Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de "outliers" para o controle do risco de evasão tributária do ICMS

Sérgio Augusto Pará Bittencourt Neto

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Ricardo Matos Chaim

Brasília
2018

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

BB624a

Bittencourt Neto, Sérgio Augusto Pará
Análise de "outliers" para o controle do risco de evasão tributária do ICMS / Sérgio Augusto Pará Bittencourt Neto; orientador Ricardo Matos Chaim. -- Brasília, 2018.
129 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2018.

1. Evasão Fiscal no ICMS. 2. Análise de Outliers. 3. Análise Envoltória de Dados Fiscais. 4. Análise de Séries Temporais Fiscais. 5. Mineração de Dados Fiscais. I. Matos Chaim, Ricardo, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de "outliers" para o controle do risco de evasão tributária do ICMS

Sérgio Augusto Pará Bittencourt Neto

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Ricardo Matos Chaim (Orientador)
CIC/UnB

Prof. Dr. Othon de Azevedo Lopes
DIR/UNB

Prof. Dr. Rosalvo Ermes Streit
Universidade Católica de Brasília

Prof.a Dr.a Aletéia Patrícia Favacho de Araújo
Coordenadora do Programa de Pós-graduação em Computação Aplicada

Brasília, 3 de julho de 2018

Dedicatória

Dedico este trabalho à minha esposa Alda Cristina e ao meu filho Sérgio como um incentivo à continuidade dos estudos por toda a vida, como me foi ensinado pelos eternos professores Randolpho e Freida Bittencourt.

Agradecimentos

Agradeço a todos os professores do Programa de Pós-Graduação em Computação Aplicada da Universidade de Brasília (PPCA-UnB) por toda a dedicação e esforços na realização do Mestrado Profissional.

Igualmente, agradeço aos colegas auditores da Gerência de Programação Fiscal e Controle de Operações (GEPRO) pelo incentivo e paciência essenciais para o desenvolvimento deste trabalho.

Resumo

Esta dissertação apresenta a aplicação associada de selecionados modelos estatísticos e de métodos de mineração de dados para a análise de *outliers* sobre as informações da Notas Fiscais Eletrônicas e do Livro Fiscal Eletrônico, proporcionando a investigação de novas modalidades de evasão fiscal no ICMS. São combinados:

1. o método de programação matemática da Análise Envoltória de Dados (DEA) para diferenciar as empresas com desempenho relativo de arrecadação ineficiente, dentro de um segmento econômico, e eleger os contribuintes suspeitos para investigação;
2. modelos de análise de séries temporais para avaliação dos dados fiscais atinentes à apuração do imposto (comparação gráfica dos valores reais e respectivas escriturações, gráficos *boxplots*, decomposição das componentes de tendência e sazonalidade e o modelo de alisamento exponencial Holtz-Winter), com o objetivo de detectar períodos de tempo anômalos (*outliers*); e
3. outras técnicas estatísticas descritivas (gráficos analíticos da distribuição de frequência), probabilísticas (Desigualdade de Chebyshev e Lei de Newcomb Benford) e o método de mineração por clusterização *K-Means* sobre as informações fiscais dos contribuintes selecionados, para identificar os registros escriturais e os documentos fiscais sob suspeição.

É proposto um recurso computacional construído em linguagem R (plataforma R Studio) para: extrair do banco de dados (ORACLE) da Receita do Distrito Federal, processar as informações aplicando-lhes os modelos e métodos designados, e em conclusão, disponibilizar os resultados em painéis analíticos que facilitam e otimizam o trabalho de auditoria. Assim, a identificação das circunstâncias anômalas, a partir de um tratamento sistemático dos dados, proporciona maior eficiência à atividade de programação fiscal de auditorias tributárias.

Palavras-chave: Evasão Fiscal no ICMS, Análise de *Outliers*, Análise Envoltória de Dados Fiscais, Análise de Séries Temporais Fiscais, Mineração de Dados Fiscais.

Abstract

This dissertation presents the associated application of selected statistical models and data mining methods for the analysis of outliers on the information of the Electronic Fiscal Notes and the Electronic Fiscal Book, providing the investigation of new types of tax evasions in ICMS.

The following methods are applied:

1. the mathematical programming method of Data Envelopment Analysis (DEA) to differentiate companies with inefficient performance relative in the tax collection within an economic segment and to choose suspected taxpayers for research;
2. the analysis of time series used in the evaluation of fiscal data related to the calculation of the ICMS tax (graphical comparison of actual values and respective deeds, boxplot graphs, the decomposition of trend and seasonality components and the Holtz-Winter method), capable of anomalous time periods (outliers) detection; and
3. descriptive statistical analysis (frequency distribution), probabilistic analysis (Chebyshev Inequality and Newcomb Benford Law) and K-Means clustering techniques on selected taxpayers' tax information to identify book entries and tax documents under suspicion.

A computational code in R language (R Studio platform) is developed for: extraction of data from the Federal District Revenue database (ORACLE), processing of the extracted information while applying the designated models and methods and generating the results in panels that facilitate and optimize audit work. Thus, in conclusion, the identification of the anomalous circumstances, based on a systematic treatment of the data, provides greater efficiency to the fiscal programming activity of tax audits.

Keywords: Tax Evasion, Analysis of Outliers, Data Envelopment Analisys in Tax Data, Analysis of Tax Time Series, Tax Data Mining.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BCC Modelo DEA - estabelecido por Banker, Charnes e Cooper.

CCR Modelo DEA - estabelecido por Charnes, Cooper e Rhodes.

CNAE Classificação Nacional de Atividades Econômicas.

COFIT Coordenação de Fiscalização Tributária.

CRS Retornos Constantes de Escala.

CTN Código Tributário Nacional.

DEA Análise Envoltória de Dados.

DF Distrito Federal - Brasil.

DMU Unidades de Tomada de Decisão.

GEPRO Gerência de Programação Fiscal e Controle de Operações.

ICM Imposto sobre Operações Relativas à Circulação de Mercadorias Realizadas por Comerciantes, Industriais e Produtores.

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.

IVC Imposto sobre Vendas e Consignações.

LFE Livro Fiscal Eletrônico.

MG Minas Gerais.

NFCE Nota Fiscal ao Consumidor Eletrônica.

NFE-E Nota Fiscal Eletrônica de Entrada.

NFE-S Nota Fiscal Eletrônica de Saída.

NFEs Notas Fiscais Eletrônicas.

PR Paraná.

RJ Rio de Janeiro.

RS Rio Grande do Sul.

SC Santa Catarina.

SEFP Secretaria de Estado de Fazenda do Distrito Federal.

SP São Paulo.

SQL Structured Query Language.

SUREC Subsecretaria da Receita do Distrito Federal.

VRS Retornos Variáveis de Escala.

Capítulo 1

Introdução

A presente pesquisa tem por proposta a construção de uma metodologia de programação das auditorias fiscais que associa a aplicação de modelos estatísticos e métodos de mineração de dados para a análise de *outliers*, objetivando a descoberta de novas hipóteses de ilícitos tributários e a consequente diminuição do risco de sucesso das condutas de evasão do Imposto sobre Operações Relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação (ICMS) no Distrito Federal - Brasil (DF).

Este capítulo dedica-se: à contextualização do tema e à demonstração da sua relevância institucional para o órgão da Receita Distrital; à definição do problema objeto da dissertação; à disposição dos objetivos almejados pela pesquisa; à apresentação de uma breve síntese da metodologia empreendida e à exposição da organização do trabalho.

1.1 Contexto Institucional e Relevância do Tema

A Gerência de Programação Fiscal e Controle de Operações (GEPRO), órgão da Coordenação de Fiscalização Tributária (COFIT) da Subsecretaria da Receita do Distrito Federal (SUREC), no contexto do Regimento Interno da Secretaria de Estado de Fazenda do Distrito Federal (SEFP), tem por missão organizacional elaborar os projetos e as ações de fiscalização tributária com o objetivo de recuperar os valores das obrigações tributárias subtraídos indevidamente pelos contribuintes (evasão fiscal).

Por outras palavras, cumpre à atividade de programação fiscal determinar quais as empresas sofrerão ação de auditoria (ou outro procedimento de fiscalização) com fundamento nos indícios de práticas evasivas (práticas de fraude ou sonegação) identificadas e apuradas pela análise sistemática das informações de interesse tributário disponíveis.

Assim, a essência do trabalho da programação fiscal é a identificação das evidências de comportamento fiscal anômalo das empresas contribuintes, a sua confirmação como

hipótese de supressão indevida e o desenvolvimento de projetos de auditoria para a recuperação do crédito tributário evadido.

Sem embargos, faz-se indispensável obedecer aos desígnios insculpidos na Portaria 130/97 - SEFP [1], que exigem a adoção de métodos científicos (estatísticos e de análise de dados) na condução dos trabalhos de programação fiscal das auditorias tributárias, *ad litteram*:

> **"PORTARIA Nº 130, DE 3 DE MARÇO DE 1997**
> **Estabelece regras e critérios para a Programação Fiscal da Auditoria Tributária da Subsecretaria da Receita da Secretaria de Fazenda e Planejamento do Distrito Federal.**
> **Art. 1º - A seleção de contribuintes a serem fiscalizados, por meio da Programação Fiscal da Auditoria Tributária da Subsecretaria da Receita da Secretaria de Fazenda e Planejamento do Distrito Federal, será constituída pela seleção científica, diligências e denúncias.**
> **Art. 2º - A seleção científica fundamentar-se-á nos princípios gerais da administração tributária e nos princípios estatísticos e empíricos de tratamento de informações e dados e obedecerá aos seguintes critérios, isolada ou conjuntamente:**
> **I - setores de atividades e dimensões econômico-financeiros dos contribuintes;**
> **II - períodos e tributos a serem fiscalizados;**
> **III - amplitude e profundidade da ação fiscal;**
> **IV - localização geográfica;**
> **V - falta de recolhimento de tributos;**
> **VI - recolhimento de tributos incompatíveis com o movimento econômico-financeiro do contribuinte;**
> **VII - falta de cumprimento de obrigação acessória;**
> **VIII - evidências de prática de evasão ou sonegação fiscal." (Grifamos)**

Em síntese, a GEPRO tem por mister a persecução (recuperação) dos créditos tributários diminuídos, ou suprimidos, indevidamente e o combate às fraudes e à sonegação fiscal por meio do desenvolvimento de projetos de auditoria tributária e de outras ações de fiscalização.

Com espeque nesse objetivo institucional, a GEPRO tem por aspiração o perficiente aproveitamento dos recursos (especialmente os recursos humanos) dedicados à atividade de auditoria, o que requesta a necessária otimização da sua utilidade produtiva (evitando o desperdío, especialmente de tempo) e exige a maior acertividade das suspeitas informadas nos papéis de trabalho disponibilizados para os auditores.

Para alcançar esse fim, essencial se faz recorrer às técnicas da Ciência de Dados (*data science*) que contribuam para o aumento da exatidão na identificação das empresas suspeitas e dos indícios a serem estudados nos exames das auditorias programadas.

Perseguindo esse propósito, este estudo deseja oferecer uma metodologia útil para o reconhecimento de anomalias indicativas de irregularidades na atividade fiscal dos contri-

buintes, o que deverá contribuir para o maior sucesso das auditorias, bem assim o melhor aproveitamento dos recursos da fiscalização tributária.

1.2 Definição do Problema

O modelo atual de programação fiscal no Distrito Federal - Brasil (DF), em grande medida, é dedicado à realização de auditorias concentradas sob reconhecidas espécies de irregularidades fiscais. Ou seja, parcela significativa das atividades de fiscalização hoje empreendidas debruça-se sobre o combate às modalidades de evasão tributária já identificadas pelo órgão distrital da Receita.

São exemplos de projetos devotados às variedades conhecidas de evasão fiscal:

a) Projetos dedicados à autuação das diferenças apuradas:

   - entre o faturamento tributável declarado e o movimento comercial obtido com as vendas realizadas com o pagamento mediado por cartão de crédito/débito;
   - entre os valores dos documentos fiscais declarados e os valores verdadeiramente destacados (emitidos);
   - pelo uso de alíquota divergentes e incompatíveis com a realidade da operação.

b) Projetos devotados à glosa de créditos registados sem a devida conciliação com os documentos fiscais destinados ao contribuinte.

c) Projetos para o cancelamento dos créditos de origem incompatível com o aproveitamento permitido (*v.g.*substituição tributária, simples nacional, empresa cancelada, uso e consumo, etc.).

Nesses projetos, por meio da aplicação de algoritmos dedicados a cada tipo conhecido de evasão do ICMS, são eleitos os contribuintes que devem ser objetos de uma ação fiscal específica sobre a espécie de conduta fiscal imprópria (evasão fiscal) praticada.

Consigne-se que, presentes na conduta fiscal dos contribuintes as hipóteses evasivas identificadas, a atual estratégia de programação fiscal otimiza sobremaneira o sucesso da atividade de auditoria, porquanto dá precisão (direcionamento) aos esforços que ali serão empreendidos.

Apesar dos resultados excepcionais que o atual modelo de programação fiscal supervisionada proporciona (modelo assentado no conhecimento prévio da conduta examinada), merece ser considerado que ele é deficiente quando se almeja a identificação de expedientes inéditos de desvios tributários (como *v.g.* novas modalidades fraudes e de sonegação), ou

seja, ele é imperfeito para o reconhecimento das novas experiências de evasão fiscal desconhecidas (ou ainda, não tratadas), em especial daquelas desenvolvidas como resposta alternativa à censura (autuações fiscais) já perpetrada.

Para o propósito de descoberta de novas formas de evasão tributária, faz-se necessária uma metodologia alternativa que extrapole o modelo supervisionado de escolha de contribuintes (modelos dedicados às práticas ilícitas já conhecidas), ou melhor, uma metodologia que supere a persecução das condutas esperadas e proponha novos indícios e padrões de suspeição e possibilite a descoberta de hipóteses de evasão fiscal subjacentes ainda inexploradas.

Objetivando satisfazer essa necessidade de inovar a programação de ações fiscais para além do prestigiado cenário de combate às já exploradas hipóteses de ilicitudes tributária, este estudo ambiciona apresentar uma sugestão diferenciada de escolha de empresas por seus dados extravagantes e suas condutas anômalas suspeitas porquanto infundadas, duvidosas e sem suporte na legislação tributária.

Isto posto, o problema a ser tratado nessa dissertação será o de oferecer uma proposta de metodologia para vencer a carência do atual modelo de programação fiscal, que possibilite a seleção de contribuintes para a fiscalização em razão do seu comportamento excêntrico (anômalo) e de um padrão de conduta tributária temerário.

1.3 Justificativa do Tema

Considerando a limitação que o vigente modelo de programação de auditorias apresenta para a identificação de novos expedientes de evasão fiscal (conforme exposto anteriormente na seção 1.2), esse estudo apresenta uma proposta de metodologia para superar esse problema, construindo um paradigma complementar (e não concorrente) de programação fiscal.

Em atenção à necessidade: a) de correção da incompletude do atual modelo de seleção de contribuintes para a fiscalização, baseado preponderantemente na persecução das modalidades ilícitas já conhecidas, b) assim como, de otimização do uso dos recursos disponíveis, pretende-se a construção de uma metodologia adicional de programação fiscal que além de possibilitar a descoberta de novas formas de empreender a evasão tributária, seja eficiente na redução do tempo e do trabalho dedicados nas atividades de auditorias.

Nesse sentido, a proposta deste trabalho é conciliar a aplicação combinada de modelos estatísticos e métodos de mineração e análise de exploração de *outliers*, que:

a) A partir de um conjunto de empresas comparáveis, permita escolher os contribuintes controversos que sustentem um resultado fiscal de arrecadação comparativamente inferior e ineficiente.

b) Selecione os períodos de tempo onde o conjunto dos documentos emitidos e escriturados se mostrem incomuns (ou inconciliáveis) e, assim, suspeitos.

c) Detecte e aponte quais os registro contábeis e os documentos fiscais que devem ser preferencialmente avaliados em exame de auditoria.

Preconiza-se a aplicação associada (contudo em etapas independentes) dos seguintes modelos e métodos de detecção de *outliers*:

a) Modelo de programação matemática Análise Envoltória de Dados (DEA) dedicado à medição das eficiências contributivas relativas das empresas, usado para distinguir os contribuintes com baixo resultado arrecadatório comparado.

b) Modelo de análise de séries temporais fiscais apropriado à decomposição analítica de suas componentes de tendência e sazonalidade e à projeção comparativa com um padrão de predição, para a identificação dos meses/anos anômalos.

c) Análise descritiva, probabilística e o método de mineração por clusterização *K-Means* para separar e dar prioridade à documentação fiscal e aos dados de escrituração contábil sob maior suspeição.

A escolha dos sobreditos modelos e métodos acontece pela sua relativa simplicidade e alta confiança na assertividade de seus resultados.

Assim, a orientação pela evolução da atividade de programação fiscal justifica a proposição da metodologia em tema, porquanto objetiva o melhor desempenho na percepção de inovadoras modalidades de evasão fiscal, qualificando-se como uma resposta para a escolha ótima dos contribuintes, períodos, circunstâncias e documentos elegíveis aos procedimentos de auditoria.

Sendo o ICMS um imposto de grande relevância financeira para o DF, assim como um tributo de alta complexidade de operacionalização, porquanto é não-cumulativo (incide sobre o valor agregado em cada operação), explica-se a escolha da aplicação da metodologia proposta sobre esse imposto em todas as etapas do trabalho. Nada obstante, registre-se que ela também poderá ser facilmente adaptada às características de outros tributos.

1.4 Objetivos da Pesquisa

O objetivo geral desse trabalho é desenvolver uma metodologia que associe selecionados modelos estatísticos e métodos de mineração e análise de dados *outliers* com o propósito de desvendar novas modalidades de evasão fiscal, proporcionando a evolução da atividade

de programação fiscal de auditorias tributárias para além do modelo atual de seleção de contribuintes fundado em hipóteses desveladas de ilícitos tributários.

Outrossim, esta dissertação tem por objetivos específicos associar e aplicar modelos e métodos que:

a) Selecionem empresas contribuintes com comportamentos fiscal duvidoso, especialmente em razão de sua baixa arrecadação em relação à expectativa de seu segmento econômico.

b) Avaliem os valores temporais das informações dos contribuintes eleitos e identifiquem os períodos (mês e ano) onde o os dados fiscais se mostrem incomuns.

c) Indiquem os documentos tributários e os valores escriturados anômalos presentes nas operações comerciais e na contabilidade fiscal das empresas investigadas.

Destarte, ambiciona-se a proposição de uma metodologia que reconheça as empresas contribuintes do ICMS que apresentem um comportamento fiscal questionável e digno de investigação, os períodos de tempo que devem ser examinados em auditoria e os valores e documentos discrepantes suspeitos de conter novas modalidades de evasão tributária.

1.5 Limitações do Estudo

Uma vez que a metodologia proposta é desenvolvida sobre as informações disponíveis para o órgão da Receita do DF, melhor dizendo informações de faturamento empresarial, dados dispostos nos documentos fiscais eletrônicos relativos às operações realizadas pelos contribuintes, bem assim os dados da escrituração contábil/fiscal de apuração do imposto pelas empresas, as conclusões sobre as condutas anômalas suspeitas não consideram as atividades praticada à margem desse ambiente de informações.

Dessa forma, pela metodologia exposta não é possível detectar as práticas de evasão (sonegação e fraudes) onde as operações comerciais de entradas e, igualmente, as de saídas aconteçam todas em espécie (dinheiro) e sem a cobertura de documentação fiscal regular.

A solução para a descoberta dessas atividades clandestinas não será objeto desse estudo.

1.6 Organização da Dissertação

Esta dissertação possui 3 (três) etapas de desenvolvimento, coincidentes com as fases de realização da metodologia em proposta, a conhecer:

- **Etapa I** - Aplicação do modelo de programação matemática de **Análise Envoltória de Dados** (DEA) - Etapa devotada à identificação e à seleção de um (ou mais) contribuinte(s) pertencente(s) a um determinado segmento econômico em razão de sua ineficiência arrecadatória relativa. Melhor dizendo, a fraca e incompatível arrecadação considerando a expectativa para seu setor econômico.

  Como regra, nessa etapa exige-se a comparação de empresas que compartilhem coincidentes características de tributação (homogeneidade), para que se evite o confronto de sujeitos passivos subjugados a diferentes incidências tributárias ou perfis de tratamento fiscal de cotejo incompossível (*v.g.* Simples Nacional *vs* Regime Normal). Nesse intento é interessante que a coleção de empresas para avaliação repercuta simultânea participação em um mesmo segmento da economia.

  Essa etapa é satisfeita com a seleção de empresas que não correspondam às expectativas de uma conduta arrecadatória regular, de acordo com parâmetros e requisitos definidos, aplicáveis ao conjunto setorial. Melhor dizendo, deverão ser reconhecidos os contribuintes que oferecem um resultado relativo atípico (discordante) digno de um aprofundamento investigativo.

  Após a avaliação comparativa do desempenho desses contribuintes coletivamente, elegem-se aqueles merecedores de uma apreciação mais contundente de seus dados fiscais procurando a confirmação dos indícios da eventual evasão fiscal praticada.

- **Etapa II - Análise das Séries Temporais Fiscais** - Nessa etapa realiza-se, sobre os contribuintes eleitos, a análise de seus dados fiscais distribuídos no tempo em busca dos momentos temporais *outliers*.

  Faz-se nesse estágio o exame do comportamento das informações tributárias contempladas nos documentos fiscais emitidos e nos lançamentos contábeis escriturados, todos disseminados ao longo do tempo. Almeja-se distinguir os períodos temporais anômalos que se apresentem discrepantes em relação aos padrões observados nos dados fiscais das empresas investigadas.

  Os resultados obtidos nessa etapa direcionam a atenção da auditoria para os meses e anos com discrepâncias relevantes, porquanto tais períodos apresentam valores e/ou resultados distorcidos em relação à regularidade esperada.

- **Etapa III** - Aplicação da **Análise Estatística** de valores *outliers* (métodos paramétricos e não paramétricos) e da mineração dos dados *outliers* por **Clusterização *K-Means*** sobre a documentação e a escrituração contábil-fiscal de cada contribuinte em pesquisa dos valores anormais ali presentes.

Nessa etapa consolida-se a identificação individualizada dos documentos emitidos (notas fiscais) e dos valores escriturais (registos lançados na contabilidade fiscal) pertinentes às operações e aos lançamentos contábeis que se anunciam como possíveis indícios de evasão fiscal perpetradas.

As etapas serão oferecidas separadamente em capítulos próprios que tratarão da apresentação dos métodos e modelos aplicados, seus contextos na literatura científica, os dados necessários e os resultados exemplificativos da sua aplicação.

Consoante o exposto graficamente no diagrama em seguimento (Figura 1.1), a metodologia a ser apresentada pressupõe o concurso das 3 etapas não concorrentes, onde serão realizadas:

a) A seleção (pelo método DEA) de empresas contribuintes suspeitas em razão dos seus resultados arrecadatórios ineficientes.

b) A análise do comportamentos tributários dessas empresas no tempo (séries temporais) em busca dos períodos que se configurem anormais e duvidosos.

c) A Mineração e a análise estatística dos dados e informações fiscais discrepantes para identificar pontualmente as operações e escriturações contábeis controversas.

Não obstante apresentarem um formato sucessivo e estruturado, as etapas da metodologia são de aplicação independente e serão desenvolvidas individualmente nesse estudo, condição que facilitará o seu entendimento.

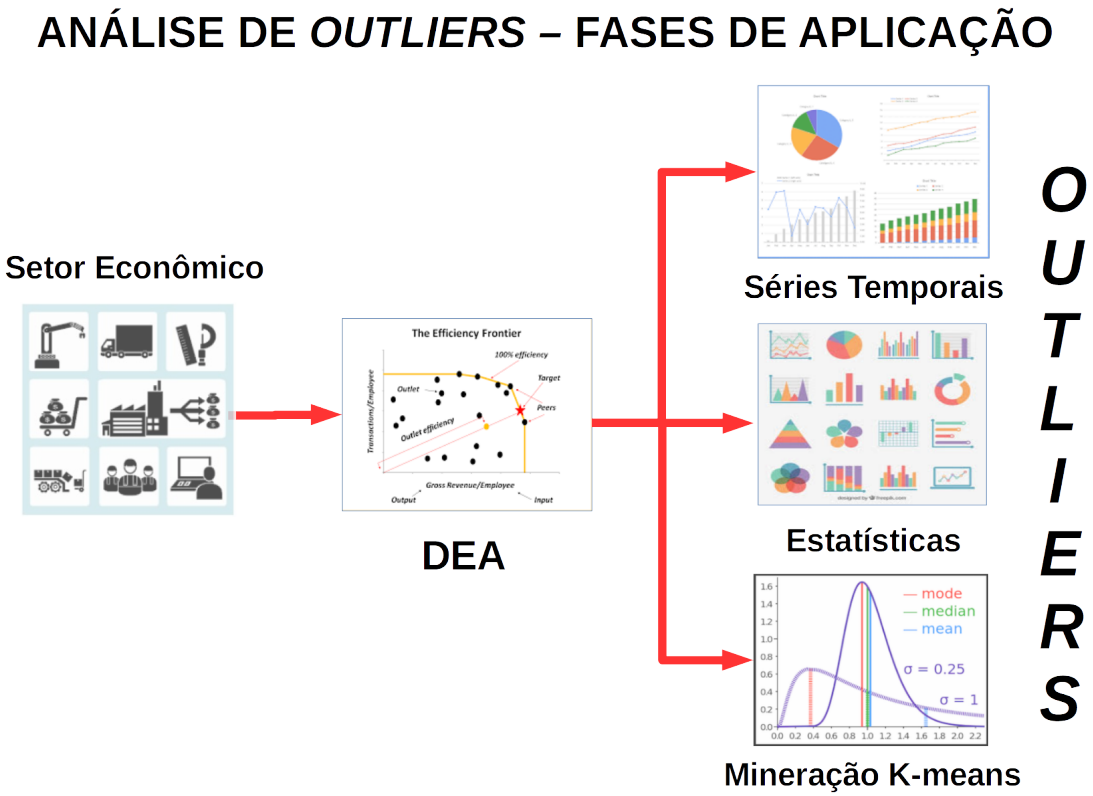

Figura 1.1: Fases de Aplicação da Metodologia de Análise de *Outliers*.

Capítulo 2

ICMS - Fraude, Sonegação e os Crimes Contra a Ordem Tributária

Este capítulo oferece um breve histórico descritivo da origem do ICMS, destacando suas principais particularidades, essenciais à percepção e ao entendimento de suas hipóteses de evasão.

O capítulo discute, entrementes, o tratamento jurídico dos tipos penais e administrativos que definem os desvios de conduta censuráveis qualificados como fraude, sonegação e crime contra a ordem tributária, expondo exemplificadamente algumas práticas usuais de tais irregularidades no ICMS.

2.1 Origem e Características do Imposto

Criado pelo art. 155, inciso II, da Constituição Federal [2], o Imposto sobre Operações Relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação (ICMS) é um imposto de competência estadual que representa a tributação brasileira sobre o valor agregado nas operações comerciais de compra e venda de mercadorias e nos serviços de transporte interestadual, intermunicipal e de comunicação.

2.1.1 Origens e Incidência do Imposto

IVC – Imposto sobre Vendas e Consignações

Originária do art. 8º da Constituição Brasileira de 1934 [3] a primeira tributação sobre operações mercantis aconteceu com o advento do Imposto sobre Vendas e Consignações (IVC), imposto de competência estadual que incidia sobre cada venda realizada (fato gerador) **cumulativamente** em todas as fases de circulação, do produtor até o consumidor.

Dessarte sua incidência acontecia “em cascata” em cada uma das operações sucessivas de venda de uma mesma mercadoria, do produtor até o consumidor final, sempre incidindo sobre a base de cálculo integral (o preço da mercadoria).

ICM – Imposto sobre Operações Relativas à Circulação de Mercadorias Realizadas por Comerciantes, Industriais e Produtores

Pelo o advento da Emenda Constitucional 18 [4] (art. 12), de 1º de dezembro de 1965, o IVC foi substituído pelo Imposto sobre Operações Relativas à Circulação de Mercadorias Realizadas por Comerciantes, Industriais e Produtores (ICM), igualmente de competência estadual. Distintamente do IVC o ICM possui a natureza de **não-cumulatividade** do tributo (parágrafo 2º), destacando-se como o primeiro imposto brasileiro sobre o valor agregado. Assim, o valor a recolher do ICM deveria ser calculado sobre o valor adicionado por cada operação sucessora.

Em adição, foram destinadas à competência da União:

a) a tributação sobre os serviços de transporte e comunicações interestaduais (art. 14, II), e

b) os Impostos Especiais (art. 16), sobre a:

   - produção, importação, circulação, distribuição ou consumo de combustíveis e lubrificantes líquidos ou gasosos de qualquer origem ou natureza;
   - produção, importação, distribuição ou consumo de energia elétrica;
   - produção, circulação ou consumo de minerais do País.

ICMS - Imposto sobre Operações Relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação

A Constituição Federal de 1988 [2] instituiu o novel sistema tributário nacional - em vigor a partir de 1º de março de 1989 - e criou (art.155, II) o atual Imposto sobre Operações Relativas à Circulação de Mercadorias e sobre Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação (ICMS), extinguindo os impostos especiais e incorporando combustíveis, lubrificantes e energia elétrica à competência estadual. Ademais, a Carta Magna ao estabelecer o campo de incidência do ICMS, nele incluiu as prestações de serviços de transporte intermunicipal, interestadual e de comunicação (estes últimos anteriormente da competência da União).

Outrossim, a novel Carta Constitucional manteve a característica de **não-cumulatividade** do imposto (tributação sobre o valor agregado).

2.1.2 Não Cumulatividade

Como um tributo incidente sobre o valor agregado, consigna o artigo 155, parágrafo segundo, inciso II, da Constituição Federal [2], que o ICMS será não-cumulativo, compensando-se o que for devido em cada operação relativa à circulação de mercadorias ou prestação de serviços com o montante cobrado nas operações anteriores pelo mesmo ou outro Estado (ou pelo Distrito Federal).

Tal modalidade de tributação constitui uma sistemática de conjugação de débitos e créditos, onde diminui-se do valor da obrigação tributária a recolher o imposto pago em etapas precedentes, pelas aquisições de bens ou serviços anteriormente tributados. O que significa dizer que o montante a pagar (devido) no período é o resultado do cotejo compensatório entre: a) o imposto apurado nas operações de vendas da empresa e b) os valores já pagos (e creditados) anteriormente, a título do ICMS sobre a formação de estoques de mercadorias para comercialização, insumos e ativos adquiridos para a produção.

Melhor dizendo, o valor a recolher do ICMS em determinado período é o resultado do confronto entre os débitos do imposto, gerados pelas vendas (saídas) de mercadorias ou prestações de serviços realizadas, diminuído dos créditos (ICMS) advindos das aquisições (entradas) acontecidas.

Assim, como regra geral, o valor da obrigação tributária do ICMS é dado pela função:

$$ICMS_{Recolher} = DÉBITOS_{Vendas} - CRÉDITOS_{Compras} \tag{2.1}$$

É importante o registro dessa característica distintiva do imposto, porquanto a análise de sua regularidade deverá considerar que os ilícitos fiscais (*v.g.* as fraudes e a sonegação) poderão acontecer por alteração imprópria de qualquer um dos componentes da equação; seja pela supressão indevida da realidade dos débitos, seja pelo incremento insustentável dos créditos acumulados.

2.2 Fraude e Sonegação Fiscal no Direito Brasileiro

Para a melhor compreensão do conceito técnico jurídico de "fraude"e "sonegação"tributárias, impõe-se uma breve invocação do domínio positivo de sua definição no direito administrativo e penal tributários.

Sobre o conceito de fraude a obra do mestre De Plácido Silva [5] ensina-nos:

> **"FRAUDE. Derivado do latim *fraus*, *fraudis* (engano, má fé, logro), entende-se geralmente como o engano malicioso ou a ação astuciosa, promovidos de má fé, para ocultação da verdade**

ou fuga do cumprimento do dever.
Nestas condições, a fraude traz consigo o sentido do engano, não como se evidencia no dolo, em que se mostra a manobra fraudulenta para induzir outrem à pratica de ato, de que lhe possa advir prejuízo, mas o engano oculto para furtar-se o fraudulento ao cumprimento do que é de sua obrigação ou para logro de terceiros. É a intenção de causar prejuízo a terceiros. Assim, a fraude sempre se funda na prática de ato lesivo a interesse de terceiros ou da coletividade, ou seja, em ato, onde se evidencia a intenção de frustrar-se a pessoa aos deveres obrigacionais ou legais.
É, por isso, indicativa de lesão de interesses individuais, ou contravenção de regra jurídica, a que se está obrigado. O dolo é astúcia empregada contra aquele com quem se contrata.
(...)
Além do sentido de contravenção à lei, notadamente fiscal, possuí o significado de contrafação, isto é reprodução imitada, adulteração, falsificação, inculcação de uma coisa por outra.
Aliás, em todas as expressões, está no seu sentido originário de engano, má fé e logro, todos fundados na intensão de trazer um prejuízo, com o qual se locupletará o fraudulento ou fraudador"

Em especial, respeitante a fraude fiscal, pontifica a lição constante do aludido Vocabulário Jurídico:

"FRAUDE FISCAL. É a contravenção as lei ou regras fiscais, com o objetivo de fugir ao pagamento do imposto devido ou o de passar mercadoria de uma qualidade ou procedência por outra."

Já pertinente ao conceito de sonegação, consigna o celebrado autor:

"SONEGAÇÃO. De sonegar, do latim *subnegare* (negar de algum modo), entende-se a ocultação, ou a subtração de alguma coisa ao destino, que lhe é reservado.
No conceito jurídico, a sonegação envolve sempre a ocultação ou a subtração dolosa de coisa, que deveriam ser mostradas, ou trazidas a certos lugares, a fim de que se satisfaçam mandos legais. Assim a sonegação importa em procedimento doloso e contrário a normas legais instituídas"

Elucidando a sonegação fiscal, esclarece:

"SONEGAÇÃO FISCAL. Em sentido fiscal, a sonegação, em princípio, designa a evasão do imposto por meio de artifícios ou manejos dolosos do contribuinte. Quer significar, pois, a falta de pagamento do imposto devido, ou a subtração ao pagamento do imposto, mediante o emprego de meios utilizados com esse objetivo.
Desse modo, a sonegação não implica numa falta de pagamento involuntária ou decorrente da falta de recursos, mas no emprego de meios para se furtar a esse pagamento.

No conceito fiscal, porém, nem toda sonegação é reputada dolosa: há a sonegação dolosa e a simples sonegação.
A sonegação simples é a que resulta da falta de pagamento do imposto, sem qualquer malícia, ou sem o emprego de ardil, ou fraude, com o que se procura subtrair ao cumprimento da imposição fiscal.
A sonegação doloda, ou a sonegação fraudulenta, é a que se gera da fraude ou da má fé do contribuinte, usando meios, manobras, ou ardis para se furtar, ou se subtrair ao pagamento do imposto."

Sem embargo da extensa e esclarecedora definição dada pela doutrina jurídica, importa saber que os conceitos positivos de sonegação e fraude fiscal estão materialmente assentados na legislação tributária brasileira e insculpidos nos arts. 71 e 72 da Lei nº 4.502/64 [6], *ad litteram*:

"LEI Nº 4.502, DE 30 DE NOVEMBRO DE 1964
(...)
Art. 71 - Sonegação é toda ação ou omissão dolosa tendente a impedir ou retardar, total ou parcialmente, o conhecimento por parte da autoridade fazendária:
I – da ocorrência do fato gerador da obrigação tributária principal, sua natureza ou circunstâncias materiais; e
II – das condições pessoais do contribuinte, suscetíveis de afetar a obrigação tributária principal ou o crédito tributário correspondente.
Art. 72 - Fraude é toda ação ou omissão dolosa tendente a impedir ou retardar, total ou parcialmente, a ocorrência do fato gerador da obrigação tributária principal, ou a excluir ou modificar as suas características essenciais, de modo a reduzir o montante do imposto devido, ou a evitar ou diferir o seu pagamento.
(...)"

De lembrar que a sobretranscrita Lei nº 4.502/64 [6] tratava originalmente do antigo Imposto sobre Consumo (de competência da União).

A posteriori, sob a égide do novo sistema tributário patrocinado pelo Código Tributário Nacional (CTN) [7] de 1966, o Decreto-Lei nº 34 [8], de 18 de novembro de 1966, aproveitou o texto da Lei nº 4.502/64 [6] adaptando seus preceitos para o recém criado Imposto sobre Produtos Industrializados (IPI).

Igualmente, o CTN [7] elegeu como hipóteses de revisão do lançamento de ofício pela autoridade administrativa o dolo, a fraude e a simulação, *ad verbum*:

"LEI Nº 5.172, DE 25 DE OUTUBRO DE 1966.
(...)
Art. 149 - O lançamento é efetuado e revisto de ofício pela autoridade administrativa nos

seguintes casos:
(...)
VII - quando se comprove que o sujeito passivo, ou terceiro em benefício daquele, agiu com dolo, fraude ou simulação;
(...)
IX - quando se comprove que, no lançamento anterior, ocorreu fraude ou falta funcional da autoridade que o efetuou, ou omissão pela mesma autoridade, de ato ou formalidade essencial."

Observe-se que o legislador pátrio adota o dolo como elemento subjetivo necessário nos conceitos e nas circunstâncias revisionais colecionados. Portanto, é forçoso concluir que os conceitos de fraude e sonegação, admitidos no direito tributário brasileiro, requestam o elemento dolo para sua configuração, melhor dizendo, a volição do agente de praticar ato ilícito na intenção de prejudicar o sujeito passivo em benefício particular.

Os conceitos definidos na Lei nº 4.502/64 [6] foram incorporados *ipsis verbis* na legislação tributária distrital, a saber o excerto do art. 62 do Código Tributário do Distrito Federal [9], à letra:

"LEI COMPLEMENTAR DISTRITAL Nº 4, DE 30 DE DEZEMBRO DE 1994
Art. 62 -
(. . .)
§ 1º - Verificando-se a ocorrência de sonegação, fraude ou conluio, aplicar-se-á multa de 200% (duzentos por cento) do valor do imposto.
§ 2º - Para os efeitos do parágrafo anterior, considera-se:
I - sonegação, toda ação ou omissão dolosa tendente a impedir ou retardar, total ou parcialmente, o conhecimento, por parte das autoridades fiscais:
a) da ocorrência do fato gerador da obrigação tributária principal, sua natureza ou suas circunstâncias materiais;
b) das condições pessoais do contribuinte, suscetíveis de afetar a obrigação tributária principal ou o crédito tributário correspondente;
II - fraude, toda ação ou omissão dolosa tendente a impedir ou retardar, total ou parcialmente, a ocorrência do fato gerador da obrigação tributária principal, a excluir ou modificar suas características essenciais, de modo a reduzir o montante do imposto devido, ou a evitar ou diferir o seu pagamento;"

Acrescente-se, que a evasão fiscal (mediante fraude ou sonegação), ao revés da elisão (planejamento tributário), representa a conduta ilícita tributária, que infringe a norma fiscal, com o objetivo de subtrair ou diminuir o cumprimento da obrigação tributária. Nessa condição o sujeito passivo (contribuinte) também praticará crime contra a Ordem Tributária.

As espécies de configuração dos crimes contra a Ordem Tributária encontram-se lançadas na Lei nº 8.137/90 [1] [10], norma que revogou tacitamente a anterior Lei dos Crimes de Sonegação Fiscal[2], a saber:

> **"LEI Nº 8.137, DE 27 DE DEZEMBRO DE 1990.**
> **Art. 1º Constitui crime contra a ordem tributária suprimir ou reduzir tributo, ou contribuição social e qualquer acessório, mediante as seguintes condutas:**
> **I - omitir informação, ou prestar declaração falsa às autoridades fazendárias;**
> **II - fraudar a fiscalização tributária, inserindo elementos inexatos, ou omitindo operação de qualquer natureza, em documento ou livro exigido pela lei fiscal;**
> **III - falsificar ou alterar nota fiscal, fatura, duplicata, nota de venda, ou qualquer outro documento relativo à operação tributável;**
> **IV - elaborar, distribuir, fornecer, emitir ou utilizar documento que saiba ou deva saber falso ou inexato;**
> **V - negar ou deixar de fornecer, quando obrigatório, nota fiscal ou documento equivalente, relativa a venda de mercadoria ou prestação de serviço, efetivamente realizada, ou fornecê-la em desacordo com a legislação.**
> **Pena - reclusão de 2 (dois) a 5 (cinco) anos, e multa.**
> **Parágrafo único. A falta de atendimento da exigência da autoridade, no prazo de 10 (dez) dias, que poderá ser convertido em horas em razão da maior ou menor complexidade da matéria ou da dificuldade quanto ao atendimento da exigência, caracteriza a infração prevista no inciso V.**
> **Art. 2º Constitui crime da mesma natureza:**
> **I - fazer declaração falsa ou omitir declaração sobre rendas, bens ou fatos, ou empregar outra fraude, para eximir-se, total ou parcialmente, de pagamento de tributo;**
> **II - deixar de recolher, no prazo legal, valor de tributo ou de contribuição social, descontado ou cobrado, na qualidade de sujeito passivo de obrigação e que deveria recolher aos cofres públicos;**
> **III - exigir, pagar ou receber, para si ou para o contribuinte beneficiário, qualquer percentagem sobre a parcela dedutível ou deduzida de imposto ou de contribuição como incentivo fiscal;**
> **IV - deixar de aplicar, ou aplicar em desacordo com o estatuído, incentivo fiscal ou parcelas de imposto liberadas por órgão ou entidade de desenvolvimento;**
> **V - utilizar ou divulgar programa de processamento de dados que permita ao sujeito passivo da obrigação tributária possuir informação contábil diversa daquela que é, por lei, fornecida à Fazenda Pública.**
> **Pena - detenção, de 6 (seis) meses a 2 (dois) anos, e multa."**

Observe-se que o art. 1º da supradita Lei 8.137/90 [10] prescreve as hipóteses de crime material, onde a elementar do tipo penal exige para a sua conformação o resultado de supressão ou de redução da obrigação tributária devida.

Entrementes, esse artigo busca garantir a certeza e a veracidade das informações fiscais prestadas, prevendo, em seu inciso I, a censura às condutas omissivas (omitir) e comissivas

[1] Atualizada pela Lei nº 9.964, de 10 de abril de 2000.
[2] Lei nº 4.729, de 14 de julho de 1965.

(prestar) que as prejudiquem e, no inciso II, espécie de norma penal em branco que combate a fraude nas informações consignadas em documentos e livros exigidos na legislação tributária.

Ademais, cuida o artigo da falsidade material consignada em documentos (e do próprio documento) fiscais obrigatórios (incisos III e IV) e da circunstância de crime instantâneo pela falta da emissão do documento obrigatório (inciso V) na operação econômica empreendida. De outro modo, o parágrafo único do art. 1º trata de crime de mera conduta omissiva (desobediência) que independente do resultado para se aperfeiçoar.

Já o art. 2º propõe as hipóteses de crime formal - onde o tipo penal exige somente a conduta do agente para sua consumação -, com a peculiariedade da previsão do crime de apropriação indevida de tributos e de contribuições (inciso II), e a tutela protetiva da correição no emprego dos incentivos fiscais (incisos III e IV).

Por essa breve exposição do tratamento jurídico atribuído aos conceitos de fraude e sonegação no direito pátrio, é necessário compreender que essas espécies de ilicitudes penal e administrativa-penal exigem rigor jurídico para a sua caracterização e especialmente para a imposição das sanções advindas de sua existência.

2.3 Exemplos das Ilicitudes Fiscais

Lembrando que a construção do valor devido à título do tributo em tema (ICMS), em um período de apuração, acontece pelo confronto do total de débitos, provenientes das operação comerciais ou de prestação de serviços do contribuinte, com o total dos créditos remanescentes e acumulados pelas aquisições efetivadas (equação 2.1), é possível asseverar que as hipóteses de ilicitudes fiscais poderão ocorrer em qualquer uma das variáveis da função de cálculo do imposto a recolher.

Destacam-se os principais métodos de evasão fiscal conhecidos e combatidos em auditoria (descrição exemplificativa, e não exaustiva)[3]:

EVASÃO TRIBUTÁRIA NOS CRÉDITOS APROPRIADOS - que se materializam pelo incremento incompossível dos valores dedutíveis por não cumulatividade, intencionando promover a redução da obrigação tributária devida por excessos insustentáveis de crédito do ICMS. Geralmente tal expediente acontece por:

a) CRÉDITOS de operações (de aquisições):

- Escriturados com valor maior que o real.
- Que nunca aconteceram (inexistentes) ou que foram canceladas.

[3]Os exemplos são derivados da experiência do autor na Auditoria Tributária do ICMS.

- Isentas (ou não tributáveis) e que não permitem creditamento.

b) CRÉDITOS de origem imprópria, posto que advindos de:

- Empresas do Simples Nacional, quando o montante está acima do limite permitido.
- Empresas baixadas ou canceladas.
- Operações de Substituição Tributária; modalidade de tributação que não está sujeita à conjugação entre débitos e créditos.
- Aquisições de bens destinados ao uso e consumo do adquirente e que não dá, até o momento, direito ao creditamento.

c) CRÉDITOS advindos do uso de alíquota:

- Interna ou interestadual maior do que a permitida.
- Com diferencial não pago no destino.

EVASÃO TRIBUTÁRIA NOS DÉBITOS ESCRITURADOS - Outrossim, as hipóteses de fraudes podem se revelar na apuração do débito do imposto devido. Desta feita, há a intenção de diminuir ou ocultar o valor real do tributo próprio nas operações realizadas, pelos seguintes artifícios:

a) DÉBITOS de operações (vendas) escriturados com o valor do ICMS:

- Menor que o real destacado na Nota Fiscal, Cupom Fiscal ou Nota Fiscal ao Consumidor Eletrônica (NFCE) emitida.
- Omitido (não escriturados), não obstante a realidade da venda.

b) DÉBITOS ocultados, originários de operações com:

- Cartão de crédito/débito não apropriados.
- O governo (vendas governamentais não escrituradas).
- Nota Legal emitida (porém, não escriturada).

c) DÉBITOS calculados com alíquota:

- Interna menor que a prevista no DF.
- Interestadual menor que 12% (doze por cento)[4].

[4]Resolução do Senado Federal número 22 de 1989.

Capítulo 3

Metodologia da Pesquisa e Dados Utilizados

Neste capítulo serão apresentados: a metodologia da pesquisa (estrutura de emprego dos métodos e modelos escolhidos) os dados úteis a sua aplicação, suas fontes de consulta, bem assim as ferramentas e estratégias indicadas para a sua extração e tratamento. Igualmente, será oferecida uma apreciação do significado tributário e a importância de cada uma das informações trabalhadas.

Posteriormente, na apresentação de cada etapa da metodologia proposta, aborda-se-á a especialidade dos dados adequados aos modelos e aos métodos aplicados naquela fase em exposição.

Em regra, as informações trabalhadas pertencem ao conjunto de 5 (cinco) anos de escrituração fiscal e de operações comerciais dispostas em documentos fiscais. Esse período é coincidente com o prazo decadencial proposto no art. 173 do CTN [7]. Não obstante, esse intervalo temporal poderá ser ampliado *v.g* para a melhor construção das séries históricas, ou reduzido, a exemplo do que se dá no estudo anual de eficiência relativa dos contribuintes.

3.1 Metodologia da Pesquisa (Aplicação)

Este estudo dedica-se à construção de uma metodologia instrumental prática desenvolvida no *Software R*[5] que selecionará e tratará as informações de interesse tributário, disponíveis nas bases de dados da Receita Distrital (GEPRO), submetendo-as a exames estatísticos e de mineração computacional de dados.

[5]R é um ambiente de software livre para computação estatística e construção de gráficos. Informações consultar: *The R Project for Statistical Computing - www.r-project.org.*

Os algoritmos analíticos adotados concluirão suas finalidades com a disponibilização das informações sobre as anomalias *outliers* percebidas, por meio de relatórios próprios à cada uma das etapas do processo de análise.

A escolha do *Software R* foi inevitável posto que resta comprovado o seu excelente desempenho em pesquisas científicas de alto nível, bem assim, por que é um software livre que não impõe custos e embaraços para a sua implementação. Ademais, é notável a sua capacidade de oferecer a desejável reprodutibilidade das soluções e dos resultados alcançados.

A metodologia preconiza um sequencial de aplicação que, em resumida descrição, consiste nos seguintes requisitos:

a) Estabelecer um mecanismo consistente de acesso à extração dos dados úteis ao funcionamento dos modelos aplicados.

   Nesse sentido é desenvolvido um método de consulta que concilia o banco de dados *ORACLE*[6], onde persistem os dados armazenados, com o *Software R*, onde as informações serão tratadas.

b) Disponibilizar um recurso para a seleção de empresas, fundado na técnica de Análise Envoltória de Dados (DEA), que proporcione o discernimento dos contribuintes dignos de observação, diante de um colegiado de pares (colecionados, *v.g.*, por sua atividade econômica principal ou por outras qualidades aglutinadoras).

   Será oferecido um painel geral contendo o resultados das eficiências relativas na arrecadação do ICMS encontrados no setor. A partir dos resultados apurados (ranqueados pelo modelo DEA) decidir-se-á a eleição dos contribuintes com baixa eficiência contributiva relativa, candidatos às demais etapas analíticas em busca de momentos e valores *outliers*.

c) Realizar a Análise das Séries Temporais Fiscais dos contribuintes escolhidos para a identificação dos períodos de mês e ano onde verifique-se excentricidades.

d) Manejar um tratamento analítico para a descoberta de *outliers* com a aplicação de modelos estatísticos (paramétricos e não-paramétricos) e do método de mineração de dados por clusterização *K-means*.

e) Proporcionar, uma conclusão analítica pertinente aos dados fiscais dos contribuintes considerados, assinalando os desvios discrepantes encontrados.

Ao fim desse processo haverá a disponibilidade de painéis interativos (*dashboards*) contendo o diagnóstico do setor e das empresas sob exame. Esses painéis são construídos

[6]ORACLE é marca registrada de um Sistema Gerenciador de Banco de Dados (SGBD) comercial. Informações consultar: *www.oracle.com/database/index.html*.

utilizando-se o pacote *Shiny*[7] (*R Markdown*[8]) e deverão indicar as situações, os momentos temporais, os documentos e os valores discrepantes encontrados.

3.2 Restrições de Sigilo Fiscal

Em respeito ao preceito de preservação do sigilo dos dados e informações econômico-financeiras dos contribuintes, inobstante os dados usados neste trabalho serem reais, a identificação das empresas participantes do estudo será omitida (art. 198 do CTN [7], combinado com art. 325 do Código Penal [11]).

3.3 Fontes das Informações

Para o desenvolvimento da metodologia proposta fez-se uso das seguintes fontes de informação extraídas das bases de dados disponíveis para a Programação Fiscal (GEPRO):

a) **Banco de Dados das Notas Fiscais Eletrônicas (NFEs)**[9], que contempla os documentos fiscais, gerados digitalmente, certificados e aprovados eletronicamente, e que explicitam o movimento de mercadorias (ou serviços) em uma operação comercial de compra e venda (ou prestação de serviços), bem assim o valor dos produtos e o ICMS pertinente à operação. Distinguem-se as seguintes espécies de NFEs :

- **Nota Fiscal Eletrônica de Entrada (NFE-E)**, aquelas que traduzem a formação dos estoques comerciais ou investimentos em ativos de produção das empresas (compras para produção ou revenda). Esta espécie de NFEs dá direito à apropriação e ao aproveitamento do ICMS pago na operação;
- **Nota Fiscal Eletrônica de Saída (NFE-S)**, aquelas que reportam as vendas realizadas entre empresas contribuintes (formação de estoques) bem como as vendas a consumidor final (pessoa física ou jurídica). Estas NFEs geram débitos do ICMS (obrigação a pagar) pelos valores do imposto nelas destacados.

As **NFEs (NFE-E e NFE-S)** podem representar:

- Operações realizadas por contribuintes locais dentro da mesma Unidade Federada (*in casu* o DF) – Operações Internas –, ou

[7] *Shiny* é um pacote do *Software R* que facilita a criação de aplicativos web interativos. Para maiores informações consultar: *shiny.rstudio.com.*

[8] *R Markdown* é uma sintaxe de formatação simples para a autoria de documentos em HTML, PDF e MS Word. Para maiores informações consultar: *rmarkdown.rstudio.com/.*

[9] Informações sobre a Nota Fiscal Eletrônica consultar: *www.nfe.fazenda.gov.br/portal/principal.aspx.*

- Operações realizadas entre empresas de Unidades Federadas distintas - Operações Interestaduais. Estas últimas sujeitam-se ao regime de repartição de receitas por diferenciação das alíquotas do imposto (7% para a origem, quando dos Estados de SP, RJ, MG, RS, SC, PR e 12% para a origem dos demais Estados)[10].

Nesse estudo somente são consideradas as NFEs válidas, ou seja, aquelas que não encontram-se canceladas, por qualquer motivo.

Outrossim, registre-se a circunstância de que grande parte das operações internas de venda a consumidor final (especialmente pessoa física) no Distrito Federal, são realizadas com o uso de Cupom Fiscal, circunstância que permanecerá até a consolidação total da Nota Fiscal ao Consumidor Eletrônica (NFCE) no fim do ano de 2018. Esse documento (cupom), assim como a NFCE não permite a apropriação de crédito pelo ICMS pago na operação.

b) **Banco de Dados do Livro Fiscal Eletrônico (LFE)**[11], que representa a escrituração fiscal do contribuinte oferecida ao Fisco Distrital, por meio eletrônico, estando registrado nele as informações contábeis de interesse tributário, em especial, as pertinentes à apuração do valor do ICMS a recolher no período.

Na escrituração fiscal realizada no LFE, conjugam-se os créditos (a recuperar) – resultantes das aquisições anteriores - com os débitos (a realizar) resultantes das operações de saídas (tributadas). O cotejo dos créditos com os débitos definirá o ICMS a ser pago na apuração mensal.

Especificamente nessa pesquisa trabalhar-se-á com o módulo E020 do LFE, posto que contém a individualização dos dados de todos os documentos fiscais escriturados, pelo contribuinte, a título de créditos e débitos do ICMS.

c) **Banco de Dados do Cartão de Crédito e Débito**, que contém as informações e os valores pertinentes às vendas realizadas pelos contribuintes com o uso do cartão como meio de pagamento, na modalidade de Crédito e/ou Débito. Essas informações são fornecidas pelas Administradoras de Cartão, consoante comanda a Lei Complementar Distrital 772, de 17 julho de 2008 [12].

[10]Resolução do Senado Federal número 22 de 1989.
[11]Informações sobre o Livro Fiscal Eletrônico do DF consultar: *www.fazenda.df.gov.br*.

3.4 Dados Utilizados na Metodologia Proposta

Os dados objetos da aplicação e da análise pelos modelos e métodos associados foram extraídos do banco de dados *ORACLE*[12] da Receita Distrital, com o uso do programa estatístico *R*[13] - sob a plataforma de desenvolvimento *RSTUDIO*[14] - e do pacote *RODBC*[15], que permite o uso da linguagem *Structured Query Language (SQL)* para a criação de consultas (*queries*) para a extração de dados diretamente de seus repositórios.

Presentes nas sobreditas fontes, os dados de interesse na aplicação dos modelos e métodos patrocinados (conjugados) são:

a) Os **Valores Contábeis** respectivos às **NFEs** e aos registros de documentos fiscais (eletrônicos e manuais) no **LFE**, correspondentes às operações de entradas (crédito) e de saídas (débitos) realizadas.

b) Os **Valores Contábeis** das operações com **Cartão** que representam o faturamento do contribuinte obtido com o uso desse meio de pagamento na função de Crédito e/ou Débito.

c) Os **Valores dos Créditos do ICMS** definidos em cada operação de entrada e que representam o montante do imposto passível de apropriação e compensação futura. Especificamente:

   - Os valores **destacados nas NFEs** emitidas pelos fornecedores (somente as válidas e não canceladas - situação 100), consoantes às compras realizadas (entradas) para formação de estoques, insumos ou como investimento em ativo imobilizado (*v.g.* máquinas e equipamentos).
   - Os valores de **ICMS escriturados** como crédito pelo contribuinte no seu **LFE**, que, por regra, devem repercutir somente os créditos passíveis de apropriação, correspondentes às respectivas NFEs recebidas em operações onde é possível o aproveitamento do imposto pago.

d) Os Valores dos **Débitos do ICMS** resultantes das vendas (débito de saída), que revelam o imposto devido nas operações/prestações de bens ou serviços tributáveis promovidas pelo contribuinte. Esses valores devem ser recolhidos ou compensados com os fortuitos créditos existentes. Tal informação está consignada:

[12]ORACLE *Database 11g, Enterprise Edition, version 11.2.0.1.0.*
[13]*Software R, version R-3.4.2, for Windows 32/64 bits.*
[14]RSTUDIO - *Integrated Development Environment (IDE) for R, version 1.1.383.*
[15]RODBC - *ODBC Database Access Interface for R, version: 1.3-15.*

- no **valor destacado nas NFEs**, a título do ICMS, emitidas pela empresa para seus clientes. Serão consideradas somente as NFEs válidas e não canceladas - situação 100.
- no **valor dos débitos do ICMS lançado no LFE** e correspondentes às saídas tributadas promovidas pelo contribuinte.

3.5 Estratégia de Extração dos Dados Utilizados

As consultas de obtenção dos dados de interesse foram escritas em *SQL* diretamente como sub-rotinas dos programas criados em linguagem do *Software R* e executadas fazendo uso da conexão *ODBC*[16] proporcionada pelo pacote *RODBC*.

Como estratégia para melhor eficiência na extração dos dados limitou-se o processo de requisição dos dados no banco à pertinência de utilidade das informações em cada etapa da metodologia. Assim, a obtenção dos dados acontece à proporção do seu uso em cada fase proposta e não de uma só vez.

Essa estrutura de requesta de dados, permite a aplicação diferenciada de cada etapa da metodologia sem o desperdício de tempo dedicado à coleta de dados que ora não serão utilizados. Igualmente, essa segmentação otimiza a concorrência no banco de dados e o fluxo da rede.

[16] *ODBC* é um acrônimo para *Open Database Connectivity* que é um padrão para acesso a Sistemas Gerenciadores de Bancos de Dados (SGBD).

Capítulo 4

Seleção de Contribuintes por suas Eficiências Contributivas Relativas

4.1 Seleção de Contribuintes de um Setor Econômico

A primeiro etapa da metodológia propugnada é a escolha de uma, ou mais, empresas que, por sua ineficiência contributiva, tornem-se merecedoras de uma maior investigação concernente ao comportamento de seus resultados fiscais (realidade contributiva).

Por eficiência contributiva deve-se entender o resultado da arrecadação (a título do ICMS) de um contribuinte em função do seu movimento comercial desenvolvido em um período de tempo.

O caminho para a escolha de contribuintes suspeitos, destinados à investigação aprofundada, poderá ser realizada pela comparação das performances arrecadatórias das empresas dentro de um mesmo setor da economia. É correto dizer que, em regra, essas empresas compartilham os mesmos parâmetros de exigências fiscais, condição que assegura a homogeneidade entre os agentes.

Esse pressuposto de homogeneidade será incorreto no cotejo de sujeitos passivos com cargas tributárias diferentes. Assim, não será prudente comparar contribuintes com regimes fiscais distintos (*v.g.* regimes: Normal, Simples Nacional ou Incentivado por eventuais benefícios fiscais). Por esta razão, a apuração da eficiência relativa sempre deve segmentar os contribuintes em função da forma de cálculo do ICMS e do setor econômico pertinente.

Para medir, e comparar, a eficiência relativa dos contribuintes introduz-se um método não paramétrico de otimização que faz uso da Análise Envoltória de Dados (DEA) - *Data Envelopment Analysis* -, que estabelecerá uma hierarquia comparativa entre as empresas estudadas, por meio da combinação dos respectivos resultados de arrecadação em função da sua realidade empresarial (compras, vendas e recebimentos).

A solução faz uso da DEA como ferramenta de mensuração e de ranqueamento da eficiência arrecadatória relativa dos contribuintes do ICMS que compartilham um mesmo segmento econômico e que estão subjugados a um mesmo tratamento tributário.

4.2 Análise Envoltória de Dados - DEA

A DEA é um método multivariado de Pesquisa Operacional, utilizado para analisar a eficiência de produtividade das unidades de decisão das Unidades de Tomada de Decisão (DMU)- *Decision Making Units* (DMU), que estabelece um indicador da eficiência relativa consoante o uso dos insumos - entradas (*inputs*) - e os resultados dos produtos realizados - saídas (*outputs*) - por essas unidades e fornece dados quantitativos sobre possíveis direções para melhorar o desempenho das DMUs, quando estas são ineficientes.

Como paradigma de avaliação o método cria uma unidade composta teórica formada pela combinação convexa das unidades de referência inerentes às DMUs. Essa unidade composta é uma unidade hipotética eficiente de comparação, que é construída a partir das unidades de referência correspondentes aos pesos duplos de ponderação (proporções) de cada DMU, e servirá como referência do ideal de eficiência.

Também conhecida como Análise de Fronteira, a DEA é estabelecida em modelos de programação matemática não paramétrica de otimização, portanto, não realiza inferências estatísticas nem se apega à medidas de tendência central, testes de coeficientes ou formalização da análise de regressão. Nesse sentido, a DEA não reivindica a determinação das relações funcionais entre os insumos (*inputs*) e os resultados (*outputs*), permitindo a utilização de variáveis discricionárias, instrumentais, de decisão, exógenas e categóricas (incluindo *dummies*) em seus modelos e aplicações.

A boa reputação dessa ferramenta vem da sua simplicidade comparativa e da ampla aplicabilidade em vários problemas encontrados no mundo real. Praticamente qualquer condição que tenha múltiplas unidades (DMUs) que funcionem de forma semelhante e que esteja preocupada com a padronização do desempenho dessas unidades pode fazer uso desta técnica.

A DEA define o posicionamento competitivo relativo de um conjunto de organizações ou atividades comparando suas eficiências técnicas, de escala alocativas ou ineficiências. Esse modelo de programação matemática permite a avaliação simultânea de múltiplos recursos e produtos para cada DMU. A capacidade dessa entidade (DMU) de gerar resultados a partir de determinados insumos define a sua eficiência. Outrossim, entende-se que as DMUs menos eficientes podem melhorar sua eficiência ao limite das melhores DMUs, cuja eficiência é definida em 100% (ou 1).

Entre os principais atributos que compõem o modelo da DEA temos:

a) A aferição da eficiência relativa de cada organização produtiva (DMU) resumida com um único número (fator) que sintetiza as interações entre múltiplas entradas e saídas.

b) A possibilidade de identificar quantitativamente as economias necessárias de insumos ou o incremento produtivo ideal que permitirá que as DMUs ineficientes se tornassem eficientes.

Para fins de investigação fiscal, na metodologia proposta, a DEA tem como predicado operacional:

a) Permitir a classificação relativa da arrecadação dos contribuintes pertencentes ao mesmo setor econômico.

b) Possibilitar a seleção de contribuintes de interesse para uma melhor investigação de sua regularidade fiscal, com base em seu baixo desempenho de eficiência relativa.

Assim, é possível dizer que a DEA corrobora com a proposta do estudo nesta etapa que busca estabelecer a seleção primária de contribuintes que têm comportamentos incompatíveis com o setor econômico a que pertencem.

Empregar-se-á o modelo clássico da DEA, em particular o modelo com Retornos Constantes de Escala (CRS), na versão dos multiplicadores e na versão envoltória, ambas com orientação ao produto, para a solução do problema de seleção de empresas consoante sua individual eficiência tributária.

4.3 DEA na Literatura Científica

A utilização da DEA para a mensuração da eficiência relativa tem sido objeto de pesquisas acadêmicas desde o final dos anos 70 quando foi desenvolvido por Charnes, Cooper e Rhodes [13] (em 1978). O trabalho de Charnes *et al.* [13] e, posteriormente, o de Banker *et al.* [14] e o de Seiford e Thrall [15], formam os alicerces do desenvolvimento do método DEA.

Como ensina Cooper *et al.* [16] a DEA é uma aplicação particular da Pesquisa Operacional, que oferece uma apropriada solução para o problema do cálculo da eficiência relativa, com base em um modelo de programação linear. Extrai-se das lições de Cooper *et al.* que a DEA pode ser explicada como uma técnica não paramétrica, construída em programação linear, para a avaliação das eficiências de organizações e mensuração de desempenho de unidades operacionais ou tomadoras de decisão DMU, que atuam em um mesmo ramo de atividade, quando a presença de múltiplas entradas e múltiplas saídas torna difícil a comparação.

De acordo com Ferreira [17], a DEA é uma abordagem de programação linear, alternativa aos métodos estatísticos paramétricos clássicos fundados em comportamentos médios ou hipotéticas eficiências máximas, que proporciona estimar a eficiência relativa por um limite de fronteira (de eficiências), que informa quais pontos limitam a produtividade sobre o qual uma unidade produtiva hipotética é tecnicamente eficiente. O desiderato da técnica DEA é construir um conjunto referencial convexo onde as DMUs podem ser classificadas em unidades eficientes e ineficientes, tendo como referencial o contorno dessa superfície limite.

No ensinamento de Casu e Molyneux [18], DEA é um método de programação matemática para a definição da fronteira de produção (maximalizada) e observação das medidas de eficiências relativas individuais em comparação com a fronteira construída.

Para Osman [19] a solução da DEA afere a eficiência relativa de cada DMU em comparação com os melhores resultados apresentados. Conforme o autor, os desempenhos máximos apurados indicam a fronteiras de produção empírica que dispõem os limites aos resultados alcançáveis em um dado conjunto de recursos. Destarte, os fatores de eficiência de uma DMU são medidos a partir das posições relativas em cotejo com a fronteira estabelecida. Cada resultado representa o descritivo das habilidades e das restrições objetivas da unidade, presumindo-se que, contornadas as restrições e ampliadas as habilidades, os resultados possam ser melhorados.

Outra definição da DEA, dada por Zhu [20], é que ela é uma ferramenta com arrimo em programação matemática sendo um método que oferece a estimativa das melhores fronteiras de produção e de avaliação comparativa em relação à eficiência de múltiplas entidades. Em uma abordagem prática da DEA, o autor fornece modelos que podem ser usados na avaliação de desempenho e na definição de *benchmarking* para as empresas (em Zhu [21]).

Ademais, Zhu [22] define formalmente a DEA como uma metodologia dirigida para fronteiras e não para limites de tendência central. Ao revés de tentar encaixar um plano através do centro dos dados como em regressões estatísticas, o modelo define uma superfície linear fracionada que se estabelece em cima das observações. Devido a esta peculiar perspectiva, a DEA mostra-se particularmente adequada para descobrir relações que permanecem ocultas a outras metodologias.

Para o sobrecitado autor, a orientação empírica da DEA e a ausência de numerosas pressuposições que acompanham outras abordagens, tais como as formas habituais de análise de regressão estatística, determinam o seu uso em diversificados estudos envolvendo a estimativa de fronteiras de eficiência no governo, em organizações sem fins lucrativos e na iniciativa privada.

Esse método hoje pode ser sistematizado e facilmente resolvido com os programas

computacionais disponíveis no mercado, sendo que a sua utilidade é atestada pelo desenvolvimento de uma grande quantidade de publicações anunciando soluções práticas desenvolvidas com o uso da ferramenta ao longo dos anos - como enunciado em Cooper *et al.* [16], Emrouznejad [23] e Cook e Seiford [24].

Kassai [25], construindo uma aplicação contábil para o método, oferece a visão da DEA na perspectiva de uma curva de eficiência (ou de produtividade maximalizada) considerando a relação ótima entre insumos e produtos. Essa curva pode ser determinada como uma fronteira de eficiência. Assim, as unidades consideradas eficientes estarão em interseção com essa curva paradigma, enquanto as ineficientes se localizarão sob ela. A fronteira de eficiência servirá de referencial para que uma empresa ineficiente busque tornar-se eficiente.

Destaca a autora os resultados advindos da aplicação da DEA, que podem ser resumidos em: a) definição de uma superfície envoltória formada pelas DMUs de melhor desempenho (eficientes), que representam o conjunto de referência para as demais unidades; b) disposição de uma medida de desempenho, que se traduz na distância de cada unidade à fronteira e; c) projeções das unidades ineficientes na fronteira, compondo metas para essas unidades.

Ademais, pontifica Kassai [25], que as DMUs podem significar grupos empresariais, empresas individuais, unidades administrativas, desde que atentem para as exigências de que as unidades em observação: a) sejam comparáveis; b) atuem sob as mesmas condições; c) e os fatores de *inputs/outputs* sejam os mesmos, diferindo apenas na intensidade e magnitude.

Consoante mostrado em Tone [26] e Zhu [20], em variados estudos aplicados, a DEA tem sido usada para fornecer novos *insights* em diversas atividades e na identificação de melhores padrões de referência (*benchmark*). Os autores acrescentam o mérito de que, desde que a DEA foi introduzida pela primeira vez em 1978, até a sua forma atual, pesquisadores de diferentes campos do conhecimento a reconhecem como uma excelente e simples, metodologia para a modelagem de processos operacionais de apreciação de desempenho.

O método é largamente aproveitada em muitas áreas de pesquisa, como: produção fabril, sistema bancário (Wanke *et al.* [27] e) , sistema educacional, sistema de saúde (Safdar *et al.* [28] e Gholami *et al.* [29]), avaliações de gestão, comércio e em outras indústrias e organizações, inclusive a de serviços (Emrouznejad [23]).

Como esclarecido em Charne *et al.* [30], o modelo inicial da DEA construído por Charnes, Cooper e Rhodes [13] e batizado por suas iniciais CCR, é até hoje o modelo mais utilizado. Esse modelo tem arrimo na definição de eficiência total da unidade, estabelecida como uma proporção, que funciona com Retornos Constantes de Escala (CRS). No modelo

CCR, a ponderação de pesos está associado às variáveis de entradas (*inputs*) e saídas (*outputs*) associadas às DMUs. Cada peso duplo estabelece a importância da DMU na composição das variáveis entrada-saída da unidade compósita. A unidade compósita é uma combinação de unidades eficientes. Dessarte, uma dada DMU é ineficiente se o modelo duplo do CCR conseguir apresentar uma unidade compósita hipotética que a supere.

Já o modelo DEA - Retornos Variáveis de Escala (VRS), estabelecido por Banker, Charnes e Cooper [14], mede a eficiência técnica com uma restrição de convexidade do tamanho de escala para a unidade composta, que é a DMU virtual. Dessa forma, diferentemente dos modelos CCR que apresentam o mesmo resultados de eficiências quando calculados, tanto pelo viés de entrada quanto pelo de saída, os modelos BCC apresentam diferentes eficiências consoante a sua orientação (pelos insumos ou pelos produtos).

Os modelos CCR e BCC configuram a representação da realidade mediante, respectivamente, a eficiência total (CCR) e a eficiência técnica (BCC) que podem ser orientadas para os *input* ou para os *output* (Ferreira [17]).

Esclarece Coelli [31] que a orientação para os *inputs* procura resolver a questão: observado o padrão de saídas (*outputs*) da unidade, qual a redução possível nas entradas *inputs*, de modo a manter o corrente nível de saídas? Concernente aos modelos orientados para os *outputs*, procura-se a resposta à questão: dado o nível de insumos (*inputs*) utilizado, qual o maior nível de *outputs* que pode ser alcançado, mantendo-se o nível de entradas constante?

Ao longo dos anos, a aplicabilidade do DEA se expandiu fazendo-se necessário o surgimento de novos modelos matemáticos para suprir essa nova gama de aplicações em diversos setores. Com essa evolução, os modelos passaram a apresentar modificações em relação ao modelo original oriundas da incorporação de novos conceitos a cada modelo. Atualmente a Análise Envoltória de Dados conta com uma variedade de modelos que abrangem desde os modelos DEA clássicos (descritos anteriormente) e suas variações, até abordagens que combinam os modelos DEA com métodos mais sofisticadso como: o de simulação de Monte Carlo e a lógica fuzzy, consoante exposto em Tone [26], Emrouznejad [23] e Ghasemi *et al.* [32].

Por proporcionar uma solução de mensuração de eficiências entre empresas que compartilham similaridade econômica, a DEA é um método ideal para a identificação e a seleção dos melhores contribuintes elegíveis à análise de anomalias, porquanto poder-se-á encontrar os contribuintes com a menor eficiência em termos de suas arrecadações de tributos.

4.4 Método DEA Aplicado

4.4.1 Descrição do Modelo

O emprego do método da DEA pode ser orientado para as entradas (insumos-*inputs*) ou para a saída (produtos-*output*) da equação econômica, e geralmente considera múltiplos componentes de insumos e produtos em espaços multidimensionais. Na metodologia proposta a orientação será para um único produto (saída) representado pela arrecadação do imposto realizada pelo contribuinte.

No modelo de tributação, é de se pressupor que os produtos de arrecadação das empresas crescem na mesma proporção que crescem os insumos (compras, vendas e recebimentos), por um fator de contribuição equivalente a uma taxa média de carga fiscal a que se submete o setor econômico. Ou seja, a perspectiva de arrecadação é correlacionada linearmente a uma taxa média de incidência do ICMS nos produtos comercializados pelo segmento.

Assim, não há variação de escala no modelo. É por esta razão que a metodologia utiliza o modelo DEA de Retornos Constantes de Escala (CRS), uma vez que a tributação setorial seguirá um padrão constante de carga tributária.

O modelo radial (convexo) CRS, também nomeado pelas iniciais dos nomes de seus desenvolvedores, Charnes, Cooper e Rhodes (CCR) [13], é o primeiro e fundamental modelo DEA, construído sobre a noção de eficiência conforme definido na engenharia clássica. Esse modelo de retornos constantes (CRS) calcula uma eficiência geral para a unidade (DMU) onde tanto a eficiência técnica pura, quanto a eficiência da escala, são agregadas em um único valor.

Na abordagem tributária, o que importa é o movimento de produção (arrecadação) para a fronteira de eficiência, tendo em conta as condições econômicas experimentadas e as atividades negociais praticadas pelos contribuintes.

A aplicação do modelo proposto será orientado para o produto, assumindo que os insumos (compras, vendas e recebimentos) não variam a critério do Fisco - eles permanecem constantes, pois são realizações econômicas das empresas (exclusivamente) - e a produção varia para atingir a fronteira da produção eficiente. Assim, presume-se que os respectivos valores do movimento econômico sejam mantidos e haja variação na contribuição tributária do ICMS pertinente a cada contribuinte, de acordo com sua eficiência relativa.

Na técnica DEA-CRS, a eficiência relativa das DMUs pode ser calculada de duas maneiras:

a) Pelo Modelo de Multiplicadores Algébricos, dedicado a estabelecer a fronteira de eficiência por otimização algébrica dos pesos de cada componente de entrada e saída.

b) Pelo Modelo de Envoltório Dual.

O método **CRS no Modelo de Multiplicadores** – Orientado ao Produto é dado pela solução da seguinte expressão de otimização:

Considerando que **y** representa os produtos, **x** os insumos, **P** a produtividades, **E** a eficiência e μ, ν os coeficientes de ponderação (pesos), minimizar:

$$Minimizar\ \frac{1}{P_0} = \frac{\sum_{i=1}^{r} \nu_i x_{i0}}{\sum_{j=1}^{s} \mu_j y_{j0}} \tag{4.1}$$

Sujeito a:

$$\frac{\sum_{i=1}^{r} \nu_i x_{ik}}{\sum_{j=1}^{s} \mu_j y_{jk}} \geq 1\ \forall\ k = 1, 2, 3, ...z$$

$$\nu_i, \mu_j \geq 0\ \forall\ i\ e\ j$$

Transformando o problema de programação fracionária em linear (considerando constante os insumos), obtemos:

$$Minimizar\ \frac{1}{E_{f0}} = \sum_{i=1}^{m} \nu_i x_{i0} \tag{4.2}$$

Sujeito a:

$$\sum_{r=1}^{s} \mu_r y_{r0} = 1$$

$$\sum_{i=1}^{m} \nu_i x_{ik} - \sum_{r=1}^{s} \mu_r y_{rk} \geq 0\ \forall\ k = 1, 2, 3...z$$

$$\nu_i, \mu_j \geq 0\ \forall\ i\ e\ j$$

O modelo permite que se escolha para cada DMU os pesos das variáveis de entrada (ν) ou de saída (μ) - da maneira mais benevolente, desde que esses pesos aplicados às outras DMUs não gerem uma proporção menor do que 1.

Já o método **CRS no Modelo Envoltório** – Orientado ao Produto é dado pela solução da seguinte expressão:

Sendo ϕ o inverso da eficiência $\frac{1}{E}$ (de forma que $1 \leq \phi \leq \infty$) e que λ represente a contribuição de cada DMU para a fronteira de eficiência do modelo (grau de importância - *benchmark*).

$$Maximizar\ \phi \tag{4.3}$$

Sujeito a:

$$x_{i0} - \sum_{j=1}^{n} x_{ij}\lambda_j \geq 0\ \forall\quad i = 1, 2, 3, ..., m$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j - \phi.y_{r0} \geq 0\ \forall\quad r = 1, 2, 3, ..., s$$

$$\phi\ and\ \lambda_k \geq 0\ \forall\quad j = 1, 2, 3, ..., n$$

Por ser um modelo *dual*, o modelo multiplicador tem o mesmo valor que a função objetivo do modelo envoltório. Nesses tipos de modelos DEA, as saídas e as entradas virtuais da DMU são os produtos e os insumos que resultam no fim do processo de minimização (multiplicadores) ou maximização (envoltória) por programação matemática linear.

Ademais, para dar maior confiabilidade aos resultados e objetivos de eficiência de cada DMU, faz-se necessário considerar eventuais folgas na projeção das fronteiras de eficiência. Portanto, o melhor modelo deve considerar essas fortuitas folgas de projeção:

Sendo **S+** a folga de saída e **S-** a folga de entrada.

$$Maximizar\ \phi + \varepsilon\ (\sum_{i=1}^{m} S_i^- + \sum_{r=1}^{s} S_r^+) \tag{4.4}$$

Sujeito a:

$$\sum_{j=1}^{n} x_{ij}\lambda_j + S_i^- = x_{i0}\ \forall\quad i = 1, 2, 3, ..., m$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j - S_r^+ = \phi y_{r0}\ \forall\quad r = 1, 2, 3, ..., s$$

$$\phi\ and\ \lambda_j \geq 0\ \forall\quad j = 1, 2, 3, ..., n$$

Sendo ε um elemento não-arquimediano ínfimo, maior que zero e menor que qualquer número real.

O modelo baseia-se no fato de que, no mesmo segmento econômico, o resultado da coleta das eficiências deve permanecer dentro de uma variação mínima razoável, atinentes ao comportamento das compras e vendas (para outras empresas ou para o consumidor final). Essas informações de insumos representam o movimento econômico da empresa diretamente relacionado ao nível de contribuição fiscal esperado.

Obviamente, vários fatores podem explicar a discrepância de um contribuinte e merecem ser apreciados por outras técnicas de auditoria fiscal pertinentes.

A robustez do modelo CRS determina que, se a DMU for ineficiente, ele será realmente relativamente ineficiente. Essa é uma condição ótima para ser utilizada como critério de identificação de contribuintes suspeitos.

Isso não quer dizer que não haja problemas fiscais nas empresas consideradas eficientes, nem que a ineficiência observada não tenha uma explicação razoável e justa. No entanto, essa eventualidade não prejudica o uso da DEA como um indicador preliminar de possíveis irregularidades dignas de atenção e exame, posto que essa é a idéia que alicerça a pretensão da metodologia em tema.

Em resumo, o método em exposição consiste em uma técnica não-paramétrica para construir uma classificação relativa de eficiências alcançadas por empresas contribuintes partícipes de um mesmo segmento econômico. Nesse sentido, a identificação de comportamento anômalo poderá otimizar as atividades de inspeção, poupando tempo na atividade de programação fiscal.

4.4.2 Dados Requeridos - Extração e Tratamento

A extração das informações do banco de dados *ORACLE* e a aplicação da técnica DEA nesta etapa acontecem com o uso do *Software R* sob a plataforma *RSTUDIO*, usando respectivamente os pacotes *RODBC* e *Benchmarking*[17].

Considerando os dados disponíveis para os órgãos de Receita e as variáveis que explicam a função econômica que produz o ICMS (movimento econômico com bens e serviços), é estabelecido fazer uso dos seguintes dados para a composição do modelo DEA em tema:

OUTPUT (saída): Como o objetivo da metodologia é definir uma lista comparativa de eficiência de arrecadação do imposto para as empresas que participam de um segmento econômico comum, o único produto (*output*) de interesse para a medição dessa eficiência a ser relativizado na equação entrada-saída é o **imposto total anual pago (ICMS)**

[17] *Benchmarcking* é um pacote para análises de fronteira e envoltória de dados (DEA). Informações: *https://cran.r-project.org/web/packages/Benchmarking/Benchmarking.pdf*.

individualmente pelas empresas participantes. Especificamente, o montante arrecadado no ano de análise sob o Código de Receita do DF número 1317 (ICMS-NORMAL).

Outrossim, a escolha dos ***INPUTS*** (entradas) baseia-se nos parâmetros de movimento econômico das empresas comerciais (suas compras e vendas), o que conduz ao uso das seguintes variáveis de insumos:

a) **Soma anual do valor contábil das NFEs**, documentos gerados, certificados e aprovados eletronicamente, que arrimam o movimento de negociação com mercadorias em uma transação comercial de compra e venda (ou serviço), bem como traduzem o valor dos bens (ou das prestações) e do ICMS da operação. Somente documentos válidos e não cancelados devem ser utilizados. Far-se-á uso separadamente dos valores atinentes às:

   - **NFE-E**, para representar a formação dos estoques comerciais da empresa e do seu ativo imobilizado produtivo – aquisições de insumos (ou recursos de máquinas e equipamentos) para a produção ou de mercadorias para revenda.
   - **NFE-S**, para definir o movimento das vendas de bens (ou serviços) entre contribuintes corporativos ou ao consumidor final.

b) **Soma anual do faturamento obtido com o uso de Cartão de Crédito ou Débito** como meio de pagamento pelas vendas, porquanto esse tipo de operação representa aproximadamente 60% a 80% das vendas ao consumidor final (especialmente pessoa física).

   O uso da informação sobre o faturamento com cartão justifica-se uma vez que boa parte das vendas para pessoas físicas (consumo final individual) não são registradas nas NFEs, acontecendo sob o amparo de outros documentos fiscais como o Cupom Fiscal (a ser substituído até o fim do ano de 2018) e a Nota Fiscal ao Consumidor Eletrônica (NFCE). Assim, será possível recuperar boa parte desse movimento de vendas com a utilização da variável cartão.

   Registre-se que, a partir do ano de 2019, a aplicação do modelo DEA ofertado deverá substituir a variável de cartão pelo valor registrado nas NFCE

   Observe-se que pode haver a sobreposição de valores nas circunstâncias de vendas pagas com cartão e acobertada por NFE-S. Este problema somente será resolvido quando da substituição total do Cupom Fiscal pela NFCE, prevista para o fim do ano de 2018, pois que será possível a utilização deste último documento como insumo no modelo.

   Diante da extrema dificuldade de identificar e separar essas informações sobrepostas, pressupõe-se que, obedecida a homogeneidade de contribuintes, a eventual proporção

de sobreposição tenha a tendência de ser uniforme no setor (repercutindo pouca variação entre as empresas), porquanto ínsita ao negócio, o que não compromete a análise.

Consigne-se ainda, que aqui não se busca a extrema precisão na apuração do desempenho das eficiências relativas e sim, almeja-se aferir indícios de comportamento tributário reprovável nas extremas discrepância dos resultados.

Ademais, no ideal de manter a homogeneidade entre as DMUs, a aplicação da DEA deve se concentrar em um momento temporal completo (período de um ano completo), com a participação de apenas contribuintes na situação ativa, todos submetidos ao mesmo regime tributário desde o período (ano) anterior ao estudo (*v.g.* o regime Normal). Finalmente, deve-se renunciar a todos os contribuintes que não possuam movimento comercial compatível com a atividade econômica (valores diminutos, própios de empresas em encerramento de atividade).

4.5 Aplicação da DEA para a Seleção de Contribuintes

4.5.1 Resultados

Trazendo para a realidade o método em exposição, extrai-se a princípio os dados necessários:

- *Input*: Arrecadação Realizada e
- *Output*: Valores das Notas de Entrada (compras), Saída (vendas) e Faturamento com Cartão (crédito e débito)

das empresas partícipes do segmento econômico alvo da análise.

Essa extração acontece por roteiros em *SQL* que buscam essas informações diretamente no Banco de Dados já no formato numérico de aplicação do DEA no *Software R*.

Em seguimento, os dados obtidos são submetidos à rotina analítica concebida com o uso do pacote *Benchmarcking* para o cálculo das eficiências relativas de cada contribuinte e a construção dos instrumentos analíticos comparativos úteis à identificação dos comportamentos suspeitos, a saber:

a) Gráfico dos índices de eficiência.

b) Tabela com os resultados coletivos das eficiências relativas encontradas

c) Os valores ideais de evolução das DMU's não eficientes, bem assim de um quadro com a distribuição dos valores de eficiência medidos como descrição geral de todo o setor.

Como produto final para análise dos resultados da aplicação do modelo DEA é gerado um painel interativo em linguagem *HTML*, por meio do pacote *Shiny* (*RStudio*), repercutindo os sobreditos instrumentos de avaliação (gráficos e tabelas).

A título de demonstração por exemplos práticos da aplicação do método descrito, elegeu-se para exame os setores econômicos:

a) **Lojas de Departamentos ou Magazines** (CNAE: G471300100) e

b) **Atacadistas de Produtos de Informática** (CNAE: G465160100)

São usados dados pertinentes ao exercício fiscal do ano de 2017, porquanto era o mais recente e completo exercício disponível quando da conclusão dessa pesquisa.

Da aplicação dos procedimentos descritos aos segmentos escolhidos, obtêm-se os seguintes resultados das eficiências relativas (arrecadação - ICMS):

Exemplo 1 - Lojas de Departamentos ou Magazines (CNAE: G471300100)

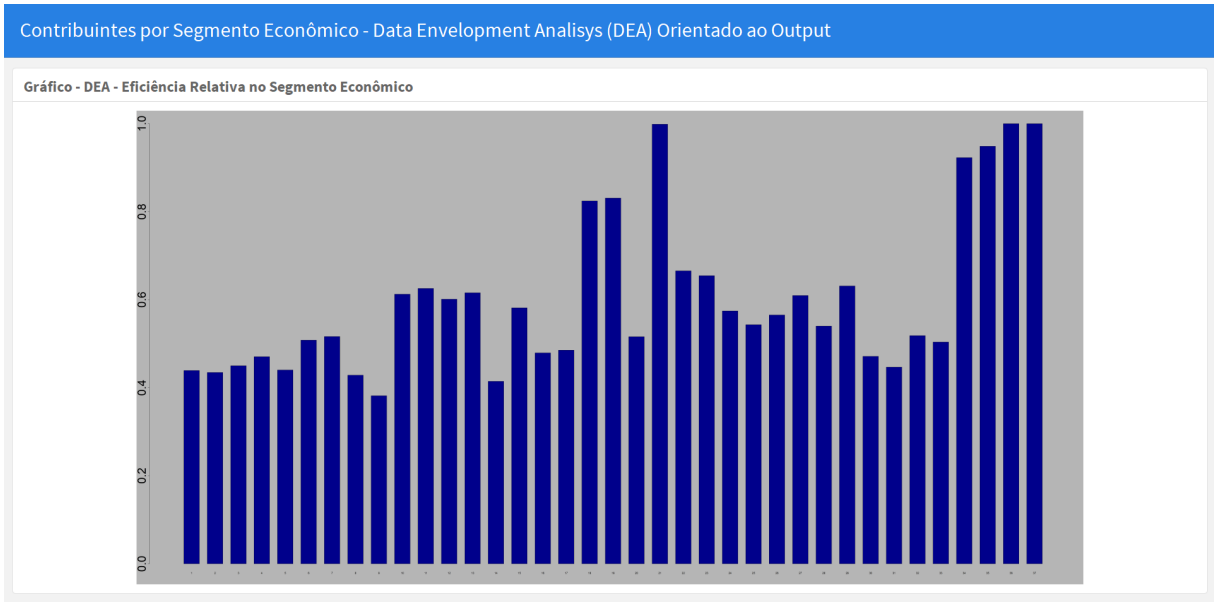

Figura 4.1: DEA - Orientado ao Output - Resultados das Eficiências Relativas - Lojas de Departamentos ou Magazines (CNAE: G471300100).

Além da exposição gráfica dos resultados das eficiências de cada contribuinte do segmento econômico em estudo (Figura 4.1), é ofertada a tabulação dos valores dos insumos

(*inputs* - Entradas, Saídas e Cartão) e produtos (*output* - Arrecadação) utilizados no modelo DEA, bem assim é disposto o cálculo do valor de incremento possível de arrecadação para a melhora do indicador relativo de ineficiência (DIF = Ideal - Real) - Tabela 4.1.

A tabela 4.1 apresenta para cada contribuinte o seu índice alcançado de eficiência relativa no setor e os valores dos insumos (entradas) e do (produto) usados para o cálculo dessa eficiência no modelo DEA (CRS). Do mesmo modo, exibe o resultado da projeção na fronteira de eficiência de cada empresa (DMU), espelhando a expectativa de arrecadação possível (ideal esperado) de se atingir, uma vez corrigida a sua ineficiência.

A partir da apreciação dos índices de eficiências relativas alcançados, é possível estabelecer o critério de limite para a escolha das empresas de interesse, detentoras dos piores resultados e candidatas a uma melhor investigação fiscal. A título de ilustração pode ser estipulado a preferência pelas empresas ranqueadas abaixo de 0,4 do índice de eficiência relativa.

Note-se que o usuário da metodologia deverá saber pautar o seu critério de decisão para a escolha das empresas objeto de fiscalização, na ponderação do resultado de sua ineficiência com a expectativa de retorno financeiro de interesse para a programação fiscal. Melhor dizendo, ainda que o índice de eficiência encontrado para determinado contribuintes esteja assaz discrepante (relativamente baixo), a opção de sua eleição para auditoria deve sopesar a capacidade de incremento de sua arrecadação em termos reais.

Utilizando o parâmetro definido acima de 0,4 identifica-se apenas 1 (um) contribuinte (empresa GM-9) deverá ser submetido a exame para a melhor apreciação de seu comportamento tributário.

Com objetivo de estabelecer um panorama do comportamento integral do setor, o painel igualmente oferece uma sumarização da distribuição de frequências das eficiências relativas apuradas e os seus valores descritivos de posição (média, mediana, primeiro e segundo quartis) - Tabela 4.2.

Para o setor econômico em avaliação, têm-se os seguinte resultados sumarizados:

Exemplo 2 - Atacadistas de Produtos de Informática (CNAE: G465160100)

Como no exemplo anterior, além da disposição gráfica dos índices obtidos, são dispostos na Tabela 4.3: o resultado da eficiência calculada no modelo DEA (CRS), os valores dos insumos e dos produtos utilizados e o resultado da projeção na fronteira de eficiência (DEA) da combinação dos *inputs* e *outputs*.

Aplicando-se igual parâmetro de seleção pelo resultado da eficiência relativa abaixo de 0,4, é possível eleger 14 (quatorze) empresas para auditoria.

Não obstante a medida adotada anteriormente, o uso do critério limite (*v.g* 0,4) de seleção poderá ser otimizado se conjugado o resultado da baixa eficiência relativa e o

Efc Perfil	#	%
$0 \leq E < 0.4$	1	2.7
$0.4 \leq E < 0.5$	11	29.7
$0.5 \leq E < 0.6$	10	27.1
$0.6 \leq E < 0.7$	8	21.6
$0.7 \leq E < 0.8$		0
$0.8 \leq E < 0.9$	2	5.4
$0.9 \leq E < 1$	2	5.4
$E = 1$	3	8.1

Min	0.3800
1º Qrt.	0.4700
Mediana	0.5400
Média	0.6014
3º Qrt.	0.6300
Max	1

Tabela 4.1: Sumário das Eficiências - Lojas de Departamentos ou Magazines (CNAE: G471300100)

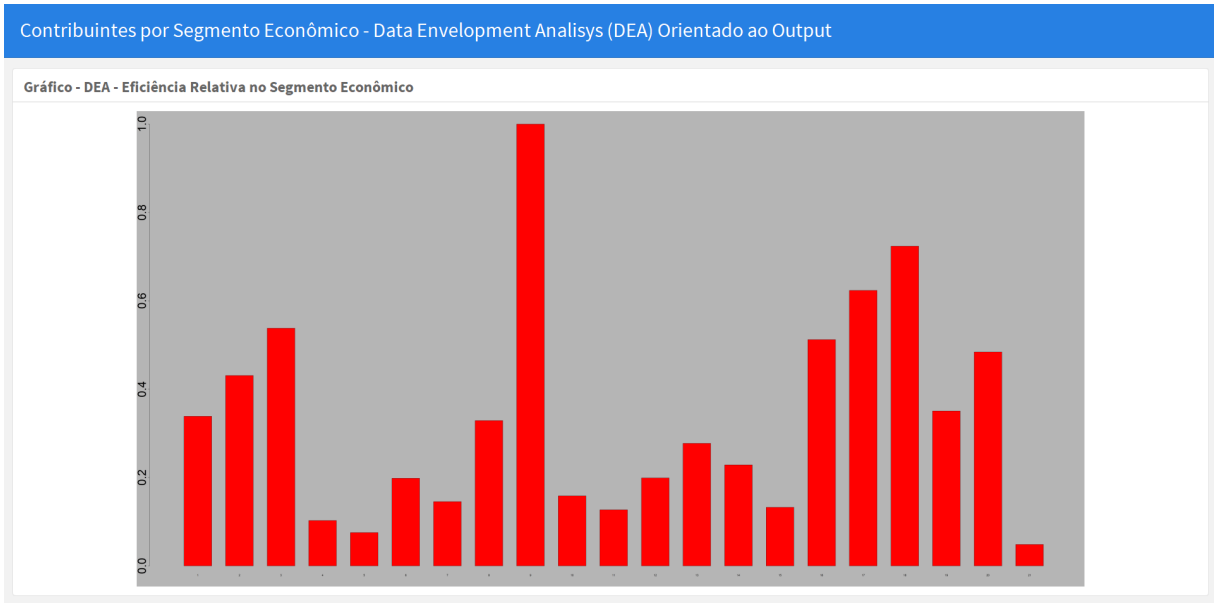

Figura 4.2: DEA - Orientado ao Output - Resultados das Eficiências Relativas - Atacadistas de Produtos de Informática (CNAE: G465160100).

valor do potencial de retorno de arrecadação (DIF), uma vez que é razoável empreender esforços direcionados a um maior incremento de arrecadação.

Para o exemplo, agrega-se (por hipótese) ao processo seletivo uma linha de corte da expectativa de retorno superior ao montante de R$ 1 milhão de arrecadação incremental, ou seja, a diferença de arrecadação possível da DMU alcançar deverá superar esse limite. Nestes termos, têm-se que somente 4 (quatro) empresas atendem a esse critério, a saber: WS-1, WS-5, WS-10 e WS-15.

É perceptível a grande discrepância presente nos resultados das empresas. Para o setor econômico em avaliação, têm-se o resultado gráfico das eficiências relativas alcançadas (Figura 4.2) e a seguinte perspectiva dos resultados sumarizados (Tabela 4.4):

CFDF	DEA	Arrecadação (output)	Ideal	Dif	%	Entradas (input)	Saídas (input)	Cartão (input)
GM-1	0.44	930	2117	1187	128	11970	1083	8600
GM-2	0.43	2086	4799	2713	130	21717	6719	19466
GM-3	0.45	2664	5919	3255	122	28355	3972	24034
GM-4	0.47	1093	2323	1230	113	11379	1136	11622
GM-5	0.44	850	1930	1080	127	9189	944	8362
GM-6	0.51	1467	2887	1420	97	16612	1412	14412
GM-7	0.52	942	1824	882	94	9278	892	7806
GM-8	0.43	726	1694	968	133	9261	860	6882
GM-9	0.38	228	597	369	162	2995	292	2889
GM-10	0.61	258	421	163	63	3227	206	3068
GM-11	0.63	261	417	156	60	3306	204	2999
GM-12	0.60	236	393	157	67	2867	192	3012
GM-13	0.62	243	395	152	63	2784	193	2849
GM-14	0.41	613	1478	865	141	6785	723	6388
GM-15	0.58	572	984	412	72	5570	481	5458
GM-16	0.48	240	501	261	109	2142	345	2034
GM-17	0.49	3470	7150	3680	106	22007	28089	33277
GM-18	0.82	1420	1722	302	21	7794	6766	6972
GM-19	0.83	989	1190	201	20	3663	1949	6123
GM-20	0.52	1562	3027	1465	94	10675	2110	12291
GM-21	1.00	159	159	0	0	1146	358	645
GM-22	0.67	131	197	66	50	1183	313	798
GM-23	0.65	139	212	73	53	1217	373	861
GM-24	0.57	8068	14046	5978	74	44319	7867	57048
GM-25	0.54	2991	5506	2515	84	16947	3080	22423
GM-26	0.57	2549	4509	1960	77	13879	2824	21248
GM-27	0.61	2451	4021	1570	64	12376	2100	19067
GM-28	0.54	1283	2376	1093	85	7312	1665	10120
GM-29	0.63	1752	2776	1024	58	9938	1847	11271
GM-30	0.47	1526	3238	1712	112	10687	2142	13150
GM-31	0.45	2676	5990	3314	124	28355	27375	24244
GM-32	0.52	501	966	465	93	2974	925	6770
GM-33	0.50	305	605	300	98	1863	639	2926
GM-34	0.92	1187	1286	99	8	3959	635	6573
GM-35	0.95	933	983	50	5	3027	512	5646
GM-36	1.00	501	501	0	0	1542	245	2035
GM-37	1.00	4427	4427	0	0	14310	12534	17919

Tabela 4.2: Resultado da Projeção na Fronteira de Eficiência - Lojas de Departamentos ou Magazines (CNAE: G471300100), Valores dos *INPUTS* e *OUTPUT* (R$ x 1000)

Em todos os dois exemplos é possível constatar que poucas empresas qualificam-se com uma eficiência relativa ótima. Contudo, como não é o objetivo do estudo analisar o fenômeno do comportamento econômico do setor, porquanto não existe a necessidade de perscrutar as causas dessas desigualdades, é suficiente a constatação dos piores resultados para possibilitar a seleção dos alvos potenciais de investigação.

Eleitos os contribuintes suspeitos, nos termos do critério de escolha preferidos, aplica-se a sequência de investigação dos dados individuais de cada empresa preferida.

CFDF	DEA	Arrecadação (output)	Ideal	Dif	%	Entradas (input)	Saídas (input)	Cartão (input)
WS-1	0.34	553	1632	1079	195	33726	41076	19
WS-2	0.43	97	225	128	132	4921	5664	0
WS-3	0.54	290	539	249	86	8232	13564	3745
WS-4	0.10	6	58	52	867	964	1469	0
WS-5	0.08	155	2053	1898	1225	25303	51647	0
WS-6	0.20	158	797	639	404	6321	33209	85
WS-7	0.15	44	302	258	586	5950	7603	309
WS-8	0.33	220	669	449	204	5305	21278	21
WS-9	1.00	247	247	0	0	1960	6215	0
WS-10	0.16	495	3121	2626	531	31530	78530	0
WS-11	0.13	34	268	234	688	7114	6732	0
WS-12	0.20	35	176	141	403	1395	6357	0
WS-13	0.28	41	148	107	261	1173	8241	0
WS-14	0.23	48	210	162	338	1666	6877	0
WS-15	0.13	177	1334	1157	654	16397	33555	0
WS-16	0.51	299	584	285	95	11096	14686	0
WS-17	0.62	119	191	72	61	2035	4802	0
WS-18	0.72	198	274	76	38	2171	14273	0
WS-19	0.35	261	744	483	185	14111	18728	0
WS-20	0.48	45	93	48	107	1493	2338	0
WS-21	0.05	40	822	782	1955	6521	35895	0

Tabela 4.3: Resultado da Projeção na Fronteira de Eficiência - Atacadistas de Produtos de Informática (CNAE: G465160100), Valores dos *INPUTS* e *OUTPUT* (R$ x 1000)

Efc Perfil	#	%
$0 \leq E < 0.1$	2	9.5
$0.1 \leq E < 0.2$	5	23.8
$0.2 \leq E < 0.3$	4	19.0
$0.3 \leq E < 0.4$	3	14.3
$0.4 \leq E < 0.5$	2	9.5
$0.5 \leq E < 0.6$	2	9.5
$0.6 \leq E < 0.7$	1	4.8
$0.7 \leq E < 0.8$	1	4.8
$0.8 \leq E < 0.9$	0	0.0
$0.9 \leq E < 1$	0	0.0
$E = 1$	1	4.8

Min	0.0500
1st Qrt.	0.1500
Median	0.2800
Mean	0.3348
3rd Qrt.	0.4800
Max	1.0000

Tabela 4.4: Sumário das Eficiências - Atacadistas de Produtos de Informática (CNAE: G465160100)

Capítulo 5

Análise do Comportamento Temporal do Contribuinte

Esta etapa propõe a aplicação de modelos clássicos de análise de séries temporais com o objetivo de identificar o acontecimento de momentos anômalos (*outliers*) ao longo do comportamento tributário dos contribuintes do ICMS no tempo.

Nesse ideal, faz-se uso de soluções clássicas da análise de séries temporais (correlação comparativa, gráficos boxplot, decomposição de componentes e estimativa Holt-Winter), em busca da percepção desses momentos anômalos nas informações fiscais do contribuinte. Depreende-se que os desvios da regularidade esperada sinalizam possíveis situações de evasão.

Não será um modelo de resultado determinístico, que permitiria a imediata autuação do contribuinte, porquanto. as anomalias detectadas significam suspeitas que podem, ou não, refletir evasão perpetrada. Essas expectativas deverão ser confirmadas em auditoria.

Trata-se de uma solução planejada para dar direcionamento às investigações durante o processo de programação fiscal, uma vez que essas circunstâncias temporais extraordinárias podem conter hipóteses de evasão fiscal.

5.1 Análise de Séries Temporais Fiscais

O objetivo fundamental da análise de séries temporais é poder estudar suas componentes básicas, visando identificar e entender o padrão de seu comportamento ao longo de de períodos de tempo.

Em apertada síntese, o propósito da análise das séries temporais tributárias é estudar a dinâmica, o padrão e a estrutura temporal dos dados fiscais dos sujeitos passivos. A partir da modelagem dos dados das compras e das vendas empreendidas, faz-se um exame dos resultados tributários da atividade econômica desenvolvida pela empresa, extraindo-

se conclusões em termos estatísticos sobre o comportamento da série para, alfim, avaliar o fiel cumprimento da obrigação tributária em razão da conduta esperada.

Ademais, a análise tradicional viabiliza-se por meio da decomposição da série de tempo nos seus principais componentes: tendência e sazonalidade. O componente de tendência explica a conduta de longo prazo da série temporal e a velocidade das mudanças percebidas. Já a componente de sazonalidade exprime as oscilações de crescimento e de queda que se afiguram continuadas em um determinado período (*v.g.* ano ou exercício), possuindo movimentos previsíveis, que ocorrem em intervalos regulares de tempo.

Logo após a seleção de um determinado contribuinte para a análise de seu comportamento tributário, extraem-se os dados fiscais respectivos ao seu movimento comercial (compras/vendas) disponíveis nas bases de dados do órgão da Receita.

Não obstante o prazo decadencial de 5 anos (previsto no art. 173 do CTN [7]) é recomendado realizar a análise de séries temporais com o máximo de informações disponíveis, o que pode significar a utilização de séries com período superior ao de decaimento, porquanto a maior amplitude da série proporciona maior acurácia na interpretação dos padrões nos dados do contribuinte e aumenta a eficiência da aplicação do modelo de alisamento exponencial proposto.

Por ser o foco da auditoria tributária, far-se-á a apreciação do contexto temporal do contribuinte pelo viés da respectiva repercussão fiscal das operações comerciais empreendidas (compras/vendas) desconsiderando-se o seu valor contábil total, o que significa dizer que a análise em tema ocorre sobre o valor do ICMS destacado em cada operação (ou prestação de serviços), montante que é uma proporção do valor econômico do negócio.

Outrossim, o uso preferencial do atributo "valor do ICMS" evita eventuais erros advindos da avaliação de documentos (e escrituração) fiscais que exprimem operações não tributadas ou tributadas por regime específico e diferenciado (como a substituição tributária), não interessantes para o propósito de análise do comportamento tributário regular dos contribuintes.

5.2 *Outliers* em Séries Temporais na Literatura Científica

Levantamentos detalhados dos principais métodos tradicionais para a detecção de desvios anormais em séries temporais podem ser encontrados nos trabalhos de Chandola *et al.* [33], [34] e Gupta *et al.* [35].

A detecção de valores anômalos (*outliers*) é bastante estudada no contexto da análise das séries temporais, em especial sob a perspectiva do tratamento de ruídos subjacentes ao conjunto dos dados e no contexto do alisamento dos períodos temporais para viabilizar

uma regressão e consequentes previsões mais precisas, nesse sentido seguem as lições de Hamilton [36], Brockwell e Davis [37], Cryer e Chan [38] Shumway e Stoffer [39] e Cowpertwait e Metcalfe [40].

Dessarte, segundo Aggarwal [41] um *outlier* é identificado como um ponto cuja remoção resulta em melhor uniformidade da distribuição de frequência para o conjunto dos dados. Pondera o citado autor, que os valores *outliers* em séries temporais podem ser percebidos como desvios da previsão realizada pelo uso dos modelos clássicos de regressão (Aggarwal [42]). Nessa perspectiva, são adotadas soluções para a modelagem de regressão e tratamento de dados extravagantes, como: a Modelagem Auto-regressiva (AR), a Média Móvel Auto-regressiva (ARMA) e a Auto-regressiva Média Móvel Integrada (ARIMA). A robustez do processo de previsão adotado é destaque para a condução de uma melhor detecção de valores extraordinários.

Quando o problema encontra-se em um contexto de um elevado número de dados, alguns métodos foram construídos para abreviar o tempo de modelagem da regressão como explicitado em Jiang *et al.* [43] e Papadimitriou *et al.* [44].

A detecção de anomalias em séries temporais pode acontecer: a) pela identificação de pontos crítico destoantes ou b) em termos da mudança de estados da série, consoante ensina Chandola *et al.* [33] e Aggarwal [42]. O primeiro caso está relacionado à análise de valores extremos, sendo o último mais sutil, posto que requer uma análise cuidadosa das regularidades na série em diferentes janelas de tempo dos dados. Para a identificação dos pontos críticos, a magnitude dos desvios pontuais percebidos é mais interessante que o formato de distribuição da série ao longo dos períodos.

Uma variedade de métodos podem ser usados para identificar formatos incomuns nas séries temporais. Os métodos mais comuns fazem o uso das medidas de distâncias em janelas fixas do tempo, como mostrado por Keogg [45]. Sendo que essas soluções podem ser aceleradas com o uso da aproximação por agregados simbólicos.

Soluções de modelos supervisionados e semi-supervisionados podem ser usados para realizar análises discriminatórias das diferentes formas de desvios e valores atípicos presentes na configuração da série temporal, consoante explicam: Aggarwal [41], Jeong [46], Mueen [47], Ye e Keogh [48]. Nesses cenários, presume-se a disponibilidade de séries temporais ideais (normais) paradigmas para a comparação.

Os modelos para identificar anomalias em sequências discretas são explorados e discutidos por Aggaward [42]. Ademais, uma pesquisa minuciosa para o caso discreto pode ser encontrada em Chandola *et al.* [34] e sob a perspectiva integrada, no contexto dos dados discretos e contínuos, pode ser explorada em Gupta *et al.* [35].

Outro método conhecido para a detecção de momentos anômalos em séries de tempo é o uso da função de distância, como o distanciamento dinâmico do tempo pesquisado em

Jeong, Jeong e Omitaomu [46] e Aggaward [42]. O uso de funções de distância possibilita o desenvolvimento de técnicas de classificação baseadas em proximidade e definição de classes.

A eficácia da detecção de anomalias pode ser aprimorada pelo uso da clusterização. Combinações de classificação e modelos de agrupamento são proposto em Al-Kaleb *et al.* [49] e Masud *et al.* [50] para distinguir entre classes normais, classes novas, raras e recorrentes.

Métodos de detecção de anomalia em séries multivariadas são discutidas por Cheng *et al.* [51], Tsay [52] e também por Baragona e Batagglia [53]. Transformações discretas podem ser usadas de forma local para realizar a detecção de anomalia diretamente na representação discreta, como anuncia Keog [54].

A detecção de *outlier* por ultrapassagem de limites, consoante adotado na metodologia proposta, foi estudada extensivamente no contexto dos dados de sensores por Budaski e Deligiannakis [55], Branch [56], Franke e Gertz [57], Yang, Nirvana e Havinga [58]. As análises de correição de fluxo em sensores são as aplicações mais comuns de identificação de anomalias em dados temporais.

Os dados temporais multivariados também podem ser representados na perspectiva de trajetórias para a detecção de valores anômalos, examinando padrões de evolução das informações, como discutido em Aggarwal [42]. Em tais casos, a trajetória pode ser tratada como dados temporais bivariados e a detecção de desvio, com o uso da predição, pode ser aplicada a essa representação.

Métodos que fazem uso de testes estatísticos para a detecção de mudanças, como a KL-distância, Wilcoxon, Kolmogorov-Smirnoff e critérios de log-verossimilhança também podem ser encontrados em Kuncheva [59] e Song [60]. Esses momentos de mudança são utilizados para o encontro de tempos *outliers* no fluxos de dados.

Especialmente, a pesquisa de Szmit e Szmit [61] sugere o uso do método de Holt-Winter de suavização exponencial para a detecção de anomalias em sequências temporais. Outra interessante aplicação da estratégia de detecção de *outliers* utilizando o método Holt-Winter pode ser encontrado no trabalho de Salem, Liu e Mehaoua [62] que apresenta um modelo de controle de defeitos em sensores médicos a partir das anormalidade detectadas nas séries temporais das medidas (dados médicos) coletada dos pacientes.

5.3 Metodologia de Análise de *Outliers* em Séries Temporais Fiscais Aplicada

5.3.1 Primeira Análise – Cotejo: Escrituração x Realidade Documental

A primeira avaliação das séries temporais do contribuinte em investigação, dar-se-á com os dados originais das séries de entrada e de saída, sobrepostas conjuntamente de maneira comparativa, onde é possível conjugar a realidade do ICMS, presente nas NFEs, com os valores de declaração consignados no LFE. Essa comparação acontecerá para as entradas e também para as saídas (NFEs vs LFE), destacando-se as eventuais divergências obtidas.

Para efeitos de interesse da auditoria devem ser observados:

a) Na comparação das entradas, especial atenção deve ser dada às diferenças NFEs menos LFE que se afigurem negativas, porquanto podem representar a escrituração de crédito incompossível, ou seja, sem a contrapartida de lastro por documento fiscal válido.

b) No confronto das saídas, devem ser observadas as diferenças *LFE - NFE* que resultem em valor menor que zero, posto que demonstram lançamento contábil de débito do imposto reduzido do valor real a recolher naquele período de tempo.

5.3.2 Segunda Análise – *Box-Plot* e a Presença de *Outliers* Mensais

Esse exame corresponde ao resultado da agregação dos valores das séries por mês, sumarizando-se o resultado desses valores em um diagrama de *box-plot* que tem por predicado apresentar graficamente a distribuição dos dados em torno da média, mediana e momentos interquartis. Essa diagramação permite a fácil identificação dos valores *outliers*, muito distintos na distribuição, e dignos de uma melhor apreciação das razões de sua ocorrência.

O *boxplot* - também chamado de gráfico de caixa - é uma solução gráfica empregada na avaliação da distribuição empírica do dados. Esse gráfico é construído pelo primeiro e terceiro quartil e pela mediana. As hastes inferiores e superiores se alongam, respectivamente:

a) sendo Q_1 o primeiro quartil e Q_3 o terceiro quartil da distribuição, a distância interquartil é dada por $IQR(\textit{interquartile range}) = (Q_3 - Q_1)$.

b) Limite inferior: $Q_1 - 1.5 * IQR$ e

c) Limite superior: $Q_3 + 1.5 * IQR$.

Os pontos fora desses limites são considerados valores discrepantes (*outliers*). Usa-se comumente como paradigma para a definição dos pontos *outliers*:

- $Outliers_{(superiores)} = 3 * IQR$, ou mais acima do terceiro quartil e
- $Outliers_{(inferiores)} = -3 * IQR$, ou menos abaixo do primeiro quartil.

Por essa solução gráfica é ofertada para avaliação os mêses de acontecimento de valores extremos e sugestivos da ocorrência de irregularidades. A percepção dessa condição possui arrimo na rara probabilidade de sua ocorrência definida pelo teorema da desigualdade de Cherbychev – que oferece uma cota de probabilidade para um valor de uma variável aleatória, com variância finita, estar a uma distância da sua esperança matemática. Assim, os valores que excederem a cota de três (ou dois) desvios padrões em relação à média (equivalente aos pontos fora dos "bigodes" da figura *box-plot*), são ocorrências candidatas à suspeição.

5.3.3 Terceira Análise - Decomposição em Componentes de Tendência e Sazonalidade

Nesse momento, as séries temporais de entradas/saídas serão decompostas para a explicação dos seus componentes de tendência e de sazonalidade.

Esses dois componentes são essenciais para a compreensão da atividade negocial da empresa e servem de referência para o entendimento da função econômica subjacente à atividade do contribuinte, bem assim a comparação de seu resultado em paralelo ao momento conjuntural da economia e do seu segmento empresarial.

Para empreender a decomposição das séries temporais fiscais, pressupõe-se o **modelo aditivo** de composição das séries, porquanto assume-se (por razoável) que o fenômeno econômico do comércio de mercadorias e, ou, a prestação de serviços, possui tendência que se movimenta exponencialmente, contudo **mantendo uma sazonalidade aditiva** que reflete a variação da demanda ínsita aos períodos do ano (12 meses), sem expansão (ou contração) exponencial no seu comportamento.

No modelo aditivo o valor da série X_{t+1} será o resultado da soma dos valores das componentes (que apresentam a mesma unidade da variável):

$$X_{t+1} = T + S + C + A \tag{5.1}$$

sendo:

T a tendência,
S a sazonalidade,
C o componente cíclico e
A o componente aleatório.

5.3.4 Quarta Análise - Estimativa Holt-Winter e a Detecção de Momentos Anômalos

Para se detectar um comportamento que fuja a um dado padrão histórico, faz-se necessária a construção de uma *baseline* para a série analisada. A estimativa de Holt-Winters (Suavização Exponencial Tripla) das séries temporais mostra-se como o modelo ideal para a construção da linha base (*baseline*) para a pesquisa de momentos temporais *outliers*.

O método de Holt-Winter para efeitos sazonais aditivos será utilizado na modelagem comparativa das séries fiscais, pressupondo-se que a amplitude do ciclo sazonal é constante e independente do nível local da série. Melhor dizendo, a variação periódica sazonal da série temporal possui comportamento estatístico que independe da taxa de crescimento (positiva ou negativa) da sua tendência.

A predição ou estimativa de Holt-Winters distribui uma série temporal em três componentes superpostos:

a) Um termo que significa a periodicidade da série.

b) Outro termo que aponta a tendência de crescimento da série.

c) E um terceiro termo que expressa a parte residual da série, resultante da dissociação das duas partes anteriores.

Todos esses três termos são tratados de forma individualizada por meio de uma suavização exponencial por média móvel ponderada (*EWMA - Exponentialy Weighted Moving Average*). O *EWMA* é um estimador aplicado às séries temporais que estabelece uma ponderação entre o valor atual da série e a estimativa anterior. Isto significa que três coeficientes, α, β e γ, devem ser atribuídos, um para cada EWMA.

A expressão para o modelo Holt-Winter, usando método aditivo, é dada por:

$$\hat{x}_{t+n} = a_t + nb_t + c_{t+n-m} \tag{5.2}$$

$$a_t = \alpha(x_t - c_{t-m}) + (1-\alpha)(a_{t-1} + b_{t-1}) \tag{5.3}$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \tag{5.4}$$

$$c_t = \gamma(x_t - a_t) + (1-\gamma)c_{t-m} \tag{5.5}$$

sendo:

$0 \leq \alpha, \beta, \gamma \leq 1$ e $a_0 = x_0$

onde:

$\hat{x}$ *é a estimação*,
x_t *é o valor corrente*,
a_t *corresponde à componente de nível*,
b_t *corresponde à componente de tendência*,
c_t *corresponde à componente periódica – sazonalidade*,
n *corresponde ao tempo de estimação* e
m *é o tamanho do período de sazonalidade.*

Dessarte, será atribuído um EWMA para cada termo e a soma dos três resulta na estimativa de Holt-Winters.

Finalmente, confrontar-se-á as séries temporais com seus correspondentes estimadores Holt-Winters, indicando as diferenças negativas (nos dados: NFEs e LFE):

a) $Previsão_{Holt-Winter} - REAL$ para as entradas e

b) $REAL - Previsão_{Holt-Winter}$ para as saídas.

Sob a perspectiva de 2 (dois) ou 3 (três) desvios-padrão, podem ser definidos os critérios para a decisão de classificação dos momentos temporais anômalos.

5.3.5 Dados Temporais Adotados

Os dados de interesse para as análises das Séries Temporais Tributárias são:

a) O total do **ICMS obtido pelas compras** (crédito de entradas), definido mensalmente, que representa o montante dos créditos mensais passíveis de apropriação para compensação. Esses valores estão representados em duas séries temporais:

   - Série da soma mensal do valor do **ICMS destacado nas NFEs** emitidas pelos fornecedores (somente as válidas e não canceladas - situação 100), consoante as compras realizadas (entradas) para formação de estoques, insumos ou como investimento em ativo imobilizado (v.g. máquinas e equipamentos).

     Não obstante, somente os créditos relativos às compras de produtos para a formação de estoques, aquisição de insumos de produção e ativo imobilizado (esse último de apropriação diferenciada em proporção), não recomenda-se *ab initio* a segregação de outras aquisições com ICMS não aproveitáveis (como, por exemplo, uso e consumo), porquanto não representam valores representativos, e se representarem, serão facilmente identificáveis como acontecimentos anômalos na série.

- Série do total mensal do valor do **ICMS escriturado** como crédito pelo contribuinte no **LFE**, que, por regra, devem repercutir somente os créditos passíveis de apropriação correspondentes às respectivas NFEs recebidas em operações onde é possível o aproveitamento do crédito do imposto.

b) O total do **ICMS resultante das vendas** (débito de saída) definido mensalmente, que revela os débitos mensais do imposto proveniente das operações/prestações de bens ou serviços tributáveis, objetos da atividade negocial do contribuinte. Esses valores devem ser recolhidos ou compensados com os fortuitos créditos existentes. Tal informação está consignada em duas séries:

- Série do total mensal do **ICMS destacado nas NFEs** (somente as válidas e não canceladas - situação 100) emitidas pela empresa para seus clientes.
- Série da quantidade mensal do valor dos débitos do **ICMS lançado no LFE** correspondentes às saídas tributadas promovidas pelo contribuinte.

As referenciadas informações estão armazenadas no banco de dados *ORACLE* da GEPRO e serão recuperadas por meio de *queries*, escritas em linguagem *SQL*, e aplicadas diretamente do *Software R* com a utilização do pacote *RODBC*.

Colhidos os dados, ordenados em formato mês e ano a que correspondem, constroem-se as séries temporais pertinentes às entradas (pelo LFE e pelas NFEs) e, igualmente, às saídas (LFE e NFEs).

Usa-se na construção dos gráficos das Séries de Tempo e nas previsões do seu comportamento os seguintes pacotes do *Software R*: *dygraphs*[18], *MTS*[19] e *tseries*[20].

5.4 Aplicação nas Séries Temporais Fiscais

5.4.1 Resultados

Em razão do Sigilo Fiscal obrigatório, não será divulgada a identificação do contribuinte em exemplo.

[18] O pacote *dygraphs* é uma interface *R* para a biblioteca de gráficos *JavaScript* dos *dygraphs*. Ele posibilita soluções sofisticadas para traçar dados de séries temporais no *Software R*. Informações: *rstudio.github.io/dygraphs/*

[19] *Multivariate Time Series (MTS)* é um pacote geral para análise de séries temporais multivariadas e para realizar estimativas em modelos de volatilidade multivariada. informações: *cran.r-project.org/web/packages/MTS/MTS.pdf*.

[20] *tseries* é um pacote do *Software R* para a análise de séries temporais fianceiras. Informações: *cran.r-project.org/web/packages/tseries/tseries.pdf*

Obtidas as informações fiscais do contribuinte escolhido para a análise, constroem-se os painéis com os gráficos das séries temporais fiscais para as Entradas e para as Saídas, a saber:

a) Os primeiros mostrarão os valores mensais do imposto destacado em NFEs e a sua correspondente escrituração no LFE, apontando as eventuais diferenças entre os valores sobrepostos. No painel das Entradas as diferenças plotadas são calculadas: $ICMS_{NFE} - ICMS_{LFE}$ e, no das Saídas: $ICMS_{LFE} - ICMS_{NFE}$.

b) Nessa fase, também são analisados os componentes de tendência e sazonalidade das séries temporais, expostos graficamente em sobreposição aos valores originais da série como forma de proporcionar o entendimento das resultantes fiscais originadas pelos movimentos de compras e de vendas tributadas.

c) Ademais, apresenta-se a descrição do comportamento mensal ao longo de toda as séries temporais. Essa análise descritiva é feita pelo painel de *Box-Plots* gerado para as séries de Entradas e Saídas (NFEs e LFE).

d) Igualmente, são produzidos os desenhos das séries temporais que correspondem ao ICMS presente mensalmente nas NFEs e no LFE, em colação justaposta com o valor (estimadores) previsto pelo método Holt-Winters (linha vermelha sólida), permitindo a construção de um paralelo de distanciamento entre os valores reais e os estimados pela suavização da série. Ademais, será possível apreciar o afastamento entre as séries por seu volume financeiro e por sua dispersão em termos de desvios padrão, tracejando-se (linhas tracejadas em vermelho) os limites de um e dois desvios-padrão para as diferenças encontradas (previsto - real).

Observe-se que todos os gráficos gerados possuem a funcionalidade de sobre-exposição dos dados relativos aos períodos selecionados pelo *mouse*, bem assim rolagem de ampliação ou compressão de períodos.

Isto posto, para o esclarecimento do modelo, adota-se um exemplo real que oferece os resultados e considerações em seguimento.

EXEMPLO: Primeira Análise - Séries Temporais Fiscais de ENTRADAS:

Aplicando-se a metodologia exposta, recorre-se às bases fiscais da Nota Fiscal Eletrônica - NFE, porquanto representa o documento fiscal próprio para certificar a realidade da operação de comercialização ou prestação, para obter o ICMS destacado nesses documentos, agregado mensalmente ao longo de todo o período disponível. De igual forma, usa-se as informações, mensais, escrituradas no Livro Fiscal Eletrônico – LFE pelo contribuinte e concernentes aos débitos apurados e aos créditos passíveis de apropriação e compensação.

Em seguimento, confronta-se a realidade disponível nas NFE com os respectivos valores escriturados no LFE.

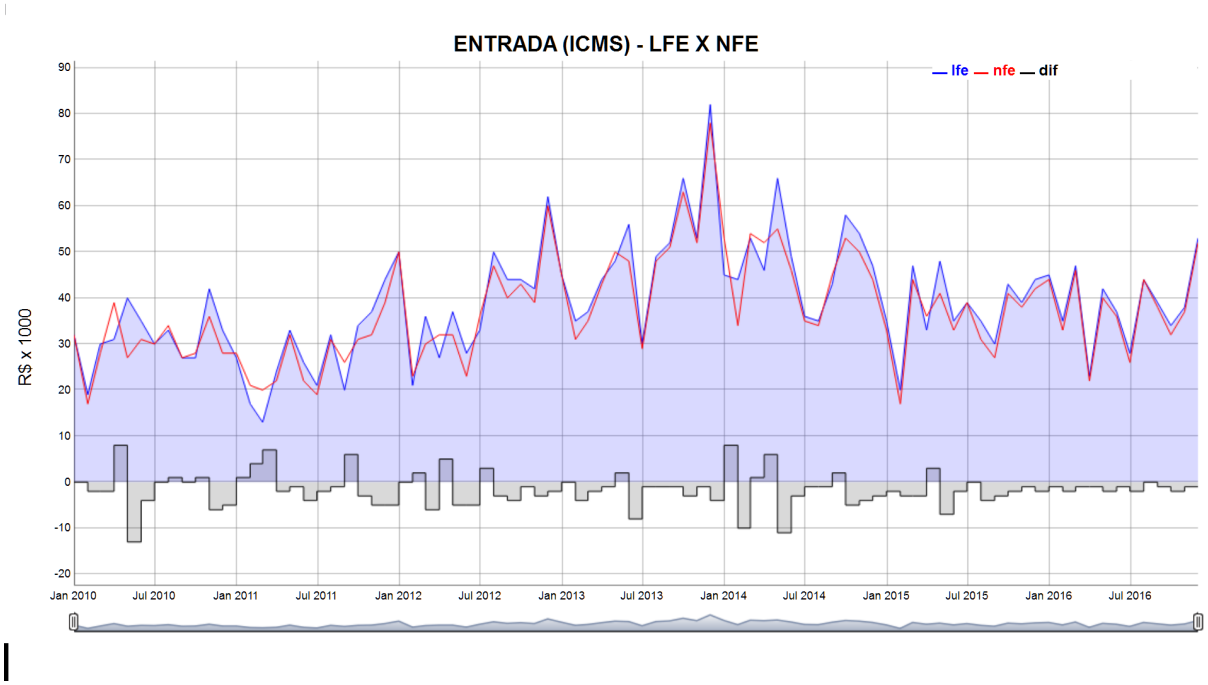

Figura 5.1: Séries Temporais Fiscais Comparadas - NFEs x LFE - ENTRADAS.

Tratando-se das séries temporais relativas ao crédito do ICMS pelas entradas acontecidas (Figura 5.1), interessa ao fisco a divergência negativa do cotejo do ICMS das NFEs menos o ICMS registrado como crédito no LFE, porquanto pressupõe o registro indevido de créditos além do sustentável pela documentação fiscal própria (NFEs) - circunstância presente nos dados informados pelo contribuinte em questão (zona cinzenta abaixo de zero).

EXEMPLO: Primeira Análise - Séries Temporais Fiscais de SAÍDAS:

De outra forma, contemplando-se as séries temporais de saída (Figura 5.2), importa considerar a diferença negativa resultado da comparação do ICMS escriturado no LFE menos o ICMS destacado nas NFEs (débitos), porquanto essa desconformidade indica possível sonegação, uma vez que nem todos os débitos de suas operações tributáveis foram registrados ou então foram registrados com valor diminuído. Essa condição de atenção não é verificada no exemplo, como é possível observar no gráfico.

Dado o perfil varejista da empresa escolhida para exemplo, é possível que o contribuinte opere com outros documentos fiscais de saída, como o Cupom Fiscal (não disponível sem uma leitura individual da memória fiscal de cada Equipamento Emissor de Cupom Fiscal -

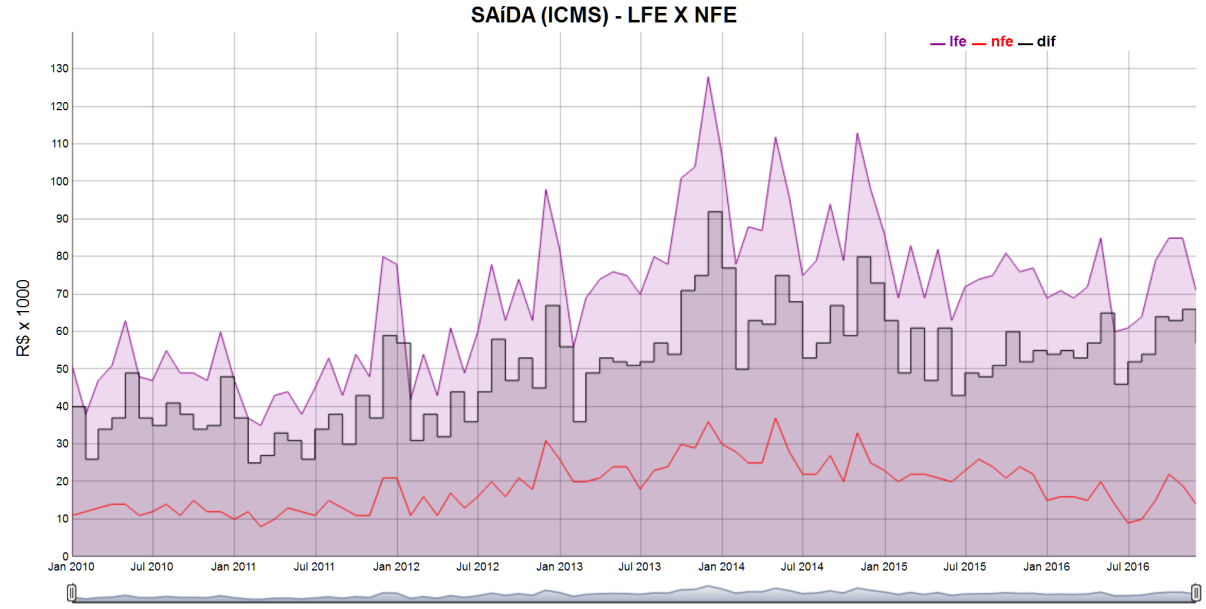

Figura 5.2: Séries Temporais Fiscais Comparadas - NFEs x LFE - SAÍDAS.

ECF) ou como a NFCE, e que estes tenham sido escriturados em suplementação às NFEs – o que explicaria a variação de débitos escriturados no LFE superiores aos presentes nas NFEs emitidas.

Em condições como a identificada pelo exemplo, a melhor maneira de verificar a condição de regularidade fiscal do contribuinte é cruzar as saídas escrituradas – por seu valor contábil - com as receitas advindas dos diversos meios de pagamentos, em especial os valores de operações faturadas com cartão de crédito ou débito e a movimentação bancária das contas da empresa. Essas conciliações, inobstante importantes, fogem ao escopo da metodologia aqui proposta.

EXEMPLO: Segunda Análise - Gráficos Box-Plots de ENTRADA e SAÍDA:

Gráficos *Box-Plots* com valores agregados em distribuição de frequência mensal espelham o acontecimento de eventuais valores dispersos nos períodos (Figura 5.3 e Figura 5.4). A partir de suas análises é possível eleger os meses onde há fortes indícios de irregularidades.

Registre-se que, para esses gráficos, somente os dados respeitantes ao período submetido à possibilidade de auditoria (5 anos) participarão da construção dessa visualização analítica de distribuição dos dados mensais das séries. Esse limite é imposto para evitar a obtenção de um valor anormal inalcançável em razão do instituto jurídico da decadência tributária.

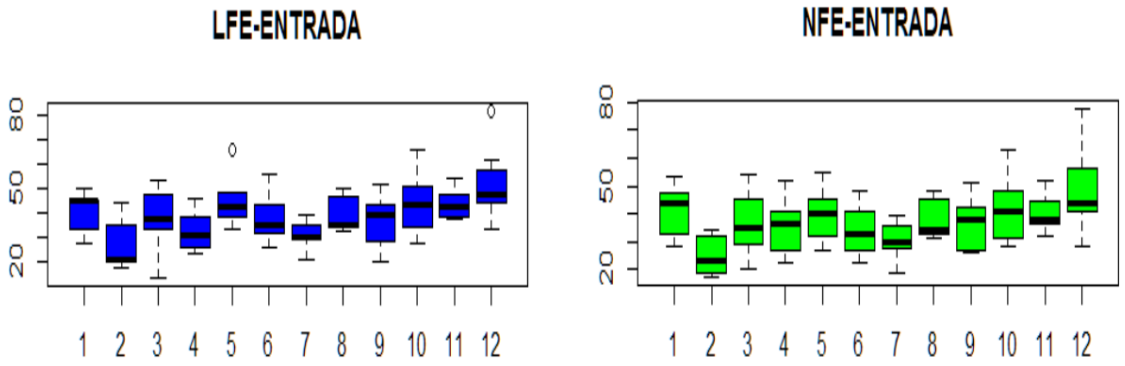

Figura 5.3: *Box-Plots* (meses) - Séries Temporais Fiscais - ENTRADAS.

Como é permitido observar nos valores de ICMS das entradas (Figura 5.3) para as NFEs e para o LFE, há a presença de valores extremos dignos de investigação na escrituração dos meses de maio e dezembro do livro eletrônico (LFE).

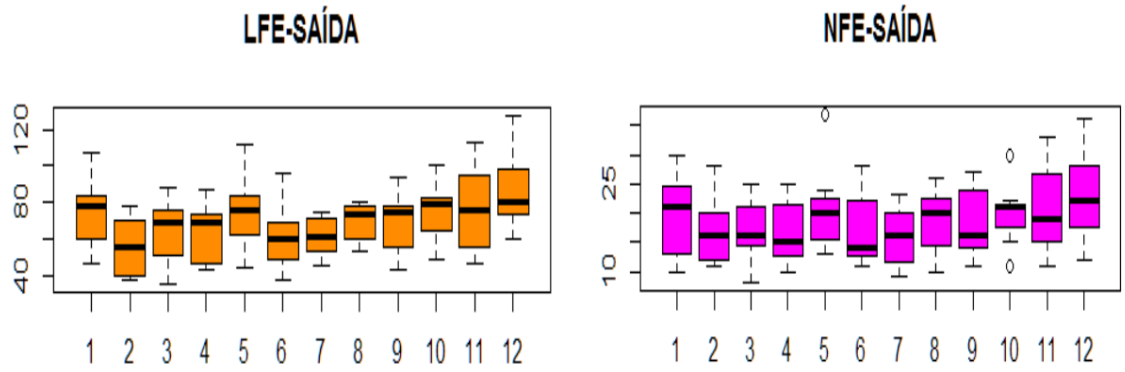

Figura 5.4: *Box-Plots* (meses) - Séries Temporais Fiscais - SAÍDAS.

Na distribuição das saídas há dispersão nas NFEs de vendas nos meses de maio e outubro (Figura 5.4).

EXEMPLO: Terceira Análise - Gráficos Tendência e Sazonalidade das ENTRADAS:

Nesse ponto, são analisados os componentes de tendência e sazonalidade das séries temporais, expostos graficamente em sobreposição aos valores originais da série como forma de proporcionar o entendimento das resultantes fiscais originadas pelos movimentos de compras e de vendas tributadas.

Respeitante às resultantes da decomposição das séries de entrada, LFE (Figura 5.5) e NFE (Figura 5.6), é possível dizer:

- TENDÊNCIA - É fácil verificar uma mudança de orientação de tendência, acentuada no ano de 2014, condição compatível com o cenário econômico de retração de demanda, que determina a diminuição sobre a formação dos estoques.
- SAZONALIDADE – Percepção de uma movimentação compatível com o setor econômico de varejo, com destaque de alta para o mês de maio e para o final de ano - períodos que exigem estoques compatíveis com a demanda do dia das mães e das festas de fim de ano.

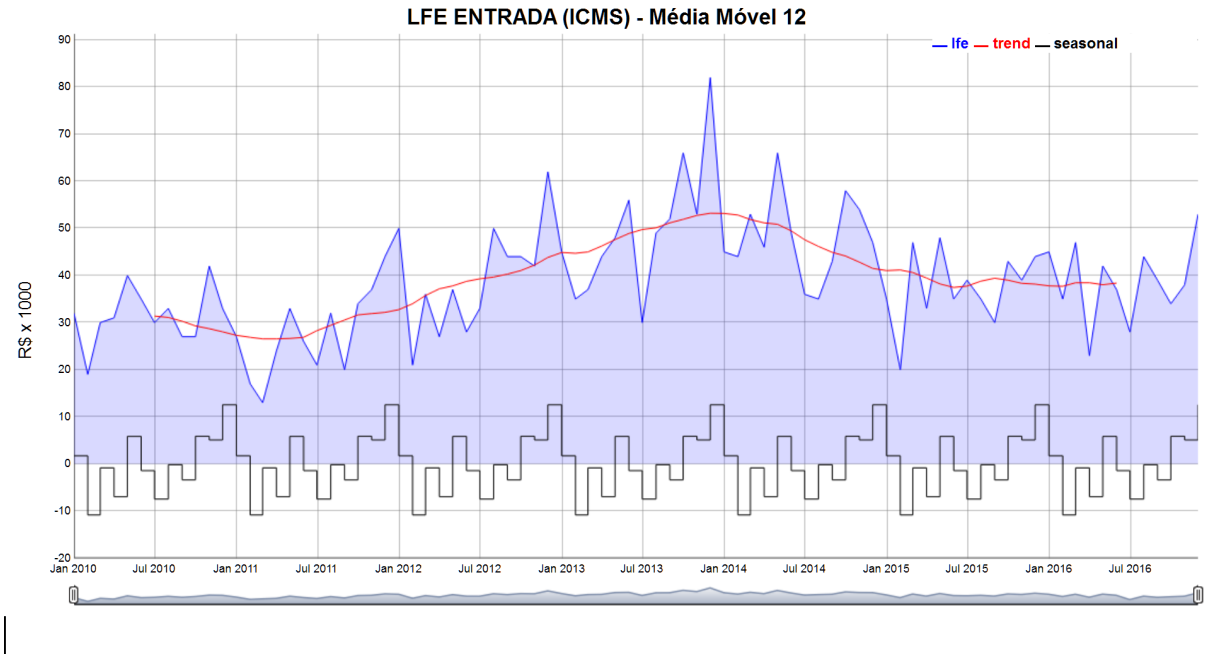

Figura 5.5: Tendência e Sazonalidade - LFE - ENTRADAS.

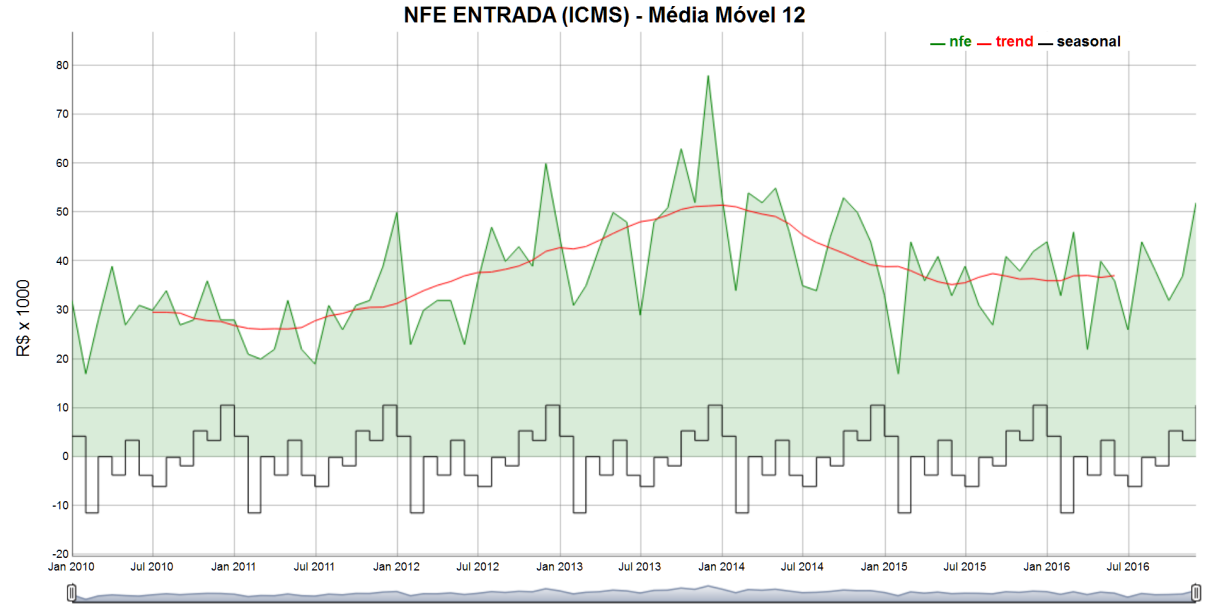

Figura 5.6: Tendência e Sazonalidade - NFEs - ENTRADAS.

As duas séries de entrada mostram semelhante movimento de sazonalidade e tendência.

Nada obstante, importa ressaltar que a coincidência dos movimentos extremos nas séries da NFE e do LFE não permite concluir pela regularidade desses momentos, posto que é possível o acontecimento de uma operação simulada para a obtenção de créditos do ICMS.

EXEMPLO: Terceira Análise - Gráficos Tendência e Sazonalidade das SAÍDAS:

Os gráficos em seguimento mostram a decomposição das séries de saída, LFE (Figura 5.7) e NFE (Figura 5.8), e confirmam uma relativa compatibilidade destas com os componentes das séries de entradas, apreciados anteriormente - oferecendo a mesma inflexão da tendência coordenada pela conjuntura da economia nacional.

- TENDÊNCIA – após um expressivo crescimento, a partir do ano de 2014, as vendas iniciam um declínio consistente.
- SAZONALIDADE – vendas altas realizadas nos meses de maio e, especialmente, em dezembro. Baixas acentuadas nos meses de fevereiro e junho.

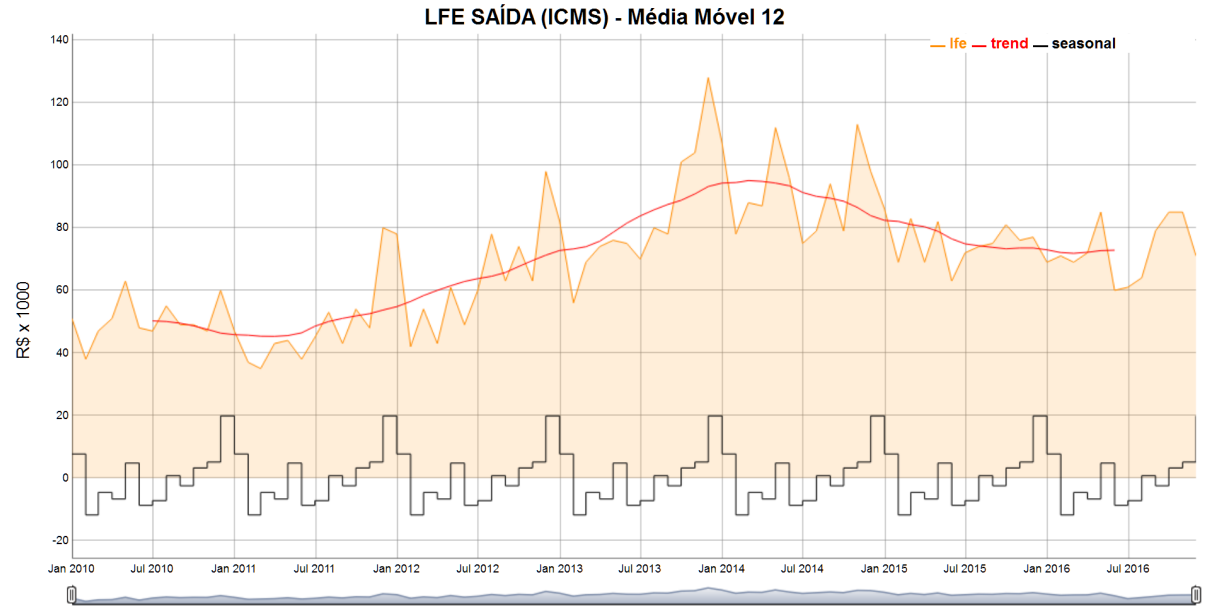

Figura 5.7: Tendência e Sazonalidade - LFE - SAÍDAS.

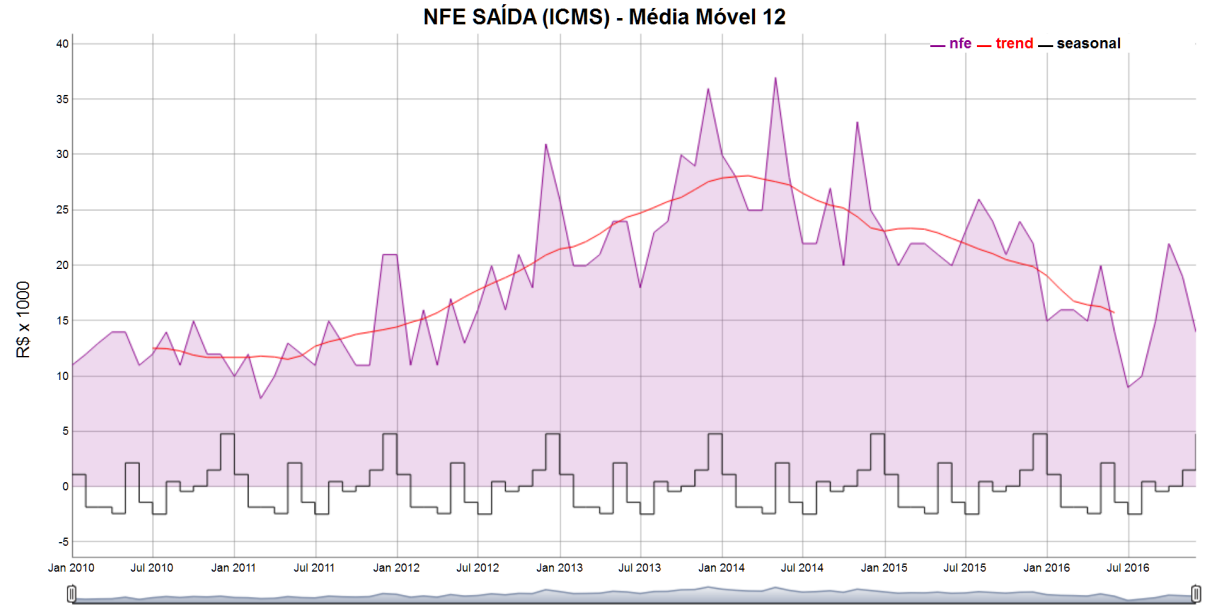

Figura 5.8: Tendência e Sazonalidade - NFEs - SAÍDAS.

EXEMPLO: Quarta Análise - estimativa Holt-Winter ENTRADAS:

Realizada a aplicação do método Holt-Wilter para alisamento e inferência das séries temporais obtêm-se, para cada série objeto de análise, a estimativa do comportamento regular

esperado proposto graficamente em sobreposição aos valores reais, assim como a diferença ponderada em desvios-padrão da variação entre o valor real e o valor estimado pelo modelo - Entradas-LFE (Figura 5.9) e Entradas-NFE (Figura 5.10).

É apresentada a composição gráfica dos estimadores Holt-Winter justapostos aos valores originais da série, permitindo a construção de um paralelo de distanciamento entre os valores reais e os estimados pela suavização da série. Ademais, será possível apreciar o afastamento entre as séries por seu volume financeiro e por sua dispersão em termos de desvios padrão.

As linhas tracejadas em vermelho representam os limites relativos a um e dois desvios-padrão das diferenças negativas e pedem atenção para os valores muito dispersos do esperado, que podem significar meses onde ocorra a apropriação imprópria de créditos. Para as Entradas a diferença relativa é calculada:

$$DifRel_{Entradas} = Previsão_{Holt-Winter} - Real$$

Nos gráficos das Entradas (Figura 5.9 e Figura 5.10) chama a atenção os valores de compras muito dispersos do esperado pelo modelo Holt-Winter, condição que pode significar a sobre-elevação indevida dos créditos fiscais - escrituração indevida de créditos inexistentes ou impróprios.

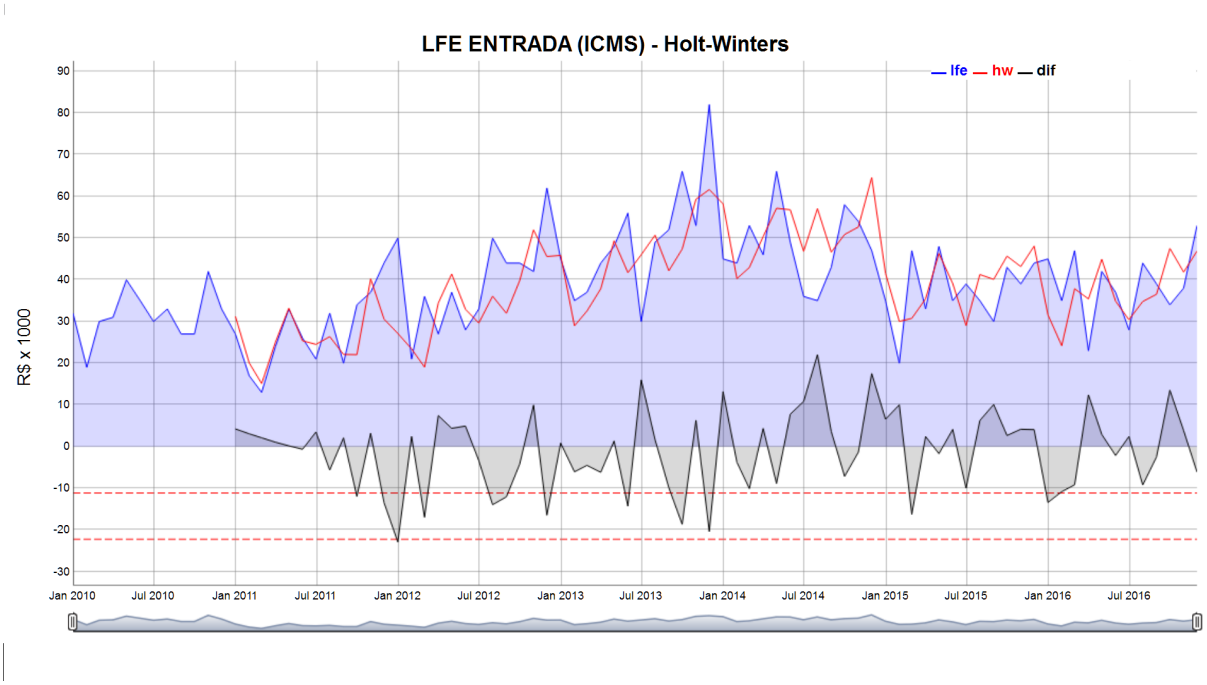

Figura 5.9: Holt-Winter - LFE - ENTRADAS.

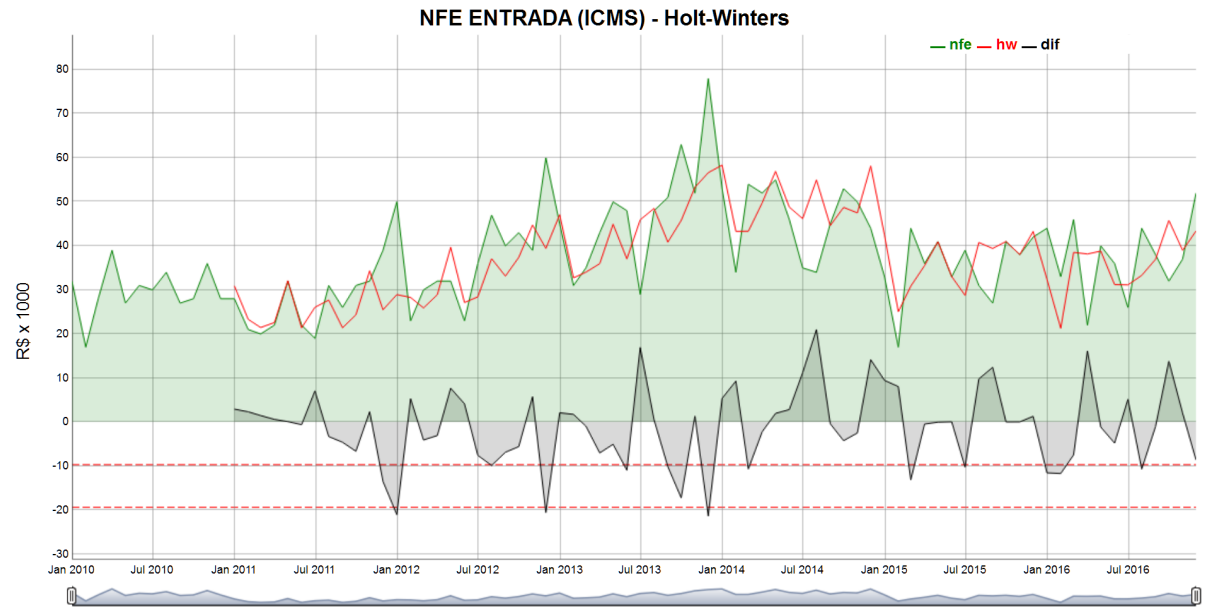

Figura 5.10: Holt-Winter - NFEs - ENTRADAS.

Caberá ao responsável pelo monitoramento decidir quais os limites de dispersão que elegerá para determinar se o período (mês) será, ou não, considerado como anômalo (*outlier*) e, assim, selecionado para uma investigação mais precisa e detalhada sobre os documentos e lançamentos nele emitidos e/ou registrados.

EXEMPLO: Quarta Análise - estimativa Holt-Winter SAÍDAS:

Igualmente são avaliadas as séries temporais pertinentes às operações de saída para apuração das divergências relativas (em desvios-padrão) entre os dados previstos pelo modelo e os dados existentes - Saídas-LFE (Figura 5.11) e Saídas-NFE (Figura 5.12).

As linhas de fronteira representam os limites relativos a um e dois desvios-padrão das diferenças, reivindicando cuidado para os valores de vendas tributáveis muito dispersos do esperado pelo modelo.

Para as Saídas a diferença relativa é calculada:

$$DifRel_{Saídas} = Real - Previsão_{Holt-Winter}$$

Destacam-se os valores de vendas muito apartados do previsto pelo método Holt-Winter, circunstância que pode indicar a omissão de saídas efetivamente realizadas.

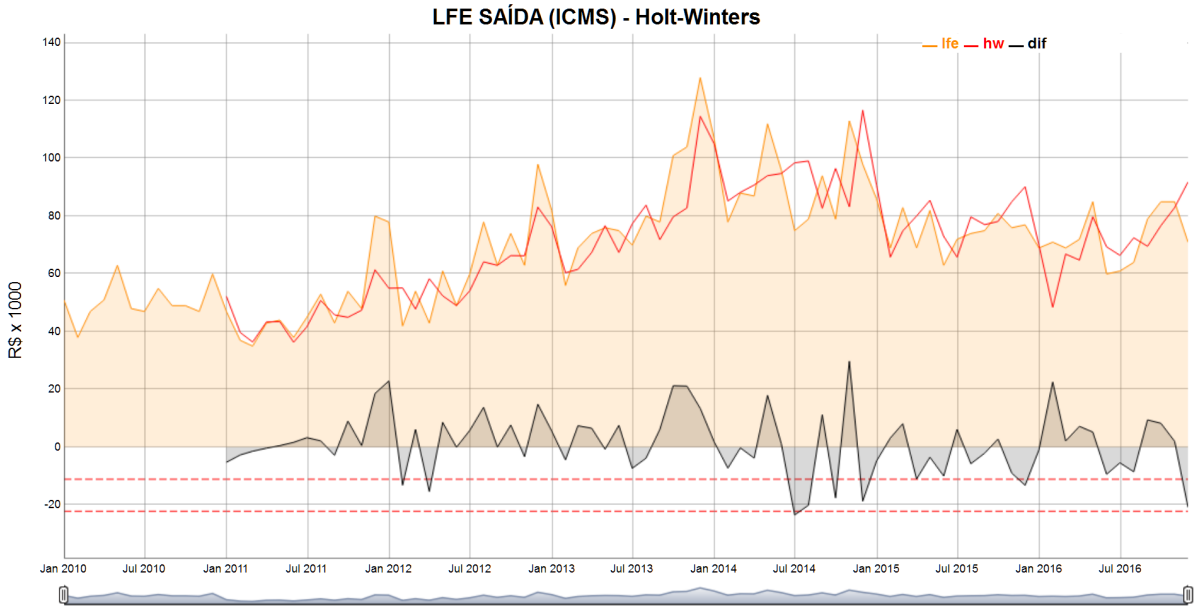

Figura 5.11: Holt-Winter - LFE - SAÍDAS.

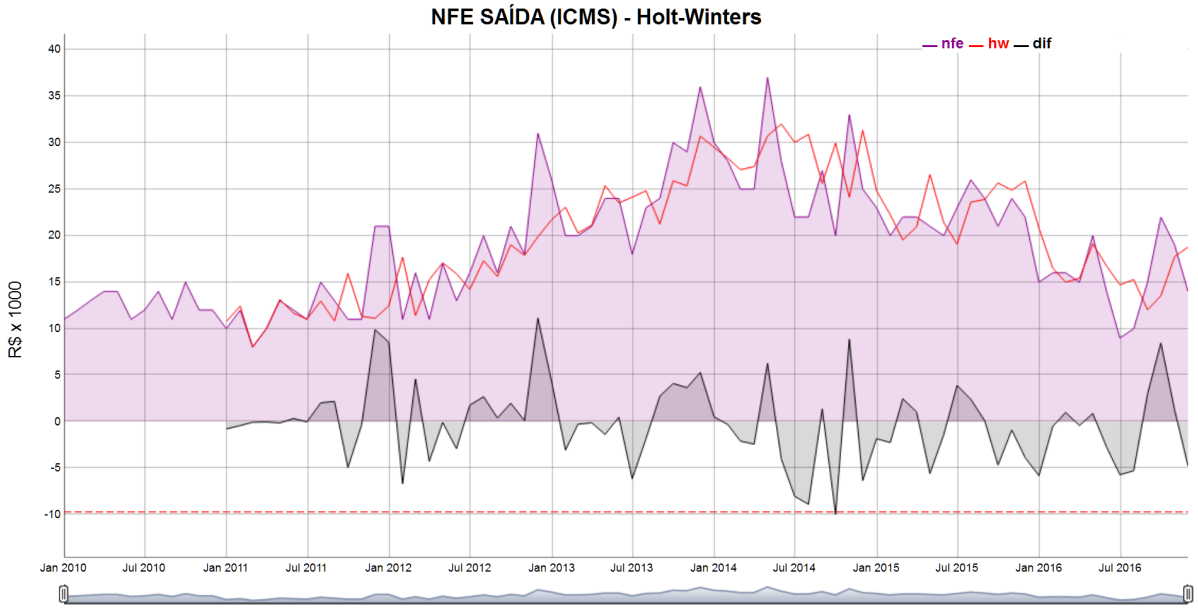

Figura 5.12: Holt-Winter - NFEs - SAÍDAS.

Com arrimo de métodos de análise de séries temporais é proposta uma forma de avaliação do comportamento tributário das empresas participantes de um processo de monitoramento e avaliação fiscal.

A partir dos dados sobre os valores do ICMS provenientes das Entradas e das Saídas (compras e vendas) realizadas pela empresa sob avaliação – informações disponíveis na escrituração do Livro Fiscal e no acervo de Notas Fiscais eletrônicas onde o contribuinte participa como destinatário ou como emitente – são construídas as séries de tempo que refletem a evolução desses valores. Sobre essas sequências temporais, aplica-se uma metodologia preditiva e comparativa para avaliar a contingência de eventuais valores anômalos suspeitos.

No exemplo aplicado foi possível identificar graficamente: a) as divergências na escrituração (créditos a maior e débitos a menor) pelo cotejo entre as séries (Entradas e Saídas – NFE vs. LFE), b) os meses que apresentam valores outliers quando comparados por suas distribuições (box-plot) ou com uma previsão de controle (Holt-Winter), e c) as componentes de tendência e sazonalidade que caracterizam a atividade econômica da empresa e que podem ser compreendidas ou cotejadas com o paradigma do setor econômico a que ela pertence na busca de dissonâncias suspeitas.

Essa etapa da metodologia sugestionada proporciona fundamento estatístico para a escolha de períodos de auditoria onde se verifique o acontecimento dos resultados anômalos de maior probabilidade de ocorrer evasão fiscal. Essa escolha pode acontecer: a) em razão da divergência entre a realidade dispostas nas NFE e os valores contabilizados no LFE, b) pelas discrepâncias em iguais períodos entre as séries ou entre as séries e o padrão esperado, e c) pela dissenção das séries com o parâmetro percebido na atividade econômica.

Assim, a metodologia é útil para otimizar o processo de monitoramento tributário, indicando os momentos com a maior possibilidade de conter a prática de ilícitos fiscais perpetrados pelo contribuinte.

Capítulo 6

Análise Estatística e Mineração K-Means

O escopo dessa terceira etapa é a aplicação de métodos de *data mining*, que incrementem a persecução às evasões fiscais ofertando a identificação precisa dos indícios das possíveis práticas de evasão do ICMS. Para atingir esse ideal far-se-á uso de modelos estatísticos (paramétricos e não-paramétricos) e do método de mineração *K-means clustering* para a definição de valores *outliers* que possam significar suspeições.

Este estágio da pesquisa intenciona propor um conjunto analítico que ajudará a maximalizar a tarefa de programação fiscal, oferecendo um procedimento analítico de pré-auditoria que identificará as discrepâncias presentes nos dados contábeis e tributários dos contribuintes, em especial, daqueles que participam do cálculo do ICMS a ser cumprido.

Nessa aspiração, serão utilizadas técnicas estatísticas descritivas e probabilísticas úteis para a detecção de anomalias, bem assim a técnica de observação de valores *outliers* pela distância média em agrupamentos (*K-means clustering*).

6.1 Análise Estatística e Mineração de Dados *Outliers* (método *K-means*)

O resultado da aplicação dos métodos enunciados em seguimento - para um contribuinte previamente selecionado - oferecerá ao auditor fiscal uma coleção de painéis analíticos mostrando as anomalias candidatas a serem inspecionadas.

É importante notar que a metodologia será igualmente aplicada aos dois momentos operacionais da tributação do ICMS, a saber:

a) **Entradas/Compras** - Neste momento, deverá ser priorizada a inspeção da escrituração (contábil) correta LFE dos créditos apropriados quando das aquisições, e a

sua confirmação nos documentos fiscais adequados (NFEs). A presença de valores extremos deve motivar a investigação da realidade dessas operações comerciais de formação de estoques, compra de ativos, etc. É possível que esses valores extraordinários representem a apropriação indevida de créditos tributários.

b) **Saídas/Vendas** - De igual forma, será considerada a exatidão dos lançamentos (contábeis) do valor tributável das operações de vendas sujeitas ao ICMS. Eventuais valores anômalos registrados inferiores podem significar a supressão (ou a redução) da tributação real das operações realizadas.

Primeiramente a avaliação gráfica da distribuição da frequência dos dados indicará visualmente a presença de valores extremos, confirmados por testes estatísticos de desigualdade probabilistica para a sua melhor precisão.

Em sequência, a análise de agrupamento dos valores contabilizados das NFEs em comparação com o correspondente imposto (ICMS) destacado nas notas, recuperará:

- Os valores anômalos pertencentes à *clusters* distantes em relação aos demais.
- Os valores discrepantes em relação aos seus respectivos agrupamentos.

Essas informações elegem as operações, documentos ou assentos contábeis como suspeitos e dignos de uma apuração em auditoria. Nesse sentido, a identificação de valores anômalos pode otimizar as atividades de inspeção, economizando o tempo da atividade de auditoria. Cumpre-se, desta forma, o objetivo de construir uma metodologia analítica útil para a atividade de investigação fiscal, facilitando e direcionando os trabalhos de inspeção para os valores com maior esperança de sucesso.

Para sustentar uma metodologia que forneça ao auditor fiscal uma discriminação das anomalias que possam constituir a ocorrência de uma hipótese de evasão do (ICMS), esta pesquisa argumenta a favor da combinação dos seguintes métodos:

6.1.1 Modelos Paramétricos e Não-Paraméticos

Medidas de Posição e de Dispersão

As medidas de tendência central (média aritmética, mediana, moda, valor mínimo, valor máximo e os quantis) e as medidas de dispersão (desvio padrão e coeficiente de variação) são calculados para os valores extraídos (por período anual) do NFEs e do LFE para comparação.

A análise comparativa dessas medidas tem o poder de esclarecer a correspondência obrigatória esperada entre a quantidade de ICMS contemplada na NFEs e seu registro necessário no LFE. Distorções podem indicar evidências de fraude.

Gráficos da Distribuição dos Dados

Histogramas, Histogramas em logaritmo, Gráficos de Distribuição *Kernel*, *Box-Plots* e *Scatter plots* serão usados para exibir o comportamento da distribuição de freqüência dos valores em teste:

- O **Histograma** é um modelo gráfico de representação de dados construído sob a pespectiva de suas frequências. Sua particular utilidade se dá em um contexto onde, trabalhando-se com um grande número de observações, necessita-se obter a probabilidade de acontecimento de um evento raro.
- O **Log-Histograma** atribuí ao eixo das ordenadas do histograma uma escala logarítmica (*in casu* base 10). Essa estratégia permite a apreciação de valores iterados e raros em uma única disposição gráfica, condição que facilita a identificação dos valores extremos *outliers* de baixa frequência.
- O **Gráfico de Densidade *Kernel*** é uma solução (não-paramétrica) para a estimação dos contornos da curva de densidade de uma variável aleatória, o que auxilia a avaliação do comportamento dos dados e proporcionando a observação dos eventos incomuns (*outliers*). Sua construção sucede a ponderação da distância de cada observação em relação a um valor central, chamado núcleo (*kernel*). Assim, é possível ter uma projeção do perimetro que conforma a distribuição de probabilidade do fenômeno e suas singularidades.
- O ***Box-Plot*** é um método não-paramético de representação diagramática da variação das observações com arrimo no uso da medida dos quartis da distribuição. As distâncias inter-quartis definem o grau de dispersão, a obliquidade nos dados e os valores *outliers*. O desenho apresenta um segmento que se estende a partir da caixa (*box*), estabelecendo as variabilidades que extrapolam os limites dos quartis superior e inferior. Os *outliers* serão destacados como pontos individuais que ultrapassam esses limites.
- O ***Scatter Plots***, ou diagrama de dispersão, é um tipo de gráfico de coordenadas que exibe valores de um conjunto de dados por pontos no espaço cartesiano. Para a melhor visualização dos dados, em especial para a melhor identificação do rompimento da fronteira limite indicada para os outliers, este gráfico é construído com os dados ordenados em ordem crescente.

A **desigualdade Chebyshev** será aplicada como parâmetro para a percepção de *outliers* estabelecendo limites de suspeição em cada gráfico. Testes *Grubbs* e Qui-quadrado também serão usados para identificação (e principalmente confirmação) de pontos distintos como anomalias.

Desigualdade de Chebyshev

Usar-se-á a desigualdade de cauda para estabelecer a probabilidade de um valor da distribuição ser considerado *outliers*. Assim:

Sendo X uma variável aleatória com média μ e variância σ^2, ambas finitas, a probabilidade de se obter um valor distante da média em uma medida acima de k desvios-padrão (σ) é menor ou igual a $1/k^2$, ou seja:

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2} \quad \forall \quad k > 0 \tag{6.1}$$

E, de maneira complementar:

$$\begin{gathered} P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \quad \Rightarrow \\ P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2} \quad \forall \quad k > 0 \end{gathered} \tag{6.2}$$

Inobstante a fraca precisão da desigualdade (estabelecida na convergência por probabilidade), porquanto nenhuma hipótese foi estabelecida sobre a da variável aleatória (somente que ela possui média e variância conhecidas), ela é muito útil para estabelecer uma fronteira de limite cuja ultrapassagem define uma anomalia.

A identificação de valores *outliers* será definida pelos excessos (superiores e inferiores) das medidas definidas pela desigualdade e repercutirá na escolha dos documentos e (ou) registros contábeis-fiscais que mereçam ser auditados, pois podem representar informações de operação impróprias ou fraudulentas.

Para fins de inspeção fiscal, melhor atenção deverá ser dedicada:

a) À verificação da formação de ativos do contribuinte e consequente acumulação de créditos, registros excessivos ou extremos no LFE, especialmente incompatíveis com as operações suportadas em NFEs válidas.

b) De outra sorte, as vendas registradas no LFE devem ser observadas quanto aos valores (débitos) registrados do imposto inferiores aos destacados no documento que suportam a operação.

Essas diferenças podem ser percebidas graficamente pela análise comparativa do perfil das distribuições de frequência dos dados (NFEs x LFE).

A justaposição comparativa entre as distribuições da NFEs e da LFE (compras e vendas) também permite a verificação visual das dissensões.

Lei de Newcomb-Benford

A lei de Newcomb-Benford será empregada como ferramenta de auxílio na pré-auditoria, atuando como instrumento de seleção de documentos e registros fiscais incomuns.

Também conhecida como Lei do Primeiro Dígito, ela prescreve a probabilidade de frequência esperada para os dígitos em dados tabulados. De modo não intuitivo, os dígitos em dados tabulados (especialmente de fenômenos financeiros) não são equiprováveis apresentando um tendência assimétrica em favor dos dígitos inferiores, a saber:

O cálculo das frequências do primeiro dígito é dado pela função:

$$P(D_1 = d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right) \forall\ d_1 \in\ (1, 2, ..., 9) \tag{6.3}$$

onde P indica a probabilidade esperada de ocorrência do primeiro dígito D_1 em exame. O zero é inadmissível como um primeiro dígito e há nove possíveis primeiros dígitos (1, 2, ..., 9).

O resultado das frequências esperadas advindas da aplicação da função do primeiro dígito é exibido a seguir na Tabela 6.1.

Dígito	Probabilidade
1	.30103
2	.17609
3	.12494
4	.09691
5	.07918
6	.06695
7	.05799
8	.05115
9	.04576

Tabela 6.1: Probabilidade Esperada para o Primeiro Dígito

Destarte é possível proceder uma análise de conformidade de um conjunto de dados, mediante a detecção de desvios em relação ao padrão esperado de ocorrência, para os dígitos analisados.

A lei do primeiro dígito será aplicada sobre os valores contábeis das NFEs e do LFE recuperados para, por meio do confronto da probabilidade logarítimica com a realidade dos valores estudados, assinalar as diferenças significativas que deverão guiar a atenção do auditor para a investigação.

6.1.2 Método *K-means Clustering*

K-means clustering é um algoritmo eficiente de aprendizagem não supervisionada que resolve o problema de agrupamento e facilita a detecção de anormalidades nesses grupos. Sua proposta é a de dividir um conjunto de dados em k segmentos ótimos, distintos e sem sobreposição de elementos.

O processo de agrupamento *K-means* deriva do seguinte problema matemático:

Sendo S_1, S_2, ..., S_k os k conjuntos contenedores das n observações associadas em *clusters*, observar que cada elemento deverá pertencer a um, e somente um, dos segmentos (*clusters*) disponíveis.

Ou seja, os conjuntos $S_{1...k}$ que particionam os n elementos devem satisfazer à duas condições:

a) $S_1 \cup S_2 \cup ... \cup S_k = (1,..., n)$ e

b) $S_k \cap S_{k'} = 0 \; \forall \; k \neq k'$.

Ademais, deve-se particionar as observações para k *clusters* de modo que a variação total dentro do cluster, somada em todos os k *clusters*, seja a menor possível. Quer dizer, sendo Δ a distância que separa um elemento de outro dentro de um *cluster*, deve-se:

$$Minimizar_{(S_1...S_k)} \left\{ \sum_{k=1}^{K} \Delta(S_k) \right\} \tag{6.4}$$

Isto é, queremos segmentar os elementos em k *clusters* de maneira que a variação total dentro do *cluster* $\Delta(S_k)$, somada em todos os *clusters* $\sum_{k=1}^{K} \Delta(S_k)$, seja tão pequena quanto possível.

Para tanto é necessário definir a variação dentro do *cluster* $\Delta(S_k)$. Dentre as formas possíveis de definir esse conceito, prefere-se o uso distância euclidiana ao quadrado (podendo ser usadas outras distâncias também).

$$\Delta(S_k) = \frac{1}{|S_k|} \sum_{(i,i' \in S_k)} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \tag{6.5}$$

sendo $|S_k|$ o número de observações em k^{th} cluster.

Assim, a variação dentro do cluster para o k^{th} *cluster* $(\Delta(S_k))$ é a soma de todas as distâncias euclidianas quadradas entre as observações no k^{th} *cluster*, dividido pelo número total de observações no *cluster* K^{th} $(|S_k|)$.

Essa medida baseia-se implicitamente na distância euclidiana linear entre dois pontos $(x_{ij} - x_{i'j})^2$. Ou seja, dado um conjunto de observações $(x_1, x_2, ..., x_n)$, onde cada observação é um verdadeiro vetor-dimensional, *K-means clustering* objetiva particionar

as n observações em k grupos ($S = S_1,\ S_2,\ ...,\ S_k$) de forma a minimizar a soma dos quadrados dentro de cada *cluster* (soma das distâncias de cada ponto no conjunto para o centro do *cluster* K^{th}) minimizando otimamente essas distâncias:

$$Minimizar \left\{ \sum_{k=1}^{K} \frac{1}{|S_k|} \sum_{i,i',S_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \tag{6.6}$$

O algoritmo *K-Means Clustering* apresenta-se da seguinte forma:

Inicialmente, o algoritmo solicita a determinação do número desejado de *clusters* K (entre 1 e k), para em seguimento atribuir a cada elemento do conjunto de observações sua exata inserção em um dos k *clusters*. Essa atribuição poderá ser aleatória e serve como parâmetro de quantidade de *clusters* iniciais para as observações.

Sequencialmente, itera-se os procedimentos abaixo até que as atribuições de cluster parem de oferecer mudanças significativas:

a) Para cada k *clusters*, calcula-se o seu centróide. O k-ésimo centróide do *cluster* K^{th} é o vetor da média característica para as observações no k^{th} *cluster*.

   K-means clustering deriva seu nome do fato de que os centróides do *cluster* são computados como a média das observações atribuídas a cada *cluster*.

b) Para cada elemento atribua-lhe o cluster cujo centróide lhe é mais próximo (onde o mais próximo é definido usando-se a distância euclidiana).

Como o algoritmo K-means encontra um ótimo local, em vez de um ótimo global, os resultados obtidos dependerão da atribuição de *cluster* inicial (aleatória) de cada observação. Por esta razão, é importante executar o algoritmo várias vezes em diferentes (aleatórias) configurações iniciais. Finalmente, seleciona-se o melhor resultado.

A investigação de anomalias (*outliers*) ocorre pelo cálculo das maiores razões de distâncias relativas. Melhor dizendo, a relação da distância absoluta do elemento para o centro do *cluster* e a distância média de todos os elementos do *cluster* em relação ao centro do respectivo *cluster*.

Realiza-se a construção de *clusters* (métodos *K-means*) para encontrar desvios anormais na associação entre: a) o montante total de NFEs e b) o ICMS sobre a transação contemplada neste documento fiscal.

O método Elbow - que investiga a porcentagem de variância explicada como função do número de clusters - é o método indicado para escolher o valor de k. Nele, é verificado o momento de inflexão da redução marginal de porcentagem da variância (diminuição explicada pelos grupos comparados por número de clusters) como indicador da quantidade ideal de *clusters*. No ponto em que a adição de um novo cluster (k + 1) não diminui

significativamente o erro (ponto de inflexão), o número de grupos k pode ser visto como o recomendado.

Após experimento sistemático realizando-se testes com uma amostra de dados de 2000 contribuintes e fazendo uso do método Elbow acima descrito, foi possível concluir que o valor ideal de k agrupamentos é igual a 6 (seis).

A formação de *clusters* pelo método de agrupamento *K-means* ajudará a investigar os valores disjuntos do ICMS declarados em relação ao valor real da operação.

6.2 Mineração de Dados *Outliers* na Literatura Científica

Existe uma boa quantidade de trabalhos científicos que tratam da análise de *outlier* no contexto estatístico e sob a perspectiva da ciência da mineração de dados. Diversos desses trabalhos oferecem soluções práticas de algoritmos computacionais para detectar esses valores atípicos em conjuntos de dados.

Hawkins [63] define um *outlier* como: *"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."*

Assim, podemos entender um *outlier* - também referenciado na literatura estatística e de *data mining* como anormalidade, valor discordante, valor desviante ou anomalia – como sendo uma observação que se desvia demasiadamente das demais do conjunto, despertando a suspeita de que ela foi constituída por um processo gerador diferente das outras.

Consoante contempla-se dos ensinamentos de Hawkins [63], um processo gerador de dados reflete a atividade de um sistema ou o comportamento observado de uma entidade. Quando esse processo de geração se comporta de forma inusitada, resulta na criação de *outliers*. O reconhecimento de tais características incomuns fornece *insights* úteis de aplicação específica, porquanto, um *outlier* geralmente contém informações de interesse sobre as condições dos sistemas e das entidades que afetam o processo de geração de dados.

Segundo Aggarwal [42] a detecção de valores atípicos encontra aplicação em numerosos domínios, onde é desejável para se determinar eventos de interesses incomuns subjacentes a um processo gerador de dados.

A lição do sobredito autor consigna que o núcleo de todos os métodos de detecção de valores extremos é a criação de um sistema probabilístico, estatístico ou de um modelo de algoritmo que caracterize os dados considerados normais. Os desvios desse padrão de "normalidade" são considerados na identificação das ocorrências anômalas.

Do mesmo modo, Tan [64] define *outliers* como fatores estranhos e apresenta a mineração de tais exceções por abordagem estatística, aplicando testes de desvios na distribuição de probabilidade dos valores univariados e indicando abordagens baseadas em grupos, para os elementos multivariados.

A discussão sobre o conceito e a utilidade dos diversos métodos de detecção de *outliers* na doutrina estatística acontece sob diferentes concepções e é objeto de uma razoável quantidade de estudos acadêmicos. Entre esses, a pesquisa escrita por Chandola *et al.* [65] destaca-se como uma das mais abrangente. Nele os autores fazem uma excelente revisão compreensiva sobre o tema da detecção de valores aberrantes a partir da perspectiva dos conjuntos uni e multivariados.

Igualmente os trabalhos de Bakar *et al.* [66], Zang [67], e Malik *et al.* [68] fornecem uma análise comparativa entre os diferentes métodos existentes para detectar anomalias. Apesar desses estudos oferecerem uma ampla gama de métodos alternativos, em todos, é enfatizado o uso dos modelos estatísticos descritivos para reconhecer valores extremos como uma indicação ótima para a análise de dados univariados.

Seguindo uma visão adequada à mineração de dados, uma multiplicidade de métodos de detecção de valores aberrantes baseados na distância, densidade ou clusters de composição são discutidos extensivamente em James [69] e Witten *et al.* [70], focando o texto especialmente na aplicação prática dos modelos dispostos.

Advogando o uso de métodos estatísticos para detectar anomalias em dados univariados combinam-se os conteúdos das lições de Han [71], Aggarwal [42] e Hastie *et al.* [72], inclusive na circunstância onde o volume dos dados seja demasiadamente grande.

Além disso, Aggarwal [73] encoraja fortemente o uso de histogramas e técnicas baseadas em *Grids*, dada a sua simplicidade de construção e exame por qualquer analista.

Em particular, os métodos não-paramétricos para a detecção de valores anormais, com aplicação direta na descoberta de fraude contábil, são extensamente tratados em Oliveira *et al.* [74].

Knorr e Ng [75] propuseram uma definição de *outlier* em função da sua distância relativa no conjunto de dados e que se encontra livre de quaisquer pressupostos de distribuição de frequência, e é generalizável também para conjuntos de dados multidimensionais. Intuitivamente, propõem que *outliers* são pontos de dados que estão longe de seus vizinhos mais próximos. Seguindo o pensamento proposto por Knorr e Ng [75], variações de algoritmos foram propostos para detectar valores posicionados à distância, como em Wu [76].

A seu turno, Aggarwal [42] conclui que os dois métodos escolhidos nessa proposta (o método estatístico-descritivo e o método *K-means*), apesar de não serem os mais recentes, ou mesmo, de não serem os modelos de detecção de *outlier*s mais sofisticados, possuem

o predicado da simplicidade e da eficiência sustentando razões que indicam seu uso como solução para mineração de dados voltada à maioria dos usuários.

As aplicações das técnicas de mineração de dados para pesquisar anomalias na tributação brasileira (ICMS) são, até agora, raras. Não obstante, existem bons trabalhos aplicados à detecção de fraude financeira e contábil, como o que foi desenvolvido no trabalho de Baesens *et al.* [77] que fornece a apresentação das técnicas de detecção de *outliers* como evidência de fraude, dando ênfase em análises paramétricas estatísticas, gráficas e de *K-means clustering*.

Uma referência notável é a pesquisa de Oliveira [74] na percepção de fraude contábil, que desenvolveu com sucesso uma análise comparativa de técnicas estatísticas para a detecção dessas anomalias, tais como: métodos *quantile-quantile*, *Hampel*, *boxplot*, distribuição t de *Student* e Testes de qui-quadrado de distribuição, *Grubbs*, *Dixon* e ESD generalizada.

Também merece menção o trabalho de Bolton [78] que utiliza com êxito a aplicação de análises estatísticas de anomalias para a distinção de fraudes como a lavagem de dinheiro, fraudes em cartão de crédito, fraudes no comércio eletrônico, fraude em telecomunicações e intrusão de computador, para nomear somente as principais.

Finalmente destacam-se os estudos de Nigrini [79] e [80] consoante a avaliação de fraudes com o uso da Lei de Newcomb-Benford, bem assim a sua aplicação em auditoria tributária proposta por Santos *et. al.* [81].

Este estudo promove a divulgação da aplicação bem-sucedida de tais técnicas, especialmente por sua relativa simplicidade em colação à sua eficiência.

6.3 Aplicação dos Modelos de Análise Estatística de *outliers* e do método *K-means Clustering*

6.3.1 Dados Utilizados

Nesta terceira etapa da metodologia proposta, serão usadas as seguintes informações extraídas das bases de dados da Receita Distrital:

a) **NFEs** Documentos fiscais de Entrada (relativos às compras) e de Saída (vendas) de mercadorias, bens ou serviços que traduzem a realidade das operações comerciais, e, em específico, fornecem o **valor dos bens** e do **ICMS destacado**.

b) **LFE** que traduz o movimento contábil-fiscal, em especial as operações tributáveis. Informações declaradas pelo contribuinte como elemento para a apuração periódica do imposto a recolher aos cofres públicos. Na contabilidade fiscal o montante dos

créditos (recuperáveis) é conjugado com os lançamentos à **débito** do imposto resultante das operações tributadas praticadas. O resultado desse cotejo compensatório é o ICMS a ser pago durante o período de apuração.

Os dados dedicados às analises ínsitas a essa fase estão armazenadas no banco de dados ORACLE da GEPRO e serão compilados mediante consultas definidas em linguagem SQL, propostas em ambiente *R Studio* e intermediadas pelo pacote RODBC.

Os gráficos dessa etapa analítica serão elaborados com o uso dos pacotes (do *Software R*): *DistributionUtils*[21], *abodOutlier*[22] e *ggplot2* [23].

Essa etapa usará os dados extraídos para identificar pontualmente os documentos e/ou os lançamentos contábeis (um a um) que se configurem como *outliers*.

6.3.2 Resultados Estatísticos

A análise de *outlier* acontece pela visualização dos diagramas gráficos analíticos pertinentes às distribuições de frequência dos dados sob exame - valores do ICMS destacados nas NFEs e escriturados no LFE, ambos sob a perspectivas das operações de Entradas e de Saídas.

A visão diversificada das distribuições dos dados colhidos, proporcionada pelos múltiplos modelos de plotagem sugestionados, assegura maior assertividade na visualização das anomalias merecedoras de preferência para verificação de regularidade.

As distribuições serão divididas pelo valor correspondente à desigualdade de Chebyshev, considerando como anomalias aqueles valores que transcendem esses limites.

Nada obstante a realização automática do cálculo e a identificação imediata dos *outlier* pelos parâmetros ínsitos ao modelo, é dado ao analista a liberdade de decisão por sua apreciação pessoal das distribuições dos valores emitidos (NFEs) e registrados (LFE).

A solução computacional disponibiliza 5 (cinco) visões gráficas das distribuições de probabilidade, expostas em painéis analíticos montados para os dados fiscais das Entradas e das Saídas, a conhecer:

a) Histogramas das Distribuição de Frequência.

b) Histogramas em Logaritmo.

c) Gráficos de Distribuição *Kernel*.

[21] *DistributionUtils* é um pacote do *Software R* que oferece funcionalidades para cálculo de v.g.: amostragem, assimetria, curtose, *log-histogram*, gráfico de calda, momentos por integração e funções de teste de distribuição. Informações: *cran.r-project.org/web/packages/DistributionUtils/DistributionUtils.pdf*

[22] *abodOutlier* é um pacote do *Software R* que implementa o algorítmo *K-means*. Informações: *cran.r-project.org/web/packages/abodOutlier/abodOutlier.pdf*

[23] *ggplot2* é um pacote do *Software R* para a criação de gráficos declarativamente com qualidade. Informações: *cran.r-project.org/web/packages/ggplot2/ggplot2.pdf*

d) Gráficos *Box-Plots*.

e) Gráficos *Scatter plots*.

Observar que:

- Para as Entradas as margens limites das anormalidades encontram-se à direita da divisa de Chebychev, apontando para a distinção de créditos anormais (*outliers*).
- Para as Saídas o limite (Chebychev) está à esquerda dos gráficos, significando valores do ICMS das operação extremamente baixos, aquém do usual.

Para o contribuinte em exemplo, logramos as subsequentes figuras consoantes aos valores do ICMS das operações de Entradas e de Saídas:

a) **Histogramas.** Representação gráfica da distribuição de frequências dos dados quantitativos do valor do ICMS - Entradas (Figura 6.1) e Saídas (Figura 6.2).

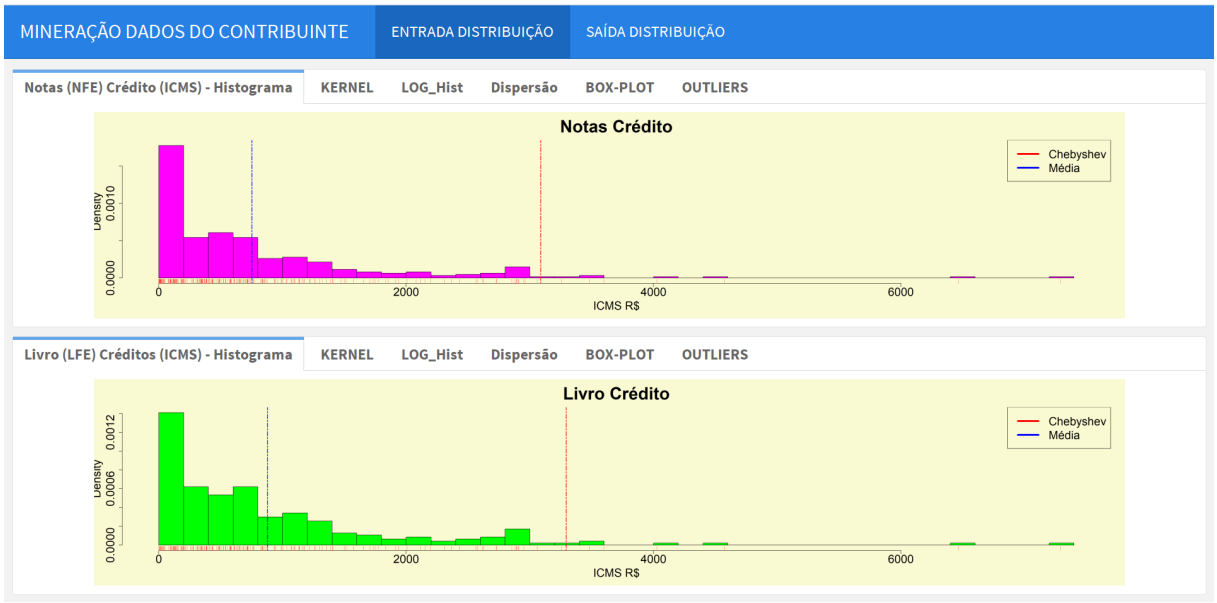

Figura 6.1: Histogramas - ENTRADAS.

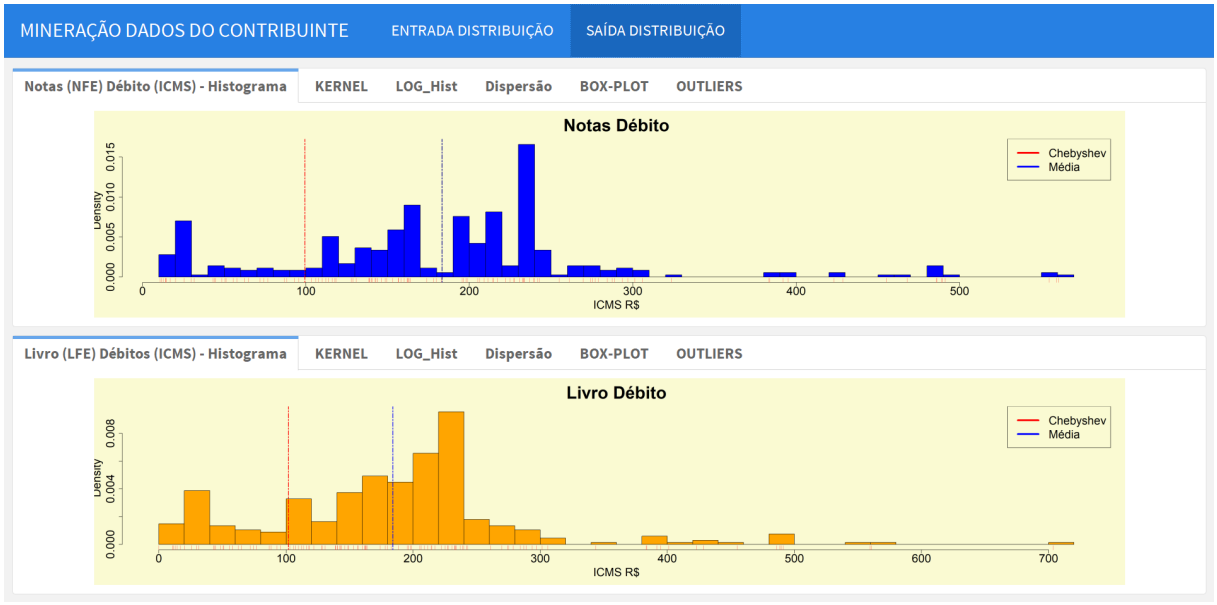

Figura 6.2: Histogramas - SAÍDAS.

b) **Histogramas em escala logarítimica**, que atenuam as altas frequências permitindo a visualização dos valores extremos (*outlier*) de baixa frequência - Entradas (Figura 6.3) e Saídas (Figura 6.4).

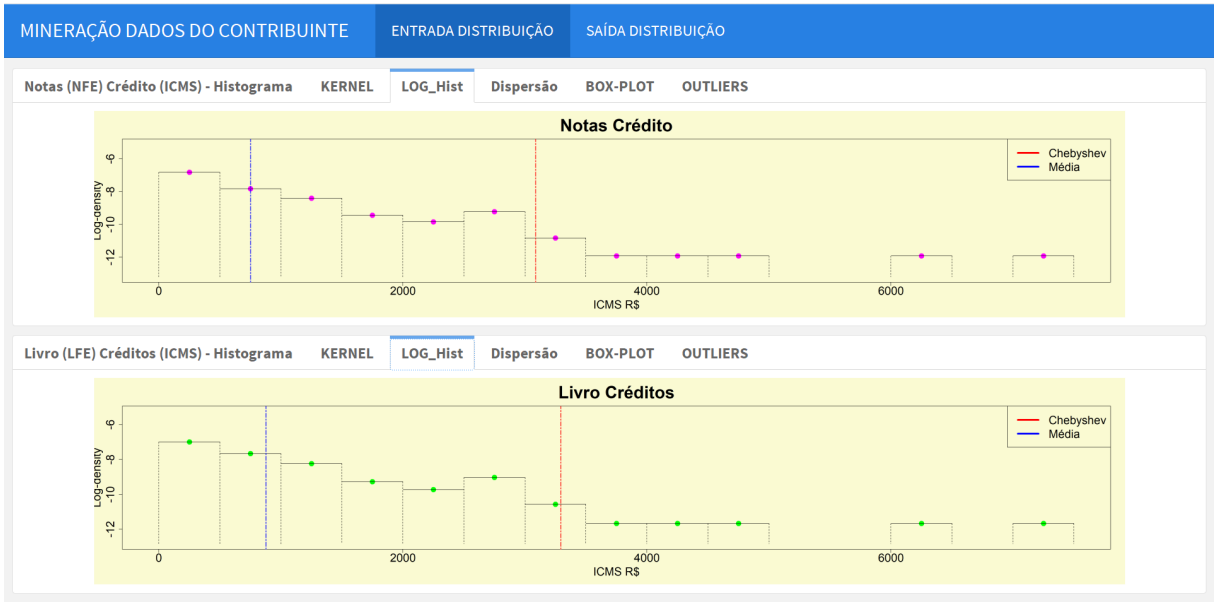

Figura 6.3: Log-Histogramas - ENTRADAS.

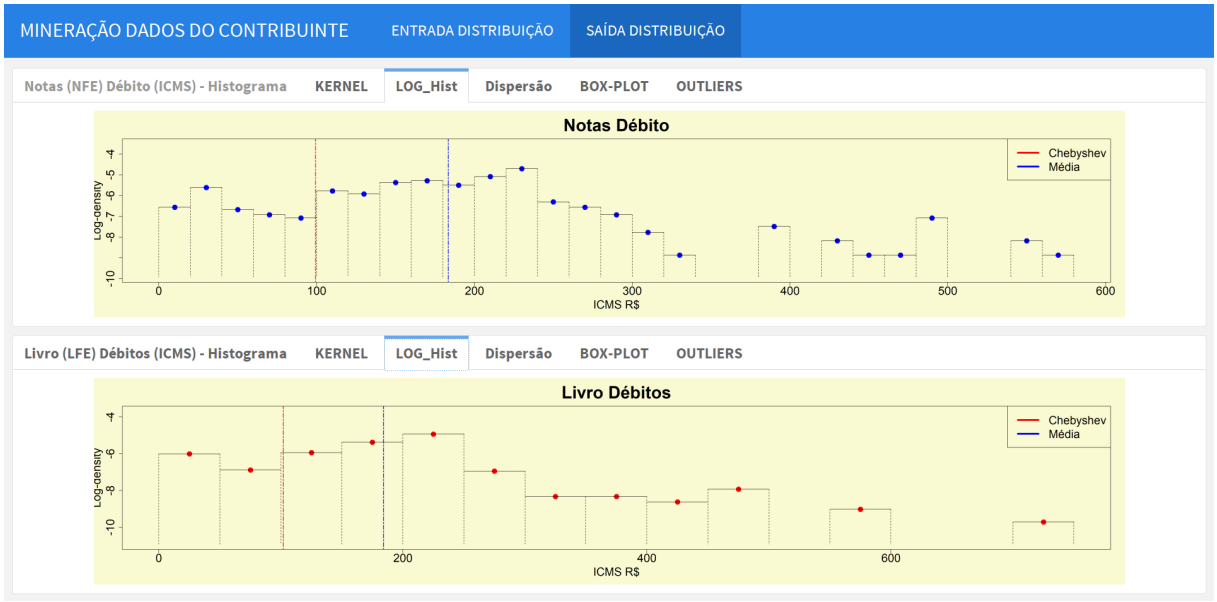

Figura 6.4: Log-Histogramas - SAÍDAS.

c) **Gráfico de Densidade *Kernel*** que é uma solução não-paramétrica para inferir a função de densidade de probabilidade da variável aleatória - Entradas (Figura 6.5) e Saídas (Figura 6.6).

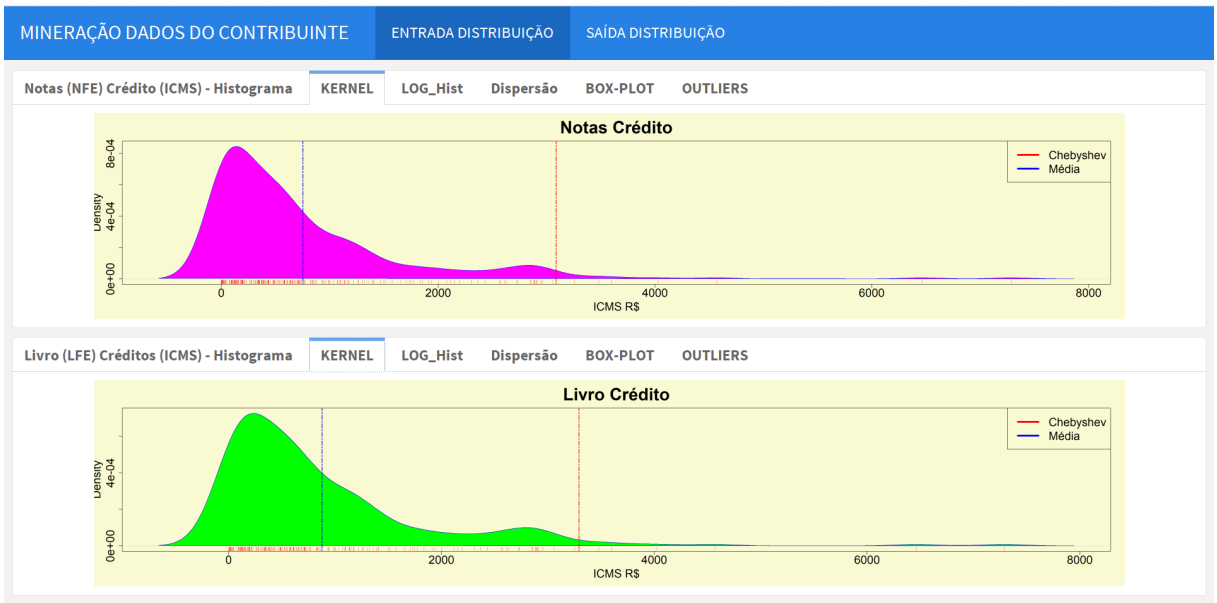

Figura 6.5: Densidades de Probabilidade *Kernel* - ENTRADAS.

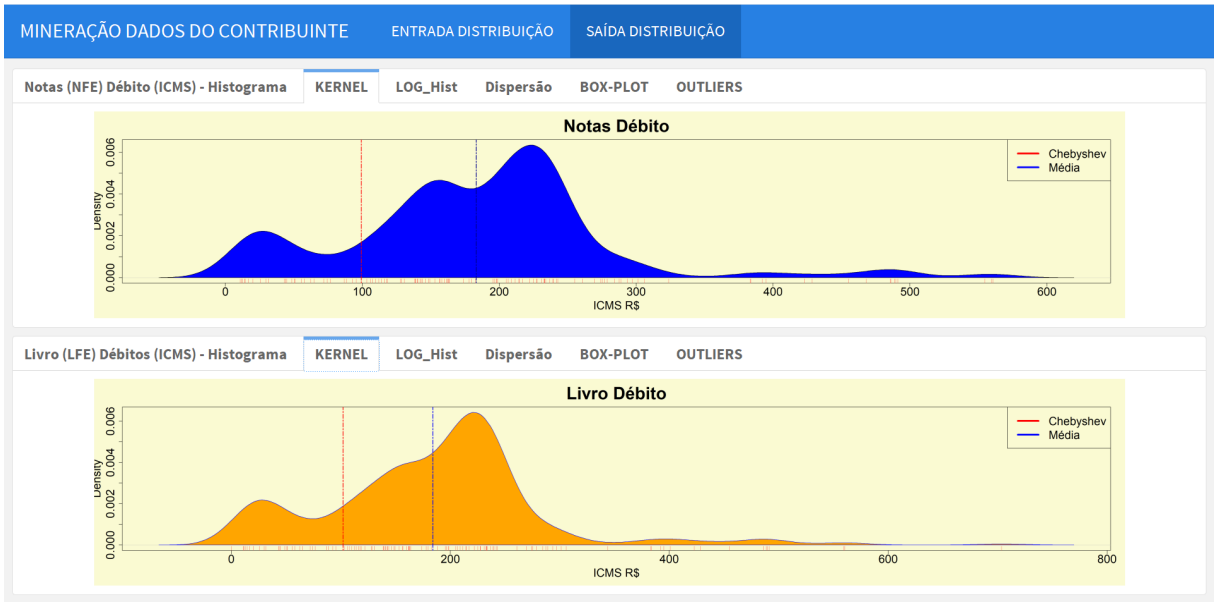

Figura 6.6: Densidades de Probabilidade *Kernel* - SAÍDAS.

d) ***Scatter Plot*** - Gráficos de Dispersão do conjunto de dados, ordenados pelo valor crescente - Entradas (Figura 6.7) e Saídas (Figura 6.8).

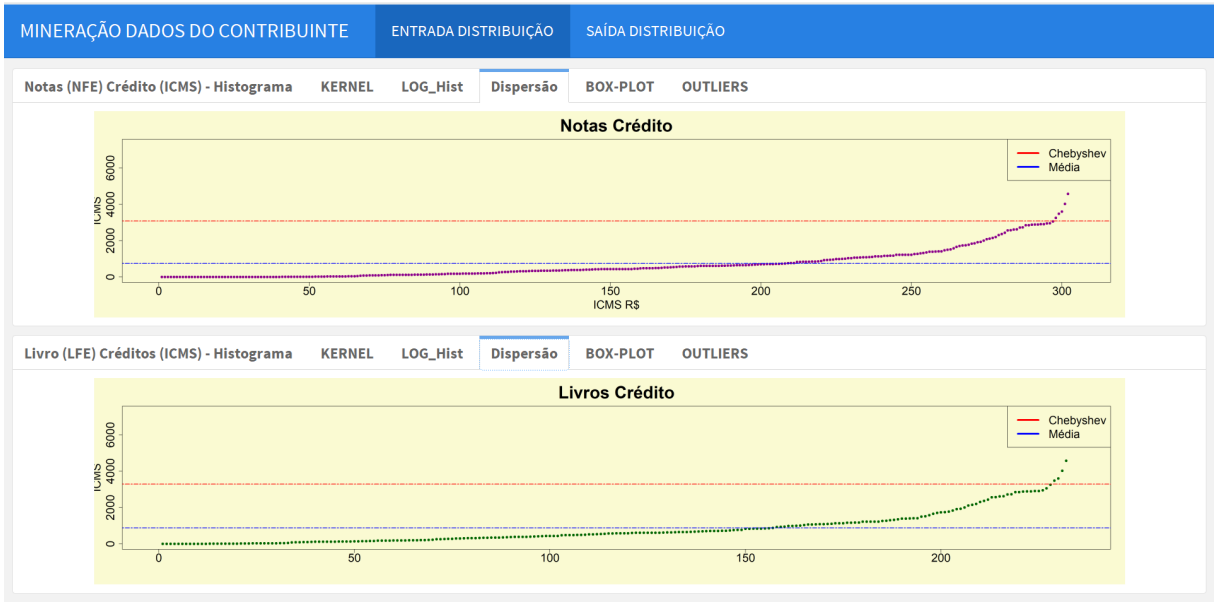

Figura 6.7: Dispersões *Scatter Plots* (ordenada) - ENTRADAS.

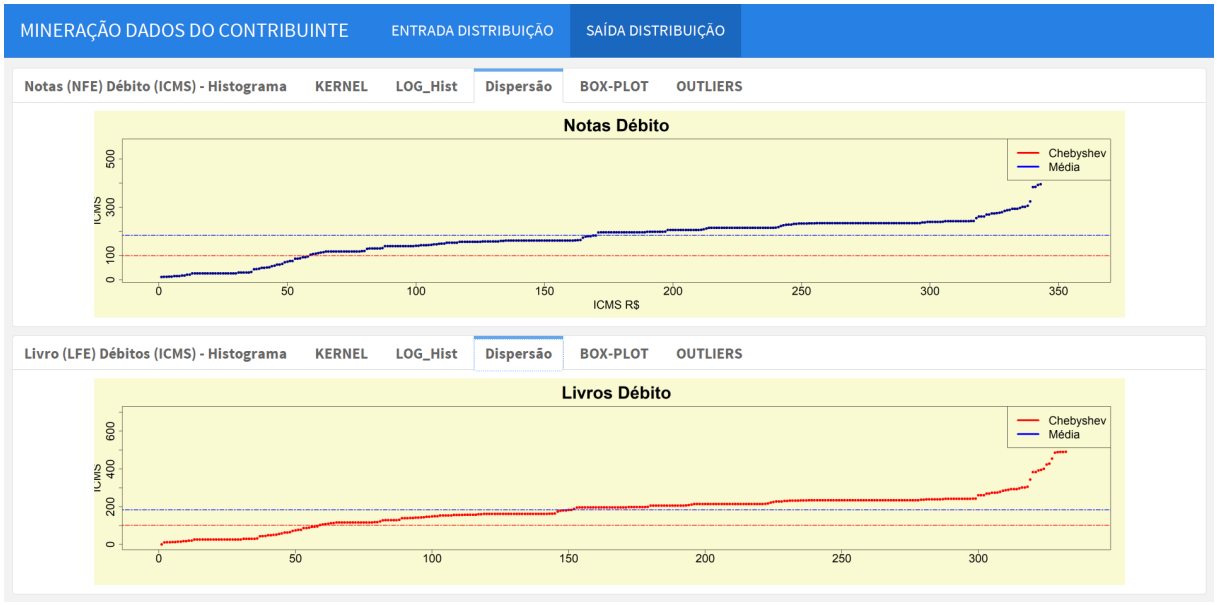

Figura 6.8: Dispersões *Scatter Plots* (ordenada) - SAÍDAS.

e) ***Box-Plot.*** Gráfico de padronização dos dados, resumido em cinco medidas: mínimo, primeiro quartil, mediana, terceiro quartil e máximo - Entradas (Figura 6.9) e Saídas (Figura 6.10).

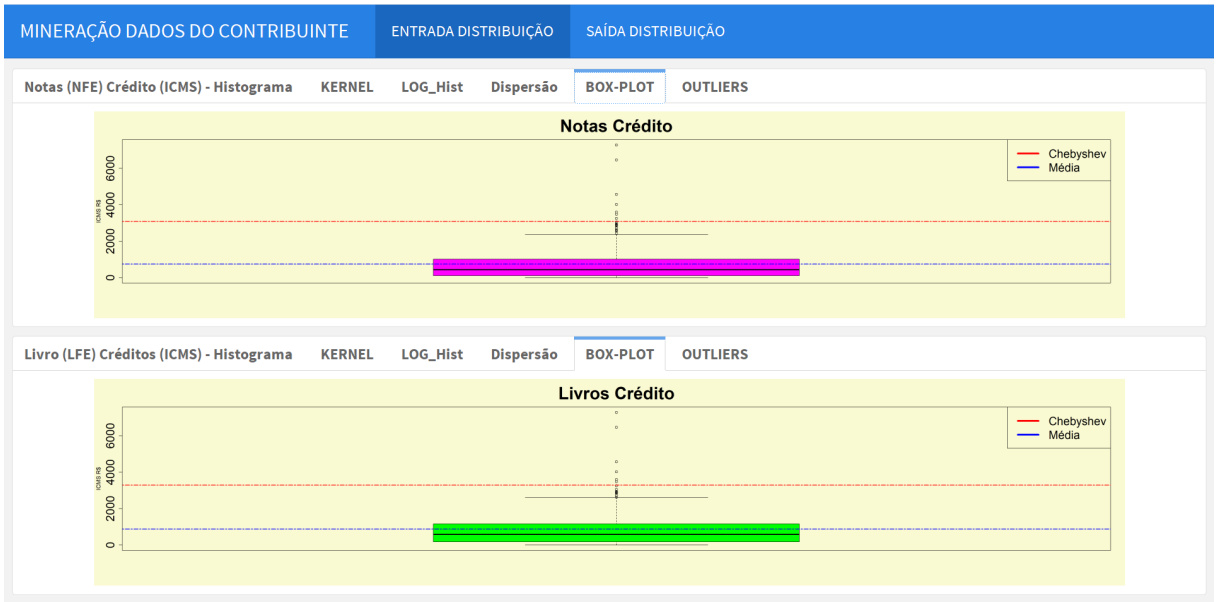

Figura 6.9: Gráficos *Box-Plots* - ENTRADAS.

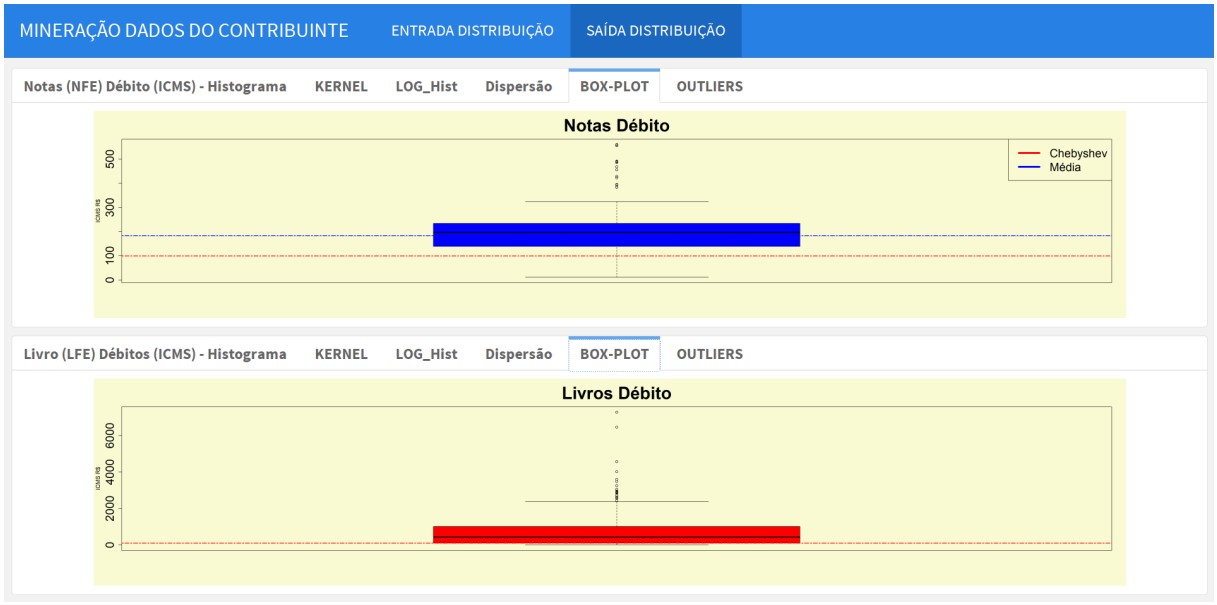

Figura 6.10: Gráficos *Box-Plots* - SAÍDAS.

f) ***OUTLIERS.*** Valores anômalos encontrados segundo o critério da Desigualdade de Chebyshev.

São indicados: o número do documento que contempla a anômalia (no LFE ou em NFEs), mês e ano de emissão e o valor da operação - Entradas (Figura 6.11) e Saídas (Figura 6.12).

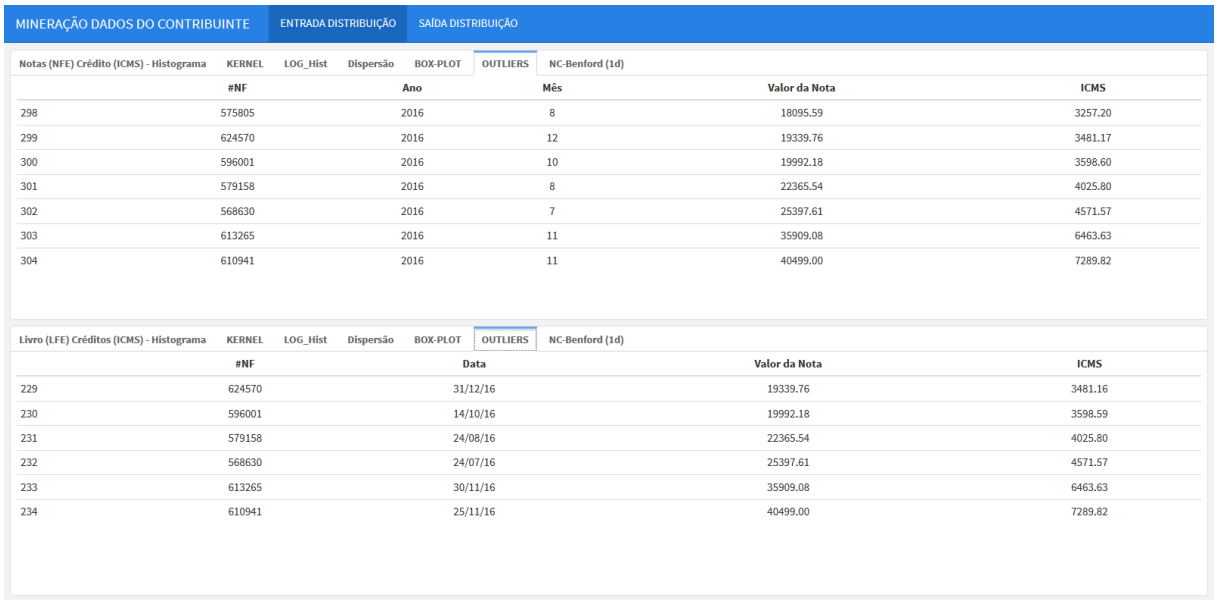

Figura 6.11: Valores *Outliers* - ENTRADAS.

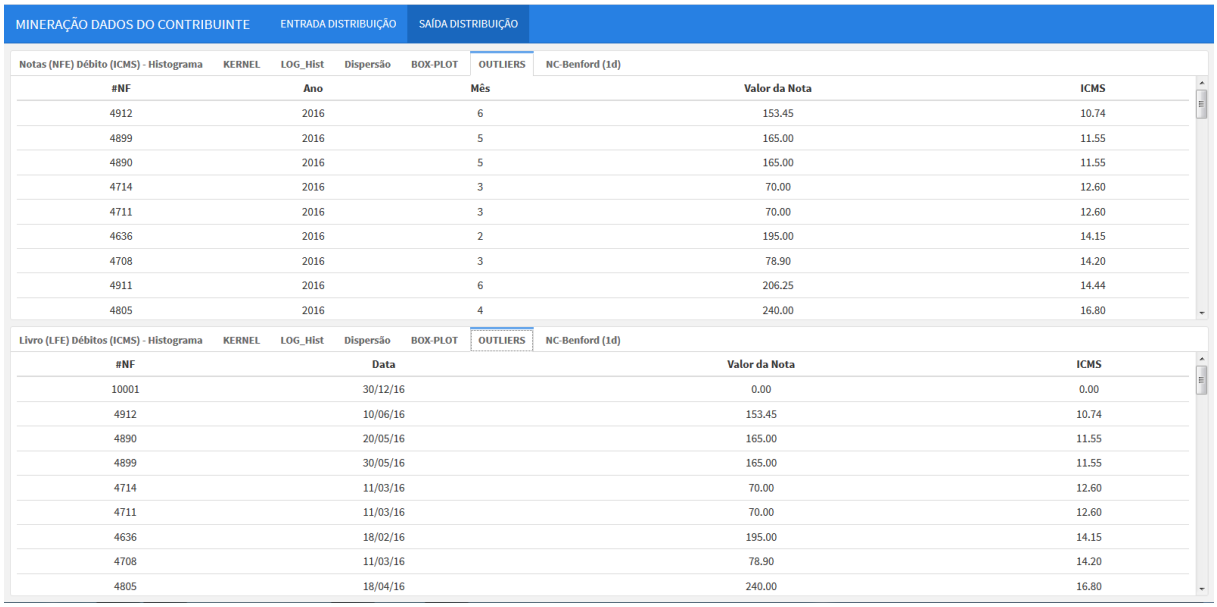

Figura 6.12: Valores *Outliers* - SAÍDAS.

6.3.3 Resultados da Lei de Newcomb-Benford

O cotejo entre a probabilidade esperada, originária da aplicação da lei do primeiro dígito, com os conjuntos formados pelos primeiros dígitos dos valores contábéis registrados nas NFEs e no LFE - Entradas (Figura 6.13) e Saídas (Figura 6.14) - indicará a existências de dissensões que exijam auditoria em determinado(s) conjunto(s) numérico(s).

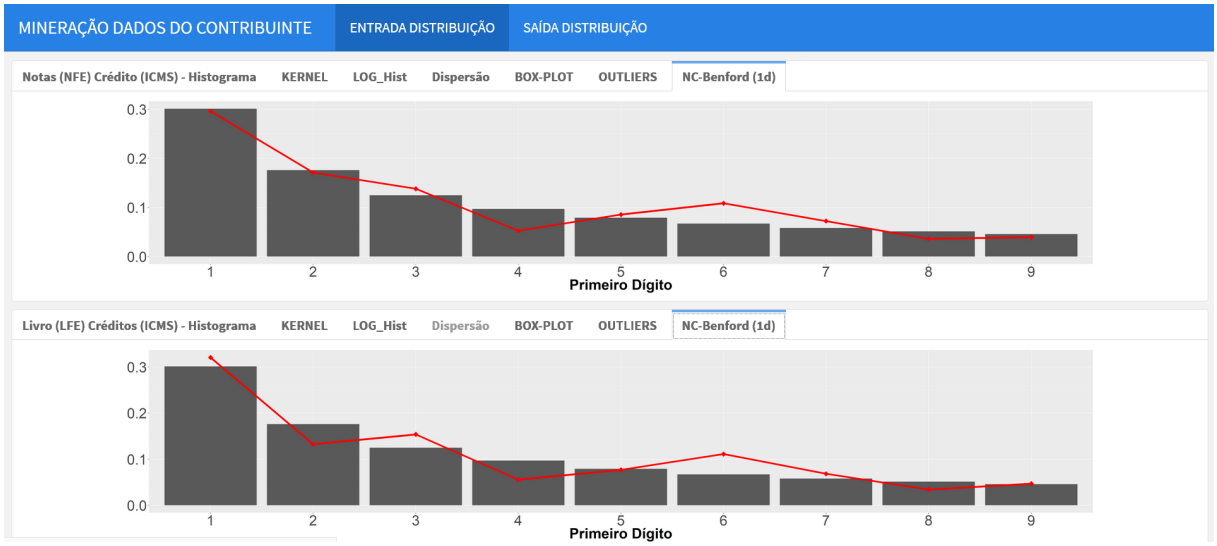

Figura 6.13: Lei do Primeiro Dígito X Valores Contábeis (NFEs e LFE) - ENTRADAS.

Nos dados analisados do exemplo em tema afiguram-se discrepantes:

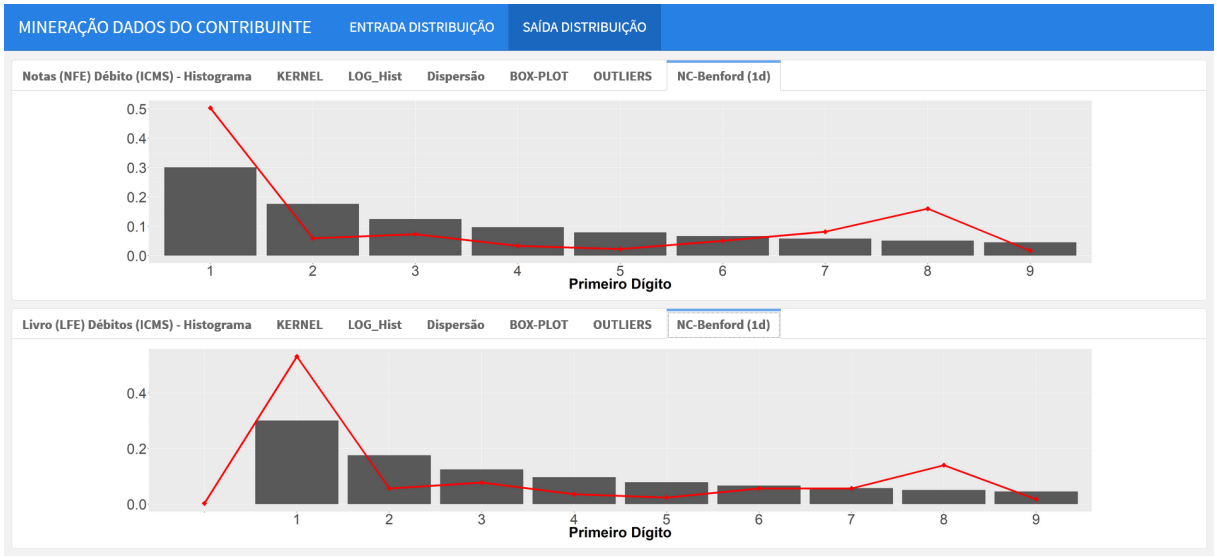

Figura 6.14: Lei do Primeiro Dígito X Valores Contábeis (NFEs e LFE) - SAÍDAS.

- Nas informações de Entradas - Os Valores Contábeis iniciados pelos dígitos 3 (três) e 6 (seis), e
- Nos dados das Saídas - Os Valores Contábeis principiados pelos dígitos 1 (um) e 8 (oito).
- Existe no LFE a escrituração de valores que correspondem a centavos (abaixo de R$1), o que instiga a verificação para a constatação de eventual registro de valores incompatíveis com os verdadeiros valores das operações comerciais (ou prestação de serviços).

Os conjuntos discordes identificados servem como critério de escolha dos documentos ou registros sobre os quais haverá a incidência da observação exploratória por parte do auditor.

6.3.4 Resultados da Clusterização *K-Means*

Aplicada a clusterização pelo método *K-mean* aos valores contábeis em confronto com os seus respectivos valores de ICMS, ambos, consoantes os documentos de NFEs e os registros consignados no LFE, obtêm-se os gráficos em seguimento com k grupos = 6:

Em todos os gráficos apresentados é possível realizar:

- A **Análise Horizontal** dos dados apurando-se os conjuntos mais extremos e/ou raros que contemplam os maiores valores das operações.

Esses sobreditos agrupamentos merecem atenção por serem extraordinários, o que levanta a suspeita sobre a sua realidade.

- A **Análise Vertical** dos dados pesquisando-se pelas eventuais incompatibilidades na combinação linear *Imposto = Base de Cálculo x Alíquota possível*. Melhor dizendo, busca-se as divergências indicadas pelos pontos que não coincidem com uma das retas que representam as alíquotas existentes.

  Os desvios da linearidade esperada, propõem a suspeita de que os documentos deslocados foram emitidos com alíquotas divergentes.

Além disso é importante atentar para a necessária simetria que deverá persistir entre os valores destacados nas NFEs e as correspondentes escriturações no LFE. Não é sustentável a distorção entre esses valores.

No exemplo propostos obteve-se os seguintes resultados para a análise dos **dados das Entradas** (Figura 6.15 e Figura 6.16):

- **Análise Horizontal** - Observa-se a nítida configuração de um *cluster* (sexto agrupamento) deslocado, de alto valor e constituído exclusivamente por dois documentos/registros. Certamente essa circunstância deverá se examinada com maior critério pelo auditor, posto que pode representar uma operação irreal ou incompatível com a atividade econômica do contribuinte, tendo sido criada especialmente para o suporte de crédito impróprio.
- **Análise Vertical** - Verifica-se, em especial no terceiro e no quarto agrupamentos, a presença de pontos destoantes das linhas de combinação linear esperadas que reclamam investigação.
- Existe uma boa simetria entre as NFEs e o LFE, condição que não invoca a atenção para providências de exame.

Figura 6.15: *K-Means Clustering k=6* - LFE - ENTRADAS.

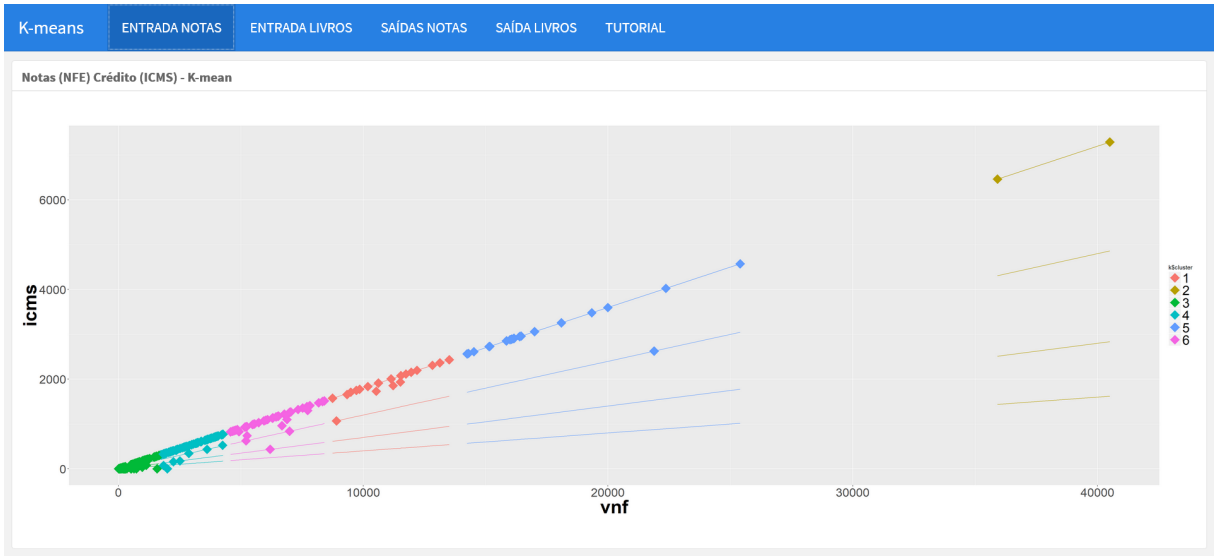

Figura 6.16: *K-Means Clustering k=6* - NFEs - ENTRADAS.

De igual sorte, alcança-se os seguintes resultados com os **dados das Saídas** (Figura 6.17 e Figura 6.18):

- **Análise Horizontal** - Sobrevêm a formação de dois agrupamentos de interesse para a fiscalização (o quinto e o sexto) por seu valor elevado e sua assimetria.
- **Análise Vertical** - Dá-se no terceiro, quarto e quinto *clusters*, a ocorrência de elementos discordes às retas de combinação linear pressupostas. Esses acontecimentos suscitam a atenção do auditor.

- Chama a atenção a falta de simetria entre os valores destacados nas NFEs e o seu respectivo registro no LFE, principalmente nos mais altos valores de operação. Essa condição requesta auditoria, já que pode representar a sonegação de vendas acontecidas e consequentemente a evasão do tributo (ICMS) devido.

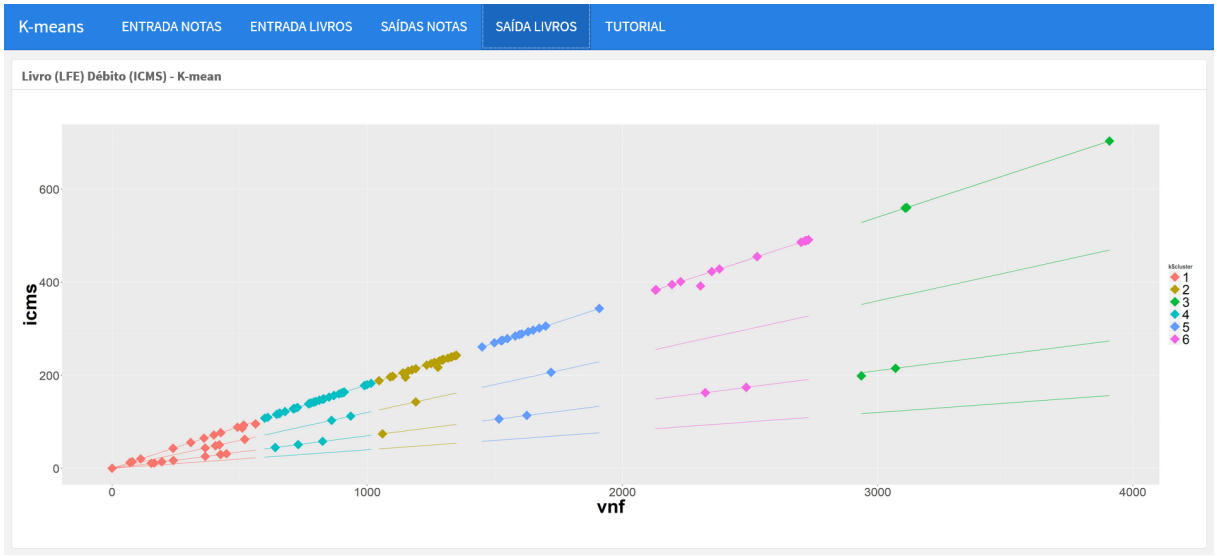

Figura 6.17: *K-Means Clustering k=6* - LFE - SAÍDAS.

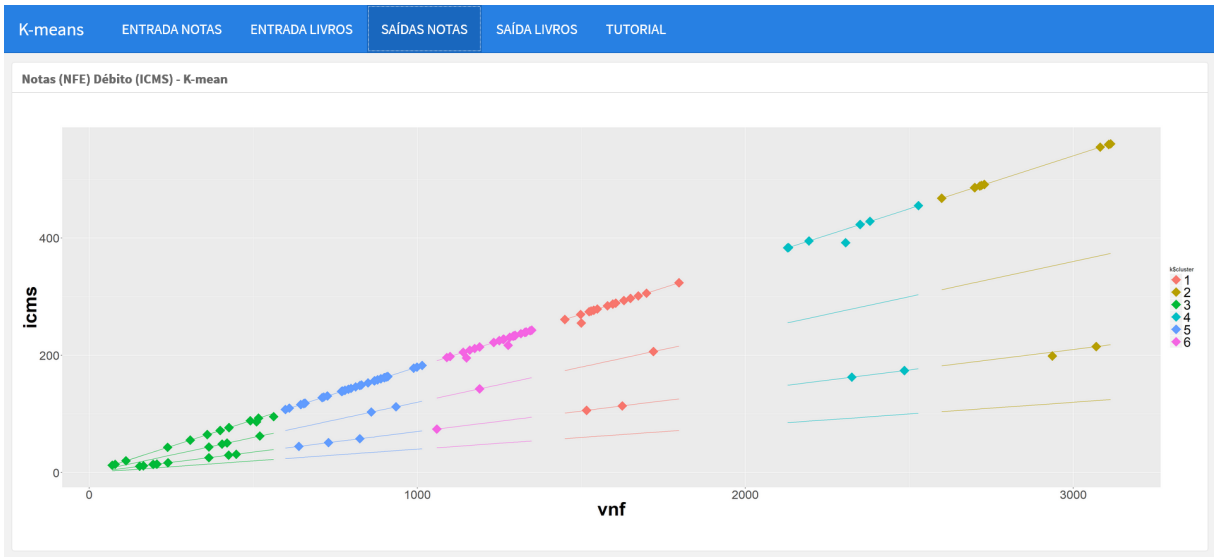

Figura 6.18: *K-Means Clustering k=6* - NFEs - SAÍDAS.

Capítulo 7

Conclusão e Resultados Alcançados

A proposta de metodologia formulada nesse trabalho demonstrou-se eficiente para a identificação (seleção) de empresas cujo comportamento fiscal merece averiguação de eventual práticas evasivas no cumprimento da obrigação tributária. Igual sucesso obteve na análise dos valores temporais dos contribuintes escolhidos, revelando os momentos temporais incomuns que solicitam investigação, bem assim os documentos e os lançamentos contábeis/fiscais que suscitam exame em sede de auditoria tributária.

Este estudo conciliou a aplicação de modelos estatístico-computacionais e métodos de mineração e análise de *outliers* úteis para a discriminação dos contribuintes no âmbito da programação fiscal, facilitando a detecção de empresas suspeitas de evasão tributária e corroborando para a melhor otimização da atividade de auditoria.

Bem assim, a proposta discutida cumpriu a realização dos objetivos geral e específicos desejados, a saber:

Satisfez o objetivo geral perseguido, porquanto construiu e implementou com êxito a aplicação de selecionados modelos estatísticos e de métodos de mineração de dados para o tratamento de *outliers*, que auxiliam a investigação de novas modalidades de evasão fiscal (sonegação e fraude) no comportamento tributário dos contribuintes do ICMS, reconhecendo e destacando as informações suspeitas.

Como resultado desse objetivo, a metodologia contribui para aperfeiçoar o *modus operandi* da atividade de programação fiscal, uma vez que ela proporciona a triagem criteriosa e objetiva dos contribuintes alvo, o tratamento de seus dados de interesse tributário e racionaliza o foco das ações fiscais.

Do mesmo modo, alcançou os objetivos específicos esperados, porquanto:

a) Aplicou, o método de programação matemática da Análise Envoltória de Dados (DEA) para aferir e diferenciar as empresas com desempenho relativo de arrecadação ineficientes, dentro de um segmento econômico, contribuintes elegíveis para um exame mais aprofundado.

b) Utilizou modelos estatísticos de análise de séries temporais para avaliação dos dados fiscais atinentes à apuração do imposto, diga-se: comparação gráfica dos valores reais sobre os escriturados; gráficos de boxplot; decomposição das componentes de tendência e sazonalidade; e método de projeção comparativa de alisamento pelo método Holtz-Winter, com o objetivo de detectar períodos de tempo (meses e anos) duvidosos (*outliers*).

c) Empregou técnicas estatísticas descritivas, probabilísticas e de clusterização *K-Means* para separar as informações, os valores, os registros escriturais e a documentação fiscal sob desconfiança (manifestamente extraordinárias) dos contribuintes selecionáveis para uma ação de fiscalização.

Para tanto, foi construído um recurso computacional que é capaz de extrair e processar os dados de interesse fiscal, aplicando-lhes os métodos e modelos designados e consolidando os resultados obtidos em painéis analíticos disponíveis para o uso do auditor. Esse engenho computacional potencializa as atividades de investigação tributária, uma vez que proporciona a disposição dos resultados das empresas examinadas, momentos temporais e valores escriturais *outliers* que devem ser avaliados em ação de fiscalização, poupando muito tempo de análise.

Nesse sentido, a identificação das circunstâncias anômalas, merecedoras de destaque, a partir de um tratamento sistemático dos dados disponíveis para a Receita, proporciona para a auditoria um valioso elenco de informações acuradas sobre prováveis desvios fiscal, condição que dará maior eficiência às investigações, porquanto as orienta os para eventos com a maior probabilidade de êxito.

Consigne-se que a metodologia apresentada já contribuiu para o incremento dos resultados de sucesso no combate à sonegação do ICMS no Distrito Federal - Brasil (DF). Como resultado de sua aplicação foi possível identificar mais de uma centena de contribuintes que ofereciam práticas fiscais extravagantes sendo as respectivas empresas devidamente selecionadas para auditoria no ano de 2018.

A redução do risco de sucesso nas tentativas de empreender novas práticas evasão fiscal, por meio da sua elucidação diligente, foi o ideal de mérito perseguidos por esse estudo. Tal competência já pode ser comemorada, porquanto hoje a metodologia é aplicada:

a) Na Gerência de Programação Fiscal e Controle de Operações (GEPRO) - setorial responsável pela programação de auditorias e ações de fiscalização -, como instrumento analítico de pesquisa das hipóteses de evasão, sonegação e fraude com resultado proficiente na construção de projetos de fiscalização.

b) Na Gerência de Monitoramento e Auditorias Especiais - divisão da Receita Distrital que monitora setores estratégicos (*v.g.* Telecomunicação, Energia, Combustíveis,

Substituição Tributária), como *baseline* das conferências analíticas sob a sua competência. Neste caso a metodologia teve de ser adaptado para prover seus resultados no *software* QlickView[24]que é o instrumento de tratamento de dados preferível pelo setor.

c) Como técnica embrionária do Grupo de Trabalho de Inteligência Artificial e Ciência de Dados (coordenado pelo autor) que tem a missão de difundir e fazer evoluir as soluções e inovações na Receita Distrital.

Além de ter cumprido o requisito de patrocinar um direcionamento otimizado para a atividade de inspeção fiscal, economizando tempo e trabalho de auditoria, esse estudo deverá sofrer continuada evolução, e motivar outros estudos e pesquisas no desiderato de incorporar novos modelos, métodos, técnicas e ferramentas analíticas aproveitáveis ao combate à evasão de tributos.

7.1 Dos Trabalhos Futuros

A metodologia apresentada merece aperfeiçoamento, permitindo a adesão de novas técnicas, modelos, métodos e soluções analíticas importantes para a evolução das atividades de auditoria e de persecução das evasões tributárias no ICMS.

Diversas outras proposições de observação de anomalias (*outliers*) dignas de notório mérito não foram tratadas nesse trabalho e poderão ser desenvolvidas como recomendação evolutiva. Consignamos alguns exemplos de técnicas que no futuro próximo deverão ser adaptadas e incluídas na metodologia apresentada:

- Redes Preditivas Neurais, aliadas a técnicas de *Deep Learning*, para monitorar o desempenho fiscal das empresas contribuintes. A avaliação probabilística das informações sobre as características cadastrais (*v.g.* geográficas e societárias), setoriais (do segmento econômico de atuação) e as fiscais de suas operações comerciais (*v.g.* alíquotas operadas, origens dos estoques, etc.) de um contribuinte pode sinalizar (não determinístico) a esperança matemática de sua eventual conduta evasiva. Em função dessa perspectiva é possível construir níveis de alerta para as intervenções da fiscalização (*v.g.* eleger uma remessa com origem/destino a contribuinte suspeito, parar o transporte e conferir a regularidade da carga).

- Regressões multivariadas aplicada como ferramenta preditiva servível para apurar o grau de descolamento entre os valores dos resultados esperados (previstos) e os dados

[24] *QlickView* uma ferramenta de *Bussiness Inteligence* desenvolvido pela empresa Qlick. É uma plataforma de *Business Discovery* que oferece BI de autoatendimento para todos os usuários de negócios.

realizados pelo contribuinte, contituindo uma medida de apuração de dispersões *outliers*.

- Mineração de texto da descrição dos produtos assentado em documentos fiscais (NFEs), para conferir a compatibilidade entre a alíquota devida para o produto e aquela aplicada na operação.
- Análise de Redes Complexas de descrição da formação (origens) dos créditos do imposto (ICMS), que ilustra graficamente todos os relacionamentos (e seus pesos) de uma empresa com os seus fornecedores, permitindo estudar as relações espúrias criadas para triangular e circularizar suportes de créditos falsos.

Referências

[1] *Portaria sefp nº 130 , de 3 de março*, 1997. Diário Oficial do Distrito Federal, Brasília. 2

[2] *Constituição da república federativa do brasil*, 1988. 10, 11, 12

[3] *Constituição da república dos estados unidos do brasil*, 1934. 10

[4] *Emenda constitucional nº 18*, 1965. 11

[5] Silva, Joseph de Plácido e: *Vocabulário Jurídico*, volume 3. Forense, 1991, ISBN 978-8-530-96060-5. 12

[6] *Lei nº 4.502, de 30 de novembro*, 1964. Diário Oficial da União, Brasília,. 14, 15

[7] *Código tributário nacional, lei nº 5.172, de 25 de outubro*, 1966. Diário Oficial da União, Brasília. 14, 19, 21, 43

[8] *Decreto-lei nº 34, de 18 de novembro*, 1966. Diário Oficial da União, Brasília. 14

[9] *Código tributário do distrito federal, lei complementar nº 4 de 30 de dezembro*, 1994. Diário Oficial do Distrito Federal, Brasília. 15

[10] *Lei nº 8.137, de 27 de dezembro*, 1990. Diário Oficial da União, Brasília. 16

[11] *Código penal brasileiro, decreto-lei nº 2.848, de 7 de dezembro*, 1940. Diário Oficial da União, Brasília. 21

[12] *Lei complementar distrital nº 772, de 17 de julho*, 2008. Diário Oficial do Distrito Federal, Brasília. 22

[13] Charnes, A., W.W. Cooper e E. Rhodes: *Measuring the efficiency of decision making units*. European Journal of Operational Research, (2):429–444, 1978. 27, 29, 31

[14] Banker, R. D., A. Charnes e W. W. Cooper: *Some models for estimating technical and scale inefficiencies in data envelopment analysis*. Management Science, 30(9):1078–1092, 1984, ISSN 0025-1909, 1526-5501. 27, 30

[15] Seiford, Lawrence M. e Robert M. Thrall: *Recent developments in DEA*. Journal of Econometrics, 46(1):7–38, 1990, ISSN 03044076. 27

[16] Cooper, William W., Lawrence M. Seiford e Joe Zhu: *Handbook on Data Envelopment Analysis*. Springer Science & Business Media, 2011, ISBN 978-1-4419-6151-8. 27, 29

[17] Ferreira, Carlos de Carvalho: *Introdução à análise envoltória de dados: teoria, modelos e aplicações*. UFV, 2009. 28, 30

[18] Casu, Barbara, Claudia Girardone e Philip Molyneux: *Productivity change in european banking: A comparison of parametric and non-parametric approaches*. Journal of Banking & Finance, 28(10):2521–2540, 2004, ISSN 03784266. 28

[19] H, Ibrahim, Osman: *Handbook of Research on Strategic Performance Management and Measurement Using Data Envelopment Analysis*. IGI Global, 2013, ISBN 978-1-4666-4475-5. 28

[20] Zhu, Joe: *Data Envelopment Analysis: A Handbook of Empirical Studies and Applications*. Springer, 2016, ISBN 978-1-4899-7684-0. 28, 29

[21] Zhu, Joe: *Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets*. Springer, 2014, ISBN 978-3-319-06647-9. 28

[22] Zhu, Joe: *Data envelopment analysis: let the data speak for themselves*. Amazon Distribution, 2014, ISBN 978-1-4975-9134-9. OCLC: 884905095. 28

[23] Emrouznejad, Ali: *Advances in data envelopment analysis*. Annals of Operations Research, 214(1):1–4, 2014, ISSN 0254-5330, 1572-9338. 29, 30

[24] Cook, Wade D. e Larry M. Seiford: *Data envelopment analysis (DEA) – thirty years on*. European Journal of Operational Research, 192(1):1–17, 2009, ISSN 03772217. 29

[25] KASSAI, Sílvia: *Utilização da análise por envoltória de dados (DEA) na análise de demonstrações contábeis*, 2002. 29

[26] Tone, Kaoru: *Advances in DEA Theory and Applications: With Extensions to Forecasting Models*. John Wiley & Sons, 2017, ISBN 978-1-118-94562-9. 29, 30

[27] Wanke, Peter, C.P. Barros e Ali Emrouznejad: *Assessing productive efficiency of banks using integrated fuzzy-DEA and bootstrapping: A case of mozambican banks*. European Journal of Operational Research, 249(1):378–389, 2016, ISSN 03772217. 29

[28] Safdar, Komal A., Ali Emrouznejad e Prasanta K. Dey: *Assessing the queuing process using data envelopment analysis: an application in health centres*. Journal of Medical Systems, 40(1), 2016, ISSN 0148-5598, 1573-689X. 29

[29] Gholami, Roya, Dolores Añón Higón e Ali Emrouznejad: *Hospital performance: Efficiency or quality? can we have both with IT?* Expert Systems with Applications, 42(12):5390–5400, 2015, ISSN 09574174. 29

[30] Charnes, Abraham, William W. Cooper, Arie Y. Lewin e Lawrence M. Seiford: *Data Envelopment Analysis: Theory, Methodology, and Applications*. Springer Science & Business Media, 2013, ISBN 978-94-011-0637-5. 29

[31] Coelli, Tim: *A multi-stage methodology for the solution of orientated DEA models.* Operations Research Letters, 23(3):143–149, 1998, ISSN 01676377. 30

[32] Ghasemi, M. R., Joshua Ignatius, Sebastián Lozano, Ali Emrouznejad e Adel Hatami-Marbini: *A fuzzy expected value approach under generalized data envelopment analysis.* Knowledge-Based Systems, 89:148–159, 2015, ISSN 09507051. 30

[33] Chandola, V., V. Mithal e V. Kumar: *Comparative evaluation of anomaly detection techniques for sequence data.* Em *2008 Eighth IEEE International Conference on Data Mining*, páginas 743–748, 2008. 43, 44

[34] Chandola, V., A. Banerjee e V. Kumar: *Anomaly detection for discrete sequences: A survey.* IEEE Transactions on Knowledge and Data Engineering, 24(5):823–839, 2012, ISSN 1041-4347. 43, 44

[35] Gupta, Manish, Jing Gao, Charu Aggarwal e Jiawei Han: *Outlier detection for temporal data.* Synthesis Lectures on Data Mining and Knowledge Discovery, 5(1):1–129, 2014, ISSN 2151-0067. 43, 44

[36] Hamilton, James Douglas: *Time Series Analysis.* Princeton University Press, 1994, ISBN 978-0-691-04289-3. 44

[37] Brockwell, Peter J. e Richard A. Davis: *Introduction to Time Series and Forecasting.* Springer Science & Business Media, 2006, ISBN 978-0-387-21657-7. 44

[38] Cryer, Jonathan D. e Kung Sik Chan: *Time Series Analysis: With Applications in R.* Springer Science & Business Media, 2008, ISBN 978-0-387-75958-6. 44

[39] Shumway, Robert H. e David S. Stoffer: *Time Series Analysis and Its Applications: With R Examples.* Springer Science & Business Media, 2010, ISBN 978-1-4419-7865-3. 44

[40] Cowpertwait, Paul S. P. e Andrew V. Metcalfe: *Introductory Time Series with R.* Springer Science & Business Media, 2009, ISBN 978-0-387-88698-5. 44

[41] Aggarwal, C.: *On abnormality detection in spuriously populated data streams.* Em *Proceedings of the 2005 SIAM International Conference on Data Mining*, Proceedings, páginas 80–91. Society for Industrial and Applied Mathematics, 2005, ISBN 978-0-89871-593-4. DOI: 10.1137/1.9781611972757.8. 44

[42] Aggarwal, Charu C.: *Outlier Analysis.* Springer, 2016, ISBN 978-3-319-47578-3. 44, 45, 69, 70

[43] Jiang, Ruoyi, Hongliang Fei e Jun Huan: *Anomaly localization for network data streams with graph joint sparse PCA.* página 886. ACM Press, 2011, ISBN 978-1-4503-0813-7. 44

[44] Papadimitriou, Spiros, Jimeng Sun e Christos Faloutsos: *Streaming pattern discovery in multiple time-series.* Em *VLDB*, 2005. 44

[45] Keogh, Eamonn, Jessica Lin, Sang Hee Lee e Helga Van Herle: *Finding the most unusual time series subsequence: algorithms and applications.* Knowledge and Information Systems, 11(1):1–27, 2006, ISSN 0219-1377, 0219-3116. 44

[46] Jeong, Young Seon, Myong K. Jeong e Olufemi A. Omitaomu: *Weighted dynamic time warping for time series classification.* Pattern Recognition, 44(9):2231–2240, 2011, ISSN 00313203. 44, 45

[47] Mueen, Abdullah, Eamonn Keogh e Neal Young: *Logical-shapelets: an expressive primitive for time series classification.* página 1154. ACM Press, 2011, ISBN 978-1-4503-0813-7. 44

[48] Ye, Lexiang e Eamonn Keogh: *Time series shapelets: a novel technique that allows accurate, interpretable and fast classification.* Data Mining and Knowledge Discovery, 22(1):149–182, 2011, ISSN 1384-5810, 1573-756X. 44

[49] Al-Khateeb, Tahseen, Mohammad M. Masud, Latifur Khan, Charu Aggarwal, Jiawei Han e Bhavani Thuraisingham: *Stream classification with recurring and novel class detection using class-based ensemble.* páginas 31–40. IEEE, 2012, ISBN 978-1-4673-4649-8 978-0-7695-4905-7. 45

[50] Masud, Mohammad M., Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, Ashok Srivastava e Nikunj C. Oza: *Classification and adaptive novel class detection of feature-evolving data streams.* IEEE Transactions on Knowledge and Data Engineering, 25(7):1484–1497, 2013, ISSN 1041-4347. 45

[51] Cheng, H., P. Tan, C. Potter e S. Klooster: *Detection and characterization of anomalies in multivariate time series.* Em *Proceedings of the 2009 SIAM International Conference on Data Mining*, Proceedings, páginas 413–424. Society for Industrial and Applied Mathematics, 2009, ISBN 978-0-89871-682-5. DOI: 10.1137/1.9781611972795.36. 45

[52] Tsay, R. S.: *Outliers in multivariate time series.* Biometrika, 87(4):789–804, 2000, ISSN 0006-3444, 1464-3510. 45

[53] Baragona, Roberto e Francesco Battaglia: *Outliers detection in multivariate time series by independent component analysis.* Neural Computation, 19(7):1962–1984, 2007. 45

[54] Keogh, Eamonn, Stefano Lonardi e Bill 'Yuan chi' Chiu: *Finding surprising patterns in a time series database in linear time and space.* página 550. ACM Press, 2002, ISBN 978-1-58113-567-1. 45

[55] Burdakis, Sabbas e Antonios Deligiannakis: *Detecting outliers in sensor networks using the geometric approach.* páginas 1108–1119. IEEE, 2012, ISBN 978-0-7695-4747-3 978-1-4673-0042-1. 45

[56] Branch, Joel W., Chris Giannella, Boleslaw Szymanski, Ran Wolff e Hillol Kargupta: *In-network outlier detection in wireless sensor networks.* Knowledge and Information Systems, 34(1):23–54, 2013, ISSN 0219-1377, 0219-3116. 45

[57] Franke, Conny e Michael Gertz: *ORDEN: outlier region detection and exploration in sensor networks.* página 1075. ACM Press, 2009, ISBN 978-1-60558-551-2. 45

[58] Zhang Y., Meratnia N., Havinga P.: *Outlier detection techniques for wireless sensor networks: A survey.* 12(2):159–170, 2010, ISSN 1553-877X. 45

[59] Kuncheva, Ludmila I.: *Change detection in streaming multivariate data using likelihood detectors.* IEEE Transactions on Knowledge and Data Engineering, 25(5):1175–1180, 2013, ISSN 1041-4347. 45

[60] Song, Xiuyao, Mingxi Wu, Christopher Jermaine e Sanjay Ranka: *Statistical change detection for multi-dimensional data.* página 667. ACM Press, 2007, ISBN 978-1-59593-609-7. 45

[61] Szmit, Maciej e Anna Szmit: *Usage of modified holt-winters method in the anomaly detection of network traffic: Case studies.* 2012. DOI: 10.1155/2012/192913. 45

[62] Salem, Osman, Yaning Liu e Ahmed Mehaoua: *Detection and isolation of faulty measurements in medical wireless sensor networks.* Future Information and Communication Technologies for Ubiquitous HealthCare (Ubi-HealthTech), 2013 First International Symposium on, 2013, ISSN 978-1-4799-0766-3. 45

[63] Hawkins, D.: *Identification of Outliers.* Springer Science & Business Media, 2013, ISBN 978-94-015-3994-4. 69

[64] Tan, Pang Ning, Michael Steinbach e Vipin Kumar: *Introduction to Data Mining.* Pearson Addison Wesley, 2006, ISBN 978-0-321-32136-7. 70

[65] Chandola, Varun, Arindam Banerjee e Vipin Kumar: *Anomaly detection: A survey.* ACM Comput. Surv., 41(3):15:1–15:58, 2009, ISSN 0360-0300. 70

[66] Bakar, Zuriana, Rosmayati Mohemad, Akbar Ahmad e Mustafa Deris: *A comparative study for outlier detection techniques in data mining.* páginas 1–6. IEEE, 2006, ISBN 978-1-4244-0022-5 978-1-4244-0023-2. 70

[67] Zhang, Ji: *Advancements of outlier detection: A survey.* ICST Transactions on Scalable Information Systems, 13(1):e2, 2013, ISSN 2032-9407. 70

[68] Malik, Kamal, H. Sadawarti e Kalra G. S: *Comparative analysis of outlier detection techniques.* International Journal of Computer Applications, 97(8):12–21, 2014, ISSN 09758887. 70

[69] James, Gareth, Daniela Witten, Trevor Hastie e Robert Tibshirani: *An Introduction to Statistical Learning: with Applications in R.* Springer Science & Business Media, 2013, ISBN 978-1-4614-7138-7. 70

[70] Witten, Ian H., Eibe Frank, Mark A. Hall e Christopher J. Pal: *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2016, ISBN 978-0-12-804357-8. 70

[71] Han, Jiawei, Jian Pei e Micheline Kamber: *Data Mining: Concepts and Techniques*. Elsevier, 2011, ISBN 978-0-12-381480-7. 70

[72] Hastie, Trevor, Robert Tibshirani e Jerome Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2013, ISBN 978-0-387-21606-5. 70

[73] Aggarwal, Charu C.: *Data Mining: The Textbook*. Springer, 2015, ISBN 978-3-319-14142-8. 70

[74] Oliveira, Cledson D., Adhemar Apparecido De Caroli, Amaury de Souza Amaral e Omar L. Vilca: *Detecção de fraudes, anomalias e erros em análise de dados contábeis: Um estudo com base em outliers*. Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos (REDECA), 1(1):102–127, 2014, ISSN 2446-9513. 70, 71

[75] Knox, Edwin M e Raymond T Ng: *Algorithms for mining distancebased outliers in large datasets*. Em *Proceedings of the International Conference on Very Large Data Bases*, páginas 392–403, 1998. 70

[76] Wu, Junjie: *Advances in K-means Clustering: A Data Mining Thinking*. Springer Science & Business Media, 2012, ISBN 978-3-642-29807-3. 70

[77] Baesens, Bart, Veronique Van Vlasselaer e Wouter Verbeke: *Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection*. John Wiley & Sons, 2015, ISBN 978-1-119-14683-4. 71

[78] Bolton, Richard J. e David J. Hand: *Statistical fraud detection: A review*. Statistical Science, 17(3):235–249, 2002, ISSN 0883-4237. 71

[79] Nigrini, Mark J.: *Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations*. John Wiley & Sons, 2011, ISBN 978-0470890462. 71

[80] Nigrini, Mark J. e Joseph T. Wells: *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. John Wiley & Sons, 2011, ISBN 978-1118152850. 71

[81] Santos, Josenildo dos, José Francisco Ribeiro Filho, Umbelina Lagioia, Bartolomeu Figueiredo Alves Filho e Ivson José Caldas de Araújo: *Aplicações da lei de newcomb-benford na auditoria tributária do imposto sobre serviços de qualquer natureza (iss)*. Revista de Contabilidade e Finanças., 20(49), 2009, ISSN 1808-057X. 71

Anexo I

Painéis Analíticos - Códigos em Linguagem SQL

PAINEL I - DEA:

```
DROP TABLE DEA_FINAL PURGE;
DROP TABLE DEA_CAD PURGE;
DROP TABLE DEA_SAI PURGE;
DROP TABLE DEA_ENT PURGE;
DROP TABLE DEA_ARR PURGE;

--- DADOS CADASTRO

CREATE TABLE DEA_CAD AS(
SELECT
    c.inscricao,
    c.cpfcnpj,
    c.nome_razao,
    c.desc_atvd_icms,
    round(months_between('31/12/2017',c.data_inscricao),0) AS meses
FROM
    zz_contribuintes c
WHERE
    c.desc_situacao = 'Ativo'
AND
    c.dsc_forma_calculo_icms = 'Normal'
AND
    substr(c.data_forma_cal_icms,7,4) < 2017
```

```
GROUP BY
    c.inscricao,
    c.cpfcnpj,
    c.nome_razao,
    c.desc_atvd_icms,
    round(months_between('31/12/2017',c.data_inscricao),0)
);

--- DADOS SAÍDAS

CREATE TABLE DEA_SAI AS(
SELECT
    u.cpfcnpj,
    SUM(n.vicms) AS icms,
    SUM(n.vnf) AS vnf
FROM
    z_nfe_2017 n
    INNER JOIN DEA_cad u ON u.cpfcnpj = n.emitente
WHERE
    n.cod_sit = '100'
GROUP BY
    u.cpfcnpj
);

--- DADOS ENTRADAS

CREATE TABLE DEA_ENT AS(
SELECT
    u.cpfcnpj,
    SUM(n.vicms) AS icms,
    SUM(n.vnf) AS vnf
FROM
    z_nfe_2017 n
    INNER JOIN DEA_cad u ON u.cpfcnpj = n.destinatario
WHERE
    n.cod_sit = '100'
GROUP BY
    u.cpfcnpj
);
```

```
--- DADOS ARRECADAÇÃO

CREATE TABLE DEA_ARR AS(
SELECT
    u.cpfcnpj,
    SUM(f.vlprinc) AS arrec
FROM
    DEA_cad u
    INNER JOIN Z_FINANCEIRO_2_18 f ON u.cpfcnpj = f.cnpj
WHERE
    f.exercicio = '2017'
AND
    f.codigo = '1317'
GROUP BY
    u.cpfcnpj
);

--- DADOS CARTÕES

CREATE TABLE DEA_CARD AS(
SELECT
    u.cpfcnpj,
    ( SUM(
        nvl(c.total_credito,0)
    ) ) + ( SUM(
        nvl(c.total_debito,0)
    ) ) AS card
FROM
    DEA_cad u
    INNER JOIN zz_cartao_contrib_full c ON u.cpfcnpj = c.num_cpf_cnpj
WHERE
    c.ano = '2017'
GROUP BY
    u.cpfcnpj
);

/* DADOS CONSOLIDADOS */

CREATE TABLE DEA_FINAL AS (
SELECT
```

```
    uc.inscricao,
    uc.cpfcnpj,
    uc.nome_razao,
    uc.desc_atvd_icms,
    uc.meses,
    round( (nvl(ua.arrec,0) / 1000),0) AS arrecadacao,
    round(
        ( (nvl(us.icms,0) / 1000) - (nvl(unb_dea_ent.icms,1) / 1000) ),
        0
    ) AS saldo_icms,
    round( (nvl(unb_dea_ent.vnf,0) / 1000),0) AS compras_nent_vnf,
    round( (nvl(us.vnf,0) / 1000),0) AS vendas_nsai_vnf,
    round( (nvl(unb_dea_card.card,0) / 1000),0) AS vendas_cartao
FROM
    DEA_cad uc
LEFT JOIN
    DEA_sai us
ON uc.cpfcnpj = us.cpfcnpj LEFT JOIN
    DEA_ent
ON uc.cpfcnpj = DEA_ent.cpfcnpj LEFT JOIN
    DEA_arr ua
ON uc.cpfcnpj = ua.cpfcnpj LEFT JOIN
    DEA_card
ON uc.cpfcnpj = DEA_card.cpfcnpj GROUP BY
    uc.inscricao,
    uc.cpfcnpj,
    uc.nome_razao,
    uc.desc_atvd_icms,
    uc.meses,
    ua.arrec,
    DEA_ent.icms,
    us.icms,
    DEA_ent.vnf,
    us.vnf,
    DEA_card.card
);

--- DADOS SEM MOVIMENTO

DELETE FROM DEA_FINAL
```

```
WHERE
(vendas_nsai_vnf +
VENDAS_CARTAO < 1000)
OR ARRECADACAO = 0;

---  CNAE <5

DELETE FROM DEA_FINAL
WHERE  desc_atvd_icms in (
SELECT
    desc_atvd_icms
FROM
    DEA_FINAL
GROUP BY
    desc_atvd_icms
HAVING
    COUNT(desc_atvd_icms) < 4);

--- DELETAR TABELAS INTERMEDIÁRIAS

DROP TABLE DEA_CAD PURGE;
DROP TABLE DEA_SAI PURGE;
DROP TABLE DEA_ENT PURGE;
DROP TABLE DEA_ARR PURGE;
DROP TABLE DEA_CARD PURGE;
```

PAINEL II - SÉRIES TEMPORAIS:

```
DROP TABLE st_lfe_e PURGE;
DROP TABLE st_lfe_s PURGE;
DROP TABLE st_nfe_s PURGE;
DROP TABLE  st_nfe_e PURGE;

--- ST LFE ENTRADA

CREATE TABLE st_lfe_e AS(
SELECT
    o000.ie,
    o000.cnpj,
    20
     ||  substr(o000.dt_ini,7,2) AS ano,
    substr(o000.dt_ini,4,2) AS mes,
    e330.ind_tot,
    e330.vl_cont,
    e330.vl_bc_icms,
    e330.vl_icms,
    e330.vl_st,
    e330.vl_compl,
    e330.vl_isnt_icms,
    e330.vl_out_icms
FROM
    admlivro.le_registro_0000@link_lfe o000
    INNER JOIN admlivro.le_registro_e001@link_lfe e001 ON (
        o000.sequencial = e001.id_pai
    )
    INNER JOIN admlivro.le_registro_e300@link_lfe e300 ON (
        e001.sequencial = e300.id_pai
    )
    INNER JOIN admlivro.le_registro_e330@link_lfe e330 ON (
        e300.sequencial = e330.id_pai
    )
    INNER JOIN dea_final c ON o000.ie = c.inscricao
WHERE
    e330.ind_tot = 4
AND
```

```
    substr(o000.dt_ini,7,2) > '09'
GROUP BY
    o000.ie,
    o000.cnpj,
    20
     ||  substr(o000.dt_ini,7,2),
    substr(o000.dt_ini,4,2),
    e330.ind_tot,
    e330.vl_cont,
    e330.vl_bc_icms,
    e330.vl_icms,
    e330.vl_st,
    e330.vl_compl,
    e330.vl_isnt_icms,
    e330.vl_out_icms)
ORDER BY
    o000.ie,
    ano,
    mes;

--- ST LFE SAÍDA

CREATE TABLE st_lfe_s AS(
SELECT
    o000.ie,
    o000.cnpj,
    20
     ||  substr(o000.dt_ini,7,2) AS ano,
    substr(o000.dt_ini,4,2) AS mes,
    e330.ind_tot,
    e330.vl_cont,
    e330.vl_bc_icms,
    e330.vl_icms,
    e330.vl_st,
    e330.vl_compl,
    e330.vl_isnt_icms,
    e330.vl_out_icms
FROM
    admlivro.le_registro_0000@link_lfe o000
    INNER JOIN admlivro.le_registro_e001@link_lfe e001 ON (
```

```
        o000.sequencial = e001.id_pai
    )
    INNER JOIN admlivro.le_registro_e300@link_lfe e300 ON (
        e001.sequencial = e300.id_pai
    )
    INNER JOIN admlivro.le_registro_e330@link_lfe e330 ON (
        e300.sequencial = e330.id_pai
    )
    INNER JOIN dea_final c ON o000.ie = c.inscricao
WHERE
    e330.ind_tot = 8
AND
    substr(o000.dt_ini,7,2) > '09'
GROUP BY
    o000.ie,
    o000.cnpj,
    20
     ||  substr(o000.dt_ini,7,2),
    substr(o000.dt_ini,4,2),
    e330.ind_tot,
    e330.vl_cont,
    e330.vl_bc_icms,
    e330.vl_icms,
    e330.vl_st,
    e330.vl_compl,
    e330.vl_isnt_icms,
    e330.vl_out_icms)
ORDER BY
    o000.ie,
    ano,
    mes;

--- ST NFE SAÍDA

CREATE TABLE st_nfe_s AS(
SELECT
    u.inscricao,
    u.cpfcnpj,
    z.ano,
    z.mes,
```

```
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2017_sum_emit z --- JÁ SÓ AS VÁLIDAS
    INNER JOIN dea_final u ON z.emitente = u.cpfcnpj
GROUP BY
    u.inscricao,
    u.cpfcnpj,
    z.ano,
    z.mes
);

INSERT INTO st_nfe_s
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
    SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2016_sum_emit z
    INNER JOIN  dea_final u ON z.emitente =  u.cpfcnpj
GROUP BY
    u.inscricao,
    u.cpfcnpj,
    z.ano,
    z.mes;

INSERT INTO st_nfe_s
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
    SELECT
     u.inscricao,
```

```
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2015_sum_emit z
    INNER JOIN  dea_final u ON z.emitente =  u.cpfcnpj
GROUP BY
    u.inscricao,
    u.cpfcnpj,
    z.ano,
    z.mes;

INSERT INTO st_nfe_s
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
    SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2014_sum_emit z
    INNER JOIN  dea_final u ON z.emitente =  u.cpfcnpj
GROUP BY
    u.inscricao,
    u.cpfcnpj,
    z.ano,
    z.mes;

INSERT INTO st_nfe_s
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
```

```
    SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2013_sum_emit z
    INNER JOIN  dea_final u ON z.emitente =  u.cpfcnpj
GROUP BY
    u.inscricao,
    u.cpfcnpj,
    z.ano,
    z.mes;

INSERT INTO st_nfe_s
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
    SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2012_sum_emit z
    INNER JOIN  dea_final u ON z.emitente =  u.cpfcnpj
GROUP BY
    u.inscricao,
    u.cpfcnpj,
    z.ano,
    z.mes;
```

```
 INSERT INTO st_nfe_s
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
    SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2011_sum_emit z
    INNER JOIN  dea_final u ON z.emitente =  u.cpfcnpj
GROUP BY
    u.inscricao,
    u.cpfcnpj,
    z.ano,
    z.mes;

 INSERT INTO st_nfe_s
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
    SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2010_sum_emit z
    INNER JOIN  dea_final u ON z.emitente =  u.cpfcnpj
GROUP BY
    u.inscricao,
    u.cpfcnpj,
    z.ano,
```

```
    z.mes;


-- DESTINATARIO

CREATE TABLE st_nfe_e AS(
SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2017_sum_dest z
    INNER JOIN  dea_final u ON z.destinatario =  u.cpfcnpj
GROUP BY
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes);

INSERT INTO st_nfe_e
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
   SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2016_sum_dest z
    INNER JOIN  dea_final u ON z.destinatario =  u.cpfcnpj
```

```
GROUP BY
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes;

INSERT INTO st_nfe_e
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
   SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2015_sum_dest z
    INNER JOIN  dea_final u ON z.destinatario =  u.cpfcnpj
GROUP BY
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes;

INSERT INTO st_nfe_e
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
   SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
```

```
    zz_nfe_2014_sum_dest z
    INNER JOIN  dea_final u ON z.destinatario =  u.cpfcnpj
GROUP BY
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes;

INSERT INTO st_nfe_e
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
    SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2013_sum_dest z
    INNER JOIN  dea_final u ON z.destinatario =  u.cpfcnpj
GROUP BY
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes;

INSERT INTO st_nfe_e
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
    SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
```

```
    SUM( z.vst) AS vst
FROM
    zz_nfe_2012_sum_dest z
    INNER JOIN  dea_final u ON z.destinatario =  u.cpfcnpj
GROUP BY
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes;

INSERT INTO st_nfe_e
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
    SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2011_sum_dest z
    INNER JOIN  dea_final u ON z.destinatario =  u.cpfcnpj
GROUP BY
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes;


INSERT INTO st_nfe_e
    (inscricao, cpfcnpj, ano, mes, vnf, vbc, vicms, vbcst, vst)
    SELECT
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes,
   SUM( z.vnf) AS vnf,
```

```
   SUM( z.vbc) AS vbc,
   SUM( z.vicms) AS vicms,
   SUM( z.vbcst) AS vbcst,
    SUM( z.vst) AS vst
FROM
    zz_nfe_2010_sum_dest z
    INNER JOIN  dea_final u ON z.destinatario =  u.cpfcnpj
GROUP BY
     u.inscricao,
     u.cpfcnpj,
    z.ano,
    z.mes;

  --- TABELA FINAL

CREATE TABLE st_FINAL AS( SELECT
    lfe_e.ie,
    lfe_e.ano,
    lfe_e.mes,
    round(nvl(lfe_e.vl_icms,0) / 1000,0) AS lfe_e_icms,
    round(nvl(lfe_s.vl_icms,0) / 1000,0) AS lfe_s_icms,
    round(nvl(nfe_e.vicms,0) / 1000,0) AS nfe_e_icms,
    round(nvl(nfe_s.vicms,0) / 1000,0) AS nfe_s_icms,
    lfe_e.cnpj
FROM
    st_lfe_e lfe_e
FULL JOIN
    st_lfe_s lfe_s
ON
    lfe_e.ie = lfe_s.ie
AND
    lfe_e.ano = lfe_s.ano
AND
    lfe_e.mes = lfe_s.mes
FULL JOIN
    st_nfe_e nfe_e
ON
    lfe_s.ie = nfe_e.inscricao
AND
    lfe_s.ano = nfe_e.ano
```

```
AND
    lfe_s.mes = nfe_e.mes
FULL JOIN
    st_nfe_s nfe_s
ON
    nfe_e.inscricao = nfe_s.inscricao
AND
    nfe_e.ano = nfe_s.ano
AND
    nfe_e.mes = nfe_s.mes
WHERE
    lfe_e.ano <= '2017'
) ORDER BY
    lfe_e.ie,
    lfe_e.ano,
    lfe_e.mes;
```

PAINEL III - MINERAÇÃO DE DADOS, NEWCOMB-BENFORD e K-MEANS:

```
DROP TABLE dea_final PURGE;
DROP TABLE MIN_nfe17_s PURGE;
DROP TABLE MIN_nfe17_e PURGE;
DROP TABLE MIN_lfe17_s PURGE;
DROP TABLE MIN_lfe17_e PURGE;

CREATE TABLE MIN_nfe17_s AS(
SELECT
   z.emitente,
    z.nnf,
    z.mes,
    z.ano,
    z.vnf,
    z.vicms
FROM
    z_nfe_2017 z
    INNER JOIN dea_final u ON z.emitente = u.cpfcnpj
WHERE
    z.cod_sit = 100
GROUP BY
    z.emitente,
    z.nnf,
    z.mes,
    z.ano,
    z.vnf,
    z.vicms
    );

CREATE TABLE MIN_nfe17_e AS(
SELECT
    z.nnf,
    z.mes,
    z.ano,
    z.vnf,
    z.vicms,
    z.destinatario
```

```
FROM
    z_nfe_2017 z
    INNER JOIN dea_final u ON z.destinatario = u.cpfcnpj
WHERE
    z.cod_sit = 100
GROUP BY
    z.nnf,
    z.mes,
    z.ano,
    z.vnf,
    z.vicms,
    z.destinatario);

CREATE TABLE MIN_lfe17_e AS(
   SELECT
   TO_CHAR( o000.cnpj) AS cnpj_dest,
    u.inscricao AS ie_dest,
    e020.num_doc,
    e020.dt_doc,
    e020.vl_cont,
    e020.vl_icms
FROM
    admlivro.le_registro_0000@link_lfe o000
    INNER JOIN admlivro.le_registro_e001@link_lfe e001 ON (
        o000.sequencial = e001.id_pai
    )
    INNER JOIN admlivro.le_registro_e020@link_lfe e020 ON (
        e001.sequencial = e020.id_pai
    )
    INNER JOIN dea_final u ON u.inscricao = o000.ie
WHERE
    ( o000.dt_ini ) >= TO_DATE('20170101000000','YYYYMMDDHH24MISS')
AND
    ( o000.dt_fin ) <= TO_DATE('20171231000000','YYYYMMDDHH24MISS')
AND
    e020.ind_oper = 0
GROUP BY
    o000.cnpj,
    u.inscricao,
    e020.num_doc,
```

```
    e020.dt_doc,
    e020.vl_cont,
    e020.vl_icms);

CREATE TABLE MIN_lfe17_s AS(
   SELECT
    TO_CHAR(o000.cnpj) AS cnpj_emit,
    u.inscricao AS ie_emit,
    e020.num_doc,
    e020.dt_doc,
    e020.vl_cont,
    e020.vl_icms
FROM
    admlivro.le_registro_0000@link_lfe o000
    INNER JOIN admlivro.le_registro_e001@link_lfe e001 ON (
        o000.sequencial = e001.id_pai
    )
    INNER JOIN admlivro.le_registro_e020@link_lfe e020 ON (
        e001.sequencial = e020.id_pai
    )
    INNER JOIN dea_final u ON u.inscricao = o000.ie
WHERE
    ( o000.dt_ini ) >= TO_DATE('20170101000000','YYYYMMDDHH24MISS')
AND
    ( o000.dt_fin ) <= TO_DATE('20171231000000','YYYYMMDDHH24MISS')
AND
    e020.ind_oper = 1
GROUP BY
    o000.cnpj,
    u.inscricao,
    e020.num_doc,
    e020.dt_doc,
    e020.vl_cont,
    e020.vl_icms);
```

Anexo II

Painéis Analíticos - Códigos em Linguagem R

PAINEL I - DEA:

```
---
title: "Contribuintes por Segmento Econômico - Data Envelopment Analisys (DEA) Orientado ao Output"
output:
  flexdashboard::flex_dashboard:
    orientation: rows
    vertical_layout: fill
---

```{r global_options, include=FALSE, results="hide", comment=FALSE}
knitr::opts_chunk$set(fig.width=60, fig.height=10, fig.path='Figs/', echo=FALSE, warning=FALSE, message=FALSE)

library(Benchmarking)
library(flexdashboard)
library(knitr)
library(RODBC)

# Connect to DB
channel <- odbcConnect("SERGIO_ORA", uid="dba_sergio", pwd="XXXXXXXXXX", believeNRows=FALSE)
odbcGetInfo(channel)

# Aquiring Data from DB
#ESCOLHIDO O SETOR ECONOMICO, BUSCAR OS DADOS

data <- sqlQuery(channel, "SELECT U.INSCRICAO,
                                  U.ARRECADACAO,
                                  U.COMPRAS_NENT_VNF,
                                  U.VENDAS_NSAI_VNF,
                                  U.VENDAS_CARTAO
                                FROM dba_sergio.UNB_DEA_GOOD U
                                  WHERE U.DESC_ATVD_ICMS =
'G471300100 - Lojas de departamentos ou magazines'
                                ORDER BY U.INSCRICAO")

x <- data.frame(data[,c(3,4,5)])
y <- data.frame(data[,2])
DMU <- data.frame(data[,1])
do <- dea(x,y, RTS="crs", ORIENTATION="out", SLACK=TRUE)
di <- dea(x,y, RTS="crs", ORIENTATION="in", SLACK=TRUE)
w <- round((do$eff*y))
k <- ((w-y)/y)
z <- data.frame(DMU,
                di$eff,
                y,
                w,
                w-y,
                round(k*100,0),
```
```

```
                data$ARRECADACAO,
                data$COMPRAS_NENT_VNF,
                data$VENDAS_NSAI_VNF,
                data$VENDAS_CARTAO)
colnames(z) <- c("CFDF","DEA", "REAL", "IDEAL", "DIF", "%","Arrecadação (OUT)", "Entradas (IN)", "Saídas (IN)", "Cartão (IN)")
s <- summary.default(di$eff)
```

```
Row  {data-height=2200}
-----------------------------------------------------------------------
### **Gráfico - DEA - Eficiência Relativa no Segmento Econômico**

```{r, results='hide'}
par(bg="gray71")
barplot(di$eff, col='darkblue',
        cex.axis = 4.4, col.axis = 'black',
        cex.lab = 4, col.lab = 'black',
        cex.main = 4, col.main = 'black',
        cex.sub = 4, col.sub = 'black',
        border = "black", space= .5)
```

Row {.tabset .tabset-fade}

Eficiências das DMUs (OUTPUT = Arrecadação / INPUTS = Entradas, Saídas, Cartão) R$x1000

```{r kable}
kable(z, digits=2, align = 'c')
```

Sumarização Histograma

```{r, fig.width=60, fig.height=20}
mn <- mean(di$eff)
md <- median(di$eff)
qnt25 <- quantile(di$eff, .25)
qnt75 <- quantile(di$eff, .75)

par(bg="gray71")
hist((di$eff), freq = TRUE, col='orange',
        cex.axis = 8, col.axis = 'black',
        lab = NULL,
        main = NULL,
        sub = NULL,
        border = "black", breaks = 5 )
abline(v = mn, col = 'black', lwd = 5)
abline(v = md, col = 'red', lwd = 5)
abline(v = qnt25, col = 'blue', lwd = 5)
abline(v = qnt75, col = 'darkgreen', lwd = 5)
legend(x = "topright", col = c('black','red', 'blue', 'darkgreen'), c("Média", "Mediana", '1_quartil', '3_quartil'),  lwd = 5, cex = 7)
```
```

PAINEL II - SÉRIES TEMPORAIS:

```
---
title: "VALOR DO ICMS - Análise de Séries Temporais (Método Holt-Winters)"
output:
  flexdashboard::flex_dashboard:
    orientation: rows
    vertical_layout: fill
---

```{r global_options, include=FALSE, results="hide", comment=FALSE}
knitr::opts_chunk$set(fig.width=30, fig.height=2,                                    fig.path='Figs/', echo=FALSE,
                      warning=FALSE, message=FALSE)

library(dygraphs)
library(MTS)
library(TTR)
library(tseries)
library(stats)
library(RODBC)

# Connect to DB
channel <- odbcConnect("SERGIO_ORA", uid="dba_sergio", pwd="XXXXXXXXXXX", believeNRows=FALSE)
odbcGetInfo(channel)

# Aquiring Data from DB
data <- sqlQuery(channel, "
                SELECT UNB_DEATS.INSCRICAO,
                UNB_DEATS.ANO,
                UNB_DEATS.MES,
                round(UNB_DEATS.LFE_E_ICMS/1000,0) as LFE_E_ICMS,
                round(UNB_DEATS.LFE_S_ICMS/1000,0) as LFE_S_ICMS,
                round(UNB_DEATS.NFE_E_ICMS/1000,0) as NFE_E_ICMS,
                round(UNB_DEATS.NFE_S_ICMS/1000,0) as NFE_S_ICMS
                FROM UNB_DEATS
                WHERE UNB_DEATS.INSCRICAO = 'XXXXXXXXXXX'
                ORDER BY UNB_DEATS.ANO,
                UNB_DEATS.MES")
```

```{r, results='hide'}
N <- dim(data)[1]
# from Jan 2010 to Dec 2016 as a time series object
LFE_E <- ts(data$LFE_E_ICMS, start=c((2016-(N/12)+1)), end=c(2016, 12), frequency=12)
NFE_E <- ts(data$NFE_E_ICMS, start=c((2016-(N/12)+1)), end=c(2016, 12), frequency=12)
DIF_E <- ts((data$NFE_E_ICMS-data$LFE_E_ICMS), start=c((2016-(N/12)+1)), end=c(2016, 12), frequency=12)
LFE_S <- ts(data$LFE_S_ICMS, start=c((2016-(N/12)+1)), end=c(2016, 12), frequency=12)
NFE_S <- ts(data$NFE_S_ICMS, start=c((2016-(N/12)+1)), end=c(2016, 12), frequency=12)
DIF_S <- ts((data$LFE_S_ICMS-data$NFE_S_ICMS), start=c((2016-(N/12)+1)), end=c(2016, 12), frequency=12)
# Decomposing TS's
d_lfe_e <- decompose(LFE_E)
d_lfe_s <- decompose(LFE_S)
d_nfe_e <- decompose(NFE_E)
d_nfe_s <- decompose(NFE_S)
# HoltWinters - Aditiva pq só a tendência é crescente
hw_lfe_e <- HoltWinters(LFE_E)
hw_lfe_s <- HoltWinters(LFE_S)
hw_nfe_e <- HoltWinters(NFE_E)
hw_nfe_s <- HoltWinters(NFE_S)
dhw_lfe_e <- (na.omit(hw_lfe_e$fitted[, 1]) - na.omit(LFE_E))
dhw_lfe_s <- (na.omit(LFE_S) - na.omit(hw_lfe_s$fitted[, 1]))
dhw_nfe_e <- (na.omit(hw_nfe_e$fitted[,1]) - na.omit(NFE_E))
dhw_nfe_s <- (na.omit(NFE_S) - na.omit(hw_nfe_s$fitted[,1]))
```

ENTRADAS
====================================
Row {data-height=450}

Séries Temporais Comparadas - Notas (NFE) x Livro (LFE)

```{r}
VICMS_E <- cbind(LFE_E, NFE_E, DIF_E )

dygraph(VICMS_E, main = "Entrada - Notas x Livro") %>%
 # dyOptions(fillGraph = TRUE, fillAlpha = 0.8) %>%
```
```

```
  dySeries('LFE_E', fillGraph = TRUE, color = "green", label = 'lfe') %>%
  dySeries('NFE_E', color = "magenta", label = 'nfe') %>%
  dySeries('DIF_E', fillGraph = TRUE, stepPlot = TRUE, color = "black", label = 'dif') %>%
  dyAxis("y", label = "ICMS (R$ x 1000)") %>%
  dyRangeSelector(height = 20) %>%
  dyHighlight(highlightCircleSize = 3,
              highlightSeriesBackgroundAlpha = 0.5,
              hideOnMouseOut = TRUE)
```

```
Row {.tabset .tabset-fade}
-------------------------------------
#####Notas (NFE) Entrada - Real X Estimado**

```{r}
HWNE <- cbind(NFE_E, hw_nfe_e$fitted[, 1], dhw_nfe_e)

dygraph(HWNE, main = "Notas Entrada - Holt-Winters") %>%
  dySeries('NFE_E', fillGraph = TRUE, color = "magenta", label = 'nfe') %>%
  dySeries('hw_nfe_e$fitted[, 1]', color = "red", label = 'hw') %>%
  dySeries('dhw_nfe_e', fillGraph = TRUE, color = "black", label = 'dif') %>%
  dyAxis("y", label = "ICMS (R$ x 1000)") %>%
  dyLimit(-(sd(hw_nfe_e$fitted[,1])), color = "red") %>%
  dyLimit(-2*(sd(hw_nfe_e$fitted[,1])), color = "red") %>%
  dyRangeSelector(height = 20)
```

###**Livro (LFE) Entrada - Real x Estimado**

```{r}
HWLFE_E <- cbind(LFE_E, hw_lfe_e$fitted[, 1], dhw_lfe_e)

dygraph(HWLFE_E, main = "Livro Entrada - Holt-Winters") %>%
  dySeries('LFE_E', fillGraph = TRUE, color = "green", label = 'lfe') %>%
  dySeries('hw_lfe_e$fitted[, 1]', color = "red", label = 'hw') %>%
  dySeries('dhw_lfe_e', fillGraph = TRUE, color = "black", label = 'dif') %>%
  dyAxis("y", label = "ICMS (R$ x 1000)") %>%
  dyLimit(-(sd(hw_lfe_e$fitted[,1])), color = "red") %>%
  dyLimit(-2*(sd(hw_lfe_e$fitted[,1])), color = "red") %>%
  dyRangeSelector(height = 20)
```

SAÍDAS {data-orientation=rows}
=====================================
Row {data-height=450}

Séries Temporais Comparadas - Notas (NFE) x Livro (LFE)

```{r}
VICMS_S <- cbind(LFE_S, NFE_S, DIF_S )

dygraph(VICMS_S, main = "Saída - Notas X Livro") %>%
  dySeries('LFE_S', fillGraph = TRUE, color = "orange", label = 'lfe') %>%
  dySeries('NFE_S', color = "blue", label = 'nfe') %>%
  dySeries('DIF_S', fillGraph = TRUE, stepPlot = TRUE,
           color = "black", label = 'dif') %>%
  dyAxis("y", label = "ICMS (R$ x 1000)") %>%
  dyRangeSelector(height = 20) %>%
  dyHighlight(highlightCircleSize = 3,
              highlightSeriesBackgroundAlpha = 0.5,
              hideOnMouseOut = TRUE)
```

```
Row {.tabset .tabset-fade}
-------------------------------------
#####Notas (NFE) Saída - Real X Estimado**

```{r}
HWNS <- cbind(NFE_S, hw_nfe_s$fitted[, 1], dhw_nfe_s)

dygraph(HWNS, main = "Notas Saída - Holt-Winters") %>%
  dySeries('NFE_S', fillGraph = TRUE, color = "blue", label = 'nfe') %>%
  dySeries('hw_nfe_s$fitted[, 1]', color = "red", label = 'hw') %>%
  dySeries('dhw_nfe_s', fillGraph = TRUE, color = "black", label = 'dif') %>%
  dyAxis("y", label = "ICMS (R$ x 1000)") %>%
  dyLimit(-(sd(hw_nfe_e$fitted[,1])), color = "red") %>%
  dyLimit(-2*(sd(hw_nfe_e$fitted[,1])), color = "red") %>%
```

```
  dyRangeSelector(height = 20)
```
```
#####**Livro (LFE) Saída - Real x Estimado**

```{r}
HWLS <- cbind(LFE_S, hw_lfe_s$fitted[, 1], dhw_lfe_s)

dygraph(HWLS, main = "Livro Saída - Holt-Winters") %>%
  dySeries('LFE_S', fillGraph = TRUE, color = "orange", label = 'lfe') %>%
  dySeries('hw_lfe_s$fitted[, 1]', color = "red", label = 'hw') %>%
  dySeries('dhw_lfe_s', fillGraph = TRUE, color = "black", label = 'dif') %>%
  dyAxis("y", label = "ICMS (R$ x 1000)") %>%
  dyLimit(-(sd(hw_lfe_e$fitted[,1])), color = "red") %>%
  dyLimit(-2*(sd(hw_lfe_e$fitted[,1])), color = "red") %>%
  dyRangeSelector(height = 20)
```

BOX-PLOT {data-orientation=rows}
=====================================
Row {data-height=450}

#####**Notas (NFE) Entrada - Ciclo de 12 Meses**

```{r, fig.width=10, fig.height=6}
par(bg="gray94")
boxplot(NFE_E-cycle(NFE_E), col='magenta', main='Notas Entrada')
```

#####**Livro (LFE) Entrada - Ciclo de 12 Meses**

```{r, fig.width=10, fig.height=6}
par(bg="gray94")
boxplot(LFE_E-cycle(LFE_E), col='green', main='Livro Entrada')
```

Row {data-height=450}

#####**Notas (NFE) Saída - Ciclo de 12 Meses**

```{r, fig.width=10, fig.height=6}
par(bg="gray94")
boxplot(NFE_S-cycle(NFE_S), col='blue', main='Notas Saída')
```

#####**Livro (LFE) Saída - Ciclo de 12 Meses**

```{r, fig.width=10, fig.height=6}
par(bg="gray94")
boxplot(LFE_S-cycle(LFE_S), col='orange', main='Livro Saída')
```
```

PAINEL III - MINERAÇÃO DE DADOS:

```
--
title: "MINERAÇÃO DADOS DO CONTRIBUINTE"
output:
  flexdashboard::flex_dashboard:
    orientation: columns
    vertical_layout: fill
---

```{r global_options, include=FALSE, results="hide", comment=FALSE}
knitr::opts_chunk$set(fig.width=30, fig.height=6,                                    fig.path='Figs/', echo=FALSE,
                      warning=FALSE, message=FALSE)

library(DistributionUtils)
library(abodOutlier)
library(datasets)
library(knitr)
library(RODBC)

# Connect to DB
channel <- odbcConnect("SERGIO_ORA", uid="dba_sergio", pwd="XXXXXXXXXXXX", believeNRows=FALSE)
odbcGetInfo(channel)
```

ENTRADA DISTRIBUIÇÃO {data-orientation=rows}
=====================================
Row {.tabset .tabset-fade}

#####Notas (NFE) Crédito (ICMS) - Histograma**

```{r}
# Aquiring data from DB

data <- sqlQuery(channel, "
                  SELECT nnf, ano, mes, vnf, VICMS
                  FROM unbmin_nfe16_e u
                  WHERE u.destinatario = 'XXXXXXXXXXXX'
                  ORDER BY U.VICMS")
nfe_e <- as.numeric(gsub(",",".",data[,5]))
vnf <- as.numeric(gsub(",",".",data[,4]))
rest <- (data[,1:3])
c <- (sqrt(var(na.omit(nfe_e)))/(sqrt(.1)))
out <- data.frame(rest,vnf,nfe_e)
out1 <- out[out$nfe_e>c,]
colnames(out1) <- c("#NF","Ano", "Mês", "Valor da Nota", "ICMS")

par(bg="gray71")
hist(na.omit(nfe_e), breaks=50, col='magenta', main='Notas Crédito', xlab='ICMS R$', pch = 19, cex = 2, freq = FALSE, cex.axis = 2, cex.main = 3, cex.lab = 2)
abline(v=c,lty=6, lwd=2, col=2)
abline(v=(mean(nfe_e)),lty=6, lwd=2, col=4)
rug(na.omit(nfe_e), col = 2)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####KERNEL**

```{r}
par(bg="gray71")
d <- density(nfe_e)
plot(d, col='black', main='Notas Crédito', xlab='ICMS R$', cex = 2, cex.axis = 2, cex.main = 3, cex.lab = 2)
polygon(d, col='magenta', border = 'blue')
abline(v=c,lty=6, lwd=2,col=2)
abline(v=(mean(nfe_e)),lty=6, lwd=2,col=4)
rug(na.omit(nfe_e), col = 2)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####LOG_Hist**

```{r}
par(bg="gray71")
logHist(na.omit(nfe_e), breaks=20, col='magenta', main='Notas Crédito', xlab='ICMS R$', pch = 19, cex = 2, freq = FALSE, cex.axis = 2, cex.main = 3, cex.lab = 2)
abline(v=c, lty=6, lwd=2, col=2)
abline(v=(mean(nfe_e)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```
```

```
```

#####Dispersão**

```{r}
par(bg="gray71")
plot((1:length(nfe_e)), nfe_e, xlim=c(0, length(na.omit(nfe_e))), col ='darkmagenta', main='Notas Crédito', xlab='ICMS R$', ylab='ICMS', pch = 19, cex.axis = 2, cex.
abline(h=c,lty=6, lwd=2,col=2)
abline(h=(mean(nfe_e)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####BOX-PLOT**

```{r}
par(bg="gray71")
boxplot(nfe_e, col ='magenta', main='Notas Crédito', cex.axis = 2, cex.main = 3, ylab='ICMS R$')
abline(h=c,lty=6, lwd=2,col=2)
abline(h=(mean(nfe_e)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####OUTLIERS**

```{r kable}
kable(out1, digits=2, align = 'c')
```
Row {.tabset .tabset-fade}

#####Livro (LFE) Créditos (ICMS) - Histograma**

```{r}
data <- sqlQuery(channel, "
                  SELECT num_doc, dt_doc, vl_cont, VL_ICMS
                  FROM UNBMIN_LFE16_E u
                  WHERE u.CNPJ_DEST = 'XXXXXXXXXXXX'
                  ORDER BY u.VL_ICMS")
lfe_e <- as.numeric(gsub(",",".", data[,4]))
vnf <- as.numeric(gsub(",",".",data[,3]))
rest <- (data[,1:2])
c <- (sqrt(var(na.omit(lfe_e)))/(sqrt(.1)))
out <- data.frame(rest,vnf,lfe_e)
out2 <- out[out$lfe_e>c,]
colnames(out2) <- c("#NF","Data", "Valor da Nota", "ICMS")

par(bg="gray71")
hist(na.omit(lfe_e), breaks=50, col="green", main='Livro Crédito', xlab='ICMS R$', pch = 19, cex = 2, freq = FALSE, cex.axis = 2, cex.main = 3, cex.lab = 2)
abline(v=c,lty=6, lwd=2,col=2)
abline(v=(mean(lfe_e)),lty=6, lwd=2,col=4)
rug(na.omit(lfe_e), col = 2)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####KERNEL**

```{r}
par(bg="gray71")
d <- density(lfe_e)
plot(d, col='black', main='Livro Crédito', xlab='ICMS R$', cex = 2, cex.axis = 2, cex.main = 3, cex.lab = 2)
polygon(d, col='green', border = 'blue')
abline(v=c,lty=6, lwd=2,col=2)
abline(v=(mean(lfe_e)),lty=6, lwd=2,col=4)
rug(na.omit(lfe_e), col = 2)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####LOG_Hist**

```{r}
par(bg="gray71")
logHist(na.omit(lfe_e), breaks=20, col="green", main='Livro Créditos', xlab='ICMS R$', pch = 19, cex = 2, cex.axis = 2, cex.main = 3, cex.lab = 2)
abline(v=c, lty=6, lwd=2, col=2)
abline(v=(mean(lfe_e)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####Dispersão**

```{r}
par(bg="gray71")
plot((1:length(lfe_e)), lfe_e, xlim=c(0, length(na.omit(lfe_e))), col ='darkgreen', main='Livros Crédito', xlab='', ylab='ICMS', pch = 19, cex.axis = 2, cex.main = 3
abline(h=c,lty=6, lwd=2,col=2)
abline(h=(mean(lfe_e)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####BOX-PLOT**

```{r}
par(bg="gray71")
boxplot(lfe_e, col ='green', main='Livros Crédito', cex.axis = 2, cex.main = 3, ylab='ICMS R$')
abline(h=c,lty=6, lwd=2,col=2)
abline(h=(mean(lfe_e)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####OUTLIERS**

```{r}
kable(out2, digits=2, align = 'c')
```

SAÍDA DISTRIBUIÇÃO {data-orientation=rows}
==========

Row {.tabset .tabset-fade}

#####Notas (NFE) Débito (ICMS) - Histograma**

```{r}
data <- sqlQuery(channel, "
                   SELECT nnf, ano, mes, vnf, VICMS
                   FROM UNBMIN_NFE16_s u
                   WHERE u.EMITENTE = 'XXXXXXXXXXXX'
                   ORDER BY u.VICMS")
nfe_s <- as.numeric(gsub(",",".",data[,5]))
vnf <- as.numeric(gsub(",",".",data[,4]))
rest <- (data[,1:3])
c <- (sqrt(var(na.omit(nfe_s)))/(sqrt(.95)))
out <- data.frame(rest,vnf,nfe_s)
out3 <- out[out$nfe_s<c,]
colnames(out3) <- c("#NF","Ano", "Mês", "Valor da Nota", "ICMS")

par(bg="gray71")
hist(na.omit(nfe_s), breaks=50, col='blue', main='Notas Débito', xlab='ICMS R$', pch = 19, cex = 2, freq = FALSE, cex.axis = 2, cex.main = 3, cex.lab = 2)
abline(v=c,lty=6, lwd=2,col=2)
abline(v=(mean(nfe_s)),lty=6, lwd=2,col='darkblue')
rug(na.omit(nfe_s), col = 2)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####KERNEL**

```{r}
par(bg="gray71")
d <- density(nfe_s)
plot(d, col='black', main='Notas Débito', xlab='ICMS R$', cex = 2, cex.axis = 2, cex.main = 3, cex.lab = 2)
polygon(d, col='blue', border = 'black')
abline(v=c,lty=6, lwd=2,col=2)
abline(v=(mean(nfe_s)),lty=6, lwd=2,col='darkblue')
rug(na.omit(nfe_s), col = 2)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####LOG_Hist**

```{r}
par(bg="gray71")
logHist(na.omit(nfe_s), breaks=20, col='blue', main='Notas Débito', xlab='ICMS R$', pch = 19, cex = 2, freq = FALSE, cex.axis = 2, cex.main = 3, cex.lab = 2)
abline(v=c, lty=6, lwd=2, col=2)
abline(v=(mean(nfe_s)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####Dispersão**

```{r}
par(bg="gray71")
plot((1:length(nfe_s)), nfe_s, xlim=c(0, length(na.omit(nfe_s))), col ='darkblue', main='Notas Débito', xlab='ICMS R$', ylab='ICMS', pch = 19, cex.axis = 2, cex.main
abline(h=c,lty=6, lwd=2,col=2)
abline(h=(mean(nfe_s)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####BOX-PLOT**

```{r}
par(bg="gray71")
boxplot(nfe_s, col ='blue', main='Notas Débito', cex.axis = 2, cex.main = 3, ylab='ICMS R$')
abline(h=c,lty=6, lwd=2,col=2)
abline(h=(mean(nfe_s)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####OUTLIERS**

```{r}
kable(out3, digits=2, align = 'c')
```

Row {.tabset .tabset-fade}

#####Livro (LFE) Débitos (ICMS) - Histograma**

```{r}
data <- sqlQuery(channel, "
                   SELECT num_doc, dt_doc, vl_cont, VL_ICMS
                   FROM UNBMIN_LFE16_S u
                   WHERE u.CNPJ_EMIT = 'XXXXXXXXXXXX'
                   ORDER BY u.VL_ICMS")
lfe_s <- as.numeric(gsub(",",".",data[,4]))
vnf <- as.numeric(gsub(",",".",data[,3]))
rest <- (data[,1:2])
c <- (sqrt(var(na.omit(lfe_s)))/(sqrt(.95)))
out <- data.frame(rest,vnf,lfe_s)
out4 <- out[out$lfe_s<c,]
colnames(out4) <- c("#NF","Data", "Valor da Nota", "ICMS")

par(bg="gray71")
hist(na.omit(lfe_s), breaks=50, col="orange", main='Livro Débito', xlab='ICMS R$', pch = 19, cex = 2, freq = FALSE, cex.axis = 2, cex.main = 3, cex.lab = 2)
abline(v=c,lty=6, lwd=2,col=2)
abline(v=(mean(lfe_s)),lty=6, lwd=2,col=4)
rug(na.omit(lfe_s), col = 2)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####KERNEL**

```{r}
par(bg="gray71")
d <- density(lfe_s)
plot(d, col='black', main='Livro Débito', xlab='ICMS R$', cex = 2, cex.axis = 2, cex.main = 3, cex.lab = 2)
polygon(d, col='orange', border = 'blue')
abline(v=c,lty=6, lwd=2,col=2)
abline(v=(mean(lfe_s)),lty=6, lwd=2,col=4)
rug(na.omit(lfe_s), col = 2)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####LOG_Hist**

```{r}
par(bg="gray71")
logHist(na.omit(lfe_s), breaks=20, col="red", main='Livro Débitos', xlab='ICMS R$', pch = 19, cex = 2, cex.axis = 2, cex.main = 3, cex.lab = 2)
abline(v=c, lty=6, lwd=2, col=2)
abline(v=(mean(lfe_s)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

#####Dispersão**

```{r}
```

```
par(bg="gray71")
plot((1:length(lfe_s)), lfe_s, xlim=c(0, length(na.omit(lfe_s))), col ='red', main='Livros Débito', xlab='', ylab='ICMS', pch = 19, cex.axis = 2, cex.main = 3, cex.la
abline(h=c,lty=6, lwd=2,col=2)
abline(h=(mean(lfe_s)),lty=6, lwd=2,col=4)
legend(x = "topright", col = c('red', 'blue'), c("Chebyshev", "Média"),  lwd = 5, cex = 2)
```

###**BOX-PLOT**

```{r}
par(bg="gray71")
boxplot(nfe_e, col ='red', main='Livros Débito', cex.axis = 2, cex.main = 3, ylab='ICMS R$')
abline(h=c,lty=6, lwd=2,col=2)
```

#####**OUTLIERS**

```{r}
kable(out4, digits=2, align = 'c')
```

PAINEL IV - NEWCOMB-BENFORD:

```
---
title: "Newcomb-Benford"
output:
  flexdashboard::flex_dashboard:
    orientation: columns
    vertical_layout: fill
---

```{r global_options, include=FALSE, results="hide", comment=FALSE}
knitr::opts_chunk$set(fig.width=20, fig.height=6,                                    fig.path='Figs/', echo=FALSE,
                      warning=FALSE, message=FALSE)

library(ggplot2)
library(knitr)
library(RODBC)

# Connect to DB
channel <- odbcConnect("SERGIO_ORA", uid="dba_sergio", pwd="XXXXXXXXXXXX", believeNRows=FALSE)
odbcGetInfo(channel)
```

ENTRADA NOTAS {data-orientation=rows}
=====================================
Row {.tabset .tabset-fade}

#####Notas (NFE) - NB (1 dígito)**

```{r}
# Aquiring data from DB

data <- sqlQuery(channel, "
                 SELECT VNF
                 FROM UNBMIN_NFE16_E
                 WHERE DESTINATARIO = 'XXXXXXXXXXXX'")
vnf <- as.numeric(gsub(",",".",data[,1]))

benlaw <- function(d) log10(1 + 1 / d)
digits <- 1:9

firstDigit <- function(x) substr(gsub('[0.]', '', x), 1, 1)

pctFirstDigit <- function(x) data.frame(table(firstDigit(x)) / length(x))

df1 <- pctFirstDigit(vnf)

df <- data.frame(x = digits, y = benlaw(digits))

ggBarplot <- ggplot(df, aes(x = factor(x), y = y, notas = "notas")) + geom_bar(stat = "identity") +
  xlab("Primeiro Dígito") + ylab(NULL)

p1 <- ggBarplot +
  geom_line(data = df1,
            aes(x = Var1, y = Freq, group = 1),
            colour = "red",
            size = 2) +
  geom_point(data = df1,
             aes(x = Var1, y = Freq, group = 1),
             colour = "red",
             size = 4, pch = 23, bg = "red")

print(p1)
```

Row {.tabset .tabset-fade}

#####Livro (LFE) - NB (1 dígito)**

```{r}
# Aquiring data from DB

data <- sqlQuery(channel, "
                 SELECT VL_CONT
                  FROM unbmin_lfe16_e
                  WHERE cnpj_dest = 'XXXXXXXXXXXX'")
vnf <- as.numeric(gsub(",",".",data[,1]))
```
```

```
benlaw <- function(d) log10(1 + 1 / d)
digits <- 1:9

firstDigit <- function(x) substr(gsub('[0.]', '', x), 1, 1)

pctFirstDigit <- function(x) data.frame(table(firstDigit(x)) / length(x))

df1 <- pctFirstDigit(vnf)

df <- data.frame(x = digits, y = benlaw(digits))
ggBarplot <- ggplot(df, aes(x = factor(x), y = y, notas = "notas")) + geom_bar(stat = "identity") +
  xlab("Primeiro Dígito") + ylab(NULL)

p1 <- ggBarplot +
  geom_line(data = df1,
            aes(x = Var1, y = Freq, group = 1),
            colour = "red",
            size = 2) +
  geom_point(data = df1,
             aes(x = Var1, y = Freq, group = 1),
             colour = "red",
             size = 4, pch = 23, bg = "red")

print(p1)
```

```
SAÍDA NOTAS {data-orientation=rows}
=====================================
Row {.tabset .tabset-fade}
-------------------------------------
#####Notas (NFE) - NB (1 dígito)**

```{r}
# Aquiring data from DB

data <- sqlQuery(channel, "
                 SELECT VNF
                 FROM UNBMIN_NFE16_S
                 WHERE EMITENTE = 'XXXXXXXXXXXX'")
vnf <- as.numeric(gsub(",",".",data[,1]))

benlaw <- function(d) log10(1 + 1 / d)
digits <- 1:9

firstDigit <- function(x) substr(gsub('[0.]', '', x), 1, 1)

pctFirstDigit <- function(x) data.frame(table(firstDigit(x)) / length(x))

df1 <- pctFirstDigit(vnf)

df <- data.frame(x = digits, y = benlaw(digits))

ggBarplot <- ggplot(df, aes(x = factor(x), y = y, notas = "notas")) + geom_bar(stat = "identity") +
  xlab("Primeiro Dígito") + ylab(NULL)

p1 <- ggBarplot +
  geom_line(data = df1,
            aes(x = Var1, y = Freq, group = 1),
            colour = "red",
            size = 2) +
  geom_point(data = df1,
             aes(x = Var1, y = Freq, group = 1),
             colour = "red",
             size = 4, pch = 23, bg = "red")

print(p1)
```

Row {.tabset .tabset-fade}

#####Livro (LFE) - NB (1 dígito)**

```{r}
# Aquiring data from DB

data <- sqlQuery(channel, "
```
```

```
                SELECT VL_CONT
                 FROM unbmin_lfe16_s
                 WHERE cnpj_emit = 'XXXXXXXXXXXX'")
vnf <- as.numeric(gsub(",",".",data[,1]))

benlaw <- function(d) log10(1 + 1 / d)
digits <- 1:9

firstDigit <- function(x) substr(gsub('[0.]', '', x), 1, 1)

pctFirstDigit <- function(x) data.frame(table(firstDigit(x)) / length(x))

df1 <- pctFirstDigit(vnf)

df <- data.frame(x = digits, y = benlaw(digits))
ggBarplot <- ggplot(df, aes(x = factor(x), y = y, notas = "notas")) + geom_bar(stat = "identity") +
  xlab("Primeiro Dígito") + ylab(NULL)


p1 <- ggBarplot +
  geom_line(data = df1,
            aes(x = Var1, y = Freq, group = 1),
            colour = "red",
            size = 2) +
  geom_point(data = df1,
             aes(x = Var1, y = Freq, group = 1),
             colour = "red",
             size = 4, pch = 23, bg = "red")

print(p1)
```

PAINEL V - *K-MEANS*:

```
---
title: "K-means"
output:
  flexdashboard::flex_dashboard:
    orientation: columns
    vertical_layout: fill
---

```{r global_options, include=FALSE, results="hide", comment=FALSE}
knitr::opts_chunk$set(fig.width=30, fig.height=10,                                    fig.path='Figs/', echo=FALSE,
                      warning=FALSE, message=FALSE)

library(DistributionUtils)
library(abodOutlier)
library(datasets)
library(ggplot2)
library(knitr)
library(RODBC)

# Connect to DB
channel <- odbcConnect("SERGIO_ORA", uid="dba_sergio", pwd="XXXXXXXXXX", believeNRows=FALSE)
odbcGetInfo(channel)
```

ENTRADA NOTAS {data-orientation=rows}
=====================================
Column {data-width=500} {data-height=2200}

###**Notas (NFE) Crédito (ICMS) - K-mean**

```{r}
# Aquiring data from DB

data <- sqlQuery(channel, "
                 SELECT VNF, VICMS
                 FROM UNBMIN_NFE16_E
                 WHERE DESTINATARIO = 'XXXXXXXXXXXX'")
icms <- as.numeric(gsub(",",".",data[,2]))
vnf <- as.numeric(gsub(",",".",data[,1]))
al <- (round((icms/vnf)*100))
nf <- data.frame(vnf,icms, al)
set.seed(20)
k <- kmeans(nf[, 1:2], 4, nstart = 20)

k$cluster <- as.factor(k$cluster)
ggplot(nf, aes(vnf, icms, color = k$cluster)) + geom_point(shape=18, size=8) +theme(axis.text=element_text(size=22),
        axis.title=element_text(size=24,face="bold")) +theme(
        legend.text=element_text(size=24))
```

ENTRADA LIVROS {data-orientation=rows}
=====================================
Column {data-width=500} {data-height=2200}

###**Livro (LFE) Crédito (ICMS) - K-mean**

```{r}
# Aquiring data from DB

data <- sqlQuery(channel, "
                 SELECT VL_CONT, VL_ICMS
                 FROM UNBMIN_LFE16_E
                 WHERE CNPJ_DEST = 'XXXXXXXXXXXX'")
icms <- as.numeric(gsub(",",".",data[,2]))
vnf <- as.numeric(gsub(",",".",data[,1]))
al <- (round((icms/vnf)*100))
nf <- data.frame(vnf,icms, al)
set.seed(20)
k <- kmeans(nf[, 1:2], 4, nstart = 20)

k$cluster <- as.factor(k$cluster)
ggplot(nf, aes(vnf, icms, color = k$cluster)) + geom_point(shape=18, size=8) +theme(axis.text=element_text(size=22),
        axis.title=element_text(size=24,face="bold")) +theme(
        legend.text=element_text(size=24))
```
```

```
SAÍDAS NOTAS {data-orientation=rows}
=====================================
Column {data-width=500} {data-height=2200}
-------------------------------------
###**Notas (NFE) Débito (ICMS) - K-mean**

'''{r}
# Aquiring data from DB

data <- sqlQuery(channel, "
                 SELECT VNF, VICMS
                 FROM UNBMIN_NFE16_S
                 WHERE EMITENTE = 'XXXXXXXXXXXX'")
icms <- as.numeric(gsub(",",".",data[,2]))
vnf <- as.numeric(gsub(",",".",data[,1]))
al <- (round((icms/vnf)*100))
nf <- data.frame(vnf,icms, al)
set.seed(20)
k <- kmeans(nf[, 1:2], 4, nstart = 20)

k$cluster <- as.factor(k$cluster)
ggplot(nf, aes(vnf, icms, color = k$cluster)) + geom_point(shape=18, size=8) +theme(axis.text=element_text(size=22),
        axis.title=element_text(size=24,face="bold")) +theme(
        legend.text=element_text(size=24))
'''

SAÍDA LIVROS {data-orientation=rows}
=====================================
Column {data-width=500} {data-height=2200}
-------------------------------------
###**Livro (LFE) Débito (ICMS) - K-mean**

'''{r}
# Aquiring data from DB

data <- sqlQuery(channel, "
                 SELECT VL_CONT, VL_ICMS
                 FROM UNBMIN_LFE16_S
                 WHERE CNPJ_EMIT = 'XXXXXXXXXXXX'")
icms <- as.numeric(gsub(",",".",data[,2]))
vnf <- as.numeric(gsub(",",".",data[,1]))
al <- (round((icms/vnf)*100))
nf <- data.frame(vnf,icms, al)
set.seed(20)
k <- kmeans(nf[, 1:2], 4, nstart = 20)

k$cluster <- as.factor(k$cluster)
ggplot(nf, aes(vnf, icms, color = k$cluster)) + geom_point(shape=18, size=8) +theme(axis.text=element_text(size=22),
        axis.title=element_text(size=24,face="bold")) +theme(
        legend.text=element_text(size=24))
'''
```