financeiros contra a variável *dilg_irre*.

Variável	Descrição	Estatística <i>t</i>	<i>p-value</i>	Média <i>0 – Regular</i>	Média <i>1 – Irregular</i>	Signif. estatística
decl_ind1	Índice de declarações	28.665	< 2.2e-16	0.262	0.037	****
decl_ind2	Índice de crédito e débito das declarações	-9.263	< 2.2e-16	-0.089	-0.019	****
decl_docf	Índice de ausência de lastro de créditos	9.593	< 2.2e-16	-0.320	-0.487	****
docf_decl	Índice de ausência de lançamento de débitos	-34.907	< 2.2e-16	0.098	0.664	****
docf_ind1	Índice de lançamento de crédito	-0.149	0.88130	-0.543	-0.540	
docf_ind2	Índice de lançamento de débito	13.607	< 2.2e-16	-0.393	-0.571	****
docf_ind3	Índice de ausência de lastro de saídas em relação a entradas	-25.546	< 2.2e-16	-0.092	0.408	****

Fonte: Elaborado pelo autor da pesquisa.

Nota: Códigos de significância estatística: 0 ****; 0.001 **; 0.01 *; 0.05 .; 0.1 ' '; 1

O *Welch Two Sample t-test* foi aplicado confrontando cada uma das variáveis da **Tabela 10** com os grupos de referência (variável *dilg_irre*) para calcular e analisar as médias de cada variável em cada grupo.

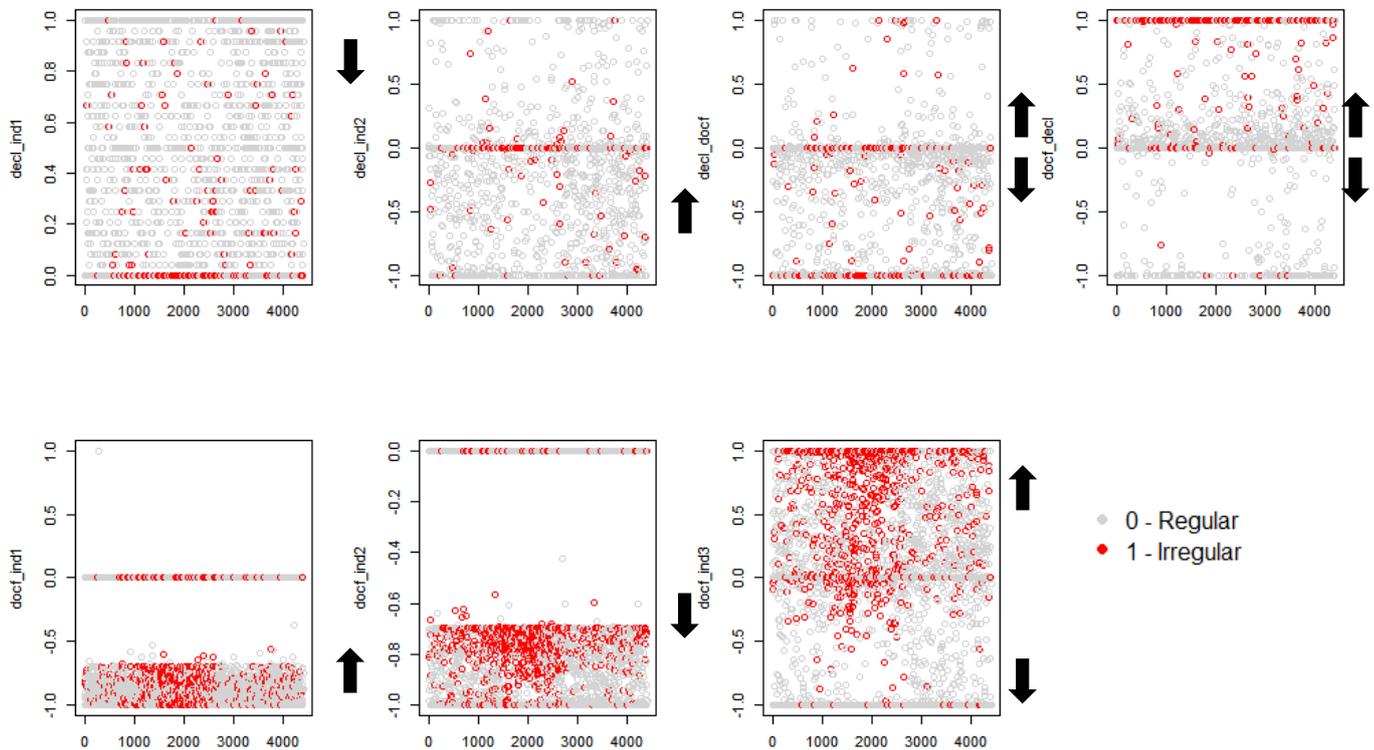
Considerando-se um **grau de confiança de 95%**, *p-values* menores que **0.05** indicam que a **hipótese nula** (de que as médias são iguais) deve ser rejeitada e a **hipótese alternativa** (médias diferentes) se torna a mais provável. Nesses casos, a coluna *Significância estatística* conterá ****, **, * ou ..

A **suposição de pesquisa** é que as **variáveis com médias estatisticamente diferentes entre os grupos tendem a exercer influência na previsão do resultado da diligência**, ou seja, na alocação do contribuinte entre os grupos *0 – Regular* e *1 – Irregular*. Essa suposição necessita, entretanto, ser confrontada com a análise visual de cada variável, feita a seguir, e também com o resultado da seleção de variáveis do modelo, que será visto na seção 2.2.3 *Modelo empírico*.

A **Tabela 10** demonstra que, das variáveis listadas, **apenas docf_ind1 – Índice de lançamento de crédito não possui médias estatisticamente diferentes entre os grupos**, o que pode indicar que essa variável não será relevante para prever o resultado da diligência.

A **Figura 17** mostra a análise visual das variáveis, através dos gráficos de dispersão dos valores de cada uma nos grupos de referência.

Figura 17: Gráficos de dispersão dos valores das variáveis de índices que representam fatores econômico-financeiros nos grupos de referência.



Fonte: Dados da pesquisa.

Nota: Gráficos de dispersão das variáveis `decl_ind1`, `decl_ind2`, `decl_docf`, `docf_decl`, `docf_ind1`, `docf_ind2` e `docf_ind3`, confrontadas com os grupos de referência (*0 - Regular* e *1 - Irregular*) da variável `dilg_irre`. O eixo y, das ordenadas, representa a escala de valores de cada variável. O eixo x, das abscissas, representa cada indivíduo da amostra. As setas \uparrow e \downarrow indicam qual seria o comportamento não conforme no âmbito da variável, independentemente da distribuição nos grupos de referência. Elas permitem verificar se o comportamento não conforme no âmbito da variável é confirmado pela distribuição dos indivíduos nos grupos de referência.

Os gráficos da **Figura 17** mostram, por cores, como se dispersam os valores de cada variável entre os grupos de referência. No primeiro gráfico, por exemplo, é possível observar que grande parte dos indivíduos classificados como *1 - Irregular* possuem valores próximos a 0.0 na variável `decl_ind1`.

Ao analisar as setas, que indicam o comportamento não conforme, e confrontá-las com os pontos de concentração de indivíduos do grupo *1 - Irregular*, podemos tecer considerações em relação à importância da variável para prever o resultado da diligência. Considerando, por exemplo, que a seta no gráfico da variável `decl_ind1` indica que o distanciamento do valor 1.0 em direção a 0.0 pode ser indício de irregularidade e que há grande concentração de indivíduos do grupo *1 - Irregular* próximos ao valor 0.0 da variável, tem-se um possível indicativo de que o resultado da diligência corrobora a suposição de que a aproximação de 0.0 indica uma probabilidade maior de que o indivíduo seja classificado como *1 - Irregular*.

Nas seções seguintes, cada gráfico da **Figura 17** será analisado e hipóteses de influência de cada variável na previsão do resultado da diligência serão discutidas.

2.2.2.2.1 Variável índice de declarações

A variável **decl_ind1 – Índice de declarações** foi criada para capturar os efeitos das variáveis **decl_entg – Quantidade de declarações entregues** e **decl_movi – Quantidade de declarações com movimento**, conforme mostrou a **Figura 16**.

De acordo com a fórmula utilizada para construção dessa variável, tem-se que **o distanciamento de 1.0 em direção a 0.0 pode indicar não conformidade**.

O primeiro gráfico da **Figura 17** mostra que há uma concentração grande de indivíduos do grupo *1 – Irregular* no quadrante próximo a 0.0. Adicionalmente, o resultado do **Welch Two Sample t-test** na **Tabela 10** mostrou que as médias dos valores da variável entre os grupos de referência são estatisticamente diferentes e que a média do grupo *1 – Irregular* (0.037) está mais próxima de 0.0 do que a do grupo *0 – Regular* (0.262).

As três informações em conjunto:

1. Formulação intrínseca à variável indicando que distanciamento de 1.0 em direção a 0.0 pode indicar irregularidade.
2. Diferença nas médias, sendo a média do grupo *1 – Irregular* mais próxima de 0.0.
3. Concentração de indivíduos classificados como *1 – Irregular* próxima ao ponto 0.0.

São possíveis evidências de que essa variável **deverá ser relevante para prever o resultado da diligência**.

2.2.2.2.2 Variável índice de crédito e débito das declarações

A variável **decl_ind2 – Índice de crédito e débito das declarações** foi criada para capturar os efeitos das variáveis **decl_cred – Valor total dos créditos nas declarações** e **decl_debi – Valor total dos débitos nas declarações**, conforme mostrou a *Figura 16*.

De acordo com a fórmula utilizada para construção dessa variável, tem-se que **o distanciamento de -1.0 pode indicar não conformidade**.

O segundo gráfico da *Figura 17* mostra que há uma concentração grande de indivíduos do grupo *1 – Irregular* no quadrante próximo a 0.0. Adicionalmente, o resultado do **Welch Two Sample t-test** na *Tabela 10* mostrou que as médias dos valores da variável entre os grupos de referência são estatisticamente diferentes e que a média do grupo *1 – Irregular* (-0.019) está mais distante de -1.0 do que a do grupo *0 – Regular* (-0.089).

As três informações em conjunto:

1. Formulação intrínseca à variável indicando que distanciamento de -1.0 pode indicar irregularidade.
2. Diferença nas médias, sendo a média do grupo *1 – Irregular* mais distante do valor -1.0.
3. Concentração de indivíduos classificados como *1 – Irregular* próxima ao ponto 0.0.

São possíveis evidências de que essa variável **deverá ser relevante para prever o resultado da diligência**.

2.2.2.2.3 Variável índice de ausência de lastro de créditos

A variável **decl_docf – Índice de ausência de lastro de créditos** foi criada para capturar os efeitos das variáveis **decl_cred – Valor total dos créditos nas declarações** e **docf_cred – Valor total dos créditos nos documentos fiscais**, conforme mostrou a *Figura 16*.

De acordo com a fórmula utilizada para construção dessa variável, tem-se que **o distanciamento do valor 0.0 pode indicar não conformidade**.

O terceiro gráfico da **Figura 17** mostra que há uma concentração grande de indivíduos do grupo *1 – Irregular* no quadrante próximo a -1.0. Adicionalmente, o resultado do **Welch Two Sample t-test** na **Tabela 10** mostrou que as médias dos valores da variável entre os grupos de referência são estatisticamente diferentes e que a média do grupo *1 – Irregular* (-0.487) está mais próxima de -1.0 do que a do grupo *0 – Regular* (-0.320).

As três informações em conjunto:

1. Formulação intrínseca à variável indicando que distanciamento de 0.0 pode indicar irregularidade.
2. Diferença nas médias, sendo a média do grupo *1 – Irregular* mais próxima do valor -1.0.
3. Concentração de indivíduos classificados como *1 – Irregular* próxima ao ponto -1.0.

São possíveis evidências de que essa variável **deverá ser relevante para prever o resultado da diligência.**

2.2.2.2.4 Variável índice de ausência de lançamento de débitos

A variável **docf_decl – Índice de ausência de lançamento de débitos** foi criada para capturar os efeitos das variáveis **decl_debi – Valor total dos débitos nas declarações** e **docf_debi – Valor total dos débitos nos documentos fiscais**, conforme mostrou a **Figura 16**.

De acordo com a fórmula utilizada para construção dessa variável, tem-se que **o distanciamento do valor 0.0 pode indicar não conformidade.**

O quarto gráfico da **Figura 17** mostra que há uma concentração grande de indivíduos do grupo *1 – Irregular* no quadrante próximo a 1.0. Adicionalmente, o resultado do **Welch Two Sample t-test** na **Tabela 10** mostrou que as médias dos valores da variável entre os grupos de referência são estatisticamente diferentes e que a média do grupo *1 – Irregular* (0.664) está mais próxima de 1.0 do que a do grupo *0 – Regular* (0.098).

As três informações em conjunto:

1. Formulação intrínseca à variável indicando que distanciamento de 0.0 pode indicar irregularidade.
2. Diferença nas médias, sendo a média do grupo *1 – Irregular* mais próxima do valor 1.0.
3. Concentração de indivíduos classificados como *1 – Irregular* próxima ao ponto 1.0.

São possíveis evidências de que essa variável **deverá ser relevante para prever o resultado da diligência.**

2.2.2.2.5 Variável índice de lançamento de crédito

A variável **docf_ind1 – Índice de lançamento de crédito** foi criada para capturar os efeitos das variáveis **docf_cred – Valor total dos créditos nos documentos fiscais** e **docf_entd – Valor total de produtos e serviços em Documentos Fiscais de entrada**, conforme mostrou a **Figura 16**.

De acordo com a fórmula utilizada para construção dessa variável, tem-se que **o distanciamento do valor -1.0 pode indicar não conformidade.**

O quinto gráfico da **Figura 17** mostra, no entanto, que a maior concentração de indivíduos do grupo *1 – Irregular* está no quadrante próximo a -1.0, o que conflita com a formulação intrínseca à variável.

Adicionalmente, o resultado do **Welch Two Sample t-test** na **Tabela 10** mostrou que as médias dos valores da variável entre os grupos de referência **não** são estatisticamente diferentes.

Isto posto, **não parece haver evidências de que a variável docf_ind1 seja relevante para prever o resultado da diligência.**

2.2.2.2.6 Variável índice de lançamento de débito

A variável **docf_ind2 – Índice de lançamento de débito** foi criada para capturar os efeitos das variáveis **docf_debi – Valor total dos débitos nos documentos fiscais** e **docf_said – Valor total de produtos e serviços em Documentos Fiscais de saída**, conforme mostrou a **Figura 16**.

De acordo com a fórmula utilizada para construção dessa variável, tem-se que **a aproximação do valor -1.0 pode indicar não conformidade**.

O resultado do *Welch Two Sample t-test* na **Tabela 10** mostrou que as médias dos valores da variável entre os grupos de referência são estatisticamente diferentes e que a média do grupo *1 – Irregular* (-0.571) está mais próxima de -1.0 do que a do grupo *0 – Regular* (-0.393).

O sexto gráfico da **Figura 17** mostra que a maior concentração de indivíduos do grupo *1 – Irregular* está no quadrante entre -0.7 e -0.9, mas há também uma concentração forte nesse quadrante de indivíduos do grupo *0 – Regular*.

Essa distribuição não confirma a formulação intrínseca à variável de que a aproximação de -1.0 poderia ser indicativo de *não conformidade*, e, portanto, **não traz evidências de que a variável docf_ind2 seja relevante para prever o resultado da diligência**.

2.2.2.2.7 Variável índice de ausência de lastro de saídas em relação a entradas

A variável **docf_ind3 – Índice de ausência de lastro de saídas em relação a entradas** foi criada para capturar os efeitos das variáveis **docf_said – Valor total de produtos e serviços em Documentos Fiscais de saída** e **docf_entd – Valor total de produtos e serviços em Documentos Fiscais de entrada**, conforme mostrou a **Figura 16**.

De acordo com a fórmula utilizada para construção dessa variável, tem-se que **a aproximação de 1.0 ou -1.0 pode indicar não conformidade**.

O sétimo gráfico da **Figura 17** mostra que há uma concentração grande de indivíduos do grupo *1 – Irregular* no quadrante próximo a 1.0. Adicionalmente, o resultado do *Welch Two Sample t-test* na **Tabela 10** mostrou que as médias dos valores da variável entre os grupos de referência são estatisticamente diferentes e que a média do grupo *1 – Irregular* (0.408) está mais próxima de 1.0 do que a do grupo *0 – Regular* (-0.092).

As três informações em conjunto:

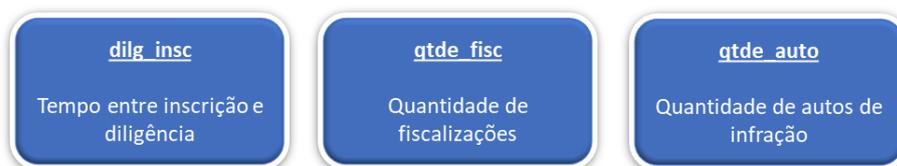
1. Formulação intrínseca à variável indicando que a aproximação de 1.0 ou -1.0 pode indicar irregularidade.
2. Diferença nas médias, sendo a média do grupo *1 – Irregular* mais próxima do valor 1.0.
3. Concentração de indivíduos classificados como 1 – Irregular próxima ao ponto 1.0.

São possíveis evidências de que essa variável **deverá ser relevante para prever o resultado da diligência.**

2.2.2.3 Variáveis que representam fatores psicológicos

As **variáveis que representam fatores psicológicos**, descritas na *Tabela 7*, são aquelas que refletem a experiência anterior com o Fisco. Como abordado na discussão teórica, essa experiência pode influenciar os níveis de conformidade. As variáveis desse grupo mapeadas na pesquisa são: **dilg_insc**, **qtde_fisc** e **qtde_auto** e são representadas na *Figura 18*.

Figura 18: Variáveis que representam fatores psicológicos.



Fonte: Elaborado pelo autor da pesquisa.

No caso dessas variáveis, não foi atribuída nenhuma formulação intrínseca que indique em qual direção os valores tenderiam à *não conformidade*. Isso ocorre porque, conforme apontado por Kirchler (2007), os efeitos psicológicos variam de indivíduo para indivíduo.

Como exemplo da argumentação, o autor cita o fato de que um auto de infração aplicado a um contribuinte pode aumentar sua aversão ao risco e estimular uma postura de conformidade. De forma inversa, um outro contribuinte pode reagir com revolta e buscar meios de sonegar para compensar a pena pecuniária.

Na mesma vertente, um contribuinte que sofreu diversas fiscalizações pode adotar uma postura de conformidade em função da percepção de que o Fisco está atento, enquanto que outro pode buscar meios mais sofisticados para esconder irregularidades.

Por fim, o tempo de existência do empreendimento também pode ter efeito diverso a depender da percepção de cada contribuinte.

Kirchler (2007) aponta que o início dos empreendimentos é normalmente tortuoso e alguns contribuintes buscam a sonegação nesse início como forma de redução de custos. Outros, por sua vez, com receio de que uma punição possa dificultar ainda mais a sobrevivência do empreendimento, buscam cumprir com todas as suas obrigações nas fases iniciais.

Na outra ponta, mas dentro da mesma argumentação, o autor explica que o aumento do tempo de existência dos empreendimentos leva alguns contribuintes à conformidade, enquanto que outros incorporam a percepção de que suas práticas irregulares não serão detectadas e buscam o não cumprimento tributário.

A ausência de formulações intrínsecas a essas variáveis não significa que elas não serão relevantes para prever o resultado da diligência, pois a análise estatística e o modelo poderão ser capazes de detectar padrões.

Analogamente à lista de variáveis da seção anterior, a exploração dos dados foi feita de forma estatística e visual. O resultado da análise estatística (aplicação do *Welch Two Sample t-test*) pode ser visualizado na **Tabela 11**.

Tabela 11: *Welch Two Sample t-test* das variáveis de índices que representam fatores psicológicos contra a variável *dilig_irre*.

Variável	Descrição	Estatística <i>t</i>	<i>p-value</i>	Média <i>0 – Regular</i>	Média <i>1 – Irregular</i>	Signif. estatística
dilig_insc	Tempo entre inscrição e diligência	19.454	< 2.2e-16	115.397	51.857	‘****’
qtde_fisc	Quantidade de fiscalizações	-8.0159	1.87e-15	0.801	1.277	‘****’
qtde_auto	Quantidade de autos de infração	2.4564	0.01412	0.102	0.059	‘.’

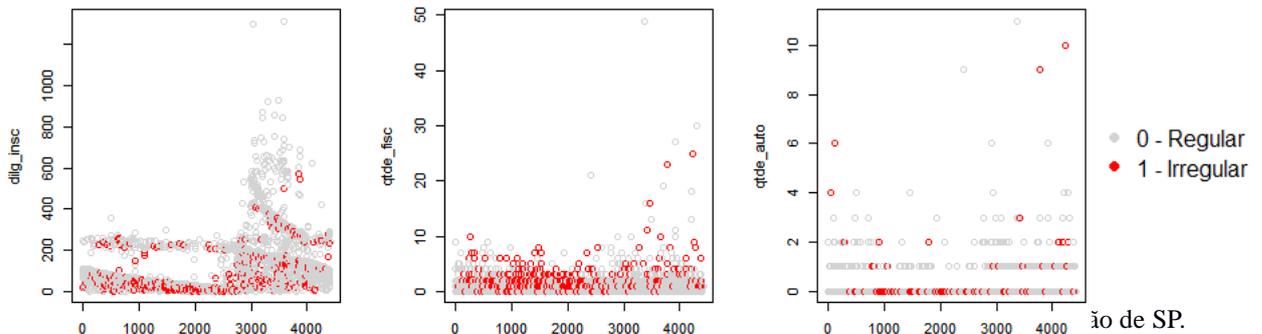
Fonte: Elaborado pelo autor da pesquisa.

Nota: Códigos de significância estatística: 0 ‘****’; 0.001 ‘***’; 0.01 ‘**’; 0.05 ‘.’; 0.1 ‘.’; 1

A **Tabela 11** demonstra que, das variáveis listadas, **todas possuem médias estatisticamente diferentes entre os grupos no intervalo de confiança de 95%**.

A **Figura 19** mostra os gráficos de dispersão dos valores de cada variável nos grupos de referência.

Figura 19: Gráficos de dispersão dos valores das variáveis que representam fatores psicológicos nos grupos de referência.



Nota: Gráficos de dispersão das variáveis *dilg_insc*, *qtde_fisc* e *qtde_auto*, confrontadas com os grupos de referência (*0 – Regular e 1 – Irregular*) da variável *dilg_irre*. O eixo y, das ordenadas, representa a escala de valores de cada variável. O eixo x, das abcissas, representa cada indivíduo da amostra. As setas \uparrow e \downarrow indicam qual seria o comportamento não conforme no âmbito da variável, independentemente da distribuição nos grupos de referência. Elas permitem verificar se o comportamento não conforme no âmbito da variável é confirmado pela distribuição dos indivíduos nos grupos de referência.

Nas seções seguintes, cada gráfico da **Figura 19** será analisado e hipóteses de influência de cada variável na previsão do resultado da diligência serão discutidas.

2.2.2.3.1 Variável tempo entre inscrição e diligência

A variável **dilg_insc** representa o tempo em meses entre a inscrição no cadastro e a diligência de apuração de inatividade efetuada no contribuinte.

O primeiro gráfico da **Figura 19** mostra que a maior concentração de indivíduos do grupo *1 – Irregular* está no quadrante entre 0 e 50, indicando que as irregularidades tributárias talvez sejam mais frequentes em empresas com menos tempo de existência ou que o Fisco promova o maior número de diligências nos meses iniciais de existência da empresa e, portanto, o grau de detecção seja maior.

A diferença entre as médias acusada pelo *Welch Two Sample t-test* e a concentração de indivíduos do grupo *1 – Irregular* no quadrante entre 0 e 50 parece indicar que esta variável **será relevante para prever o resultado da diligência.**

2.2.2.3.2 Variável quantidade de fiscalizações

A variável **qtde_fisc** representa a quantidade de fiscalizações nos doze meses imediatamente anteriores à diligência de apuração de inatividade.

O segundo gráfico da **Figura 19** mostra que a maior concentração de indivíduos do grupo *1 – Irregular* ocorre a partir de valores maiores que 0 até 5. Ocorre, no entanto, que há também uma grande concentração de indivíduos do grupo *0 – Regular* no mesmo local.

Embora a diferença entre as médias tenha sido acusada pelo *Welch Two Sample t-test*, a falta de confirmação de um padrão no gráfico de dispersão **parece não permitir supor que essa variável será relevante para prever o resultado da diligência.**

2.2.2.3.3 Variável quantidade de autos de infração

A variável **qtde_auto** representa a quantidade de autos de infração sofridos nos doze meses imediatamente anteriores à diligência de apuração de inatividade.

O terceiro gráfico da **Figura 19** mostra que a maior concentração de indivíduos do grupo *1 – Irregular* ocorre no quadrante próximo a 0, mas isso também é verdade em relação ao grupo *0 – Regular*, **o que não traz evidências de que essa variável será relevante para previsão do resultado diligência.**

2.2.2.4 Variáveis que representam fatores de ambiente

As variáveis que representam **fatores de ambiente**, descritas na **Tabela 8**, são **aquelas que refletem as características demográficas dos contribuintes**, que podem ter efeito sobre seu nível de conformidade.

De acordo com Kirchler (2007), o ambiente de inserção dos contribuintes pode influenciar fortemente o cumprimento tributário. O autor aponta, por exemplo, que um contribuinte inserido em uma região ou em uma atividade econômica em que a sonegação é prática comum, poderá ter uma maior propensão à *não conformidade*.

Assim como as variáveis que representam fatores psicológicos, não serão estabelecidas formulações intrínsecas a essas variáveis, o que não significa que elas não sejam relevantes para prever o resultado da diligência, pois o modelo poderá ser capaz de detectar

padrões. As variáveis desse grupo mapeadas na pesquisa são: **cad_t_natj**, **cad_t_port**, **cad_t_cnae**, **cad_t_subt** e **cad_t_delg** e são apresentadas na **Figura 20**.

Figura 20: Variáveis que representam fatores de ambiente.



Fonte: Elaborado pelo autor da pesquisa.

Por serem variáveis categóricas, a análise estatística de diferença de médias não se aplica a essas variáveis, que terão análise visual, como se vê nas seções a seguir.

2.2.2.4.1 Variável natureza jurídica

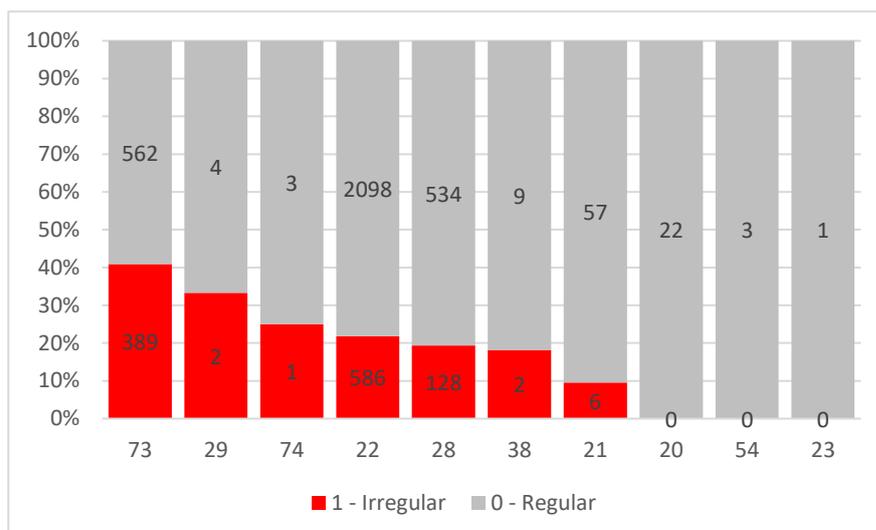
A variável **cad_t_natj** representa a Natureza Jurídica – NJ do estabelecimento, ou seja, sua natureza empresarial. A **Tabela 12** apresenta a lista de naturezas jurídicas presentes na amostra e suas denominações.

Tabela 12: Lista de naturezas jurídicas.

ID	Natureza Jurídica
20	Sociedade Anônima Aberta
21	Sociedade Anônima Fechada
22	Sociedade Empresária Limitada
23	Sociedade Empresária em Nome Coletivo
28	Empresário (Individual)
29	Cooperativa
38	Sociedade Simples Limitada
54	Associação Privada
73	Empresa Individual De Responsabilidade Limitada (de Natureza Empresária)
74	Empresa Individual De Responsabilidade Limitada (de Natureza Simples)

Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.

O **Gráfico 5** mostra a distribuição de valores da variável **cad_t_natj** em relação aos grupos de referência.

Gráfico 5: Distribuição de valores de cadt_natj.

Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.
 Nota: o eixo y, das ordenadas, representa o código de cada natureza jurídica, ao passo que o eixo x, das abcissas, mostra os totais de indivíduos nos grupos 0 – Regular e 1 – Irregular.

Considerando-se a distribuição percentual e os valores absolutos, o **Gráfico 5** mostra que a maior concentração de indivíduos do grupo *1 – Irregular* está nas naturezas jurídicas:

- 73 – Empresa Individual de Responsabilidade;
- 22 – Sociedade Empresária Limitada e
- 28 – Empresário (Individual).

É possível que essa variável seja relevante na previsão do resultado da diligência em função da maior concentração de indivíduos com irregularidade nas naturezas jurídicas citadas. Se for o caso e houver algum padrão estatístico significativo, este deverá ser detectado pelo modelo.

2.2.2.4.2 Variável porte empresarial

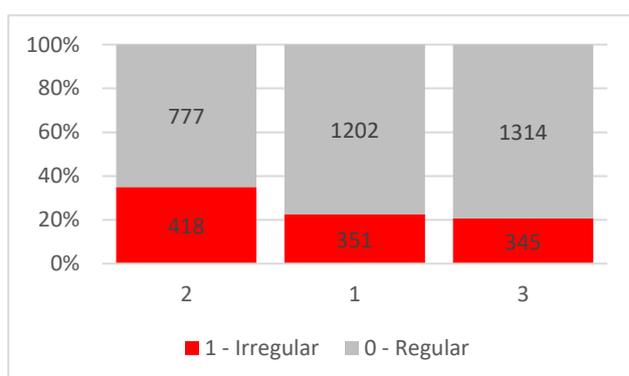
A variável **cad_t_port** representa o Porte Empresarial do estabelecimento. A **Tabela 13** apresenta a lista de portes empresariais presentes na amostra e suas denominações.

Tabela 13: Lista de portes empresariais.

ID	Porte Empresarial
1	Microempresa
2	Empresa de Pequeno Porte
3	Demais

Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.

O **Gráfico 6** mostra a distribuição de valores da variável `cad_t_port` em relação aos grupos de referência.

Gráfico 6: Distribuição de valores de `cad_t_port`.

Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.

Nota: o eixo y, das ordenadas, representa o código de cada porte empresarial, ao passo que o eixo x, das abcissas, mostra os totais de indivíduos nos grupos 0 – Regular e 1 – Irregular.

Considerando-se a distribuição percentual e os valores absolutos, o **Gráfico 6** mostra que os portes **1 – Microempresas** e **3 – Demais tipos de porte de empresas** possuem uma distribuição semelhante em relação à concentração de indivíduos do grupo **1 – Irregular**.

As empresas do porte **2 – Empresa de Pequeno Porte**, por sua vez, parecem ter uma concentração maior de indivíduos do grupo **1 – Irregular** e **é possível que tal concentração seja relevante para a predição do resultado da diligência**.

2.2.2.4.3 Variável código de atividade econômica principal

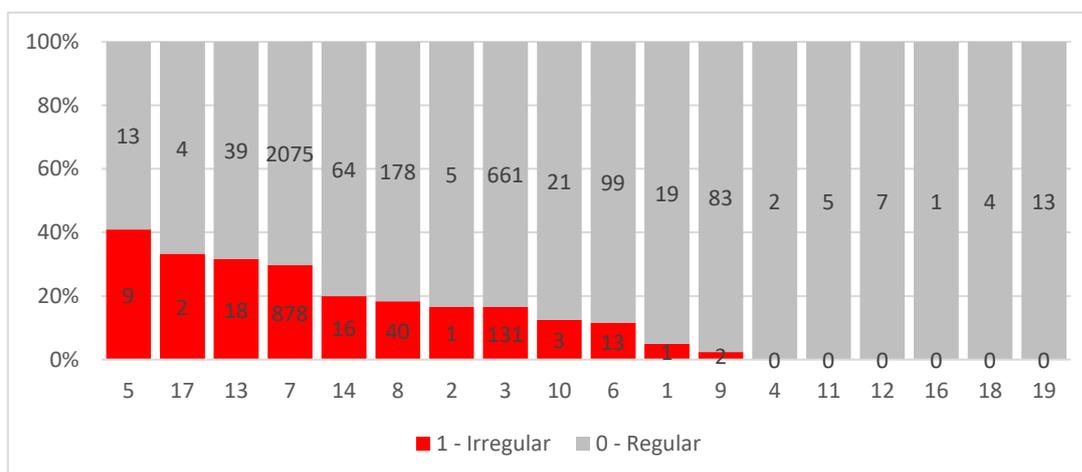
A variável `cad_t_cnae` representa o Código de Atividade Econômica Principal – CNAE do estabelecimento. A **Tabela 14** apresenta a lista de atividades econômicas presentes na amostra e suas denominações.

Tabela 14: Lista de códigos de atividade econômica principal.

ID	Atividade Econômica
1	AGRICULTURA, PECUÁRIA, PRODUÇÃO FLORESTAL, PESCA E AQUICULTURA
2	INDÚSTRIAS EXTRATIVAS
3	INDÚSTRIAS DE TRANSFORMAÇÃO
4	ELETRICIDADE E GÁS
5	ÁGUA, ESGOTO, ATIVIDADES DE GESTÃO DE RESÍDUOS E DESCONTAMINAÇÃO
6	CONSTRUÇÃO
7	COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS
8	TRANSPORTE, ARMAZENAGEM E CORREIO
9	ALOJAMENTO E ALIMENTAÇÃO
10	INFORMAÇÃO E COMUNICAÇÃO
11	ATIVIDADES FINANCEIRAS, DE SEGUROS E SERVIÇOS RELACIONADOS
12	ATIVIDADES IMOBILIÁRIAS
13	ATIVIDADES PROFISSIONAIS, CIENTÍFICAS E TÉCNICAS
14	ATIVIDADES ADMINISTRATIVAS E SERVIÇOS COMPLEMENTARES
16	EDUCAÇÃO
17	SAÚDE HUMANA E SERVIÇOS SOCIAIS
18	ARTES, CULTURA, ESPORTE E RECREAÇÃO
19	OUTRAS ATIVIDADES DE SERVIÇOS

Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.

O **Gráfico 7** mostra a distribuição de valores da variável `cad_t_cnae` em relação aos grupos de referência.

Gráfico 7: Distribuição de valores de `cad_t_cnae`.

Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.

Nota: o eixo y, das ordenadas, representa o código de cada CNAE, ao passo que o eixo x, das abscissas, mostra os totais de indivíduos nos grupos 0 – Regular e 1 – Irregular.

Considerando-se a distribuição percentual e os valores absolutos, o **Gráfico 7** mostra que a maior concentração de indivíduos do grupo 1 – Irregular está nos CNAE:

- 7 – COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS;
- 3 – INDÚSTRIAS DE TRANSFORMAÇÃO e
- 8 – TRANSPORTE, ARMAZENAGEM E CORREIO.

Como a área de atuação do contribuinte tem grande influência na definição de quais indivíduos se relacionam com ele, **é possível que a área – juntamente com as variáveis que representam fatores sociológicos – possa ter influência no comportamento tributário e, conseqüentemente, na predição do resultado da diligência.**

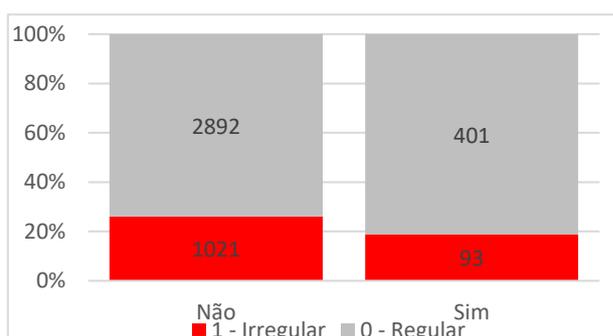
2.2.2.4.4 Variável substituto tributário

A variável **cad_t_subt** indica se o contribuinte é (1 – Sim) ou não (0 – Não) Substituto Tributário – ST. O substituto tributário é um terceiro que a lei obriga ao pagamento do tributo, embora não tenha praticado o fato gerador (COELHO, 2001). Esse terceiro não arcará com o ônus financeiro, que é repassado ao contribuinte de fato, mas será responsável pelo recolhimento aos cofres públicos.

É o que pode ocorrer, por exemplo, com venda de combustíveis. Ao invés de determinar que a arrecadação seja feita na venda ao consumidor final, em cada posto de combustível, a lei pode estabelecer que a distribuidora recolha antecipadamente os montantes devidos em cada etapa subsequente da cadeia e repasse os montantes ao Estado de uma única vez.

O **Gráfico 8** mostra a distribuição de valores da variável **cad_t_subt** em relação aos grupos de referência.

Gráfico 8: Distribuição de valores de cad_t_subt.



Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.
Nota: o eixo y, das ordenadas, representa cada valor da variável Substituto Tributário, ao passo que o eixo x, das abscissas, mostra os totais de indivíduos nos grupos 0 – Regular e 1 – Irregular.

O **Gráfico 8** mostra que a distribuição de indivíduos do grupo 1 – Irregular não é muito diferente nos dois grupos, sendo aproximadamente 22% no caso de ST = Não e 19% no caso de ST = Sim.

Esse pode ser um indicativo de que a variável não será relevante para prever o resultado da diligência.

2.2.2.4.5 Variável delegacia regional tributária

A variável **cad_t_delg** representa a Delegacia Regional Tributária – DRT com jurisdição sobre o contribuinte. De forma geral, ela remete à localização geográfica do estabelecimento.

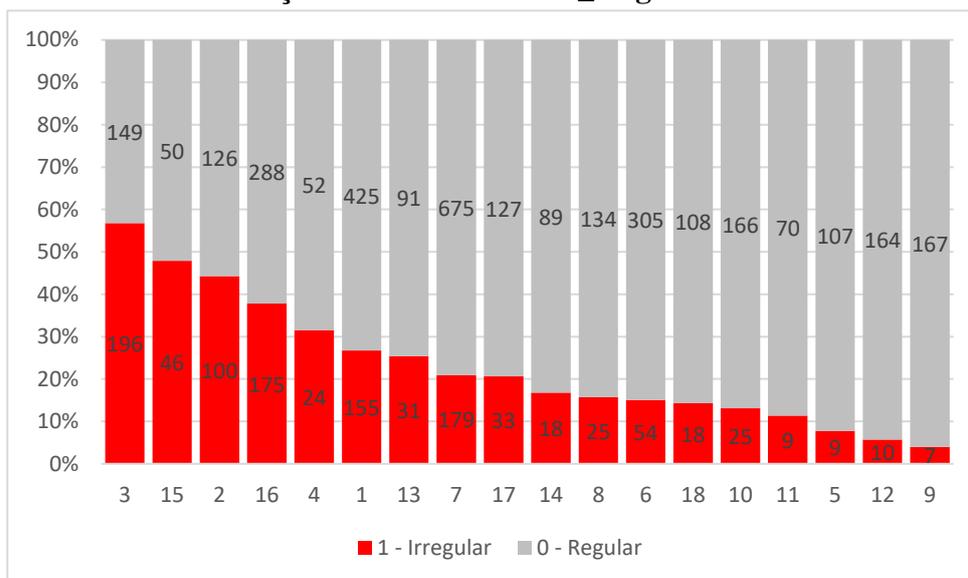
A **Tabela 15** apresenta a lista das delegacias presentes na amostra e suas denominações.

Tabela 15: Lista de delegacias regionais tributárias.

ID	Atividade Econômica
1	DRTC-I - SÃO PAULO
2	DRTC-II - SÃO PAULO
3	DRTC-III - SÃO PAULO
4	DRT-02 - LITORAL
5	DRT-03 - VALE DO PARAÍBA
6	DRT-04 - SOROCABA
7	DRT-05 - CAMPINAS
8	DRT-06 - RIBEIRÃO PRETO
9	DRT-07 - BAURU
10	DRT-08 - SÃO JOSÉ DO RIO PRETO
11	DRT-09 - ARAÇATUBA
12	DRT-10 - PRESIDENTE PRUDENTE
13	DRT-11 - MARÍLIA
14	DRT-12 - ABCD
15	DRT-13 - GUARULHOS
16	DRT-14 - OSASCO
17	DRT-15 - ARARAQUARA

Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.

O **Gráfico 9** mostra a distribuição de valores da variável **cad_t_delg** em relação aos grupos de referência.

Gráfico 9: Distribuição de valores de cadt_delg.

Fonte: Adaptado de resposta a consulta do autor junto ao Sistema Integrado de Informações ao Cidadão de SP.
 Nota: o eixo y, das ordenadas, representa o código de cada delegacia regional tributária, ao passo que o eixo x, das abscissas, mostra os totais de indivíduos nos grupos 0 – Regular e 1 – Irregular.

Considerando-se a distribuição percentual e os valores absolutos, o **Gráfico 9** mostra que a maior concentração de indivíduos do grupo *1 – Irregular* está nas DRT:

- 3 – DRTC-III - SÃO PAULO;
- 16 – DRT-14 – OSASCO;
- 1 – DRTC-I - SÃO PAULO;
- 7 – DRT-05 - CAMPINAS e
- 2 – DRTC-II - SÃO PAULO.

Essa variável reflete a área de localização geográfica do contribuinte. Na discussão teórica foi visto que a área em que o contribuinte é inserido pode influenciar seu comportamento tributário e, por isso, **é possível que essa variável auxilie da predição do resultado da diligência.**

2.2.2.5 Variáveis que representam fatores sociológicos

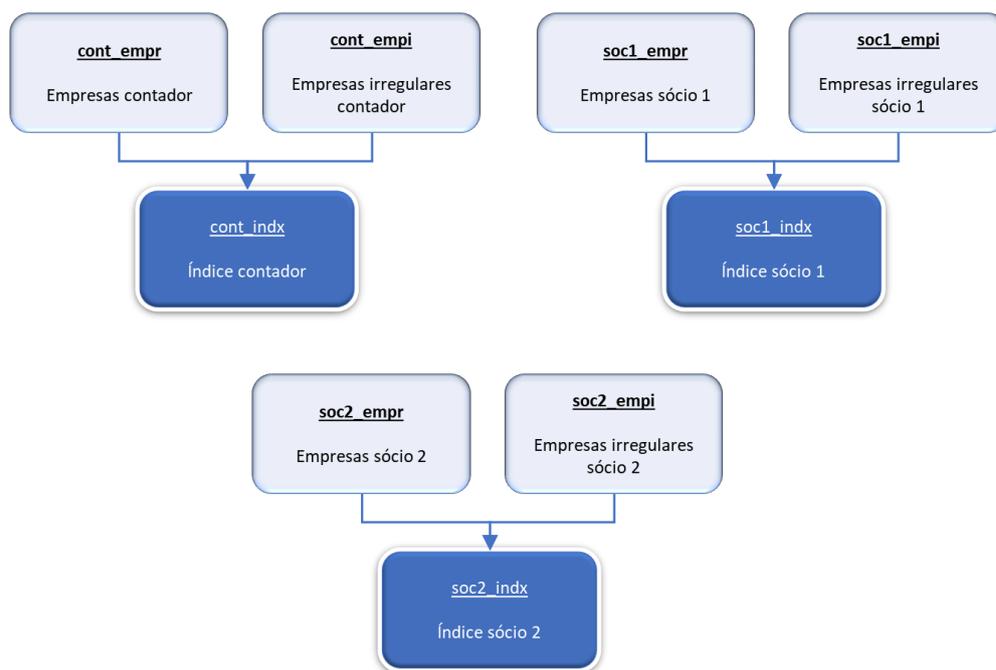
Dentre as variáveis que representam os fatores sociológicos descritas na **Tabela 9**, serão detalhas aquelas criadas na pesquisa, quais sejam: **cont_indx**, **soc1_indx** e **soc2_indx**.

Fatores sociológicos são aqueles ligados à relação direta do contribuinte com outros indivíduos. Kirchler (2007) aponta que a convivência do contribuinte com outros indivíduos pode exercer grande influência sobre sua postura de conformidade.

Se o contribuinte convive com pessoas propensas à sonegação, poderá ser influenciado à *não conformidade*. Se, por outro lado, se relacionar majoritariamente com pessoas que cumprem suas obrigações tributárias, a tendência à conformidade poderá ser elevada.

As fórmulas utilizadas na criação das variáveis-índices não serão exibidas para resguardar o sigilo de algumas informações, mas suas variáveis de origem estão descritas na **Figura 21**.

Figura 21: Composição das variáveis de índices que representam fatores sociológicos.



Fonte: Elaborado pelo autor da pesquisa.

Como é possível observar, **há pelo menos um índice relacionado a cada uma das variáveis de origem** (variáveis base). Em função disso, **supõe-se que os índices sejam suficientes para mapear as influências de tais variáveis na previsão do resultado da**

diligência. A confirmação ou não dessa suposição será efetuada na seção 2.2.3 *Modelo empírico*.

A **Tabela 16** detalha o resultado da análise estatística (aplicação do *Welch Two Sample t-test*) das variáveis *cont_idx*, *soc1_idx* e *soc2_idx*. Nela é possível observar que, das três variáveis, **todas possuem médias estatisticamente diferentes entre os grupos no intervalo de confiança de 95%.**

Tabela 16: Welch Two Sample t-test das variáveis de índices que representam fatores sociológicos contra a variável *dilg_irre*.

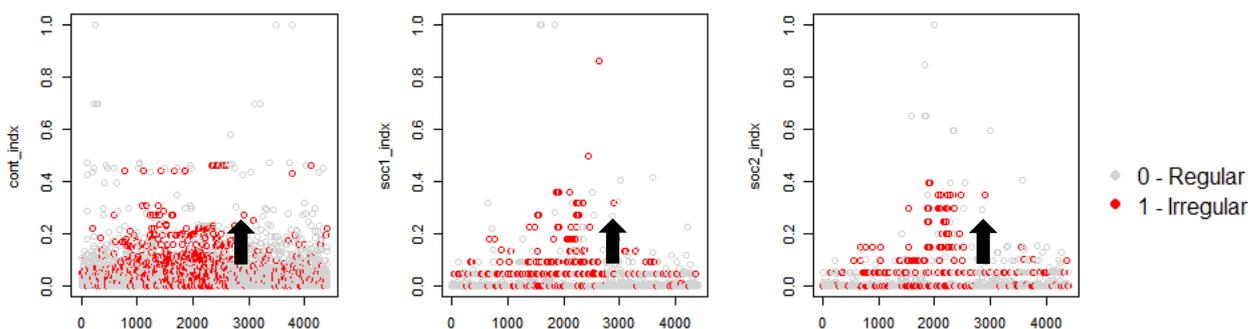
Variável	Descrição	Estatística <i>t</i>	<i>p-value</i>	Média 0 – Regular	Média 1 – Irregular	Signif. estatística
<i>cont_idx</i>	Índice contador	-5.6918	1.459e-08	0.047	0.063	‘***’
<i>soc1_idx</i>	Índice sócio 1	-5.7076	1.34e-08	0.011	0.024	‘***’
<i>soc2_idx</i>	Índice sócio 2	-5.1388	3.073e-07	0.010	0.021	‘***’

Fonte: Elaborado pelo autor da pesquisa.

Nota: Códigos de significância estatística: 0 ‘***’; 0.001 ‘**’; 0.01 ‘*’; 0.05 ‘.’; 0.1 ‘ ’; 1

A **Figura 22** mostra os gráficos de dispersão dos valores de cada variável nos grupos de referência.

Figura 22: Gráficos de dispersão dos valores das variáveis que representam fatores sociológicos nos grupos de referência.



Fonte: Dados da pesquisa.

Nota: Gráficos de dispersão das variáveis *cont_idx*, *soc1_idx* e *soc2_idx*, confrontadas com os grupos de referência (0 – Regular e 1 – Irregular) da variável *dilg_irre*. O eixo y, das ordenadas, representa a escala de valores de cada variável. O eixo x, das abcissas, representa cada indivíduo da amostra. As setas ↑ e ↓ indicam qual seria o comportamento não conforme no âmbito da variável, independentemente da distribuição nos grupos de referência. Elas permitem verificar se o comportamento não conforme no âmbito da variável é confirmado pela distribuição dos indivíduos nos grupos de referência.

Nas seções seguintes, cada gráfico da **Figura 22** será analisado e hipóteses de influência de cada variável na previsão do resultado da diligência serão discutidas.

2.2.2.5.1 Variável índice contador

A variável **cont_indx** – **Índice contador** foi criada para capturar os efeitos das variáveis **cont_empr** – **Empresas contador** e **cont_empi** – **Empresas irregulares contador**, conforme mostra a **Figura 21**.

De acordo com a fórmula utilizada para construção dessa variável, tem-se que **o aumento no valor do índice pode indicar não conformidade**.

O primeiro gráfico da **Figura 22** mostra que indivíduos do grupo *1 – Irregular* possuem valores maiores no índice. Adicionalmente, o resultado do **Welch Two Sample t-test** na **Tabela 16** mostrou que as médias dos valores da variável entre os grupos de referência são estatisticamente diferentes e que a média do grupo *1 – Irregular* (0.063) é maior do que a do grupo *0 – Regular* (0.047).

As três informações em conjunto:

1. Formulação intrínseca à variável indicando que o aumento do valor do índice pode indicar irregularidade.
2. Diferença nas médias, sendo a média do grupo *1 – Irregular* maior.
3. Demonstração visual de que indivíduos classificados como *1 – Irregular* possuem valores maiores no índice.

São possíveis evidências de que essa variável **deverá ser relevante para prever o resultado da diligência**.

2.2.2.5.2 Variável índice sócio 1

A variável **soc1_indx** – **Índice sócio 1** foi criada para capturar os efeitos das variáveis **soc1_empr** – **Empresas sócio 1** e **soc1_empi** – **Empresas irregulares sócio 1**, conforme mostra a **Figura 21**.

O sócio 1 é o primeiro sócio com maior participação no capital social da empresa. Nos casos em que mais de um sócio possuem a maior participação no capital social, foi priorizado aquele com funções de administração da empresa.

De acordo com a fórmula utilizada para construção dessa variável, tem-se que **o aumento no valor do índice pode indicar não conformidade**.

O segundo gráfico da **Figura 22** mostra que indivíduos do grupo *1 – Irregular* possuem valores maiores no índice. Adicionalmente, o resultado do **Welch Two Sample t-test** na **Tabela 16** mostrou que as médias dos valores da variável entre os grupos de referência são estatisticamente diferentes e que a média do grupo *1 – Irregular* (0.024) é maior do que a do grupo *0 – Regular* (0.011).

As três informações em conjunto:

1. Formulação intrínseca à variável indicando que o aumento do valor do índice pode indicar irregularidade.
2. Diferença nas médias, sendo a média do grupo *1 – Irregular* maior.
3. Demonstração visual de que indivíduos classificados como *1 – Irregular* possuem valores maiores no índice.

São possíveis evidências de que essa variável **deverá ser relevante para prever o resultado da diligência.**

2.2.2.5.3 Variável índice sócio 2

A variável **soc2_idx – Índice sócio 2** foi criada para capturar os efeitos das variáveis **soc2_empr – Empresas sócio 2** e **soc2_empi – Empresas irregulares sócio 2**, conforme mostra a **Figura 21**.

O sócio 2 é o segundo sócio com maior participação no capital social da empresa. Nos casos em que mais de um sócio possuem a segunda maior participação no capital social, foi priorizado aquele com funções de administração da empresa.

De acordo com a fórmula utilizada para construção dessa variável, tem-se que **o aumento no valor do índice pode indicar não conformidade.**

O terceiro gráfico da **Figura 22** mostra que indivíduos do grupo *1 – Irregular* possuem valores maiores no índice. Adicionalmente, o resultado do **Welch Two Sample t-test** na **Tabela 16** mostrou que as médias dos valores da variável entre os grupos de referência são estatisticamente diferentes e que a média do grupo *1 – Irregular* (0.021) é maior do que a do grupo *0 – Regular* (0.010).

As três informações em conjunto:

1. Formulação intrínseca à variável indicando que o aumento do valor do índice pode indicar irregularidade.
2. Diferença nas médias, sendo a média do grupo *1 – Irregular* maior.
3. Demonstração visual de que indivíduos classificados como *1 – Irregular* possuem valores maiores no índice.

São possíveis evidências de que essa variável **deverá ser relevante para prever o resultado da diligência.**

2.2.2.6 Síntese das hipóteses de influência no evento de interesse

A **Tabela 17** demonstra uma síntese das hipóteses de influências das variáveis elencadas nas seções anteriores em relação à predição do resultado da diligência.

Tabela 17: Síntese das hipóteses de influência no resultado da diligência

Grupo	Variável	Descrição	Influência esperada?	Direção da influência	
Fatores econômico-financeiros	decl_ind1	Índice de declarações	Sim	Distanciamento de 1.0 em direção a 0.0	↓
	decl_ind2	Índice de crédito/débito das declarações	Sim	Distanciamento de -1.0	↑
	decl_docf	Índice de ausência de lastro de créditos	Sim	Distanciamento de 0.0 em direção a -1.0	↓
	docf_decl	Índice de ausência de lançamento de débitos	Sim	Distanciamento de 0.0 em direção a 1.0	↑
	docf_ind1	Índice de lançamento de crédito	Não	-	
	docf_ind2	Índice de lançamento de débito	Não	-	
	docf_ind3	Índice de ausência de lastro de saídas	Sim	Distanciamento de 0.0 em direção a 1.0	↑
Fatores psicológicos	dilg_insc	Tempo entre inscrição e diligência	Possível	-	
	qtde_fisc	Quantidade de fiscalizações	Possível	-	
	qtde_auto	Quantidade de autos de infração	Possível	-	
Fatores de ambiente	cad_t_natj	Natureza jurídica do estabelecimento	Possível	-	
	cad_t_port	Porte Empresarial	Possível	-	
	cad_t_cnae	Código de atividade econômica principal	Possível	-	
	cad_t_subt	Substituto Tributário?	Possível	-	
Fatores sociológicos	cad_t_delg	Delegacia Regional Tributária de jurisdição	Possível	-	
	cont_indx	Índice de irregularidade contador	Sim	Aumento no índice	↑
	soc1_indx	Índice de irregularidade sócio 1	Sim	Aumento no índice	↑
	soc2_indx	Índice de irregularidade sócio 2	Sim	Aumento no índice	↑

Fonte: Elaborado pelo autor da pesquisa.

2.2.3 Modelo empírico

Para o tratamento dos dados, foi criado um modelo estatístico utilizando-se a **técnica de Análise de Dados Multivariados denominada Regressão Logística – Logit**.

De acordo com Hosmer e Lemeshow (2000), trata-se de uma **técnica de Aprendizado de Máquina (*Machine Learning*)** e foi escolhida para a pesquisa porque é, segundo os autores, capaz de criar modelos extremamente flexíveis e que oferecem um ótimo ajuste frente às mais diversas técnicas de **Aprendizado de Máquina e Inteligência Artificial**, como Árvores de Decisão (*Decision Tree*), Floresta Aleatória (*Random Forest*), Máquina de Vetores de Suporte (*Support Vector Machine*), Redes Neurais Artificiais (*Artificial Neural Networks*) e *Boosting*.

Barboza, Kimura e Altman (2017) argumentam que a Regressão Logística é superior a técnicas de Análise de Dados Multivariados, como Análise Discriminante, por sua maior flexibilidade em termos de pressupostos e também pelo fato de produzir resultados no intervalo entre 0 e 1, que representam a probabilidade de ocorrência do evento de interesse.

A aplicação da Regressão Logística **objetiva encontrar relações de dependência entre uma variável dependente e outras variáveis explicativas (independentes)**.

Fávero (2009) aponta que o método *Logit* é utilizado para encontrar relações de dependência quando a variável dependente for qualitativa (não contínua) e as variáveis independentes forem quantitativas (métricas) ou qualitativas (não contínuas). Segundo o autor, a função logística assume valores entre 0 e 1 para qualquer Z entre $-\infty$ e $+\infty$, o que a permite **prever** não apenas a **possibilidade** de ocorrência de determinado evento de interesse, mas também a **probabilidade** de sua ocorrência.

Nessa esteira, Lattin, Carroll e Green (2011) descrevem o método *Logit* como uma técnica multivariada de dependência destinada a **identificar as variáveis mais significativas para previsão de ocorrência de determinado evento de interesse, provendo a probabilidade de sua ocorrência**. No estudo em questão, o **evento de interesse é a ocorrência de irregularidade tributária além da simples inatividade**.

No tocante aos tipos do método *Logit*, foi utilizado o Binomial (**Binário**). Fávero (2009) ensina que o *Logit* pode ser Binomial ou Multinomial. No caso Binomial, a **variável dependente** é uma *dummy*, ou seja, uma variável dicotômica, que **somente pode assumir dois valores**. Na Logística Multinomial, a variável dependente pode assumir mais de dois valores.

A fórmula matemática para a Regressão Logística Binomial do presente estudo é a seguinte.

$$\text{logit}(p) = \ln \left[\frac{P(\text{Empresa em situação irregular} = 1)}{1 - P(\text{Empresa em situação irregular} = 1)} \right]$$

Em que $P(\text{Empresa em situação irregular}=1)$ indica a probabilidade de ocorrência do evento de interesse, qual seja, a existência de possível irregularidade tributária além da simples inatividade e $1 - P(\text{Empresa em situação irregular}=1)$ indica a probabilidade de não ocorrência desse evento.

Ferreira, Celso e Barbosa Neto (2012) apontam que a variável dependente da equação é o logaritmo relacionado à probabilidade de ocorrência de determinado evento. Os autores apontam, ainda, que em modelos com variáveis contínuas relacionadas a atributos que servem como variáveis explicativas, é necessário estimar um modelo *Logit* no qual somente uma alternativa esteja associada a cada conjunto de valores assumidos pelas variáveis independentes. A **estimação dos parâmetros** é feita pelo **método de máxima verossimilhança**, que busca maximizar a probabilidade (verossimilhança) de que um evento (o evento de interesse) ocorra.

O modelo *Logit Binomial* pressupõe uma série de variáveis explicativas independentes, que são utilizadas para se prever a ocorrência do evento de interesse refletido na variável dependente.

A variável dependente, no caso deste estudo, é **dilg_irre**, e conforme detalhado na **Tabela 5**, é binária ao assumir dois grupos de referência de resultados da **diligência fiscal (0 – Regular e 1 – Irregular)**. O **evento de interesse** é o retratado pelo grupo **1 – Irregular**.

As variáveis explicativas, por sua vez, são as características que buscam explicar e refletir os fatores que determinam o perfil de cumprimento no âmbito da amostra de pesquisa e foram retratadas nas **Tabelas 6, 7, 8 e 9**.

É possível, no entanto, que nem todas as variáveis explicativas coletadas sejam relevantes para o modelo. Para definir quais são relevantes, será utilizada a abordagem sugerida por Collett (2003) e Hosmer e Lemeshow (2000), que visa buscar a construção mais parcimoniosa possível entre aquelas capazes de explicar os dados.

Isso significa minimizar o número de variáveis explicativas tanto quanto possível, ou seja, incluir apenas aquelas que realmente contribuam com os resultados do modelo.

A sugestão dos autores decorre do fato de que modelos com menos variáveis são mais estáveis e, portanto, mais generalizáveis. O aumento do número de variáveis aumenta o

desvio padrão e torna o modelo mais dependente dos dados observados, diminuindo seu potencial de generalização.

Collett (2003) e Hosmer e Lemeshow (2000) apontam, ainda, que a escolha das variáveis não deve ser feita unicamente com base em resultados estatísticos. O conhecimento do pesquisador da realidade que os dados retratam pode levar à escolha de variáveis que não necessariamente possuem indicação estatística para compor o modelo final.

Isso decorre do fato de que a realidade refletida pelos dados não se resume apenas à amostra, que normalmente é a base de dados disponível. Por isso, é possível que uma variável sem indicação estatística para compor o modelo seja de grande valia para garantir seu poder explicativo e de previsão quando o modelo é aplicado em toda a população, ou até mesmo, generalizado para outras bases de dados.

2.2.3.1 Procedimento para seleção de variáveis

Para a condução da análise estatística na seleção de variáveis, foi necessário determinar a medida a ser utilizada e o nível de significância de referência. No caso da Regressão Logística, Fávero (2009) explica que a medida geral do ajuste do modelo é dada pela verossimilhança.

Também chamada de desvio residual, a fórmula dessa medida é *-2 vezes o logaritmo do valor da verossimilhança* e é chamada de **-2LL**. Quanto menor o valor de **-2LL**, melhor, pois quando a verossimilhança for igual a 1, indicando o hipotético ajuste perfeito em que o modelo explica totalmente a realidade, o valor de **-2LL** será zero.

Quanto ao nível de significância, Bendel e Afif (1977), em estudo sobre Regressão Linear, e Mickey e Greenland (1987), em estudo sobre Regressão Logística, demonstraram que o uso de valores de significância tradicionais, como 0.05, comumente ocasiona falhas na identificação de variáveis importantes. Por outro lado, o uso de valores maiores, como 0.25, pode causar o efeito de incluir no modelo variáveis de importância questionável. Por esse motivo, foi adotado neste trabalho o nível de significância de 0.10 (90% de confiança) para seleção das variáveis.

Quanto ao método de seleção das variáveis, Collett (2003) argumenta que rotinas automatizadas, como *forward selection*, *backward elimination* ou a combinação das duas, conhecida como *stepwise procedure*, podem parecer uma boa alternativa em caso de construção de modelos com grande número de variáveis explicativas. O problema de tais

métodos, segundo o autor, é que eles normalmente levam à identificação de um modelo específico de variáveis, ao invés de um conjunto com diversos modelos igualmente bons.

Além disso, os resultados podem variar grandemente dependendo das rotinas automatizadas adotadas e estas não levam em conta o conhecimento do pesquisador a respeito da realidade que os dados retratam.

Em função das desvantagens das ferramentas automatizadas, a seleção de variáveis neste estudo foi feita com base em seis passos não automatizados, adaptados das recomendações de Collett (2003) e Hosmer e Lemeshow (2000).

No *Passo 1* cada variável é analisada individualmente contra o modelo de referência. O modelo de referência neste passo é aquele que não contém nenhuma variável explicativa, mas apenas a variável dependente. As variáveis que geram modelos com $p\text{-value} < 0.1$ passam a compor o modelo de referência do passo seguinte.

No *Passo 2*, cria-se um modelo de referência com as variáveis retidas no *Passo 1* e ajustam-se modelos menores, retirando-se uma variável a uma para verificar se o modelo piora sem ela. Se o modelo piorar ($p\text{-value} < 0.1$) ou se, mesmo não piorando, o pesquisador considerar a variável importante para o modelo, ela é mantida, caso contrário é retirada.

No *Passo 3*, cria-se um modelo de referência com as variáveis retidas no *Passo 2*. Uma a uma, cada variável descartada no *Passo 2* é retornada para confirmar se não é mesmo estatisticamente significativa. Se $p\text{-value}$ for < 0.1 , a variável volta ao modelo.

No *Passo 4*, cria-se um modelo de referência com as variáveis retidas no *Passo 3*. As variáveis descartadas no *Passo 1* são retornadas para que se confirme se não são mesmo estatisticamente significantes. Caso sejam, devem retornar ao modelo.

No *Passo 5*, cria-se um modelo de referência com as variáveis retidas no *Passo 4*. Uma a uma, cada variável é retirada a fim de se verificar se o modelo piora sem ela. As variáveis que retiradas do modelo o fazem piorar estatisticamente e aquelas consideradas de relevância para a pesquisa permanecem no modelo. As demais são retiradas. Neste passo, caso uma variável seja mantida em função de seu resultado estatístico, mas seja considerada menos relevante pelo pesquisador por estar representada por outra variável, esta pode ser eliminada do modelo.

No *Passo 6*, cria-se um modelo de referência com as variáveis retidas no *Passo 5*. Neste passo, são incluídas interações entre essas variáveis para ver se melhoram o modelo ($p\text{-value} < 0.1$). Caso melhorem, têm potencial de permanecer no modelo final a depender da análise do pesquisador, caso contrário, são desconsideradas.

Os resultados dos seis passos são detalhados a seguir. Em cada um foram analisadas informações estatísticas e a importância percebida da variável pelo pesquisador frente ao contexto que os dados refletem.

2.2.3.1.1 Passos

A *Tabela 18* mostra o resultado do *Passo 1*.

Tabela 18: Resultado do Passo 1

Modelo	-2LL	p-value	Significativo a 0.1?
Referência: NULL = sem variável explicativa	4983,1568	N/A	N/A
docf_decl	3900,3599	1,81E-237	Sim
cont_indx	4952,3535	2,86E-08	Sim
decl_ind1	4487,5873	8,75E-110	Sim
decl_ind2	4939,0627	3,13E-11	Sim
decl_docf	4888,5598	2,33E-22	Sim
docf_ind3	4396,3442	1,24E-129	Sim
soc1_indx	4952,1821	2,61E-08	Sim
soc2_indx	4957,831	4,84E-07	Sim
decl_cred	4959,4872	1,14E-06	Sim
decl_debi	4946,7187	1,58E-09	Sim
docf_cred	4981,7933	0,242931674	Não
docf_debi	4896,1039	1,06E-20	Sim
docf_entd	4982,7579	0,52767653	Não
docf_said	4981,0729	0,148858495	Não
docf_ind1	4983,1355	0,884030248	Não
docf_ind2	4835,8286	6,65E-34	Sim
decl_entg	4462,1716	2,58E-115	Sim
decl_movi	4600,0965	2,68E-85	Sim
cont_empr	4910,0074	1,20E-17	Sim
cont_empi	4878,8028	1,69E-24	Sim
soc1_empr	4831,1881	6,44E-35	Sim
soc1_empi	4939,0032	3,04E-11	Sim
soc2_empr	4862,7138	5,06E-28	Sim
soc2_empi	4946,8866	1,72E-09	Sim
qtde_fisc	4923,2978	1,02E-14	Sim
qtde_auto	4975,2788	0,005004077	Sim
dilg_insc	4661,1996	5,43E-72	Sim
cad_t_natj	4816,1199	2,55E-31	Sim
cad_t_port	4903,2646	4,48E-18	Sim
cad_t_cnae	4831,9226	1,40E-23	Sim
cad_t_subt	4970,1834	0,000315942	Sim
cad_t_delg	4520,8260	1,59E-87	Sim

Fonte: Elaborado pelo autor da pesquisa.

As variáveis destacadas em cinza e letra negrito não tiveram sua significância estatística confirmada e, por isso, não serão retidas no *Passo 1*. As demais tiveram e serão mantidas.

A *Tabela 19* mostra o resultado do *Passo 2*.

Tabela 19: Resultado do Passo 2

Modelo	-2LL	<i>p-value</i>	Significativo a 0.1?
Referência:	2909,4778	N/A	N/A
docf_decl+cont_idx+decl_ind1+decl_ind2+decl_docf+docf_ind3+soc1_idx+soc2_idx+decl_cred+decl_debi+docf_debi+docf_ind2+decl_entg+decl_movi+cont_empr+cont_empi+soc1_empr+soc1_empi+soc2_empr+soc2_empi+qtde_fisc+qtde_auto+dilg_insc+cadt_natj+cadt_port+cadt_cnae+cadt_subt+cadt_delg			
Excluída: docf_decl	2970,6127	5,33E-15	Sim
Excluída: cont_idx	2916,2506	0,009255816	Sim
Excluída: decl_ind1	2909,8507	0,541450255	Não
Excluída: decl_ind2	2910,2463	0,380683127	Não
Excluída: decl_docf	2910,5666	0,296740195	Não
Excluída: docf_ind3	2939,8196	3,62E-08	Sim
Excluída: soc1_idx	2909,6940	0,642009076	Não
Excluída: soc2_idx	2914,5749	0,023966468	Sim
Excluída: decl_cred	2911,0360	0,211928583	Não
Excluída: decl_debi	2910,7069	0,267588338	Não
Excluída: docf_debi	2916,6375	0,007456246	Sim
Excluída: docf_ind2	2910,8418	0,242857653	Não
Excluída: decl_entg	2909,8507	0,541450181	Não
Excluída: decl_movi	2909,8507	0,541450324	Não
Excluída: cont_empr	2916,2506	0,009255835	Sim
Excluída: cont_empi	2916,2506	0,00925581	Sim
Excluída: soc1_empr	2909,6940	0,641998098	Não
Excluída: soc1_empi	2909,6940	0,642009255	Não
Excluída: soc2_empr	2914,5749	0,02396623	Sim
Excluída: soc2_empi	2914,5749	0,023966469	Sim
Excluída: qtde_fisc	2968,7716	1,36E-14	Sim
Excluída: qtde_auto	2940,2897	2,84E-08	Sim
Excluída: dilg_insc	2936,9097	1,63E-07	Sim
Excluída: cadt_natj	2921,8701	0,192084808	Não
Excluída: cadt_port	2911,6079	0,344713797	Não
Excluída: cadt_cnae	2947,1885	0,002692888	Sim
Excluída: cadt_subt	2909,5001	0,881281358	Não
Excluída: cadt_delg	3143,9952	2,99E-40	Sim

Fonte: Elaborado pelo autor da pesquisa.

As variáveis destacadas em cinza e letra negrito não tiveram sua significância estatística confirmada e, por isso, não serão retidas no *Passo 2*. Além delas, mais quatro variáveis (decl_ind1, decl_ind2, decl_docf, soc1_idx) foram indicadas como estatisticamente não significantes neste passo. Estas, no entanto, tiveram a importância percebida pelo

pesquisador confirmada na seção 2.2.2.6 *Síntese das hipóteses de influência no evento de interesse* e, por isso, serão mantidas.

A *Tabela 20* mostra o resultado do *Passo 3*.

Tabela 20: Resultado do Passo 3

Modelo	-2LL	p-value	Significativo a 0.1?
Referência: docf_decl+cont_indx+decl_ind1+decl_ind2+decl_docf+docf_ind3+soc1_indx+soc2_indx+docf_debi+cont_empr+cont_empi+soc2_empr+soc2_empi+qtde_fisc+qtde_auto+dilg_insc+cadt_cnae+cadt_delgcadt_delg	3000,3099	N/A	N/A
Incluída: decl_cred	3000,2951	0,903083141	Não
Incluída: decl_debi	3000,2415	0,793664814	Não
Incluída: docf_ind2	3000,3098	0,991545712	Não
Incluída: decl_entg	2955,4752	2,14E-11	Sim
Incluída: decl_movi	2955,4752	2,14E-11	Sim
Incluída: soc1_empr	2974,5174	3,80E-07	Sim
Incluída: soc1_empi	2974,5175	3,80E-07	Sim
Incluída: cadt_natj	2985,2330	0,088845035	Sim
Incluída: cadt_port	2996,8080	0,173605368	Não
Incluída: cadt_subt	3000,2008	0,741102649	Não

Fonte: Elaborado pelo autor da pesquisa.

As variáveis destacadas em cinza e letra negrito não tiveram sua significância estatística confirmada e, por isso, não serão retidas no *Passo 3*. As demais foram estatisticamente significantes e retornarão ao modelo neste passo.

A *Tabela 21* mostra o resultado do *Passo 4*.

Tabela 21: Resultado do Passo 4

Modelo	-2LL	p-value	Significativo a 0.1?
Referência: docf_decl+cont_indx+decl_ind1+decl_ind2+decl_docf+docf_ind3+soc1_indx+soc2_indx+docf_debi+qtde_fisc+qtde_auto+dilg_insc+cadt_cnae+cadt_delg+decl_entg+decl_movi+cont_empr+cont_empi+soc1_empr+soc1_empi+soc2_empr+soc2_empi+cadt_natj	2914,7517	N/A	N/A
Incluída: docf_cred	2912,1898	0,109462863	Não
Incluída: docf_entd	2914,0318	0,396157517	Não
Incluída: docf_said	2914,5327	0,639795792	Não
Incluída: docf_ind1	2914,0917	0,416540076	Não

Fonte: Elaborado pelo autor da pesquisa.

Nenhuma das variáveis excluídas no *Passo 1* foi significativa no *Passo 4* para retornar ao modelo.

A *Tabela 22* mostra o resultado do *Passo 5*.

Tabela 22: Resultado do *Passo 5*

Modelo	-2LL	<i>p-value</i>	Significativo a 0.1?
Referência:	2914,7517	N/A	N/A
docf_decl+cont_indx+decl_ind1+decl_ind2+decl_docf+docf_ind3+soc1_indx+soc2_indx+docf_debi+qtde_fisc+qtde_auto+dilg_insc+cadt_cnae+cadt_delg+decl_entg+decl_movi+cont_empr+cont_empi+soc1_empr+soc1_empi+soc2_empr+soc2_empi+cadt_natj			
Excluída: docf_decl	2981,6035	2,93E-16	Sim
Excluída: cont_indx	2921,8903	0,007544295	Sim
Excluída: decl_ind1	2915,1415	0,532428678	Não
Excluída: decl_ind2	2915,1221	0,542787928	Não
Excluída: decl_docf	2915,2468	0,481677155	Não
Excluída: docf_ind3	2945,4194	3,06E-08	Sim
Excluída: soc1_indx	2914,9328	0,670484416	Não
Excluída: soc2_indx	2919,8896	0,023409699	Sim
Excluída: docf_debi	2922,7329	0,004726579	Sim
Excluída: qtde_fisc	2975,7861	5,61E-15	Sim
Excluída: qtde_auto	2945,438	3,03E-08	Sim
Excluída: dilg_insc	2942,6192	1,30E-07	Sim
Excluída: cadt_cnae	2952,4444	0,002708254	Sim
Excluída: cadt_delg	3162,1356	7,14E-43	Sim
Excluída: decl_entg	2915,1415	0,532428526	Não
Excluída: decl_movi	2915,1415	0,532428778	Não
Excluída: cont_empr	2921,8903	0,00754431	Sim
Excluída: cont_empi	2921,8903	0,00754429	Sim
Excluída: soc1_empr	2914,9328	0,670473211	Não
Excluída: soc1_empi	2914,9328	0,670484711	Não
Excluída: soc2_empr	2919,8896	0,023409474	Sim
Excluída: soc2_empi	2919,8896	0,023409701	Sim
Excluída: cadt_natj	2928,7467	0,122506337	Não

Fonte: Elaborado pelo autor da pesquisa.

As variáveis destacadas em cinza e letra negrito não tiveram sua significância estatística confirmada e, por isso, não serão retidas no *Passo 5*. Além delas, há cinco variáveis com marca cinza e letra normal (docf_debi, cont_empr, cont_empi, soc2_empr, soc2_empi) que tiveram significância estatística confirmada, mas que serão excluídas do modelo por já terem seus valores refletidos em outras variáveis, como é possível observar nas *Figuras 16 e 21*.

Por fim, três variáveis com letra em negrito e sem marca cinza (decl_ind1, decl_ind2, decl_docf) foram indicadas como estatisticamente não significantes neste passo. Estas, no entanto, tiveram a importância esperada pelo pesquisador confirmada na seção 2.2.2.6 *Síntese das hipóteses de influência no evento de interesse* e, por isso, serão mantidas.

A *Tabela 23* mostra o resultado do *Passo 6*.

Tabela 23: Resultado do *Passo 6*

Modelo	-2LL	<i>p-value</i>	Significativo a 0.1?
Referência: docf_decl+cont_idx+decl_ind1+decl_ind2+decl_docf+docf_ind3+soc1_idx+soc2_idx+qtde_fisc+qtde_auto+dilg_insc+cadt_cnae+cadt_delg	3122,0182	N/A	N/A
Interação: docf_decl*decl_ind1	3095,9615	3,32E-07	Sim
Interação: soc1_idx*soc2_idx	3101,9580	7,50E-06	Sim
Interação: soc2_idx*cadt_delg	3068,0045	9,86E-06	Sim
Interação: qtde_auto*cadt_delg	3069,5320	1,72E-05	Sim
Interação: soc1_idx*cadt_delg	3071,2832	3,25E-05	Sim
Interação: qtde_fisc*cadt_delg	3072,1446	4,42E-05	Sim
Interação: dilg_insc*cadt_delg	3073,2827	6,62E-05	Sim
Interação: qtde_fisc*qtde_auto	3108,5715	0,000245434	Sim
Interação: decl_ind1*decl_docf	3109,7144	0,000452033	Sim
Interação: decl_ind2*docf_ind3	3110,8518	0,000832893	Sim
Interação: decl_docf*soc1_idx	3111,5053	0,001185396	Sim
Interação: decl_ind1*dilg_insc	3112,2703	0,001795229	Sim
Interação: cont_idx*decl_ind1	3113,1409	0,002887344	Sim
Interação: docf_decl*decl_ind2	3113,4068	0,003340522	Sim
Interação: decl_ind1*soc1_idx	3113,8657	0,004300009	Sim
Interação: cont_idx*qtde_auto	3114,2651	0,005362051	Sim
Interação: docf_decl*cadt_delg	3086,9280	0,006053572	Sim
Interação: decl_ind1*cadt_delg	3087,0215	0,006227444	Sim
Interação: decl_ind1*cadt_cnae	3090,7322	0,008042116	Sim
Interação: decl_ind2*cadt_cnae	3093,7439	0,008284981	Sim
Interação: decl_ind2*soc2_idx	3115,2819	0,009447117	Sim
Interação: decl_ind2*dilg_insc	3115,5163	0,010775675	Sim
Interação: soc2_idx*qtde_fisc	3115,5367	0,010900046	Sim
Interação: docf_ind3*cadt_delg	3089,2137	0,011932017	Sim
Interação: cont_idx*qtde_fisc	3115,8603	0,013082368	Sim
Interação: decl_docf*cadt_delg	3090,7548	0,018547699	Sim
Interação: decl_ind1*decl_ind2	3116,7835	0,022140736	Sim
Interação: cont_idx*cadt_delg	3092,5994	0,030853655	Sim
Interação: decl_ind2*soc1_idx	3117,4528	0,032622919	Sim
Interação: decl_ind1*qtde_fisc	3117,5510	0,034550569	Sim
Interação: soc2_idx*qtde_auto	3117,8300	0,040705848	Sim
Interação: soc1_idx*dilg_insc	3118,1431	0,049006847	Sim
Interação: qtde_auto*cadt_cnae	3104,0476	0,055461462	Sim
Interação: decl_ind2*qtde_fisc	3118,4968	0,060579154	Sim
Interação: docf_decl*decl_docf	3118,6133	0,065000096	Sim
Interação: docf_decl*dilg_insc	3118,7586	0,071005996	Sim

Fonte: Elaborado pelo autor da pesquisa.

Das 13 variáveis, foram produzidas 78 interações. A *Tabela 23* está ordenada de forma ascendente pelo *p-value* e mostra as interações que tiveram significância estatística a 0.1.

As interações destacadas em cinza e letra negrito foram incluídas na pesquisa. Considerando a recomendação de Hosmer e Lemeshow (2000), de que a inclusão de variáveis no modelo deve ser feita com parcimônia para não trazer instabilidade a ele, optou-se por

incluir apenas as interações com o maior nível de significância estatística. O limite adotado foi o número de variáveis do modelo (treze) dividido por dois, o que acarretou em sete interações após arredondamento.

A interação **dilg_insc*cadt_delg** seria incluída por esse critério, mas o teste Hosmer-Lemeshow, que será apresentado adiante, teve um resultado pior com ela, indicando que sua presença no modelo teria o potencial de piorar seu ajuste e, por isso, ela foi descartada e a interação seguinte (**qtde_fisc*qtde_auto**) incluída.

2.2.3.2 O modelo estatístico

Após a execução dos passos de seleção de variáveis, criou-se o modelo com as variáveis e interações selecionadas.

A **Tabela 24** lista as variáveis e as interações selecionadas para o modelo.

Tabela 24: Variáveis e interações selecionadas

Variáveis	Interações
docf_decl	docf_decl*decl_ind1
cont_indx	soc1_indx*soc2_indx
decl_ind1	soc2_indx*cadt_delg
decl_ind2	qtde_auto*cadt_delg
decl_docf	soc1_indx*cadt_delg
docf_ind3	qtde_fisc*cadt_delg
soc1_indx	qtde_fisc*qtde_auto
soc2_indx	
qtde_fisc	
qtde_auto	
dilg_insc	
cadt_cnae	
cadt_delg	

Fonte: Elaborado pelo autor da pesquisa.

O modelo *Logit* foi criado no R através da função **glm** do pacote **stats** (R CORE TEAM, 2018). A chamada a essa função requer pelo menos três argumentos:

- A fórmula da função, que indica qual é a variável dependente (**dilg_irre**) e quais as variáveis explicativas (incluindo as interações). A função detectará automaticamente o evento de interesse, que no caso em questão é o valor 1 (Irregularidade) da variável **dilg_irre**.

- O tipo de método *Logit* empregado, o qual foi informado como binomial ao parâmetro *family*.
- Os dados base para o modelo, que no estudo em questão são compostos da amostra de pesquisa com as 4.407 observações.

A chamada à função **glm** com os respectivos parâmetros pode ser visualizada no **Quadro 3**.

Quadro 3: Fórmula de criação do modelo

```
Call: glm(

formula = dilg_irre ~
docf_decl + cont_indx + decl_ind1 + decl_ind2 + decl_docf + docf_ind3 +
soc1_indx + soc2_indx + qtde_fisc + qtde_auto + dilg_insc + cadt_cnae +
cadt_delg +
docf_decl * decl_ind1 + soc1_indx * soc2_indx + soc2_indx * cadt_delg +
qtde_auto * cadt_delg + soc1_indx * cadt_delg + qtde_fisc * cadt_delg +
qtde_fisc * qtde_auto,

family = binomial,
data = amostra)
```

Fonte: Elaborado pelo autor da pesquisa.

O **Quadro 4** mostra o resultado completo do modelo criado. Nele é possível perceber os coeficientes gerados. Estes são utilizados para o cálculo da probabilidade de ocorrência do evento de interesse. O (Intercept) representa a variável dependente (**dilg_irre**). As variáveis contínuas aparecem uma única vez, com seus respectivos coeficientes. As variáveis categóricas são expandidas para cada valor existente na amostra. Como exemplo, percebe-se que a variável *cadt_cnae* foi acrescida dos **ID's** que aparecem na **Tabela 14**.

Ao final da criação do modelo são apresentadas algumas informações de cunho geral, como *Null Deviance* e *Residual Deviance*, que serão discutidas na seção 2.2.4.2 *Qualidade do ajuste do modelo*.

Quadro 4: O modelo criado

```

Call: glm(formula = dilg_irre ~ docf_decl + cont_indx + decl_ind1
+decl_ind2 + decl_docf + docf_ind3 + soc1_indx + soc2_indx +
qtde_fisc + qtde_auto + dilg_insc + cadt_cnae + cadt_delg +
docf_decl * decl_ind1 + soc1_indx * soc2_indx + soc2_indx *
cadt_delg + qtde_auto * cadt_delg + soc1_indx * cadt_delg +
qtde_fisc * cadt_delg + qtde_fisc * qtde_auto,
family = binomial, data = amostra)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3747  -0.5193  -0.2256   0.1613   3.1435

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.419e+00  1.125e+00  -2.151  0.031511 *
docf_decl    1.685e+00  1.645e-01  10.243 < 2e-16 ***
cont_indx    8.104e-01  5.895e-01   1.375  0.169254
decl_ind1   -3.230e+00  3.118e-01 -10.362 < 2e-16 ***
decl_ind2    1.157e-01  2.562e-01   0.452  0.651463
decl_docf    4.672e-02  1.281e-01   0.365  0.715345
docf_ind3    6.798e-01  1.163e-01   5.847  5.02e-09 ***
soc1_indx    9.201e+00  5.864e+00   1.569  0.116589
soc2_indx    1.799e+00  5.305e+00   0.339  0.734497
qtde_fisc    4.550e-01  1.094e-01   4.159  3.20e-05 ***
qtde_auto    1.286e+00  4.389e-01   2.931  0.003382 **
dilg_insc   -3.610e-03  5.914e-04  -6.105  1.03e-09 ***
cadt_cnae2    2.770e+00  1.629e+00   1.700  0.089084 .
cadt_cnae3    6.609e-01  1.120e+00   0.590  0.555165
cadt_cnae4   -1.683e+01  2.795e+03  -0.006  0.995196
cadt_cnae5    1.508e+00  1.298e+00   1.162  0.245156
cadt_cnae6    1.089e-01  1.165e+00   0.093  0.925534
cadt_cnae7    7.302e-01  1.114e+00   0.656  0.512138
cadt_cnae8    1.100e+00  1.136e+00   0.969  0.332775
cadt_cnae9   -1.321e+00  1.345e+00  -0.982  0.326015
cadt_cnae10   3.804e-01  1.297e+00   0.293  0.769293
cadt_cnae11  -1.577e+01  1.378e+03  -0.011  0.990866
cadt_cnae12  -1.433e+01  1.322e+03  -0.011  0.991347
cadt_cnae13   3.023e-01  1.175e+00   0.257  0.796874
cadt_cnae14   6.380e-01  1.169e+00   0.546  0.585309
cadt_cnae17   1.833e+00  1.466e+00   1.251  0.211018
cadt_cnae18  -1.601e+01  1.626e+03  -0.010  0.992141
cadt_cnae19  -1.548e+01  9.984e+02  -0.016  0.987628
cadt_delg2    9.969e-01  2.694e-01   3.700  0.000216 ***
cadt_delg3    1.401e+00  2.601e-01   5.385  7.23e-08 ***
cadt_delg4    3.394e-01  5.432e-01   0.625  0.532116
cadt_delg5   -4.618e-01  6.348e-01  -0.727  0.466935
cadt_delg6   -5.212e-01  3.060e-01  -1.703  0.088550 .
cadt_delg7    8.907e-02  2.160e-01   0.412  0.680063
cadt_delg8   -4.756e-01  4.171e-01  -1.140  0.254254
cadt_delg9   -1.817e+00  6.121e-01  -2.968  0.002996 **
cadt_delg10  -1.282e+00  5.407e-01  -2.371  0.017756 *
cadt_delg11  -1.293e+00  5.923e-01  -2.183  0.029003 *
cadt_delg12  -9.179e-01  6.557e-01  -1.400  0.161581
cadt_delg13  -6.543e-02  4.136e-01  -0.158  0.874317
cadt_delg14   4.781e-01  4.394e-01   1.088  0.276528
cadt_delg15   1.038e+00  3.998e-01   2.596  0.009431 **
cadt_delg16   1.427e-01  2.560e-01   0.557  0.577239
cadt_delg17  -1.356e-01  3.842e-01  -0.353  0.724031
cadt_delg18  -1.133e+00  5.015e-01  -2.259  0.023853 *
docf_decl:decl_ind1 -2.253e+00  5.191e-01  -4.340  1.42e-05 ***
soc1_indx:soc2_indx -6.439e+00  3.959e+00  -1.626  0.103846
soc2_indx:cadt_delg2 -1.039e+00  1.285e+01  -0.081  0.935566
soc2_indx:cadt_delg3  1.096e+01  8.769e+00   1.250  0.211317
soc2_indx:cadt_delg4  -8.763e+00  1.839e+01  -0.476  0.633750
soc2_indx:cadt_delg5  -1.077e+03  7.109e+02  -1.516  0.129629

soc2_indx:cadt_delg6  -2.559e+01  1.803e+01  -1.419  0.155891
soc2_indx:cadt_delg7  -5.529e+00  9.050e+00  -0.611  0.541262
soc2_indx:cadt_delg8  -5.725e+00  1.542e+01  -0.371  0.710434
soc2_indx:cadt_delg9  -2.181e+02  4.667e+02  -0.467  0.640302
soc2_indx:cadt_delg10 -2.319e+02  4.034e+02  -0.575  0.565355
soc2_indx:cadt_delg11 -4.579e+02  1.611e+03  -0.284  0.776190
soc2_indx:cadt_delg12 -5.400e+00  4.578e+01  -0.118  0.906090
soc2_indx:cadt_delg13  9.643e+01  1.497e+02  0.644  0.519516
soc2_indx:cadt_delg14  2.481e+01  1.861e+01  1.333  0.182459
soc2_indx:cadt_delg15  2.284e+01  3.126e+01  0.731  0.464942
soc2_indx:cadt_delg16 -1.095e+00  5.412e+00  -0.202  0.839585
soc2_indx:cadt_delg17  2.786e+01  1.324e+01  2.105  0.035311 *
soc2_indx:cadt_delg18 -2.993e+00  7.747e+00  -0.386  0.699217
qtde_auto:cadt_delg2  -1.587e+00  8.555e-01  -1.855  0.063575 .
qtde_auto:cadt_delg3  -9.175e-01  5.688e-01  -1.613  0.106776
qtde_auto:cadt_delg4  -1.269e-01  1.244e+00  -0.102  0.918756
qtde_auto:cadt_delg5  -1.504e+01  6.406e+02  -0.023  0.981273
qtde_auto:cadt_delg6  -1.079e+00  7.342e-01  -1.470  0.141463
qtde_auto:cadt_delg7  -1.031e+00  5.804e-01  -1.776  0.075727 .
qtde_auto:cadt_delg8  -1.521e+01  9.922e+02  -0.015  0.987769
qtde_auto:cadt_delg9  -1.468e+01  1.326e+03  -0.014  0.988699
qtde_auto:cadt_delg10 -1.512e+01  1.082e+03  -0.014  0.988848
qtde_auto:cadt_delg11 -1.263e+01  7.434e+02  -0.017  0.986442
qtde_auto:cadt_delg12 -1.382e+01  5.024e+02  -0.028  0.978051
qtde_auto:cadt_delg13 -1.092e+00  1.078e+00  -1.014  0.310733
qtde_auto:cadt_delg14 -1.479e+01  8.539e+02  -0.017  0.986178
qtde_auto:cadt_delg15 -2.404e+00  7.911e-01  -3.039  0.002373 **
qtde_auto:cadt_delg16 -1.761e+00  7.314e-01  -2.407  0.016067 *
qtde_auto:cadt_delg17 -2.096e-01  7.940e-01  -0.264  0.791795
qtde_auto:cadt_delg18 -1.495e+01  8.254e+02  -0.018  0.985552
soc1_indx:cadt_delg2  -1.384e+01  1.004e+01  -1.379  0.167798
soc1_indx:cadt_delg3  -2.209e+00  7.083e+00  -0.312  0.755095
soc1_indx:cadt_delg4  -4.173e+01  2.281e+01  -1.829  0.067357 .
soc1_indx:cadt_delg5  5.499e+01  2.443e+01  2.251  0.024377 *
soc1_indx:cadt_delg6  3.073e+01  1.450e+01  2.119  0.034114 *
soc1_indx:cadt_delg7  -3.071e+00  7.069e+00  -0.434  0.663988
soc1_indx:cadt_delg8  9.080e+00  1.165e+01  0.779  0.435842
soc1_indx:cadt_delg9  3.262e+01  3.470e+01  0.940  0.347178
soc1_indx:cadt_delg10 -1.532e+01  2.066e+01  -0.742  0.458340
soc1_indx:cadt_delg11 -5.119e+01  1.313e+02  -0.390  0.696548
soc1_indx:cadt_delg12  1.934e+01  2.028e+01  0.954  0.340267
soc1_indx:cadt_delg13  1.204e+01  2.371e+01  0.508  0.611427
soc1_indx:cadt_delg14 -2.143e+01  1.752e+01  -1.223  0.221414
soc1_indx:cadt_delg15  2.667e+00  2.001e+01  0.133  0.893958
soc1_indx:cadt_delg16 -7.196e+00  6.198e+00  -1.161  0.245686
soc1_indx:cadt_delg17  5.702e+00  1.403e+01  0.406  0.684528
soc1_indx:cadt_delg18 -1.563e+01  1.292e+01  -1.210  0.226334
qtde_fisc:cadt_delg2  7.321e-02  1.597e-01  0.458  0.646619
qtde_fisc:cadt_delg3  -2.086e-02  1.706e-01  -0.122  0.902668
qtde_fisc:cadt_delg4  5.953e-01  3.259e-01  1.826  0.067793 .
qtde_fisc:cadt_delg5  3.650e-02  6.043e-01  0.060  0.951842
qtde_fisc:cadt_delg6  -1.062e-01  1.884e-01  -0.564  0.572914
qtde_fisc:cadt_delg7  -5.831e-03  1.483e-01  -0.039  0.968627
qtde_fisc:cadt_delg8  2.125e-01  2.998e-01  0.709  0.478380
qtde_fisc:cadt_delg9  2.880e-01  2.428e-01  1.186  0.235497
qtde_fisc:cadt_delg10  1.152e+00  4.648e-01  2.478  0.013211 *
qtde_fisc:cadt_delg11 -4.460e-01  4.641e-01  -0.961  0.336525
qtde_fisc:cadt_delg12 -3.163e-01  6.480e-01  -0.488  0.625472
qtde_fisc:cadt_delg13 -1.699e-01  3.603e-01  -0.471  0.637343
qtde_fisc:cadt_delg14 -2.155e-01  2.684e-01  -0.803  0.421972
qtde_fisc:cadt_delg15  5.860e-01  3.283e-01  1.785  0.074275 .
qtde_fisc:cadt_delg16  7.853e-02  1.582e-01  0.496  0.619679
qtde_fisc:cadt_delg17 -1.340e-01  3.719e-01  -0.360  0.718733
qtde_fisc:cadt_delg18  2.834e-01  2.245e-01  1.262  0.206784
qtde_fisc:qtde_auto  -2.389e-01  6.365e-02  -3.752  0.000175
***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

AIC: 3113.3
Number of Fisher Scoring iterations: 16

Null deviance: 4972.0 on 4398 degrees of freedom
Residual deviance: 2881.3 on 4283 degrees of freedom

```

Fonte: Elaborado pelo autor da pesquisa.

2.2.3.3 Resultados da seleção de variáveis

Construído o modelo, é possível verificar quais variáveis possuem maior influência na predição do resultado da diligência. Para o cálculo dessa influência, foi utilizada como referência a diminuição que cada variável acarretou no Desvio Residual ($-2LL$) e a significância estatística de cada uma delas.

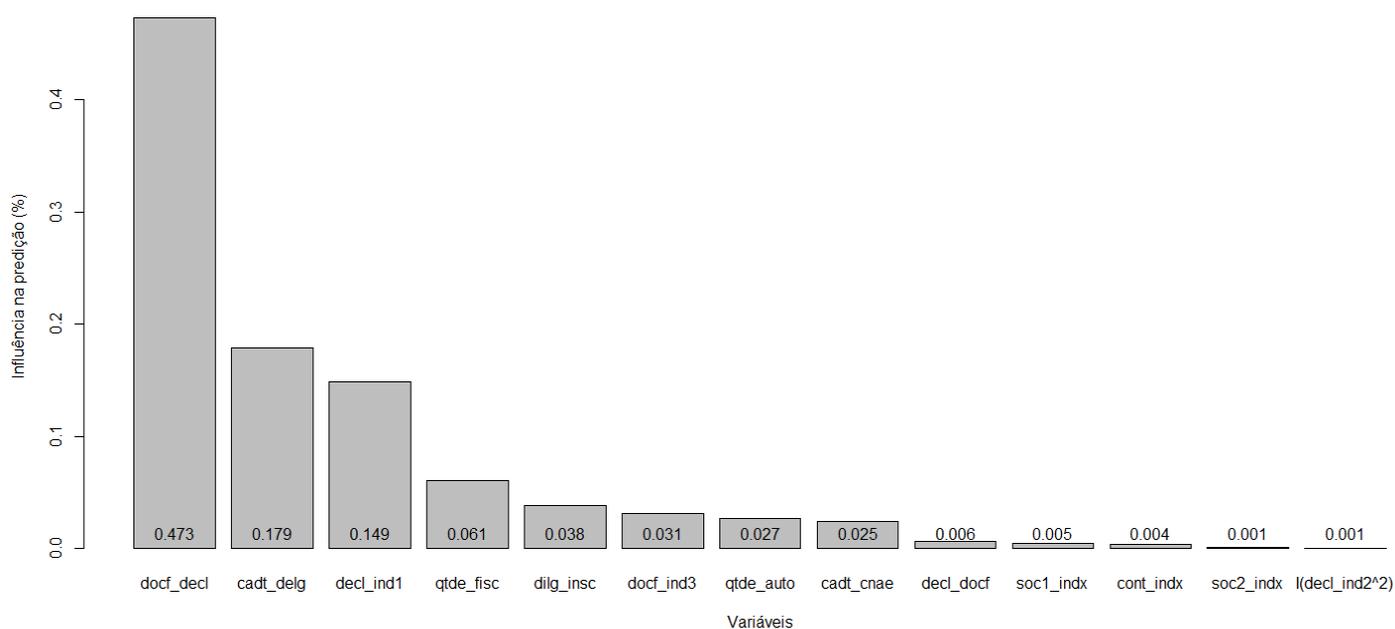
A diminuição no Desvio Residual foi obtida com a aplicação de um teste Qui-quadrado do modelo contra o modelo *null* (com apenas a variável dependente). A significância estatística, por sua vez, foi obtida com a aplicação do teste de Wald que, de acordo com Fávero (2009), é capaz de listar a significância estatística de cada variável do modelo.

Uma vez obtidas as medidas, a influência foi calculada com a fórmula abaixo (em que i representa cada variável). O resultado foi ponderado pelo total para se obter a influência em pontos percentuais.

$$\text{Influência} = (\text{Diminuição Desvio Residual}_i + \text{Significância}_i)/2$$

A influência das variáveis na predição pode ser observada no **Gráfico 10**.

Gráfico 10: Influência das variáveis na predição.



Fonte: Elaborado pelo autor da pesquisa.

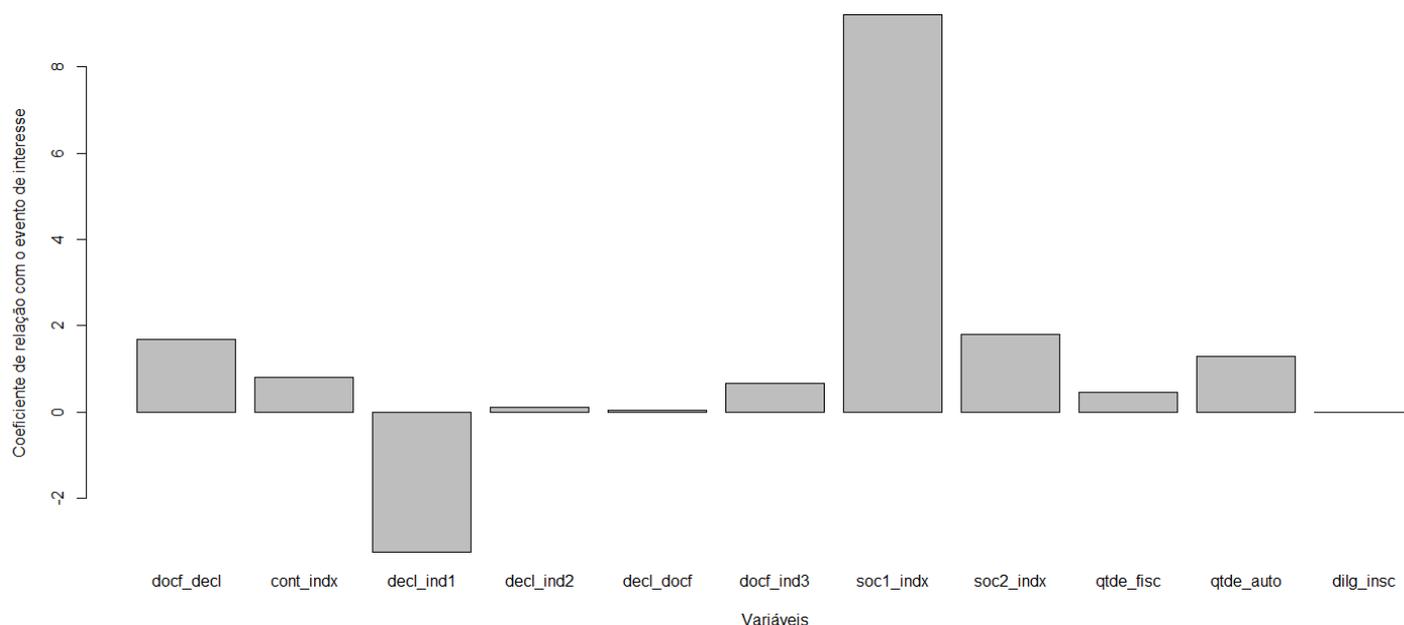
De acordo com o *Gráfico 10*, a variável que mais influenciou no resultado da predição foi a **docf_decl** – Índice de ausência de lançamento de débitos. Essa variável reflete os débitos lançados em declarações e em Notas Fiscais e procura mapear discrepâncias entre ambos, pois estas podem implicar em diminuição irregular dos impostos devidos.

A segunda variável mais importante foi a **cad_t_delg** – Delegacia Regional Tributária. Essa variável tem estreita relação com a localização geográfica do contribuinte e, de acordo com os resultados da pesquisa, pode ser útil à Administração Tributária no monitoramento de irregularidades e práticas sonegatórias. Uma possível utilidade, por exemplo, seria a criação de modelos como o desta dissertação, porém refinados por Delegacia. Tais modelos permitiriam identificar diferenças na ocorrência do evento de interesse nas várias localidades.

A variável seguinte, **decl_ind1** – Índice de declarações, também teve um peso relativamente elevado na predição da ocorrência do evento de interesse. Essa variável mapeia a regularidade na entrega de Declarações e a ocorrência de Declarações sem movimento. Uma declaração sem movimento pode indicar que o contribuinte deixou de operar, mas pode indicar também casos em que omite inteiramente as operações para não pagar os tributos devidos.

As demais variáveis tiveram um peso menor, mas não nulo na predição. O desempenho das três primeiras pode ter sido influenciado por elas, o que tende a ratificar a importância de todas as variáveis selecionadas e justificar sua manutenção do modelo.

Uma vez analisada a influência das variáveis na predição, torna-se importante considerar também a direção dessa influência, conforme demonstrado no *Gráfico 11*. A direção da influência foi obtida através dos coeficientes listados no *Quadro 4*.

Gráfico 11: Direção da influência das variáveis na predição.

Fonte: Elaborado pelo autor da pesquisa.

Em linhas gerais, o *Gráfico 11* demonstra o sentido da relação da variável com o evento de interesse. Tomando o caso de **soc1_indx**, por exemplo, é possível observar que um aumento nos valores da variável provoca também um aumento na probabilidade de ocorrência do evento de interesse.

No caso da variável **decl_ind1**, a relação é inversa, o que indica que é a diminuição no índice representado pela variável que aumenta a probabilidade de ocorrência do evento de interesse.

Esse sentido inversamente relacionado era o esperado para a variável **decl_ind1**, pois conforme se observou na *Tabela 17*, o distanciamento de seus valores de 1.0 em direção a 0.0 (diminuição) foi apontado como hipótese de aumento da probabilidade de ocorrência de irregularidade.

Comparando-se a *Tabela 17* com o *Gráfico 11*, observa-se que a hipótese de influência se confirmou para todas as variáveis, à exceção de **decl_docf**. No caso dessa variável, a hipótese de influência levantada foi que o distanciamento do índice de 0.0 em direção a -1.0 (diminuição) aumentaria a probabilidade de ocorrência de irregularidade, mas o *Gráfico 11* mostra o contrário.

Como justificativa, tem-se que a hipótese apontada na *Tabela 17* foi assim formulada por causa da percepção visual de concentração de indivíduos do grupo 1 – *Irregular* no quadrante próximo a -1.0, mostrada na *Figura 17*. Se for considerada, no entanto, a

formulação intrínseca à variável, discutida na 2.2.2.2.3 *Variável Índice de ausência de lastro de créditos* (que indica que o distanciamento de 0.0 [em qualquer direção] pode indicar irregularidade), o sentido de influência da variável identificado pelo modelo se torna plausível.

2.2.4 *Validação do modelo*

Uma vez definidas as variáveis e criado o modelo, é preciso validá-lo, para garantir que seus resultados poderão ser aplicados além da amostra, ou seja, confirmar que não estão restritos aos dados que embasaram sua criação.

Para isso, o primeiro passo, segundo Fávero (2009), é verificar se os pressupostos subjacentes à técnica estatística empregada estão sendo atendidos.

O passo seguinte é a verificação da qualidade do ajuste do modelo. Allisson (2014) explica que a qualidade do ajuste busca responder se o modelo se ajusta ou não aos dados. Para isso, são efetuados diversos testes, chamados de *Goodness-of-Fit (GOF) Tests*, que permitem verificar se o modelo está aderente aos dados que o embasam.

As próximas seções abordarão esses tópicos através da discussão dos pressupostos do modelo e dos testes de ajuste.

2.2.4.1 *Os pressupostos do modelo Logit*

A Regressão Logística é uma técnica estatística bastante flexível, que não impõe diversos dos pressupostos de outras técnicas de classificação e predição, tais como normalidade de resíduos e homogeneidade de variância. É o que argumentam Barboza, Kimura e Altman (2017).

Nada obstante, alguns requisitos precisam ser atendidos para que seja possível generalizar o modelo para além da amostra. Kassambara (2018) e Fávero (2009) apontam que um modelo *Logit* deve observar os seguintes pressupostos:

1. Ausência de Multicolinearidade;
2. Ausência de *Influential Values*;
3. Relação linear entre as variáveis explicativas contínuas e a variável dependente.

2.2.4.1.1 Ausência de Multicolinearidade

De acordo com Kennedy (2003), Multicolinearidade corresponde à situação em que variáveis independentes são altamente correlacionadas. Em outras palavras, ocorre com variáveis que apresentam comportamento semelhante em função da alta linearidade existente entre elas. Ainda, segundo o autor, a existência de Multicolinearidade pode indicar um problema na amostra de pesquisa utilizada e prejudicar os resultados obtidos pelo modelo.

Bello (2008) aponta que na presença de Multicolinearidade, a variância dos parâmetros estimados se eleva, as estimativas se tornam não confiáveis e o teste de hipóteses passa a ter pouco poder.

A detecção de Multicolinearidade pode ser feita através das chamadas estatísticas VIF (*Variance Inflation Factors*), que no caso de Regressão Logística são chamadas de *General Variance Inflation Factors* (O'BRIEN, 2007).

A tolerância de uma variável explicativa (independente) indica a proporção de sua variação que independe das outras variáveis explicativas. Se a tolerância for baixa, a variável explicativa compartilhará um percentual elevado de sua variância com as demais variáveis. O VIF é o inverso do nível de tolerância e sua fórmula é $1/Tolerância$ (FÁVERO, 2009).

A Tabela 25 mostra vários níveis de tolerância e seus VIF correspondentes.

Tabela 25: Relação entre tolerância e VIF.

Tolerância	VIF	Tolerância	VIF
0,05	10,00	0,55	1,67
0,10	6,67	0,60	1,54
0,15	5,00	0,65	1,43
0,20	4,00	0,70	1,33
0,25	3,33	0,75	1,25
0,30	2,86	0,80	1,18
0,35	2,50	0,85	1,11
0,40	2,22	0,90	1,05
0,45	2,00	0,95	1,00
0,50	1,82	1,00	0,95

Fonte: Elaborado pelo autor da pesquisa.

Segundo Menard (2002), uma tolerância de menos que 0.20 é causa de preocupação e uma tolerância de menos que 0.10 quase que certamente indica um problema sério de colinearidade e, por isso, a regra geral é que um VIF igual ou menor que 4 seria suficiente para afastar preocupações com Multicolinearidade.

Para verificação do GVIF das variáveis do modelo foi utilizada a função **vif** do pacote **car** do Software R (FOX e WEISBERG, 2011). O resultado pode ser visto na **Tabela 26**.

Tabela 26: Resultado da verificação de Multicolinearidade.

Variável	GVIF
docf_decl	2.718766
cont_indx	1.047443
decl_ind1	1.147588
decl_ind2	1.192594
decl_docf	1.793936
docf_ind3	1.927668
soc1_indx	2.609275
soc2_indx	2.768044
qtde_fisc	1.869836
qtde_auto	1.864170
dilg_insc	1.179580
cad_t_cnae	1.340058
cad_t_delg	1.561458

Fonte: Elaborado pelo autor da pesquisa.

O maior GVIF da **Tabela 26** foi 2.768044, da variável *soc2_indx*, o que significa que todas as variáveis do modelo possuem um nível de tolerância maior que 0.30 e isso afastaria a preocupação com Multicolinearidade no modelo.

Diante do exposto, é possível afirmar que **o modelo criado passou no teste de Multicolinearidade**.

2.2.4.1.2 Ausência de Influential Values

Observações que possuem valores surpreendentemente distantes das outras da amostra são chamadas de *Outlying values* ou, de forma simplificada, *Outliers*. De acordo com Osborne (2004), tais observações podem ter origem em erros na captura dos dados (erros de medição, erros de cálculo e erros metodológicos) ou podem simplesmente refletir uma manifestação natural de variabilidade nos dados.

Por sua natureza e origem nem sempre serem claras, a remoção de um *Outlier* deve ser cuidadosamente analisada, conforme defendem Hosmer e Lemeshow (2000). Sua presença no modelo pode melhorar seu ajuste ou ser desejada pelo pesquisador por refletir a realidade intrínseca de parte da população.

Um conceito relacionado e não menos importante é o de *Influential Values*. Ao contrário de *Outliers*, que podem ser benéficos na adequação do modelo à realidade, *Influential Values* são observações que causam distorções no modelo e prejudicam a qualidade do seu ajuste, e que por isso, devem ter sua remoção considerada.

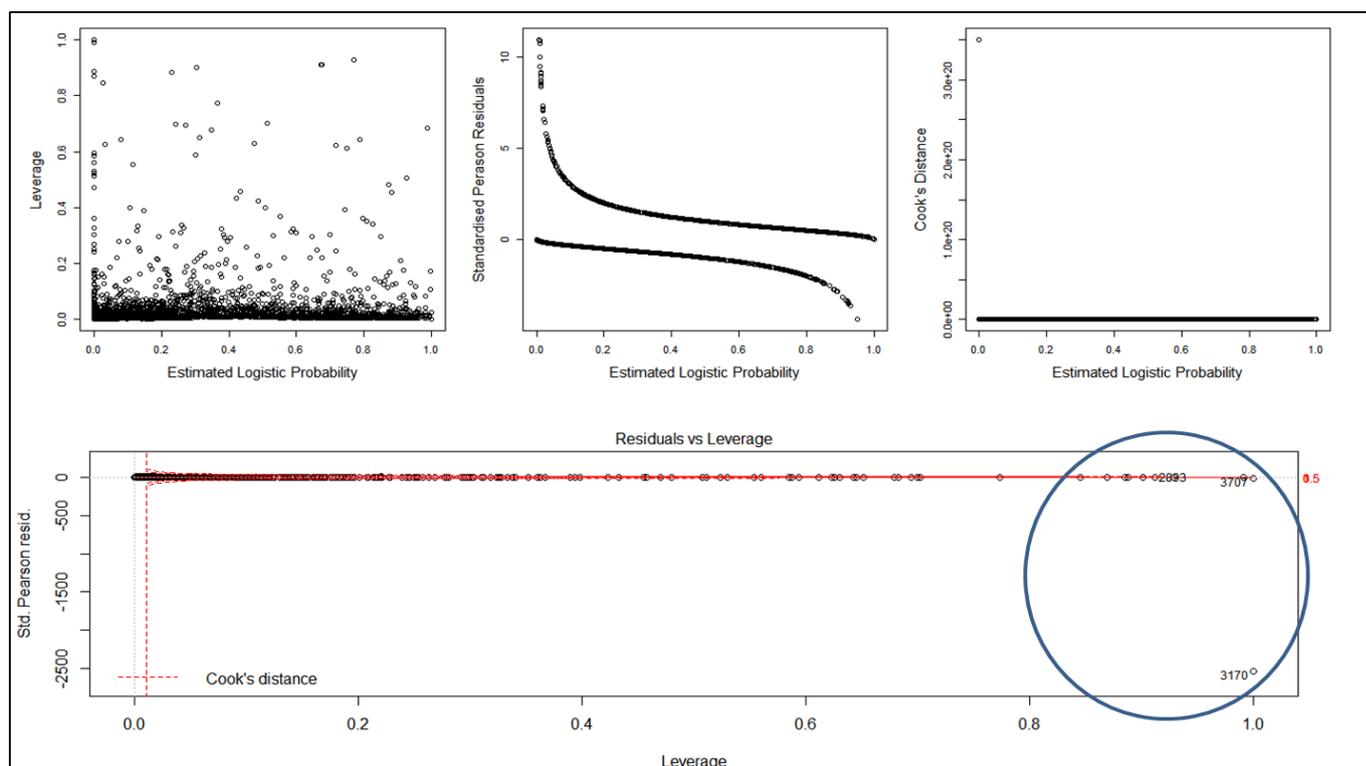
Collett (2003) aponta que nem todos *Outliers* são *Influential Values* e vice versa, e por isso, Hosmer e Lemeshow (2000), Collett (2003), Fávero (2009) e Kassambara (2018) recomendam que antes da remoção de um valor da mostra, sua classificação como *Influential Value* seja constatada através dos seguintes passos:

1. Análise das medidas *Leverage*, *Standardised Pearson Residuals* e *Cook's Distance*.
2. Análise individual pelo pesquisador dos valores para confirmar que sua remoção do conjunto de dados não prejudicaria a amostra.
3. Verificação do ajuste do modelo sem os *Influential Values*.

A medida *Cook's Distance* é função das medidas *Leverage* e *Standardised Pearson Residuals* e detecta a influência de determinada observação no modelo. Observações com elevados *Standardised Pearson Residuals* ou grandes *Leverages* aumentam a *Cook's Distance*, indicando maior influência individual dessas observações no modelo (FÁVERO, 2009).

Cada uma das três medidas, plotadas pelo Software R, é apresentada na **Figura 23**, que contém, também, um gráfico com todas juntas.

Figura 23: *Leverage, Standardised Person Residuals e Cook's Distance* do modelo.



Fonte: Elaborado pelo autor da pesquisa.

Os três primeiros gráficos da **Figura 23** mostram os valores de cada medida e o último gráfico é uma junção de todas elas. Analisando-se o último gráfico, é possível observar que alguns valores parecem não se ajustar e ainda causar distorção no modelo.

A análise individual demonstrou se tratar de 8 observações, as de posições 3170, 3707, 2893, 3767, 4230, 2001, 2514 e 4195 na amostra. Considerando o total de indivíduos na amostra (4.407), essas observações representam 0,18% do total.

Para preservar o sigilo de algumas informações, os valores dessas observações nas variáveis explicativas não serão exibidos neste texto, mas é possível dizer que aparentemente essas observações possuem valores normais, ou seja, não refletem erros de captura. Nada obstante, os valores também não aparentam ter nenhuma particularidade que torne sua presença indispensável na amostra, o que sugere que sua remoção não trará prejuízos ao modelo.

Para confirmar essa suposição, foi analisado o ajuste do modelo sem essas observações. Para tanto, foram utilizadas as medidas **-2LL**, discutida anteriormente, **Pseudo R^2 McFadden**, **Accuracy** e **Area Under the ROC Curve – AUC**. As três últimas medidas serão discutidas na seção 2.2.5 *Análise dos resultados do modelo*. Por ora, é importante ressaltar que enquanto **-2LL** é melhor quando diminui, elas são melhores quando aumentam.

A *Tabela 27* mostra a comparação das medidas antes e depois da remoção das observações.

Tabela 27: Medidas após remoção das observações.

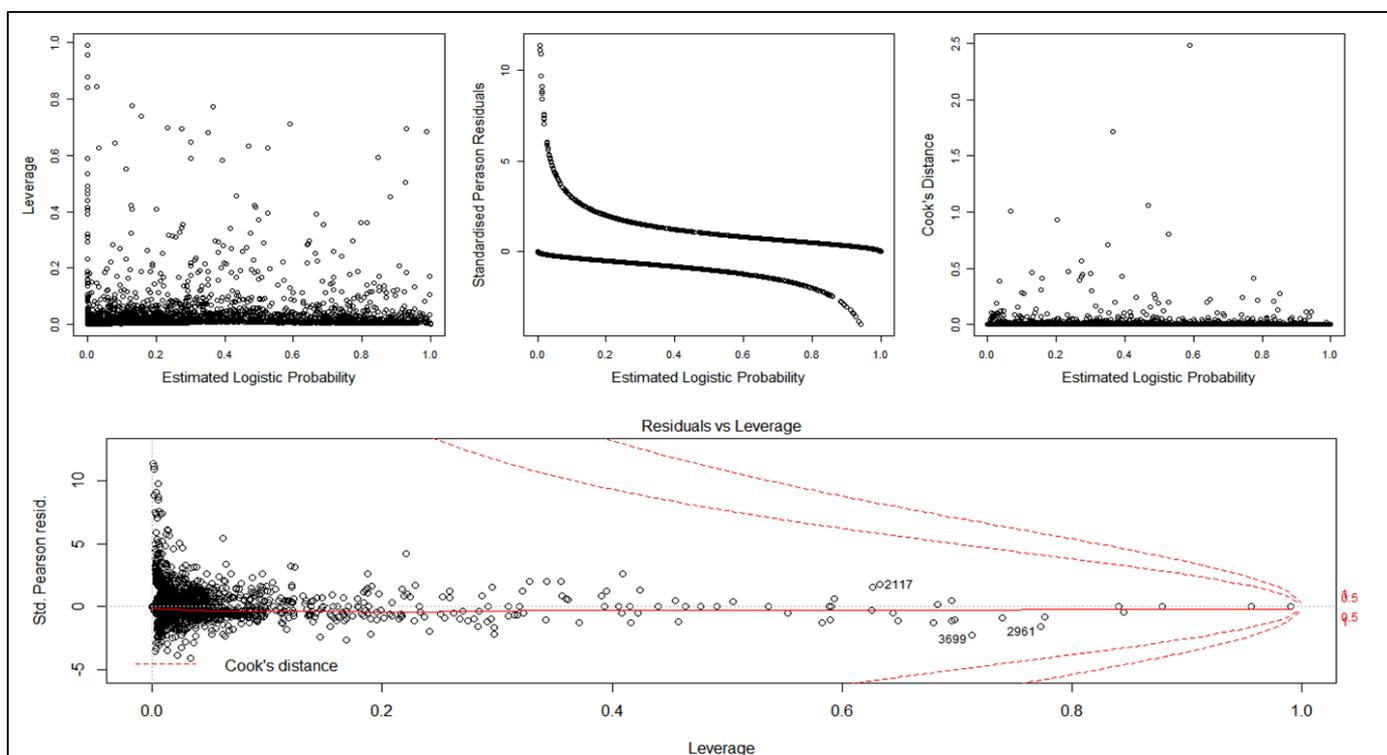
	<i>-2LL</i>	<i>Pseudo R² McFadden</i>	<i>Accuracy</i>	<i>AUC</i>
Com as observações	2922,75	0,41347	0,8580	0,902
Sem as observações	2879,53	0,42085	0.8581	0.904

Fonte: Elaborado pelo autor da pesquisa.

A *Tabela 27* mostra que em todas os casos o ajuste do modelo se torna melhor sem as observações, o que parece indicar que se tratam, de fato, de *Influential Values*.

A *Figura 24* refaz os gráficos apresentados na *Figura 23*, mas desta vez sem os *Influential Values*.

Figura 24: Leverage, Standardised Person Residuals e Cook's Distance sem Influential Values



Fonte: Elaborado pelo autor da pesquisa.

Visualmente, a mudança parece ser discreta no primeiro e segundo gráficos, mas bem expressiva no terceiro e no quarto, o que tende a confirmar que as observações são *Influential Values*, justificando sua remoção do modelo.

Por fim, aplicou-se o teste de significância de Bonferroni, capaz de detectar *outliers* e *Influential Values* (COOK e PRESCOTT, 1981). Para isso, foi utilizada a função **outlierTest** do pacote **car** do Software R.

O resultado, apresentado no **Quadro 5**, mostra que não há mais nenhum *outlier* ou *Influential Value* no modelo de acordo com esse teste. Esse **parece ser um indicativo de que o modelo está bem ajustado sem as observações e atende ao pressuposto de ausência de *Influential Values***. Na seção 2.2.4.2 *Qualidade do ajuste do modelo*, a qualidade do ajuste do modelo sem as observações será confirmada.

Quadro 5: Resultado do outlierTest.

No Studentized residuals with Bonferonni p < 0.05				
Largest rstudent :				
	rstudent	unadjusted	p-value	Bonferonni p
845	3.13831		0.0016992	NA

Fonte: Elaborado pelo autor da pesquisa.

2.2.4.1.3 Relação linear entre as variáveis explicativas contínuas e a variável dependente

Em técnicas de classificação e predição nem todas as variáveis de influência na determinação de ocorrência do evento de interesse são conhecidas. Um exemplo dado na seção 2.2.1.6 *Limitações da pesquisa*, foram a crenças religiosas.

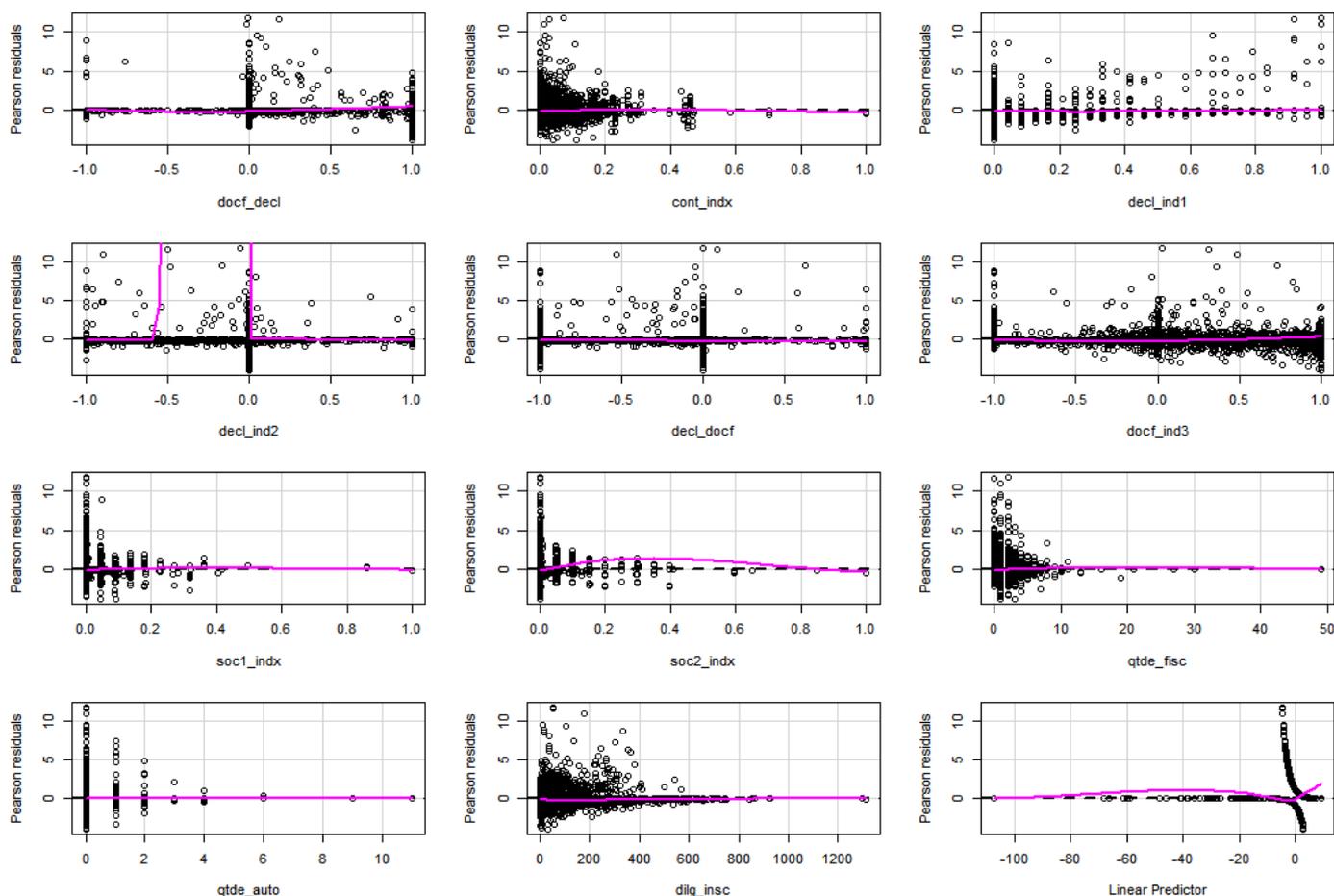
Fávero (2009) aponta que, em função disso, pressupõe-se que os estimadores (variáveis) ausentes sejam constantes em relação aos dados na amostra. Esse pressuposto pode ser violado caso a relação entre a variável dependente e alguma variável explicativa seja não linear.

Hosmer e Lemeshow (2000) argumentam que a verificação de linearidade entre variáveis contínuas e a variável dependente é boa prática na construção de modelos *Logit*. Zhang (2016a) também defende a importância da verificação e indica que caso alguma relação de não linearidade seja encontrada, a variável explicativa pode ser transformada através da aplicação de potência de 2 ou 3 (*2 or 3 power-terms*), polinômios fracionais (*fractional polynomials*) ou funções *spline*.

Para verificar a ocorrência de linearidade nas variáveis explicativas, foi utilizada a função **residualPlots** do pacote **car** do Software R.

A *Figura 25* mostra o resultado dos gráficos gerados.

Figura 25: Resultado da verificação de linearidade das variáveis contínuas



Fonte: Elaborado pelo autor da pesquisa.

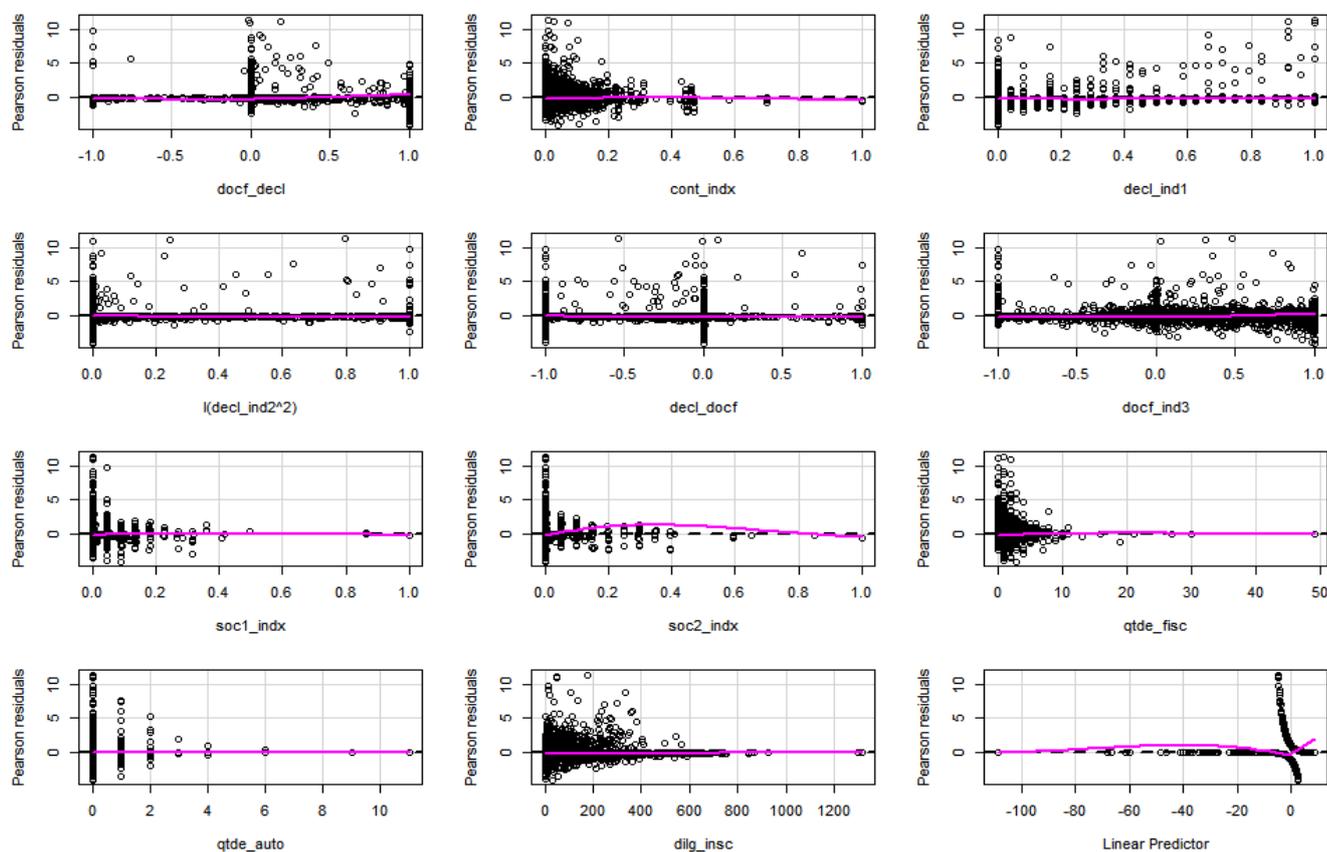
A *Figura 25* apresenta os valores confrontados com os respectivos resíduos de cada variável explicativa contínua. O último gráfico da figura confronta os valores da predição do modelo com seus resíduos gerais.

Zhang (2016b) indica que quanto mais reta for a linha em cada gráfico, mais linear é a relação dos resíduos com a variável explicativa e a relação desta com a variável dependente. Inspeccionando-se visualmente, é possível observar que a variável *decl_ind2* não possui relação de linearidade com a variável dependente e requer transformação.

Para isso, optou-se por aplicar a potência de 2 nos valores da variável através da seguinte fórmula: $I(\text{decl_ind2}^2)$. O símbolo “^” indica **potência de** (*power of*), e a função “I” foi utilizada para preservar os valores sem efetuar qualquer interpretação ou conversão.

A *Figura 26* refaz os gráficos da *Figura 25*, mas desta vez com a transformação aplicada em *decl_ind2*.

Figura 26: Resultado da verificação de linearidade das variáveis contínuas com *decl_ind2* transformada



Fonte: Elaborado pelo autor da pesquisa.

Como é possível observar, após a transformação de *decl_ind2*, seus resíduos aparentam ter linearidade com seus valores e conseqüentemente, a respectiva variável explicativa aparenta relação de linearidade com a variável dependente, **indicando que o modelo passou a atender o pressuposto de linearidade das variáveis contínuas.**

Na próxima seção serão efetuados diversos testes no modelo originado para que se verifique o grau de qualidade do seu ajuste.

2.2.4.2 Qualidade do ajuste do modelo

A aferição da qualidade do ajuste do modelo não objetiva avaliar seu poder de predição, mas sim buscar indicativos de que o modelo está bem ajustado e não se restringe à amostra no qual se baseia.

Tal aferição pode ser efetuada levando-se em consideração o modelo como um todo ou considerando cada termo individualmente. Nos próximos dois tópicos serão abordadas ambas as aferições.

2.2.4.2.1 Medidas de ajustes individuais

Hosmer e Lemeshow (2000) argumentam que as medidas de ajustes individuais são fundamentais na verificação do ajuste do modelo, porque permitem analisar como cada observação se comporta em relação às outras.

Para tanto, os autores sugerem três medidas²⁴.

1. ΔX^2 – *Delta chi-square*;
2. ΔD – *Delta deviance*;
3. $\Delta \hat{\beta}$ – *Delta beta*.

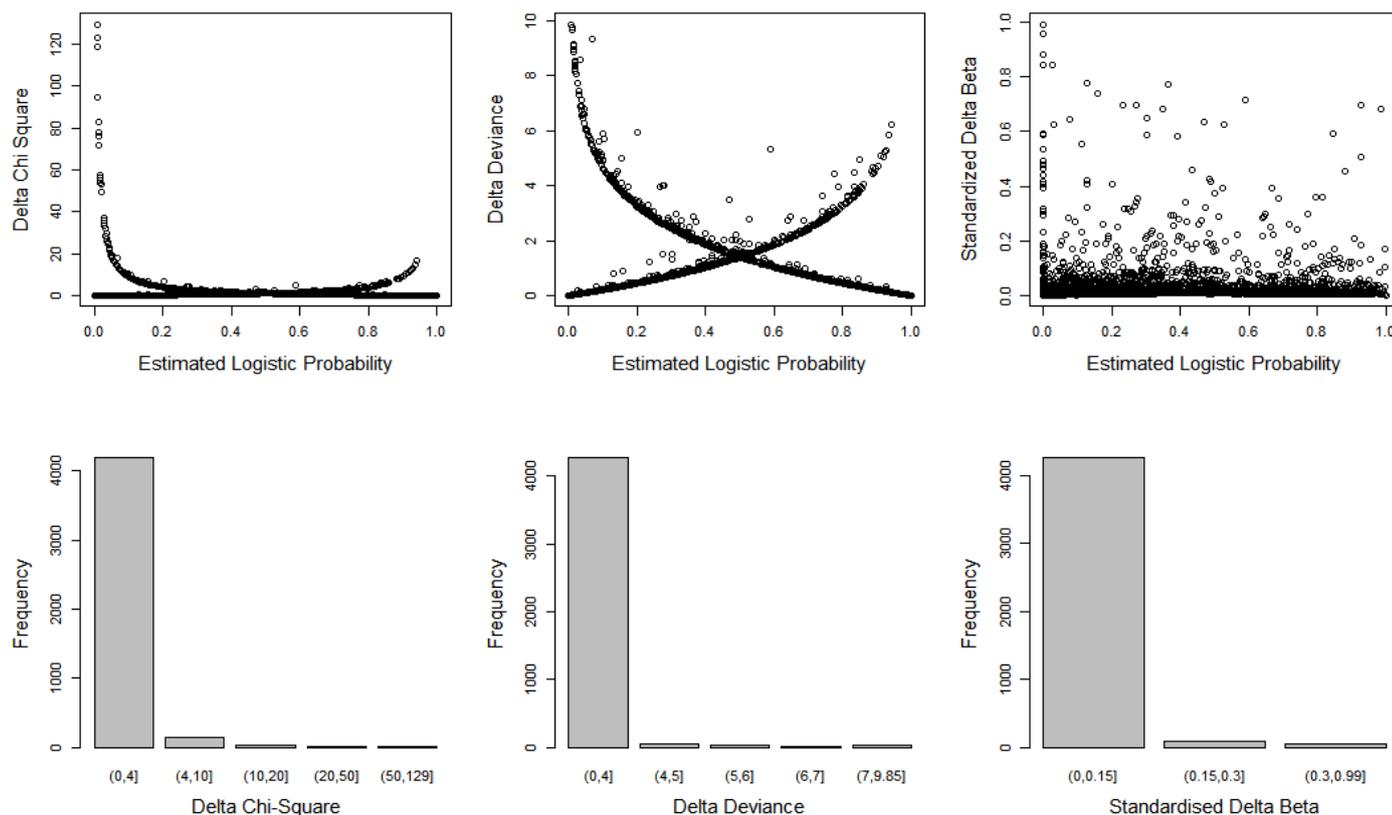
As estatísticas *Delta chi-square* e *Delta deviance* representam resíduos de *Pearson* e de *Deviance*, respectivamente, e ajudam a identificar observações que não se ajustaram corretamente ao modelo. Usualmente, elevados valores em ambas as estatísticas indicam tais situações.

A estatística *Delta beta*, por sua vez, ajuda a identificar valores que exercem influência muito elevada no modelo a ponto de prejudicar seu ajuste. Essa estatística também é uma maneira de detectar *Influential Values*, assim como as medidas vistas na seção 2.2.4.1.2 *Ausência de Influential Values*.

A **Figura 27** apresenta três gráficos, um para cada estatística, e outros três com a distribuição de frequência dos valores de cada uma.

²⁴ Como não se trata de uma pesquisa com foco principal em matemática ou estatística, para efeitos de simplificação, as fórmulas de tais medidas serão omitidas do texto. Leitores interessados são encorajados a verificarem Hosmer e Lemeshow (2000, pp 173 e 174).

Figura 27: Resultado da verificação das medidas de ajustes individuais.



Fonte: Elaborado pelo autor da pesquisa.

Os primeiros três gráficos apresentam as estatísticas confrontadas com o resultado do modelo, ou seja, a probabilidade estimada de cada observação pertencer a um dos grupos de interesse 0 – *Regular* e 1 – *Irregular*.

Embora o intervalo de alcance dos dois primeiros gráficos apresente diferenças significativas – o que é uma característica das estatísticas utilizadas – a forma de ambos é similar e mostra curvas quadráticas. De acordo com Hosmer e Lemeshow (2000), a curva quadrática é um dos tipos de relações possíveis e esperadas entre o resultado do modelo *Logit* e suas variáveis contínuas.

Alguns pontos exibidos nos dois primeiros gráficos possuem valores mais elevados, indicando um ajuste inferior em relação aos outros, mas esse comportamento tende a ser normal, já que modelos usualmente não são capazes de se ajustarem de forma plena a todas as observações. O importante nos gráficos, segundo Hosmer e Lemeshow (2000), é analisar os pontos como um todo, para que se chegue à conclusão de que o modelo está ou não bem ajustado.

Para isso, segundo os autores, é importante observar se a maioria dos valores de ΔX^2 e ΔD são menores ou pelo menos não muito maiores que 4, pois isso indicaria um bom ajuste do modelo.

Nos gráficos de distribuição de frequência de cada uma das duas medidas é possível observar que a maioria dos valores tanto de ΔX^2 quanto de ΔD estão concentrados entre 0 e 4, o que tende a indicar que o ajuste do modelo esteja correto.

Com relação ao terceiro gráfico de distribuição, é possível observar que a maioria das observações possui valores menores que 0.15 e quase a totalidade, valores menores que 0.3. Hosmer e Lemeshow (2000) defendem que observações com valores menores que 0.3 certamente não prejudicam o modelo e que aquelas com valores menores que 1.0 possuem grande probabilidade de não exercer influência suficiente para trazerem qualquer prejuízo nos coeficientes deste.

Há exceções a essa regra e por isso, outras medidas de detecção de *Influential Values* devem ser utilizadas. Foi exatamente o que se fez na seção 2.2.4.1.2 *Ausência de Influential Values*, em que foram identificados e removidos *Influential Values* que prejudicavam o ajuste do modelo.

Diante do exposto, **pode-se inferir que o modelo está bem ajustado em relação às medidas individuais das observações.**

Na próxima seção serão verificados os valores globais de ajuste do modelo.

2.2.4.2.2 Testes de ajuste geral do modelo

A verificação do ajuste geral do modelo é feita através de testes que geram um número que pode ser comparado com valores de referência.

Allison (2014) explica que esse número gerado é um *p-value* associado à hipótese de verificação, que nesse caso é a de que o modelo está bem ajustado. Se o *p-value* for maior que o nível de significância estabelecido, o modelo passou no teste, caso contrário, a hipótese de que o ajuste está correto foi rejeitada.

O autor sugere que o nível de significância para os testes seja de 0.05, que geraria um nível de confiança de 95%. Como existem diferentes testes disponíveis para o modelo *Logit*, a sugestão é que mais de um teste seja feito, a fim de que se obtenha uma precisão maior em relação à qualidade do ajuste.

Com base nas discussões apresentadas por Hosmer e Lemeshow (2000), Collett (2003), Fávero (2009), Allisson (2014) e Zhang (2016a), optou-se pela utilização de 5 testes de verificação da qualidade do ajuste do modelo:

1. *Pearson Chi-square Statistic and Deviance.*
2. *Hosmer-Lemeshow decile of risk test*, também chamado de *Hosmer-Lemeshow goodness of fit (GOF) test.*
3. *Osius and Rojek normal approximation to the distribution of the Pearson chi-square statistic.*
4. *Stukel's test.*
5. *Kolmogorov-Smirnov test.*

Cada um deles será discutido e aplicado ao modelo nas seções seguintes. Os itens 1 a 4 foram calculados através da função **gof** do pacote **LogisticDx** do Software R (DARDIS, 2015), ao passo que o 5 foi calculado com a função **ks.text** do pacote **stats** do mesmo software (R CORE TEAM, 2018),

Deviance e Pearson Chi-square Statistic

As medidas clássicas para avaliação do ajuste do modelo logístico são a *Deviance* e a *Person Chi-square Statistic* (ALLISSON, 2014).

A *Deviance*, ou desvio residual, é a medida **-2LL** discutida na seção 2.2.3.1 *Procedimento para seleção de variáveis* e indica o quanto o modelo está melhor ajustado do que modelo *null* (com apenas a variável dependente).

O **Quadro 4** demonstrou que o desvio residual do modelo *null* é de 4972.0, ao passo que o modelo construído neste trabalho possui um desvio residual de 2881.3, o que indica uma **melhora de 2090.7** pontos. Após a retirada dos *Influential Values*, o desvio residual diminuiu ainda mais, indo a 2879.53.

A *Pearson Chi-square Statistic*, ou *Standardised Pearson Residuals*, por sua vez, busca comparar os valores observados com aqueles calculados pelo modelo. Essa medida pode ser calculada por observação (individual), por variável explicativa (grupo) ou por tabela de contingência²⁵.

²⁵ Mais detalhes sobre tabelas de contingência em testes de ajuste podem ser obtidos em Fávero (2009).

A **Tabela 28** mostra os resultados dos testes de *Pearson Chi-square* e de *Deviance*. A hipótese nula é que o ajuste do modelo está correto e, nesse caso, *p-values* maiores que 0.05 confirmam a qualidade do ajuste do modelo em um intervalo de confiança de 95%.

Tabela 28: *Pearson Chi-square Statistic* e *Deviance* tests

Teste	Base	<i>p-value</i>	Modelo está bem ajustado?
<i>Pearson Chi-square</i>	Individual	0.395750	Sim
<i>Pearson Chi-square</i>	Grupo	0.096009	Sim
<i>Pearson Chi-square</i>	Tabela de contingência	0.096009	Sim
<i>Deviance</i>	Individual	1.000000	Sim
<i>Deviance</i>	Grupo	1.000000	Sim
<i>Deviance</i>	Tabela de contingência	1.000000	Sim

Fonte: Elaborado pelo autor da pesquisa.

Em todos os casos o *p-value* foi bem acima de 0.05, indicando que o **modelo está bem ajustado segundo as medidas *Deviance* e *Pearson Chi-square Statistic***.

Hosmer-Lemeshow Test

O teste Hosmer-Lemeshow agrupa as observações por faixas de valores das probabilidades estimadas de ocorrência do evento de interesse (*I – Irregular*) e, através da aplicação de um teste Qui-quadrado (*Pearson's chi-square test*), verifica se há diferenças estatisticamente significativas entre as classificações previstas em cada faixa e as reais da variável dependente (*dilg_irre*). (HOSMER e LEMESHOW, 2000).

A hipótese de verificação é que não há diferenças significativas e nesse caso, *p-values* maiores que 0.05 indicam que o modelo passou no teste em um intervalo de confiança de 95%.

Um dos parâmetros do teste, e componente fundamental de influência no seu resultado, é o número de faixas em que os dados serão divididos. O número de faixas sugerido por Hosmer e Lemeshow (2000) é 10, mas Paul et al. (2013) argumentam que para amostras grandes, esse valor é possivelmente muito pequeno, e em função disso, propuseram a seguinte fórmula para calcular o número de faixas com base no tamanho da amostra.

$$g = \max \left(10, \min \left\{ \frac{m}{2}, \frac{n - m}{2}, 2 + 8 \left(\frac{n}{1000} \right)^2 \right\} \right)$$

Em que *n* é o tamanho da amostra e *m* é o número de ocorrências do evento de interesse. Aplicando-se a fórmula à amostra de pesquisa, com 4.407 observações e 1.114

ocorrências do evento de interesse (vide *Gráfico 4*), o número de faixas obtido foi entre 157 e 158 (157,37).

O teste foi, então, aplicado com as três parametrizações (números de faixas iguais a 10, 157 e 158) e o resultado pode ser visto na *Tabela 29*.

Tabela 29: Resultado do Hosmer-Lemeshow Test

Número de faixas	<i>p-value</i>	Modelo está bem ajustado?
10	0.1161	Sim
157	0.8540	Sim
158	0.1425	Sim

Fonte: Elaborado pelo autor da pesquisa.

Em todos os casos o *p-value* foi bem acima de 0.05, indicando que o **modelo está bem ajustado segundo o Hosmer-Lemeshow Test**.

Osius and Rojek normal approximation to the distribution of the Pearson chi-square statistic

Osius e Rojek (1992) demonstraram que o *Pearson's Chi-square* tem uma distribuição normal assintótica em relação à média e ao desvio padrão dos quais deriva. Subtraindo-se a média e dividindo-se pelo desvio padrão, obtém-se um teste estatístico que tem aproximadamente uma distribuição normal padrão sob a hipótese nula, o que significa que *p-values* elevados não permitem rejeitar a hipótese nula, de que o modelo está corretamente ajustado.

Para efetuar essa verificação, propuseram o *Osius-Rojek Test*, que foi aplicado ao modelo desta pesquisa. O *p-value* de retorno foi **0.9999934**, o que indica que o **modelo está bem ajustado segundo o Osius-Rojek Test**.

Stukel's test

O teste de Stukel é conduzido através da adição de dois parâmetros ao modelo, derivando outro modelo com probabilidades de ocorrência do evento de interesse diferentes do modelo original (mais acentuadas ou menos acentuadas). Uma comparação dos dois modelos é então efetuada e a não rejeição da hipótese nula (*p-value* maior que 0.05) indica que o modelo original está bem ajustado (STUKEL, 1988).

Hosmer e Lemeshow (2000) apontam que o teste pode ser feito de duas formas: como *Score test* ou como *Log-likelihood test*. A **Tabela 30** mostra os resultados de ambos.

Tabela 30: Resultado do Stukel's Test

Parâmetro	<i>p-value</i>	Modelo está bem ajustado?
Score test: parâmetro 1	0.1860911	Sim
Score test: parâmetro 2	0.1359749	Sim
Score test: parâmetros 1 e 2 juntos	0.1372965	Sim
Log-likelihood test: parâmetro 1	0.1689396	Sim
Log-likelihood test: parâmetro 2	0.1590053	Sim
Log-likelihood test: parâmetros 1 e 2 juntos	0.2862762	Sim

Fonte: Elaborado pelo autor da pesquisa.

Em todos os casos o *p-value* foi bem acima de 0.05, indicando que o **modelo está bem ajustado segundo o *Stukel's Test***.

Kolmogorov-Smirnov test

De acordo com Fávero (2009), o teste de Kolmogorov-Smirnov é uma medida de qualidade de ajuste do modelo que mede o grau de segregação entre os grupos de referência (*0 – Regular e 1 – Irregular*).

Ainda, de acordo com os autores, o resultado do teste pode ser interpretado da seguinte forma:

- Menor 0.30: indica baixa discriminação.
- Entre 0.30 e 0.50: indica boa discriminação.
- Maior que 0.50: indica ótima discriminação.

O teste foi aplicado ao modelo desta pesquisa e retornou o valor **0.75**, indicando que o **modelo está com um ótimo ajuste o segundo o *Kolmogorov-Smirnov Test***.

2.2.4.2.3 Qualidade do ajuste do modelo: síntese

As verificações das medidas ΔX^2 – *Delta chi-square*, ΔD – *Delta deviance* e $\Delta \hat{\beta}$ – *Delta beta* demonstraram que o modelo está bem ajustado no que concerne às observações individualmente.

Os testes *Pearson Chi-square Statistic and Deviance*, *Hosmer-Lemeshow*, *Osius-Rojek*, *Stukel* e *Kolmogorov-Smirnov*, por sua vez, demonstraram que o modelo está bem ajustado como um todo.

Diante disto, **pode-se inferir que o modelo apresentado na pesquisa é capaz de ser estendido para além da amostra e ser utilizado no restante da população** para prever o resultado da diligência entre os contribuintes que foram submetidos ao procedimento de Cassação por Inatividade Presumida.

Uma vez concluídas as verificações da qualidade do ajuste e obtidos os indícios necessários de que o modelo está bem ajustado, é necessário aferir o poder de predição do modelo, ou seja, o grau de acerto em classificar as observações no grupo correto. É o que se apresenta na seção a seguir.

2.2.5 Análise dos resultados do modelo

Enquanto as medidas e testes de qualidade do ajuste procuram apresentar informações que ajudam a determinar se o modelo está ou não corretamente ajustado, as medidas de capacidade de predição (*predictive power*) procuram demonstrar a capacidade do modelo de classificar corretamente as observações.

Posto de outra forma, essas medidas representam o potencial de um modelo *Logit* de atender ao objetivo precípua de sua criação, qual seja, de prever corretamente a ocorrência do evento de interesse.

Em relação a esse potencial de predição, a distribuição não homogênea dos indivíduos na amostra impõe uma exigência adicional. O **Gráfico 4** demonstrou a não homogeneidade da amostra em relação ao evento de interesse, o que significa que ao se selecionar aleatoriamente um indivíduo do conjunto, ele terá aproximadamente três vezes mais chances de pertencer ao grupo 0 – *Regular* do que ao grupo 1 – *Irregular*. Posto de outra forma, selecionando-se aleatoriamente um indivíduo da amostra, este terá 74,72% de chances de pertencer ao grupo 0 – *Regular* em função da distribuição de indivíduos na amostra.

Nesse diapasão, Hosmer e Lemeshow (2000) argumentam que, para ser útil, um modelo deve ter uma taxa de acerto na predição maior que a probabilidade de ocorrências de indivíduos no maior grupo. Diante disso, o modelo da presente pesquisa deve ter uma taxa de acerto na predição (*Accuracy*) maior que 74,72% para que seja considerado útil.

Além da *Accuracy*, a capacidade de predição do modelo pode ser aferida com outras medidas, diferentes e complementares entre si. Cada uma delas é discutida nas seções a seguir.

2.2.5.1 *Accuracy*

Hosmer e Lemeshow (2000) descrevem a *Accuracy* como uma medida que indica a capacidade de acerto do modelo, ou seja, o percentual de classificação correta dos indivíduos entre dois grupos possíveis.

Considerando que o resultado do *Logit* é a probabilidade de ocorrência do evento de interesse, para se calcular a *Accuracy* é necessário determinar qual será o ponto de corte da classificação (*cut-off* ou *cut-point*). Se for utilizado, por exemplo, um *cut-off* de 50%, indivíduos com probabilidade associada maior que 0.5 serão classificados no grupo *1 - Irregular*, caso contrário, serão classificados no grupo *0 - Regular*.

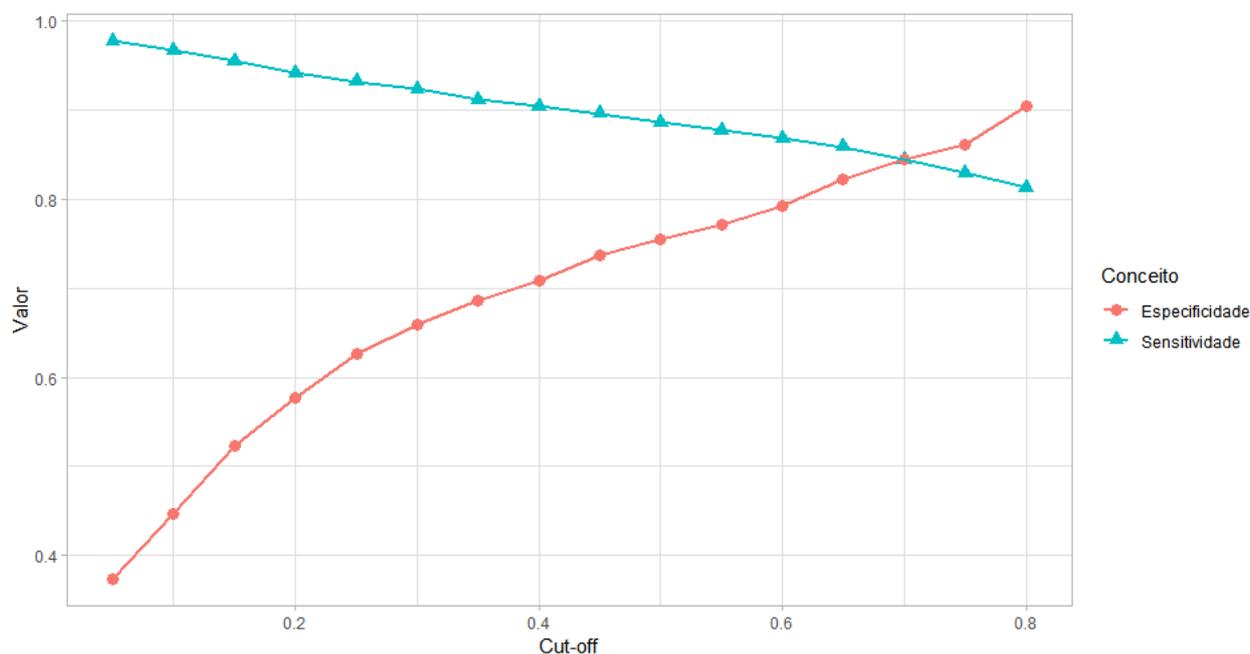
Para se determinar o *cut-off*, é necessário considerar os conceitos de **Sensitividade** (*Sensitivity*) e **Especificidade** (*Specificity*).

Sensitividade é a probabilidade de detecção do sinal positivo verdadeiro (*true positive rate*), ou seja, classificar como *1 - Irregular* um indivíduo que efetivamente pertence a esse grupo.

Especificidade, por sua vez, é a probabilidade de detecção do sinal negativo verdadeiro (*true negative rate*), ou de não classificar como *1 - Irregular* um indivíduo que não pertence a esse grupo.

O **Gráfico 12** mostra os níveis de Sensitividade e Especificidade para diversos pontos de corte possíveis do modelo.

Gráfico 12: Valores de Sensitividade e Especificidade por *cut-off*.



Fonte: Elaborado pelo autor da pesquisa.

A determinação do *cut-off* ideal requer a análise do grau de importância de cada um dos conceitos. Pelo gráfico é possível observar que se os dois conceitos tiverem a mesma importância, o *cut-off* ideal, que maximiza ambos, seria 0.7.

Ocorre, no entanto, que nem sempre se deve atribuir o mesmo peso aos conceitos. Imagine-se, por exemplo, o cadastro biométrico para votação nas eleições. A não detecção de uma digital correta (baixa Sensitividade) faz com que o eleitor precise provar por outros meios que é quem diz ser, como mostrar documento com foto, por exemplo. Por outro lado, a não detecção de uma digital incorreta (baixa Especificidade) pode ter como resultado a liberação do voto de uma pessoa que se passa por outra.

Parece ficar claro que em um caso como esse é melhor buscar um *cut-off* cuja Especificidade seja mais alta, mesmo que ao custo de uma menor Sensitividade.

No caso da presente pesquisa, ao contrário do exemplo do cadastro biométrico anterior, a **baixa Especificidade parece não ser um problema**. Isso ocorre porque o proposto na pesquisa é que a classificação de um indivíduo como *I – Irregular* não tenha o efeito direto de gerar qualquer sanção sobre ele, mas tão somente apontar que é necessária uma análise mais criteriosa dos pedidos de abertura de novos estabelecimentos.

Portanto, a classificação errônea do indivíduo como *I – Irregular* seria mitigada com a análise do pedido de nova inscrição, como será visto na seção 3 *PLANO DE AÇÃO*. A falha

na classificação, por sua vez, como *1 - Irregular*, de um indivíduo que sonegou tributos (**baixa Sensitividade**), parece ter um potencial de dano maior, pois não provocaria uma análise humana do pedido de nova inscrição.

Em função dessa argumentação, optou-se por não utilizar o *cut-off* de 0.7 e mantê-lo em **0.5**, caso em que a **Sensitividade é maior**.

Uma vez definido o *cut-off*, para se calcular a *Accuracy* é necessário criar uma tabela de classificação (*Classification Table*), também chamada de *Confusion Matrix*. Fávero (2009) aponta que a *Confusion Matrix* é criada da seguinte forma:

- Calculam-se valores de probabilidade para cada item da observação.
- Classificam-se os valores da probabilidade obtida considerando o ponto de corte.
- Comparam-se os valores classificados com os observados (reais).

A **Tabela 31** mostra a *Confusion Matrix* gerada pelo modelo considerando-se o *cut-off* de 0.5.

Tabela 31: Confusion Matrix com cut-off igual a 0.5.

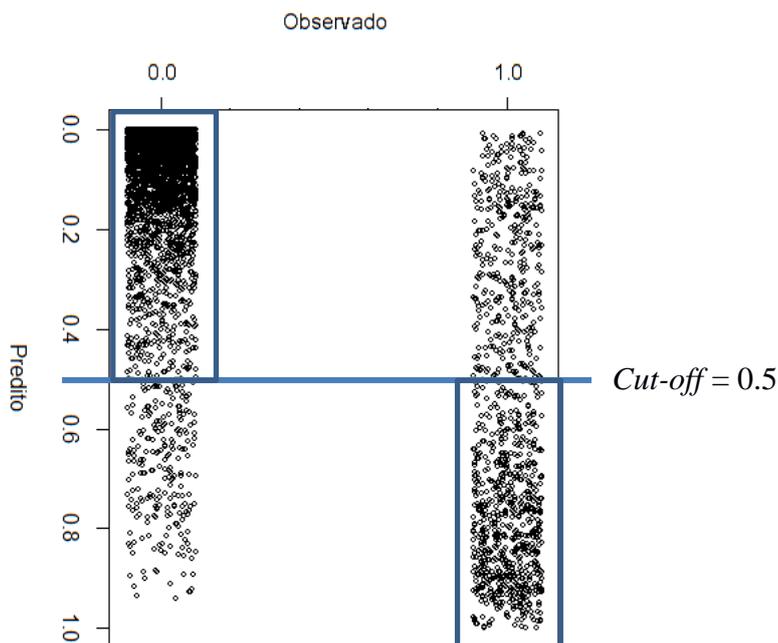
Predito	Observado (referência)	
	0 - Regular	1 - Irregular
0 - Regular	3055	391
1 - Irregular	233	720

Fonte: Dados da pesquisa.

Como é possível observar na **Tabela 31**, a *Classification Table* é uma matriz cuja diagonal marcada em negrito representa os acertos na predição, nesse caso $3.055 + 720 = 3.775$. Considerando um total de 4.399 indivíduos da amostra, já que 8 *Influential Values* foram excluídos, chega-se a uma *Accuracy* de 85,81%.

A **Figura 28** apresenta uma visualização gráfica da *Confusion Matrix*. As observações dentro dos retângulos foram as corretamente classificadas.

Figura 28: Visualização gráfica da *Classification Table*.



Fonte: Elaborado pelo autor da pesquisa.

Embora não haja uma medida de referência para dizer se a *Accuracy* está boa ou não, Hosmer e Lemeshow (2000) argumentam que ela deve ser maior que a probabilidade de ocorrências de indivíduos no maior grupo, que é de 74,72% no caso da presente pesquisa. Isso posto, a ***Accuracy* alcançada pelo modelo parece bastante elevada**, já que ampliou o poder de predição em mais de 11 pontos percentuais em relação à seleção aleatória e diminuiu a **quantidade de erros para 14,19%**.

2.2.5.2 Repeated K-fold Cross Validation

Uma pergunta que se coloca é se a *Accuracy* seria mantida se o modelo fosse aplicado a dados novos, ou seja, dados com os quais o modelo não tenha sido “treinado”. Essa questão surge porque técnicas de *Machine Learning* “aprendem” com os dados e ao efetuar a predição sobre os mesmos dados com os quais aprenderam, podem gerar uma *Accuracy* que não necessariamente seria observada na prática.

Uma forma de mitigar essa limitação é testar o modelo com dados novos, para os quais também se conhece o resultado da ocorrência do evento de interesse, mas que não foram utilizados na aprendizagem. O problema dessa estratégia é que nem sempre o resultado da

ocorrência do evento de interesse está disponível no restante da população, como é o caso da presente pesquisa.

Outra alternativa, normalmente mais viável, é utilizar a própria amostra. Para isso, torna-se parte dos dados oculta ao modelo e outra parte visível. Depois de treinado com a parte visível, o modelo é aplicado à parte oculta e a *Accuracy* da predição calculada.

Para aumentar a confiabilidade do resultado, o processo é repetido n vezes com a amostra dividida aleatoriamente em k segmentos (*folds*) mutuamente exclusivos. Um *fold* é ocultado do modelo e depois utilizado na predição. Os demais são utilizados para a aprendizagem. O resultado final consiste da média das *Accuracies* obtidas em cada repetição. Essa técnica é chamada de *Repeated K-fold Cross Validation* (KOHAVI, 1995).

A implementação da técnica no R foi feita através da função **train** do pacote **caret** (KUHN, 2018). O parâmetro **trainControl** foi definido de forma que o processo se repetisse 10 vezes com 5 *folds* cada.

O resultado pode ser observado no *Quadro 6*.

Quadro 6: Resultado do *Repeated K-fold Cross Validation*.

```
Generalized Linear Model

4399 samples
 13 predictor
 2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 10 times)
Summary of sample sizes: 3519, 3520, 3518, 3520, 3519, 3518, ...
Resampling results:

  Accuracy   Kappa
 0.8444636  0.5700386
```

Fonte: Elaborado pelo autor da pesquisa.

No *Quadro 6* é possível observar que **mesmo com a predição executada sobre bases desconhecidas pelo modelo, a *Accuracy* se manteve elevada, em 84,45%.**

Além da *Accuracy*, foi gerado também o valor *Kappa* (*Cohen's kappa coefficient*). Segundo Vieira (2005), o valor *Kappa* pode ser interpretado da seguinte forma:

- < 0 Ruim;
- 0.01 - 0.20 Pequeno;
- 0.21 - 0.40 Razoável;
- **0.41 - 0.60 Bom (*moderate*);**
- 0.61 - 0.80 Substancial;
- 0.81 - 0.99 Quase perfeito.

No tocante ao modelo da pesquisa o *Kappa* de 0.57 pode ser considerado um bom resultado, segundo a escala acima.

2.2.5.3 Area under the ROC Curve

Como visto anteriormente, a *Accuracy* depende da definição de um único ponto de corte, que mapeia um valor específico de Sensitividade e Especificidade. Se para cada ponto de corte possível forem mapeados os valores respectivos de Sensitividade e Especificidade, é possível construir uma medida mais robusta, chamada ROC (*Receiver Operating Characteristic*).

A área sob a curva ROC (*Area under the ROC Curve – AUC*) representa a capacidade do modelo distinguir, em diferentes pontos de corte, entre os indivíduos para os quais o evento de interesse ocorreu e aqueles para os quais não ocorreu.

Por ter essa característica de mapear diversos pontos de corte ao mesmo tempo, a AUC é considerada uma medida de poder preditivo mais robusta que a *Accuracy* (HOSMER e LEMESHOW, 2000).

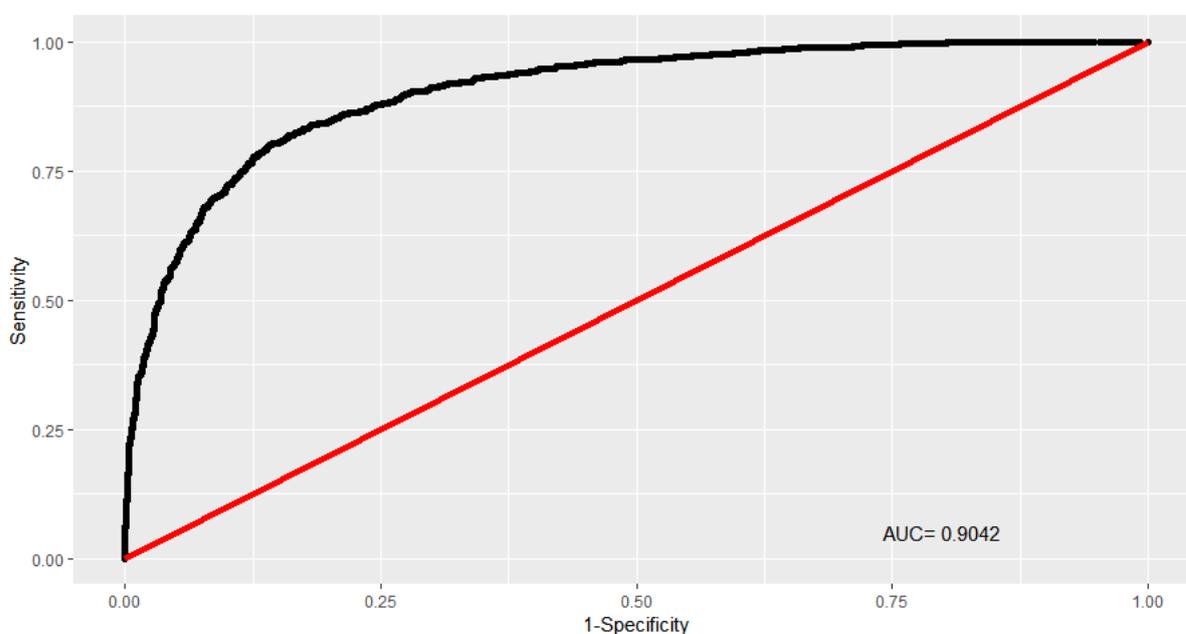
Fávero (2009) explica que quanto maior a área abaixo da curva ROC, que vai de 0 a 1, maior é a capacidade do modelo discriminar o evento de interesse. De outro lado, quanto mais próxima a curva ROC estiver da reta diagonal, pior é seu poder preditivo.

Hosmer e Lemeshow (2000) apresentam os seguintes valores de referência para interpretação da AUC:

- AUC \leq 0.5 Nenhuma discriminação.
- AUC entre 0.7 e 0.8 Discriminação aceitável
- AUC entre 0.8 e 0.9 Discriminação excelente
- AUC \geq 0.9 **Discriminação excepcional**

O *Gráfico 13* apresenta a curva ROC do modelo da pesquisa, criada através da função `rocplot` do pacote **Deducer** do Software R (FELLOWS, 2012).

Gráfico 13: ROC Curve do modelo.



Fonte: Elaborado pelo autor da pesquisa.

Como é possível observar, a *AUC* calculada para o modelo foi de **0.9042**, e de acordo a referência, o modelo é capaz de discriminar de forma excepcional a ocorrência do evento de interesse.

2.2.5.4 Pseudo R^2

Além da *Accuracy* e da *Area under the ROC Curve*, uma medida muito utilizada para aferição da capacidade preditiva de um modelo *Logit* é a chamada *Pseudo R^2* .

De acordo com Lattin, Carroll e Green (2011), a *Pseudo R^2* é uma medida interpretável como a fração da incerteza explicada por um modelo em relação ao modelo com incerteza máxima (aquele sem nenhuma variável explicativa).

Os autores apontam que diferentemente da Regressão Linear, na Regressão Logística é incomum haver valores de *Pseudo R^2* próximos de 1.0 e que **valores no intervalo entre 0.3 e 0.5 são frequentemente descritos como ajustes excelentes**.

Diferentes autores propuseram fórmulas de cálculo do *Pseudo R^2* na Regressão Logística. Allison (2014) lista as fórmulas mais utilizadas: McFadden, McFaddenAdjusted, CoxSnell, Nagelkerke e Tjur²⁶.

A **Tabela 32** apresenta as medidas de *Pseudo R^2* do modelo da pesquisa, obtidas através da função **PseudoR2** do pacote **DescTools** do Software R (SIGNORELL et al., 2018).

Tabela 32: Pseudo R^2 do modelo.

<i>Pseudo R^2</i>	Resultado	Acima de 0.3?
McFadden	0.421	Sim
McFaddenAdj	0.374	Sim
CoxSnell	0.378	Sim
Nagelkerke	0.559	Sim
Tjur	0.460	Sim

Fonte: Elaborado pelo autor da pesquisa.

Como é possível observar, **em todos os casos o resultado pode ser considerado excelente, segundo a referência de Lattin, Carroll e Green (2011)**.

²⁶ Mais detalhes sobre essas medidas e suas fórmulas podem ser obtidos em Allison (2014).

2.2.5.5 Síntese das medidas do modelo

O resultado da *Accuracy* foi de 85,81%, que implica apenas 14,19% de erro de classificação (*Misclassification rate*).

Na técnica *Repeated K-fold Cross Validation*, em que o processo de divisão da amostra em 5 grupos – com um ocultado do modelo – ocorreu 10 vezes, a *Accuracy* se manteve elevada em 84,45%.

A *Area under the ROC Curve* do modelo alcançou 90,42%, o que pode ser considerado como discriminação excepcional, segundo Hosmer e Lemeshow (2000).

Por fim, nas 5 diferentes fórmulas mais utilizadas de *Pseudo R²*, o menor resultado foi de 37,4%, considerado excelente na referência apontada por Lattin, Carroll e Green (2011).

Diante disto, **pode-se inferir que o modelo apresentado na pesquisa é capaz de ser estendido para além da amostra, ser utilizado no restante da população, e identificar com alto grau de acerto a ocorrência do evento de interesse**, o que poderia ser de grande auxílio na ampliação da capacidade do Processo de Cassação por Inatividade.

2.2.6 Resultados e discussão

Os principais estudos abordados na discussão teórica defendem que a decisão de sonegar tributos não é puramente econômico-racional, pois diversos fatores a influenciam. De acordo com o modelo conceitual da **Pirâmide de Conformidade**, os fatores de influência podem ser agrupados como:

- **Econômico-financeiros.**
- **Psicológicos.**
- **Sociológicos.**
- **De ambiente.**

Braithwaite (2003a e 2003b) e Kirchler (2007) defendem que os contribuintes devem ser segmentados levando-se tais fatores em consideração, para que o tratamento correto possa ser dado a cada um. Segundo os autores, tratar contribuintes mal intencionados com benevolência ou contribuintes bem intencionados com rigor desmedido pode, em ambos os

casos, aumentar a *não conformidade*. A chave para o incremento do cumprimento tributário seria, portanto, aplicar o tratamento adequado a cada contribuinte de acordo com seu perfil.

Nessa esteira, a presente pesquisa argumentou que os contribuintes do ICMS do Estado de São Paulo, quanto submetidos a apuração de inatividade pelo processo de Cassação por Inatividade Presumida, não são segmentados entre aqueles que apenas estão inativos e aqueles que além da inatividade cometeram outros ilícitos tributários.

A não segmentação dificulta a análise de solicitações de novas inscrições, pois não gera informações do comportamento tributário pregresso dos sócios e contadores em relação aos estabelecimentos cassados. Em função disso, o mesmo tratamento é dado a todos os contribuintes quando da análise de novos pedidos de inscrição.

Na esperança de encontrar uma solução que mantivesse as vantagens do processo de Cassação por Inatividade Presumida, especialmente aquela relacionada à capacidade de processamento de dados em grandes volumes, e ainda o tornasse capaz de segmentar os contribuintes de acordo com seu comportamento tributário, foi conduzida a pesquisa em questão, que se propôs a criar um modelo estatístico capaz de segmentar esses contribuintes.

Para que tal segmentação fosse possível, com base nos fatores de influência no comportamento tributário apontados pela teoria, procurou-se identificar quais características (variáveis) poderiam refletir tais fatores, de forma que a questão de pesquisa **“quais características permitem segmentar em perfis de conformidade os contribuintes do ICMS do estado de São Paulo que foram submetidos a procedimento de apuração de inatividade?”** pudesse ser respondida.

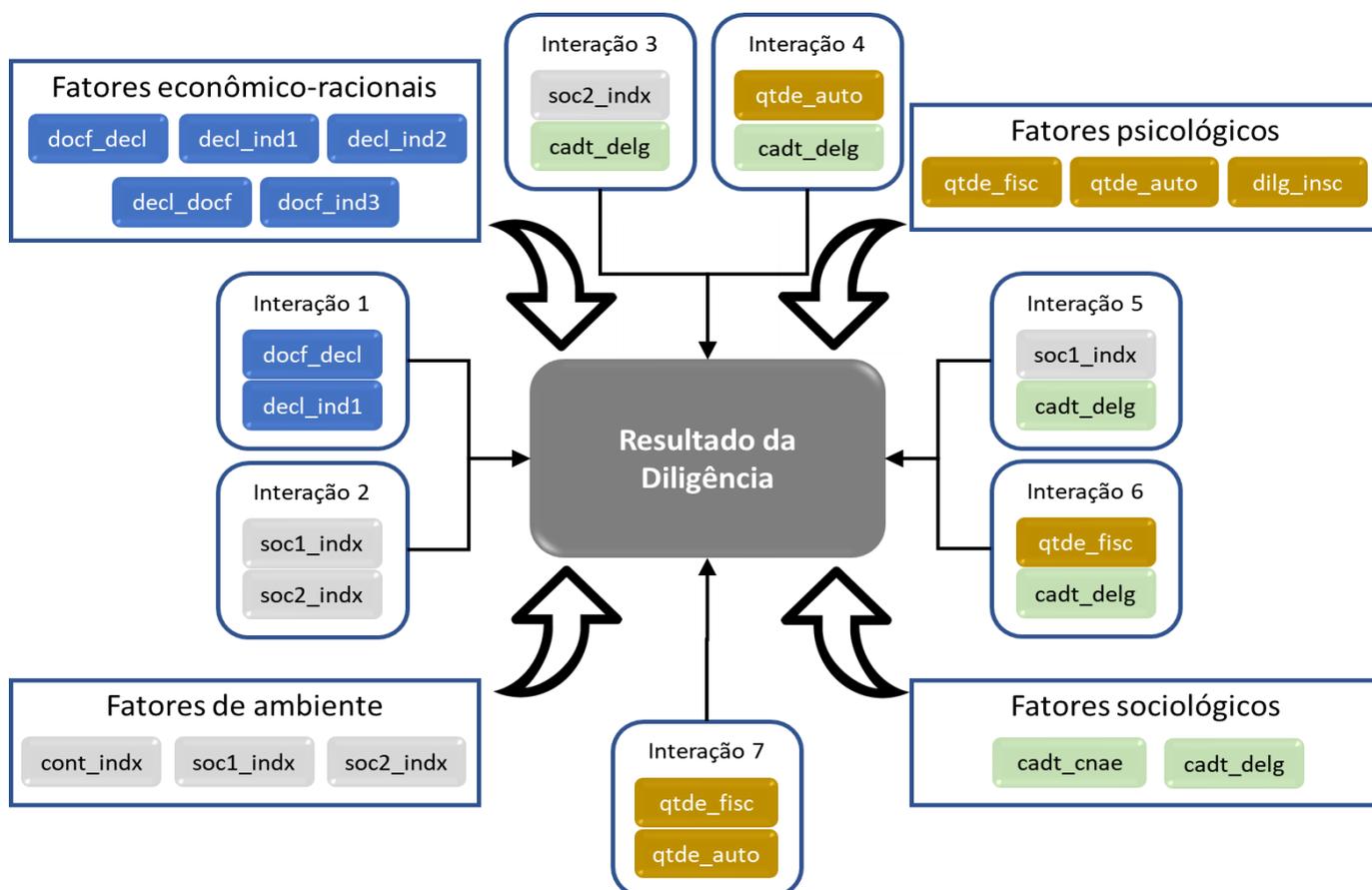
Para isso, buscou-se uma amostra de contribuintes que, assim como aqueles Cassados por Inatividade Presumida, também tivessem sido submetidos a procedimento de apuração de inatividade, mas que, adicionalmente ao resultado da apuração, tivessem também informações relativas à ocorrência ou não de outras ilicitudes tributárias.

Selecionou-se, então, 4.407 contribuintes do ICMS que tiveram apuração de inatividade conduzida por Diligência Fiscal da Secretaria da Fazenda do Estado de São Paulo, entre os anos de 2015 e 2017.

A partir do resultado da diligência, buscou-se construir um modelo estatístico que, dentre as variáveis (características) disponíveis na amostra, encontrasse aquelas com a maior capacidade de explicar e predizer o respectivo resultado. Para tanto, optou-se pela **técnica de Análise de Dados Multivariados denominada Regressão Logística – *Logit***, que, segundo Hosmer e Lemeshow (2000), é capaz de criar modelos extremamente flexíveis e que oferecem um ótimo ajuste frente a outras técnicas de *Machine Learning*.

O modelo apontou como variáveis e interações de maior influência na predição do resultado da diligência aquelas exibidas anteriormente na *Tabela 24*. A relação dessas variáveis e interações com os respectivos fatores de influência pode ser vista na *Figura 29*.

Figura 29: Variáveis e interações do modelo



Fonte: Elaborado pelo autor da pesquisa.

Pela *Figura 29* é possível observar que os resultados do modelo estatístico criado na pesquisa corroboram os estudos elencados por Kirchler (2007), que afirmam que a decisão de sonegar está ligada a diferentes fatores. Por isso, foram mapeadas variáveis representativas de cada um daqueles apontados como relevantes pelo modelo conceitual da Pirâmide de Conformidade.

O grupo de **fatores econômico-financeiros** representa a decisão racional do indivíduo quando confrontado com a opção de sonegar ou não sonegar. Para representa-lo, foram mapeadas diversas variáveis que representam informações fiscais e financeiras, que possivelmente refletem ou influenciam o comportamento do contribuinte frente à sonegação.

A **Figura 29** mostra que as seguintes variáveis foram relevantes para o modelo estatístico:

- docf_decl – Índice de ausência de lançamento de débitos
- decl_ind1 – Índice de declarações
- decl_ind2 – Índice de crédito e débito das declarações
- decl_docf – Índice de ausência de lastro de créditos
- docf_ind3 – Índice de ausência de lastro de saídas em relação a entradas

O **Índice de ausência de lançamento de débitos** e o **Índice de declarações** buscaram refletir o comportamento fiscal do contribuinte no tocante aos Atos de Omissão, definidos por Kirchler (2007). Segundo o autor, Atos de Omissão implicam em ocultar informações das autoridades fiscais para diminuir o imposto devido. Nesse sentido, as variáveis em questão buscam mapear, por exemplo, a falta de lançamento de operações de venda nas declarações de imposto, ou até mesmo a completa omissão em declarar.

O **Índice de ausência de lastro de créditos** buscou refletir o comportamento fiscal do contribuinte no tocante aos Atos de Comissão, também definidos por Kirchler (2007). Segundo o autor, Atos de Comissão implicam em gerar valores de despesa sem lastro real para diminuir o imposto devido. Nessa esteira, essa variável busca mapear o lançamento de créditos nas declarações que não tenham lastro em despesas reais com compras de produtos.

O **Índice de crédito e débito das declarações** e o **Índice de ausência de lastro de saídas em relação a entradas** buscaram, por fim, refletir tanto Atos de Omissão quanto Atos de Comissão. O primeiro busca mapear, por exemplo, a relação entre créditos e débitos nas declarações, em busca de padrões suspeitos. O último, por sua vez, busca o mesmo em relação às entradas (compras) e saídas (vendas) de mercadorias.

Em ambos os casos, podem ser identificados padrões de Omissão, em que vendas são omitidas, ou Comissão, em que compras inexistentes são declaradas. Adicionalmente, podem ser identificados padrões de vendas inexistentes declaradas ou compras omitidas.

A relevância dessas variáveis, confirmada pelo modelo, é um indicativo de que houve evidência estatística suficiente para encontrar padrões capazes de auxiliar na predição do comportamento tributário.

De forma geral, a relevância das variáveis ligadas a fatores econômico-financeiros oferece evidências de que quando fazem cálculos utilitaristas em relação ao custo do montante

devido de impostos, alguns contribuintes podem optar pela sonegação se tiverem a impressão de que não serão punidos ou que o risco da punição compensaria. A opção pela conformidade, mesmo quando o cálculo racional aponta a sonegação como melhor opção é, de acordo com a teoria adotada nesta dissertação, devida aos outros fatores, que são analisados a seguir.

O grupo de **fatores psicológicos** representa o grau de aversão ao risco do contribuinte e pode servir como contraponto ao ímpeto sonegatório. Um contribuinte que possui um alto grau de aversão ao risco pode ser dissuadido da decisão de sonegar mesmo que o cálculo racional a aponte como compensadora.

Os estudos explorados na discussão teórica mostraram que o grau de aversão ao risco no tocante à sonegação depende grandemente da percepção do contribuinte em relação à atuação do Fisco. Se o contribuinte tiver a percepção que qualquer deslize será detectado e punido, seu grau de aversão ao risco deverá aumentar, da mesma forma que poderá diminuir se houver percepção crescente de impunidade.

A **Figura 29** mostrou que as seguintes variáveis foram relevantes para o modelo estatístico no tocante aos fatores psicológicos:

- qtde_fisc – Quantidade de fiscalizações
- qtde_auto – Quantidade de autos de infração
- dilig_insc – Tempo entre inscrição e diligência

A **Quantidade de fiscalizações** e a **Quantidade de autos de infração** buscaram refletir a reação do contribuinte em relação à atuação do Fisco com diligências efetuadas e multas aplicadas.

Um contribuinte decidido a sonegar (baixíssimo grau de aversão ao risco), como aqueles no topo da Pirâmide de Conformidade da **Figura 10**, poderá continuar ou até aumentar as práticas sonegatórias quando confrontado pelo fisco com fiscalizações e punições.

Nesses casos, o modelo da Pirâmide de Conformidade sugere a aplicação de todo o rigor possível da lei para que o contribuinte seja impedido de continuar com as práticas lesivas. Exatamente para isso, o presente estudo busca segregar esses contribuintes para que sua inscrição no cadastro possa ser detalhadamente analisada e possivelmente negada quando comprovadas as intenções fraudulentárias.

O **Gráfico 11** demonstrou que ambas as variáveis estão positivamente relacionadas à ocorrência do evento de interesse, ou seja, há evidências de que contribuintes reiteradamente punidos pelo fisco após a detecção de ilicitudes em diligências possuem uma maior

probabilidade de ter sua inscrição no cadastro eventualmente cassada após diligência de apuração de inatividade.

A detecção e utilização de padrões como esse nos contribuintes Cassados por Inatividade Presumida através do modelo aqui proposto poderia, portanto, impedir a continuidade das práticas ilícitas através da não concessão de inscrições de novos estabelecimentos dos mesmos indivíduos.

O **Tempo entre inscrição e diligência** teve correlação positiva com a ocorrência do evento de interesse, indicando que houve maior probabilidade de sonegação com o aumento do tempo de operação do estabelecimento.

O grupo de **fatores de ambiente** representa as influências recebidas do meio em que o contribuinte atua. Alguns dos estudos citados nesse texto trouxeram evidências de que contribuintes inseridos em atividades ou locais em que a sonegação é maior, serão mais propensos a ela.

Para identificar tais influências, foram mapeadas diversas variáveis demográficas para verificação pelo modelo. As que tiveram influência estatisticamente confirmada na amostra de estudo foram as seguintes:

- `cdt_cnae` – Código de atividade econômica principal
- `cdt_delg` – Delegacia Regional Tributária de jurisdição

O **Código de atividade econômica principal** reflete o principal ramo de atividade do contribuinte. Como apontado na seção 2.2.2.4.3 *Variável código de atividade econômica principal*, os ramos de atividade em que a ocorrência do evento de interesse foi maior, foram:

- 7 – COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS;
- 3 – INDÚSTRIAS DE TRANSFORMAÇÃO e
- 8 – TRANSPORTE, ARMAZENAGEM E CORREIO.

Embora a busca de explicações para a ocorrência maior nas atividades acima fuja ao escopo da pesquisa, o fato dessa variável ter tido sua relevância confirmada pelo modelo sugere que há evidências de que em algumas atividades, a ocorrência de práticas sonegatórias parece ser maior e estas deveriam, portanto, ser observadas.

A **Delegacia Regional Tributária de jurisdição – DRT** busca refletir a localização geográfica do contribuinte. Essa foi uma das principais variáveis de influência no modelo,

conforme apontou o **Gráfico 10**. Adicionalmente, foi a variável com maior número de interações estatisticamente significantes.

Em virtude disso, seria interessante a criação de modelos por DRT em estudos futuros, de forma que os padrões de influência das outras variáveis no evento de interesse pudessem ser comparados e modelos específicos otimizados para cada uma delas.

Por fim, o grupo de **fatores sociológicos** representa as influências recebidas de outros indivíduos. Como discutido na abordagem teórica, contribuintes cuja rede de relações seja composta primordialmente por indivíduos com tendência à sonegação, terão maior propensão para refletir essa tendência.

Por isso, foram selecionados para cada estabelecimento o contador e os dois sócios principais. Adicionalmente, buscou-se mapear o histórico destes indivíduos em outros estabelecimentos através das variáveis:

- cont_indx – Índice contador
- soc1_indx – Índice de irregularidade sócio 1
- soc2_indx – Índice de irregularidade sócio 2

O peso de influência das três variáveis no modelo foi relativamente pequeno se comparado a outras variáveis, conforme se viu no **Gráfico 10**. Nada obstante, a significância estatística constatada indica que a presença delas no modelo pode influenciar positivamente as outras variáveis, razão pela qual sua importância não pode ser minimizada.

O **Gráfico 11** demonstrou que as três variáveis estão positivamente relacionadas à ocorrência do evento de interesse, ou seja, há evidências de que estabelecimentos expostos a contadores e sócios com mais irregularidades possuem maior probabilidade de incorrer em ilicitudes tributárias.

Além das variáveis de influência discutidas, é possível verificar na **Figura 29** que **algumas interações entre variáveis foram significantes e contribuíram** com o modelo estatístico. Na **Interação 1** é possível verificar que o modelo identificou que a interação entre as variáveis **Índice de crédito e débito das declarações** e o **Índice de ausência de lastro de saídas em relação a entradas**, ambas relacionadas a fatores econômico-financeiros e à identificação de atos de omissão, contribuíram para o modelo.

Na mesma linha, a **Interação 2**, entre os dois sócios principais de cada estabelecimento, ambos representativos de fatores sociológicos, contribuíram para a qualidade do ajuste final do modelo através da diminuição de seu desvio residual.

A *Interação 7*, por sua vez, abriga duas variáveis do grupo de fatores psicológicos, quais sejam: **Quantidade de fiscalizações** e a **Quantidade de autos de infração**. Como ambas estão intimamente ligadas à atuação do Fisco, foram detectados padrões na interação entre as duas que podem refletir a postura do contribuinte frente às Autoridades Tributárias.

As demais interações estão todas relacionadas à variável **Delegacia Regional Tributária de jurisdição – DRT**, o que reforça a hipótese de que um modelo estatístico por delegacia poderia trazer grandes contribuições na detecção de padrões adicionais entre as outras variáveis da pesquisa.

3 PLANO DE AÇÃO

A partir da discussão teórica apresentada e do estudo empírico conduzido, foi montada esta proposta de intervenção no âmbito da Secretaria da Fazenda do Estado de São Paulo.

O objetivo principal dessa proposta é oferecer ideias e sugestões de utilização do modelo criado na pesquisa para segmentar os contribuintes Cassados por Inatividade Presumida.

Na seção 2.2.1 *Metodologia* foi apresentada a população deste estudo, que compreende os **contribuintes do ICMS do Estado de São Paulo enquadrados no Regime Periódico de Apuração e que tenham sido submetidos a procedimento de apuração de inatividade entre os anos de 2015 e 2017**.

Tais contribuintes se dividem em dois grupos: aqueles cuja apuração de inatividade se deu através de Diligência Fiscal e aqueles cuja apuração se deu através do Processo de Cassação por Inatividade Presumida.

Como discutido, somente os contribuintes que tiveram apuração por Diligência Fiscal possuem informações relativas a irregularidades tributárias cometidas e, portanto, são os únicos para os quais o processo atual permite segmentação em perfis, na forma apresentada nesta dissertação. Por esse motivo, esses contribuintes compuseram a amostra de pesquisa, para servirem de guia para o modelo estatístico construído.

Os contribuintes Cassados por Inatividade Presumida, que integram o restante da população, ainda não são segmentados pela falta de informações relativas à ocorrência de irregularidades tributárias. A falta de segmentação no âmbito desses contribuintes não permite que pedidos de inscrição de novos estabelecimentos dos mesmos sócios e contabilistas sejam automaticamente segregados para que a análise detalhada seja feita.

Essa segmentação seria de grande auxílio na diminuição da brecha tributária no âmbito do ICMS no Estado de São Paulo, pois como argumentado por Viol (2015), a maneira mais eficiente de impedir o comportamento tributário lesivo é no momento da inscrição no cadastro, porque impede que contribuintes com atitudes fraudadoras tenham os registros cadastrais de suas empresas deferidos e prossigam com suas práticas dolosas.

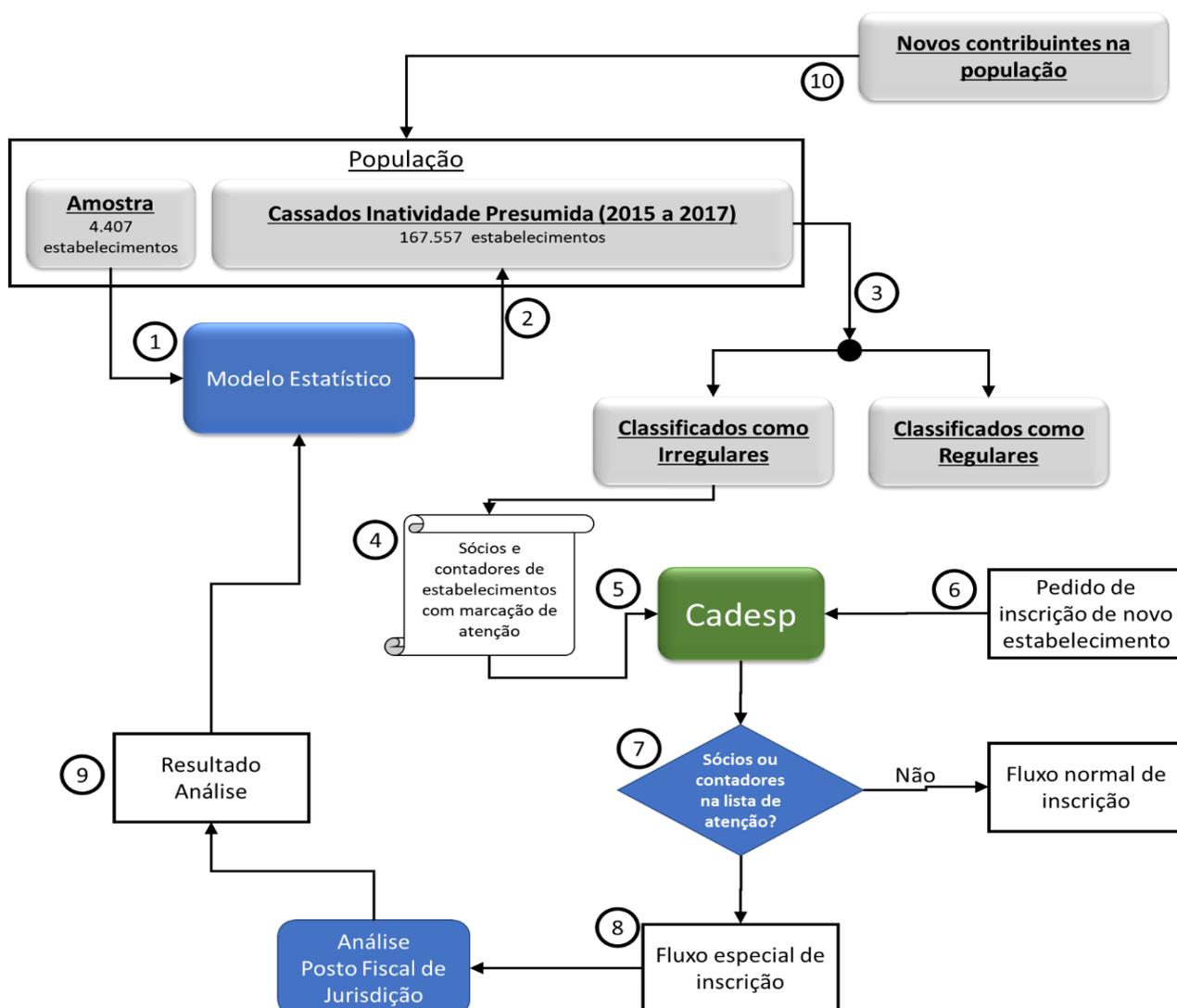
A pesquisa demonstrou que o modelo *Logit* criado oferece evidências estatísticas de que sua capacidade predição de mais de 84% de acerto seria mantida se aplicada ao restante da população. Adicionalmente, o erro de classificação, que se restringe a menos de 16% dos casos, poderia ser mitigado através da não recusa automática de pedidos de inscrição de indivíduos classificados pelo modelo como irregulares. Ao contrário, esses pedidos seriam

direcionados para análise detalhada, que buscaria subsídios adicionais para confirmar a situação de intenção fraudulenta.

3.1 Solução proposta

Um desenho esquemático da solução proposta pela pesquisa pode ser visualizado na *Figura 30*.

Figura 30: Fluxo da solução proposta



Fonte: Elaborado pelo autor da pesquisa.

Como é possível observar na *Figura 30*, o modelo construído através do aprendizado com a amostra (1) seria aplicado ao restante da população (2). O modelo geraria como saída a

categorização dos indivíduos da população em dois grupos (3): aqueles com indícios de sonegação (*I – Irregular*) e aqueles sem indícios de sonegação (*0 – Regular*).

As partes relacionadas (sócios e contadores) aos estabelecimentos categorizados como possivelmente irregulares receberiam uma marcação de atenção (4), que alimentaria o Sistema de Cadastro de Contribuintes de ICMS do Estado de São Paulo – Cadesp (5).

Ao receber pedido de inscrição de novo estabelecimento (6), o sistema Cadesp verificaria se os sócios e contador do estabelecimento fazem parte da lista de atenção (7). Em caso negativo, o processo de inscrição seguiria seu curso normal, como é atualmente, de eventual deferimento automático. Em caso positivo, no entanto, o Cadesp encaminharia o pedido para análise fiscal (8), informando ao Posto Fiscal de jurisdição se tratar de pedido de inscrição de estabelecimento que possui sócio ou contador com marcação de atenção atribuída pelo modelo estatístico criado nesta dissertação.

O Posto Fiscal se manifestaria a respeito do pedido de inscrição (9) e o resultado da análise seria retroalimentado no modelo estatístico, que se aperfeiçoaria através do aprendizado das novas informações.

O modelo seria também retroalimentado com novos indivíduos inseridos na população (10). Os indivíduos que entrassem na população com resultado da diligência fiscal seriam utilizados para aperfeiçoar o aprendizado do modelo. Os demais, novos Cassados por Inatividade Presumida, seriam categorizados pelo modelo aprimorado e passariam a integrar a lista de atenção em caso de classificação como possíveis irregulares.

Dessa forma, o modelo seria interativo e em constante evolução, e considerando a discussão teórica apresentada de que a percepção que o contribuinte possui em relação à atuação do Fisco também é vetor de influência em seu comportamento tributário, acredita-se que tal medida teria efeito duplo na diminuição da brecha tributária. Esse efeito se daria em uma vertente, pela não concessão de novas inscrições a contribuintes com intenção dolosa, e em outra, pelo aumento da percepção de um Fisco presente, que busca a justiça fiscal e a concorrência leal através do combate à sonegação.

3.2 Operacionalização da solução

Para detalhar os passos necessários para implementação da solução proposta neste capítulo, optou-se pela utilização da ferramenta japonesa denominada **5W2H**.

Daychouw (2007) e Meira (2003) ensinam que essa ferramenta auxilia na elaboração de planos de ação, pois determina de forma detalhada:

- O que será feito (objetivo) – **What?**
- Por que será feito (justificativa) – **Why?**
- Onde será feito (local) – **Where?**
- Quem fará (responsável) – **Who?**
- Quando será feito (cronograma) – **When?**
- Como será feito (método) – **How?**
- Quanto custará (custo dos recursos) – **How much?**

Nesse sentido, utilizando-se como os passos listados na *Figura 30*, propõe-se as seguintes ações para implementação da solução aqui sugerida. Cabe frisar que se tratam apenas de sugestões, que podem ser ajustadas para melhor se adequarem à realidade da Secretaria da Fazenda do Estado de São Paulo.

Quadro 7: Passos para implantação do plano de ação

What? O que?	Why? Por que?	Where? Onde?	Who? Quem?	When? Quando?	How? Como?	How much? Quanto?
Suporte normativo ao processo.	Para verificar a necessidade de alteração de resoluções ou portarias para servir de apoio nos casos de indeferimento de novas inscrições.	Jurídico Sefaz/SP	Pelos servidores das áreas do jurídico.	Início da implantação.	Através da emissão de pareceres e sugestões de redação.	Custo de 2 servidores por 2 semanas.
Carga periódica do modelo com os dados de aprendizado (amostra)	Para alimentar o modelo periodicamente com dados já existentes e novos.	Diretoria de Informações – DI.	Autor da pesquisa, que é servidor lotado na DI.	No início da implantação e depois 1x a cada 3 meses.	Através da seleção e preparação dos dados, e sua carga no Software R.	Custo de 1 servidor por 2 semanas a cada 3 meses.
Execução do modelo na população de Cassados por Inatividade Presumida	Para gerar, periodicamente, a lista de atenção.	Diretoria de Informações – DI.	Autor da pesquisa, que é servidor lotado na DI.	No início da implantação e depois 1x a cada 3 meses.	Através da seleção e preparação dos dados, e sua carga no Software R.	Custo de 1 servidor por 2 semanas a cada 3 meses.
Adaptação do Sistema Cadesp para ser alimentado pela lista de atenção e encaminhar mensagem personalizada ao Posto Fiscal.	Para que a lista de atenção possa ser processada corretamente no sistema Cadesp e uma mensagem personalizada seja encaminhada ao Posto Fiscal de Jurisdição quando o pedido de inscrição for encaminhado para análise especial.	Diretoria de Informações – DI e Departamento de Tecnologia da Informação.	Servidores de ambos os departamentos, responsáveis pelo sistema Cadesp.	Início da implantação.	Através da programação do sistema.	Custo de 2 servidores por 4 semanas.
Análise de pedidos de inscrição no fluxo especial.	Para que o pedido de inscrição possa ser analisado no Posto Fiscal de jurisdição levando-se em consideração a lista de atenção gerada pelo modelo.	Posto Fiscal de jurisdição	Servidores responsáveis pela análise de pedidos cadastrais	Sempre que houver pedido de inscrição direcionado para o fluxo especial	Através da análise detalhada dos estabelecimentos ativos e inativos dos sócios e contador que solicitaram nova inscrição.	Custo de 1 servidor por 1 dia a cada semana.
Geração de relatório para mostrar resultado da análise do Posto Fiscal.	Para que seja possível retroalimentar o modelo com o resultado da análise do Posto Fiscal.	Diretoria de Informações – DI.	Autor da pesquisa, que é servidor lotado na DI.	No início da implantação e depois 1x a cada 3 meses.	Através da seleção e preparação dos dados, e sua carga no Software R.	Custo de 1 servidor por 1 semana a cada 3 meses.

Fonte: Elaborado pelo autor da pesquisa.

CONSIDERAÇÕES FINAIS

O presente estudo apresentou uma pesquisa que buscou mecanismos de segmentação em uma população pouco estudada, qual seja: a de contribuintes do ICMS que sofreram apuração de inatividade.

O benefício potencial da segmentação dos indivíduos dessa população reside no fato de que futuros pedidos de inscrições no Cadastro do ICMS feitos pelos sócios ou contabilistas dos estabelecimentos cassados por inatividade podem ser direcionados para análise criteriosa, com informações que permitam a identificação de indivíduos com intenção fraudulenta.

Para tornar possível essa segmentação, buscou-se na literatura os principais fatores que influenciam o comportamento tributário. Com base em tais fatores, identificou-se na amostra de pesquisa características associadas a eles e estas compuseram um modelo capaz de prever com grau suficiente de acerto o resultado de uma Diligência Fiscal que aponte irregularidades tributárias além da simples inatividade do estabelecimento.

Identificadas as variáveis (características) relevantes para o modelo, foi necessário verificar se os resultados da predição eram confiáveis e não estariam restritos apenas aos dados a partir dos quais o modelo foi construído, ou seja, que não estariam restritos à amostra. Para isso, a qualidade do ajuste do modelo foi verificada utilizando-se os principais testes estatísticos aplicáveis à Regressão Logística, quais sejam:

- **Validação de pressupostos**
 - Ausência de Multicolinearidade,
 - Ausência de Influential Values e
 - Relação linear entre as variáveis explicativas contínuas e a variável dependente.

- **Medidas de ajustes individuais**
 - ΔX^2 – Delta chi-square,
 - ΔD – Delta deviance e
 - $\Delta \hat{\beta}$ – Delta beta.

- **Testes de ajuste geral**
 - Deviance e Pearson Chi-square Statistic,
 - Hosmer-Lemeshow test,
 - Osius and Rojek test,
 - Stukel's test e
 - Kolmogorov-Smirnov test.

O modelo foi bem sucedido em todos os testes e os resultados, discutidos na seção 2.2.4 *Validação do modelo*, permitem afirmar que existem evidências estatísticas de que a qualidade do ajuste possibilita sua utilização no restante da população.

Encontradas as evidências necessárias de que o ajuste do modelo estava correto, passou-se à análise de qual seria seu poder de predição do resultado da diligência, que é o que permitiria, em última análise, segmentar com o grau necessário de acuracidade os contribuintes Cassados por Inatividade Presumida.

Para isso, foram analisadas as principais métricas de capacidade de predição de um modelo *Logit*, quais sejam:

- **Accuracy,**
- **Accuracy em Repeated K-fold Cross Validation,**
- **Area under the ROC Curve e**
- **Pseudo R².**

Em todas as métricas, o modelo alcançou excelentes resultados, com **destaque para a Accuracy em Repeated K-fold Cross Validation**, que mede o grau de precisão com que o modelo consegue acertar o resultado da diligência fiscal em uma base de dados na qual não foi treinado. O resultado obtido, de **84,45%**, limitou a **taxa de erros a 15,55%**.

Esse resultado permitiria elevar a capacidade de segmentação dos contribuintes Cassados por Inatividade Presumida de 0% de precisão, visto que atualmente não é possível segmentá-los pelos motivos expostos durante o estudo, para 84,45%, o que na visão do autor da pesquisa, seria de grande valia para a diminuição da brecha tributária no ICMS no Estado de São Paulo.

Em vista dos resultados obtidos, considera-se que a questão de pesquisa foi respondida, conforme amplamente detalhado na seção 2.2.6 *Resultados e discussão*, e que todos os objetivos traçados na pesquisa foram alcançados, conforme se discute a seguir.

Objetivos da pesquisa:

- Determinar que modelo de conformidade proposto na teoria seria apropriado para segmentar corretamente a população estudada.
 - O modelo conceitual considerado o mais apropriado na discussão teórica foi o da Pirâmide de Conformidade.
- Identificar quais fatores são significativos para determinação do perfil do contribuinte de acordo com o modelo escolhido.
 - Os fatores significativos identificados foram: Econômico-financeiros, Psicológicos, Sociológicos e de Ambiente.
- Encontrar, a partir de análise estatística, quais características identificadas na população de estudo poderiam melhor refletir tais fatores.
 - As características identificadas são aquelas representadas pelas variáveis listadas na *Figura 29*.
- Construir um modelo estatístico que contribua para a definição do perfil de conformidade de contribuintes que foram inabilitados do Cadastro do ICMS.
 - O modelo estatístico foi construído com sucesso e seus aspectos e resultados amplamente discutidos nesse texto.

Para que as análises e resultados alcançados na pesquisa possam ser utilizados de forma efetiva pela Secretaria da Fazenda do Estado de São Paulo e, eventualmente outras Administrações Tributárias, apresentou-se uma sugestão de intervenção na seção 3 *PLANO DE AÇÃO*.

A discussão teórica e a pesquisa empírica, associadas à proposta de intervenção, tiveram como princípio basilar a busca da justiça tributária, através do combate à sonegação e diminuição da brecha tributária, e estímulo ao aumento da conformidade no âmbito daqueles contribuintes dispostos a contribuir com uma sociedade mais justa e isonômica.

Como sugestão para estudos futuros, reforça-se a discussão apresentada na seção 2.2.6 *Resultados e discussão* de necessidade de construção de um modelo estatístico por Delegacia Regional Tributária e comparação de seus resultados e padrões detectados.

Adicionalmente, propõe-se também a criação de modelos com outras técnicas de **Machine Learning** e **Inteligência Artificial** como Árvores de Decisão (*Decision Tree*), Floresta Aleatória (*Random Forest*), Máquina de Vetores de Suporte (*Support Vector Machine*), Redes Neurais Artificiais (*Artificial Neural Networks*) e *Boosting*. Esses modelos poderiam ser comparados com o criado nesta dissertação, com vistas ao seu aperfeiçoamento.

Uma outra possibilidade de extensão da pesquisa seria sua condução em outros estados da federação, a fim de que os resultados entre eles pudessem ser comparados.

Por fim, no caso de implementação do plano de ação apresentado, sugere-se a documentação dos resultados das análises de pedidos de inscrição que tenham sido originadas a partir da classificação efetuada pelo modelo e a comparação dos resultados com as previsões efetuadas.

REFERÊNCIAS

- ADAMS, Caroline; WEBLEY, Paul. **Small business owners' attitudes on VAT compliance in the UK**. *Journal of Economic Psychology*, v. 22, n. 2, p. 195-216, 2001.
- AFONSO, José Roberto (Coord.). **A renúncia tributária do ICMS no Brasil**. Washington: BID, IDB-DP-327, 2014.
- ALLINGHAM, Michael G.; SANDMO, Agnar. **Income tax evasion: A theoretical analysis**. *Journal of public economics* 1.3-4 (1972): 323-338.
- ALLISON, Paul D. **Measures of fit for logistic regression**. In: *Proceedings of the SAS Global Forum 2014 Conference*. 2014. p. 1-13.
- ALM, James. **Expanding the Theory of Tax Compliance from Individual to Group Motivations**. Tulane University: Tulane Economics Working Paper Series. Working Paper 1309, February 2013.
- ALM, James. **Measuring, explaining, and controlling tax evasion: lessons from theory, experiments, and field studies**. *International Tax and Public Finance*. February 2012, Volume 19, Issue 1, p. 54–77.
- ALM, James; MARTINEZ-VAZQUEZ, J; SCHNEIDER, F. **'Sizing' the problem of the hard-to tax**. In J. Alm, J. Martinez-Vazquez and S. Wallace (eds.), *Taxing the hard-to-tax: Lessons from theory and practice*. Amsterdam: Elsevier. 2004.
- ALM, James; TORGLER, Benno. **Culture differences and tax morale in the United States and in Europe**. *Journal of economic psychology*, v. 27, n. 2, p. 224-246, 2006.
- ANDREONI, James; ERARD, Brian; FEINSTEIN, Jonathan. **Tax compliance**. *Journal of economic literature*, v. 36, n. 2, p. 818-860, 1998.
- ARMITSTEAD, Louise. **Starbucks should face tax inquiry, say MPs**. *Telegraph*. co. uk, 2012.
- ARROW, Kenneth J. **Essays in the theory of risk-bearing**. North-Holland, ch. 3, 1970.
- BALESTRINO, Alessandro; GALMARINI, Umberto. **Imperfect tax compliance and the optimal provision of public goods**. *Bulletin of Economic Research*, v. 55, n. 1, p. 37-52, 2003.
- BARBOZA, Flavio; KIMURA, Herbert e ALTMAN, Edward. **Machine learning models and bankruptcy prediction**. *Expert Systems with Applications*, v. 83, p. 405-417, 2017.
- BECKER, Gary S. **Crime and punishment: An economic approach. The economic dimensions of crime**. Palgrave Macmillan UK, 1968. 13-68.
- BELLO, Zakri. **A statistical comparison of the CAPM to the Fama-French Three Factor Model and the Carhart's Model**. 2008.

BENDAL, R. B.; AFIFI, A. A. **Comparison of stopping rules in forward regression.** *Journal of the American Statistical Association*, v. 72, p. 46-53, 1977.

BEST, D. J.; RAYNER, J. C. W. **Welch's approximate solution for the Behrens–Fisher problem.** *Technometrics*, v. 29, n. 2, p. 205-210, 1987.

BNDES. Banco Nacional de Desenvolvimento. **Guia do financiamento.** 2018. Disponível em: <<https://goo.gl/EWiyne>>. Acesso em: 16 fev. 2018.

BRAITHWAITE, John. **Markets in vice. Markets in virtue.** Leichhardt: The Federation Press. 2005.

BRAITHWAITE, Valerie. **A new approach to tax compliance.** In V. Braithwaite (ed.), *Taxing democracy*, p. 1 – 11. Aldershot: Ashgate. 2003a.

BRAITHWAITE, Valerie. **Dancing with tax authorities: Motivational postures and non-compliant actions.** In V. Braithwaite (ed.), *Taxing democracy*, p. 15 – 39. Aldershot: Ashgate. 2003b.

BRASIL. **Constituição da República Federativa do Brasil (1988).** Brasília: 1988. Disponível em: <<https://goo.gl/X5iMf1>>. Acesso em: 16 fev. 2018.

BRASIL. **Lei Complementar nr. 123, de 14 de dezembro de 2006.** Institui o Estatuto Nacional da Microempresa e da Empresa de Pequeno Porte. Disponível em: <<https://goo.gl/4cePTS>>. Acesso em: 16 fev. 2018.

BRASIL. **Lei nr. 5.172, de 25 de outubro de 1966.** Dispõe sobre o Sistema Tributário Nacional e institui normas gerais de direito tributário aplicáveis à União, Estados e Municípios. Disponível em: <<https://goo.gl/jVxrG1>>. Acesso em: 16 fev. 2018.

BRASIL. **Lei nr. 10.406, de 10 de janeiro de 2002.** Institui o Código Civil. Disponível em: <<https://goo.gl/qrMXvL>>. Acesso em: 16 fev. 2018.

BRASIL. Superior Tribunal de Justiça. Súmula nº 435. **Diário de Justiça da União**, Brasília, 13 maio 2017.

CABRAL, Ana Rita Nascimento. **O Princípio da Capacidade Contributiva: Eficácia e Efetividade. Uma Análise a partir do Conceito Jurídico.** XVIII Congresso Nacional do CONPEDI. São Paulo – SP, 04, 05, 06 e 07 de novembro de 2009. Disponível em: <<https://goo.gl/ivi100>>. Acesso em: 16 fev. 2018.

CARLOS, Américo Fernando Brás. **Impostos - Teoria Geral.** 5 ed. Edições Almedina, 2016.

CARRAZA, Roque Antônio. **ICMS.** 15 ed. São Paulo: Malheiros, 2015.

CHARLET, Alain; OWENS Jeffrey. **An International Perspective on VAT.** *Tax Notes International*. Volume 59, Number 12. September 20, 2010. Disponível em: <<https://goo.gl/ii2GPJ>>. Acesso em: 16 fev. 2018.

COBHAM, Alex. **Tax evasion, tax avoidance and development finance**. Queen Elizabeth House Working Paper Series 129, 2005.

COELHO, Ricardo da S.; Bernardo, Heloísa Pinna; Maia, Rodrigo Guedes. **GUERRA FISCAL: uma análise da evolução da renúncia fiscal praticada pelos Estados Brasileiros**. X Congresso de Administração, Sociedade e Inovação. Petrópolis. 2017.

COELHO, Ricardo da S. et al. **GUERRA FISCAL: um estudo comparativo entre a renúncia fiscal e os indicadores financeiros e econômicos dos estados de Minas Gerais, Rio de Janeiro e Mato Grosso**. XX Semead – Seminários em Administração. São Paulo. 2017

COELHO, Sacha Calmon Navarro. **Curso de direito tributário: (comentários à Constituição e ao Código tributário nacional, artigo por artigo)**. Rio de Janeiro: Forense. 2001

COLE, Alan. **Estate and Inheritance Taxes around the World**. 2015. Disponível em: <<https://goo.gl/HVKpdW>>. Acesso em: 16 fev. 2018.

COLLETT, David. **Modelling binary data**. Chapman & Hall, 2003.

COOK, R. Dennis; PRESCOTT, Philip. **On the accuracy of Bonferroni significance levels for detecting outliers in linear models**. *Technometrics*, v. 23, n. 1, p. 59-63, 1981.

DARDIS, Chris. LogisticDx: **Diagnostic Tests for Models with a Binomial Response**. R package version 0.2. 2015. <https://CRAN.R-project.org/package=LogisticDx>.

DAYCHOUW, M. **40 ferramentas e técnicas de gerenciamento**. Rio de Janeiro: Brasport, 2007.

DOS SANTOS, Aécio Flávio Quintela et al. **Eventos cadastrais e consequências na tributação do ICMS do estado da Bahia**. Disponível em <<https://goo.gl/PUyKDy>>. Acesso em: 16 fev. 2018.

EBRAHIMI, Helia. **Starbucks faces further criticism over its tax practices**. Telegraph. co. uk, 2012.

EBRILL, Liam et al. **The Modern VAT**. 2001. Washington: International Monetary Fund.

ERIKSEN, Knut; FALLAN, Lars. **Tax knowledge and attitudes towards taxation; A report on a quasi-experiment**. *Journal of economic psychology*, v. 17, n. 3, p. 387-402, 1996.

ETZIONI, Amitai. **Moral dimension: Toward a new economics**. Simon and Schuster, 2010.

EUROPEAN COMMISSION. **Taxation and Customs Union: VAT GAP Report**. September, 2016. Disponível em: <<https://goo.gl/8PTekD>>. Acesso em: 16 fev. 2018.

FÁVERO, Luiz Paulo. **Análise de dados: modelagem multivariada para tomada de decisões**. Rio de Janeiro: Elsevier, 2009.

FERRAGUT, Maria Rita. **Responsabilidade tributária e o Código Civil de 2002**. Direito, v. 2, n. 9, 2015.

FERREIRA, Marco A. M.; CELSO, Alex S. S.; NETO, João E. B. **Aplicação do modelo logit binomial na análise do risco de crédito em instituições bancárias**. Revista de Negócios. Blumenau, v17, n.1, p.41 – 59, Janeiro/ Março 2012.

FIELDHOUSE, Andrew. **A review of the economic research on the effects of raising ordinary income tax rates**. New York: Economic Policy Institute, 2013.

FOX, John; WEISBERG. **An {R} Companion to Applied Regression**, Second Edition. Thousand Oaks CA: Sage. Sanford. 2011.

GERHARDT, T. E., SILVEIRA, D. T. **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009.

GIAMBIAGI, Fábio; ALÉM, Ana Cláudia. **Finanças Públicas**. 2011. 4 ed. Rio de Janeiro: Elsevier.

GIL, Antonio Carlos. **Como Elaborar Projetos de Pesquisa**. São Paulo: Atlas, 1991.

HMRC. **Measuring tax gaps 2016 edition: Tax gap estimates for 2014-15**. Official Statistics release 20 October 2016. Disponível em: <<https://goo.gl/Yc1s6C>>. Acesso em: 16 fev. 2018.

HOSMER, David W.; Lemeshow, Stanley. **Applied Logistic Regression**. 2. ed. John Wiley & Sons, 2000.

JAMES, Simon; ALLEY, Clinton. **Tax compliance, self-assessment and tax administration**. Journal of Finance and Management in Public Services, p. 27– 42. 2002.

KASSAMBARA, Alboukadel. **Machine Learning Essentials: Practical Guide in R**. STHDA, 2018.

KENNEDY, Peter. **A guide to econometrics**. MIT press, 2003.

KING, Sharmila; SHEFFRIN, Steven M. **Tax evasion and equity theory: An investigative approach**. International Tax and Public Finance, v. 9, n. 4, p. 505-521, 2002.

KIRCHLER, Erich. **The Economic Psychology of Tax Behaviour**. Cambridge University Press, 2007.

KIRCHLER, Erich; MACIEJOVSKY, Boris; SCHNEIDER, Friedrich. **Everyday representations of tax avoidance, tax evasion, and tax flight: Do legal differences matter?** Journal of Economic Psychology, v. 24, n. 4, p. 535-553, 2003.

KOHAVI, Ron et al. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. In: Ijcai. 1995. p. 1137-1145.

KUHN, MAX. **caret: Classification and Regression Training**. R package version 6.0-80. 2018. <https://CRAN.R-project.org/package=caret>

LATTIN, James; Carroll, J. Douglas; Green, Paul E. **Análise de dados multivariados**. São Paulo: Cengage Learning, 2011.

LAWLER, John J. **Not just for the money: An economic theory of personal motivation**. *Personnel Psychology*, v. 51, n. 3, p. 740, 1998.

LONG, Susan B.; SWINGEN, Judyth A. **Taxpayer compliance: Setting new agendas for research**. 1991.

MENARD, Scott. **Applied logistic regression analysis**. Sage, 2002.

MEIRA, R. C. **As ferramentas para a melhoria da qualidade**. Porto Alegre: SEBRAE, 2003.

MICKEY, R. M.; GREENLAND, S. **A study of the impact of confounder-selection criteria on effect estimation**. In: *American Journal of Epidemiology*. 111 Market Place, STE 840, Baltimore, MD 21202-6709: Johns Hopkins Univ School Hygiene Pub Health, 1987. p. 737-737.

O'BRIEN, Robert M. **A caution regarding rules of thumb for variance inflation factors**. *Quality & quantity*, v. 41, n. 5, p. 673-690, 2007.

OECD. **Choosing a Broad Base-Low Rate Approach to Taxation**. OECD Tax Policy Studies N° 19, OECD Publishing, 2010.

OECD. **Compliance Risk Management: Managing and Improving Tax Compliance**. Center for Tax Policy and Administration. October, 2004.

OECD. **Glossary of Tax Terms**. 2017b. Disponível em: <<https://goo.gl/CjyEeG>>. Acesso em: 16 fev. 2018.

OECD. **OECD Secretary-General Report to G20 Leaders**. November, 2015. Disponível em: <<https://goo.gl/5fvyuJ>>. Acesso em: 16 fev. 2018.

OSBORNE, Jason W.; OVERBAY, Amy. **The power of outliers (and why researchers should always check for them)**. *Practical assessment, research & evaluation*, v. 9, n. 6, p. 1-12, 2004.

MANKIW, N. Gregory. **How Inherited Wealth Helps the Economy**. 2014. Disponível em: <<https://goo.gl/Fjn68h>>. Acesso em: 16 fev. 2018.

MCBARNET, Doreen et al. **When compliance is not the solution but the problem: From changes in law to changes in attitude**. Canberra: Australian National University, Centre for Tax System Integrity. 2001.

MEDEIROS, Tarcízio Dinoá. **Mensuração da economia informal e da evasão fiscal no Brasil**. Série Política Fiscal, 42. Comissão Econômica para América Latina y el Caribe, Santiago, 1993.

MIGUEL, Luciano Garcia. **A hipótese de incidência do ICMS e a evolução dos conceitos tradicionais de mercadoria e serviço de comunicação**. Disponível em: <<https://tede2.pucsp.br/handle/handle/7069>>. Acesso em: 16 fev. 2018.

MOSSIN, Jan. **Taxation and risk-taking: an expected utility approach**. *Economica*, v. 35, n. 137, p. 74-82, 1968.

MYLES, Gareth. **Economic Growth and the Role of Taxation-Theory**. OECD Economics Department Working Papers, No. 713, OECD Publishing, Paris. 2009.

MUSGRAVE, Richard A.; MUSGRAVE, Peggy B. **Public Finance in Theory and Practice**. 2004. 5 ed. Tata McGraw-Hill Edition 2004.

NEUMANN, Jv; MORGENSTERN, Oskar; KUHN, H. W. **Theory of games and economic behavior**. Princeton University Press, Commemorative Edition (1 May 2007), 1953.

NISHIOKA, Alexandre Naoki. **Planejamento fiscal e elusão tributária na constituição e gestão de sociedades: os limites da requalificação dos atos e negócios jurídicos pela administração**. 2010. Tese (Doutorado em Direito Econômico e Financeiro) – Faculdade de Direito, Universidade de São Paulo, São Paulo, 2010.

OSIUS, Gerhard; ROJEK, Dieter. **Normal goodness-of-fit tests for multinomial models with large degrees of freedom**. *Journal of the American Statistical Association*, v. 87, n. 420, p. 1145-1152, 1992.

PAUL, Prabasaj; PENNELL, Michael L.; LEMESHOW, Stanley. **Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets**. *Statistics in medicine*, v. 32, n. 1, p. 67-80, 2013.

PIKETTY, Thomas. **O capital no século XXI**. 2014. Rio de Janeiro: Intrínseca.

R Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL <https://www.R-project.org/>.

SANDMO, Agnar. **Three decades of tax evasion: A perspective on the literature**. In: Skatteforum (The Research Forum on Taxation), Rosendal, Norway. 2003.

SÃO PAULO. **Decreto 45.490, de 30 de novembro de 2000**. Regulamento do Imposto sobre Operações Relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação – RICMS. Disponível em: <<https://goo.gl/EhWuZA>>. Acesso em: 16 fev. 2018.

SÃO PAULO. **Lei 6.374, de 01 de março de 1993**. Dispõe sobre a instituição do Imposto sobre Operações Relativas à Circulação de Mercadorias e sobre Prestação de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação – ICMS. Disponível em: <<https://goo.gl/EhWuZA>>. Acesso em: 16 fev. 2018.

SÃO PAULO. **Portaria CAT 95, de 24 de novembro de 2006.** Dispõe sobre a suspensão, cassação e nulidade da eficácia da inscrição no Cadastro de Contribuintes do ICMS e dá outras providências. Disponível em: <<https://goo.gl/EhWuZA>>. Acesso em: 16 fev. 2018.

SÃO PAULO. **Portaria CAT 92, de 23 de dezembro de 1998.** Implanta e uniformiza procedimentos relativos ao sistema eletrônico de serviços dos Postos Fiscais Administrativos do Estado. Disponível em: <<https://goo.gl/2jat2x>>. Acesso em: 16 fev. 2018.

Sefaz/SP. Secretaria da Fazenda do Estado de São Paulo. **Indicadores de Efetivo de Fiscalização - Relação número de contribuintes inscritos no ICMS - número total de fiscais.** 2017. Disponível em: <<https://goo.gl/c2PhPc>>. Acesso em: 16 fev. 2018.

Sefaz/SP. Secretaria da Fazenda do Estado de São Paulo. **Estrutura Organizacional.** 2018. Disponível em: <<https://goo.gl/rzR4gX>>. Acesso em: 02 dez. 2018.

SCHNEIDER, Friedrich. **Shadow economies around the world: what do we really know?** European Journal of Political Economy, Vol. 21, N°. 3, p. 598-642, 2005.

SHEFRIN, Hersh M.; THALER, Richard H. **The behavioral life-cycle hypothesis.** Economic inquiry, v. 26, n. 4, p. 609-643, 1988.

SHELAK, Benjamin J. **The impact of the US underground economy A note relating to the impact on state finances.** Journal of Government Information, v. 24, n. 2, p. 113-117, 1997.

SIC.SP. **Sistema Integrado de Informações ao Cidadão.** 2018. Disponível em: <<https://goo.gl/ojMThp>>. Acesso em: 05 out. 2018.

SIMPSON, Sally S. **Corporate crime, law, and social control.** Cambridge University Press, 2002.

SIQUEIRA, Marcelo Lettieri; FRANCISCO S. Ramos. **A economia da sonegação: teorias e evidências empíricas.** Revista de Economia Contemporânea. Rio de Janeiro, 9(3): 555-581, set./dez. 2005.

SNOW, Arthur; WARREN, Ronald S. **Tax evasion under random audits with uncertain detection.** Economics Letters, v. 88, n. 1, p. 97-100, 2005.

STAKE, Robert E. **The art of case study research.** Thousand Oaks, CA: Sage, 1995.

STUKEL, Thérèse A. **Generalized logistic models.** Journal of the American Statistical Association, v. 83, n. 402, p. 426-431, 1988.

SWISTAK, Artur. **VAT in Federal Countries: international experience.** Seminar: ICMS and the Future of the States Guarujá, São Paulo, Brazil, Set., 2015. Disponível em: <<https://goo.gl/OEhyzi>>. Acesso em: 16 fev. 2018.

TANZI, Vito; SCHNEIDER, Friedrich, BARBOSA FILHO, Fernando de Holanda. **Economia subterrânea: uma visão contemporânea da economia informal no Brasil.** ETCO, Instituto Brasileiro de Ética Concorrencial. Rio de Janeiro: Elsevier, 2009.

TAYLOR, Natalie. **Understanding taxpayer attitudes through understanding taxpayer identities**. In V. Braithwaite (ed.), *Taxing democracy*, p. 71 – 89. Aldershot: Ashgate. 2003

THACKRAY, Mick. **Applied Tax Gap Analysis in the United Kingdom: Its use in tax administration, and future research**. January, 2012. Disponível em: <<https://goo.gl/DshrGp>>. Acesso em: 14 ago. 2017.

TORGLER, Benno. **Tax compliance and tax morale: a theoretical and empirical analysis**. Cheltenham: Edward Elgar Publishing, 2007.

TULKENS, Henry; JACQUEMIN, Alex. **The cost of delinquency: a problem of optimal allocation of private and public expenditure**. Kathol. Univ. CORE Discussion Paper 7133, 1971.

VIOL, Andrea Lemgruber. **A administração tributária moderna e a maximização do cumprimento tributário**. Revista da Receita Federal: estudos tributários e aduaneiros. Brasília-DF, v.1, n.2, p.50-82, jan./jul. 2015.

WEBER, Till; FOOKEN, Olaf; HERRMANN, Benedikt. **Behavioural economics and taxation**. Nº. 41. Directorate General Taxation and Customs Union, European Commission, 2014.

WEBLEY, Paul; COLE, Michaela; EIDJAR, Ole-Petter. **The prediction of self-reported and hypothetical tax-evasion: Evidence from England, France and Norway**. Journal of Economic Psychology, v. 22, n. 2, p. 141-155, 2001.

WORLD BANK GROUP; PWC. **Paying Taxes 2017**. Disponível em: <<https://goo.gl/Qv4ixn>>. Acesso em: 16 fev. 2018.

YIN, Robert K. **Case study research: Design and methods**. 3 ed. Thousand Oaks, CA: Sage, 2003.

YUEN, Karen K. **The two-sample trimmed t for unequal population variances**. Biometrika, v. 61, n. 1, p. 165-170, 1974.

ZHANG, Zhongheng. **Model building strategy for logistic regression: purposeful selection**. Annals of translational medicine, v. 4, n. 6, 2016a.

ZHANG, Zhongheng. **Residuals and regression diagnostics: focusing on logistic regression**. Annals of translational medicine, v. 4, n. 10, 2016b.

ZODROW, George R. **The Sales Tax, the VAT, and Taxes in Between—or, Is the Only Good NRST a "VAT in Drag"?** 1999. National Tax Journal. Vol. 52, No. 3, TAX POLICY IN A TIME OF SURPLUS (September, 1999), p. 429-442.

APÊNDICES

APÊNDICE A – COMANDOS NO R PARA CONSTRUÇÃO, ANÁLISE E VALIDAÇÃO DO MODELO

```

#Regressão Logística

library(ROCR)
library(reportROC)
library(car)

#Carrega os dados
amostra <- read.csv(file.choose(), sep=";")

#Informa ao R que as colunas são categóricas (fatores)
amostra$cadt_natj <- as.factor(amostra$cadt_natj)
amostra$cadt_port <- as.factor(amostra$cadt_port)
amostra$cadt_cnae <- as.factor(amostra$cadt_cnae)
amostra$cadt_subt <- as.factor(amostra$cadt_subt)
amostra$cadt_delg <- as.factor(amostra$cadt_delg)

attach(amostra)

#Apenas as variáveis que geraram o melhor modelo na base de Variable Selection
var <- c("docf_decl", "cont_indx", "decl_ind1", "I(decl_ind2^2)", "decl_docf",
        "docf_ind3", "soc1_indx", "soc2_indx", "qtde_fisc", "qtde_auto",
        "dilg_insc", "cadt_cnae", "cadt_delg"

        ,"docf_decl*decl_ind1", "soc1_indx*soc2_indx"
        ,"soc2_indx*cadt_delg", "qtde_auto*cadt_delg"
        ,"soc1_indx*cadt_delg", "qtde_fisc*cadt_delg",
        "qtde_fisc*qtde_auto")

var <- paste(var, collapse = "+")

#Variáveis com decl_ind2 sem transformação
varNoStd <- c("docf_decl", "cont_indx", "decl_ind1", "decl_ind2", "decl_docf",
             "docf_ind3", "soc1_indx", "soc2_indx", "qtde_fisc", "qtde_auto",
             "dilg_insc", "cadt_cnae", "cadt_delg"

             ,"docf_decl*decl_ind1", "soc1_indx*soc2_indx"
             ,"soc2_indx*cadt_delg", "qtde_auto*cadt_delg"
             ,"soc1_indx*cadt_delg", "qtde_fisc*cadt_delg",
             "qtde_fisc*qtde_auto")

varNoStd <- paste(varNoStd, collapse = "+")

#Executa a Regressão Logística na base de aprendizado com todas as variáveis escolhidas na
seleção
modelo <- glm(as.formula(paste0("as.factor(dilg_irre) ~", var)), data = amostra, family =
binomial)
pred <- predict(modelo, amostra, type="response")

#####
#2.2.3.3 Resultados da seleção de variáveis
#####

#Influência das variáveis na predição.

#Diminuição Desvio Residual
(a <- anova(modelo, test = "Chisq"))

#Significância#
#Modelo sem função quadrática
modelos <- glm(as.formula(paste0("as.factor(dilg_irre) ~", varNoStd)), data = amostra, family
= binomial)
preds <- predict(modelos, amostra, type="response")
(wald <- Anova(modelos, type="II", test="Wald"))

variaveis = rownames(a[2:14,])

```

```

peso = (a$Deviance[2:14] + wald$Chisq[1:13])/2
total <- sum(peso)
peso <- peso/total
peso <- data.frame(variaveis, peso)

peso <- peso[order(-peso$peso),]
peso <- data.frame(peso)

bb <- barplot(peso$peso, names.arg = peso$variaveis, xlab='Variáveis', ylab = 'Influência na
predição (%)')
text(bb, 0, round(peso$peso,3), pos = 3)

#####
#Direção da influência das variáveis na predição.
#####

sm <- data.frame((summary(modelos)$coefficients)
variaveis <- rownames(sm[2:12,])
direcao <- sm$Estimate[2:12]

relacao <- data.frame(variaveis, direcao)
bb <- barplot(relacao$direcao, names.arg = relacao$variaveis, xlab='Variáveis', ylab =
'Coeficiente de relação com o evento de interesse')

#Modelo com decl_ind2 transformada
modelo <- glm(as.formula(paste0("as.factor(dilg_irre) ~", var)), data = amostra, family =
binomial)
pred <- predict(modelo, amostra, type="response")

#####
#2.2.4.1 Os pressupostos do modelo Logit
#####

#####
#2.2.4.1.1 Ausência de Multicolinearidade
#####

#Apenas as variáveis sem as interações

mvar <- c("docf_decl", "cont_indx", "decl_ind1", "decl_ind2", "decl_docf",
"docf_ind3", "soc1_indx", "soc2_indx", "qtde_fisc", "qtde_auto",
"dilg_insc", "cadt_cnae", "cadt_delg")

mvar <- paste(mvar, collapse = "+")

mmodelo <- glm(as.formula(paste0("dilg_irre ~", mvar)), data = amostra, family = binomial)
car::vif(mmodelo)

#####
#2.2.4.1.2 Ausência de Influential Values
#####

library(dplyr)
library(broom)

l = lm.influence(modelo)$hat
Rp = residuals.glm(modelo, "pearson")
D = Rp^2 * lm.influence(modelo)$hat / (5 * ((1-lm.influence(modelo)$hat)^2)) #Cook's distance.

#Gráficos Leverage Stand. Pearson Residuals e Cooks'Distance individualmente
par(mfrow = c(1, 3))
plot(l ~ pred, xlab = "Estimated Logistic Probability", ylab = "Leverage", cex.lab = 1.4) #,
ylim=c(0,10))
plot(Rp ~ pred, xlab = "Estimated Logistic Probability", ylab = "Standardised Perason
Residuals", cex.lab = 1.4) #, ylim=c(0,10))
plot(D ~ pred, xlab = "Estimated Logistic Probability", ylab = "Cook's Distance", cex.lab =
1.4) #, ylim=c(0,3))

#Gráficos Leverage Stand. Pearson Residuals e Cooks'Distance individualmente juntos
par(mfrow = c(1, 1))
plot(modelo,5)

```

```

#Filtrar Influential Values potenciais
outliers <- augment(modelo) %>% mutate(index = 1:n())
(out <- outliers %>% filter(abs(.std.resid) > 3.5 | abs(.cooks) > 0.2))

outGeral <- out$index

#Análise individual dos Influential Values potenciais
View(outGeral)

#Remove Influential Values da amostra
amostra <- amostra[-outGeral,]

#Refaz o modelo sem os Influential Values
modelo <- glm(as.formula(paste0("as.factor(dilg_irre) ~", var)), data = amostra, family =
binomial)
pred <- predict(modelo, amostra, type="response")

#Refaz os gráficos

l = lm.influence(modelo)$hat
Rp = residuals.glm(modelo, "pearson")
D = Rp^2 * lm.influence(modelo)$hat/(5 * ((1-lm.influence(modelo)$hat)^2)) #Cook's distance.

#Gráficos Leverage Stand. Pearson Residuals e Cooks'Distance individualmente
par(mfrow = c(1, 3))
plot(l ~ pred, xlab = "Estimated Logistic Probability", ylab = "Leverage", cex.lab = 1.4) #,
ylim=c(0,10))
plot(Rp ~ pred, xlab = "Estimated Logistic Probability", ylab = "Standardised Perason
Residuals", cex.lab = 1.4) #, ylim=c(0,10))
plot(D ~ pred, xlab = "Estimated Logistic Probability", ylab = "Cook's Distance", cex.lab =
1.4) #, ylim=c(0,3))

#Gráficos Leverage Stand. Pearson Residuals e Cooks'Distance individualmente juntos
par(mfrow = c(1, 1))
plot(modelo,5)

#Teste de Bonferonni
outlierTest(modelo)

#####
#2.2.4.1.3 Relação linear entre as variáveis explicativas contínuas e a variável dependente
#####

#Modelo sem função quadrática
modelo <- glm(as.formula(paste0("as.factor(dilg_irre) ~", varNoStd)), data = amostra, family =
binomial)
pred <- predict(modelo, amostra, type="response")

#Plota o modelo sem ajustes quadráticos para verificar linearidade
residualPlots(modelo, layout = c(4, 3), terms = ~ . -cadt_cnae -cadt_delg)

#Modelo com decl_ind2 transformada
modelo <- glm(as.formula(paste0("as.factor(dilg_irre) ~", var)), data = amostra, family =
binomial)
pred <- predict(modelo, amostra, type="response")

#Plota o modelo sem ajustes quadráticos para verificar linearidade
residualPlots(modelo, layout = c(4, 3), terms = ~ . -cadt_cnae -cadt_delg)

#####
#2.2.4.2 Qualidade do ajuste do modelo
#####

#####
#2.2.4.2.1 Medidas de ajustes individuais
#####

deltaBeta <- lm.influence(modelo)$hat #hat values
Rp = residuals.glm(modelo, "pearson")
Rsp = residuals.glm(modelo, "pearson")/sqrt(1 - lm.influence(modelo)$hat)#studentized Pearson
residuals
dev = residuals.glm(modelo, "deviance")

```

```

deltaChisqstat = Rsp^2 #Delta chi-square statistic
deltadev = lm.influence(modelo)$hat * (Rsp^2) + dev^2 #Delta deviance statistic
D = Rsp^2 * lm.influence(modelo)$hat / (5 * ((1-lm.influence(modelo)$hat)^2)) #Cook's distance.
l = lm.influence(modelo)$hat

par(mfrow = c(2, 3))
plot(deltaChisqstat ~ pred, xlab = "Estimated Logistic Probability", ylab = "Delta Chi
Square", cex.lab = 1.4) #, ylim=c(0,10))
plot(deltadev ~ pred, xlab = "Estimated Logistic Probability", ylab = "Delta Deviance",
cex.lab = 1.4)
plot(deltaBeta ~ pred, xlab = "Estimated Logistic Probability", ylab = "Standardized Delta
Beta", cex.lab = 1.4) #, ylim=c(0,10))

deltaChisqstatSuma <- summary(as.factor(cut(deltaChisqstat, breaks=c(0, 4, 10, 20, 50,
max(deltaChisqstat))))))
barplot(deltaChisqstatSuma,xlab = "Delta Chi-Square", ylab = "Frequency", cex.lab = 1.4)

deltadevSuma <- summary(as.factor(cut(deltadev, breaks=c(0, 4, 5, 6, 7, max(deltadev))))))
barplot(deltadevSuma, xlab = "Delta Deviance", ylab = "Frequency", cex.lab = 1.4)

deltaBetaSuma <- summary(as.factor(cut(deltaBeta, breaks=c(0, 0.15, 0.3, max(deltaBeta))))))
barplot(deltaBetaSuma, xlab = "Standardised Delta Beta", ylab = "Frequency", cex.lab = 1.4)

#####
#2.2.4.2.2 Testes de ajuste geral do modelo
#####

#####
#Deviance e Pearson Chi-square Statistic
#####

library(ResourceSelection)
library(LogisticDx)

measures <- gof(modelo, g = 157)

#PrI test of the Pearsons residuals, calculated by individual
#drI test of the deviance residuals, calculated by individual
#PrG test of the Pearsons residuals, calculated by covariate group
#drG test of the deviance residuals, calculated by covariate group
#PrCT test of the Pearsons residuals, calculated from the contingency table
#drCT test of the deviance residuals, calculated from the contingency table

measures$chiSq

#####
#Hosmer-Lemeshow Test
#####

n <- nrow(amostra)
m <- nrow(amostra[which(amostra$dilg_irre==1), ])

termo1 <- m/2
termo2 <- (n-m)/2
termo3 <- (n/1000)^2*8+2

(g <- round((max(c(10, min(c(termo1, termo2, termo3))))),0))
(h <- hoslem.test(amostra$dilg_irre, pred))
(h <- hoslem.test(amostra$dilg_irre, pred, g=g))
(h <- hoslem.test(amostra$dilg_irre, pred, g=g+1))

#####
#Osius and Rojek normal approximation to the distribution of the Pearson chi-square statistic
#####

measures$gof[3]

#####
#Stukel's test
#####

measures$gof[4] #Score test: parâmetro 1
measures$gof[5] #Score test: parâmetro 2
measures$gof[6] #Score test: parâmetros 1 e 2 juntos

```

```

measures$gof[7]      #Log-likelihood test: parâmetro 1
measures$gof[8]      #Log-likelihood test: parâmetro 2
measures$gof[9]      #Log-likelihood test: parâmetros 1 e 2 juntos

#####
#Kolmogorov-Smirnov test#
#####

ks.test(pred, amostra$dilg_irre, alternative='two.sided')

#####
#2.2.5 Análise dos resultados do modelo
#####

#####
#2.2.5.1 Accuracy#
#####

library(tidyr)
library(ggplot2)
library(caret)
library(e1071)

#Prediz a classificação
pred <- predict(modelo, amostra, type="response")

#Busca o valor do cut off
cutoff <- data.frame(matrix(numeric(0), ncol = 3, nrow = 16))
colnames(cutoff) <- c("Cutoff", "Sensitividade", "Especificidade")

inc <- 0.05
for(i in 1:16)
{
  cutoff[i,]$Cutoff <- inc
  p <- ifelse(pred > inc, 1, 0)
  cm <- table(Predicted = p, Actual = amostra$dilg_irre)
  cutoff[i,]$Sensitividade <- cm[1,1]/(cm[1,1]+cm[1,2])
  cutoff[i,]$Especificidade <- cm[2,2]/(cm[2,1]+cm[2,2])
  inc <- inc + 0.05
}

par(mfrow = c(1, 1))

(cutoffplot <- cutoff %>% gather(Conceito, Valor, -Cutoff))
p <- ggplot(cutoffplot, mapping = aes(x = Cutoff, y = Valor, color = Conceito)) + geom_line()
+
  theme_bw() + theme_light() +
  #geom_point(size = 5, shape = 20)
  geom_point(aes(shape=Conceito, color=Conceito), size = 3) +
  geom_line(size = 1)

p + labs(x = "Cut-off")

#Confusion Matrix com cut-off = 0.5

modelot <- train(as.formula(paste0("as.factor(dilg_irre) ~", var)), data=amostra,
method="glm", family="binomial")
predt <- predict(modelot, newdata = amostra)
confusionMatrix(data=predt, as.factor(amostra$dilg_irre))

#Visualização gráfica da Classification Table.
predj <- predict(modelot, newdata = amostra, type="prob")
plot(predj$"1", jitter(as.numeric(amostra$dilg_irre), 0.5), cex=0.5, ylab = "Observado", xlab =
"Predito")

#####
#2.2.5.2 Repeated K-fold Cross Validation
#####

ctrl <- trainControl(method = "repeatedcv", number = 5, repeats=10, savePredictions = TRUE)

```

```
mod_fit <- train(as.formula(paste0("as.factor(dilg_irre) ~", var)), data=amostra,
method="glm", family="binomial",
               trControl = ctrl, tuneLength = 5)
print (mod_fit)

#####
#2.2.5.3 Area under the ROC Curve
#####

library(caTools)
library(Deducer)
rocplot(modelo)

#####
#2.2.5.4 Pseudo R2
#####

library(DescTools)
PseudoR2(modelo, c("McFadden", "McFaddenAdj", "CoxSnell", "Nagelkerke", "Tjur"))
```