

FUNDAÇÃO GETULIO VARGAS  
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

CAROLINA MOURÃO DE PAULA

**OTIMIZAÇÃO DO PORTFÓLIO DE CLIENTES: UMA NOVA PROPOSIÇÃO  
ATRAVÉS DE DESIGN DE PROCESSOS BASEADO EM DADOS**

SÃO PAULO

2021

CAROLINA MOURÃO DE PAULA

**OTIMIZAÇÃO DO PORTFÓLIO DE CLIENTES: UMA NOVA PROPOSIÇÃO  
ATRAVÉS DE DESIGN DE PROCESSOS BASEADO EM DADOS**

Trabalho Aplicado apresentado à Escola de Administração de Empresas de São Paulo da Fundação Getulio Vargas como requisito para a obtenção do título de Mestre em Gestão para a Competitividade.

Linha de Pesquisa: Tecnologia da Informação.

Orientador: Prof. Dr. Eduardo de Rezende Francisco

SÃO PAULO

2021

Paula, Carolina Mourão de.

Otimização do portfólio de clientes : uma nova proposição através de design de processos baseado em dados / Carolina Mourão de Paula. - 2021.

135 f.

Orientador: Eduardo de Rezende Francisco.

Dissertação (mestrado profissional MPGC) – Fundação Getulio Vargas, Escola de Administração de Empresas de São Paulo.

1. Clientes - Contatos. 2. Mineração de dados (Computação). 3. Aquisição de conhecimento (Sistemas especialistas). 4. Algoritmos de computador. I. Francisco, Eduardo de Rezende. II. Dissertação (mestrado profissional MPGC) – Escola de Administração de Empresas de São Paulo. III. Fundação Getulio Vargas. IV. Título.

CDU 658.8::007.522.2

Ficha Catalográfica elaborada por: Raphael Figueiredo Xavier CRB SP-009987/O

Biblioteca Karl A. Boedeker da Fundação Getulio Vargas - SP

CAROLINA MOURÃO DE PAULA

**OTIMIZAÇÃO DO PORTFÓLIO DE CLIENTES: UMA NOVA PROPOSIÇÃO  
ATRAVÉS DE DESIGN DE PROCESSOS BASEADO EM DADOS**

Trabalho Aplicado apresentado à Escola de Administração de Empresas de São Paulo da Fundação Getúlio Vargas como requisito para a obtenção do título de Mestre em Gestão para a Competitividade.

Linha de Pesquisa: Tecnologia da Informação.

Data de Aprovação:

19/03/2021

Banca Examinadora:

---

Prof. Dr. Eduardo de Rezende Francisco  
(Orientador)  
FGV EAESP

---

Prof. Dr. José Luiz Carlos Kugler  
FGV EAESP

---

Prof. Dr. Gustavo Correa Mirapalheta  
FGV EAESP

---

Prof. Dr. Fabiano Rodrigues  
ESPM

## **AGRADECIMENTO**

Fui eu quem dei o primeiro passo para começar meu caminho pelo mestrado, mas minha jornada foi marcada por pessoas iluminadas que me deram muito incentivo para continuar. Meus agradecimentos são muito sinceros.

Ao querido professor Eduardo, que tem sido um grande companheiro na minha busca pelo conhecimento. Por me apresentar ao mundo acadêmico, me orientar e me acolher, não só com este trabalho, mas durante todo o curso com o seu conhecimento. E ao seu verdadeiro interesse em me fazer continuar a estudar.

À minha fiel companheira, Sara, que sempre me apoiou a ser uma pessoa melhor. E sua profunda compreensão em momentos em que eu estive distante.

Aos dedicados professores da FGV, que enriqueceram nossas aulas com tanta energia e sabedoria, tornando nossos encontros verdadeiramente enriquecedores e inspiradores. Deles, levo muito aprendizado para a minha vida.

## RESUMO

Dentre as principais práticas de modelagem de processos, a técnica *Process Mining* captura e analisa o comportamento real de processos a partir do uso de *Analytics* no contexto das soluções de Mineração de Dados. Não obstante, apesar dos avanços dessa área, ainda há espaço para inclusão de outras técnicas analíticas que alcancem melhorias na performance do processo a partir dos dados. Este trabalho propõe o uso de análises probabilísticas com Cadeias de Markov para a descoberta do real processo a partir dos *logs* do sistema e aplica em estudo de caso empírico a uma empresa multinacional de serviços logísticos. Como principais resultados, destaca-se que a aplicação dessa técnica revelou um processo com variações muito mais probabilísticas do que determinísticas formalizadas na instituição. Mais ainda, a dissimilaridade entre os processos dos clientes foi notada de forma significativa, o que implica em um distanciamento da ideia de um processo único institucionalizado. Com os dados derivados da descoberta dos diferentes processos, foi possível executar simulações que produzissem resultados para melhorar a produtividade do departamento. Dentre as soluções para testar cenários que produzissem melhor resultado entre receita e custos derivados dos processos, um algoritmo baseado em alocações aleatórias de volume de negócio por cliente testou a performance de diferentes composições desses portfólios para conseguir melhor receita média comparado ao desvio padrão do *lead time* do processo. Para otimizar os resultados das simulações, foi empregado algoritmo genético para resolver o modelo proposto de forma eficiente. Assim, foi formulada uma função objetivo para diminuir o desvio padrão do *lead time* do processo, aumentar a receita média do departamento, diminuir o desvio padrão da receita, diminuir o *lead time* do processo e diminuir o escore de desvios do processo, empregando uma nova carteira de clientes. Esse novo modelo, deriva da abordagem matemática da teoria de portfólio criada por H. Markowitz (1952), diminuindo o desvio padrão de 344,26 horas para 297,54 horas e aumentando a receita média por transação por cliente de R\$ 702,29 para R\$ 729,75, potencializando um significativo aumento anual de receita para a companhia. A mineração de processos foi a fundação para um tipo de procedimento que analisa custos dos processos como base para a obtenção de uma carteira de clientes com melhores resultados e abre um horizonte de usos combinados entre a Gestão de Processos e os métodos analíticos de otimização.

Palavras-chave: Mineração de dados; *Process Mining*; Algoritmos Genéticos; Portfólio de Clientes; *Big Data Analytics*

## **ABSTRACT**

Among the main process modelling practices, the Process Mining technique captures and analyses the actual behaviour of processes using Analytics in the context of Data Mining solutions. Nevertheless, despite the advances in this area, there is still room for the inclusion of other analytical techniques that achieve improvements in the performance of the process from the data. This work proposes the use of probabilistic analysis with Markov Chains to discover the process from the system logs and apply it to an empirical case study to a multinational logistics services company. Among the main results, it is highlighted that the application of this technique revealed a process with variations characteristics of much more probabilistic nature than deterministic formalized in the institution. Furthermore, the dissimilarity between the processes of the clients was significantly noticed, which implies a distance from the idea of a single institutionalized process. With the data derived from the discovery of the different processes, it was possible to run simulations that would produce results to improve the productivity of the department. Among the solutions to test scenarios that would produce better results between revenue and costs derived from processes, an algorithm based on random allocations of work load per client tested the performance of different compositions of these portfolios to achieve better average revenue compared to the standard deviation of the process lead time. To optimize the results of the simulations, a genetic algorithm was used to solve the proposed model efficiently. Thus, an objective function was formulated to decrease the standard deviation of the process lead time, increase the average revenue of the department, decrease the standard deviation of the revenue, decrease the process lead time and decrease the process deviation score, using a new customer portfolio. This new model derives from the mathematical approach of portfolio theory created by H. Markowitz (1952), decreasing the standard deviation from 344.26 hours to 297.54 hours and increasing the average revenue from BRL 702.29 to BRL 729.75, leveraging a significant annual revenue increase for the company. Therefore, Process Mining was the foundation for a type of procedure that analyses process costs as a basis for obtaining a client portfolio with better results and it expands a horizon of combined uses between Process Management and analytical methods of optimization.

**Keywords:** Data Mining; Process Mining; Genetic Algorithms; Customer Portfolio; Big Data Analytics

## Lista de Figuras

<b>Figura 1:</b> Diagrama de Venn da combinação de linhas de pensamento.....	19
<b>Figura 2:</b> Exemplo do grafo com as probabilidades de transição dos três tipos de clima.....	33
<b>Figura 3:</b> Fronteira eficiente de um portfólio de investimentos .....	38
<b>Figura 4:</b> Matriz de classificação dos clientes .....	40
<b>Figura 5:</b> Representação do processo de seleção, cruzamento e mutação .....	43
<b>Figura 6:</b> Processos do Desembaraço Aduaneiro.....	48
<b>Figura 7:</b> Processo operacional institucionalizado pela empresa e seus eventos .....	49
<b>Figura 8:</b> Etapas Metodológicas do Estudo.....	58
<b>Figura 9:</b> Grafo que representa o processo estruturalmente institucionalizado .....	73
<b>Figura 10:</b> Grafo das interconexões do processo na prática, a partir dos dados .....	74
<b>Figura 11:</b> Processo descoberto para o cliente 1 .....	76
<b>Figura 12:</b> Processo descoberto para o cliente 2.....	77

## Lista de Gráficos

<b>Gráfico 1:</b> Quantidade de declarações por ano.....	50
<b>Gráfico 2:</b> Média e desvio padrão do <i>lead time</i> dos eventos por ano.....	50
<b>Gráfico 3:</b> Cruzamento do Desvio Padrão do Processo e Receita entre os clientes .....	80
<b>Gráfico 4:</b> Portfólios com métodos aleatórios: Desvio Padrão versus Receita .....	82
<b>Gráfico 5:</b> Portfólios com métodos aleatórios: Objetivo geral versus Receita.....	83
<b>Gráfico 6:</b> Rebalanceamento da carteira de clientes .....	85
<b>Gráfico 7:</b> Desempenho por método aleatório e por Algoritmo Genético .....	86

## Lista de Tabelas

<b>Tabela 1:</b> Exemplo das probabilidades de transição dos três tipos de clima.....	33
<b>Tabela 2:</b> Top-3 sequência de eventos mais frequentes .....	71
<b>Tabela 3:</b> Probabilidades de eventos após a chegada (ADD). Atual versus Esperada .....	72
<b>Tabela 4:</b> Resultados dos valores obtidos através da otimização do portfólio de clientes ....	87
<b>Tabela 5:</b> Resultados dos valores obtidos na escala de 1 ano .....	92

## Lista de Quadros

<b>Quadro 1:</b> Resumo da verificação das hipóteses do estudo .....	88
--	----

## Lista de Acrônimos e Siglas

B2B	<i>Business to Business</i>
BDA	<i>Big Data Analytics</i>
BPM	<i>Business Process Management</i>
BPR	<i>Business Process Reengineering</i>
CFOP	Código Fiscal de Operação e Prestação
COFINS	Contribuição para o Financiamento de Seguridade Social
CPA	<i>Customer Profitability Analysis</i>
CRAN	<i>Comprehensive R Archive Network</i>
DI	Declaração de Importação
DTMC	<i>Discrete Time Markov Chain</i>
EDI	<i>Electronic Data Interchange</i>
GA	<i>Genetic Algorithm</i>
ICMS	Imposto sobre Mercadorias e Serviços
ID	<i>Identifier</i>
IDE	<i>Integrated Development Environment</i>
II	Imposto de Importação
IPI	Imposto de Produtos Industrializados
KDD	<i>Knowledge Discovery from Data</i>
ML	<i>Machine Learning</i>
MPT	<i>Modern Portfolio Theory</i>
NCM	Nomenclatura Comum do Mercosul
NP	<i>Non-Deterministic Polynomial time</i>
P&D	Pesquisa e Desenvolvimento
PIS	Programa de Integração Social
PSL	Prestadora de Serviços Logísticos
SEFAZ	Secretaria da Fazenda
SI	Sistema de Informação

SOP	<i>Service Operations Procedures</i>
SQL	<i>Structured Query Language</i>
ST	Situação Tributária
TI	Tecnologia da Informação
TMP	Teoria Moderna de Portfólio
TQM	<i>Total Quality Management</i>

# SUMÁRIO

1	INTRODUÇÃO .....	15
	1.1 Relevância.....	20
	1.2 Objetivo da pesquisa .....	20
	1.3 Organização do trabalho .....	21
2	REVISÃO DA LITERATURA .....	23
	2.1 Gestão de processos .....	23
	2.2 <i>Big Data Analytics</i> .....	25
	2.2.1 Data Science .....	27
	2.2.2 <i>Data Mining</i> .....	28
	2.2.3 Process Mining .....	28
	2.3 Processos estocásticos e Cadeia de Markov .....	30
	2.4 Processos no setor de serviços .....	34
	2.5 Classificação de custos com foco no cliente .....	35
	2.6 A gestão de portfólio de clientes.....	36
	2.7 Algoritmos Heurísticos .....	41
	2.7.1 Algoritmos Genéticos.....	41
	2.8 CASO: Empresa de serviços do setor logístico .....	43
	2.8.1 Processos do desembaraço aduaneiro.....	45
3	MÉTODO .....	53
	3.1 ESTUDO DE CASO.....	54
	3.1.1 Coleta de dados .....	58
	3.1.2 Os softwares .....	61
	3.1.3 O desenho do processo .....	62
	3.1.4 O desenho do processo por cliente .....	62
	3.1.5 As variáveis de análise dos clientes.....	63
	3.1.6 As simulações de portfólio de clientes .....	66
	3.1.7 As simulações por modelos de otimização.....	67
4	ANÁLISE DOS RESULTADOS .....	71
5	CONCLUSÕES .....	90
	5.1 Implicações Gerenciais .....	94
6	REFERÊNCIAS.....	96

# 1 INTRODUÇÃO

O Big Data tem um grande potencial para transformar processos de negócio (Wamba et al., 2015). Em todos os setores, em todas as partes do mundo, os líderes executivos se perguntam se estão obtendo o maior valor possível das enormes quantidades de informações que já possuem em suas organizações (LaValle et al., 2011). Com esta quantidade enorme de dados complexos e heterogêneos fluindo de qualquer lugar, a qualquer hora e qualquer dispositivo, é fato dizer que estamos na era do Big Data (BD) (Sivarajah et al., 2017). Independentemente de onde o BD é gerado e compartilhado, junto com a realidade de BD vem o desafio de analisá-lo de uma forma que agregue Grande Valor (*Big Value*) (Sivarajah et al., 2017). O BD apresenta tantas oportunidades de ganho de valor que já é considerado o novo petróleo (Yi, Liu, Liu e Jin, 2014). O processamento e a gestão correta dos dados podem criar novos conhecimentos e facilitar a resposta às oportunidades e desafios emergentes (Chen et al., 2013).

Existe um discurso cada vez maior sobre o BD oferecer grandes oportunidades por meio da abundância de fontes de dados de diferentes áreas. Como exemplo, pode-se citar a criação de valor (Brown, Chui e Manyika, 2011), a geração de inteligência de negócios para tomadas de decisões mais embasadas (Chen e Zhang, 2014) e o aumento de visibilidade e flexibilidade da cadeia de suprimentos e uma melhor alocação de recursos (Kumar, Niu e Ré, 2013).

Novas oportunidades e desafios também criam oportunidades para pesquisar melhores algoritmos e sistemas (Agarwal e Dhar, 2014). Nesse cenário, um dos processos de obtenção de conhecimento que tem crescido e se tornado um paradigma emergente é a gestão de processos através de dados (Shmueli e Koppius, 2011; Chen e Storey 2012; Breuker et al., 2016).

Por sua vez, o desenho de processo é um problema crítico de negócio. As possibilidades de ganho de vantagem competitiva através de novos processos mais eficientes geram interesse em ferramentas que criem e reinventem processos (Bernstein et al., 1999; Malone et al., 1999). Com os processos sendo um importante ativo de uma empresa (Accorsi et al., 2015), a gestão de processos já foi identificada como prioridade já que tem uma função importante para o sucesso de uma organização que visa adquirir vantagem competitiva pela eficiência (Fantinato et al., 2010). Devido a esta relevância em desempenho, implementar o gerenciamento de processos de negócios

tornou-se cada vez mais popular dentro das empresas. Desse movimento surgiu o *Business Process Management* (BPM), que é a gestão de todo o ciclo de vida dos processos de uma organização.

Algumas abordagens para melhoria de processos se baseiam na ideia de um modelo formal de processo. Levitt (1972) considerou a “abordagem da linha de produção” para melhorar significativamente o desempenho da organização em custos, principalmente por meio de maior eficiência e qualidade (Bowen e Youngdahl, 1998). Assim, entre as propostas de Levitt (1997), uma foi a de padronização de serviço. A padronização permite previsibilidade, pré-planejamento e controle de processo que, por sua vez, fornece uniformidade na qualidade do serviço. Infelizmente, o desenvolvimento de um modelo formal para um processo, muitas vezes em ambientes mais complexos, tem apresentado falhas na prática. Os participantes de um negócio podem consultar a referência formal, mas possuem espaço para interpretação e adesão (Lee et al., 2008). Para entender o paradoxo da padronização e da liberdade de adesão, este trabalho visa usar os dados, chamados aqui de *logs*, capturados de um Sistema de Informação (SI), para gerar um modelo formal do comportamento real desse processo. A premissa deste trabalho é de que essa visão permite visualizar transições não previstas no desenho institucional e de que os *logs* dos sistemas contam uma história diferente do processo institucionalizado, ou seja, não refletem absolutamente o que o processo estabelece.

Há uma tendência crescente em automatizar o estudo de processos, com base em dados (Lakshmanan et al., 2012; Chiu et al., 2019; Ibanez-Sanchez et al., 2019; van der Aalst, 2011; Weijters et al., 2006; Goedertier et al., 2009). Sob essa técnica, os *logs* são capturados de um SI responsável pela gestão das atividades, no qual existe uma interação viva entre eles e operações e, em seguida, usados para gerar um modelo. Com técnicas de Mineração de Dados é possível identificar padrões nos dados através de *Knowledge Discovery from Data* (KDD) (Fayyad et al., 1996) criando valor a partir dos dados. Diante dessas possibilidades, este trabalho estuda a capacidade de combinar o BPM com técnicas de Mineração de Dados, para facilitar a análise de processos.

Desse conjunto de técnicas, derivou-se o *Process Mining*. Breuker, Matzner, Delfmann e Becker (2016) mencionaram em sua pesquisa que o estudo de *Process Mining* na literatura de BPM cresceu de 0 para 31,4% de 2000 a 2011, o que o torna um conceito chave em BPM.

Entre tantos algoritmos, este estudo optou pelo uso de modelos probabilísticos. As técnicas probabilísticas são populares em muitos campos devido a sua capacidade de lidar com dados

ruidosos e possivelmente incompletos. Neste trabalho foram utilizadas Cadeias de Markov para a análise de processos. Cadeias de Markov apresentam o benefício de mostrar facilmente às pessoas de negócio uma rede de possibilidades das diferentes transições de estados do processo (Jónás et al., 2014).

Para dar sequência a esse estudo, com algumas variáveis obtidas da análise probabilística do processo, foi proposto um modelo matemático que buscasse encontrar a melhor combinação da distribuição de esforço de trabalho entre os clientes com maior valor para o negócio, baseando-se no processo descoberto para cada cliente. Sob a estrutura teórica da economia dos custos de transação (Williamson, 1981), as organizações buscam minimizar os custos necessários para realizar uma transação econômica. Harmon (2011) identificou que os clientes têm necessidades específicas e essas especificidades influenciam os processos da empresa prestadora de serviços. Ter informações sobre os custos da prestação de serviços ao cliente pode modificar o relacionamento de uma empresa com clientes que não são rentáveis (Kaplan e Narayanan, 2001).

O objetivo foi a minimização de custos provenientes de um processo mal executado, representado aqui pelo tempo de processamento (*lead time*) e desvio padrão do *lead time*, em comparação à receita média de um portfólio de clientes. O valor monetário do custo foi substituído pelo *lead time*, partindo da premissa de que as horas necessárias para a execução das etapas do processo representam um custo para a empresa. A abordagem dos custos de transação para o estudo da organização econômica considera a transação como a unidade básica de análise e sustenta que uma compreensão da economia dos custos é central para o estudo das organizações (Williamson, 1981).

Foi desenvolvido um algoritmo genético (AG, *Genetic Algorithm*) com bases heurísticas para resolver o modelo proposto de forma eficiente. Os AGs são uma família de modelos computacionais para resolver problemas de otimização inspirados na evolução, que incorporam uma solução matemática para um problema numa estrutura semelhante a de um cromossomo e aplicam operadores de seleção e cruzamento dos genes (Holland, 1992). Normalmente os AGs são vistos como otimizadores de funções, embora a quantidade de problemas para o qual os AGs se aplicam seja bastante abrangente. Este modelo derivou da abordagem matemática da teoria de portfólio criada por H. Markowitz (1952), integrando a variável risco representada pelo desvio padrão do *lead time* com as receitas que esses clientes proporcionam ao negócio. Este trabalho

incorpora essas análises através de portfólios de clientes que consideram o aspecto do processo como ponto de partida.

Para implementar os modelos, foi executado um estudo de caso em uma empresa do setor de serviços. Ponsignon, Smart, e Maull (2012) buscaram entender como o contexto do setor em que a empresa atua tem um papel importante na abordagem em processos. A empresa estudada é uma provedora de serviços de logística de terceiros (PSL) que possui como clientes empresas que enfrentam desafios na sua cadeia de suprimentos.

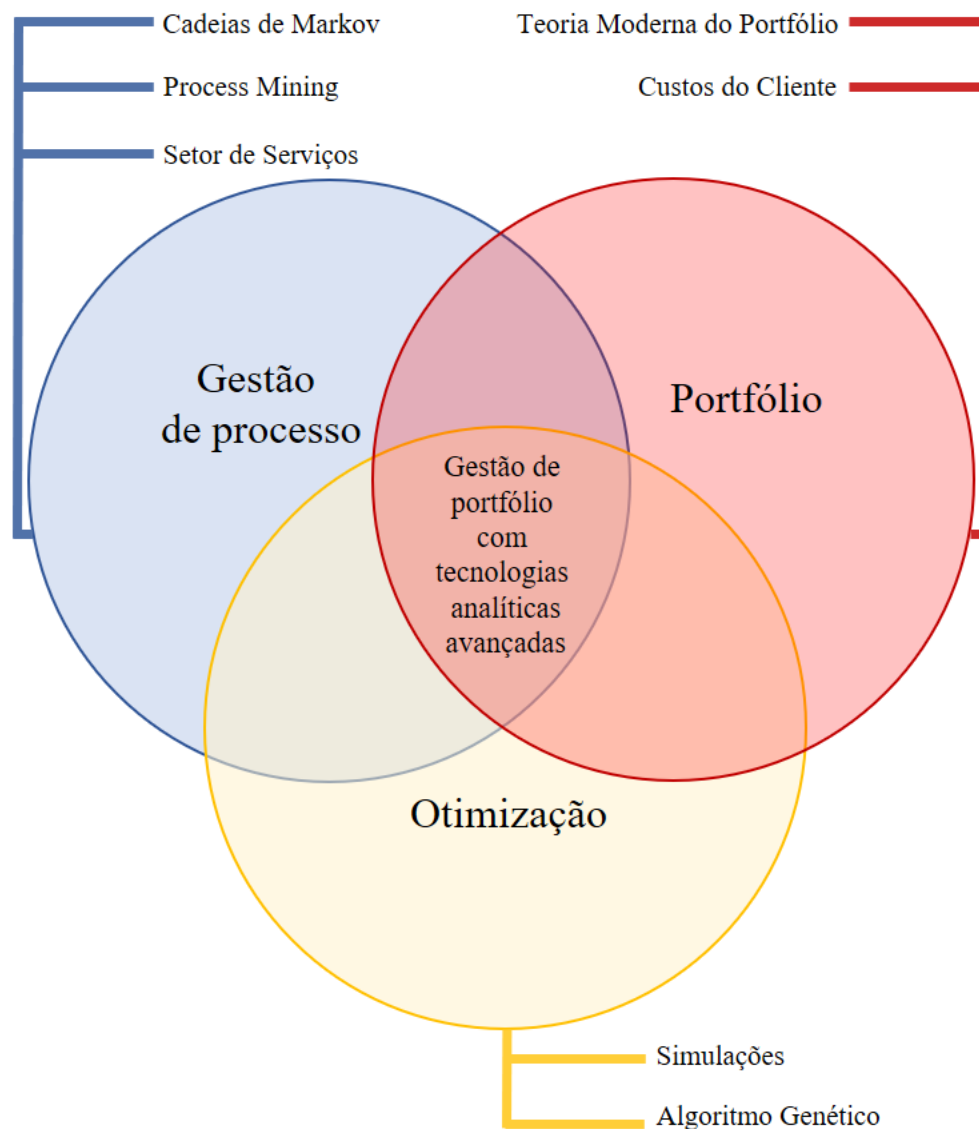
A natureza do trabalho é exploratória, já que investiga evidências para testar as seguintes hipóteses:

*H1: Os dados (logs) do sistema apontam as diferenças dos processos e suas variâncias presentes na prestação de serviço por cliente.*

*H2: O lead time e os desvios dos processos por cliente são dimensões idiossincráticas do custo/benefício do cliente.*

*H3: A governança das estruturas de custos pode ser realizada através de um rebalanceamento da carteira de clientes utilizando variáveis derivadas da análise de processos e receita, através da criação de um portfólio de clientes ótimo.*

Este estudo busca contribuir com a descoberta de conhecimento analítico integrada para o desenvolvimento das organizações em tomarem decisões mais assertivas e encontrarem novos caminhos para enfrentar os constantes desafios, não apenas explorando a teoria, mas gerando experimentos na prática e ofertar também novos caminhos mais conectados com o fenômeno do BD. Segundo Agarwal e Dhar (2014), esta pesquisa pode ser considerada "não tradicional", na medida em que visa desenvolver modelos de fenômenos específicos, conforme a combinação das três técnicas e paradigmas destacados na Figura 1, ou ser um passo para a construção de uma teoria.



**Figura 1:** Diagrama de Venn da combinação de linhas de pensamento do trabalho

Fonte: Elaborado pela Autora.

Neste contexto, a pergunta de pesquisa é: **é possível utilizar o processo operacional, desapegado do senso comum formalizado institucionalmente, utilizando outrossim dados oriundos dos processos, e assim possibilitar novas interpretações, pela modelagem de sua característica probabilística, como direcionador para melhorar as operações de negócio, maximizando o retorno de uma empresa?**

## 1.1 Relevância

Esse trabalho é importante para auxiliar as empresas que enfrentam barreiras no entendimento de seus processos e possuem dificuldades em estabelecer melhorias quando desconhecem como os processos são aplicados na prática. Mais ainda, visa auxiliar suas estratégias de negócio que dependam de um processo bem estabelecido e buscam na gestão destes e no reúso de dados uma forma de encontrar novos caminhos para antigos problemas.

A segunda contribuição é para o desenvolvimento acadêmico do estudo de otimização de processos e visa convidar a participação de pesquisadores de Sistemas de Informação (SI) na exploração de dados para promover análises estatísticas robustas que possam auxiliar negócios e mostrar o potencial que o *Analytics* tem de inspirar novos caminhos.

A terceira contribuição é poder desafiar os modelos tradicionais de modelagem de processos que usualmente requerem a participação de múltiplos agentes em entrevistas de mapeamento, o que se provou ser custoso, lento e incompleto (Colombo et al., 2006). A subjetividade inerente à participação de pessoas pode provocar ruídos de informação que podem pôr à prova a qualidade da produção do trabalho.

Ao posicionar a ciência de dados como um diferenciador que promove melhor diálogo entre academia e o mundo prático, a estrutura metodológica quantitativa usada na prática da resolução de um problema de negócio, conforme apresentado neste trabalho, pode contribuir para o desenvolvimento de melhores técnicas de *Big Data Analytics* (BDA) e inspirações para o uso de diferentes algoritmos combinados, para que tenham cada vez maior encaixe dentro dos projetos das empresas.

## 1.2 Objetivo da pesquisa

O objetivo desta pesquisa é a construção de um conjunto de ferramentas analíticas através da captura dos dados e análise do comportamento real dos processos embasada nas características de probabilidades, obtidas dos *logs* do sistema, para melhor identificar o verdadeiro processo, levando em conta que os clientes têm necessidades específicas e essas especificidades influenciam os processos da empresa. Conseqüentemente, melhorar a carteira de clientes com o *lead time* do processo como entrada. Além disso, reequilibrar o portfólio com algoritmos de otimização com base na teoria moderna de portfólio.

O estudo definiu a construção de um conjunto de ferramentas analíticas estruturadas para capturar e analisar o comportamento real dos processos a partir da captura e uso dos dados em que existe uma interação viva e combinações que podem ser descobertas e representadas dentro dos modelos analíticos propostos. O universo de combinações de estruturas analíticas propostas possibilita novas interpretações e novos resultados para compreender fenômenos que então não seriam possíveis sem a evidenciação obtida por uma gestão de informação moderna e transformativa. Com formulações que tragam novas visões fundadas pelo poder da ciência analítica para sustentar tomadas de decisões que possam ser diferentes e despegadas dos modelos prévios e das abordagens de senso comum para melhorar as operações dos negócios e dar novos destinos aos dados armazenados pelas organizações, que não tenham um uso pré-definido.

### **1.3 Organização do trabalho**

A organização deste trabalho operacionalizou os modelos propostos por meio de um estudo de caso para caracterizar a viabilização prática da teoria matemática, estatística e computacional estudada. A pesquisa busca relatar, no nível organizacional, a condução das análises em uma empresa que busca excelência operacional em seus processos. A organização se inspira em desafiar a prática da proposta de Levitt (1972) de “abordagem da linha de produção” adotada pelas empresas. Para melhorar seu desempenho em custos, as empresas implementariam processos padronizados para alcançar maior eficiência e qualidade (Bowen e Youngdahl, 1998).

A empresa utilizada como caso neste estudo é uma empresa de serviços no setor de Prestação de Serviços Logísticos. A situação de um departamento da empresa fica ainda mais impactada pela busca da excelência operacional em seus processos já que enfrenta desafios de improdutividade e muitas vezes resultados financeiros aquém do desejado (apresentada no Apêndice A), portanto busca a implementação de um processo que impacte positivamente a produção dos serviços. Tal proposta, portanto, está coerente com a estratégia de aplicação dos métodos analíticos adotada no trabalho, a qual foi operacionalizada por meio de estudo de caso. Esse trabalho foi produzido em três etapas de análises que visam segregar a técnica implementada para melhor compreensão da necessidade dela, porém todas elas estão conectadas e uma etapa é dependente da anterior pois trata-se de um incremento e evolução da técnica previamente adotada.

Este trabalho é dividido em 6 etapas.

Nesta introdução é apresentado o tema, a relevância e o objetivo. No capítulo 2, o referencial teórico e o caso, descrevendo pontos relevantes na literatura sobre os instrumentais usados na aplicação do método e é apresentada a empresa a ser estudada, em que o método é aplicado para demonstrar os resultados obtidos na prática. O caso apresenta uma empresa que busca redesenhar seus processos e enfrenta desafios associados a eficiência operacional em um departamento caracterizado por um grande volume de pessoas e um resultado econômico aquém do desejado pelas metas corporativas. Ele reflete o desafio de mapeamento de processos em um ambiente inspirado por Lee, Wyner e Pentland (2008) sobre um negócio em que a complexidade está associada a um grande espaço de interpretação e liberdade quanto ao processo desenhado como referência.

No capítulo 3 é apresentado o método, com alguma antecipação de resultados preliminares, fundamentais para o entendimento do método *per si*. Após isso, no capítulo 4 são expostas as análises dos resultados. O capítulo 5 apresenta a conclusão, implicações gerenciais, limitações e sugestões de estudos decorrentes deste.

## 2 REVISÃO DA LITERATURA

Nesta revisão, são apresentadas 8 seções relacionadas ao tema. Primeiramente, a gestão de processos no panorama organizacional e a evolução dessa disciplina que busca nos dados maior eficácia. Na seção 2.2, aborda-se a disciplina de *Big Data Analytics*. Na seção 2.3, a modelagem de processos estocásticos e probabilísticos que explicam a abordagem matemática para a resolução do problema de modelagem de processos. Na seção 2.4, são expostos alguns desafios encontrados no mapeamento de processos no setor de serviços. Na seção 2.5 é apresentado um outro desafio organizacional encontrado nas análises de custos, já que no setor de serviços, em particular, a lucratividade por cliente é um fator crítico de análise. Essa seção teórica deriva da perspectiva que o trabalho propõe de modelar processos a nível dos clientes e elabora a necessidade de um olhar atento para as variações no negócio devido à influência deles sobre o processo. A seção 2.6 apresenta a importância de uma gestão de portfólio de clientes, para otimizar a alocação dos recursos entre os clientes para balancear os riscos. Na seção 2.7, é apresentada a teoria de algoritmos de otimização. Por fim, na seção 2.8, é apresentado o caso.

### 2.1 Gestão de processos

A busca por melhorias está presente em todas as organizações. Ao olhar para uma organização existem vários aspectos que podem ser melhorados, um deles são os seus processos (Janssen et al., 2015). Um exemplo disso é o *Six Sigma*, criado pela Motorola em 1986. *Six Sigma* se propõe a melhorar a qualidade da produção, identificando e removendo as causas dos defeitos e minimizando a variabilidade durante a fabricação; é amplamente utilizado no setor industrial (Janssen et al., 2015; Motorola University, 2005). Outro exemplo igualmente conhecido é o *Lean Manufacturing*, que é uma "filosofia" amplamente criada com base no sistema de produção da Toyota. Aqui, a ênfase está nas partes do processo de produção que agregam valor ao cliente final, fazendo com que qualquer aspecto que não agregue valor seja considerado um desperdício e, portanto, deve ser eliminado (Janssen et al., 2015).

O *Business Process Management* (BPM) é uma metodologia geral que oferece suporte ao *design*, gerenciamento e melhoria dos processos de negócios, a fim de aumentar a produtividade de uma empresa (Smith e Fingar, 2003). BPM também envolve um sistema de software que

implementa o conceito. O sistema BPM gerencia processos de negócios usando modelos pré-definidos. Um modelo de processo é um modelo legível por computador, que representa atividades, relações de precedência entre eles e recursos (Bae e Seo, 2007).

A reengenharia de processos de negócios (BPR, *Business Process Re-engineering*) é um paradigma que promove um redesenho radical nos processos através da Tecnologia da Informação (TI) para poder alcançar melhorias substanciais na performance operacional (Hammer, 1990). O conceito de BPR tem suas origens na década de 1990 como uma consequência do movimento de gestão da qualidade total (TQM, *Total Quality Management*) da década de 1980 (Gulden e Reck, 1992, DeCock e Hiplin, 1997). Hammer (1990) ressaltou a importância de romper com a velha maneira de fazer negócios e a necessidade de redesenhar o processo de negócios completamente para alcançar melhorias de ordem de magnitude.

BPR foi estudado abundantemente em artigos de gestão (Davenport e Short, 1990; Earl, 1994; Hall et al., 1993; Hammer, 1990; Kaplan e Murdock, 1991). A modelagem das características dos processos de negócios fornece uma visão qualitativa sobre a maneira como os clientes são atendidos. No entanto, a maneira como as atividades internas são executadas para fornecer esse serviço permite a otimização quantitativa com base nas métricas do processo (Franken e Weger, 1997).

Wu (2003) discutiu que uma área de grande preocupação tem sido a dificuldade de convencer a alta administração no sentido de promover o papel estratégico da tecnologia da informação (TI) na reengenharia de processos. Além disso, Davenport (1993) examinou o importante papel que a TI desempenha como um facilitador dos projetos de inovação de processos.

A mineração de processos encontrou um lugar na aplicação da análise de dados no contexto de BPM (Chen e Storey, 2012). Esse novo campo de pesquisa resultante da combinação das áreas de gestão de processos e mineração de dados, é chamado de *Process Mining* (van der Aalst, 1998; van der Aalst et al., 2003). Ele implica no desenvolvimento de ferramentas para gerar *insights* com eventos coletados durante a execução de um processo de negócio (van der Aalst, 2011). Sendo também considerado um campo da Ciência de Dados, a técnica engloba um grande número de modelos analíticos de *Big Data*, focada na análise dos eventos gerados nos *logs* de sistemas durante as atividades dos processos das organizações (Chen e Storey, 2012).

Desde que o trabalho de mineração de processos começou em meados da década de 1990, o objetivo foi principalmente apoiar a análise retrospectiva, por exemplo, construindo

representações visuais dos processos de negócios a partir de dados ou verificando se os processos são executados conforme o esperado (van der Aalst, 2011). Com a crescente popularidade da análise preditiva, surgiu o interesse em usar a mineração de processos para analisar não apenas o passado, mas também o presente e o futuro para obter uma visão abrangente de um processo (Grigori et al., 2004).

## 2.2 *Big Data Analytics*

O interesse em avaliar o valor do negócio, o desempenho da empresa e as soluções de análise de *Big Data* (BD) está aumentando (McAfee et al., 2012; Kiron, Prentice e Ferguson, 2014; Akter et al., 2016). A atenção ao BD vem crescendo significativamente na academia e na prática nos últimos anos (Luvizan, Meirelles & Diniz, 2015; Park & Leydesdorff, 2013). Embora a academia tenha se engajado no tema, com um significativo aumento no volume de publicações, diversos pesquisadores ainda destacam a necessidade de estudos sobre várias questões e aspectos relevantes do BD (Abbasi, Sarker & Chiang, 2016; Boyd & Crawford, 2012; Martin, 2015; Müller et al., 2016; Newell & Marabelli, 2015; Park & Leydesdorff, 2013).

O papel do BD dentro dos negócios foi destacado, especialmente quando se considerou a sua capacidade de lidar com os elementos de volume, velocidade e variedade (Zhou et al., 2014; Duan e Xiong, 2015), ou, de acordo com Wamba, Akter, Edwards, Chopin e Gnanzou (2015), volume, velocidade, variedade, veracidade e valor. O volume tem a ver com a quantidade de dados gerados. A velocidade se refere à rapidez em que os dados são produzidos e sua arbitrariedade de geração. A variedade é um conceito que descreve os diferentes formatos de dados. A veracidade se refere a imprevisibilidade em relação a previsões confiáveis. O valor são os benefícios a serem obtidos por meio do uso de BD.

Gregor, Martin, Fernandez, Stern, e Vitale (2006) apresentaram quatro valores aplicados para soluções de análise de BD: transacionais, estratégicas, transformacionais e informacionais. O valor transacional se refere a a capacidade das soluções de análise de fornecer benefícios operacionais, como redução dos custos operacionais, aumento da produtividade dos funcionários ou economia no gerenciamento da cadeia de suprimentos. O valor estratégico refere-se ao aprimoramento de uma oferta da empresa, por exemplo, em termos de serviço ao cliente ou inovações de produto. O valor transformacional mede a capacidade de uma organização de mudar

a fim de aproveitar as oportunidades de negócios ou transformar seus modelos de negócio. O valor informativo refere-se à capacidade de uma organização de melhorar o fluxo de informações que por sua vez permite um acesso mais rápido e fácil aos dados e os fornece em formatos mais utilizáveis.

Clemons, Dewan, Kauffman e Weber (2017) exemplificaram a obtenção de *insights* gerados através do BD, aplicados a processos de negócios, que podem ser usados para melhorar eficiência, produtividade, acessibilidade e disponibilidade. Segundo Grover, Chiang, Liang e Zhang (2018), muitas empresas têm usado o BD para obter *insights* sobre o desempenho da organização, o que pode enriquecer a inteligência organizacional para tomada de decisão, desenvolver uma estrutura organizacional dinâmica para responder ao mercado e mudanças ambientais, melhorar a utilização da capacidade e aumentar o retorno sobre ativos.

O BD é uma tecnologia que trabalha com informações estruturadas e não estruturadas em grande escala. O dicionário de Tecnologia da Informação da Gartner (2020) define *Big Data* como ativos de informação de alto volume, alta velocidade, alta variedade e que demandam formas inovadoras e eficientes de processamento para uma visão e tomadas de decisão aprimoradas, e automação de processos. Entre os trabalhos existentes, o conceito de BD não é unânime, tornando seu perímetro confuso e sua abordagem acadêmica ainda mais desafiadora (Luvizan, Meirelles & Diniz, 2015; Martin, 2015). Há diversas definições de seu conceito: são ferramentas capazes de manejar volumes de dados superiores aos limites das ferramentas tradicionais; é a capacidade analítica de construir inteligência a partir de grandes volumes de dados; é um novo cenário em que há disponibilidade de dados e de ferramentas para manejá-los sem precedentes; entre outras (Demchenko et al., 2013; Letouzé, 2012; Manyika et al., 2011; Martin, 2015; Park & Leydesdorff, 2013). Martin (2015) propõem abordar o BD como uma indústria para destacar a complexidade do ecossistema no qual ele está inserido. Tal ecossistema envolve não apenas as ferramentas em si, mas também aspectos relacionados à sua infraestrutura, aos dados e à obtenção destes, bem como aos impactos para aqueles que as implementam e as usam — seus parceiros e modelos de negócios, entre outros atores e elementos do cenário geral (Martin, 2015). Segundo Arunachalam, Kumar e Kawalek (2018), BD é frequentemente relacionado ao avanço tecnológico e associado a novos tipos de arquiteturas de banco de dados, como Hadoop, NoSQL e processamento paralelo distribuído de dados.

Segundo Francisco (2015), um termo que melhor referencia tal conceito seria *Big Data Analytics* (BDA), porque o eleva ao potencial de ir além da quantidade de dados apenas.

BDA pode ser dividido em quatro categorias dependendo do resultado entregue. Erl, Khattak e Buhler (2016) e Souza (2014) descreveram as categorias em:

*Analytics* descritivo: é realizada para responder a perguntas sobre eventos que já ocorreram. Esta forma de análise contextualiza os dados para gerar informações (Joseph e Johnson, 2013).

*Analytics* diagnóstico: determinar a causa de um fenômeno que ocorreu no passado. Possui foco no motivo por trás do evento (Bihani e Patil, 2014).

*Analytics* preditivo: é realizada na tentativa de determinar o resultado de um evento que pode ocorrer no futuro. As informações são aprimoradas para gerar conhecimento que transmita como essas informações estão relacionadas. A força e a magnitude das associações formam a base dos modelos usados para gerar previsões futuras baseadas em eventos passados. É importante entender que os modelos usados para análises preditivas têm dependências implícitas nas condições em que ocorreram os eventos passados. Se essas condições mudarem, os modelos que fazem as previsões precisam ser atualizados (Waller e Fawcett, 2013).

*Analytics* prescritivo: baseia-se nos resultados da análise preditiva ao prescrever ações que devem ser tomadas. O foco não é apenas em qual opção prescrita é melhor seguir, mas por quê. Em outras palavras, a análise prescritiva fornece resultados que se fundamentam em uma situação, em um contexto. Assim, este tipo de análise pode ser usado para obter uma vantagem ou mitigar um risco (Szungott, Henne, e von Voigt, 2012).

### 2.2.1 Data Science

Ozdemir (2016) explicou que a ciência de dados é o uso de dados para obter novos *insights* que, de outra forma, teriam sido perdidos. Segundo John e Berndan (2018), abrange um conjunto de princípios, definições de problemas, algoritmos e processos para extrair padrões não óbvios e úteis de grandes conjuntos de dados. Muitos dos elementos da ciência de dados foram desenvolvidos em campos relacionados, como aprendizado de máquina e mineração de dados. Na verdade, os termos ciência de dados, aprendizado de máquina e mineração de dados costumam ser usados como sinônimos. A semelhança entre essas disciplinas é um foco em melhorar a tomada de decisão por meio da análise de dados. No entanto, embora a ciência de dados aproveite esses

outros campos, é mais amplo em escopo. O aprendizado de máquina se concentra no projeto e na avaliação de algoritmos para extrair padrões de dados. A mineração de dados geralmente lida com a análise de dados estruturados e muitas vezes implica uma ênfase em aplicativos comerciais.

### **2.2.2 Data Mining**

Provost e Fawcett (2013) afirmaram que um princípio importante da ciência de dados é que a mineração de dados é um processo com estágios razoavelmente bem compreendidos. Alguns envolvem a aplicação de tecnologia da informação, como a descoberta automatizada e avaliação de padrões de dados, enquanto outros requerem principalmente a criatividade do analista, conhecimento de negócios e senso comum.

Segundo Witten, Frank, Hall e Pal (2017), a mineração de dados é definida como o processo de descoberta de padrões nos dados. É uma técnica pertencente ao processo de KDD (Fayyad et al., 1996) em que dados são coletados, transformados e organizados antes de serem analisados. O resultado desse processo permite a visualização de estruturas categoricamente organizadas para a interpretação humana e a partir desse ponto, permitem criação de valor a partir dos dados.

Os padrões descobertos devem ter sentido para a condução de alguma vantagem, geralmente, vantagem econômica. Os padrões ajudam a explicar algo sobre os dados e geram *insights* ou permitem processos de decisões mais rápidos e com maior precisão. A mineração de dados pode resolver problemas analisando dados que estejam presentes em banco de dados. Algumas das primeiras aplicações foram no varejo, especialmente na forma de análise de cesta de mercado (Voican, 2020).

### **2.2.3 Process Mining**

Algumas abordagens para melhoria de processos se baseiam na ideia de uma representação abstrata de processo, na forma de um diagrama de fluxo de processos. No entanto, esse formato tem apresentado certas falhas quando executado na prática. Por exemplo, os participantes de um negócio podem consultar um fluxo como referência, mas ainda assim possuem a liberdade para interpretar esse modelo formalizado e liberdade para aderirem ou não (Lee et al., 2008). Para conferir maior rigor na descoberta de um processo, originou-se a possibilidade de automatizar o

estudo de processos, com base em dados (Lakshmanan et al., 2012; Chiu et al., 2019; Ibanez-Sanchez et al., 2019; van der Aalst, 2011; Weijters et al., 2006; Goedertier et al., 2009). Essa é uma abordagem alternativa para as técnicas de mapeamento que requerem questionários e técnicas de mediação e envolvimento dos atores no relato de suas atividades (Colombo et al., 2006).

Franken e Weger (1997) estreitaram esse conceito ao mencionar que as modelagens das características comportamentais além de proporcionar informações qualitativas sobre um processo também geram informações quantitativas para otimizações. Sob essa técnica, os dados que descrevem eventos de processo são primeiro capturados de um processo em andamento e, em seguida, usados para gerar um modelo formal do comportamento desse processo.

*Process Mining* foi aplicado dentro do campo médico usando informações de banco de dados para criar modelos gráficos que expliquem em uma linguagem mais amigável o processo clínico (Ibanez-Sanchez et al., 2019). A mineração de processos é uma ferramenta promissora para a execução de auditorias, especialmente testes de controles internos (Chiu et al., 2019). Os autores afirmam que os dados armazenados em *logs* de eventos fornecem aos auditores informações ricas que poderiam servir como evidência de auditoria adicional ao realizar testes de controles ou outros procedimentos de auditoria. Além disso, *logs* de eventos são automaticamente registrados no sistema de TI quando as atividades ou processos de negócios são executados. Portanto, são menos propensos a serem alterados ou distorcidos.

A literatura apresenta um grande número de técnicas de descoberta de processo (Chiu e Jans, 2019; Colombo et al., 2006; González López de Murillas, Reijers e van der Aalst, 2019; Jónás et al., 2014; Lakshmanan et al., 2015; van der Aalst et al., 2011; van der Aalst, 2016). Retratando os diferentes algoritmos obtidos ao longo do tempo, primeiramente na década de 1990, em que as ideias em torno de mineração de dados consideravam apenas a existência de um processo linear e, portanto, a ausência de concorrência (van der Aalst et al., 2003). Em seguida foi estudado o algoritmo do tipo  $\alpha$  que constrói Redes de Petri capazes de expressar processos concorrentes (van der Aalst et al., 2004). Logo em seguida, foi criado uma evolução do anterior, o algoritmo  $\alpha++$ , que podia analisar enlaçamentos. O algoritmo  $\alpha++$  suporta vários tipos de eventos enlaçados e também traz consigo a capacidade de aprender sobre as dependências entre as tarefas que não estão obviamente relacionadas nos dados de *log* (Wen et al., 2007).

Apesar de diversos algoritmos, os desafios presentes nos exercícios de descoberta de processos ainda residem em elementos básicos como a qualidade dos dados. No caso da mineração

de processos, os desafios estão na qualidade e na falta de informação dos *logs* (Breuker et al., 2016). A falta de informação, dada sua complexidade, requer mais atenção para que seja resolvida. Uma possibilidade é projetar técnicas de descoberta que deduzam os comportamentos que estão ausentes nos dados analisados (van der Aalst, 2011). Porém deduções e generalizações produzem modelos incorretos. Muitas técnicas de descoberta de processos, por exemplo, o *Heuristics Miner* (Weijters et al., 2006) e o *AGNEs Miner* (Goedertier et al., 2009), consideram também as anomalias nos dados. Eles analisam a frequência dos padrões de comportamento dos dados e deduzem como normais aqueles que são encontrados com bastante frequência. Outras técnicas de descoberta incluem o *Fuzzy Miner* (Günther e van der Aalst, 2007), que não produz Redes de Petri, mas uma representação visual da importância de tipos de eventos individuais e as correlações entre eles (Bose e van der Aalst, 2012).

A dinâmica de processos sustentados por sistemas é estocástica por natureza; portanto, é importante introduzir o não determinismo como um ingrediente-chave do modelo (Colombo et al., 2006). Cadeias de Markov representam uma classe de processos estocásticos que desperta grande interesse para uma variedade de aplicações práticas. A contextualização de uma situação prática, dada por Jónás, Kalló e Tóth (2014) retratou que em um ambiente de rápidas mudanças com tarefas tão complexas, ferramentas de modelagem foram imprescindíveis para ajudar a análise de processos e Cadeias de Markov ajudaram nesse propósito.

Cadeias de Markov podem ser usadas para o estudo de diferentes casos de negócio; se provaram eficazes na resolução de muitos problemas, desde gestão empresarial até problemas de finanças (Wu and Chuang, 2010), passando por processos (Guerry, 2011) e questões de conformidade (Verlinden et al., 2012).

### **2.3 Processos estocásticos e Cadeia de Markov**

Processos estocásticos ou probabilísticos são considerados processos similares àqueles que ocorrem em ambientes de negócio, em que as atividades são complexas, com múltiplos estados possíveis e muitos caminhos possíveis (Jónás et al., 2014).

Segundo o trabalho de Nogueira (2016), o conceitual matemático é: um processo estocástico é definido com uma coleção de variáveis randômicas  $X(t)$  indexadas por um parâmetro  $t$  pertencente a um conjunto  $T$ . Frequentemente  $T$  é um conjunto de inteiros não-negativos.  $X(t)$

representa uma característica mensurável de interesse no tempo  $t$ . Os processos estocásticos são usados para descrever o procedimento de um sistema sobre um período de tempo, sendo assim, a variável randômica  $X(t)$  representa o estado do sistema no parâmetro  $t$ . Podem ser classificados em:

*Em relação ao Estado:*

- Estado Discreto (também denominado Cadeia):  $X(t)$  é definido sobre um conjunto de estados enumeráveis e finitos.
- Estado Contínuo (conhecido como Sequência):  $X(t)$  quando o conjunto de estados é infinito.

*Em relação ao Tempo:*

- Tempo Discreto:  $t$  é finito e enumerável.
- Tempo Contínuo:  $t$  é infinito.

A propriedade de Markov afirma que a distribuição do estado futuro  $X_{k+1}$  depende apenas do estado atual  $X_k$  e não depende dos estados anteriores  $X_{k-1}, X_{k-2}, \dots, X_1$ . Um processo estocástico é dito ser Markoviano se:

$$P\{X(t_{k+1}) \leq x_{k+1} | X(t_k) = x_k, X(t_{k-1}) = x_{k-1}, \dots, X(t_1) = x_1, X(t_0) = x_0\} = P\{X(t_{k+1}) \leq x_{k+1} | X(t_k) = x_k\} \quad (1)$$

onde:

$t_n \leq t_1 \leq \dots \leq t_k \leq t_{k+1} = 0, 1, \dots$  para a sequência  $k_n, k_1, \dots, k_{t-1}, k_1, k_{t+1}$  ;

$x_k$  é um estado no instante  $t_k$  ;

$x_{k+1}$  é um estado no tempo  $t_{k+1}$  ; e

$P(X_{k+1})$  é a probabilidade de  $X(t_{k+1})$  ser  $x_{k+1}$  no instante  $t_{k+1}$  .

A equação principal expressa que a probabilidade condicional de qualquer evento futuro, dado qualquer evento passado e o estado presente  $X(t_k) = x_k$  é independente do evento passado e dependente somente do estado presente.

Um tipo de processo Markoviano é a Cadeia de Markov em Tempo Discreto (DTMC, *Discrete-time Markov chain*) que permite a modelagem de probabilidades de transição entre estados discretos, geralmente no tempo. Embora sua teoria básica não seja excessivamente complexa, eles são extremamente eficazes para modelar sequências de dados categóricos (Ching et al., 2008). Para ilustrar, algumas aplicações conhecidas desse algoritmo, podem ser encontradas em linguística (Markov, 1907), teoria da informação, com o exemplo do algoritmo original do Google baseado na teoria das cadeias de Markov (Lawrence Page et al., 1999), medicina, com o exemplo da transição entre os estados de gravidade do HIV (Craig e Sendi, 2002), economia e sociologia em que Jones (1997) mostra uma aplicação das Cadeias de Markov para modelar a mobilidade social.

As probabilidades para que os estados sejam transicionados a outros são denominadas Probabilidades de Transição, que representam a probabilidade do estado  $X(t_{k+1})$  ser  $x_{k+1}$  no instante  $t_{k+1}$  dado que o estado  $X(t_k)$  é  $x_k$  no instante  $t_k$ .

Uma matriz estocástica é uma matriz quadrada que descreve as transições em uma Cadeia de Markov. É conhecida como matriz de probabilidades, matriz de transições, ou matriz de Markov (Cayley, 1858). Uma matriz estocástica descreve uma Cadeia de Markov  $X_i$  em um espaço de estados finito  $S$  com cardinalidade  $S$ . Se a probabilidade de passar de  $i$  para  $j$  é  $Pr(j|i) = P_{ij}$ , a matriz  $P$  é fornecida usando  $P_i$  como a linha  $i$  e  $P_j$  como a coluna  $j$ .

$$P = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,j} & \cdots & P_{1,S} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,j} & \cdots & P_{2,S} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{i,1} & P_{i,2} & \cdots & P_{i,j} & \cdots & P_{i,S} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ P_{S,1} & P_{S,2} & \cdots & P_{S,j} & \cdots & P_{S,S} \end{bmatrix} \quad (2)$$

Baseado nas propriedades de Cadeias de Markov, a soma das probabilidades de transição de um estado para todos os outros é 1 para cada uma das linhas, assim:

$$\sum_{j=1}^S P_{i,j} = 1 \quad (3)$$

Uma vez que o estado  $x_k$  é alcançado, a transação continua apenas nele e o processo como um todo termina. Assim o estado  $x_k$  é o convergente, ou seja  $p_{k,k} = 1$  e portanto  $p_{kj} = 0$  para cada  $j \neq k$ .

Para exemplificar, considere probabilidades de transição de estados de temperatura ensolarado, nublado e chuvoso, conforma a Tabela 1.

**Tabela 1:** Exemplo das probabilidades de transição dos três tipos de clima

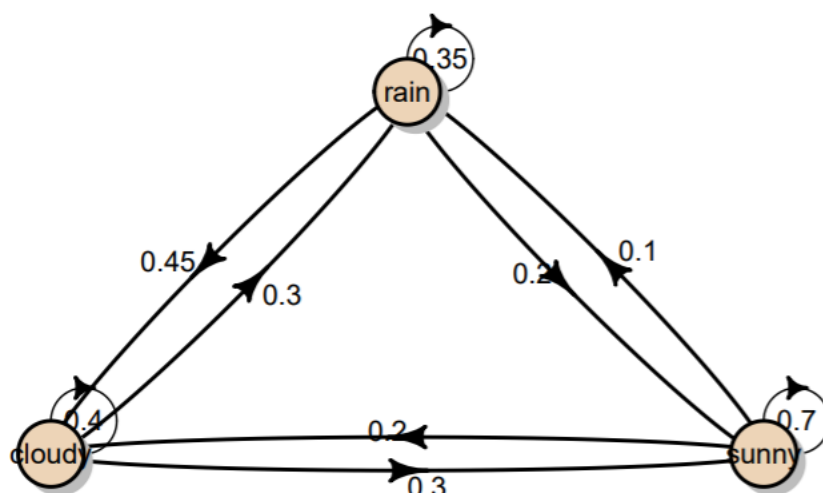
Clima $X_0$	Clima $X_1$		
	Ensolarado	Nublado	Chuvoso
Ensolarado	0,70	0,20	0,10
Nublado	0,30	0,40	0,30
Chuvoso	0,20	0,45	0,35

**Fonte:** Spedicato et al. (2017).

**Notas:** Matriz quadrada que descreve a transição de estados de clima. Por exemplo: se o estado  $X_0$  for ensolarado, a probabilidade de que  $X_1$  seja nublado é de 0,20.

Em  $P_1$ ,  $p_{1,1} = 0,70$  é a probabilidade de que o clima em  $X_1 = \textit{ensolarado}$  já que o clima em  $X_0 = \textit{ensolarado}$  é 0,70.

Esta matriz de transição é representada pelo grafo na Figura 2.



**Figura 2:** Exemplo do grafo com as probabilidades de transição dos três tipos de clima

**Fonte:** Spedicato et al. (2017).

**Notas:** A imagem ilustra o grafo obtido da matriz de transição da Tabela 1. Os nós são os diferentes climas como: nublado (*cloud*), chuvoso (*rain*) e ensolarado (*sunny*). As arestas são as probabilidades de transições de um clima para o outro. De *cloud* para *cloud* a probabilidade é de 0,4; de *cloud* para *sunny* a probabilidade é 0,3; de *cloud* para *rain* a probabilidade é de 0,3. De *rain* para *rain* a probabilidade é de 0,35; de *rain* para *cloud* é de 0,45; de *rain* para *sunny* é 0,2. De *sunny* para *sunny* a probabilidade é de 0,7; de *sunny* para *cloud* a probabilidade é de 0,2; de *sunny* para *rain* a probabilidade é de 0,1

## 2.4 Processos no setor de serviços

A maioria dos conceitos relacionados a processos e qualidade foram projetados para organizações de manufatura e precisam ser modificados para atender às organizações que estão cada vez mais focadas na prestação de serviços (Harmon, 2011).

Ponsignon, Smart e Maull (2012) em estudo que buscava entender sobre como o contexto do setor em que a empresa atua tem um papel importante na abordagem em processos, afirmaram que especificamente no setor de serviços há inconsistência na forma como o desenho de processos é feito. Sousa e Voss (2008, p. 697) relataram (como citado em Ponsignon et al., 2012) que há uma concordância quase que universal sobre como as melhores práticas de melhoria de processos foram feitas por estudo de casos no setor de manufatura somente.

O setor de serviços apresenta características básicas que influenciam o processo produtivo, como o grau de participação do cliente e do funcionário nos processos (Jorge et al., 2011). Brundage, George e Bowen (1995) apontaram em seu modelo de lógica do serviço o papel do consumidor, do funcionário e da técnica de trabalho. Entende-se por lógica do funcionário, segundo Zeithaml, Berry e Parasuraman (1988), a capacidade de os funcionários entenderem suas atividades, a habilidade de desempenharem seu trabalho de forma eficiente e sua motivação em relação ao mesmo.

Como as metodologias de melhoria, qualidade e otimização foram bastante exauridas no setor de indústria e manufatura, Levitt (1972) considerou a “abordagem da linha de produção” como aquela pela qual os serviços poderiam também melhorar significativamente seu desempenho em custos, principalmente por meio de maior eficiência e qualidade (Bowen e Youngdahl, 1998). Na década de 1970, os serviços estavam em ascensão, assim como as frustrações com as ineficiências, baixa qualidade e baixa produtividade característica do setor. Infelizmente, os modelos de gerenciamento de serviços estavam surgindo muito mais lentamente do que o próprio setor (Bowen e Youngdahl, 1998). Assim, entre as propostas de Levitt (1997), uma foi a de

padronização de serviço. A padronização permite previsibilidade, pré-planejamento e controle de processo que, por sua vez, fornece uniformidade na qualidade do serviço. Até o momento da conclusão deste estudo, os métodos de qualidade aplicados à manufatura influenciam as propostas de padronização do setor de serviços e sugerem que um desenho de processo pode ser único e aplicado na produção total da empresa. Este trabalho visa desafiar essa massificação da produção na prestação de serviços.

Para contrabalançar a ideia de padronização de produtos e serviços, foi identificado que o cliente tem necessidades específicas, com um modelo chamado Proposta de Valor do Cliente em que o foco é totalmente nas necessidades e desejos do cliente (Harmon, 2011).

## **2.5 Classificação de custos com foco no cliente**

A estimativa de custos é uma atividade fundamental de muitos processos de engenharia e de negócios (Qian e Ben-Arieh, 2008). O valor do custo da produção é uma informação frequentemente investigada e dificilmente obtida (Braithwaite e Samakh, 1998). Essa questão é um problema já que a lucratividade de uma empresa é dependente do custo gerado na prestação de serviços aos clientes (Guerreiro et al., 2008). No setor de serviços, em particular, a lucratividade por cliente é um fator crítico de análise, porque o lucro na prestação de serviços depende do comportamento do próprio cliente (Guerreiro et al., 2008). Albalaki e Majeed (2018) abordaram a Análise de Rentabilidade do Cliente (CPA, *Customer Profitability Analysis*) como foco de estudo e foco de atenção para a gestão da contabilidade, partindo da premissa de que os clientes são os principais impulsionadores das atividades de custo, e parte deles os critérios mais importantes para explicar os custos e as receitas geradas por uma empresa (None e Griffin, 1997; Howell e Soucy, 1990; Van Raaij, 2005; Kuchta e Troska, 2007).

Kaplan e Narayanan (2001) forneceram um exemplo de um caso de como as informações sobre os custos da prestação de serviços ao cliente podem modificar o relacionamento de uma empresa com clientes que não são rentáveis. Nesse caso, os dois clientes de maior volume foram considerados os menos rentáveis para a empresa. Os autores ainda argumentaram que os clientes podem ser divididos com base no custo do serviço que recebem. Em outras palavras, os clientes podem ser clientes de "alto custo" e "baixo custo para atender". Portanto, diferentes modelos de

negócios funcionarão para diferentes clientes, dependendo de sua posição na divisão dos custos de serviço. Essa prática corrobora os objetivos do CPA.

A literatura revela estudos relacionados à compreensão do custo associado ao atendimento ao cliente na cadeia de suprimentos com base em modelagem de dados (Kone e Karwan, 2011; Sun et al., 2015). Kone e Karwan (2011) modelaram a predição do custo para atender novos clientes em uma rede de distribuição de gás industrial. Eles desenvolveram um novo método para agrupar clientes e aplicaram regressões. No setor de transporte e logística, uma comparação detalhada entre a modelagem com o uso de regressão e de redes neurais foi analisada por Karlaftis e Vlahogianni (2011) para modelagem de custos. Regressões e aprendizado de máquina foram aplicados para estimativa de custo de injeção plástica (Wang, 2007) e estimativa de preço do petróleo (Azadeh et al., 2012).

De forma objetiva, os clientes devem contribuir financeiramente para a empresa prestadora de serviços. Para as empresas operarem com certa vantagem competitiva sustentável, sua função de relacionamento com clientes deve identificar e reter clientes com alto valor ou potencial de lucro (Niraj et al., 2001). Em busca de informações estratégicas e relevantes para as decisões de gerenciamento de operações e recursos para melhorar a lucratividade, o CPA é uma ferramenta essencial de apoio à decisão (Cardinaels et al., 2004).

## **2.6 A gestão de portfólio de clientes**

O conceito de diversificação existe há algum tempo e é bastante intuitivo. Entretanto, a teoria que a sustenta e sua aplicação na área de investimentos só veio surgir em 1952, com a publicação de Harry Markowitz e seu famoso artigo de seleção de portfólio. Além de evitar problemas relacionados com o investimento direcionado em apenas um produto, portfólios também ofertam retornos com menor volatilidade, representado por uma medida estatística conhecida como desvio padrão. Foi o primeiro a considerar o desejo da diversificação dos investimentos (Rubinstein, 2002). Markowitz foi o primeiro a aplicar variância ou desvio padrão como medida de risco.

Para identificar qual o melhor portfólio, Markowitz estabeleceu uma fundação matemática. Considere um portfólio com  $n$  ativos diferentes, no qual o ativo  $i$  dará o retorno  $R_i$ . Consideremos  $\mu_i$  e  $\sigma_i^2$  serem respectivamente a média e a variância e  $\sigma_{ij}$  a covariância entre  $R_i$  e  $R_j$ . Agora

suponha que a quantia relativa dos valores investidos no ativo  $i$  é  $x_i$ . Se  $R$  é o retorno do portfólio todo então:

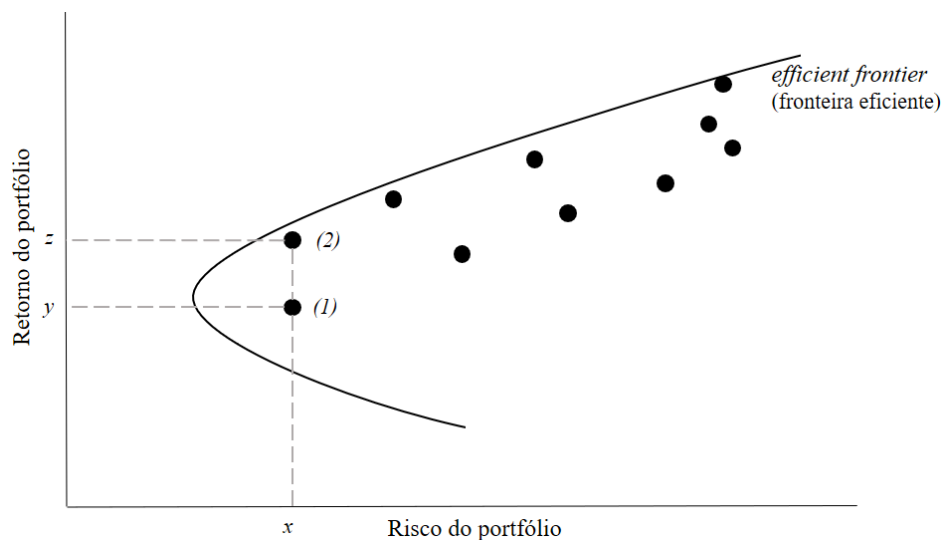
$$\mu = E[R] = \sum_{i=1}^n \mu_i x_i \quad (4)$$

$$\sigma^2 = Var[R] = \sum_{i=1}^n \sum_{j=1}^n \sigma_{ij} x_i x_j \quad (5)$$

$$\sum_{i=1}^n x_i = 1 \quad (6)$$

Para quantidades diferentes de  $x_1, \dots, x_n$  o resultado das combinações para  $\mu$  e  $\sigma^2$  será diferente também. Um conjunto com todas as possibilidades de  $\mu$  e  $\sigma^2$  é chamado de conjunto atingível (*attainable set*). O espaço com o menor  $\sigma^2$  para uma  $\mu$  ou maior e máximo  $\mu$  para um  $\sigma^2$  ou menor é chamado de conjunto eficiente (*efficient frontier*). Um investidor que busca maior lucro com o menor risco, quer maximizar  $\mu$  e minimizar  $\sigma^2$  e portanto ele deveria escolher um portfólio que dê algumas combinações de resultados dentro do espaço eficiente (Figura 3).

A utilização da teoria referente a portfólio é utilizada em diferentes áreas do conhecimento, como na matemática, na saúde e na administração. Utiliza-se a gestão de portfólio em Pesquisa e Desenvolvimento (P&D), para otimizar a alocação dos recursos entre os projetos para balancear os riscos, benefícios e para alinhar os projetos com a estratégia da empresa (Dickinson, Thornton e Graves, 2001). Existem múltiplos projetos nas organizações com riscos e retornos diversos para serem escolhidos pelos executivos das empresas, e a gestão desse portfólio pode auxiliar os gestores. Outra utilização do conceito de portfólio são os portfólios de novos produtos em desenvolvimento, que são compostos por projetos que resultam em novos itens que podem ser comercializados (Kumar, Ajjan e Niu, 2008).



**Figura 3:** Fronteira eficiente de um portfólio de investimentos

**Fonte:** Adaptado e traduzido de Markowitz, 1952.

**Notas:** Os pontos 1 e 2 possuem o mesmo valor de risco  $x$ , porém retornos diferentes. O ponto 1 tem o retorno  $y$  e o ponto 2 o retorno  $z$ , no qual  $z > y$ , tornando este último mais atraente.

A análise de portfólio de clientes é um conceito de relacionamento *Business to Business* (B2B) usado para analisar as relações das organizações, vendedora e compradora, a fim de ajudar os gerentes a alocar recursos organizacionais escassos e garantir a lucratividade de longo prazo dos relacionamentos com o cliente. Essa teoria confronta em especial a visão míope do desenvolvimento de fortes relacionamentos sob o princípio estrito de pura retenção do cliente. Autores sugeriram que a empresa deveria ajustar seu sistema de relacionamento com o cliente de acordo com o valor do cliente e concentrar seus esforços em gerenciar com diferenciação a variação dos diferentes perfis (Johnson e Selnes, 2005).

Modelos de portfólio de clientes foram propostos para que a composição ótima dos relacionamentos com clientes fosse determinada avaliando-se a atratividade relativa dos relacionamentos com clientes em relação à posição competitiva da empresa vendedora (Fiocca, 1982; Campbell e Cunningham, 1983). Outros pesquisadores se concentraram no estudo do custo para servir (*cost to serve*) os clientes sustentados pela lucratividade dos relacionamentos (Shapiro et al., 1987; Krapfel et al., 1991). A estratégia pode corresponder a uma variável que influencia também a composição de uma carteira de clientes (Pardo e Salle, 1995). Fiocca (1982), ao revisar as dimensões que influenciavam as carteiras de clientes existentes, identificou a abordagem estratégica da conta para o desenvolvimento e análise da carteira. As sub variáveis relevantes para

sugerir um determinado modelo de carteira de clientes dependem principalmente do comportamento do cliente, da natureza do produto, da indústria, das características da empresa e da preferência de sua gestão (Wind e Mahajan, 1981).

Terho (2008) afirmou que desde o início da década de 90 muitas propostas de modelagens de portfólio de clientes foram apresentadas como ferramenta para a gestão estratégica de relacionamentos com base na premissa de que os recursos da empresa são limitados e, portanto, não podem ser usados na mesma dosagem para todos os clientes. Seria irracional tratar todos os relacionamentos da mesma forma, sendo assim preferível diferenciar essa alocação de recursos com base no valor do relacionamento. Turnbull (1990) afirmou que a empresa deveria então organizar todo o portfólio de relacionamentos e analisar se possui um portfólio seguro para manter uma performance de longo prazo.

Há estudos que mostraram dispersão da distribuição do lucro das empresas entre seus clientes (Jacquelyn, Reinartz e Kumar, 2004; Niraj et al., 2001; Reinartz e Kumar, 2000, 2003; Storbacka, 1997), o que evidencia a relevância de mais estudos na gestão de portfólios. Eng (2004) realizou um teste empírico sobre algumas dimensões mais analíticas para análise de portfólio como lucro, Retorno sobre o Investimento (ROI, *Return on Investment*) e taxa de crescimento. Os resultados mostraram que havia a necessidade de criar um conjunto de dimensões de variáveis que pudessem melhor descrever o portfólio de relacionamento.

Shapiro, Rangan, Moriarty e Ross (1987) sugeriram analisar os clientes sob duas dimensões: receita e o custo para servir (*cost to serve*). Associou essas duas variáveis a um quadrante que retratou a dinâmica em que atuavam a parte vendedora e a parte compradora.

Preço	Alto	Passive (passivo)	Carriage trade (classe alta)
	Baixo	Bargain basement (barganheiro)	Aggressive (agressivo)
		Baixo	Alto
		Custo	

**Figura 4:** Matriz de classificação dos clientes

**Fonte:** Adaptado e traduzido de Shapiro et al., 1987.

**Notas:** Matriz construída com o comparativo do preço do produto ou serviço que o cliente paga à empresa versus o custo da empresa em manter o cliente. Dessa matriz, são encontrados quatro (4) tipos de clientes: Passivo, Classe Alta, Barganheiro e Agressivo.

O eixo vertical representado no modelo é a receita, composta pelas características de receitas baixas, na parte inferior do gráfico e altas para a parte superior. O eixo horizontal é o custo para servir (*cost to serve*) o cliente, sendo este baixo na parte esquerda e alta na direita. Deste quadrante surgiram 4 perfis chamados de: *Passive*, *Bargain basement*, *Carriage trade* e *Agressivo*. O *Carriage trade* custa muito para servir, mas está disposto a pagar mais caro também. O perfil totalmente oposto a ele, são os clientes de *Bargain basement* - sensíveis ao preço e insensíveis ao serviço e qualidade. Eles podem ter um serviço mais barato do que o *Carriage trade*. Similarmente, os clientes *Passive* também custam menos à empresa, mas eles estão dispostos a aceitar preços altos. Essas contas geram resultados altamente lucrativos. Existem várias razões para que isso aconteça. Em alguns casos, o produto é insignificante demais para justificar uma posição de negociação difícil sobre o preço. Outros clientes são insensíveis ao preço porque o produto é

crucial para sua operação. Outros ainda permanecem com seu fornecedor atual, mais ou menos independentemente do preço, por causa do custo da troca. Os clientes *Agressivos*, por outro lado, exigem (e geralmente recebem) a mais alta qualidade de produto, o melhor serviço e preços baixos. Compradores agressivos geralmente são poderosos. A sua prática de comprar em grandes quantidades lhes dá vantagem junto aos fornecedores na busca de acordos de preços e mais serviços.

## 2.7 Algoritmos Heurísticos

Mesmo que as empresas usem programas de computador para programar o processo de tomada de decisão, o mecanismo é muitas vezes repleto de erros causados pelo fator humano e mudanças frequentes e rápidas nos fenômenos reais. Uma solução para este problema é, entre outros, uma aplicação de métodos de inteligência (heurísticos e metaheurísticos) que podem encontrar em um tempo relativamente baixo a solução que está perto do ideal e aceitável do ponto de vista do problema (Burduk et al., 2019). A prática confirmou que métodos heurísticos podem ser implementados com sucesso em processos de produção (Kotowska et al., 2018; Górnicka et al., 2018; Burgy e Bulbul, 2018; Paprocka et al., 2017). O uso de algoritmos em produção e processos logísticos são mostrados em muitas publicações. Algoritmos estão sendo usados em otimização de processos, como por exemplo, no gerenciamento do transporte (Bożejko et al., 2017), agendamento (Chen et al., 2018; Bożejko et al., 2008), sincronização de entrega (Gdowska e Książek, 2017) ou planejando e ajudando na tomada de decisão (Ahmad et al., 2017).

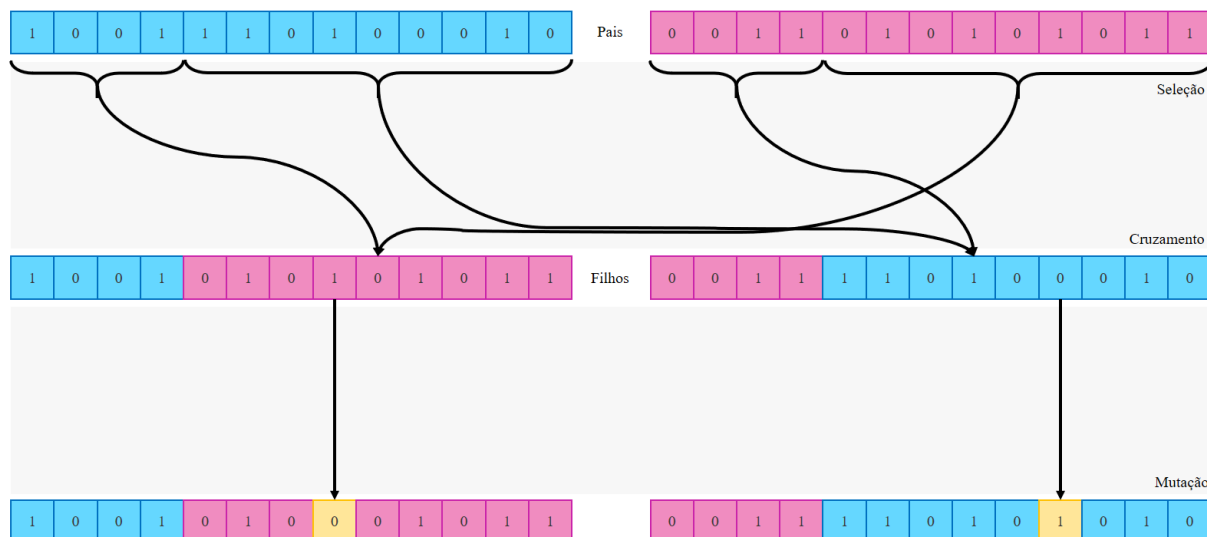
### 2.7.1 Algoritmos Genéticos

Algoritmos genéticos (GA, *Genetic Algorithms*) são técnicas de pesquisa amplamente utilizadas para aproximar soluções ótimas na análise de problemas complexos (Goldberg, 1989; Gen e Cheng, 2000; Lim, 2014). Mimetizando o processo evolutivo da natureza, a técnica é baseada na teoria da evolução, na qual a aptidão de um indivíduo determina sua capacidade de sobreviver. Ele espelha a evolução das populações na natureza com os operadores genéticos de seleção, cruzamento (*crossover*) e mutação para melhorar a qualidade da solução à medida que o processo de otimização prossegue (Goldberg, 1989).

Akpınar (2019) explicou que os algoritmos genéticos foram primeiramente aplicados nas áreas de redes neurais, em seguida, em aprendizado de máquina e recentemente no que agora é chamado de computação evolucionária. O objetivo principal em todos esses campos de aplicações é desenvolver uma série de soluções para um determinado problema de cálculos inspirados na variação genética e seleção natural (Akpınar, 2019). Os GAs são abordagens que tentam encontrar a solução ideal e são baseados em metodologias evolutivas como a biologia evolutiva, que estuda herança, aprendizagem, seleção e mutação.

O algoritmo genético é simples e tem uma excelente capacidade de pesquisa global, da qual o controle dos parâmetros tem uma influência significativa no seu desempenho. Além da definição de parâmetros de controle, o algoritmo em si é a chave para decidir se a computação genética é bem-sucedida ou não. Há três principais parâmetros de controle de genética no algoritmo: tamanho da população, probabilidade de cruzamento e probabilidade de mutação (Xue et al., 2019).

O GA começa com uma população inicial de indivíduos gerados aleatoriamente. Cada indivíduo na população representa uma solução potencial para o problema. Os indivíduos evoluem por meio de iterações sucessivas, chamadas gerações. A aptidão de um indivíduo é avaliada pelo valor retornado de uma função objetivo. Todas as gerações são criadas através dos parâmetros genéticos como seleção, cruzamento (*crossover*) e mutação. A seleção permite a propagação das características de alta aptidão na população, eliminando os cromossomos mais fracos desta. O cruzamento (*crossover*) combina as informações provenientes de diferentes indivíduos da população. A mutação permite que o GA explore o espaço de pesquisa não acessível pelo operador de *crossover* (İnkaya e Akansel, 2017; Urso et al., 2019).



**Figura 5:** Representação do processo de seleção, cruzamento e mutação

**Fonte:** Adaptado e traduzido de Urso et al., 2019.

**Notas:** Dois indivíduos aleatórios de uma população, chamados Pais, possuem um conjunto de genes, denominados cromossomos. Eles são selecionados para passar seus genes para a nova geração, os Filhos, por um cruzamento. Um gene aleatório dos filhos sofre mutação.

## 2.8 CASO: Empresa de serviços do setor logístico

Para o caso, foi avaliado um departamento de uma empresa multinacional no setor de logística e serviços, uma PSL, que possui como clientes empresas que enfrentam desafios na sua cadeia de suprimentos e buscam constantemente eficiência operacional. Lieb, Millen e Wassenhove (1993) utilizaram a seguinte definição para as PSLs: “A logística de terceiros envolve o uso de empresas externas para executar funções de logística que tradicionalmente são executadas em uma organização. As funções desempenhadas pelo terceiro podem abranger todo o processo logístico ou atividades selecionadas dentro desse processo”.

Para esse trabalho, foi avaliado um caso de uma empresa multinacional no setor de logística e serviços. É uma empresa de logística global com sede em Seattle, Washington. A empresa emprega profissionais treinados em 177 escritórios distritais e em várias filiais localizadas em seis continentes, vinculadas a uma rede mundial integrada por meio de um sistema integrado de gerenciamento de informações. Os serviços incluem a consolidação ou o transporte de frete aéreo e marítimo, declaração alfandegária, consolidação de fornecedores, seguro de carga, transporte com horário definido, gerenciamento de pedidos, armazenamento e distribuição e soluções

logísticas personalizadas. A autora obteve facilidade no acesso aos dados transacionais e portanto, a sua escolha se deu por convêniência.

Segundo Porter (1998), as forças para se manter a vantagem competitiva abrangem não apenas as atividades de entrega de valor da empresa, mas também de seus fornecedores, clientes, fornecedores dos fornecedores e clientes dos clientes. Este entendimento ajuda a determinar quais as atividades da cadeia logística serão executadas pela empresa em questão e quais serão executadas pelos seus parceiros. Desde sua origem na década de 1980, as PSLs vêm surgindo como uma opção para ajudar a aumentar a eficácia da cadeia de suprimentos (Maloni e Carter, 2006). E com o seu crescimento, o mercado tem pressionado essas empresas a terem maior amplitude e aumentarem o escopo de serviço, ganhando maior valor e inclusive incorporando um papel estratégico na cadeia de suprimentos (Zacharia et al., 2011). De acordo com Cohen e Roussel (2014), os serviços logísticos têm um papel muito importante para gerar diferenciais estratégicos sustentáveis para as empresas, especialmente um diferencial na eficiência operacional, em que o serviço logístico promove eficiência com a centralização ou compartilhamento de ativos e consequente redução de custos. Porém ainda existe um potencial em um diferencial mais atuante, em que o serviço logístico auxilia a reduzir *lead times* e garantir flexibilidade e velocidade nas entregas. Além de estratégia, Sahay e Mohan (2006) avaliaram estudos econômicos para lidar com as atividades logísticas de forma eficaz e eficiente e, portanto, consideraram abarcar a função logística internamente, com prestadores do serviço, ou possuir subsidiárias de logística através da criação ou compra de uma empresa de logística, ou poder terceirizar a função e comprar o serviço. Até o momento da conclusão deste estudo, cresce o interesse pela terceira opção, ou seja, terceirização de funções logística para PSL. Um exemplo de tal benefício de natureza econômica, que propõe terceirizar as atividades de transporte, é uma alternativa que reduz o investimento de capital em instalações, equipamentos, tecnologia e mão de obra (Razzaque e Sheng, 1998).

Lieb, Millen e Wassenhove (1993) identificaram serviços mais comumente executados por essas organizações: operação ou gerenciamento de armazéns; consolidação de cargas; gestão da informação; operação ou gerenciamento de frotas; negociação de frete; seleção de transportadoras; emissão e gerenciamento de pedidos; importação e exportação; montagem ou instalação de produtos; segregação de produtos para clientes; e reposição de estoque.

### 2.8.1 Processos do desembaraço aduaneiro

As empresas que oferecem serviços de desembaraço aduaneiro como parte do processo de importação precisam passar por vários estágios para gerar uma Declaração de Importação (DI) antes de esta ser liberada pela alfândega. A DI precisa ser confeccionada a partir de documentos importantes para a legislação local. O despacho de importação é o procedimento mediante o qual é verificada a exatidão dos dados declarados pelo importador em relação à mercadoria importada, aos documentos apresentados e à legislação específica, com vistas ao seu desembaraço aduaneiro (Receita Federal, 2020).

Outro fator considerado no comércio mundial é o número cada vez maior de acordos comerciais regionais para reduzir as restrições comerciais entre dois ou mais países de uma região. Para tirar proveito de tal acordo especial, uma empresa precisa fornecer a documentação necessária (Hausman et al., 2010).

Globalmente os governos requerem uma fatura comercial, listando os itens adquiridos pela entidade importadora. Muitas vezes este documento é emitido no idioma inglês e não é suficiente para que os analistas da Receita Federal tenham total compreensão dos itens embarcados e portanto, a confecção de uma DI precisa trazer a tradução das mercadorias embarcadas, listadas na fatura comercial (Receita Federal, 2019). Outra razão para a emissão de um DI é a tratativa de outros atributos, essenciais para identificar o conteúdo embarcado, como quantidade importada, valor comercial, fabricante, país de origem, aeroporto ou porto de chegada, lembrando que existem dispositivos legais que sustentam esse modelo (Lee e Silverman, 2008).

Nesse aspecto, alguns clientes se antecipam e proativamente organizam seus produtos que venham a ser importados, de forma a estarem traduzidos e dimensionados para que a empresa possa trabalhar com um banco de dados limpo e que possa influenciar em sua produtividade. Enquanto que outros agregam uma nova camada de complexidade e compartilham informações conforme a demanda.

Além disso, os países possuem normalmente um sistema de classificação de mercadorias com o fim de organizar, em um sistema de codificação, o que é a mercadoria (Meixell e Gargeya, 2005). A Nomenclatura é um sistema ordenado que permite, pela aplicação de regras e procedimentos próprios, determinar um único código numérico para uma dada mercadoria. Esse código, uma vez conhecido, passa a representar a própria mercadoria (Receita Federal, 2021). No Brasil é utilizado o código de Nomenclatura Comum do Mercosul (NCM). Há uma preocupação

para que essa classificação, bem como a tradução para o português, seja feita pelos clientes, devido aos riscos que a empresa não quer assumir e arcar com o ônus de uma penalidade aplicada a classificações e traduções incorretas, e também para delimitar claramente onde as responsabilidades da empresa e de seus clientes estão, dentro de um processo nem sempre trivial.

Para confeccionar uma DI, os dados são atualizados no sistema por meio de digitação, o que representa um trabalho extremamente manual e moroso. Esse método, mesmo que antiquado para os padrões atuais de dinamismo digital e tecnológico, visto que grande parte das empresas possuem dados transacionando em formato eletrônico, acaba por ser ainda muito comum para o departamento analisado. Informações necessárias para a digitação da declaração ainda são compartilhadas em formato documental. Alguns clientes buscaram evoluir através de automações que geram planilhas eletrônicas ou arquivos texto que, ao serem compartilhados com a empresa, contribuam com a produtividade de todos os atores envolvidos. Outros ainda buscaram uma solução para troca eletrônica de arquivos (EDI, *Electronic Data Interchange*), inteligentemente garantindo maior velocidade ao trabalho, porém não é frequentemente encontrado (Grainger, 2008). A adoção de EDI é importante para aprimorar a comunicação, coordenação e colaboração além das fronteiras organizacionais (Lin, 2014).

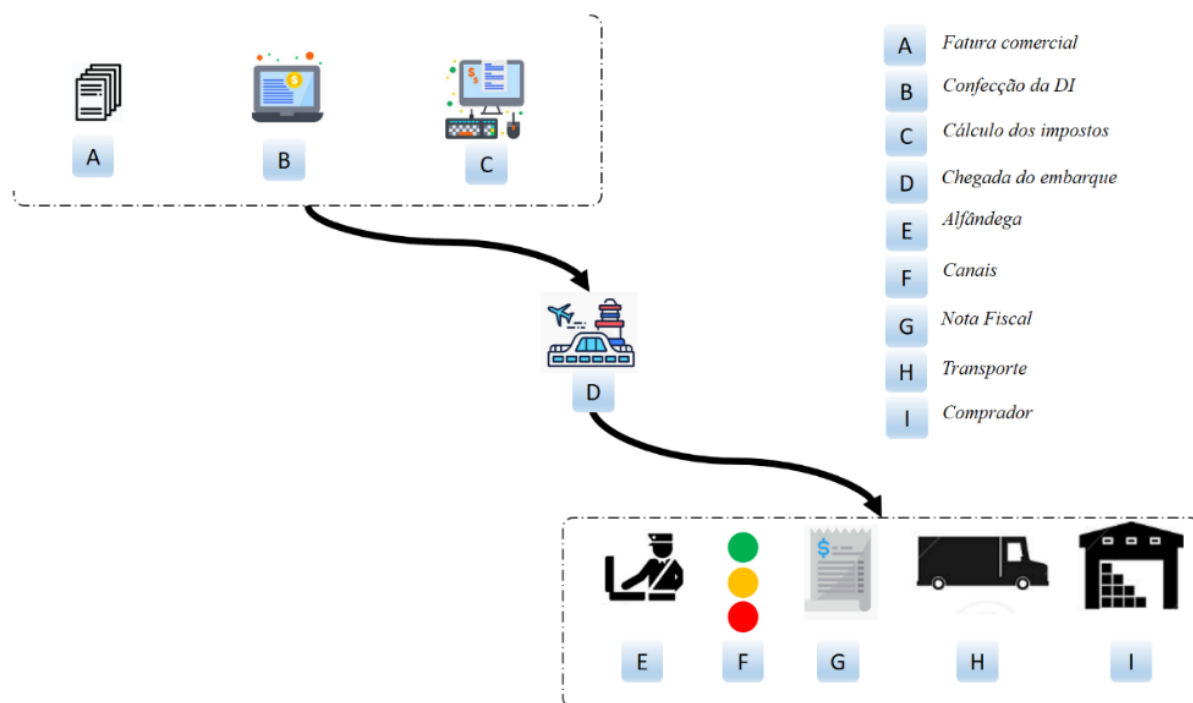
Um outro agravante dessa atividade decorre do volume de itens importados que acaba por gerar uma fatura comercial com uma grande quantidade de itens. Esse cenário em que a quantidade de itens é alta e o cliente não fornece arquivos eletrônicos corrobora a extrema intervenção humana e ineficiências que tornam todo o trabalho repleto de ameaças quanto à acuracidade das informações (Teo, Tan e Wei, 1997).

Após digitada, uma DI precisa ter seus impostos calculados que precisarão ser pagos pelo importador para a Receita Federal. Os impostos calculados são compostos de: Imposto de Importação (II), Imposto de Produtos Industrializados (IPI) quando aplicável, Programa de Integração Social (PIS), Contribuição para o Financiamento de Seguridade Social (COFINS) e o Imposto sobre Mercadoria e Serviços (ICMS). Adicionalmente, o cálculo retorna alguns dispositivos legais para que os impostos possam ser reduzidos ou mesmo suspensos devido a certos NCMs com incentivos fiscais. Alguns NCMs podem gerar também a necessidade de licenças especiais junto a órgãos como o INMETRO e a Anvisa. Todos esses resultados são conferidos e auditados após a digitação e cálculos da DI.

Na sequência, é feita a transmissão da declaração para a Receita. Um passo de extrema importância é o gerenciamento do dinheiro para o pagamento dos impostos. Alguns clientes deixam disponível o valor monetário e as suas contas bancárias para que o débito ocorra. Para que haja conformidade, é preciso uma troca de comunicação contendo a clara autorização dos departamentos financeiros. Dessas práticas, são necessários acompanhamentos diários e pontuais. Outros clientes já adiantam o dinheiro para a conta bancária da empresa e dessa conta de crédito são debitados os valores das declarações. Outros acompanhamentos com autorizações e clareza entre os atores ocorre também.

A partir desse ponto, espera-se a resposta da liberação da declaração por meio de canais aduaneiros. Esses canais são definidos por um sistema de cores, que são: verde, significando que a mercadoria pode ser nacionalizada; amarelo, indicando que é necessária a inspeção de documentos; e vermelho, que exige além da inspeção de documentos, a inspeção física da carga. Os dois últimos canais adicionam trabalho extra resultado de múltiplas interações e intervenções, por serem necessários a coordenação de documentos adicionais e também a coordenação de uma equipe de inspeção física.

Após a liberação da mercadoria, é preciso tratar o pagamento do ICMS junto à Secretaria da Fazenda do Estado (SEFAZ), que de forma descentralizada oferta para cada estado uma plataforma desenvolvida para a submissão do pagamento. Feito isso, o transporte da mercadoria é coordenado com uma empresa transportadora e é feita a emissão da Nota Fiscal de Nacionalização com os valores da Declaração e outros novos atributos necessários para a nota, como a Situação Tributária (ST), Código Fiscal de Operação e Prestação (CFOP) entre outros, para acompanhar o carregamento. Assim, a mercadoria pode então circular nacionalmente em cumprimento à legislação fiscal até a localização física do comprador.



**Figura 6:** Processos do Desembaraço Aduaneiro

**Fonte:** Elaborado pela Autora.

**Notas:** A fatura comercial (A), a confecção da DI (B) e o cálculo dos impostos (C) são passos executados na sequência e podem ser feitos antes da chegada da carga física no aeroporto (D). Depois, podem ser feitos o registro na Alfândega (E), a classificação dos canais (F), a emissão da Nota Fiscal (G) e a chegada na fábrica do comprador (I).

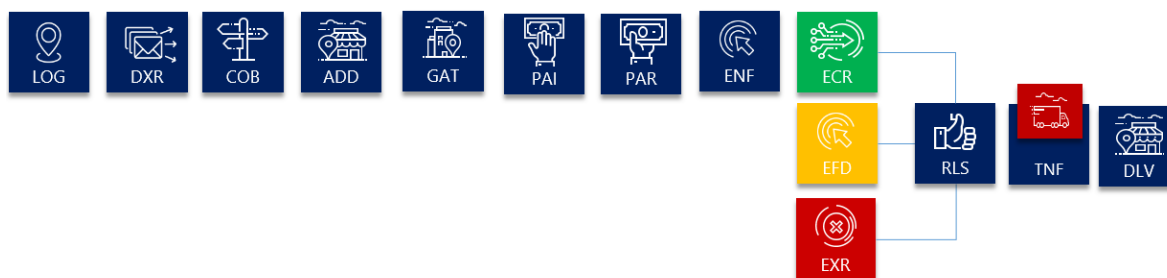
Essa descrição é apenas uma visão geral de atividades muito comuns para o dia-a-dia operacional do departamento. Para a gestão do departamento, essas atividades demandam a atualização do sistema de informação utilizado. Elas ocorrem por meio de **eventos** gerados dentro do sistema, sendo que alguns ocorrem de forma automática e outros manualmente, através da intervenção humana. Além do processo padrão, cada cliente geralmente adiciona mais tarefas ao processo por alguma razão particular, como por exemplo, a necessidade de documentações extras, notificações específicas, coordenação com o seu departamento de logística e também financeiro. Alguns precisam da confecção de relatórios de acompanhamento. Outros requerem que a digitação da DI seja feita de forma personalizada, preenchendo certos campos do sistema, visando com que a informação gerada seja utilizada para novas finalidades.

O processo é composto por etapas, que representam a sequência do trabalho a ser prestado, para conformidade operacional. É importante que os processos sejam bem gerenciados no

comércio internacional. Não é apenas para controlar os prazos de entrega de serviços e os custos operacionais; erros em processos também podem resultar em penalidades ou multas significativas e causar atrasos adicionais. Processos transacionais mal gerenciados também podem significar que uma empresa pode deixar de obter receita (Hausman et al., 2010).

Cada etapa é representada por um evento (apresentada no Apêndice B).

O processo publicado na empresa é:



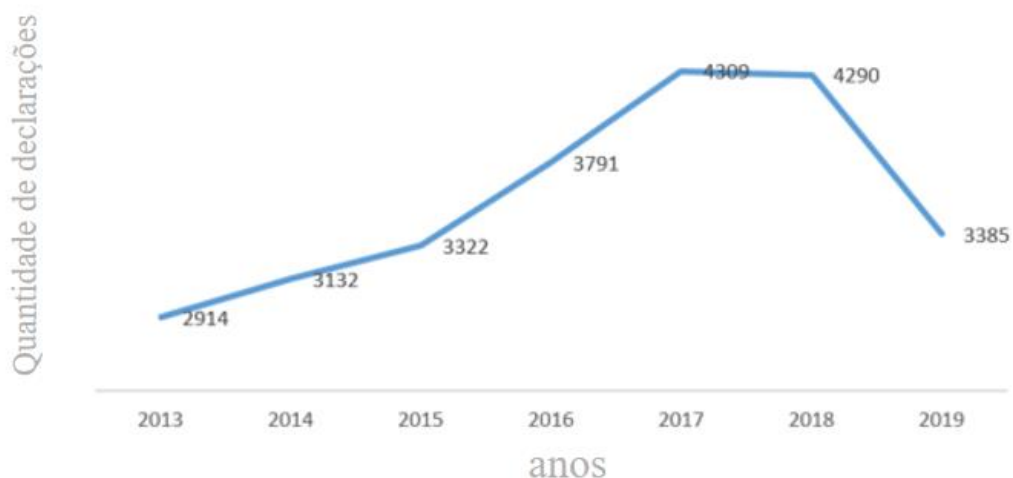
**Figura 7:** Processo operacional institucionalizado pela empresa e seus eventos

**Fonte:** Empresa analisada.

**Notas:** Cada caixa representa um evento, codificado em três (3) letras para representar o fenômeno. O significado dos códigos está disponível no Apêndice B.

Explorando mais a fundo a gestão do departamento, vale mencionar que, ao final do ciclo mensal existem indicadores de resultado financeiro para o departamento que são afetados pela dinâmica dos próprios processos. Além da receita gerada, são considerados os custos gerados por meio de folha de pagamento, espaço utilizado pela equipe relativo ao aluguel do andar, multas absorvidas; esses números contribuem para a identificação do custo operacional agregado do departamento. Quanto à medição de performance, em termos de crescimento, é levado em consideração a quantidade de declarações realizadas por mês, por cliente, para derivar análises, como por exemplo, encontrar os clientes de maior importância. A argumentação atual é de que os clientes mais importantes são aqueles com maior volume de declarações. Outros indicadores internos e operacionais estão relacionados ao tempo levado para execução de cada etapa do processo.

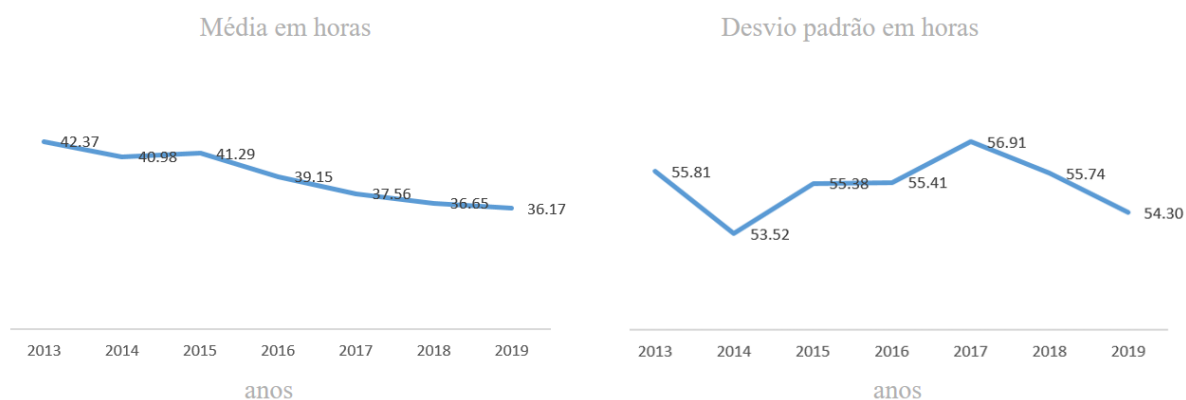
O volume de declarações de importação anuais tem mostrado crescimento, como mostra o Gráfico 1.



**Gráfico 1:** Quantidade de declarações por ano

**Fonte:** Elaborado pela Autora.

O tempo é contado em horas ou dias e além de ser monitorado internamente, é também um medidor externo, avaliado pelos clientes com relação à agilidade da prestação de serviços da empresa. A média e o desvio padrão em horas das transições de eventos, por ano, estão no Gráfico 2.



**Gráfico 2:** Média e desvio padrão do *lead time* dos eventos por ano

**Fonte:** Elaborado pela Autora.

**Notas:** Para cálculo dos valores sem interferência de *outliers*, os elementos da amostra estão dentro do Percentil de 5 a 95.

Algumas revisões de performance junto aos clientes se dão com base no tempo médio para executar alguns macro-processos, como o tempo decorrido entre o recebimento da documentação até a transmissão para a Receita, ou o tempo entre a transmissão para a Receita e a liberação total da carga, ou ainda o tempo entre liberação da carga e a chegada à planta do cliente. Sob a estrutura teórica da economia dos custos de transação (Williamson, 1981), as organizações buscam minimizar os custos necessários para realizar uma transação econômica.

Ao final de tudo, o faturamento é feito com base nos valores dos impostos calculados, além das taxas de honorários. Então é feito o envio do faturamento ao cliente. O departamento financeiro, então, fica responsável pelo monitoramento das contas contábeis para se assegurar de que os pagamentos foram feitos e que a contabilidade está em dia.

Até o momento da conclusão deste estudo, existem frentes de projetos que buscam melhorar a eficiência operacional por meio de melhoria de processos. A maioria delas visa diminuir o tempo total de execução (*lead time*) dos serviços, a fim de que o mesmo número de funcionários possa confeccionar um maior número de declarações. Em meio a isso, o processo institucional tem sido revisitado e rearranjado para que o caminho ótimo seja encontrado. Algumas ações resultantes dessas iniciativas resultaram em projetos que incluem automação de trabalhos manuais e repetitivos.

Em termos estratégicos, a existência desse setor é justificada pela necessidade de apresentar esse tipo de serviço para clientes que a empresa tem em sua carteira. A empresa está disposta a abrir mão da lucratividade dessa prestação de serviços especificamente, para obter lucro em outros. Também há o interesse, por uma questão de retenção, que os clientes usem o maior número de serviços para que o seu custo de troca desmotive a sua saída. Mas há aqueles que têm um único interesse pelo serviço de desembaraço, dispensando as outras linhas de serviços.

A concorrência emprega a oferta de muitas soluções logísticas, incluindo o desembaraço e pressiona a empresa a se manter párea na entrega de valor. Possui muitos desafios que cruzam múltiplas áreas funcionais, como processos, tecnologia, taxa de retenção de funcionários, temas jurídicos e regulatórios, que dificultam o crescimento em escala. Além de pressões internas, o mercado pressiona o setor para que as tarifas sejam cada vez menores, exercendo influência sobre a constante busca de otimização nos processos para diminuição de custos. Isso torna a busca da excelência operacional um objetivo. Segundo Hausman, Lee, Napier, Thompson e Zheng (2010),

o desembaraço aduaneiro se beneficia muito das análises baseadas em modelos matemáticos, especialmente modelos analíticos que quantifiquem a redução do tempo do ciclo do processo.

O departamento é conhecido por ser altamente improdutivo, operando frente a grandes ineficiências de processo e com resultados financeiros negativos. Nesse cenário repleto de desafios e oportunidades, há o desejo de que floresçam soluções inovadoras para obter sucesso na execução de seus negócios.

Nesta oportunidade se posiciona esta dissertação de mestrado. Para o estudo de caso, foram analisadas diferentes transações de Declaração de Importação da filial de São Paulo, no departamento de Desembaraço Aduaneiro, produzidas durante os anos completos de 2013 a 2020. Esse recorte foi devido à empresa apresentar eventos diferentes antes de 2013 e também por ter um processo com menos passos mapeados em seu manual de operações. O desembaraço aduaneiro vem ganhando maior atenção quanto à eficiência e com o passar dos anos, mais passos têm sido agregados ao processo, para garantir maior visibilidade do tempo em cada uma das etapas. Assim, durante o período analisado, foi possível agrupar um conjunto de dados com *logs* com os mesmos eventos.

### 3 MÉTODO

Este trabalho aplicado foi elaborado por um processo metodológico de abordagem quantitativa, pois utiliza dados na coleta de informações e técnicas estatísticas para o tratamento e análise destes dados. O método utilizado é a aplicação das técnicas analíticas com base em toda a referência teórica da melhoria de processos, BDA, algoritmos estocásticos, custos e portfólio. O estudo investiga a possibilidade de contribuir, suportado por evidências empíricas na avaliação dos dados de uma empresa multinacional no setor de logística e serviços, com a gestão de processos como base para otimização de carteira de clientes, empregando as técnicas e analisando os resultados.

Os métodos aplicados e a comprovação dos resultados podem contribuir para a evolução da gestão de projetos nas empresas e alavancar o potencial do processo como uma nova dimensão para a análise de clientes, em especial empresas do setor de serviços logísticos. Sem a aplicação das técnicas em um estudo de caso, não seria possível consagrar as teorias propostas neste trabalho.

A natureza do trabalho é exploratória, já que investiga evidências para testar as seguintes hipóteses:

- H1: Os dados (logs) do sistema apontam as diferenças dos processos e suas variâncias presentes na prestação de serviço por cliente.*
- H2: O lead time e os desvios dos processos por cliente são dimensões idiossincráticas do custo/benefício do cliente.*
- H3: A governança das estruturas de custos pode ser realizada através de um rebalanceamento da carteira de clientes utilizando variáveis derivadas da análise de processos e receita, através da criação de um portfólio de clientes ótimo.*

Para este estudo, a informação sobre custos não pôde ser obtida por seu caráter confidencial. Por esta razão, toda variável dependente do custo, como requer a própria definição de lucratividade, será adaptada. O valor monetário do custo será substituído pelo *lead time*, partindo da premissa de que as horas necessárias para a execução das etapas do processo representam um custo para a empresa. O tempo do trabalho parado à espera da próxima atividade,

pela impossibilidade de seguir os próximos passos, também será medido como tempo de processamento.

Qualquer que seja o atributo utilizado para medir o custo, o tempo na empresa estudada é o fator mais determinante para justificar o custo do processo e o custo de um cliente. Os funcionários precisam dedicar maior tempo processando as etapas, desde a primeira ação até o faturamento final e, portanto, um cliente com maior uso do tempo operacional, tende a custar mais para a empresa. É importante mencionar que, ainda que neste estudo a variável representativa de custo tenha sido o tempo, o modelo desenhado pode processar a mesma análise de custo utilizando outras representações quantitativas em aplicações futuras, como o valor monetário, o número de funcionários para determinada atividade, semanas de trabalho, ou qualquer outro recurso que faça sentido analisar dentro de um outro contexto. Portanto, nos modelos aplicados no método para análise da H2 e H3, a falta de uma informação monetária para representar o custo foi representada pelas horas.

### **3.1 ESTUDO DE CASO**

A unidade de análise será o departamento de desembaraço aduaneiro da empresa.

Para explorar desafios em outras áreas da economia, a área de serviços foi escolhida por sua crescente importância na economia global e por permitir à pesquisadora a operacionalização dos modelos. O setor decorre de grande necessidade de otimização e eficiência para angariar margens de lucro, ganhar eficiência e produtividade. Devido à própria natureza inerente à produção de serviços, há bastante dificuldade em mapear os custos associados às etapas do processo.

Quanto aos instrumentos de pesquisa, para coleta de dados primários, foi possível a extração de dados da base de dados transacional. Quinze (15) horas foram investidas no refinamento dos dados obtidos do banco para servir de dados primários para os modelos, e cento e sessenta (160) horas na construção do código com os comandos para aplicação dos algoritmos. Os dados secundários foram obtidos em documentos institucionais que descreviam os procedimentos operacionais (SOP, *Standard Operating Procedure*) do departamento para coletar o desenho de processo publicado na empresa. O estudo será dividido em 3 etapas e utilizará os seguintes procedimentos metodológicos:

**Etapa 1:** Os dados serão coletados da base de dados transacional.

O objetivo é utilizar um algoritmo para conduzir uma descoberta do processo de uma área de negócio, para eliminar deduções e premissas e saber como ele se dá na prática, face aos dados produzidos pelo uso do sistema. Encontrar qual é o processo real não é uma tarefa trivial e muitas vezes assume-se qual ele é em trabalhos de reengenharia. As técnicas de reengenharia de processos são formuladas na presunção de que o processo já é conhecido e portanto, é um ponto de crítica deste trabalho. A motivação é o uso de algoritmos que descubram os processos e possam ser aplicados para resolução de problemas reais de negócios e contrapor a ideia de que o processo institucionalizado serve seu propósito de ser um modelo para a execução.

Na proposta de Franken e Weger (1997), o processo é apenas um modelo de abstração de um processo de negócio imaginário e abstrair significa ignorar os detalhes, simplificar. A abstração é um formato apropriado para conceituações e discussões de níveis mais superficiais, não é a verdadeira realização do processo de negócio no mundo real. Enquanto isso é válido em certos contextos, este trabalho identifica a necessidade de que essa abstração não seja usada de forma indevida e, portanto, propõe caracterizar o processo realizado na prática como formato para conversas que tragam profundidade na gestão de processos. Franken e Weger (1997) definem esse último como processo realizável, que é a implementação do processo imaginário.

Entre tantos algoritmos, descobrir o processo realizável caminhará com o uso de modelos estocásticos. Com Cadeias de Markov, o não determinismo será levado em consideração, explorando a correlação local, entre dois eventos, que denotam o passo a passo, naturalmente presente em muitos processos.

**Etapa 2:** Serão elaboradas análises estatísticas envolvendo a descoberta de processos via Cadeias de Markov, para ilustrar a característica probabilística do processo na prática.

Uma vez feita a descoberta do processo por mecanismos analíticos, o objetivo desta fase é estender os resultados para configurar formatos de análise que levem a melhores tomadas de decisão com valores calculados depois de descoberto o processo realizável. Um indicador a ser utilizado nesta fase é o tempo entre cada etapa do processo, em horas. O tempo decorrido para a execução das etapas é um fator de custo para a organização. O tempo é também uma medida de performance da prestação de serviço e da produtividade. Encontrar as variâncias nos tempos de processamento para servir a um cliente tem implicações na medida de produtividade que refletem

no resultado financeiro, já que um departamento improdutivo tende a utilizar mais tempo para a execução de tarefas e tende a auferir resultados inferiores. Além do tempo, as não conformidades dos processos servem para o cálculo de um escore que aumenta conforme o número de desvios encontrados. Sendo assim, partindo da etapa 1 em que o processo realizado na prática é descoberto, o cálculo de tempo entre cada etapa produzida é a matéria-prima utilizada para saber, com maior precisão, o tempo que o processo leva para ser executado e finalizado, sendo esta variável um fator importante para diferenciar o custo, em tempo, do cliente. Saber custos além de ser uma atividade fundamental, retratado por Qian e Ben-Arieh (2008), é também complexo e por isso mesmo bastante investigado. O trabalho ainda considera o que é exposto por Guerreiro, Bio, Vazquez e Merschmann (2008) – os custos são dependentes das atividades prestadas para diferentes clientes. Assim esta etapa é movida pelo desejo de contribuir com a análise do processo resultante do comportamento do cliente e portanto, pelo desejo de construir instrumentos que ajudem na compreensão desse fenômeno.

A análise seguirá com as seguintes fases:

1. Criação da matriz de transição da Cadeia de Markov, que gerará uma matriz quadrada em que o não determinismo é levado em consideração, explorando a correlação local, entre dois eventos;
2. Criação da Cadeia de Markov; e
3. Criação da Cadeia de Markov por cliente.

**Etapa 3:** Com os desenhos de processos descobertos, serão calculadas as variáveis que servirão de insumo para identificar o perfil dos clientes.

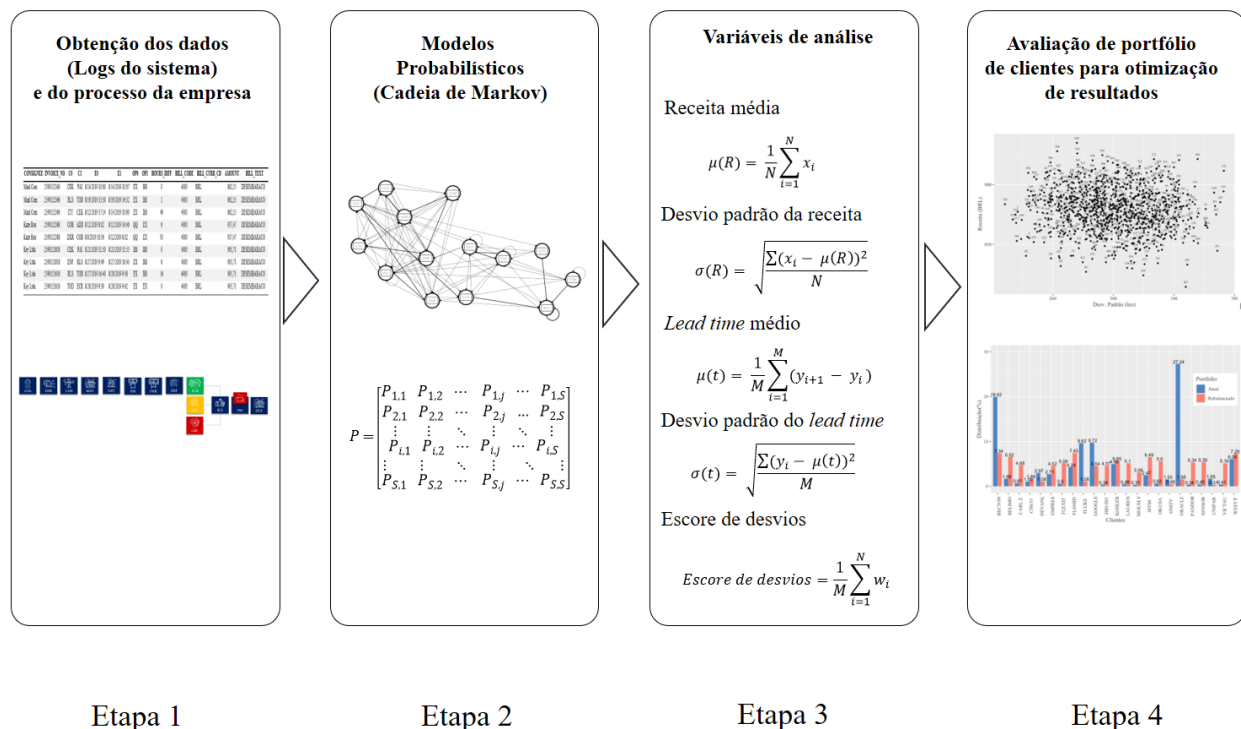
O objetivo desta fase é contribuir, ainda que modestamente, com a criação de instrumentos que possibilitem a construção de um portfólio de clientes fornecido por um modelo que utiliza as variáveis obtidas na etapa 2. São variáveis que indicam problemas no processo, como o *lead time* por etapa do processo, a variância do *lead time* por etapa do processo e o escore do processo. Elas são detratoras da eficiência da execução do negócio e portanto, devem ser minimizadas. A variável receita disponível nos dados originais, é o valor a ser maximizado. E a variância da receita foi utilizada como variável a ser minimizada, independentemente da descoberta do processo, já que o intuito é adicionar estabilidade ao processo e previsibilidade da receita.

A forma aplicada ainda requer um certo debate, pois não é trivial, já que a teoria é originada no campo de estudos de portfólio de investimentos financeiros, instituída por Markowitz (1952). O uso de algoritmos genéticos como forma de buscar otimizações é usado para executar simulações de melhor carteira de clientes de acordo com a função. Tais possibilidades de cruzamento e conexão de informações também abrem caminhos para automação de processos de análise e novas aplicações que são inviáveis se os dados não são usados. Terho (2008) afirma que nas décadas de 1990 e 2000 muitas propostas de modelagens de portfólio de clientes foram apresentadas como ferramentas para a gestão estratégica de relacionamentos. Ele contextualiza que os recursos de uma empresa são limitados e, portanto, não podem ser usados na mesma dosagem para todos os clientes.

Essa parte do estudo, portanto, além dos resultados obtidos pela própria aplicação do instrumento, possibilitará a obtenção de informações sobre diferentes formatos de processamento de análises, cruzamento de informações e paradigmas que a disponibilidade de dados, ciência e tecnologia podem convergir na administração de empresas.

Neste trabalho, escolhemos uma empresa que viabilizou a aplicação desses modelos para um estudo de processos. O modelo teórico apresentado nas três etapas contribuiu para a compreensão do caso e mostrou-se uma boa conjunção de ferramentas para explorar processos na prática.

**Etapa 4:** A teoria moderna de portfólio será a base teórica usada para a construção de um portfólio de clientes otimizado pelas variáveis anteriores, que visa diminuir a variância do *lead time* do processo, aumentar a receita, diminuir a variância da receita, diminuir o *lead time* de execução do processo e diminuir o escore de desvios do processo. Nessa abordagem, os ativos tratados serão os clientes, como analogia aos ativos financeiros, que de forma similar, são compostos por risco, considerado a variância do processo para servi-lo e retorno, considerando-se a receita que eles ofertam. O portfólio será o resultado de um método computacional executado a partir do uso de algoritmos genéticos. O uso de algoritmos genéticos como forma de buscar otimizações será usado para executar simulações de melhor carteira de clientes com base na função minimizadora.



**Figura 8:** Etapas Metodológicas do Estudo

**Fonte:** Elaborado pela Autora.

**Notas:** As 4 etapas do método. Na Etapa 1 os dados do processo institucionalizado são obtidos. Na etapa 2, os dados são analisados por meio do algoritmo de Cadeias de Markov e os processos por cliente são mapeados. Na etapa 3 o *lead time* e outras variáveis dos processos descobertos por clientes são computadas. Na etapa 4 os algoritmos de otimização de portfólio de clientes são calculados.

### 3.1.1 Coleta de dados

Os dados coletados serão obtidos da base transacional da empresa, que é alimentada diariamente pela operação. Através do acesso ao banco e com uma rotina escrita em *Structured Query Language* (SQL), serão obtidos dados que descrevem uma sequência lógica para as transações, análoga ao conceito de *log*, que representam, para cada unidade de serviço prestado, chamada pela empresa de serviço de Desembaraço Aduaneiro, a evolução até que o processo seja concluído, denominados aqui de eventos (apresentados no Apêndice B).

O SQL, apresentado no Apêndice C, servirá para estruturar o resultado com a sequência dos eventos na ordem em que aconteceram na prática. Essa ordem é importante para que o algoritmo que analise os dados possa ser capaz de identificar o evento e seu subsequente. Serão

escolhidos os eventos que estão documentados dentro do processo institucional referente ao negócio analisado e seus significados estão apresentados no Apêndice B.

Os *logs* serão transferidos para um arquivo – o Apêndice D apresenta o nome das variáveis extraídas. Para poder apresentar a estrutura dos *logs* obtidos e facilitar a compreensão do leitor, a autora decidiu antecipar neste capítulo alguns resultados da coleta de dados. Foram retornadas 250.298 linhas no arquivo, das quais, são formados por 25.315 DIs processadas de 01 de janeiro de 2013 até 28 de janeiro de 2020 e o Apêndice E traz um exemplo da estrutura dos *logs*. Os dados poderiam ter sido coletados para os anos inteiros, de 2013 até 2019, porém em favor de utilizar mais dados para gerar análises melhores, a data corte foi 28 de janeiro de 2020, quando este estudo foi iniciado.

Os dados obtidos para a análise deste trabalho serão processados para narrar qual a ordem em que os eventos ocorrem, deixando estruturado qual foi o evento inicial de uma transação e o evento que ocorreu logo em seguida. A proposta é organizar os dados para poder identificar todos os tipos de transições que envolvam um evento e o próximo e assim, servir de base à obtenção das probabilidades de transição de eventos. Com essa estrutura, a intenção é descobrir se há coerência na ordem dos eventos, afetada pela dinâmica dos próprios dados, em comparação à ordem da representação abstrata de processo, comunicado na forma de um diagrama de fluxo de processos. Com isso, ao combinar a informação gerada em uma matriz de eventos e eventos da sequência, será possível refletir todas as transições do processo e a distribuição da frequência em que elas ocorrem.

A partir dos caminhos traçados e partindo da premissa de que o tempo utilizado nesse percurso é um reflexo de horas de trabalho investido pelas pessoas na prestação de serviço, entende-se que essas horas podem ser traduzidas em custos operacionais. Sendo assim, é proposto fazer um mapeamento de custos com base no *lead time* do processo. A seguir, para refletir a influência do processo descoberto especificamente para um cliente e a evidência de não conformidades encontradas, segundo o fluxo organizacional, ajudar a criar um escore que avalia a qualidade do processo e a distância entre ele e o publicado.

Esse nível de reflexão pode contribuir para a identificação da lucratividade a nível de transação e clientes, que durante a execução deste trabalho, não se sabe. A lucratividade por cliente é um fator crítico de análise, porque permite identificar se o relacionamento de negócio entre as partes cria valor. Como foco de atenção para a gestão, parte-se da premissa de que os clientes são

os principais impulsionadores das atividades de custo, e parte deles os critérios mais importantes para explicar os custos e as receitas geradas por uma empresa (Guerreiro et al., 2008; Albalaki e Majeed, 2018; None e Griffin, 1997; Howell e Soucy, 1990; Van Raaij, 2005; Kuchta e Troska, 2007).

Usando técnicas e métodos não tão triviais, é possível formular uma análise que demonstre o valor do cliente. Pode ser chamado de retorno do cliente. Esse retorno pode demonstrar o custo de servir um cliente (*cost to serve*) através de diferentes facetas, incluindo o *lead time* e o escore de desvios calculado do processo. Assim, segundo Shapiro, Rangan, Moriarty e Ross (1987) que sugeriu analisar os clientes sob duas dimensões: receita e o custo para servir, o departamento tendo em mãos algumas variáveis que representem os custos da prestação de serviços ao cliente, pode traçar estratégias para modificar o relacionamento com aqueles que não sejam rentáveis. Dessa forma, sugere problematizar a argumentação atual da empresa de que clientes mais importantes são aqueles com maior volume.

Propondo essa reconsideração, o trabalho visa sugerir abrir mão de padrões tradicionais e apresentar uma alternativa para alcançar melhores resultados operacionais. Através da reconsideração da carteira de clientes e a diversificação entre aqueles que sejam capazes de diminuir as assimetrias de processos e que gerem desequilíbrios, inspirado por Shapiro, Rangan, Moriarty e Ross (1987), de clientes com o perfil mais *Passive*, que custam menos à empresa e estão dispostos a aceitar preços altos.

Em última instância, ter todo o negócio sustentado por um grupo pequeno de clientes, ainda que sejam rentáveis, põe em risco a sustentação dos negócios. Para mitigar esse risco e alinhado à ideia da teoria de portfólio moderna, a diversificação é um fator crítico e, analogamente, podemos aplicar seu conceito, na construção de uma carteira de clientes diversificada, que possa diminuir a dependência do negócio de um grupo pequeno.

Durante este estudo de caso, o risco de um processo caro influenciado pelos clientes e o retorno que eles oferecem não são tratados dentro da organização. O uso dessa teoria também confronta a noção de fortes relacionamentos sob o princípio estrito de pura retenção, que hoje é aplicado na empresa. Para ajudar a empresa a criar um novo sistema de relacionamento com o cliente de acordo com o peso processual que ele gera e concentrar seus esforços em gerenciar uma carteira de clientes mais vantajosa, uma técnica analítica capaz de encontrar um estado ótimo em meios a algoritmos de otimização, pode ser aplicada.

### 3.1.2 Os softwares

O método para manipular os dados será feito com o software R em sua versão 4.0.2 (R Development Core Team, 2020) em máquina Apple® MacBook Air® com macOS® High Sierra versão 10.13 de 2 GHz Intel® Core® i7 e 8 GBytes de RAM, por possuir bibliotecas para análise dos principais algoritmos necessários para conduzir as análises. Para a interface de execução, será utilizado um ambiente de desenvolvimento integrado (IDE, *Integrated Development Environment*), o RStudio em sua versão 1.0.143 (RStudio Team, 2020), que utiliza a linguagem de programação R para interpretar os comandos de programação e os resultados. O R é uma linguagem de computação e também uma ferramenta para a análise de dados. Foi escolhida a linguagem R com a sua execução na IDE RStudio, dada a flexibilidade que vem do uso de uma linguagem de programação, pois o R proporciona a liberdade para a mescla de diferentes algoritmos e experimentações, devido a grande quantidade de pacotes que encapsulam a lógica matemática necessária, garantindo reuso e a produtividade da pesquisa. Os pacotes disponibilizados pelo R são gratuitos e de acesso público e as informações sobre sua utilização assim como principais pacotes podem ser encontradas no site da CRAN (*Comprehensive R Archive Network*). Os pacotes utilizados serão MARKOVCHAIN 0.8.5-2 (Spedicato, 2017), READXL 1.3.1 (Wickham e Bryan, 2019), GGPLOT2 3.3.2 (Wickham, 2016), GGTHEMES 4.2.0 (Arnold, 2019), DPLYR 1.0.2 (Wickham et al., 2020), IGRAPH 1.2.6 (Csardi e Nepusz, 2006) e GA 3.2 (Scrucca, 2013).

A agilidade também vem do fato de que os resultados são prontamente disponibilizados na sua interface gráfica, sem necessidade de investimento de tempo em instalações de ambientes de execução complexos.

Para o desenho da Cadeia de Markov, será utilizado o software Gephi em sua versão 0.9.2 (Bastian, Heymann e Jacomy, 2009), em máquina Apple® MacBook Air® com macOS® High Sierra versão 10.13 de 2 GHz Intel® Core® i7 e 8 GBytes de RAM, por permitir adicionar camadas gráficas e mais intuitivas ao grafo obtido no R. O Gephi permite manipular as cores dos elementos, o tamanho dos vértices e dos nós com maior usabilidade. A visualização da cadeia se mostrará mais compreensível no Gephi do que no R.

Também será utilizado o Microsoft Excel em sua versão 2016 em máquina Intel® Core® i5 de 2,60 GHz com Windows 10 Enterprise™ e 16 Gbytes de RAM, para melhor compreensão dos dados, executar filtros e prepará-los para a leitura no R.

### **3.1.3 O desenho do processo**

Com os dados do sistema, será possível obter a sequência de eventos ocorridos a partir de uma matriz de transição para todas as DIs do período analisado. O departamento possui um sistema de gestão que registra os caminhos em que cada DI transita até ser completada. Desde o recebimento dos documentos que permitem que o processo se inicie, até a liberação na alfândega, para que a carga de fato entre no Brasil.

Com isso será representada uma matriz de transição, apresentada no Apêndice F, que é o resultado do processamento dos dados disponibilizados nos *logs* do sistema. Cada linha da matriz representa o evento de origem e cada coluna o evento que ocorre na sequência. Essa matriz estocástica é uma matriz quadrada que descreve as transições em uma Cadeia de Markov, dos eventos. Os valores da matriz são expressos em porcentagem, em que cada linha totaliza 100%, ou o número real 1. Ela revela as chances de transição do evento original para todos os outros disponíveis.

Baseado nas propriedades de cadeias de Markov, a soma das probabilidades de transição de um estado para todos os outros é 1 para cada uma das linhas.

### **3.1.4 O desenho do processo por cliente**

Segundo a literatura, o comportamento do cliente e outras características deste tem alta influência no custo da prestação de serviço. Dessa perspectiva, o estudo buscará a análise de processos por cliente para identificar o modelo de processos descoberto para servir cada cliente especificamente. Aqueles que possuem um tempo de processamento maior tendem a elevar os custos da prestação do serviço e em virtude dessa maior servitude. O estado ótimo de um cliente custoso seria obter maior receita para justificar a permanência deste na carteira de clientes.

O princípio dessa abordagem é o reconhecimento de que diferentes clientes possuem diferentes famílias de produtos importados, diferentes segmentos e canais de mercado e, com isso, necessidades distintas. Por exemplo, o número de pedidos de compras importadas ou o número de

itens em cada um desses pedidos aumentam o esforço do trabalho. O tipo de mercadoria importada e a necessidade ou não de licenças adicionam mais tarefas. Os valores mais altos de certas mercadorias que oferecem maior risco se algum erro é transmitido à alfândega, exigem maiores necessidades de conferências. A frequência de comunicação necessária exigida na prestação do serviço e o formato, seja com e-mails, ou através de manutenção de arquivos de *follow-up* requerem tempo para criá-los. O fornecimento ou não de informações relevantes à declaração da importação, por parte do cliente, influenciam a quantidade de esforço para obtê-las, geralmente envolvendo maior quantidade de troca de e-mails até que o cliente tenha esclarecido todas as dúvidas. E o cliente, dependendo de seu poder de barganha, pode delegar tarefas extras para a empresa, gerando atividades que não agregam valor à prestação de serviço em si, mas apenas geram à empresa maior desperdício de tempo em atividades operacionais e por essa demanda extra, não pagam honorários que reflitam essa sobrecarga de trabalho.

Assim, dentre os processos descobertos pelos dados, o estudo busca observar se os clientes apresentam desenhos de processos similares, ainda que exista evidências sobre a natureza da flexibilidade que esse setor demonstra para atender necessidades específicas de seus clientes.

### 3.1.5 As variáveis de análise dos clientes

Ainda que o volume de transações fosse um indicador de produtividade do departamento, o indicador de receita também contribuía para analisar as finanças da empresa. A receita, no entanto, como métrica, apenas narrava as entradas na empresa e sozinha não refletia o valor do resultado real, do lucro, pois necessitava das saídas, que a empresa não tinha mecanismos para calcular. Como o custo conforme visto é uma variável de difícil obtenção dentro do setor analisado conforme Braithwaite e Samakh (1998), neste estudo serão utilizados o *lead time* do processo e o desvio padrão do *lead time* como a representação dos custos.

No entanto, vale aqui explicar todas as variáveis que serão consideradas durante a descoberta do processo por cliente: Receita média, Desvio padrão da receita, *Lead time* médio, Desvio padrão do *lead time* e Escore do processo.

**Receita média:** a média das receitas de todas as declarações de importação do cliente, conforme a equação 7:

$$\mu(R) = \frac{1}{N} \sum_{i=1}^N x_i \quad (7)$$

onde:

$\mu(R)$  é a receita média do cliente;

$N$  é o número de declarações de importação para o cliente; e

$x_i$  é o valor da receita para cada declaração de importação  $i$  ( $i = 1, \dots, N$ ).

**Desvio padrão da receita:** o desvio padrão das receitas de todas as declarações de importação do cliente, conforme a equação 8.

$$\sigma(R) = \sqrt{\frac{\sum (x_i - \mu(R))^2}{N}} \quad (8)$$

onde:

$\sigma(R)$  é o desvio padrão das receitas do cliente;

$\mu(R)$  é a receita média do cliente;

$N$  é o número de declarações de importação para o cliente; e

$x_i$  é o valor da receita para cada declaração de importação  $i$  ( $i = 1, \dots, N$ ).

**Lead time médio:** a média do *lead time* de todas as declarações de importação do cliente, conforme a equação 9.

$$\mu(t) = \frac{1}{M} \sum_{i=1}^M (y_{i+1} - y_i) \quad (9)$$

onde:

$\mu(t)$  é o *lead time* médio do cliente;

$M$  é o número de transições encontradas para o cliente;

$y_i$  é a data do evento  $i$  ( $i = 1, \dots, M$ ); e

$y_{i+1}$  é a data do evento  $i + 1$  ( $i = 1, \dots, M$ ).

**Desvio padrão do *lead time*:** o desvio padrão do *lead time* de todas as declarações de importação do cliente, conforme a equação 10.

$$\sigma(t) = \sqrt{\frac{\sum (y_i - \mu(t))^2}{M}} \quad (10)$$

onde:

$\sigma(t)$  é o desvio padrão do *lead time* do cliente;

$\mu(T)$  é o *lead time* médio do cliente;

$M$  é o número de transições encontradas para o cliente; e

$y_i$  é a data de um evento  $i$  ( $i = 1, \dots, M$ ).

**Escore de desvios processo:** a quantidade de sequências não correspondentes ao processo institucionalizado encontradas no processo descoberto do cliente, dividida pela quantidade total de sequências encontradas. Assim como mostra a equação 10.

$$Escore\ de\ desvios = \frac{1}{M} \sum_{i=1}^N w_i \quad (11)$$

onde:

$M$  é o número de transições encontradas para o cliente;

$N$  é o número de transições não esperadas no processo institucionalizado;

$w_i$  é uma transição não esperada no processo institucionalizado  $i$  ( $i = 1, \dots, M$ ).

Essas variáveis serão tratadas como medidas de estudo de otimização de resultados. Durante a execução deste estudo, o risco, sob a ótica de processos e o retorno, juntos, não têm sido tratados dentro da organização. Inclusive é parte da proposta deste estudo ofertar a governança de resultados ótimos para a empresa, partindo de uma visão que combine essas variáveis. Além disso, baseado na teoria de portfólio moderna (TPM), portfólios podem ofertar retornos com menor volatilidade. Nesta teoria, a diversificação também é um fator crítico, pois ajuda na redução de

riscos. Ao manter mais ativos de investimento na carteira, é possível balancear os retornos entre ativos diversos e diminuir a dependência do negócio de um único ativo, que no contexto deste trabalho, é o cliente. O objetivo dessa análise será aplicar a formulação da teoria para a composição da carteira de clientes. Na empresa analisada, o conceito de diversificação é usualmente gerenciado por intuição.

Além de evitar problemas relacionados com o investimento direcionado em apenas um produto, a aplicação da técnica permite maximizar resultados, recombinação a distribuição de esforço entre os clientes, uma vez que é capaz de reduzir custos, encorajar a empresa na busca de maximização de resultados junto aos clientes, ou revendo o processo, ou aumentando as tarifas cobradas, ou ainda, revendo a carteira de clientes e alocando os recursos da empresa de forma mais eficiente, atendendo de forma diferenciada aqueles que são vistos como mais rentáveis.

O gestor responsável por dar melhor direcionamento aos recursos pode ter que enfrentar alguns dilemas, como, qual conta ele precisa manter a qualquer custo? Qual conta poderia sair da lista sem afetar tanto seus resultados? Dentre as contas desejadas, qual a quantidade de trabalho que o departamento certamente poderia absorver com os mesmos recursos? Segundo o desenho atual do processo que o departamento executa, quais os perfis de clientes mais se enquadram em sua linha de produção? A partir daí, o gestor pode querer reajustar parte do processo a fim de se preparar melhor para certos tipos de clientes. Ou se perceber que esse caminho seja menos desejável, talvez focar mais nas qualidades atuais e trabalhar com maior foco em criar uma campanha de vendas que convença clientes a aumentarem o volume de transações confiados ao seu departamento.

### **3.1.6 As simulações de portfólio de clientes**

As considerações tradicionais na empresa estudada eram feitas com base na receita obtida para cada um dos clientes. De acordo com o conceito de diversificação, na possibilidade de maiores retornos com menor volatilidade e alinhado à ideia de que se possa fazer uma análise que vá além da receita e com o uso de maior quantidade de variáveis, serão feitas simulações para poder obter uma visualização de cenários de diferentes portfólios de clientes.

Simular é uma prática que pode aliviar o peso do uso de tarefas manuais e ineficientes para a tomada de decisões gerenciais. Seria difícil fazer esse tipo de estudo de combinação de processos por cliente de forma manual para testar diferentes cenários.

Os cenários serão calculados com o uso da aleatoriedade. Eles não serão criados com uma estrutura que otimize o cenário global, como é natural da aleatoriedade. Porém será um modelo que introduzirá as oportunidades que a empresa possui, ao ajudar a visualizar que duas ou mais carteiras podem ofertar uma média de variação de lead time similar e média de receita, ou retorno, diferentes. Nesta etapa, a função criada, utilizará as variáveis de análise dos clientes. Juntas, o resultado da função será formado por construtos que possam diminuir o desvio padrão do *lead time* do processo, aumentar a receita média do departamento, diminuir o desvio padrão da receita, diminuir o *lead time* do processo e diminuir o escore de desvios do processo.

A função será formatada da seguinte maneira:

$$\sum_{i=1}^N w_i \sigma(t)_i + w_i \frac{1}{R_i} + w_i \sigma(R)_i + w_i \mu(t)_i + w_i \alpha_i \quad (12)$$

onde:

$N$  é o número de clientes ativos;

$w_i$  é a proporção ( $0 \leq w_i \leq 1$ ) do portfólio de DIs para cada cliente  $i$  ( $i = 1, \dots, N$ );

$\sigma(t)_i$  é o desvio padrão do *lead time* do processo para cada cliente  $i$  ( $i = 1, \dots, N$ );

$R_i$  é a receita média para cada cliente  $i$  ( $i = 1, \dots, N$ );

$\sigma(R)_i$  é o desvio padrão da receita para cada cliente  $i$  ( $i = 1, \dots, N$ );

$\mu(t)_i$  é a média do *lead time* do processo para cada cliente  $i$  ( $i = 1, \dots, N$ ); e

$\alpha_i$  é o escore de desvios do processo para cada cliente  $i$  ( $i = 1, \dots, N$ ).

### 3.1.7 As simulações por modelos de otimização

Neste artigo, a técnica de algoritmos genéticos (Goldberg, 1989) será aplicada para resolver a otimização do portfólio de clientes. A escolha do algoritmo genético foi devida à familiaridade da autora com o conceito.

No algoritmo de aleatoriedade, os diversos portfólios gerados serão resultados de simulações aleatórias dentro do espaço total de possíveis cenários. Para otimizar o resultado da carteira, o método parte da perspectiva de que os dados auxiliam a função matemática de um algoritmo de otimização que pudesse encontrar a melhor combinação de clientes na carteira para melhorar os resultados do negócio.

Os algoritmos genéticos possuem capacidade computacional para testar uma variedade de soluções através de um processo de evolução que funciona com base em uma metáfora de sobrevivência da natureza, daqueles que têm maior capacidade de adaptação, de mutações, cruzamentos genéticos e sobrevivam a seleção natural. Nesta etapa, com o uso do algoritmo genético será possível otimizar os resultados soluções, adicionando na sua função objetivo, novas camadas de variáveis que pudessem diminuir o desvio padrão do *lead time* do processo, aumentar a receita média do departamento, diminuir o desvio padrão da receita, diminuir o *lead time* do processo e diminuir o escore de desvios do processo.

A fórmula precisa ser bem definida para refletir a função objetivo com suas devidas restrições. É preciso rigor para desenhar a fórmula para que tenha os parâmetros suficientes e gerem resultados coerentes e que possam gerar discussões mais conectadas com a realidade do negócio. Portanto, para colaborar na formatação matemática, a base lógica que definirá o que a função precisará descobrir, se originou da teoria de portfólio. O objetivo de um portfólio ótimo é ter um conjunto de ativos que maximize valores positivos para o negócio e que diminua os negativos.

Durante a organização dos dados para este método, foram separados os dados de clientes de 2019 apenas. Se fossem usadas as informações de anos anteriores, os resultados não refletiriam o conjunto correto de clientes atuais, pois devido à rotatividade de clientes, os dados estariam obsoletos e considerando clientes que não mais estão na empresa.

A função foi formatada da seguinte maneira:

$$\text{Minimize } \sum_{i=1}^N w_i \sigma(t)_i + w_i \frac{1}{R_i} + w_i \sigma(R)_i + w_i \mu(t)_i + w_i \alpha_i \quad (13)$$

Outras equações a serem consideradas:

$$\sum_{i=1}^N w_i = 1 \quad (14)$$

$$\sum_{i=1}^N z_i = N \quad (15)$$

$$\varepsilon_i w_i \leq w_i \leq \delta_i w_i \quad i = 1, \dots, N \quad (16)$$

onde:

$N$  é o número de clientes ativos;

$w_i$  é a proporção ( $0 \leq w_i \leq 1$ ) do portfólio de DIs para cada cliente  $i$  ( $i = 1, \dots, N$ );

$\varepsilon_i$  é a proporção mínima de DIs para cada cliente  $i$  ( $i = 1, \dots, N$ );

$\delta_i$  é a proporção máxima de DIs para cada cliente  $i$  ( $i = 1, \dots, N$ );

$z_i = 1$  para cada cliente  $i$  ( $i = 1, \dots, N$ );

$\sigma(t)_i$  é o desvio padrão do tempo de execução do processo para cada cliente  $i$  ( $i = 1, \dots, N$ );

$R_i$  é a receita média para cada cliente  $i$  ( $i = 1, \dots, N$ );

$\sigma(R)_i$  é o desvio padrão da receita para cada cliente  $i$  ( $i = 1, \dots, N$ );

$\mu(t)_i$  é a média do tempo de execução do processo para cada cliente  $i$  ( $i = 1, \dots, N$ );

$\alpha_i$  é o escore de desvios do processo para cada cliente  $i$  ( $i = 1, \dots, N$ );

A equação 13 expressou que o desvio padrão do *lead time* do processo, o desvio padrão da receita, o *lead time* do processo e o escore de desvios do processo precisavam ser minimizados, enquanto a a receita média maximizada. A equação 15 garantiu que todos os clientes no conjunto de dados, precisariam estar presentes. A equação 16 garantiu uma proporção mínima de declarações para cada cliente e garantiu uma proporção máxima de declarações para cada cliente.

As definições de limites foram feitas para permitir mudanças realizáveis. Os problemas de otimização possuem limitações para manter os resultados retornados dentro dos limites de possibilidades de implementação. Neste caso, o algoritmo, ao tentar otimizar, poderia levar a concentração de todos os processos do negócio para um cliente apenas, possivelmente o que possui o menor valor para uma variável que a função objetivo busque minimizar ou o contrário. As

limitações tornam um problema mais difícil de solucionar, porque além de buscar a melhor combinação de valores para maximizar ou minimizar um valor, é preciso respeitá-las.

Para estabelecer um limite à distribuição de volume de negócio para os diferentes clientes, foi criada uma função que desse uma penalidade para o algoritmo, toda vez que ele tentasse mudar a distribuição de declarações por cliente além de uma mudança muito agressiva de valor, conforme a equação 17. O limite estabelecido será de uma mudança de 5% para mais ou menos ao número atual. Esse valor foi escolhido pela autora, não tendo qualquer relação com objetivos da empresa. É um valor arbitrário. Ao estabelecer esse limite de mudança, a premissa é de propor mudanças, porém realistas dentro de uma margem de crescimento possível. Outros números que façam sentido dentro do negócio aplicado podem ser escolhidos.

Ao processar uma mudança maior do que esses limites, o valor retornado da função aumentaria, tornando-o assim, uma penalidade de ordem positiva, que faria o algoritmo se distanciar do objetivo de minimização.

$$\Delta = w_j - w_i \quad (17)$$

onde:

$$\Delta < 5$$

$w_i$  é a proporção ( $0 \leq w_i \leq 1$ ) do portfólio de DIs para cada cliente  $i$  ( $i = 1, \dots, N$ );

$w_j$  é a nova proporção ( $0 \leq w_i \leq 1$ ) do portfólio de DIs para cada cliente  $i$  ( $i = 1, \dots, N$ );

O código reproduzível completo do trabalho está disponível no Apêndice K.

## 4 ANÁLISE DOS RESULTADOS

Através do método e resultados apresentados no Capítulo 3, foi possível descobrir o processo, com representações visuais, a partir dos dados disponíveis (van der Aalst, 2011).

Primeiramente, foram identificadas as transições dos eventos do processo, que revelaram a diferença da prática comparada ao processo formalizado pela instituição. Com base na rede obtida pelo algoritmo de Cadeia de Markov, o que se pôde observar é que o processo em um ambiente ambíguo e complexo, pode ser melhor representado por meio de probabilidades, incluindo o formato não determinístico como ingrediente chave do modelo (Colombo et al., 2006). A princípio, as probabilidades foram geradas por meio da tabela de transição (Apêndice F), de cada caminho percorrido, que resultaram em 113 diferentes transições de eventos do processo. No entanto, algumas das transições encontradas não espelharam nenhum dos caminhos esperados do processo, dessa forma sendo tratadas como não conformidades.

A Tabela 1 apresenta as 3 transições com maior frequência. A transição mais frequente foi aquela entre o ECR e o RLS, com 8,33% em relação a todo o conjunto de transições. Essa transição é o momento em que a Receita Federal classifica a nacionalização da mercadoria com o canal verde e na sequência acontece a liberação do desembaraço. A segunda transição mais frequente, com 8,24%, foi entre o ENF e o ECR, que na ordem desenhada pelo processo, é a transição anterior à que foi descrita. Representa a transmissão dos dados para a Receita e na sequência a classificação do canal verde. Em terceiro lugar, a transição mais frequente foi a do ADD e do GAT, com 8,10% de frequência, que representa a chegada física da mercadoria nos portos e aeroportos, seguida da carga liberada pelo terminal.

**Tabela 2:** Top-3 sequência de eventos mais frequentes

Sequência	Média	Mediana	Desvio Padrão	%
ECR RLS	13,74 h	0,00 h	48,77 h	8,33
ENF ECR	16,63 h	16,00 h	26,63 h	8,24
ADD GAT	22,15 h	6,00 h	50,86 h	8,10

**Fonte:** Elaborado pela Autora.

**Notas:** As sequências com maior frequência são as ECR – RLS, que traduz a transição de uma declaração de importação de *canal verde* para *declaração desembaraçada*, com 8,33% de representatividade do total de transições. Em segundo lugar está a transição ENF – ECR que indica a transição de *registro da declaração*

para o *canal verde*, com 8,24% de frequência. Em terceiro está a transição entre ADD – GAT que representa a *chegada física da carga* e a *carga liberada pelo terminal*.

Do *lead time* de todas as transições, apresentado de forma completa no Apêndice G, foram obtidos o tempo médio por transição, a mediana e o desvio padrão a nível departamental. Para o departamento, para cada transição de eventos, a média é de 79,54 horas, a mediana de 18,00 horas e o desvio padrão de 868,86 horas. Os valores foram calculados para toda a população. Para cada transição foi criado o desenho da curva da distribuição do tempo. O Apêndice H apresenta quatro (4) exemplos de curvas para diferentes transições.

As probabilidades foram calculadas com base nas transições de eventos encontradas a partir do evento original.

Ex.: Probabilidades de eventos após a chegada (ADD) descritas na Tabela 3:

**Tabela 3:** Probabilidades de eventos após a chegada (ADD). Atual versus Esperada

	Eventos em $X_1$ quando o ADD é o evento $X_0$													
	ADD	COB	DLV	DXR	ECR	EFD	ENF	EXR	GAT	LOG	PAI	PAR	RLS	TNF
Atual	0,00	0,01	0,00	0,04	0,00	0,00	0,03	0,00	0,75	0,02	0,09	0,05	0,00	0,00
Esperado	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00

**Fonte:** Elaborado pela Autora.

**Notas:** As probabilidades encontradas nos dados foram: 0,01 para o embarque da carga (COB); 0,04 para o recebimento dos documentos (DXR); 0,03 para o registro da Declaração (ENF); 0,75 para a carga liberada no terminal (GAT); 0,02 para a referência criada no Sistema (LOG); 0,09 para os valores financeiros pré-calculados (PAI); 0,05 para o recebimento dos valores financeiros (PAR). O processo institucionalizado declara que após o ADD, o evento deveria ser para 100% dos casos, o GAT.

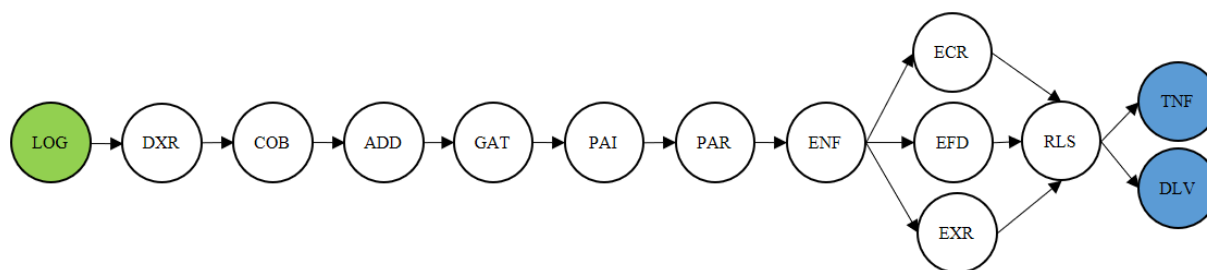
A partir dessa análise, foi possível identificar algumas transições em não conformidade, como o embarque da carga (COB), que de forma alguma pode acontecer depois da chegada. O registro da declaração (ENF), que ocorre apenas após a chegada e uma série de outros passos intermediários, também não estava alinhada ao fluxo de processos. O caminho mais provável após a chegada, de fato era a carga liberada no terminal (GAT). Porém não obtivemos 100% de transição entre o ADD e o GAT, que resultaram em 75% de chances.

Em resumo, a partir da distribuição de probabilidades, foi possível notar que houve uma variedade de combinações de eventos, com múltiplos estados possíveis (Jónás et al., 2014), tornando o desenho linearmente representado em forma de diagrama de fluxo da empresa, uma

representação não realista do processo e que, portanto, essas dissimilaridades entre os processos colocaram à prova o processo publicado internamente.

A matriz de transição (Cayley, 1858) foi o insumo para gerar o grafo, que traduz todos os possíveis estados do processo e a ordem como aconteceram em sequências de dados categóricos (Ching et al., 2008), baseado em como os eventos foram de fato inseridos no sistema (Lakshmanan et al., 2012; Chiu et al., 2019; Ibanez-Sanchez et al., 2019; van der Aalst, 2011; Weijters et al. et al., 2006; Goedertier et al. et al., 2009 ). O desenho do processo resultante das transições de probabilidades foi representado em uma forma de grafo que é um modelo gráfico que explica de forma mais amigável o processo (Ibanez-Sanchez et al., 2019), apresentado na Figura 8, ilustrando a variedade de combinações transitórias.

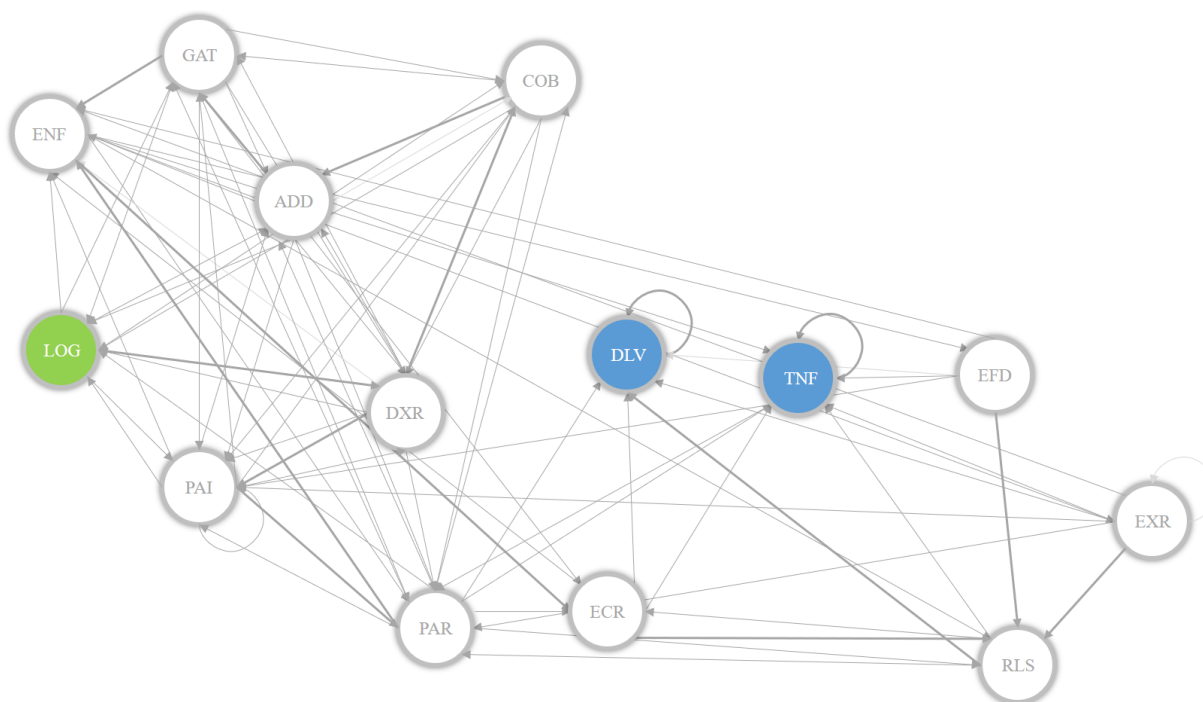
É importante lembrar que o presente trabalho se propõe a demonstrar a análise de processos pela perspectiva dos dados, isto é, dos *logs* do sistema (Chen e Storey, 2012). É possível notar desde o princípio, pela característica estrutural do grafo, que a ordem não é definitivamente linear conforme proposta pelo desenho de processo oficial da empresa, como mostra a Figura 9. Na prática, as transações passam por eventos em uma ordem probabilística. O grafo apresenta característica concorrentes (van der Aalst et al. et al., 2004) e enlaçamentos (Wen et al. et al., 2007).



**Figura 9:** Grafo que representa o processo estruturalmente institucionalizado

**Fonte:** Elaborado pela Autora.

**Notas:** Cada vértice do grafo é um evento do processo da empresa. Cada vértice representa um evento, codificado em três (3) letras para representar o fenômeno. O significado dos códigos está disponível no Apêndice B. A vértice de cor verde é o início (LOG) e as vértices em azul os finais (TNF ou DLV) do processo.



**Figura 10:** Grafo das interconexões do processo na prática, a partir dos dados

**Fonte:** Elaborado pela Autora, utilizando as ferramentas R 4.0.2 e Gephi 0.9.2.

**Notas:** A vértice de cor verde é o início (LOG) e as vértices em azul os finais (TNF ou DLV) do processo.

O grafo obtido na Figura 10, pela modelagem apresenta uma das características de Cadeias de Markov: a quantidade de estados é finita. Nele, temos a representação gráfica de catorze (14) estados. Cada transição de estado é dependente do estado anterior. O fenômeno estudado é um processo Markoviano, dada a modelagem das probabilidades de transição entre estados discretos.

Verificando se os processos são executados conforme o esperado (van der Aalst, 2011), não houve coerência na ordem dos eventos em comparação à ordem da representação abstrata de processo, oferecendo uma capacidade de auditoria (Chiu et al., 2019). Segundo Lee et al., (2008), a representação abstrata do processo em formato de diagrama de fluxo se reduz a um modelo de referência e portanto, não configura uma restrição, permitindo liberdade para adesão. Assim o processo na prática refletido pela dinâmica dos próprios dados, fornecem aos auditores informações ricas e não distorcidas.

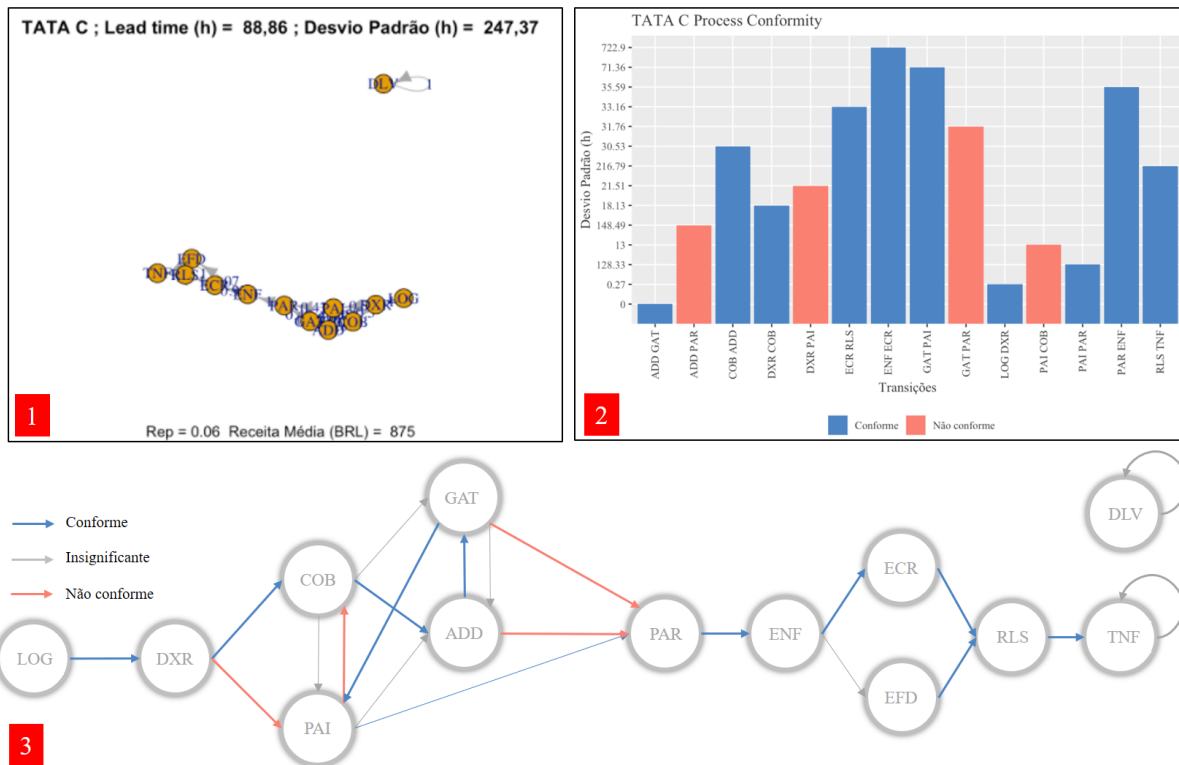
Além disso, esse contexto não favorece o controle e a gestão do processo (Janssen et al., 2015). Vale ressaltar, que eliminar essas variações pode ser uma consideração gerencial, pois

segundo Gregor, Martin, Fernandez, Stern e Vitale (2006), o BDA à nível transacional fornece benefícios operacionais, como redução dos custos operacionais, aumento da produtividade dos funcionários ou economia no gerenciamento da cadeia de suprimentos.

Na sequência, movido pelo desejo de contribuir com a análise do processo resultante do comportamento do cliente e portanto, pelo desejo de construir instrumentos que ajudem na compreensão desse fenômeno, derivamos a descoberta do processo por cliente. O setor de serviços, da empresa estudada, apresenta características básicas do setor, em que o grau de participação do cliente influencia o processo (Jorge et al., 2011).

Para cada um deles, foi obtido uma peculiaridade no processo em forma de grafo.

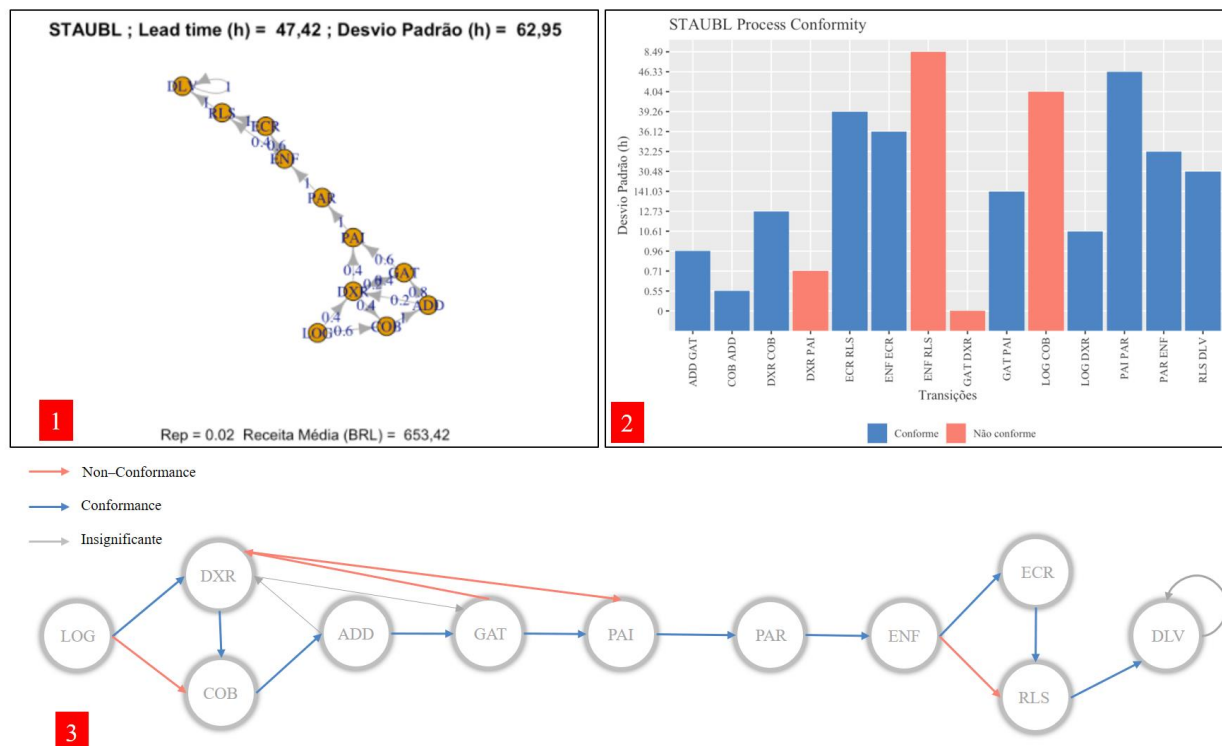
O cliente 1 possuiu um processo representado apresentado na Figura 11. Algumas sequências não seguem o processo institucionalizado. Uma delas é o ADD – PAR. Esta sequência descreve que logo após a chegada, os valores para pagamento dos impostos foram recebidos. No entanto, o cliente só poderia ter pago se houvesse sido informado dos valores a serem pagos (PAI). Outra não conformidade é o DXR – PAI. Esta sequência descreve que os valores foram calculados logo após o recebimento dos documentos. No entanto, os valores só podem ser calculados após a carga ter sido liberada pelo terminal (GAT). Essas não conformidades, adicionam *lead time* ao processo e causam desvios, representados no gráfico 2 na Figura 11.



**Figura 11:** Processo descoberto para o cliente 1

**Fonte:** Elaborado pela Autora, utilizando as ferramentas R 4.0.2 e Gephi 0.9.2.

**Notas:** Cliente 1: Volume de negócios: 0.06%; Lead time: 88,86 horas; Receita média: R\$ 875,00; Desvio padrão: 247,37 horas. Esta figura tem 3 partes. O grafo em 1, gerado pela ferramenta R e logo ao lado o gráfico 2 com as sequências conformes e não-conformes e o desvio padrão de cada uma. O grafo 3 foi a combinação do grafo encontrado com as cores das sequências conformes e não-conformes.



**Figura 12:** Processo descoberto para o cliente 2

**Fonte:** Elaborado pela Autora, utilizando as ferramentas R 4.0.2 e Gephi 0.9.2.

**Notas:** Cliente 2: Volume de negócios: 0.02%; Lead time: 47,42 horas; Receita média: R\$ 653,42; Desvio padrão: 62,95 horas. Esta figura tem 3 partes. O grafo em 1, gerado pela ferramenta R e logo ao lado do gráfico 2 com as sequências conformes e não-conformes e do desvio padrão de cada uma. O grafo 3 foi a combinação do grafo encontrado com as cores das sequências conformes e não-conformes.

Para o cliente 2, algumas sequências não seguem o processo institucionalizado, como mostra a Figura 12. Assim como o cliente 1, também está não conforme com DXR – PAI. Uma outra sequência incorreta é o GAT – DXR. Esta sequência descreve que logo após a carga liberada pelo terminal, os documentos são recebidos. O ideal é que o cliente envie os documentos o quanto antes para que o processo de digitação se inicie. Tardar o envio de documentos apenas piora a performance do departamento. Uma outra sequência não conforme é LOG – COB. Justamente por não enviar os documentos, o embarque acontece sem documentos, que seriam retratados pela sequência LOG – DXR – COB. Essas não conformidades, adicionam *lead time* ao processo e causam desvios, representados no gráfico na Figura 12.

Essas variações retrataram erros no processo que deveriam ser ajustados para fins de controle de qualidade. Essas não conformidades foram evidenciadas por meio do cálculo de um

escore do processo. Novamente, emergiu-se as assimetrias e a falta de padronização na prestação do serviço. Levitt (1972) considerou a “abordagem da linha de produção” como aquela pela qual os serviços poderiam também melhorar significativamente seu desempenho em custos. Infelizmente, no caso estudado, a abordagem do autor não teve aderência.

Assim, dentre os processos descobertos pelos dados, o estudo mostrou que os clientes apresentaram um desenho em formato de grafo diferentes, apresentados no Apêndice I, fornecendo maiores evidências sobre a natureza da flexibilidade que esse setor demonstra para atender necessidades específicas de seus clientes. Enquanto essa análise evidenciou a natureza da flexibilidade no atendimento, também apresentou a distância dos processos executados daquele processo único institucionalizado a todo o departamento.

Portanto, essa análise conferiu evidências das dissimilaridades entre os processos dos clientes que colocaram à prova a existência de um processo único institucionalizado. Esse *gap* entre a expectativa e realidade foi acentuado quando analisado a partir do histórico do serviço prestado a um cliente específico, resultando em uma diversidade de desenho de processos, juntamente com variabilidade nos indicadores, como média, desvio padrão, receita e escore do processo.

Dessa forma, no contexto deste estudo de caso, a H1 é verdadeira.

*H1: Os dados (logs) do sistema apontam as diferenças dos processos e suas variâncias presentes na prestação de serviço por cliente.*

Com os dados separados por cliente, o grafo com o comportamento para cada um corroborou o aprendizado sobre o desenho do processo. Essas diferenças e particularidades que cada cliente apresentou com relação a ordem dos eventos nos seus processos foi insumo para executar um cálculo que pudesse expressar o indicador de qualidade do processo em relação ao processo oficial, o escore do processo. Certas transições de eventos que apresentaram não conformidade foram obtidas para que a proporção de desvios fosse calculada relativamente ao conjunto total de transições efetuadas no histórico do processo do cliente. Com os dados derivados da descoberta dos diferentes processos, tais como o desvio padrão do *lead time* e o *lead time* do

processo, e o escore de desvios do processo, foi possível estender análises que foram além dos métodos tradicionais de avaliação do valor do cliente.

Segundo Shapiro, Rangan, Moriarty e Ross (1987), que sugeriu analisar os clientes sob duas dimensões: receita e o custo para servir, foram selecionadas a combinação de variáveis relacionadas aos pontos negativos dos clientes, como o desvio padrão do *lead time* do processo e variáveis positivas, como a receita. Foi um método não trivial, motivado pela dinâmica dos processos e seus desvios, como base para avaliação do custo. O uso do desvio padrão como métrica foi fortemente inspirado pela TPM de Markowitz (1952).

Conforme foram cruzadas essas duas dimensões, foi possível observar um posicionamento dos clientes comparados entre si. A partir dessa modelagem gráfica e segundo o modelo de Shapiro, Rangan, Moriarty e Ross (1987) foi possível encontrar os clientes *Carriage trade* (*Classe Alta*), *Bargain basement* (*Barganheiro*), *Passive* (*Passivo*) e *Aggressive* (*Agressivo*), conforme Gráfico 3.



Nesta comparação foi apresentado um retrato do posicionamento do cliente em relação às variáveis de desvio padrão do tempo de execução do processo e receita ofertada. A maior parte dos clientes estão na parte inferior do gráfico, indicando baixas receitas, enquanto que no eixo horizontal, estão dispersas, indicando bastante variabilidade do desvio padrão do processo. Isso demonstrou como a empresa tem variações da execução do trabalho enquanto que a receita se mantém menor. Informações sobre os custos da prestação de serviços ao cliente podem modificar o relacionamento de uma empresa com clientes que não são rentáveis (Kaplan e Narayanan, 2001).

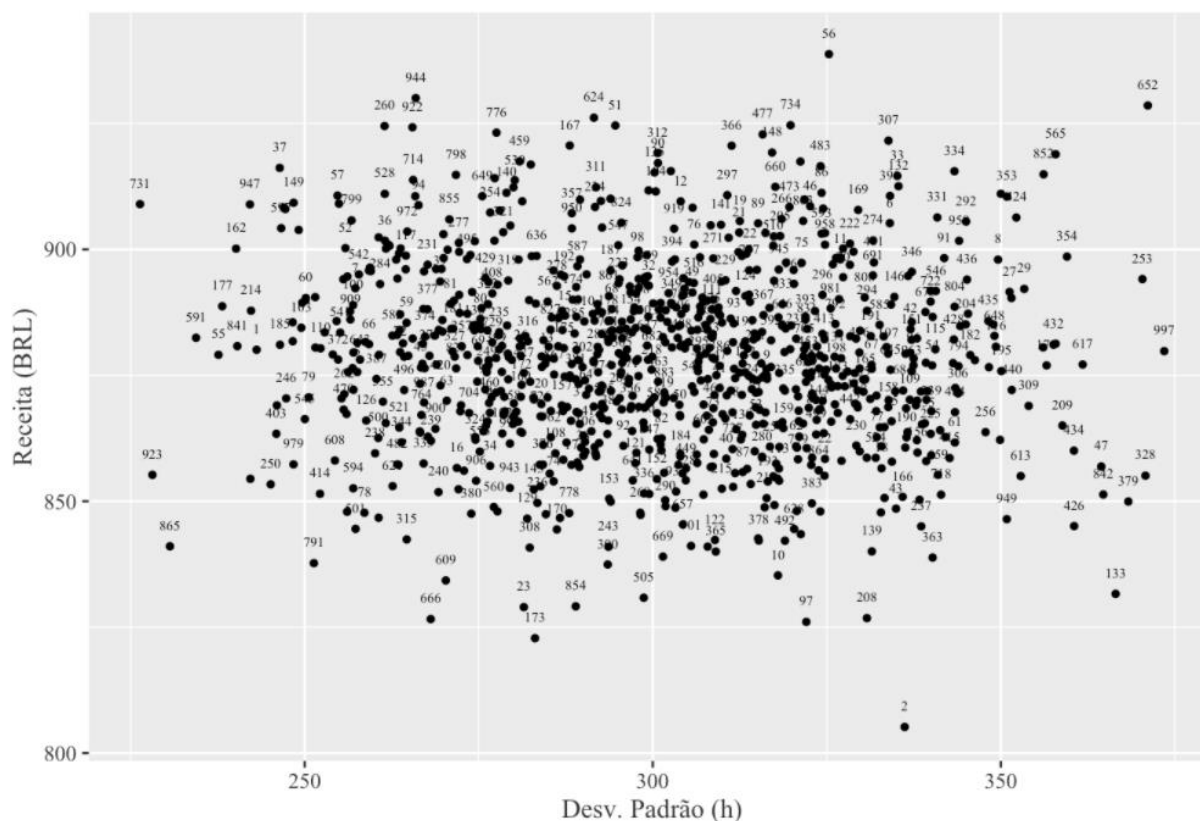
O posicionamento ideal de um cliente é que este maximize os resultados de receita, com o menor desvio padrão do tempo de processamento. Essa questão é relevante já que a lucratividade de uma empresa é dependente do custo gerado na prestação de serviços aos clientes (Guerreiro et al., 2008). Esse cruzamento permitiu visualizar como os vários clientes estavam em relação aos outros. Dentre as possíveis opções que explicassem um maior desvio padrão, estavam as diferentes tarefas personalizadas para atender o serviço do cliente e tarefas de exceção que não geravam retorno para a empresa. Para atender a um cliente, o desvio pode ser analisado como um fator de risco no controle dos limites de qualidade do processo.

Todos os clientes agrupados, assim como em diferentes áreas, puderam ser análogos ao conjunto de ativos que faz parte do gerenciamento de um portfólio de investimentos, em que são combinados os riscos e os retornos para garantir maximização de ganhos maiores com menores riscos (Dickinson, Thornton e Graves, 2001; Kumar, Ajjan e Niu, 2008). Os clientes puderam ser gerenciados como um portfólio, combinando o risco atrelado ao processo necessário para atendê-lo e o retorno em formato de receita. Para maximizar os benefícios e auxiliar na seleção de clientes específicos ou perfis de clientes específicos, uma combinação de diferentes volumes de negócio por clientes pôde trazer análise de cenários que trouxessem resultados mais atrativos para o negócio.

Com o uso de simulações de variáveis a fim de encontrar a melhor redistribuição de volume de trabalho entre os diferentes clientes e então calcular os novos valores encontrados para o desvio padrão do *lead time* e a receita média do departamento, as simulações, uma ferramenta que possibilita análise de cenários, permitiu a comparação de diferentes carteiras.

Ao utilizar o desvio padrão do *lead time* e a receitas geradas, recombinação das quantidades de declarações processadas pelos clientes, as simulações geraram uma grande quantidade de cenários de portfólio representados pelos pontos no gráfico. No Gráfico 4, para cada cada ponto,

os cálculos da média da receita dessa carteira no eixo y e do total de variação do processo no serviço prestado ao cliente no eixo x foram comparados. A cada ponto foi atribuído um número de identificação (ID, *Identification*) da simulação.



**Gráfico 4:** Portfólios com métodos aleatórios: Desvio Padrão versus Receita

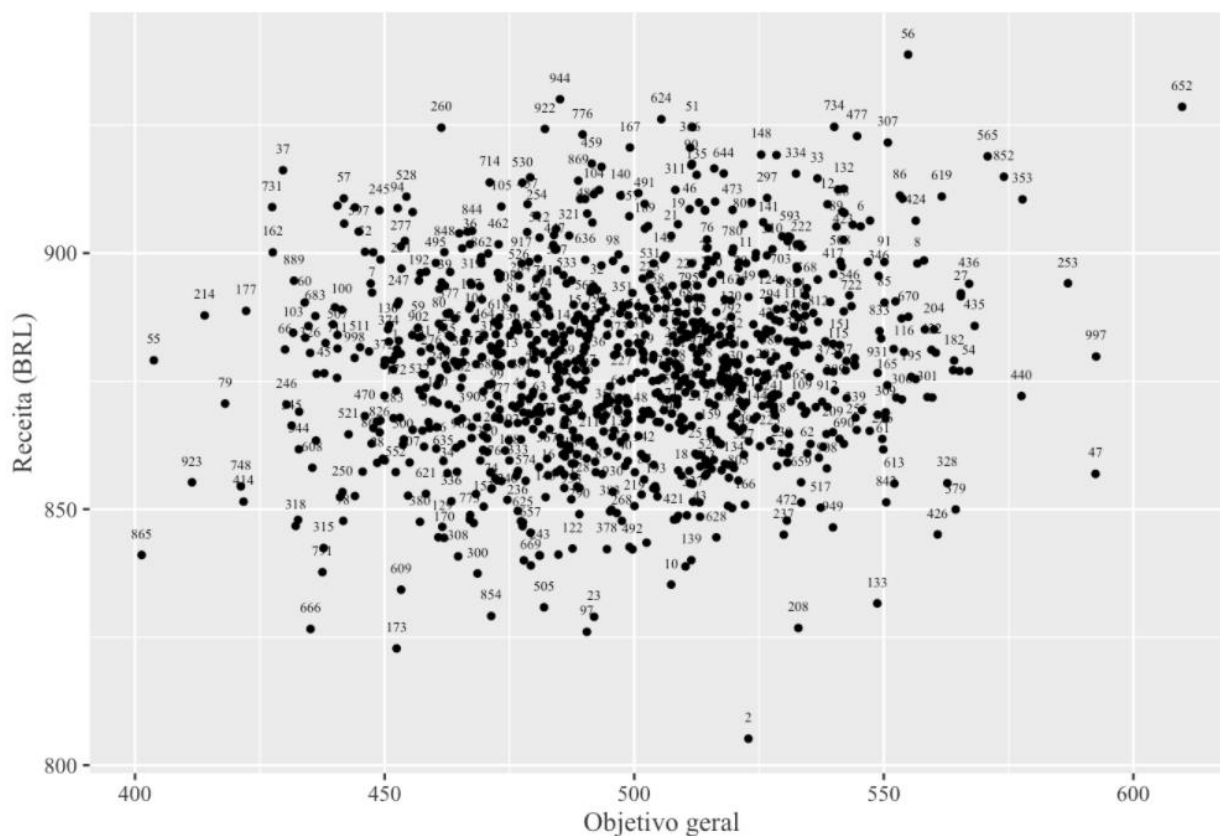
**Fonte:** Elaborado pela Autora, utilizando a ferramenta R 4.0.2.

**Notas:** Cada ponto no gráfico representa um portfólio de clientes com diferentes volumes de declarações de importação distribuídos entre eles. O posicionamento do portfólio indica a receita média que o departamento obterá no mês contra o desvio padrão médio para dada carteira

As simulações identificadas nessa parte do estudo ajudaram a visualizar que duas ou mais carteiras podem ofertar uma média de variação de *lead time* similar e média de receita, ou retorno, diferentes.

No entanto, esses resultados foram calculados com o uso da aleatoriedade para cada cenário, que não permitiu definir qual a melhor carteira. Porém foi um modelo que introduziu as oportunidades que a empresa possuía.

Nas simulações, uma outra análise feita, foi a de cruzar o valor obtido da função geral, da equação 12, que combina construtos que são interessantes para que o negócio tenha um melhor desempenho, como mostra o Gráfico 5.



**Gráfico 5:** Portfólios com métodos aleatórios: Objetivo geral versus Receita

**Fonte:** Elaborado pela Autora, utilizando a ferramenta R 4.0.2.

**Notas:** Cada ponto no gráfico representa um portfólio de clientes com diferentes volumes de declarações de importação distribuídos entre eles. O posicionamento do portfólio indica a receita média que o departamento obterá no mês contra a combinação de construtos que são interessantes para que o negócio, para dada carteira

Dessa forma, as simulações identificadas nesse passo ajudaram a visualizar que uma mesma carteira pode ofertar um resultado de função, que otimiza vários construtos do departamento, diferente.

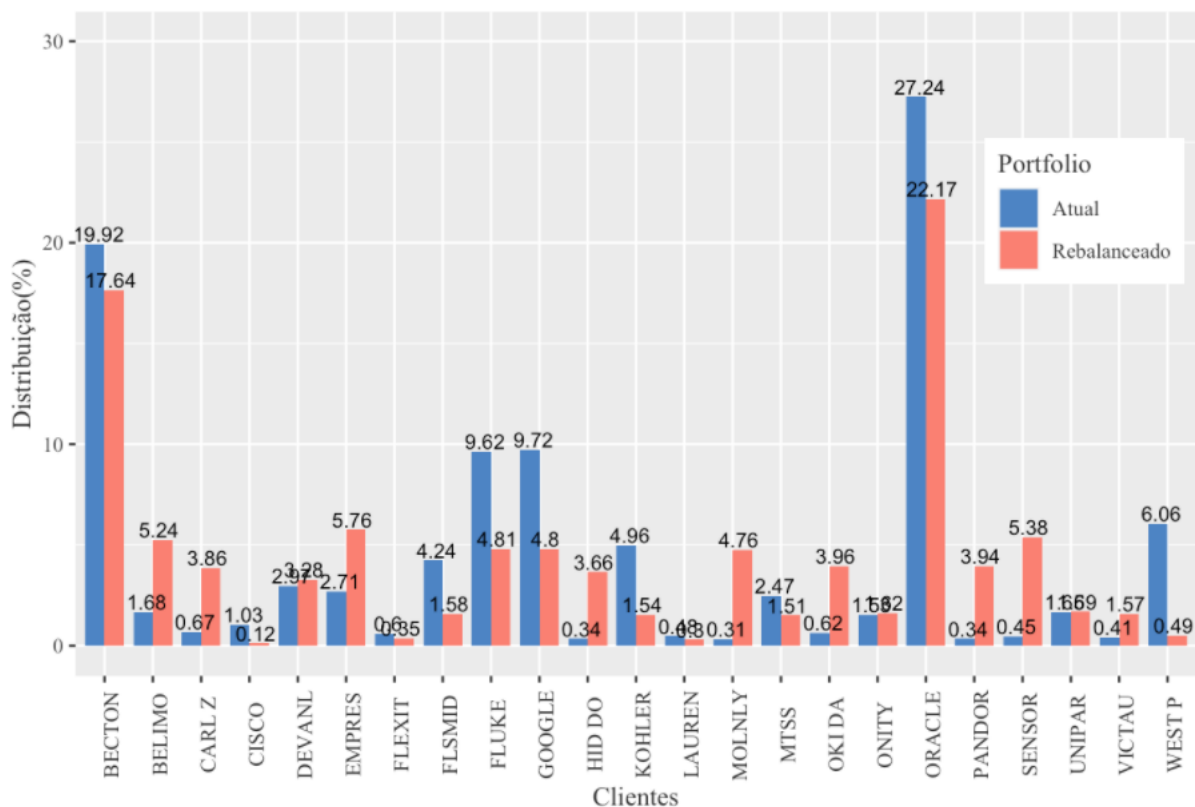
O sistema de simulações permitiu que um conjunto de combinações de distribuição do volume de trabalho, entre clientes com processos diferenciados em caminhos e tempos, fosse examinado de forma holística. Dessa forma, no contexto deste estudo de caso, a H2 é verdadeira.

*H2: O lead time e os desvios dos processos por cliente são dimensões idiossincráticas do custo/benefício do cliente.*

Visando melhorar a busca pelo melhor portfólio e otimizar o algoritmo, já que as simulações não possuem heurísticas que ofertem a melhor carteira, a técnica de algoritmos genéticos foi aplicada para resolver a otimização da carteira devido ao seu desempenho relativamente melhor na busca da solução ótima do portfólio de clientes, sustentado pela ótica do custo para servir (Shapiro et al., 1987; Krapfel et al., 1991). Assim, visando diminuir o desvio padrão do *lead time* do processo, aumentar a receita média do departamento, diminuir o desvio padrão da receita, diminuir o *lead time* do processo e diminuir o escore de desvios do processo, o algoritmo genético teve maior performance quando comparado a simulação de cenários usando um método aleatório.

A análise incluiu a geração de 10.000 simulações e o tempo de processamento do algoritmo foi de 14,73 segundos.

Foi gerado pelo algoritmo um novo rebalanceamento do portfólio com os clientes da carteira, apresentado no Apêndice J. O Gráfico 6 mostra a carteira otimizada e a atual para entender em quais proporções o rebalanceamento foi feito.

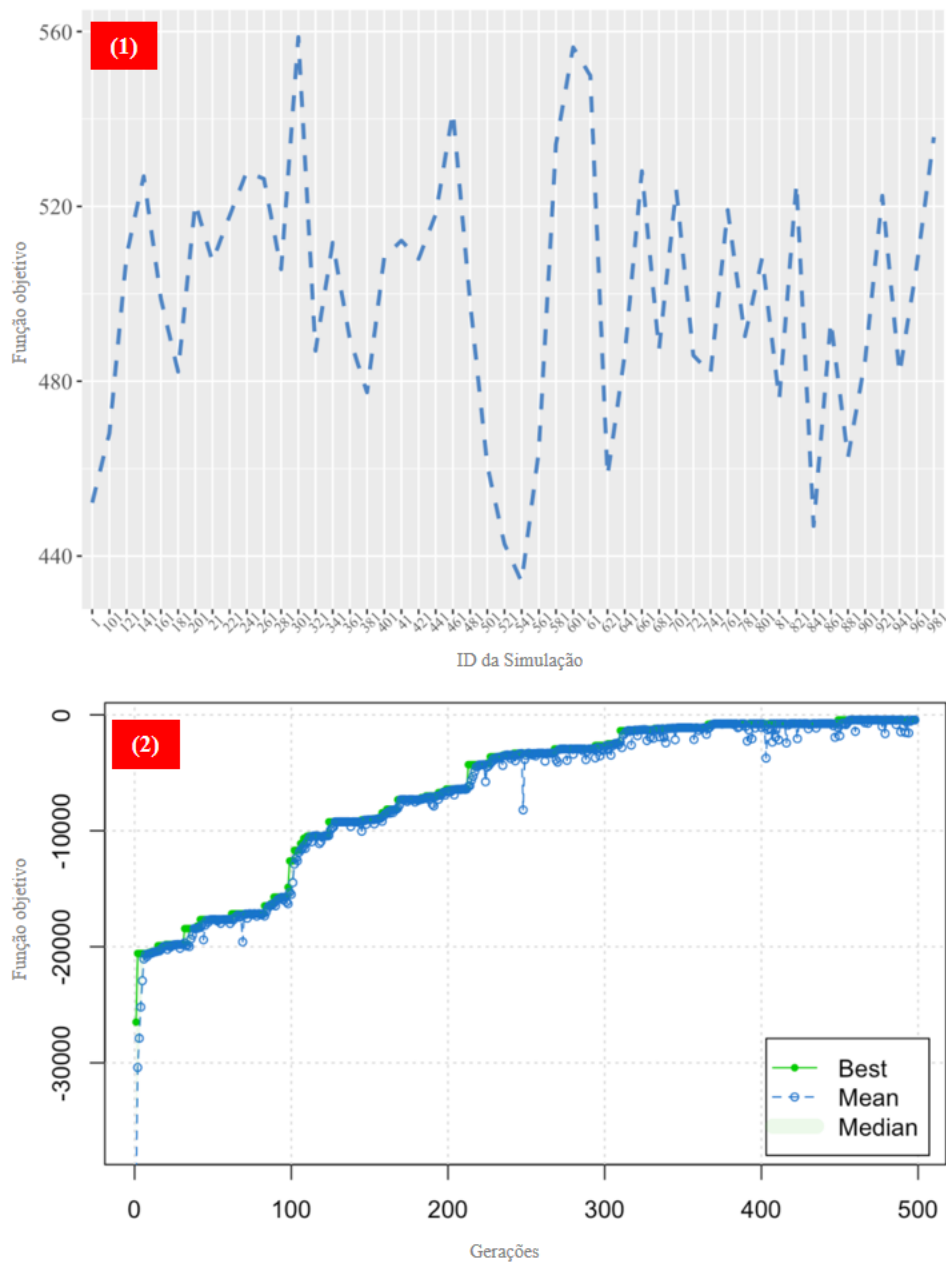


**Gráfico 6:** Rebalanceamento da carteira de clientes

**Fonte:** Elaborado pela Autora, utilizando a ferramenta R 4.0.2.

**Notas:** O resultado da distribuição de declarações de importação por cliente, no comparativo de como é atualmente e de como o algoritmo sugere que esteja para otimizar o resultado do departamento.

Como o algoritmo genético buscou minimizar a função objetivo, ele possuiu uma tendência a obter resultados melhores do que aqueles obtidos previamente durante seu processamento. Além da obtenção do valor mínimo ótimo global, a performance ao longo do tempo foi mais eficaz quando comparada com o algoritmo aleatório, como mostra o Gráfico 7.



**Gráfico 7:** Desempenho por método aleatório e por Algoritmo Genético

**Fonte:** Elaborado pela Autora, utilizando a ferramenta R 4.0.2.

**Notas:** Desempenha na minimização do problema. No gráfico (1), os resultados da função objetivo do eixo Y são aleatórios e a curva obtida não apresenta tendência ao longo das simulações executadas no eixo X. A curva não tem um padrão. No gráfico (2), resultado do algoritmo genético, a fitness function ou função objetivo tem uma tendência a ir diminuindo ao longo das diferentes gerações no eixo X. O AG por essência busca o melhor resultado possível dentre todas as simulações, portanto é uma otimização do algoritmo do gráfico (1).

Portanto, levou-se em conta uma redistribuição da quantidade de processos entre os clientes para uma estratégia de rebalanceamento em termos de retornos ajustados ao risco de manter clientes com processos ineficientes na carteira. Assim sendo, foi aplicada uma estratégia de rebalanceamento ideal que buscou minimizar variações indesejadas ao mesmo tempo que permitisse o aumento da receita. Esse resultado confronta a visão do desenvolvimento de fortes relacionamentos sob o estrito princípio de retenção do cliente. Assim, sugere a empresa ajustar seu sistema de relacionamento com o cliente de acordo com o seu valor e concentrar seus esforços em gerenciar com diferenciação essa nova lista, que tem um perfil diferenciado (Johnson e Selnes, 2005).

**Tabela 4:** Resultados dos valores obtidos através da otimização do portfólio de clientes

Métricas	Antes	Depois	Diferença
Desvio padrão do <i>lead time</i> do processo	344,26 h	297,54 h	-13,57 % ↓
<i>Lead time</i> do processo	90,30 h	89,76 h	-00,60 % ↓
Receita média	R\$ 702,29	R\$ 729,75	+03,91 % ↑
Desvio padrão da Receita	R\$ 38,67	R\$ 25,03	-35,27 % ↓
Escore de desvios	0,29	0,31	+06,90 % ↑

**Fonte:** Elaborado pela Autora.

O valor médio do desvio padrão no novo portfólio foi de 297,54 horas, a receita média foi de R\$ 729,75, o desvio padrão da receita foi de R\$ 25,03, o *lead time* do processo de 89,76 h e o escore de desvios de 0.31.

Esses foram melhores resultados comparados aos valores médios atuais que são de 344,26 horas para o desvio padrão do processo, R\$ 702,29 em média para a receita, R\$ 38,67 para o desvio padrão da receita, 90,30 horas para o *lead time* do processo, porém o escore de desvios atualmente é melhor com 0,29. Por fim, no contexto deste estudo de caso, a H3 é verdadeira.

*H3: A governança das estruturas de custos pode ser realizada através de um rebalanceamento da carteira de clientes utilizando variáveis derivadas da análise de processos e receita, através da criação de um portfólio de clientes ótimo.*

O Quadro 1 resume as hipóteses com seus resultados.

**Quadro 1:** Resumo da verificação das hipóteses do estudo

Hipóteses	Resultado	Observações
<p><i>H1: Os dados (logs) do sistema apontam as diferenças dos processos e suas variâncias presentes na prestação de serviço por cliente.</i></p>	<p><b>Conforme esperado</b></p>	<p>Por Cadeia de Markov, descobrir o processo a partir dos <i>logs</i>, Os processos são executados sem coerência com a representação institucionalizada do processo.</p> <hr/> <p>O grau de participação do cliente influencia o processo. A verdadeira realização do processo por cliente não foi coerente com a abstração da organização.</p>
<p><i>H2: O lead time e os desvios dos processos por cliente são dimensões idiossincráticas do custo/benefício do cliente.</i></p>	<p><b>Conforme esperado</b></p>	<p>Com dados sob a ótica do processo, foi possível modelar graficamente os perfis dos clientes.</p> <hr/> <p>Saber os custos da prestação de serviços ao cliente pode modificar o relacionamento de uma empresa com clientes que não são rentáveis.</p> <hr/> <p>Maximizar os resultados com criações de carteira de clientes. Simulações ajudaram a visualizar que duas carteiras podem ofertar uma média de variação de <i>lead time</i> similar (custos) enquanto uma delas traz maior retorno.</p>
<p><i>H3: A governança das estruturas de custos pode ser realizada através de um rebalanceamento da carteira de clientes utilizando variáveis derivadas da análise de processos e receita, através da criação de um portfólio de clientes ótimo.</i></p>	<p><b>Conforme esperado</b></p>	<p>A técnica de algoritmos genéticos foi aplicada para resolver a otimização da carteira devido ao seu desempenho relativamente melhor na busca da solução ótima do portfólio de clientes.</p> <hr/> <p>O algoritmo genético teve maior performance quando comparado a simulação de cenários usando um método aleatório.</p>

**Fonte:** Elaborado pela Autora.



## 5 CONCLUSÕES

Esta pesquisa trouxe uma análise de múltiplas aplicações de tecnologias analíticas que visaram facilitar a gestão de processos e geração de *insights* nas organizações. Para contribuir com a descoberta de conhecimento analítico, para o desenvolvimento das organizações em tomarem decisões mais assertivas e encontrarem novos caminhos para enfrentar os constantes desafios, não apenas explorando a teoria, mas gerando experimentos na prática. Como estudo de caso, foram usados os dados do sistema de gestão de uma empresa do setor logístico.

O trabalho concluiu que os modelos obtidos no estudo puderam contribuir como uma opção de técnica a ser empregada na análise de processos, usando dados e contribuindo para a disciplina de BPM que cada vez ganha maior espaço entre as organizações. Estendeu-se a partir da descoberta do processo a descoberta a nível de cliente, em um cenário em que este tem grande influência nos processos de uma empresa. Na análise do desempenho do cliente, foi encontrado uma grande variação no caminho percorrido pelo processo, a fim de executar a prestação de serviço. Com os dados sobre performance do cliente, foram utilizados modelos de simulação e otimização para a criação de um portfólio de clientes otimizado a ofertar a possibilidade de diminuir o desvio padrão do *lead time* do processo, aumentar a receita média do departamento, diminuir o desvio padrão da receita, diminuir o *lead time* do processo e diminuir o escore de desvios do processo, buscando eficiência. Apresentou novos usos para algoritmos já conhecidos.

Os modelos propostos forneceram à organização novos caminhos mais conectados com o fenômeno do *Big Data*. As ferramentas apresentadas puderam ser usadas para conduzir análises integradas e um conjunto de soluções baseadas em dados que mostrou outros caminhos na obtenção de resultados através de tomadas de decisões direcionadas por modelos analíticos. As soluções tiveram um potencial de auxiliar no desempenho da empresa através do ponto de partida de análise de processos de forma mais eclética, desafiando os modelos tradicionais. Além disso, apresentaram novas formas de evoluir partindo dos modelos de processos, passando por análises mais profundas orientadas aos perfis dos clientes e chegando nas simulações de portfólios, orientadas pela teoria.

Este trabalho ofereceu maior conhecimento que pode levar a melhor tomada de decisão para desenvolver novas soluções para problemas, novas oportunidades de negócios e diminuição de riscos (Baesens et al., 2016; Ketter et al., 2016; Martin, 2015; Müller et al., 2016).

Os resultados apontaram que a empresa teve a sua disposição um conjunto de ferramentas que mostraram como criar valor e considerável ganho de vantagem competitiva, na busca de eficiência operacional e otimização de recursos através do uso do *Analytics* em sua forma mais ousada, com a combinação de diferentes modelos, formando todo um *framework* de técnicas. Esses recursos são raros dentro do setor estudado e de difícil imitação. Isso levaria provavelmente a uma vantagem competitiva sustentável.

Os modelos apresentados fomentaram também o princípio de frugalidade, já que fez uso e incentivou o software livre, desmistificando a ideia de que para obter *insights* através dos dados é preciso grandes investimentos em plataformas analíticas. Outra característica frugal foi a reutilização dos dados já presentes dentro do banco de dados, em um nível de baixa granularidade, que, mesmo sendo de natureza transacional e operacional, foram reutilizados para objetivos mais estratégicos e proporcionaram a definição de um modelo de processo de alto nível. Uma outra grande vantagem desta abordagem é que o estudo ocorreu enquanto os atores humanos trabalhavam sem serem perturbados, praticamente inconsciente do mapeamento de processo feito pelos *logs* de seus registros, sem que fosse necessário despender tempo com consultorias, entrevistas com as equipes de negócio ou a aquisição de um número de funcionários da organização para o estudo de processos.

Assim, com a aplicação desses modelos, a empresa poderia dedicar seus recursos em trabalhar de forma mais estratégica mediante os resultados de otimização obtidos. É importante dar, no entanto, aos gestores, a oportunidade de compreender os diferentes mecanismos para que se sintam à vontade em tomar decisões outrora formatadas por intuição ou experiência profissional prévia.

Primeiramente, a técnica de processos estocásticos, em específico Cadeias de Markov, foi apresentada como possibilidade de análise de descoberta de processo. Ferramentas de modelagem são imprescindíveis para ajudar na análise de processos. Ambientes que apresentam complexidade e chances de que certas etapas do negócio aconteçam merecem atenção mais probabilística para a sua modelagem. Cadeias de Markov serviram para esse propósito.

Em continuidade, foram computadas variáveis de processos por cliente, a fim de capturar a complexidade para a prestação de serviços. A premissa foi de que o tempo decorrido para a execução das etapas e os desvios são um fator de custo para a organização. Foi preciso estudar o tempo necessário para a execução de cada fase do processo associada a prestação de serviço de

cada cliente. Encontrar as variâncias nos tempos de processamento para servir a um cliente tem implicações na medida de produtividade que refletem no resultado financeiro, já que um departamento improdutivo tende a utilizar mais tempo para a execução de tarefas e tende a auferir resultados inferiores.

Por último, foram usados algoritmos para o problema de rebalanceamento do portfólio de clientes, como forma de otimizar os cálculos. Primeiramente foi apresentado o algoritmo de simulações com métodos aleatórios, seguido de algoritmo genético, trazendo velocidade na obtenção das informações. Os algoritmos propostos estavam integrados no ambiente da organização, obtiveram dados do banco de dados transacional e operacional do negócio e usaram essas informações em análises de cenários.

Com os novos valores obtidos a partir das otimizações, pode-se inferir quais teriam sido os valores do negócio quando aplicados na escala de um ano de trabalho. O departamento poderia diminuir o desvio padrão em 158.147,20 horas, o *lead time* em 1.827,90 horas, o desvio padrão da receita em R\$ 46.171,40 e aumentar a receita em R\$ 92.852,10, conforme Tabela 5. Se esse método for aplicado em todas as outras filiais do Brasil, esse ganho de receita poderia chegar em até R\$ 464.260,50 para a empresa em todo o país.

**Tabela 5:** Resultados dos valores obtidos na escala de 1 ano

Métricas	Antes	Depois	Diferença
Desvio padrão do <i>lead time</i> do processo	1.165.320,10 h	1.007.172,90 h	-158.147,20 h
<i>Lead time</i> do processo	305.665,50 h	303.837,60 h	-1.827,90 h
Receita média	R\$ 2.377.251,65	R\$ 2.470.203,75	R\$ 92.852,10
Desvio padrão da Receita	R\$ 130.897,95	R\$ 84.726,55	- R\$ 46.171,40

**Fonte:** Elaborado pela Autora.

**Notas:** Valores considerados utilizando os volumes de 2018 que totalizaram 3.385 declarações de importação.

Uma das limitações deste estudo é que o acesso aos dados não foi em tempo real, o que ainda não torna a solução um produto integrado para a gestão e requer a intermediação de analistas que possam manipular os dados e os códigos. Devido à natureza dinâmica das empresas, elas exigem recursos de análise de dados em tempo real, o que permite o processamento de dados também em tempo real para monitorar processos e eventos (Manyika et al., 2011). A abordagem deste estudo na descoberta do processo não utilizou os passos característicos de modelagem, em

que os dados são executados em um conjunto de dados de treino e depois no conjunto de teste, portanto, está sujeita a *overfitting* e pode gerar resultados inferenciais inconsistentes. Foi usado um estudo de caso para avaliar a eficácia dos modelos, em que algumas restrições do próprio ambiente específico desse negócio foram satisfeitas, não utilizando outros segmentos para um estudo comparativo. Além disso, apenas um tipo de algoritmo heurístico foi utilizado como método, o algoritmo genético. Há uma variedade de algoritmos que podem ser usados para comparar melhores desempenhos. Cabe, futuramente, a possibilidade de saber qual técnica deve ser escolhida devido ao seu desempenho na solução de um problema.

Para estudos futuros, podem ser analisados outros setores a fim de validar se esses modelos não precisam de adaptações e calibrações. A questão organizacional também poderia ser abordada, para avaliar se a aplicação desses métodos requer uma mudança de cultura da empresa e as possíveis fricções entre os atores envolvidos. Alinhar as pessoas, a tecnologia e os recursos organizacionais para se tornar uma empresa orientada a dados é problemático e desafiador (Weill e Ross, 2009)

Arunachalam et al. (2018) ressaltaram que para lidar efetivamente com a difusão de tecnologias BDA em empresas, questões organizacionais e comportamentais relacionadas à adoção e prática devem ser abordadas. Apesar de sua popularidade no setor, muitas organizações relutam em investir em tecnologia BDA devido à ambigüidade em reconhecer benefícios potenciais. De acordo com Kiron e Shockley (2011), três características principais constituem a cultura orientada a dados: a análise deve ser tratada como um ativo valioso ou estratégico, a alta administração deve apoiar e emponderar a análise em toda a organização e a acessibilidade de *insights* baseados em dados para tomadores de decisão. Lavalle, Lesser, Shockley, Hopkins e Kruschwitz (2017) realizaram uma pesquisa para identificar as barreiras significativas para a adoção de análises nas organizações. Apesar dos desafios da tecnologia e da qualidade dos dados, as organizações consideraram a cultura e as questões administrativas como a barreira significativa para a adoção de análises. Além disso, Cao, Duan e Li (2015) argumentaram que a presença da cultura baseada em dados facilitaria as organizações a tomar decisões baseadas em dados e confiar em decisões baseadas em fatos para desenvolver novos produtos e serviços. Certamente, organizações que possuem recursos analíticos avançados não poderiam extrair valor total se não fossem efetivamente integradas ao processo de tomada de decisão de negócios e não fossem aceitas como uma ferramenta de tomada de decisão (Blackburn et al., 2015). Schoenherr e Speier-Pero

(2015) identificaram barreiras potenciais como funcionários inexperientes, restrições de tempo, falta de integração, falta de solução de análise preditiva apropriada e problemas com gerenciamento de mudanças.

## 5.1 Implicações Gerenciais

Através do método apresentado, foi possível modelar o processo, a partir dos dados disponíveis para a modelagem e a informação gerada foi utilizada para uma finalidade mais analítica. Através dos resultados apresentados, é esperado que decisões gerenciais possam ser feitas com maior precisão. Foi apresentada a possibilidade de automatizar o estudo de processos, com base em dados (*logs*) dos sistemas de operação, como uma abordagem alternativa para as técnicas de mapeamento comuns, com questionários e envolvimento dos atores no relato de suas atividades.

Os processos para execução de serviços representados em forma de Cadeias de Markov no caso estudado, de desembaraço aduaneiro, podem ser o ponto de partida para muitas questões relacionadas a diretrizes de gestão e otimização de processos.

As análises quantitativas originadas permitem entender a existência de probabilidades dentro de um processo, que podem ter várias naturezas de tomadas de decisão e tendem a ter maior precisão quando baseadas em dados do próprio sistema usado para traçar o processo real.

Neste estudo de caso, colocamos em prova um negócio que contém intrinsecamente muitas etapas e caminhos para poder entregar ao cliente o serviço esperado. Sendo assim, em um ambiente complexo, muitas vezes não é fácil saber quais etapas estão tomando mais tempo para serem executadas e se o processo é um elemento que de fato auxilia. Usando o método apresentado, tornou-se possível determinar as chances de que um serviço prestado tenha várias facetas.

Os números da estatística descritiva, como médias e desvio padrão, adicionam informações sobre a consistência do processo. Se as etapas possuem muitas variações na sua execução, as causas precisam ser analisadas.

Além de permitir maior controle sobre a execução dos processos e oferecer *insights* sobre as melhorias necessárias, os resultados apresentados pelos modelos permitem outras considerações gerenciais. Identificar se gargalos em certas áreas podem ser devidos a necessidade de mais recursos, ou de treinamentos aos recursos existentes ou de baixa performance de qualquer outra

natureza. Nos grafos obtidos, é possível observar os caminhos mais predominantes de forma geral e os caminhos coadjuvantes, e se o caminho predominante vier a ser o que possui maior indicador de *lead time*, então este passa a ser parte do foco de atenção.

O negócio analisado requer grande esforço humano em tarefas manuais e repetitivas e caso, parte do processo com maior gargalo venha a ser um que tenha muito uso de tarefas manuais, estudos sobre automação em lugares específicos podem ser levados em consideração também.

Além da análise de processos, os dados resultantes dessa primeira fase de análise permitem ao gestor combinar algumas variáveis a fim de obter respostas estratégicas com relação ao processo que está intimamente ligado com os clientes que compõem a sua carteira. Uma delas é a análise da produtividade da carteira, definida pelos diferentes processos customizados para cada cliente. Os clientes são os principais impulsionadores das atividades do processo, já que passos importantes dependem de revisões e aprovações que eles enviam à empresa, e com esta questão surge um ponto de partida para critérios importantes para explicar custos gerados pelos processos.

Algumas considerações econômicas podem ser estudadas caso o modelo venha acompanhado de valores relacionados ao custo, ou algum valor que represente as saídas da empresa para executar o processo. Usando semânticas de mais fácil compreensão para o corpo executivo da empresa, muitas decisões podem ser mais facilmente tomadas tendo em mãos um indicador que demonstre o que a empresa perde caso não dê a devida atenção aos processos descobertos e que lhe causem prejuízo, tais como os custos do erro, os custos de tarefas repetidas, os custos de tarefas manuais, os custos de espera ou custos de atividades específicas atreladas a um cliente com demandas peculiares. Alguns métodos de melhor precificação podem ser estudados e desenvolvidos para eliminar distorções na precificação que o gestor faz sem variáveis de performance de processos que este estudo revela.

Outra opção é tomar decisões sobre o próprio faturamento que podem afetar o fluxo de caixa, no que tange ao recebimento de contas. Se o processo leva um tempo indesejável para ser executado, todo um impacto negativo é gerado na parte financeira, já que os pagamentos são feitos ao final do processo, portanto acompanham a lentidão da prestação de serviços.

Ajustando o processo internamente e eliminando desperdícios de qualquer ordem, a empresa pode focar esforços em melhorar a sua eficiência operacional e se tornar mais competitiva em setores em que a guerra de preços é fator crucial para alcançar uma participação no mercado de maior relevância.

## 6 REFERÊNCIAS

- Abbasi, A., Sarker, S. & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, 17(2).
- Accorsi, R., Lehmann, A. & Lohmann, N. (2015). Information leak detection in business process models: Theory, application, and tool support. *Information Systems*, 47, 244-257.
- Agarwal R. & Dhar V. (2014). Editorial – big data, data science, and analytics: the opportunity and challenge for IS research. *Information Systems Research*, 25(3), 443-448. Acessado em 06 de fevereiro, 2021 <https://pubsonline.informs.org/doi/pdf/10.1287/isre.2014.0546>
- Aguilar-Savén E. S. (2004). Business process modelling: Review and framework. *International Journal of Production Economics*, 90, 129-149. DOI:10.1016/S0925-5273(03)00102-6.
- Akter, S., Wamba S. F., Gunasekaran A., Dubey R., & Childe S. J.. (2016). How to Improve Firm Performance Using *Big Data* Analytics Capability and Business Strategy Alignment? *International Journal of Production Economics*, 182, 113–131.
- Akpınar, M. (2019). Application of Genetic Algorithm for Optimization of Heat-Transfer Parameters. *Sakarya University Journal of Science*, 23(6), 1123–1130. <https://doi.org/10.16984/aufenbilder.500643>
- Albalaki, F. & Majeed R. J. (2018). Customer Profitability Analysis, Cost System Purposes and Decision Making Process: A Research Framework. *Account and Financial Management Journal*, 03. 10.31142/afmj/v3i5.03.
- Ahmad R., Tichadou S. & Hascoet JY. (2017). A knowledge-based intelligent decision system for production planning. *The International Journal of Advanced Manufacturing Technology*, 89(5-8), 1717–172, <http://doi.org/10.1007/s00170-016-9214-z>.
- Arnold J. B. (2019). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 4.2.0. <https://CRAN.R-project.org/package=ggthemes>.
- Arunachalam D., Kumar N. & Kawalek J.P. (2018). Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice, *Transportation Research Part E: Logistics and Transportation Review*, 114, 416-436. <https://doi.org/10.1016/j.tre.2017.04.001>.
- Azadeh, A., Moghaddam, M., Khakzad, M. & Ebrahimipour, V. (2012). A flexible neural network-fuzzy mathematical programming algorithm for improvement of oil price estimation and forecasting. *Computers e Industrial Engineering*, 62(2), 421–430. <https://doi.org/10.1016/j.cie.2011.06.019>.

- Bae, H. & Seo, Y. (2007). BPM-based integration of supply chain process modeling, executing and monitoring. *International Journal of Production Research*, 45(11), 2545–2566. <https://doi.org/10.1080/00207540601020502>.
- Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J. & Zhao, J. L. (2016). Transformational Issues of Big Data and Analytics in Networked Business. *MIS quarterly*, 38(2), 629-631.
- Bastian M., Heymann S. & Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*. Acessado em 7 de Julho, 2020 de <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.
- Bernstein, A., Klein, M., & Malone, T. W. (1999). The Process Recombinator: A Tool for Generating New Business Process Ideas. *Proceedings of the 20th International Conference on Information Systems*, P. De and J. I. DeGross (eds.), Charlotte, NC, 179-192.
- Bihani P. & Patil S.T. (2014). A comparative study of data analysis techniques. *International Journal of Emerging Trends & Technology in Computer Science*, 3(2), 95-101.
- Blackburn R., Lurz K., Priese B., Göb R. & Darkow I.L. (2015). A predictive analytics approach for demand forecasting in the process industry. *International Transactions in Operational Research*, 22(3), 407-428. <https://doi.org/10.1111/itor.12122>.
- Bowen, D.E. & Youngdahl, W.E. (1998), “Lean” service: in defense of a production-line approach. *International Journal of Service Industry Management*, 9(3), 207-225. <https://doi.org/10.1108/09564239810223510>.
- Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society*, 15, 662-679.
- Bożejko, W., Grymin R. & Pempera J. (2017). Scheduling and Routing Algorithms for Rail Freight Transportation. *Procedia Engineering*, 178, 206-212, <http://doi.org/10.1016/j.proeng.2017.01.098>
- Bożejko, W., Pempera J. & Smutnicki A. (2008). Parallel Single-Thread Strategies in Scheduling, in Rutkowski L., Tadeusiewicz R., Zadeh L.A., Zurada J.M. (Ed.), *Artificial Intelligence and Soft Computing*, ICAISC 2008, Lecture Notes in Computer Science, 5097, Springer, Berlin, Heidelberg, 995-1006, [http://doi.org/10.1007/978-3-540-69731-2\\_94](http://doi.org/10.1007/978-3-540-69731-2_94).
- Braithwaite, A. & Samakh, E. (1998). The Cost-to-Serve Method. *International Journal of Logistics Management*, 9(1), 69-84. <https://doi.org/10.1108/09574099810805753>
- Breuker, D., Matzner, M., Delfmann, P. & Becker, J. (2016). Comprehensible Predictive Models for Business Processes. *MIS quarterly*, 40. 2016. 10.25300/MISQ/2016/40.4.10.
- Brown, B., Chui M. & Manyika J. (2011). Are you ready for the era of Big Data? *The McKinsey Quarterly*, 4(2011), 24-35.

- Burduk, A., Musiał, K., Kochańska, J., Górnicka, D. & Stetsenko, A. (2019). Tabu Search and Genetic Algorithm for Production Process Scheduling Problem. *LogForum*, 15(2), 181–189. <https://doi.org/10.17270/J.LOG.2019.315>
- Burgy, R. & Bulbul K., (2018). The job shop scheduling problem with convex costs, *European Journal of Operational Research*, 268(1), 82-100, <http://doi.org/10.1016/j.ejor.2018.01.027>.
- Cayley, A. (1858). A Memoir on the Theory of Matrices. *Philosophical Transactions of the Royal Society of London*, 148, 17-37. Acessado em 6 de Fevereiro, 2021, de <http://www.jstor.org/stable/108649>.
- Campbell, N.G. e Cunningham, M.T. (1983), Customer analysis for strategy development in industrial marketing. *Strategic Management Journal*, 4, 369-80.
- Cao, G., Duan, Y. & Li, G. (2015). Linking Business Analytics to Decision Making Effectiveness: A Path Model Analysis. *IEEE Transactions on Engineering Management*. 62(3), 384-395. <https://ieeexplore.ieee.org/document/7132744>.
- Cardinaels, E., Roodhooft, F. & Warlop, L. (2004). Customer profitability analysis reports for resource allocation: the role of complex marketing environments. *Abacus*, 40(2), 238-258.
- Chen, H. & Storey, V. C. (2012). Business Intelligence and Analytics: From *Big Data* to Big Impact. *MIS quarterly*, 36(4), 1165-1188.
- Chen, L., Jansen, K. & Zhang, G. (2018). On the optimality of exact and approximation algorithms for scheduling problems, *Journal of Computer and System Sciences*, 96, 1-32. <http://doi.org/10.1016/j.jcss.2018.03.005>.
- Chen J., Chen Y., Du X., Li C., Lu J., Zhao S. & Zhou X. (2013). Big data challenge: a data management perspective. *Frontiers of Computer Science*, 7(2), 157-164.
- Chen C.L.P. & Zhang C.Y. (2014). Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Information Sciences*, 275, 314-347.
- Ching W.-K., Ng M. K. & Fung E. S. (2008). Higher-order multivariate markov chains and their applications. *Linear Algebra and its Applications*, 428(2-3), 492–507.
- Chiu, T., Brown-Liburd, H. & Vasarhelyi, M. A. (2019). Performing Tests of Internal Controls Using *Process Mining*. *CPA Journal*, 89(6), 54–57.
- Chiu, T., & Jans, M. (2019). Process Mining of Event Logs: A Case Study Evaluating Internal Control Effectiveness. *Accounting Horizons*, 33(3), 141–156. <https://doi.org/10.2308/acch-52458>.
- Clemons, E.K., Dewan, R.M., Kauffman, R.J. & Weber, T.A. (2017). Understanding the information-based transformation of strategy and society. *Journal of Management Information Systems*, 34(2), 425–456.

- Cohen S. & Roussel J. (2014). *Strategic Supply Chain Management: The Five Core Disciplines for Top Performance*. McGraw-Hill Education; 2 edition.
- Colombo, A., Damiani, E. & Gianini, G. (2006). Discovering the software process by means of stochastic workflow analysis. *Journal of Systems Architecture*, 52(11), 684–692. <https://doi.org/10.1016/j.sysarc.2006.06.012>.
- Craig B. A. & Sendi P. P. (2002). Estimation of the transition matrix of a discrete-time markov chain. *Health Economics*, 11(1).
- Csardi G. & Nepusz T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695. <https://igraph.org>.
- Davenport, T. H. & Short, J. E. (1990) The new industrial engineering: information technology and business process redesign, *Sloan Management Review*, 31(4), 11-27.
- Davenport, T.H., (1993). *Process innovation*. Boston, MA. Harvard Business School Press.
- DeCock, C. & Hiplin, I., (1997). TQM and BPR: beyond the beyond myth. *Journal of Management Studies*, 34(5), 659–674.
- Demchenko, Y., Grosso, P., De Laat, C. & Membrey, P. (2013). Addressing Big Data Issues in Scientific Data Infrastructure. In: First International Symposium on Big Data and Data Analytics in Collaboration (BDDAC 2013). Part of The 2013 International Conference on Collaboration Technologies and Systems (CTS 2013), San Diego, USA.
- Dickinson, M., Thornton, A. & Graves, S. (2001). Technology portfólio management: Optimizing interdependent projects over multiple time periods. *IEEE Transactions on Engineering Management*, 48(4), 518–527.
- Duan, L. & Y. Xiong. (2015). *Big Data Analytics and Business Analytics*. *Journal of Management Analytics*, 2(1), 1–21.
- Earl, M. J. (1994). The new and old of business process redesign. *Journal of Strategic Information Systems*, 3(1), 5-22.
- Eng, Teck-Yong (2004). Does customer portfólio analysis relate to customer performance? An empirical analysis of alternative strategic perspective. *Journal of Business e Industrial Marketing*, 19(1), 49-67.
- Erl, T., Khattak, W. & Buhler, P., (2016). *Big Data Fundamentals - Concepts, Drivers e Techniques*. Indiana: Pearson.
- Fantinato, M., I. M. S. Gimenes & M. B. F. Toledo. (2010). Product Line in the Business Process Management Domain. In *Applied Software Product Line Engineering*, edited by K. Kang, V. Sugumaran, and S. Park, 497–530. Boston, MA: Auerbach Publications.

- Fayyad, U., G. Piatetsky-Shapiro & P. Smyth. (1996). From *Data Mining* to Knowledge Discovery in Databases. *AI Magazine*, 17, 37–54.
- Fiocca, R. (1982). Account portfólio analysis for strategy development. *Industrial Marketing Management*, 11, 53-62.
- Franken, H.M., & Weger, M.K. (1997). A modelling framework for capturing business processes dynamics. *Knowledge and Process Management*, 4(3), 153-162.
- Francisco, E. de R. (2015). A tendência do *Big Data*. *GV-executivo*, 14(2),. Acessado em 22 de agosto, 2020. [https://rae.fgv.br/sites/rae.fgv.br/files/coluna\\_tecnologia.pdf](https://rae.fgv.br/sites/rae.fgv.br/files/coluna_tecnologia.pdf).
- Gartner. (2020). “Gartner IT Glossary.” Acessado em 22 de agosto, 2020 <http://www.gartner.com/it-glossary/big-data/>.
- Gdowska K., Książek R. (2017). A random search algorithm for cyclic delivery synchronization problem, *Logforum*, 13(3), 263-272. <http://doi.org/10.17270/J.LOG.2017.3.2>.
- Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. New York. Addison-Wesley.
- González López de Murillas, E., Reijers, H.A. & van der Aalst, W.M.P. (2019). Connecting databases with process mining: a meta model and toolset. *Softw Syst Model*, 18, 1209–1247. <https://doi.org/10.1007/s10270-018-0664-7>.
- Górnicka D., Markowski M. & Burduk A. (2018). Optimization of production organization in a packaging company by ant colony algorithm, in Burduk A., Mazurkiewicz D. (Ed.), *ISPEM 2017, AISC*, 637, Springer, Cham, 336-346. [http://doi.org/10.1007/978-3-319-64465-3\\_33](http://doi.org/10.1007/978-3-319-64465-3_33).
- Grainger A. (2008). Customs and trade facilitation: from concepts to implementation. *World Customs Journal*, 2(1), 11 – 14.
- Gregor, S., Martin M., Fernandez W., Stern S., e Vitale M. (2006). The Transformational Dimension in the Realization of Business Value from Information Technology. *The Journal of Strategic Information Systems*, 15(3), 17-30.
- Grigori, D., Casati, F., Castellanos, M. & Dayal, U. (2004). “Business Process Intelligence,” *Computers in Industry*, 53(3), 321-343
- Grover, V., Chiang, R. H. L., Liang, T.-P. & Zhang, D. (2018). Creating Strategic Business Value from *Big Data* Analytics: A Research Framework. *Journal of Management Information Systems*, 35(2), 388–423. <https://doi.org/10.1080/07421222.2018.1451951>.
- Guerreiro, R., Bio, S. R., Vazquez, E. & Merschmann, V. (2008). Cost-to-serve measurement and customer profitability analysis. *International Journal of Logistics Management*, 19(3), 389–407. <https://doi.org/10.1108/09574090810919215>.

- Guerry M. A. (2011). Hidden heterogeneity in manpower systems: A Markov-switching model approach. *European Journal of Operational Research*, 210(1), 106-113. DOI: 10.1016/j.ejor.2010.10.039.
- Gulden, G.K. & Reck, R.H., (1992). Combining quality and reengineering efforts for process excellence. *Information Strategy: The Executive's Journal*, 8(3), 10–16.
- Hall, G. H., Rosenthal, J. & Wade, J. (1993). How to make reengineering really work. *Harvard Business Review*, November-December, 119-131.
- Hammer, M. (1990). Reengineering work: don't automate, obliterate. *Harvard Business Review*, July-August, 104-112.
- Harmon, Paul. (2011). How Do Processes Create Value? Acessado em 07 de julho, 2020 <http://www.bptrends.com/publicationfiles/02-01-11-ARTHow%20do%20Processes%20Create%20Value-Harmon-Final.pdf>.
- Hausman, W.H., Lee, H.L., Napier, G.R.F., Thompson, A. & Zheng, Y. (2010). A Process Analysis of Global Trade Management: An Inductive Approach\*. *Journal of Supply Chain Management*, 46, 5-29. <https://doi.org/10.1111/j.1745-493X.2010.03187.x>
- Holland, J. (1992). Genetic Algorithms. *Scientific American*, 267(1), 66-73. Acessado em 25 de janeiro, 2021 <http://www.jstor.org/stable/24939139>
- Howell, R. A. & Soucy, S. (1990). Customer profitability. *Management Accounting*, 72(4), 43-47.
- Ibanez-Sanchez, G., Fernandez-Llatas, C., Martinez-Millana, A., Aparici-Tortajada, L., Valero-Ramon, Z., Traver, V., Celda, A., Mandingorra, J., Munoz-Gama, J., Sepúlveda, M., Gálvez, V., Rojas, E., & Capurro, D. (2019.). Toward value-based healthcare through interactive *Process Mining* in emergency rooms: The stroke case. *International Journal of Environmental Research and Public Health*, 16(10). <https://doi.org/10.3390/ijerph16101783>.
- İnkaya, T. & Akansel, M. (2017). Coordinated scheduling of the transfer lots in an assembly-type supply chain: a genetic algorithm approach. *Journal of Intelligent Manufacturing*, 28(4), 1005–1015. <https://doi.org/10.1007/s10845-015-1041-9>.
- Jacquelyn, T., Reinartz, W. & Kumar, V. (2004). Getting the most out of all your customers. *Harvard Business Review*, 82(7-8), 116-123.
- Janssen, K. J., Nendels, F. C. W., Smit, S. L. & Ravesteyn, P. (2015). Business Processes Management in the Netherlands and Portugal: The Effect of BPM Maturity on BPM Performance. *Journal of International Technology e Information Management*, 24(1), 33–52.
- John D. Kelleher & Brendan Tierney. (2018). Data Science. The MIT Press.
- Johnson M. & Selnes F. (2005). Diversifying your customer portfólio. *MIT Sloan Management Review*, 46(3), 11 - 14.

- Jónás, T., Kalló, N. & Tóth, Z. E. (2014). Application of Markov Chains for Modeling and Managing Industrial Electronic Repair Processes. *Periodica Polytechnica Social and Management Sciences*, 22, 87-98. 10.3311/PPso.7438.
- Jones C. I. (1997). On the evolution of the world income distribution. Available at SSRN 59412.
- Jorge, Gabriela e Miyake, Dario e Silva, Marcia. (2011). APLICAÇÃO DA ABORDAGEM LEAN NUM PROCESSO DE SERVIÇO DO SETOR FINANCEIRO: UM ESTUDO DE CASO. *Conference: XIV Simpósio de Administração da Produção, Logística e Operações Internacionais, At São Paulo*. Acessado em 07 de julho, 2020 [https://www.researchgate.net/publication/296706166\\_APLICACAO\\_DA\\_ABORDAGEM\\_LEAN\\_NUM\\_PROCESSO\\_DE\\_SERVICO\\_DO\\_SETOR\\_FINANCEIRO\\_UM\\_ESTUDO\\_DE\\_CASO](https://www.researchgate.net/publication/296706166_APLICACAO_DA_ABORDAGEM_LEAN_NUM_PROCESSO_DE_SERVICO_DO_SETOR_FINANCEIRO_UM_ESTUDO_DE_CASO)
- Joseph R.C. & Johnson N.A. (2013). Big data and transformational government. *IT Professional*, 15(6), 43-48.
- Kaplan, R. B. & Murdock, L. (1991). Core process redesign. *The McKinsey Quarterly*, 2, 27-43.
- Kaplan, Robert & Narayanan, V. (2001). Customer Profitability Measurement and Management Customer Profitability Measurement and Management. *Journal of cost management*.
- Karlaftis, M. G. & Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some *insights*. *Transportation Research: Part C*, 19(3), 387–399. <https://doi.org/10.1016/j.trc.2010.10.004>
- Ketter, W., Peters, M., Collins, J. & Gupta, A. (2016). Competitive Benchmarking: An IS Research Approach To Address Wicked Problems With *Big Data* And Analytics. *MIS quarterly*, 40(4), 1057.
- Kingman-Brundage, J., George, W. R. & Bowen, D. E. (1995). “Service logic”: achieving service system integration. *International Journal of Service Industry Management*, 6(4), 20-39.
- Kiron, D., P. K. Prentice & R. B. Ferguson. (2014). The Analytics Mandate. *MIT Sloan Management Review*, 55(4), 1.
- Kiron, D. & Shockley R. (2011). Creating Business Value with Analytics. *MIT Sloan Management Review*, 53, 57–63.
- Kone, E. R. S. & Karwan, M. H. (2011). Combining a new data classification technique and regression analysis to predict the Cost-To-Serve new customers. *Computers e Industrial Engineering*, 61(1), 184–197. <https://doi.org/10.1016/j.cie.2011.03.009>.
- Kotowska J., Markowski M. & Burduk A., (2018). Optimization of the Supply of Components for Mass Production with the Use of the Ant Colony Algorithm, in Burduk A., Mazurkiewicz D. (Ed.), ISPEM 2017, AISC, 637, Springer, Cham, 347-357, [http://doi.org/10.1007/978-3-319-4465-3\\_34](http://doi.org/10.1007/978-3-319-4465-3_34).

- Krapfel, R., Salmond, D. e Spekman, R. (1991). A strategic approach to managing buyer-seller relationships. *European Journal Of Marketing*, 25(9), 22-37.
- Kuchta, D., & Troska, M. (2007). Activity-based costing and customer profitability. *Journal of cost management*, 21(3), 18-25.
- Kumar, R., Ajjan, H. & Niu, Y. (2018). Information technology Portfólio Management: literature review, framework, and research issues. *Information Resource Management Journal*, 21(3), 64-87.
- Kumar A., Niu F. & Ré C. (2013). Hazy: making it easier to build and maintain big-data analytics. *Communications of the ACM*, 56(3), 40-49.
- Lakshmanan, G.T., Shamsi, D., Doganata, Y.N. et al. (2015). A markov prediction model for data-driven semi-structured business processes. *Knowl Inf Syst*, 42, 97–126. <https://doi.org/10.1007/s10115-013-0697-8>.
- Lavalle S., Lesser E., Shockley R., Hopkins M.S. & Kruschwitz N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, 52, 20-32.
- Lavalle S., Lesser E., Shockley R., Hopkins M.S. & Kruschwitz N. (2017). Analytics: The new path to value: How the smartest organizations are embedding analytics to transform insights into action. *MIT Sloan Management Review*, 12, 1-28.
- Lee, H.L. & Silverman A. (2008). Renault's Logan Car: Managing Customs Duties for a Global Product. *Stanford Graduate School of Business Case GS*, 62.
- Lee, J., Wyner, G. M. & Pentland, B. T. (2008). Process Grammar as a Tool for Business Process Design. *MIS quarterly*, 32(4), 757–778. <https://doi.org/10.2307/25148871>.
- Letouzé, E. (2012). Big Data for Development: Challenges & Opportunities. *UN Global Pulse*.
- Levitt, T. (1972). Production-line approach to service. *Harvard Business Review*, 50(5), 20-31.
- Lieb R.C., Millen R.A. & Wassenhove L.V. (1993). Third-party logistics services: a comparison of experienced American and European manufacturers. *International Journal of Physical Distribution e Logistics Management*, 6(23), 35-44.
- Lin H.-F. (2014). Understanding the determinants of electronic supply chain management system adoption: using the technology–organization–environment framework. *Technological Forecasting and Social Change*, 86, 80-92.
- Luvizan, S., Meirelles, F., & Diniz, E. (2015). Big Data: Evolução Das Publicações E Oportunidades De Pesquisa. *Revista Eletrônica de Sistemas de Informação*, 14(3).
- Malone, T. W., Crowston, K., Lee, J., Pentland, B., Dellarocas, C., Wyner, G., Quimby, J., Osborne, C., Bernstein, A., Herman, G. & Klein, M. (1999). Tools for Inventing Organizations: Toward a Handbook of Organizational Processes. *Management Science*, 45(3), 425-443.

- Maloni, M. J., & C. R. Carter. (2006). Opportunities for Research in Third Party Logistics. *Transportation Journal*, 45(2), 23–38.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity. *Report by McKinsey Global Institute*.
- Markov A. A. (1907). Issledovanie zamechatel nogo sluchaya zavisimyh ispytanij. *Izvestiya Akademii Nauk*, SPb, VI seriya, 1(93), 61–80.
- Markowitz, H. (1952). Portfólio Selection. *The Journal of Finance*, 7(1), 77-9.
- Martin, K. (2015). Ethical issues in the *Big Data* industry. *MIS quarterly Executive*, 14(2), 6785.
- Meixell, M.J. & Gargeya V.B. (2005). Global Supply Chain Design: A Literature Review and Critique. *Transportation Research Part E*, 41, 531-550.
- McAfee, A., E. Brynjolfsson, T. H. Davenport, D. J. Patil, & D. Barton. (2012). “*Big Data*.” The Management Revolution. *Harvard Business Review*, 90(10), 61–67.
- Müller, O., Junglas, I., Brocke, J. V. & Debortoli, S. (2016). Utilizing *Big Data* analytics for *Information Systems Research*: challenges, promises and guidelines. *European Journal of Information Systems*, 25(4), 289-302.
- Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of ‘datification’. *The Journal of Strategic Information Systems*, 24(1), 3-14.
- Niraj, R., Gupta, M. & Narasimhan, C. (2001). Customer profitability in a supply chain. *Journal Of Marketing*, 65(3), 1-16.
- Nogueira, F. (2016). Modelagem e Simulação - Cadeias de Markov. *Notas de aula*. Acessado em 01 de outubro, 2020 <https://www.ufjf.br/epd042/files/2009/02/cadeiaMarkov1.pdf>.
- Noone, B., & Griffin, P. (1997). Enhancing yield management with customer profitability analysis. *International Journal of Contemporary Hospitality Management*, 9(2), 75-79.
- Ozdemir, S. (2016). Principles of Data Science. Packt Publishing.
- Page L., Brin S., Motwani R., & Winograd T. (1999). The pagerank citation ranking: Bringing order to the web. *Technical Report 1999-66, Stanford InfoLabURL*. Acessado em 6 de Fevereiro, 2021 de <http://ilpubs.stanford.edu:8090/422/>.
- Paprocka I., Gwiazda A. & Baczkowicz M. (2017) . Scheduling of an assembly process of a chosen technical mean using the critical chain approach. *MATEC Web of Conferences*, 94, <http://doi.org/10.1051/mateconf/20179406015>.
- Pardo, C. & Salle, R. (1995), “Defining customer boundaries: the first step in customer portfólio management”, in Turnbull, P.W., Yorke, D. e Naude, P. (Eds), *Interaction, Relationships and Networks*:

- Past-Present-Future, *IMP 11th International Conference, Manchester Federal School of Business and Management*, Manchester, 2, 962-78.
- Park, H., & Leydesdorff, L. (2013). Decomposing Social And Semantic Networks In Emerging “Big Data” Research. *Journal of Informetrics*, 7, 756-765.
- Pil Han, S., Park, S., & Oh, W. (2016). Mobile App Analytics: A Multiple Discrete-Continuous Choice Framework. *Management Information Systems Quarterly*. 40(4).
- Ponsignon, F., Smart, P. A., & Maull, R. S. (2012). Process design principles in service firms: Universal or context dependent? A literature review and new research directions. *Total Quality Management e Business Excellence*, 23(11/12), 1273–1296. <https://doi.org/10.1080/14783363.2011.637797>.
- Porter M., *Competitive Advantage*, New York: Free Press, 1998.
- Provost, F., & Fawcett, T. (2013). *Data Science for Business : What You Need to Know About Data Mining and Data-Analytic Thinking: Vol. 1st ed.* O’Reilly Média.
- Qian, L., & Ben-Arieh, D. (2008). Parametric cost estimation based on activity-based costing: A case study for design and development of rotational parts. *International Journal of Production Economics*, 113(2), 805–818. <https://doi.org/10.1016/j.ijpe.2007.08.010>.
- R 4.0.2. Linguagem e ambiente de desenvolvimento integrados, para cálculos estatísticos e gráficos. [S.I.] R Development Core Team, 2020. Acessado em 7 de Fevereiro, 2021 de <http://www.R-project.org/>.
- R Development Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Acessado em 7 de Fevereiro, 2021 de <http://www.R-project.org/>.
- Razzaque, M., & C. Sheng. (1998). Outsourcing of Logistics Functions: A Literature Survey. *International Journal of Physical Distribution and Logistics Management*, 30(2), 112–127.
- Receita Federal (2019). A descrição detalhada da mercadoria na DI pode ser feita em língua inglesa ou em outro idioma estrangeiro? *Ministério da Economia*. Acessado em 25 de janeiro, 2021 <https://www.gov.br/receitafederal/pt-br/assuntos/aduana-e-comercio-exterior/manuais/despacho-de-importacao/sistemas/siscomex-importacao-web/declaracao-de-importacao/funcionalidades/elaborar-uma-nova-solicitacao-de-di/preenchimento-da-di-1/formularios-de-dados-especificos-da-adicao/a-descricao-detalhada-da-mercadoria-na-di-pode-ser-feita-em-lingua-inglesa>.
- Receita Federal (2020). Despacho de Importação – Introdução. *Ministério da Economia*. Acessado em 25 de janeiro, 2021 <https://www.gov.br/receitafederal/pt-br/assuntos/aduana-e-comercio-exterior/manuais/despacho-de-importacao/topicos-1/conceitos-e-definicoes/despacho-de-importacao>.
- Receita Federal. NCM. *Ministério da Economia*. Acessado em 25 de janeiro, 2021 <https://receita.economia.gov.br/orientacao/aduaneira/classificacao-fiscal-de-mercadorias/ncm>.

- Reinartz, W. & Kumar, V. (2000) n the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal Of Marketing*, 41(3), 293-305.
- RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. Acessado em 7 de Fevereiro, 2021 de <http://www.rstudio.com/>.
- Rubinstein M. (2002). M. Markowitz's "Portfólio Selection": A Fifty-Year Retrospective. *The Journal of Finance*. 57(2), 1041-1045.
- Sahay, B. S., & R. Mohan. (2006). 3PL Practices: An Indian Perspective. *International Journal of Physical Distribution e Logistics Management*, 36(9), 666–689.
- Schoenherr T. & Speier-Pero C. (2015). Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential. *Journal of Business Logistics*. 36(1), 120-132. <https://doi.org/10.1111/jbl.12082>.
- Scrucca L. (2013). GA: A Package for Genetic Algorithms in R. *Journal of Statistical Software*, 53(4), 1-37. <http://www.jstatsoft.org/v53/i04/>.
- Shapiro, B., Rangan, K., Moriarty, R. & Ross, E. (1987), Manage customers for profits (not just sales). *Harvard Business Review*, September-October, 101-8.
- Shmueli, G., & Koppius, O. R. (2011). Predictive Analytics in *Information Systems Research*. *MIS quarterly* 35(3), 553-572.
- Sivarajah U., Kamal M.M., Irani Z. & Weerakkody V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263-286. <https://doi.org/10.1016/j.jbusres.2016.08.001>.
- Smith, H. & Fingar, P. (2003). *Business Process Management – The Third Wave*. Tampa, FL. MeghanKiffer Press.
- Spedicato G.A. (2017). Discrete Time Markov Chains with R. *The R Journal*. R package version 0.6.9.7. <https://journal.r-project.org/archive/2017/RJ-2017-036/index.html>.
- Souza C. G., (2014). Supply chain analytics. *Business Horizons*, 57(5), 595-605, ISSN 0007-6813. <https://doi.org/10.1016/j.bushor.2014.06.004>.
- Spedicato G. A., Kang T. S., Yalamanchi S. B., Yadav D. & Cordón I. (2017). The markovchain Package: A Package for Easily Handling Discrete Markov Chains in R.
- Sun, L., Karwan, M. H., Gemici-Ozkan, B., & Pinto, J. M. (2015). Estimating the long-term cost to serve new customers in joint distribution. *Computers e Industrial Engineering*, 80, 1–11. <https://doi.org/10.1016/j.cie.2014.11.012>.
- Szongott C., Henne B. & von Voigt G. (2012). Big data privacy issues in public social media. *6th IEEE international conference on digital ecosystems technologies (DEST)*, 1-6.

- Teo, H-H, Tan, BCY & Wei, K-K. (1997), Organizational transformation using electronic data interchange: the case of TradeNet in Singapore. *Journal of Management Information Systems*, 13(4), 139-166.
- Terho, H. (2008). Customer Portfólio Management – The Construct and Performance. Acessado em 28 de julho, 2020 [https://www.impgroup.org/uploads/dissertations/dissertation\\_31.pdf](https://www.impgroup.org/uploads/dissertations/dissertation_31.pdf)
- Turnbull, P. (1990). A review of portfólio planning models for industrial marketing and purchasing management. *European Journal Of Marketing*, 24(3), 7-22.
- Urso A., Fiannaca A., La Rosa M., Valentina R. & Rizzo R. (2019). Data Mining: Prediction Methods. *Encyclopedia of Bioinformatics and Computational Biology*, 1, 413-430. <https://doi.org/10.1016/B978-0-12-809633-8.20462-7>
- van der Aalst, W. M. P., Schonenberg, M. H., & Song, M. (2011). Time Prediction Based on *Process Mining*. *Information Systems Journal*, 36(2), 450-475.
- van der Aalst, W. M. P., A. J. M. M. Weijters, & L. Maruster. (2003). “Workflow Mining: Discovering Process Models from Event Logs.” *IEEE Transactions on Knowledge and Data Engineering*, 16, 1128–1142. doi:10.1109/TKDE.2004.47.
- van der Aalst, W. M. P., van Dongen, B., Herbst, J., Maruster, L., Schimm, G., & Weijters, A. (2003). Workflow Mining: A Survey of Issues and Approaches. *Data and Knowledge Engineering*, 47(2), 237-267.
- van der Aalst & W. M. P. (1998). “The Application of Petri Nets to Workflow Management.” *Journal of Circuits, Systems and Computers*, 8(1), 21–66. doi:10.1142/S0218126698000043.
- van der Aalst, W. M. P. (2013). Business Process Management: A Comprehensive Survey. *ISRN Software Engineering*, 1-37. <http://dx.doi.org/10.1155/2013/507984>.
- van der Aalst, W. M. (2016). *Process Mining – Data Science in Action*. 1st ed. Berlin Heidelberg: Springer.
- van der Aalst, W. M. P., Weijters, T., & Maruster, L. (2004). Workflow Mining: Discovering Process Models from Event Logs. *IEEE Transactions on Knowledge and Data Engineering* 16(9), 1128-1142.
- Van Raaij, & M. (2005). The strategic value of customer profitability analysis. *Marketing Intelligence and Planning*, 23(4), 372-381.
- Voican, O. (2020). Using *Data Mining* Methods to Solve Classification Problems in Financial-Banking Institutions. *Economic Computation and Economic Cybernetics Studies and Research*, 54(1), 159–176. <https://doi.org/10.24818/18423264/54.1.20.11>.
- Yi X., Liu F., Liu J. & Jin H (2014). Building a network highway for big data: architecture and challenges. *IEEE Network*, 28(4), 5-13.
- Waller M.A. & Fawcett S.E. (2013) Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2), 77-84.

- Wamba, S. F., S. Akter, A. Edwards, G. Chopin, & D. Gnanzou. (2015). How 'Big Data' Can Make Big Impact: Findings from a Systematic Review and a Longitudinal Case Study. *International Journal of Production Economics* 165, 234–246.
- Wang, H. S. (2007). Application of BPN with feature-based models on cost estimation of plastic injection products. *Computers and Industrial Engineering*, 53(1), 79–94. <https://doi.org/10.1016/j.cie.2007.04.005>.
- Weill P., Ross J.W. (2009). IT savvy: What top executives must know to go from pain to gain. *Harvard Business Press*.
- Wen, L., van der Aalst, W. M. P., Wang, J., & Sun, J. (2007). Mining Process Models with Non-Free-Choice Constructs. *Data Mining and Knowledge Discovery*, 15(2), 145-180.
- Wickham H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. Springer-Verlag New York. ISBN 978-3-319-24277-4. <https://ggplot2.tidyverse.org>.
- Wickham H. & Bryan J. (2019). readxl: Read Excel Files. R package version 1.3.1. <https://CRAN.R-project.org/package=readxl>.
- Wickham H., François R., Henry L. e Müller K. (2020). dplyr: A Grammar of Data Manipulation. R package version. 1.0.2. <https://CRAN.R-project.org/package=dplyr>.
- Williamson, O.E. (1981). The Economics of Organization: The Transaction Cost Approach. *The American Journal of Sociology*, 87, 548-577.
- Wind, Y. & Mahajan, V. (1981). "Designing product and business portfolios", *Harvard Business Review*, 155-65.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining : Practical Machine Learning Tools and Techniques: Vol. Fourth edition*. Morgan Kaufmann.
- Wu C. H. & Chuang Y. T. (2010). An innovative approach for strategic capacity portfolio planning under uncertainties. *European Journal of Operational Research*, 207(2), 1002-1013. DOI: 10.1016/j.ejor.2010.05.015.
- Wu, I.-L. (2003). Understanding senior management's behavior in promoting the strategic role of IT in process reengineering: use of the theory of reasoned action. *Information and Management*, 41(1), 1–11.
- Xue, S., Yuan, X., & Elhoseny, M. (2019). Intelligent system for products personalization and design using genetic algorithm. *Journal of Intelligent and Fuzzy Systems*, 37(1), 63–70. <https://doi.org/10.3233/JIFS-179064>

- Zacharia, Zach e Sanders, Nada & Nix, Nancy. (2011). The Emerging Role of the Third-Party Logistics Provider (3PL) as an Orchestrator. *Journal of Business Logistics*. 32, 40 - 54. 10.1111/j.2158-1592.2011.01004.x.
- Zeithaml, V.A., Berry L.L. & Parasuraman, A. (1988). Communication and control processes in the delivery of service quality. *Journal Of Marketing*, 52(2), 35-48.
- Zhou, Z. H., N. V. Chawla, Y. Jin, & G. J. Williams. (2014). *Big Data* Opportunities and Challenges: Discussions from Data Analytics Perspectives [Discussion Forum]. *IEEE Computational Intelligence Magazine*, 9(4), 62–74.

**APÊNDICE A – Receita em USD por Declaração de Importação durante os meses de 2013 a 2020.**

Este apêndice lista os valores monetários de receita média por Declaração de Importação do departamento analisado desde 2013 até 2020. A cada novo ano em uma coluna, são colocados o valor médio em dólar e a variação em porcentagem do mês com o mesmo mês do ano anterior. No último ano, a pesquisadora pode capturar os dados até o mês de abril.

Mês	Receita no ano/mês e variação em relação ao ano anterior														
	2013	2014		2015	2016	2017	2018		2019	2020					
Jan	\$ 400,00	\$ 360,00	-10,00%	\$ 337,00	-6,60%	\$ 297,00	-11,60%	\$ 281,00	-5,40%	\$ 286,00	1,70%	\$ 266,00	-7,00%	\$ 202,00	-23,90%
Feb	\$ 572,00	\$ 505,00	-11,80%	\$ 319,00	-36,70%	\$ 274,00	-14,00%	\$ 343,00	25,00%	\$ 296,00	-13,60%	\$ 293,00	-1,20%	\$ 215,00	-26,50%
Mar	\$ 415,00	\$ 310,00	-25,40%	\$ -82,00	-	\$ 215,00	+	\$ 286,00	33,10%	\$ 285,00	-0,40%	\$ 287,00	0,50%	\$ 147,00	-48,50%
Apr	\$ 392,00	\$ 582,00	48,60%	\$ 296,00	-49,20%	\$ 247,00	-16,40%	\$ 278,00	12,40%	\$ 228,00	-17,80%	\$ 148,00	-35,00%	\$ 193,00	30,30%
May	\$ 427,00	\$ 430,00	0,60%	\$ 176,00	-59,10%	\$ 285,00	62,20%	\$ 265,00	-7,00%	\$ 302,00	14,00%	\$ 204,00	-32,40%		
Jun	\$ 416,00	\$ 423,00	1,70%	\$ 492,00	16,20%	\$ 304,00	-38,10%	\$ 239,00	-21,50%	\$ 229,00	-3,90%	\$ 213,00	-7,10%		
Jul	\$ 370,00	\$ 416,00	12,20%	\$ 325,00	-21,80%	\$ 316,00	-2,60%	\$ 296,00	-6,50%	\$ 318,00	7,30%	\$ 180,00	-43,20%		
Aug	\$ 382,00	\$ 443,00	16,00%	\$ 230,00	-48,10%	\$ 313,00	36,10%	\$ 290,00	-7,40%	\$ 298,00	3,00%	\$ 222,00	-25,60%		
Sep	\$ 359,00	\$ 382,00	6,50%	\$ 255,00	-33,20%	\$ 269,00	5,50%	\$ 414,00	53,80%	\$ 364,00	-12,20%	\$ 248,00	-31,70%		
Oct	\$ 439,00	\$ 389,00	-11,30%	\$ 165,00	-57,40%	\$ 256,00	54,80%	\$ 294,00	14,90%	\$ 363,00	23,30%	\$ 209,00	-42,30%		
Nov	\$ 332,00	\$ 331,00	-0,30%	\$ 254,00	-23,40%	\$ 302,00	19,00%	\$ 243,00	-19,50%	\$ 327,00	34,50%	\$ 205,00	-37,30%		
Dec	\$ 422,00	\$ 377,00	-10,70%	\$ 218,00	-41,90%	\$ 251,00	14,70%	\$ 386,00	53,70%	\$ 268,00	-30,60%	\$ 202,00	-24,60%		
YTD	\$ 441,00	\$ 441,00	-0,10%	\$ 229,00	-48,00%	\$ 256,00	11,80%	\$ 296,00	15,40%	\$ 271,00	-8,40%	\$ 232,00	-14,20%	\$ 188,00	-19,00%
Total	\$ 408,00	\$ 413,00	1,20%	\$ 265,00	-35,80%	\$ 277,00	4,60%	\$ 299,00	7,90%	\$ 293,00	-2,00%	\$ 215,00	-26,60%	\$ 188,00	-12,50%

## APÊNDICE B – Descrição do significado dos eventos do processo.

Este apêndice descreve o código dos eventos utilizados para descrever em qual etapa do processo uma Declaração de Importação se encontra. Esses códigos são uma definição da empresa estudada e seguem o padrão de três letras.

<b>Evento</b>	<b>Significado</b>
LOG	Referência criada no sistema
DXR	Documentos recebidos
COB	Embarque confirmado
ADD	Chegada da carga
GAT	Carga liberada pelo terminal
PAI	Valores financeiros pré-calculados e necessidade de receber valores adiantados do cliente
PAR	Recebimento dos valores
ENF	Registro da declaração
ECR	Declaração em canal verde, indicando que tudo está em conformidade
EFD	Declaração em canal amarelo, indicando necessidade de inspeção de documentos
EXR	Declaração em canal vermelho, indicando necessidade de inspeção física da mercadoria
RLS	Declaração desembaraçada
TNF	Transportadora notificada
DLV	Mercadoria entregue

## APÊNDICE C – SQL usado na extração dos dados.

Este apêndice apresenta os comandos executados na base transacional do sistema de acompanhamento de despacho aduaneiro, codificados segundo linguagem SQL, padrão amplamente empregado por ferramentas de gestão de bancos de dados relacionais. Trata-se de um sistema global, distribuído por filial.

```

WITH raw_data
  AS (SELECT x.*,
            Row_number()
              OVER (
                partition BY x.invoice_no
                ORDER BY x.invoice_no DESC) AS RN_INVOICE_NO
  FROM   (SELECT Max(sl.consignee) AS CONSIGNEE,
              Max(( CASE
                WHEN EV.event_time_1 > 19000101 THEN Timestamp(
                  Rtrim(Char(EV.event_time_1))
                  || Substr(Digits(
                    CASE
                      WHEN EV.event_time_2 >=
                        11100 THEN
                        EV.event_time_2
                      ELSE 12000000
                    END), 3, 6))
                END )) AS Data_Evento,
              ev.event_code AS Cod_Evento,
              ev.operator,
              Max(ev.invoice_no) AS Invoice_No
  FROM   iasdb.event ev
  INNER JOIN iasdb.ship_log sl
            ON ev.invoice_no = sl.invoice_no
  WHERE  sl.invoice_no IN (SELECT ev_enf.invoice_no
                          FROM   iasdb.event ev_enf
                          WHERE  ev_enf.event_code = 'LOG'
                          AND ev_enf.event_time_1 BETWEEN
                              20130101 AND 20200128)
  AND sl.invoice_type IN ( '2', '7' )
  AND sl.ship_type IN ( 'F' )
  AND ev.operator IS NOT NULL
  AND ev.event_code IN ( 'COB', 'ADD', 'ENF', 'RLS',
                        'PAI', 'PAR', 'DXR', 'GAT',
                        'LOG', 'ECR', 'EFD', 'EXR',
                        'TNF', 'DLV' )

  GROUP BY ev.event_code,
           sl.invoice_type,
           sl.ship_type,
           ev.operator,
           ev.invoice_no
  ORDER BY ev.invoice_no,
           data_evento,

```

```

                                ev.operator) x),
raw_data_t_2
AS (SELECT RD.*,
      CASE
        WHEN Mod(RD.rn_invoice_no, 2) = 1 THEN 0
        WHEN Mod(RD.rn_invoice_no, 2) = 0 THEN 1
      END AS T
  FROM   raw_data RD),
raw_data_t_t_next
AS (SELECT RD2.consignee,
          RD2.data_evento,
          RD2.cod_evento,
          RD2.operator,
          RD2.invoice_no,
          RD2.rn_invoice_no,
          RD2.t,
          CASE
            WHEN ( RD2.t = 0 ) THEN RD2.cod_evento
          END AS C0,
          CASE
            WHEN ( RD2.t = 1 ) THEN RD2.cod_evento
          END AS C1,
          CASE
            WHEN ( RD2.t = 0 ) THEN RD2.data_evento
          END AS E0,
          CASE
            WHEN ( RD2.t = 1 ) THEN RD2.data_evento
          END AS E1,
          CASE
            WHEN ( RD2.t = 0 ) THEN RD2.operator
          END AS OP0,
          CASE
            WHEN ( RD2.t = 1 ) THEN RD2.operator
          END AS OP1
  FROM   raw_data_t_2 RD2),
raw_data_control_e0_e1
AS (SELECT RDTN.consignee,
          RDTN.invoice_no,
          RDTN.c0 AS C0,
          RDTN.c1 AS C1,
          RDTN.e0 AS E0,
          RDTN.e1 AS E1,
          RDTN.op0 AS OP0,
          RDTN.op1 AS OP1,
          RDTN.rn_invoice_no,
          RDTN.t,
          CASE
            WHEN ( RDTN.e0 IS NOT NULL
                  AND RDTN.t = 0 ) THEN RDTN.rn_invoice_no + RDTN.t
            WHEN ( RDTN.e1 IS NOT NULL
                  AND RDTN.t = 1 ) THEN RDTN.rn_invoice_no - RDTN.t
          END
          AS CONTROL_E0_E1,
          CASE
            WHEN ( RDTN.e0 IS NULL
                  AND RDTN.t = 1 ) THEN RDTN.rn_invoice_no + RDTN.t
            WHEN ( RDTN.e1 IS NULL

```

```

                AND RDTN.t = 0 ) THEN RDTN.rn_invoice_no - RDTN.t
            END          AS CONTROL_E1_E0
    FROM      raw_data_t_t_next RDTN),
e0_e1_matrix
AS (SELECT EE.consignee,
          EE.invoice_no,
          Max(EE.c0)      AS C0,
          Max(EE.c1)      AS C1,
          Max(EE.e0)      AS E0,
          Max(EE.e1)      AS E1,
          Max(EE.op0)     AS OP0,
          Max(EE.op1)     AS OP1,
          EE.control_e0_e1 AS CONTROL,
          'FIRST UNION'   AS SOURCE
    FROM      raw_data_control_e0_e1 EE
    GROUP BY EE.consignee,
            EE.invoice_no,
            EE.control_e0_e1
    HAVING Max(EE.c1) IS NOT NULL
           AND Max(EE.c0) IS NOT NULL
    UNION
    SELECT EE.consignee,
          EE.invoice_no,
          Max(EE.c1)      AS C0,
          Max(EE.c0)      AS C1,
          Max(EE.e1)      AS E0,
          Max(EE.e0)      AS E1,
          Max(EE.op0)     AS OP0,
          Max(EE.op1)     AS OP1,
          EE.control_e1_e0 AS CONTROL,
          'SECOND UNION'  AS SOURCE
    FROM      raw_data_control_e0_e1 EE
    GROUP BY EE.consignee,
            EE.invoice_no,
            EE.control_e1_e0
    HAVING Max(EE.c1) IS NOT NULL
           AND Max(EE.c0) IS NOT NULL),
invoices (invoice_no_ref, cancel_ref)
AS (SELECT DISTINCT BH.invoice_no_ref,
                  Cast(COALESCE(BH.cancel_ref, '') AS VARCHAR(10)) AS
                  CANCEL_REF
    FROM      iasdb.ship_log SL
            INNER JOIN iasdb.bill_header BH
                    ON SL.invoice_no = BH.invoice_no
            INNER JOIN e0_e1_matrix M
                    ON SL.invoice_no = M.invoice_no),
valid_invoices (invoice_no, invoice_no_ref, print_no)
AS (SELECT SL.invoice_no,
          BH.invoice_no_ref,
          BH.print_no
    FROM      iasdb.ship_log SL
            INNER JOIN iasdb.bill_header BH
                    ON SL.invoice_no = BH.invoice_no
    WHERE     BH.invoice_no_ref IN (SELECT invoice_no_ref
                                   FROM invoices)
           AND BH.invoice_no_ref NOT IN (SELECT cancel_ref

```

```

                                FROM invoices)
                                AND LEFT(BH.invoice_no_ref, 1) <> 'C')
SELECT M.*,
       Timestampdiff(8, Char(M.e1 - M.e0)) AS HOURS_DIFF,
       II.bill_code,
       II.bill_curr_cd,
       Cast(II.bill_amount / 100 AS DECIMAL(10, 2)) AS AMOUNT,
       II.bill_text
FROM   e0_e1_matrix M
       INNER JOIN valid_invoices VI
              ON M.invoice_no = VI.invoice_no
       INNER JOIN iasdb.invoice_item II
              ON VI.invoice_no = II.invoice_no
              AND VI.print_no = II.print_no
WHERE  II.bill_code <> '0000'
       AND II.bill_amount <> 0
       AND II.bill_code IN ( '4003', '4005', '4388', '4388', '4429' )

```

## APÊNDICE D – Variáveis utilizadas nas análises.

Este apêndice descreve os nomes das variáveis da fonte de dados utilizada nas análises e o que elas significam.

<b>Variável</b>	<b>Descrição</b>
CONSIGNEE	Identificação do cliente
INVOICE_NO	Identificação da transação do negócio
C0	Código do evento inicial
C1	Código do evento secundário
E0	Data e hora do evento inicial C1
E2	Data e hora do evento secundário C2
OP0	Sigla do sistema do operador responsável pelo evento inicial C1
OP2	Sigla do sistema do operador responsável pela evento secundário C2
HOURS_DIFF	Horas entre o evento inicial e o evento secundário
BILL_CODE	Código contábil para o faturamento da receita associada a transação
BILL_CURR_CD	Código da moeda usada para o faturamento da transação
AMOUNT	Quantia faturada ao cliente
BILL_TEXT	Descrição do tipo de item faturado ao cliente

## APÊNDICE E – Exemplo do formato dos dados coletados

Este apêndice lista exemplo dos dados do arquivo fonte utilizado nas análises. A descrição de cada coluna está no apêndice D.

CONSIGNEE	INVOICE_NO	C0	C1	E0	E1	OP0	OP1	HOURS_ DIFF	BILL_ CODE	BILL_ CURR_ CD	AMOUNT	BILL_TEXT
Manl Com	2590132366	CEK	PAI	8/14/2019 10:00	8/14/2019 10:07	EX	BB	0	4003	BRL	862,15	DESEMBARACO
Manl Com	2590132366	RLS	TOD	8/19/2019 13:30	8/19/2019 14:32	EX	BB	1	4003	BRL	862,15	DESEMBARACO
Manl Com	2590132366	ITY	CEK	8/12/2019 17:14	8/14/2019 10:00	EX	BB	40	4003	BRL	862,15	DESEMBARACO
Kaze Rou	2590132388	COB	ADD	8/12/2019 8:02	8/12/2019 14:46	QQ	EX	6	4003	BRL	937,47	DESEMBARACO
Kaze Rou	2590132388	DXR	COB	8/8/2019 10:56	8/12/2019 8:02	QQ	EX	93	4003	BRL	937,47	DESEMBARACO
Key Ltda	2590132618	CEK	PAI	8/23/2019 11:50	8/23/2019 11:53	BB	BB	0	4003	BRL	905,71	DESEMBARACO
Key Ltda	2590132618	ENF	RLS	8/27/2019 9:49	8/27/2019 16:40	EX	BB	6	4003	BRL	905,71	DESEMBARACO
Key Ltda	2590132618	RLS	TOD	8/27/2019 16:40	8/28/2019 9:30	EX	BB	16	4003	BRL	905,71	DESEMBARACO
Key Ltda	2590132618	TOD	ECR	8/28/2019 9:30	8/28/2019 9:42	EX	EX	0	4003	BRL	905,71	DESEMBARACO



## APÊNDICE G – Estatística descritiva das sequências de eventos

Este apêndice apresenta a estatística descritiva de todas as 113 transições encontradas nos dados.

A ordem da tabela está em ordem alfabética das sequências.

<b>Sequência</b>	<b>Média</b>	<b>Mediana</b>	<b>Desvio padrão</b>	<b>%</b>
ADD COB	7,63	3,00	10,52	0,11
ADD DXR	40,10	11,00	293,85	0,38
ADD ECR	139,31	59,00	202,33	0,01
ADD ENF	200,87	120,00	249,13	0,27
ADD GAT	22,15	6,00	50,86	8,10
ADD LOG	545,53	10,00	7.468,32	0,17
ADD PAI	54,35	23,00	119,99	0,97
ADD PAR	69,79	32,00	99,56	0,59
COB ADD	35,87	10,00	106,65	6,52
COB COB	121,44	23,00	206,21	0,01
COB DXR	118,98	52,00	162,79	0,77
COB ENF	100,51	71,00	114,56	0,03
COB GAT	26,43	11,00	82,63	0,77
COB LOG	60,91	43,00	68,37	0,82
COB PAI	155,81	81,00	183,22	0,73
COB PAR	124,01	55,00	176,76	0,15
DLV DLV	267,02	88,00	632,13	0,51
DXR ADD	120,70	41,00	185,03	0,95
DXR COB	51,68	34,00	78,72	5,57
DXR ECR	124,89	38,50	262,91	0,01
DXR ENF	129,15	30,00	248,84	0,13
DXR GAT	47,20	15,00	101,53	0,29
DXR LOG	67,15	15,00	377,36	1,15
DXR PAI	58,47	21,00	124,55	2,40
DXR PAR	168,69	65,00	321,17	0,11
DXR RLS	49,76	0,00	173,71	0,02
ECR DLV	422,28	144,00	1.008,52	0,44

<b>Sequência</b>	<b>Média</b>	<b>Mediana</b>	<b>Desvio padrão</b>	<b>%</b>
ECR DXR	471,88	0,00	2.703,52	0,02
ECR ENF	14,61	2,00	125,14	0,12
ECR GAT	1,54	2,00	1,98	0,01
ECR PAI	284,23	102,50	426,03	0,02
ECR PAR	21,10	5,00	37,99	0,10
ECR RLS	13,74	0,00	48,77	8,33
ECR TNF	73,02	23,00	231,50	0,42
EFD DLV	290,50	158,00	311,29	0,01
EFD PAR	58,96	6,00	129,00	0,02
EFD RLS	195,73	97,00	344,15	0,58
EFD TNF	292,28	171,00	392,58	0,01
ENF DLV	148,27	128,50	125,38	0,01
ENF ECR	16,63	16,00	26,63	8,24
ENF EFD	18,14	13,00	30,26	0,60
ENF ENF	4,19	0,00	14,18	0,01
ENF EXR	15,85	7,50	26,20	0,39
ENF PAI	8,94	2,00	19,71	0,02
ENF PAR	2,19	1,00	6,34	0,29
ENF RLS	18,18	5,00	49,27	0,96
ENF TNF	9,77	3,00	25,71	0,12
EXR PAR	80,96	8,00	188,43	0,02
EXR RLS	309,66	170,00	506,72	0,38
EXR TNF	254,54	164,00	262,95	0,01
GAT ADD	2,18	0,00	25,08	0,75
GAT COB	15,49	10,00	23,14	0,07
GAT DXR	110,46	17,00	534,73	0,29
GAT ECR	82,48	60,00	104,68	0,07
GAT ENF	111,68	65,00	179,20	4,80
GAT LOG	259,90	20,00	2.829,00	0,29
GAT PAI	96,78	49,00	170,94	2,33
GAT PAR	120,55	74,00	178,96	1,89
GAT RLS	98,81	70,50	80,57	0,02

<b>Sequência</b>	<b>Média</b>	<b>Mediana</b>	<b>Desvio padrão</b>	<b>%</b>
LOG ADD	81,86	37,00	123,88	0,39
LOG COB	49,22	30,00	63,12	1,68
LOG DXR	9,70	0,00	41,04	7,64
LOG ENF	140,64	30,00	274,87	0,08
LOG GAT	30,71	10,50	81,56	0,07
LOG PAI	35,04	4,00	100,82	0,73
LOG PAR	64,28	29,50	97,00	0,05
PAI ADD	109,65	50,00	147,92	1,08
PAI COB	33,10	18,00	80,75	1,12
PAI DLV	314,80	50,00	684,62	0,02
PAI DXR	69,53	3,00	277,15	0,07
PAI ECR	55,64	4,50	150,76	0,02
PAI ENF	58,96	8,00	146,26	0,78
PAI GAT	30,63	18,00	47,85	0,69
PAI LOG	115,68	43,00	337,74	0,44
PAI PAI	20,58	0,00	64,89	0,05
PAI PAR	122,35	72,00	164,39	3,64
PAI RLS	61,26	18,00	82,09	0,01
PAI TNF	99,41	27,00	221,07	0,01
PAR ADD	91,10	50,00	112,96	0,80
PAR COB	232,47	15,00	4.330,61	0,25
PAR DLV	141,66	87,00	334,17	0,23
PAR DXR	73,74	24,00	124,84	0,02
PAR ECR	30,72	17,00	57,84	0,30
PAR EFD	50,47	18,00	143,14	0,02
PAR ENF	91,93	42,00	167,99	4,38
PAR EXR	65,80	19,50	121,06	0,01
PAR GAT	69,81	39,00	111,12	0,59
PAR LOG	171,11	47,00	1.086,54	0,13
PAR PAI	356,67	64,50	1.084,23	0,06
PAR PAR	378,71	204,00	472,73	0,01
PAR RLS	149,79	23,00	343,88	0,12
PAR TNF	70,45	47,00	82,52	0,11

<b>Sequência</b>	<b>Média</b>	<b>Mediana</b>	<b>Desvio padrão</b>	<b>%</b>
RLS COB	839,17	102,00	1.168,89	0,01
RLS DLV	259,71	117,00	649,56	4,91
RLS ECR	7,28	0,00	193,90	0,83
RLS EFD	2,58	2,00	3,48	0,01
RLS ENF	62,22	2,50	415,00	0,03
RLS PAI	424,33	75,50	1.086,17	0,04
RLS PAR	44,64	7,00	134,42	0,29
RLS TNF	99,69	38,00	344,55	3,04
TNF TNF	219,95	0,00	684,52	2,00

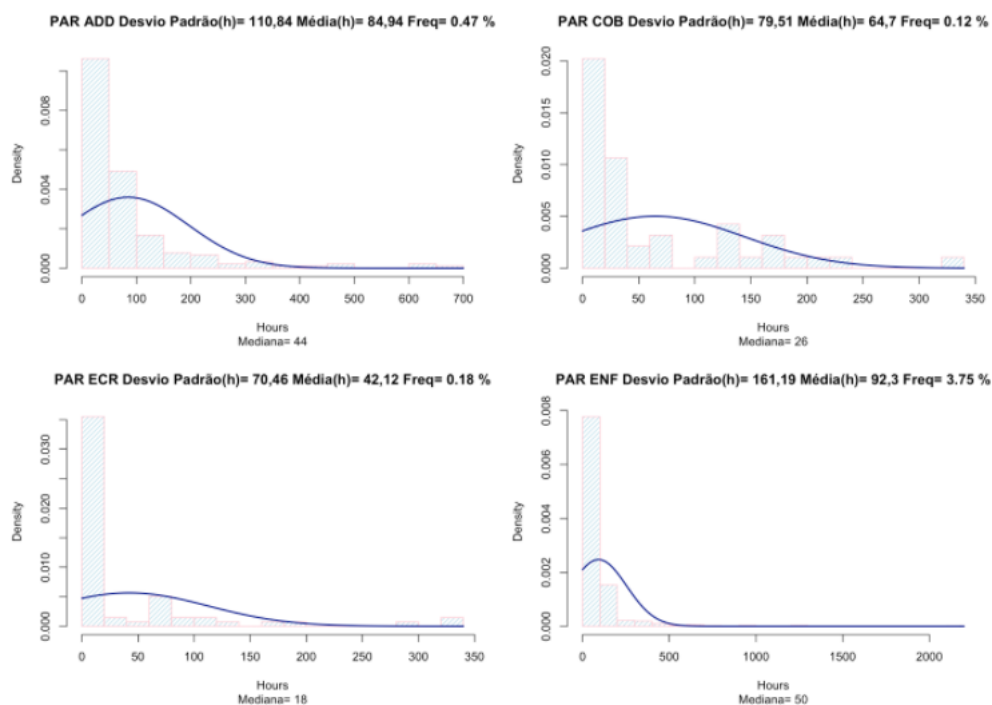
## APÊNDICE H – Análise da curva da transição de estados

Este Apêndice apresenta a distribuição da curva do tempo em horas entre os dois eventos para quatro (4) sequências diferentes. O primeiro gráfico apresenta a sequência PAR ADD com desvio padrão de 110,84 horas e média de 84,94 horas. O segundo gráfico apresenta a sequência PAR COB com desvio padrão de 79,51 horas; média de 64,70 horas. O terceiro gráfico apresenta a sequência PAR ECR com desvio padrão de 70,46 horas; média de 42,12 horas. E o quarto gráfico apresenta a sequência PAR ENF com desvio padrão de 161,19 horas; média de 92,30 horas.

Estas diferentes distribuições mostram como cada transição tem dissimilaridades, portanto, influenciam a eficiência do processo de formas diferentes.

O código para gerar estes gráficos foi criado na linguagem de programação (script interpretado) do programa R, em sua versão 4.0.2. R é um software criado em ambiente livre para computação e programação estatística e geração de saídas gráficas, bastante extenso em funcionalidade e amplamente utilizado no meio acadêmico (R Development Core Team, 2020).

As extensões (*packages*) que foram utilizadas nesta implementação foram: MARKOVCHAIN 0.8.5-2 (Spedicato, 2017), READXL 1.3.1 (Wickham e Bryan, 2019), GGPLOT2 3.3.2 (Wickham, 2016), GGTHEMES 4.2.0 (Arnold, 2019), DPLYR 1.0.2 (Wickham et al., 2020), IGRAPH 1.2.6 (Csardi e Nepusz, 2006) e GA 3.2 (Scrucca, 2013).



## APÊNDICE I – Cadeias de Markov de alguns outros clientes.

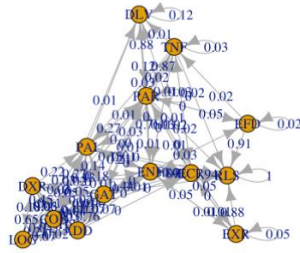
O Apêndice I apresenta exemplos de nove (9) diferentes grafos, cada um deles descrevendo o desenho do processo desses diferentes clientes.

O código para gerar estes gráficos foi criado na linguagem de programação (script interpretado) do programa R, em sua versão 4.0.2. R é um software criado em ambiente livre para computação e programação estatística e geração de saídas gráficas, bastante extenso em funcionalidade e amplamente utilizado no meio acadêmico (R Development Core Team, 2020).

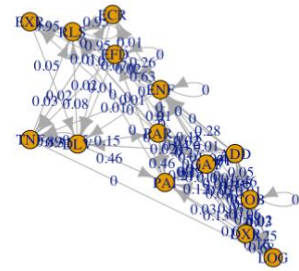
As extensões (*packages*) que foram utilizadas nesta implementação foram: MARKOVCHAIN 0.8.5-2 (Spedicato, 2017), READXL 1.3.1 (Wickham e Bryan, 2019), GGPLOT2 3.3.2 (Wickham, 2016), GGTHEMES 4.2.0 (Arnold, 2019), DPLYR 1.0.2 (Wickham et al., 2020), IGRAPH 1.2.6 (Csardi e Nepusz, 2006) e GA 3.2 (Scrucca, 2013).



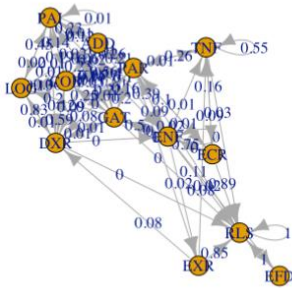
ADT SE  
 Receita média R\$ 91,70  
 Desvio padrão: 269,99 horas



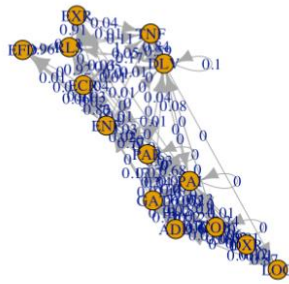
BECTON  
 Receita média R\$ 126,91  
 Desvio padrão 1157,86 horas



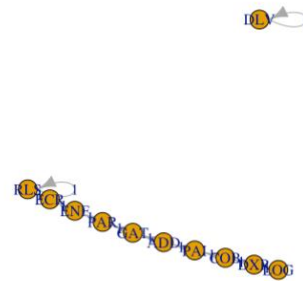
CISCO  
 Receita média R\$ 61,32  
 Desvio padrão 944,31 horas



DEVANL  
 Receita média R\$ 306,78  
 Desvio padrão: 269,99 horas



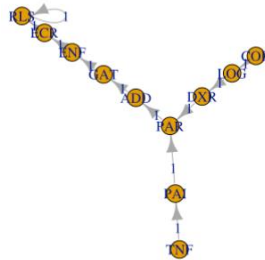
FLUKE  
 Receita média R\$ 203,27  
 Desvio padrão: 166,89 horas



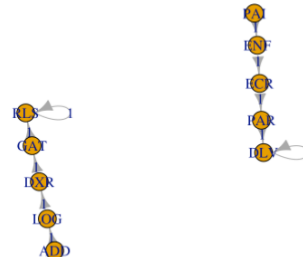
GOLD F  
 Receita média R\$ 217,81  
 Desvio padrão: 70,89 horas



ORACLE  
 Receita média R\$ 136,13  
 Desvio padrão: 159,06 horas



SOMA S  
 Receita média R\$ 140,00  
 Desvio padrão: 118,61 horas



USPHARMA  
 Receita média R\$ 210,00  
 Desvio padrão: 194,05 horas

**APÊNDICE J** – Tabela de clientes rebalanceada.

O Apêndice J apresenta os valores de distribuição de frequência do volume de trabalho do departamento entre os clientes que resultem em maior ganho.

<b>Clientes</b>	<b>Frequência Ótima</b>	<b>Frequência Atual</b>	<b>Delta</b>
BECTON	17,64%	19,92%	-2,28%
BELIMO	5,24%	1,68%	3,56%
CARL Z	3,86%	0,67%	3,19%
CISCO	0,12%	1,03%	-0,91%
DEVANL	3,28%	2,97%	0,31%
EMPRES	5,76%	2,71%	3,05%
FLEXIT	0,35%	0,60%	-0,25%
FLSMID	1,58%	4,24%	-2,66%
FLUKE	4,81%	9,62%	-4,81%
GOOGLE	4,80%	9,72%	-4,92%
HID DO	3,66%	0,34%	3,32%
KOHLER	1,54%	4,96%	-3,42%
LAUREN	0,30%	0,48%	-0,18%
MOLNLY	4,76%	0,31%	4,45%
MTSS	1,51%	2,47%	-0,96%
OKI DA	3,96%	0,62%	3,34%
ONITY	1,62%	1,53	0,09%
ORACLE	22,17%	27,24%	-5,07%
PANDOR	3,94%	0,34%	3,60%
SENSOR	5,38%	0,45%	4,93%
UNIPAR	1,69%	1,65	0,04%
VICTAU	1,57%	0,41%	1,16%
WEST P	0,49%	6,06%	-5,57%

## APÊNDICE K – Algoritmos aplicados no trabalho.

O Apêndice K deste trabalho reproduz o código desenvolvido para a implementação dos algoritmos aplicados no trabalho.

O código foi criado na linguagem de programação (script interpretado) do programa R, em sua versão 4.0.2. R é um software criado em ambiente livre para computação e programação estatística e geração de saídas gráficas, bastante extenso em funcionalidade e amplamente utilizado no meio acadêmico (R Development Core Team, 2020).

As extensões (*packages*) que foram utilizadas nesta implementação foram: MARKOVCHAIN 0.8.5-2 (Spedicato, 2017), READXL 1.3.1 (Wickham e Bryan, 2019), GGPLOT2 3.3.2 (Wickham, 2016), GGTHEMES 4.2.0 (Arnold, 2019), DPLYR 1.0.2 (Wickham et al., 2020), IGRAPH 1.2.6 (Csardi e Nepusz, 2006) e GA 3.2 (Scrucca, 2013).

```
#####
#####
## ===== ##
## == Multiple techniques: Markov Chain, Genetic Algorithms and Simulations == ##
## ===== ##
## ##
## Carolina Mourao de Paula - FGV EAESP ##
## ##
## Code in R (www.r-project.com) - Applied in R 4.0.1 ##
## Dissertation: ##
##     OTIMIZAÇÃO DO PORTFÓLIO DE CLIENTES: UMA NOVA PROPOSIÇÃO ##
##     ATRAVÉS DE DESIGN DE PROCESSOS BASEADO EM DADOS ##
## First version: 17 February 2021 ##
## ##
#####
#####

library(knitr)
library(kableExtra)
library(markovchain)
library(readxl)
library(ggplot2)
library(ggthemes)
library(dplyr)
library(GA)
library(igraph)
library(parallel)
library(doParallel)

#Initial variables
convergenceStates <- c("DLV", "TNF")
sequencesAllowed <- c("LOG DXR", "DXR COB", "COB ADD", "ADD GAT", "GAT PAI", "PAI
PAR", "PAR ENF", "ENF ECR", "ENF EFD", "ENF EXR", "ECR RLS", "EFD RLS", "EXR RLS", "RLS
TNF", "RLS DLV")

#Read file
```

```

Data_Source <- read_excel("~/Documents/TA FGV
2020/Data/Data_Source_2013_Jan2020_SAO v big data CORE EVENTS client coded brl
honorarios.xlsx")
Data_Source <- Data_Source[complete.cases(Data_Source),]

rows <- levels(factor(Data_Source$C0))
cols <- levels(factor(Data_Source$C1))
rows <- unique(c(rows,cols))
cols <- rows
tmValues <- c()
for(i in 1:length(rows)) {
  row <- rows[i]
  rowData <- Data_Source$C1[Data_Source$C0==row]
  row_total <- length(rowData)
  if(row_total ==0) {
    tmValues <- c(tmValues, rep.int(0, length(rows)-1), 1)
  }
  for(j in 1:length(cols)) {
    col <- cols[j]
    if(row %in% convergenceStates) {
      if(col == row) {
        tmValues <- c(tmValues, 1)
      } else {
        tmValues <- c(tmValues, 0)
      }
    } else {
      col_total <- ifelse(is.na(sum(rowData==col)), 0, sum(rowData==col))
      tmValues <- c(tmValues, col_total / row_total)
    }
  }
}
tm <- matrix(tmValues, nrow = length(rows), ncol = length(cols), byrow = TRUE )
dimnames(tm) <- list(rows, cols)
rnames <- rownames(tm)

#Creating a DTMC.
dtmc <- new("markovchain",transitionMatrix=tm,
           name="MarkovChain")

#Plot Markov
plot(dtmc)

#Descriptive Statistics
seq <- paste(Data_Source$C0, Data_Source$C1)
Data_Source$Seq <- seq
factors <- factors <- levels(factor(Data_Source$Seq))
Data_Source_Summary <- data.frame()
for(i in 1:length(factors)) {
  x<- Data_Source$HOURS_DIFF[Data_Source$Seq==factors[i]]
  m<-ifelse(is.na(mean(x)),0,round(mean(x),2))
  md<-ifelse(is.na(median(x)),0,round(median(x),2))
  std<-ifelse(is.na(sd(x)),0,round(sd(x),2))
  c <- round(length(x) / length(seq) * 100, 2)
  percent_representation <- rep(c, length(x))
  Data_Source$Percent_Representation[Data_Source$Seq==factors[i]] <-
percent_representation

  if(c >0) {
    Data_Source_Summary <- rbind(Data_Source_Summary,
data.frame(cbind(Seq=factors[i], Mean=m, Median=md, Sd=std, Rep=c)))
  }
}

# Overall values

```

```

avg_hour_diff <- round(mean(Data_Source$HOURS_DIFF), 2)
md_hour_diff <- round(median(Data_Source$HOURS_DIFF), 2)
sd_hour_diff <- round(sd(Data_Source$HOURS_DIFF), 2)

# Build transition matrix by customer
tot_files_by_client <- nrow(Data_Source %>% group_by(CONSIGNEE, INVOICE_NO) %>%
summarise(total = n()))
clients <- levels(factor(Data_Source$CONSIGNEE))
cost_revenue_df <- data.frame()
process_conformity_df <- data.frame()
for(a in 1:length(clients)) {
  escore <- 0
  client <- clients[a]
  cDataSource <- Data_Source[Data_Source$CONSIGNEE==client,]
  tot_files <- nrow(cDataSource %>% group_by(INVOICE_NO) %>% summarise(total =
n()))
  representativity <- round((tot_files/tot_files_by_client) * 100,2)
  cDataSource %>% group_by(INVOICE_NO) %>% summarise(avg=mean(AMOUNT)) %>%
select(avg) -> dfRev
  avg_rev <- round(mean(cDataSource$AMOUNT), 2)
  sd_rev <- round(sd(cDataSource$AMOUNT), 2)
  avg_hourdiff<- round(mean(cDataSource$HOURS_DIFF), 2)
  sd_hourdiff <- round(sd(cDataSource$HOURS_DIFF), 2)
  rows <- levels(factor(cDataSource$C0))
  cols <- levels(factor(cDataSource$C1))
  allcombinatios <- unique(c(rows, cols))
  tmValues <- c()
  for(i in 1:length(allcombinatios)) {
    row <- allcombinatios[i]
    rowData <- cDataSource$C1[cDataSource$C0==row]
    row_total <- length(rowData)
    for(j in 1:length(allcombinatios)) {
      col <- allcombinatios[j]
      if(row %in% convergenceStates) {
        if(col == row) {
          tmValues <- c(tmValues, 1)
        } else {
          tmValues <- c(tmValues, 0)
        }
      } else {
        col_total <- ifelse(is.na(sum(rowData==col)), 0, sum(rowData==col))
        seq <- paste(row,col)
        avg_hourdiff_seq <- round(mean(cDataSource$HOURS_DIFF[cDataSource$C0==row &
cDataSource$C1==col]), 2)
        sd_hourdiff_seq <- round(sd(cDataSource$HOURS_DIFF[cDataSource$C0==row &
cDataSource$C1==col]), 2)
        seq_count <- length(cDataSource$HOURS_DIFF[cDataSource$C0==row &
cDataSource$C1==col])
        if (seq_count > 0) {
          escore <- escore + ifelse(sum(sequencesAllowed==seq)==0, seq_count, 0)
          conformity_desc <- ifelse(sum(sequencesAllowed==seq)==0, "Non-
conformance", "Conformance")
          process_conformity_df <- rbind(process_conformity_df,
data.frame(cbind(Client=client, Seq=seq, SeqCount=seq_count, AvgHour=avg_hourdiff_seq,
SdHour=sd_hourdiff_seq, ProcessEscore=escore, ConformanceDesc=conformity_desc)))
        }
        value <- ifelse(row_total==0, 0, col_total / row_total)
        tmValues <- c(tmValues, value)
      }
    }
  }
}
tmValues

```

```

tm <- matrix(tmValues, nrow = length(allcombinatios), ncol =
length(allcombinatios), byrow = TRUE )
dimnames(tm) <- list(allcombinatios, allcombinatios)

#Adjust 0 values in TM
row.sums <- (apply(tm,1,sum))
zerodRows <- names(row.sums[row.sums > 1.05 | row.sums < 0.99])
for(i in 1:length(zerodRows)) {
  row <- zerodRows[i]
  tm[rownames(tm) == row, colnames(tm) == row] <- 1
}

if(dim(tm)[1] == dim(tm)[2]) {
  #Creating a DTMC.
  write.csv(tm, paste(client, 'tm_core.csv'))
  dtmc <- new("markovchain",transitionMatrix=tm,
             name="MarkovChain para o Depto")

  #Plot o Markov
  title <- paste(client, "; Lead time (h) = ",format(avg_hour_diff,
decimal.mark=","), "; Desvio Padrão (h) = ",format(sd hour_diff, decimal.mark=","))
  subTitle <- paste("Rep =",representativity," Receita Média (BRL) =
",format(avg_rev, decimal.mark=","))
  plot(dtmc, main=title, sub = subTitle, cex.axis=0.75 )

  p <- ggplot(process_conformity_df[process_conformity_df$Client == client &
!is.na(process_conformity_df$SdHour),],aes(x=Seq,y=SdHour,fill=factor(ConformanceDe
sc)))+
  geom_bar(stat="identity")+
  ggtitle(paste(client,"Process Conformity")) +
  scale_fill_manual(values = c("#4E84C4", "salmon"),name="",
                    labels=c("Conforme", "Não conforme"))+
  xlab("Transições")+ylab("Desvio Padrão (h)") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size = 9,
family="Times New Roman", color="grey20"),
        text = element_text(family="Times New Roman",color="grey20"),
        legend.position = "bottom")

  print(p)
}

escore <- round(escore / nrow(cDataSource),2)
cost_revenue_df <- rbind(cost_revenue_df,
data.frame(cbind(Client=client,AvgRev=avg_rev,SdHour=sd_hourdiff,AvgHour=avg_hourdi
ff,Rep=representativity,ProcessEscore=escore,SdRev=sd_rev)))

}

cost_revenue_df$AvgRevN <- as.numeric(cost_revenue_df$AvgRev)
cost_revenue_df$RepN <- as.numeric(cost_revenue_df$Rep)
cost_revenue_df$SdHourN <- as.numeric(cost_revenue_df$SdHour)
cost_revenue_df$AvgHourN <- as.numeric(cost_revenue_df$AvgHour)
cost_revenue_df$ProcessEscoreN <- as.numeric(cost_revenue_df$ProcessEscore)
cost_revenue_df$SdRevN <- as.numeric(cost_revenue_df$SdRev)

#Plot the current portfolio: desv padrao vs receita
qtyInChart <- nrow(cost_revenue_df)
loops <- ceiling(nrow(cost_revenue_df) / qtyInChart)
first <- 1
len <- qtyInChart
end <- first + len
portfolio <-list()
index <- 1
for(i in 1:loops) {

```

```

p<- ggplot(data = cost_revenue_df[first:end,], aes(x = SdHourN, y = AvgRevN)) +
  geom_point(data = cost_revenue_df[first:end,],size=1) +
  geom_text(data = cost_revenue_df[first:end,],aes(label=Client),vjust=-
1,nudge_y=1.5,check_overlap=TRUE,size=2,family = "Times New Roman") +
  ggtitle("Desvio Padrão h X Receita Média") +
  labs(x = "Desvio Padrão (horas)", y = "Receita (BRL)") +
  theme(axis.text = element_text(size = 9,family="Times New
Roman",color="grey20"),
        text = element_text(family="Times New Roman",color="grey20"))

portfolio[[index]] <- p

index <- index + 1
first <- first + len
len <- ifelse(nrow(cost_revenue_df)-first > qtyInChart, qtyInChart,
nrow(cost_revenue_df)-first)
end <- first + len
}

for(i in 1:length(portfolio)) {

  print(portfolio[[i]])

}

#Plot the current portfolio: process score vs receita
qtyInChart <- nrow(cost_revenue_df)
loops <- ceiling(nrow(cost_revenue_df) / qtyInChart)
first <- 1
len <- qtyInChart
end <- first + len
portfolio <-list()
index <- 1
for(i in 1:loops) {

  p<- ggplot(data = cost_revenue_df[first:end,], aes(x = ProcessScoreN, y =
AvgRevN)) +
    geom_point(data = cost_revenue_df[first:end,],size=1) +
    geom_text(data = cost_revenue_df[first:end,],aes(label=Client),vjust=-
1,nudge_y=1.5,check_overlap=TRUE,size=2,family = "Times New Roman") +
    ggtitle("Process Score X Receita Média") +
    labs(x = "Process Score", y = "Receita (BRL)") +
    theme(axis.text = element_text(size = 9,family="Times New
Roman",color="grey20"),
          text = element_text(family="Times New Roman",color="grey20"))

  portfolio[[index]] <- p

  index <- index + 1
  first <- first + len
  len <- ifelse(nrow(cost_revenue_df)-first > qtyInChart, qtyInChart,
nrow(cost_revenue_df)-first)
  end <- first + len
}

for(i in 1:length(portfolio)) {

  print(portfolio[[i]])

}

#Plot the current portfolio: lead time vs receita
qtyInChart <- nrow(cost_revenue_df)

```

```

loops <- ceiling(nrow(cost_revenue_df) / qtyInChart)
first <- 1
len <- qtyInChart
end <- first + len
portfolio <- list()
index <- 1
for(i in 1:loops) {

  p<- ggplot(data = cost_revenue_df[first:end,], aes(x = AvgHourN, y = AvgRevN)) +
    geom_point(data = cost_revenue_df[first:end,],size=1) +
    geom_text(data = cost_revenue_df[first:end,],aes(label=Client),vjust=-
1,nudge_y=1.5,check_overlap=TRUE,size=2,family = "Times New Roman") +
  ggtitle("Lead Time (h) X Receita Média") +
  labs(x = "Lead Time (h)", y = "Receita (BRL)") +
  theme(axis.text = element_text(size = 9,family="Times New
Roman",color="grey20"),
        text = element_text(family="Times New Roman",color="grey20"))

  portfolio[[index]] <- p

  index <- index + 1
  first <- first + len
  len <- ifelse(nrow(cost_revenue_df)-first > qtyInChart, qtyInChart,
nrow(cost_revenue_df)-first)
  end <- first + len
}

for(i in 1:length(portfolio)) {

  print(portfolio[[i]])

}

# Simulating random portfolios
cost_revenue_df.transp <- t(cost_revenue_df)
sims <- matrix(0,ncol = 7 + ncol(cost_revenue_df.transp))
dimnames(sims) <- list(NULL,
c("sim.portfolio",clients,"SdLead","AvgRev","SdRev","AvgLead","Score","Function"))
sim.portfolio <- 0
for (i in 1:1000) {
  sim.portfolio <- sim.portfolio + 1
  w <- c(runif(ncol(cost_revenue_df.transp)))
  if (sum(w) == 0) { w <- w + 1e-2 }
  w <- w / sum(w)
  SdLead <- w %*%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='SdHourN',])
  AvgRev <- w %*%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='AvgRevN',])
  SdRev <- w %*%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='SdRevN',])
  AvgLead <- w %*%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='AvgHourN',])
  Score <- w %*%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='ProcessScoreN
',])

  obj <- (Score + (1 / AvgRev) + SdLead + SdRev + AvgLead)

  w <- matrix(w,nrow = 1)
  newline <- cbind(sim.portfolio, w, SdLead, AvgRev, SdRev, AvgLead, Score, obj)
  sims <- rbind(sims, newline)
}
sims <- sims[-1,]
sims.portfolio <- sims[,1]
sims.metrics <- sims[, (ncol(sims)-5):ncol(sims)]

```

```

risk_returns.sims <- cbind(sims.portfolio, sims.metrics)
sim.cond <- data.frame(matrix(0, nrow = nrow(sims)))
risk_returns.sims_df <- data.frame(risk_returns.sims)
risk_returns.sims_df <- cbind(risk_returns.sims_df, sim.cond)

ggplot(data = risk_returns.sims_df, aes(x = SdLead, y = AvgRev)) +
  geom_point(data = risk_returns.sims_df, size=1) +
  geom_text(data = risk_returns.sims_df, aes(label=sims.portfolio), vjust=-
1, nudge_y=1.5, check_overlap=TRUE, size=2, family="Times New Roman") +
  ggtitle("Desv. Padrão X Receita em simulações") +
  labs(x = "Desv. Padrão (h)", y = "Receita (BRL)") +
  theme(axis.text = element_text(size = 9, family="Times New
Roman", color="grey20"),
        text = element_text(family="Times New Roman", color="grey20"))

ggplot(data = risk_returns.sims_df, aes(x = Function, y = AvgRev)) +
  geom_point(data = risk_returns.sims_df, size=1) +
  geom_text(data = risk_returns.sims_df, aes(label=sims.portfolio), vjust=-
1, nudge_y=1.5, check_overlap=TRUE, size=2, family="Times New Roman") +
  ggtitle("Objetivo geral X Receita em simulações") +
  labs(x = "Objetivo geral", y = "Receita (BRL)") +
  theme(axis.text = element_text(size = 9, family="Times New
Roman", color="grey20"),
        text = element_text(family="Times New Roman", color="grey20"))
rnames <- rownames(risk_returns.sims_df)
risk_returns.sims_red_df <- risk_returns.sims_df[seq(from=1,
to=nrow(risk_returns.sims_df), by=20),]
ggplot(data = risk_returns.sims_red_df, aes(x=rownames(risk_returns.sims_red_df),
y=Function, group=1)) +
  geom_line(linetype = "dashed", size=1, color="#4E84C4")+

  theme(plot.title = element_text(size=16, lineheight=2, face="bold"),
        legend.text=element_text(size=14),
        legend.title=element_text(size=14, face="bold"),
        axis.text.y = element_text(size=12),
        axis.text.x = element_text(size=8, angle=45),
        strip.background = element_rect(fill = "White"),
        text = element_text(family="Times New Roman", color="grey20")) +
  labs(x = "Simulações", y = "Objetivo geral")

# Genetic Algorithm Simulations
# Read file
Data_Source <- read_excel("~/Documents/TA FGV
2020/Data/Data_Source_2013_Jan2020_SAO v big data CORE EVENTS client coded brl
honorarios 2018b.xlsx")
Data_Source <- Data_Source[complete.cases(Data_Source),]

tot_files_by_client <- nrow(Data_Source %>% group_by(CONSIGNEE, INVOICE_NO) %>%
summarise(total = n()))
clients <- levels(factor(Data_Source$CONSIGNEE))
cost_revenue_df <- data.frame()
process_conformity_df <- data.frame()
for(a in 1:length(clients)) {
  escore <- 0
  client <- clients[a]
  cDataSource <- Data_Source[Data_Source$CONSIGNEE==client,]
  tot_files <- nrow(cDataSource %>% group_by(INVOICE_NO) %>% summarise(total =
n()))
  representativity <- round((tot_files/tot_files_by_client) * 100, 2)
  cDataSource %>% group_by(INVOICE_NO) %>% summarise(avg=mean(AMOUNT)) %>%
select(avg) -> dfRev
  avg_rev <- round(mean(dfRev$avg), 2)
  sd_rev <- round(sd(cDataSource$AMOUNT))
  avg_hourdiff <- round(mean(cDataSource$HOURS_DIFF), 2)
  sd_hourdiff <- round(sd(cDataSource$HOURS_DIFF), 2)
}

```

```

rows <- levels(factor(cDataSource$C0))
cols <- levels(factor(cDataSource$C1))
allcombinatios <- unique(c(rows, cols))
tmValues <- c()
for(i in 1:length(allcombinatios)) {
  row <- allcombinatios[i]
  rowData <- cDataSource$C1[cDataSource$C0==row]
  row_total <- length(rowData)
  for(j in 1:length(allcombinatios)) {
    col <- allcombinatios[j]
    if(row %in% convergenceStates) {
      if(col == row) {
        tmValues <- c(tmValues, 1)
      } else {
        tmValues <- c(tmValues, 0)
      }
    } else {
      col_total <- ifelse(is.na(sum(rowData==col)), 0, sum(rowData==col))
      seq <- paste(row,col)
      avg_hourdiff_seq <- round(mean(cDataSource$HOURS_DIFF[cDataSource$C0==row &
cDataSource$C1==col]), 2)
      sd_hourdiff_seq <- round(sd(cDataSource$HOURS_DIFF[cDataSource$C0==row &
cDataSource$C1==col]), 2)
      seq_count <- length(cDataSource$HOURS_DIFF[cDataSource$C0==row &
cDataSource$C1==col])
      if (seq_count > 0) {
        escore <- escore + ifelse(sum(sequencesAllowed==seq)==0, seq_count, 0)
        conformity_desc <- ifelse(sum(sequencesAllowed==seq)==0, "Non-
conformance", "Conformance")
        process_conformity_df <- rbind(process_conformity_df,
data.frame(cbind(Client=client, Seq=seq, SeqCount=seq_count, AvgHour=avg_hourdiff_seq,
SdHour=sd_hourdiff_seq, ProcessEscore=escore, ConformanceDesc=conformity_desc)))
      }
      value <- ifelse(row_total==0, 0, col_total / row_total)
      tmValues <- c(tmValues, value)
    }
  }
}
tmValues
tm <- matrix(tmValues, nrow = length(allcombinatios), ncol =
length(allcombinatios), byrow = TRUE )
dimnames(tm) <- list(allcombinatios, allcombinatios)

#Adjust 0 values in TM
row.sums <- (apply(tm,1,sum))
zerodRows <- names(row.sums[row.sums > 1.05 | row.sums < 0.99])
for(i in 1:length(zerodRows)) {
  row <- zerodRows[i]
  tm[rownames(tm) == row, colnames(tm) == row] <- 1
}

escore <- round(escore / nrow(cDataSource), 2)

cost_revenue_df <- rbind(cost_revenue_df,
data.frame(cbind(Client=client, AvgRev=avg_rev, SdHour=sd_hourdiff, AvgHour=avg_hourdi
ff, Rep=representativity, ProcessEscore=escore, SdRev=sd_rev)))
}

cost_revenue_df$AvgRevN <- as.numeric(cost_revenue_df$AvgRev)
cost_revenue_df$RepN <- as.numeric(cost_revenue_df$Rep)
cost_revenue_df$SdHourN <- as.numeric(cost_revenue_df$SdHour)
cost_revenue_df$SdRevN <- as.numeric(cost_revenue_df$SdRev)
cost_revenue_df$AvgHourN <- as.numeric(cost_revenue_df$AvgHour)
cost_revenue_df$Score <- as.numeric(cost_revenue_df$ProcessEscore)

```

```

cost_revenue_df =
cost_revenue_df[c("Client", "AvgRevN", "RepN", "SdHourN", "SdRevN", "AvgHourN", "Score")]

write.csv(cost_revenue_df, 'cost_revenue_df.csv')
cost_revenue_df.transp <- t(cost_revenue_df)

popsize_ <- 10
pcrossover_ <- 0.75
pmutation_ <- 0.05

constraint <- function(x) {
  boundary_constr = (sum(x)-1)**2 # "sum x = 1" constraint
  y <-
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='RepN',])
  for (i in 1:length(x)) {
    deltaRep <- abs((x[i]*100) - y[i])
    boundary_constr = boundary_constr +
      max(c(0, (deltaRep-5))**2) # "delta <= 5" constraint
  }
  return (boundary_constr)
}

function_obj <- function(x) {
  fn.call <- fn.call + 1
  if (sum(x) == 0) { x <- x + 1e-2 }
  x <- x / sum(x)
  SdLead <-
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='SdHourN',])
  AvgRev <-
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='AvgRevN',])
  SdRev <-
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='SdRevN',])
  AvgLead <-
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='AvgHourN',])
  Score <-
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='Score',])
  obj <- ((x %>% Score) + (x %>% (1 / AvgRev)) + (x %>% SdLead) + (x %>% SdRev) +
(x %>% AvgLead)) + (100*constraint(x))

  return(obj)
}

set.seed(1234)
fn.call <- 0
initialTime <- Sys.time()
# GA function performs maximization so we must multiply the objective function by -
1
fitness <- function(x) {- function_obj(x)}
GA <- ga(type = "real-valued", fitness = fitness, lower = rep(0,
ncol(cost_revenue_df.transp)), upper = rep(1, ncol(cost_revenue_df.transp)),
popSize = popsize_, pcrossover = pcrossover_, pmutation = pmutation_, maxiter =
10000, run = 50, keepBest = TRUE, parallel = TRUE, monitor = FALSE)
endTime <- Sys.time()
totalTimeSA <- endTime - initialTime
resultado_GA <- summary(GA)
w.optimum_GA <- resultado_GA$solution
w.optimum_GA <- t(w.optimum_GA)
w.optimum_GA <- w.optimum_GA[,1] / sum(w.optimum_GA[,1])
w.optimum_GA_df <- cbind(clients, w.optimum = w.optimum_GA)

# GA results receita media
optimum_return <- w.optimum_GA %>%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='AvgRevN',])
round(optimum_return, 2)

```

```

# Receita media antes
optimum_return <-
(as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='RepN',])/100)
%%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='AvgRevN',])
round(optimum_return, 2)
# GA results desvio padrao
optimum_risk <- w.optimum_GA %%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='SdHourN',])
round(optimum_risk, 2)

# Desvio padrao antes
optimum_risk <-
(as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='RepN',])/100)
%%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='SdHourN',])
round(optimum_risk, 2)

# GA results lead time
optimum_return <- w.optimum_GA %%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='AvgHourN',])
round(optimum_return, 2)

# Lead time antes
optimum_return <-
(as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='RepN',])/100)
%%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='AvgHourN',])
round(optimum_return, 2)

# GA results desvio padrao da receita
optimum_return <- w.optimum_GA %%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='SdRevN',])
round(optimum_return, 2)

# Desvio padrao da receita antes
optimum_return <-
(as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='RepN',])/100)
%% as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='SdRevN',])
round(optimum_return, 2)

# GA results escore de desvios
optimum_return <- w.optimum_GA %%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='Score',])
round(optimum_return, 2)

# Escore de desvios antes
optimum_return <-
(as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='RepN',])/100)
%% as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='Score',])
round(optimum_return, 2)
# Plotar o grafico de evolucao do GA
plot(GA)

# Novo portfolio otimo
opt_portfolio <- data.frame( Client = clients,
AvgRevN = round(w.optimum_GA %%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='AvgRevN',]),0)
,
RepN = round(100 * w.optimum_GA, 2),
SdHourN = round(w.optimum_GA %%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='SdHourN',]),2)
,
SdRevN = round(w.optimum_GA %%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='SdRevN',]),2)
,

```

```

AvgHourN = round(w.optimum_GA %*%
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='AvgHourN',]),2
),
Score =
as.numeric(cost_revenue_df.transp[rownames(cost_revenue_df.transp)=='Score',]),
Status = "Rebalanceado"
)
cost_revenue_df$Status = "Atual"

opt_portfolio <- rbind(opt_portfolio,cost_revenue_df)

p <- ggplot(opt_portfolio, aes(x=Client,y=RepN,fill=factor(Status)))+
geom_bar(stat="identity", position=position_dodge(), width = 0.8)+
geom_text(aes(label = RepN), vjust=-0.2,
position = position_dodge(0.5), size=3)+
ylim(0, max(opt_portfolio$RepN)*1.1)+
ggtitle("Portfolio atual vs rebalanceado") +
scale_fill_manual(values = c("#4E84C4", "salmon"),name="Portfolio",
labels=c("Atual", "Rebalanceado"))+
xlab("Clientes")+ylab("Distribuição(%)") +
theme(axis.text.x = element_text(angle = 90, size = 9, family="Times New Roman",
color="grey20"),
text = element_text(family="Times New Roman",color="grey20"),
legend.position = c(0.9, 0.7))

print(p)

# Apendice, Histograms
seq <- paste(Data_Source$C0, Data_Source$C1)
Data_Source$Seq <- seq
factors <- factors <- levels(factor(Data_Source$Seq))
Data_Source_Summary <- data.frame()
for(i in 1:length(factors)) {
x<- Data_Source$HOURS_DIFF[Data_Source$Seq==factors[i]]
m<-mean(x)
md<-median(x)
std<-sqrt(var(x))
c <- round(length(x) / length(seq) * 100, 2)
percent_representation <- rep(c, length(x))

Data_Source$Percent_Representation[Data_Source$Seq==factors[i]] <-
percent_representation

r <- hist(x, density=20, breaks=20, main=paste(factors[i], "Desvio Padrão(h)=",
format(round(std,2), decimal.mark=","), "Média(h)=", format(round(m,2),
decimal.mark=","), "Freq=", c, "%") , xlab="Hours", col = "lightblue", border =
"pink", prob=TRUE, sub=paste("Mediana=", format(md, decimal.mark=",")))
curve(dnorm(x, mean=m, sd=std),
col="darkblue", lwd=2, add=TRUE, yaxt="n")
}

```