UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO MECÂNICA

LEONARDO ANDRÉ COLARES DANTAS

UTILIZAÇÃO DA ANÁLISE ENVOLTÓRIA DE DADOS NA CONCEPÇÃO DE UM SISTEMA DE APOIO A DECISÃO PARA GESTÃO DE UMA DISTRIBUIDORA DE RECARGAS DE DISPOSITIVOS MÓVEIS

FORTALEZA

2016

LEONARDO ANDRÉ COLARES DANTAS

UTILIZAÇÃO DA ANÁLISE ENVOLTÓRIA DE DADOS NA CONCEPÇÃO DE UM SISTEMA DE APOIO A DECISÃO PARA GESTÃO DE UMA DISTRIBUIDORA DE RECARGAS DE DISPOSITIVOS MÓVEIS

Monografia apresentada ao Curso de Engenharia de Produção Mecânica do Departamento de Engenharia de Produção da Universidade Federal do Ceará, como requisito parcial para a obtenção do título de Engenheiro de Produção Mecânica.

Orientador: Prof. Dr. Heráclito Lopes Jaguaribe Pontes.

FORTALEZA
2016

LEONARDO ANDRÉ COLARES DANTAS

UTILIZAÇÃO DA ANÁLISE ENVOLTÓRIA DE DADOS NA CONCEPÇÃO DE UM SISTEMA DE APOIO A DECISÃO PARA GESTÃO DE UMA DISTRIBUIDORA DE RECARGAS DE DISPOSITIVOS MÓVEIS

Monografia apresentada ao Curso de Engenharia de Produção Mecânica do Departamento de Engenharia de Produção da Universidade Federal do Ceará, como requisito parcial para a obtenção do título de Engenheiro de Produção Mecânica.

Aprovada em ____/_____/_____.

BANCA EXAMINADORA

_____

Profº. Dr. Heráclito Lopes Jaguaribe Pontes (Orientador)

Universidade Federal do Ceará (UFC)

_____

Universidade Federal do Ceará (UFC)

_____

Universidade Federal do Ceará (UFC)

A meus pais, que com todo amor, nunca deixaram de acreditar na minha pessoa.

AGRADECIMENTOS

Primeiramente a minha família, pais e irmão, pelo amor, incentivo e apoio incondicional.

A todos os professores por me proporcionarem o conhecimento não apenas racional, mas a manifestação do caráter e afetividade da educação no processo de formação profissional, por tanto que se dedicaram a mim, não somente por terem me ensinado, mas por terem me feito aprender.

Em especial ao Prof. Rogério Masîh, pela orientação desde os tempos de bolsista do PET, incluindo todo apoio e confiança depositado em mim ao longo da trajetória acadêmica.

Ao meu orientador, Prof. Heráclito Jaguaribe, pelo suporte no pouco tempo que lhe coube, pelas suas correções e incentivos, que diante de um semestre extremamente conturbado conseguiu me inspirar a buscar a importância do projeto final de curso.

A todos os meus amigos, em especial, Adriana Farias Melo, que forneceu suporte não só emocional, mas técnico, trazendo foco e disciplina para finalização de todo projeto.

“Não ganhe o mundo e perca sua alma;
sabedoria é melhor que prata e ouro”.
Bob Marley

RESUMO

A indústria de telecomunicações vem apresentando dificuldades nos últimos anos para conseguir elevar sua lucratividade. Dessa forma, grandes empresas do mercado têm adotado uma forte estratégia de redução de custos, promovendo cortes em diversos segmentos do negócio na tentativa de alavancar sua margem. Assim, partindo-se da necessidade de melhorias de performance e assertividade na alocação dos recursos, foi estudada a possibilidade de implementar um sistema de apoio a decisão com auxílio da metodologia Análise Envoltória de Dados (Data Envelopment Analysis), em uma empresa de distribuição de recargas pré-pagas situada na cidade de Fortaleza. Para isso, deve-se levar em consideração fatores tecnológicos, com o uso de ferramentas consideradas inovadoras para o mercado. Inicialmente foi mapeada as fontes de dados necessárias para o trabalho e foi avaliada a qualidade dos dados recebidos. Após o mapeamento, foi idealizada a arquitetura do projeto validada em um segundo momento por prova de conceito aplicada a uma base- amostra. O desenvolvimento deu-se por criação de processo ETL, realizando tratamentos nos campos principais relacionados aos bairros e às coordenadas geográficas, aplicação de metodologia DEA para priorização de ações de acordo com escala de eficiência e concepção de uma metodologia para classificar os pontos de vendas segundo nível de faturamento. A funcionalidade do sistema criado foi validada pela aplicação prática à base mensal de uma operadora, onde conseguiu-se obter uma redução de 1.247 pontos de vendas, ou 19,15%, sem impactar de forma significativa o faturamento total, além de prover recursos visuais permitindo análise dos pontos de vendas sob óticas distintas.

Palavras-chave: Sistemas de apoio a decisão, *Big Data*, Processo ETL

ABSTRACT

The telecommunications industry has been suffering in recent years with their ability to increase its profitability. In this way, large companies in the market have adopted a strong strategy of reducing costs, promoting cuts in several segments of the business in an attempt to leverage its margin. Therefore, considering the need for performance improvements and assertiveness in the allocation of resources, the possibility of implementing a decision support system with the aid of a Data Envelopment Analysis (DEA) methodology was studied in a telecommunications company located in the city of Fortaleza. For this, one must take into account technological factors, using tools considered innovative for the market. Initially, the data sources needed to work were mapped and the quality of the data received was evaluated. After mapping, the architecture of the project was idealized in a second moment and validated by proof of concept (PoC) applied to a sample base. The development took place with the ETL process creation, performing treatments in the main fields related to neighborhoods and geographic coordinates, application of DEA methodology for prioritization of actions according to scale of efficiency and design of a methodology to classify the points of sales (POS) according to the level of revenue generated. The functionality of the system was validated by the practical application on the monthly basis of an operator, where it achieved a reduction of 1,247 points of sale, or 19.15%, without significantly impacting the total revenue, besides providing visual aids allowing analysis of each POS under different optics.

Keywords: Decision support systems, Big Data, ETL process

LISTA DE FIGURAS

Figura 1– Pesquisa on-line *Big Data* 26
Figura 2– Divisão estrutura de dados 27
Figura 3 – As 5 fases principais do *Big Data* 30
Figura 4– Métodos de análise preditiva 32
Figura 5 – Metodologia *Agile* para *Business Intelligence* 38
Figura 6– Exemplo geral de sistema 44
Figura 7 - Modelagem CCR 50
Figura 8– Modelo BCC 50
Figura 9– Campos String em Repositório Geral (Interna) 55
Figura 10–Campos numéricos em Repositório Geral (Interna) 55
Figura 11– Campos String em BI PDV (Interna) 55
Figura 12 - Campos numéricos em BI PDV (Interna) 56
Figura 13 - Campos String em Censo Bairros Fortaleza (IBGE) 57
Figura 14 - Campos numéricos em Censo Bairros Fortaleza (IBGE) 57
Figura 15 - Campos String em base Bairros Shape (Portal Fortaleza Dados Abertos) .. 57
Figura 16 - Campos numéricos em Bairros Shape (Portal Fortaleza Dados Abertos) ... 58
Figura 17 - Campos String da base Bairros IDH (Portal Fortaleza Dados Abertos) 58
Figura 18 - Campos numéricos da base Bairros IDH (Portal Fortaleza Dados Abertos) 58
Figura 19 - Prova de conceito campo bairros sem tratamento 62
Figura 20 - Prova de conceito campo bairros com tratamento 63
Figura 21 - Código para implementação simples de análise envoltória por dados 64
Figura 22 - Dashboard histórico de eficiência I 65
Figura 23 - Dashboard histórico de eficiência II 66
Figura 24 - Processo de extração das bases internas 67
Figura 25 - Primeira etapa do processo ETL 68
Figura 26 - Consolidação bases externas 69
Figura 27 - Exemplo preenchimento incorretos 70
Figura 28 - Tratamento Bairros 70
Figura 29 - Exemplo de grupo gerado pela lógica fuzzy 71
Figura 30 - Exemplo de tratamento manual bairros 72
Figura 31 - Tratamento campos Latitude/Longitude 73
Figura 32 - Parte final Processo ETL 74

Figura 33 - Metodologia DEA 75
Figura 34 - Código para implementação de modelagem DEA 76
Figura 35 - Análise Pontos de Vendas 77
Figura 36 - Categorias de faturamento 78
Figura 37 - Métricas de teste 79
Figura 38 - Exemplo estrutura de output 81
Figura 39 - Interface usuário 81
Figura 40 - Situação Janeiro/2015 82
Figura 41 - Clusters Janeiro/2015 83
Figura 42 - Matriz correlação Pearson 84
Figura 43 - Planilha correlação Pearson 84
Figura 44 - Faturamento x Quantidade de domicílios 85
Figura 45 - Matriz correlação Spearman 85
Figura 46 - Planilha correlação Spearman 86
Figura 47 - Coeficiente de variação 86
Figura 48 - Resultado aplicação DEA 87
Figura 49 - Visualização grau de eficiência 88
Figura 50 - Situação proposta pelo modelo 89
Figura 51 - PDV's removidos 90
Figura 52 - Bairro Aldeota situação inicial 90
Figura 53 - Bairro Aldeota situação final 91
Figura 54 - Avaliação métricas performance 92

LISTA DE QUADROS

Quadro 1 - 5 V's do Big Data 29
Quadro 2 - Fatores chave em BI 38
Quadro 3 - Conceitos de eficiência 44
Quadro 4 - Definição de métricas teste 79
Quadro 5 - Benefícios gerados no trabalho 94

SUMÁRIO

1. INTRODUÇÃO 14
1.1 Contextualização 14
1.2 Justificativa 15
1.3 Objetivos 15
1.3.1 Objetivo geral 15
1.3.2 Objetivos específicos 16
1.4 Metodologia da pesquisa 16
1.5 Estrutura do trabalho 17
2. FUNDAMENTAÇÃO TEÓRICA 18
2.1 Teoria da decisão 18
2.1.1 Teoria da decisão especial 21
2.1.2 Sistemas de apoio a decisão espacial 23
2.1.3 Construção de sistemas SADE 24
2.2 Definindo *Big Data* 25
2.2.1 Big Data Analytics 29
2.2.1 Extração, transformação e carregamento 34
2.3 *Business Inteligence* 36
2.3.1 Infra-estrutura de Business Inteligence 37
2.3.2 Ciclo de vida de um projeto de Business Inteligence 38
2.4 Análise por Envoltória de Dados 42
2.4.1 Eficiência Técnica 44
2.4.2 Etapas de aplicação dos modelos DEA 46
3. ESTUDO DE CASO 52
3.1 Caracterização da empresa 52
3.2 Etapas do estudo 52
3.3 Descoberta 54

3.4 Arquitetura 58
3.4.1 Caracterização da arquitetura 59
3.4.2 Mapeamento de ferramentas necessárias 59
3.4.3 Prova de conceito 60
3.5 Concepção/Desenvolvimento 67
3.5.1 Processo ETL 67
3.5.2 Aplicação da metodologia DEA 74
3.5.3 Análise dos pontos de vendas passíveis de remoção 76
3.6 Teste/Produção 78
3.7 Aplicação prática 82
3.7.1 Situação atual 82
3.7.2 Situação proposta 83
3.7.3 Comparação de resultados 91
4. CONCLUSÃO 95
REFERÊNCIAS 97
APÊNDICE A – FLUXO VERSÃO FINAL 101

1. INTRODUÇÃO

1.1 Contextualização

A indústria de telecomunicações vem apresentando dificuldades nos últimos anos para conseguir elevar sua lucratividade. Dessa forma, grandes empresas do mercado têm adotado uma forte estratégia de redução de custos, promovendo cortes em diversos segmentos do negócio na tentativa de alavancar sua margem.

De modo semelhante, as distribuidoras responsáveis pelas recargas de aparelhos celulares têm sofrido grandes impactos, visto o baixo percentual que é repassado pelas operadoras para as distribuidoras como fonte de receita das recargas realizadas, o que implica na dificuldade de conseguir manter-se saudável no mercado. Além do baixo percentual que é recebido por parte dessas empresas, ainda existe uma comissão de 30% do valor recebido que deve ser repassado aos pontos de vendas, responsáveis locais pela recarga ao usuário final. Outros custos associados ao negócio de distribuição de recarga, envolve o aluguel das máquinas de recarga assim como o salário dos responsáveis da empresa por supervisionar os pontos de vendas, perfazendo rotas todas as semanas que devem abranger os diversos comércios representados pela distribuição.

Para fins práticos, também deve se levar em consideração a possibilidade que algumas operadoras oferecem de programas de excelência, os quais dão oportunidade ao distribuidor de aumentar a margem recebida, caso atinja certos critérios pré-estabelecidos pelas operadoras, sendo um dos critérios a capacidade, por meio da quantidade de pontos de vendas, para atingir a população como um todo.

Sendo assim, na busca por tornar suas operações mais eficientes, as distribuidoras de recarga estão cada vez mais procurando se modernizar e assim fazer uso de novas tecnologias que possam impactar suas decisões de forma positiva, trazendo embasamento científico para a gestão de suas operações.

O presente trabalho foi desenvolvido em uma empresa de médio porte com sede na cidade de Fortaleza, atuante no segmento de distribuição de recargas pré-pagas e possuindo abrangência nacional.

1.2 Justificativa

A crescente massa de dados que são trabalhadas diariamente pelas empresas trazem consigo imensas oportunidades de alavancar as decisões relacionadas ao negócio, podendo oferecer vantagem competitiva (TAURION, 2013). A difusão de tecnologias *Big Data* ainda é assunto relativamente recente no Brasil e pouco é utilizado quando restringimos ao contexto de pequenas e médias empresas. Desta forma, busca-se explorar estas aplicações dentro de áreas da Engenharia de Produção, com auxílio de Pesquisa Operacional.

O problema proposto no trabalho foi baseado na intenção de contribuir para o aumento da lucratividade das empresas de distribuição de recarga, responsáveis pela alocação dos pontos de venda, atacando a componente custos do negócio. Este estudo se justifica pela necessidade de apresentar embasamento científico para analisar a performance dos pontos de vendas, atendendo a demanda existente, com o menor custo possível, alavancando oportunidades frente a concorrência.

A Análise Envoltória de Dados (DEA) é uma ferramenta poderosíssima de Pesquisa Operacional, que já apresenta centenas de publicações em diversas áreas de estudo como setor público, saúde, engenharia reversa, energias renováveis, entre outros.

O uso de *softwares* de ponta como *Alteryx* e *Tableau* para processamento e análise de dados, *softwares* de processamento e visualização de dados georreferenciados como *CartoDB*, além de linguagens de programação como *R* para aplicação dos modelos de Pesquisa Operacional é parte fundamental do desenvolvimento *Big Data* proposto neste projeto.

No presente trabalho, há a aplicação de conceitos relacionados a *Big Data*, para a criação de um sistema de apoio a decisão espacial que possa auxiliar os tomadores de decisão a entender o atual cenário em que se encontra alocados seus recursos, diagnosticar possíveis problemas relacionados a essa estrutura e compreender aonde e como esses recursos podem ser melhor utilizados.

1.3 Objetivos

1.3.1 Objetivo geral

O objetivo geral consiste em conceber um modelo de apoio a decisão espacial, para avaliar a eficiência dos pontos de venda através da análise por envoltória de dados, visando aprimorar a performance geral do sistema, através de melhor uso dos recursos disponíveis.

1.3.2 Objetivos específicos

Os objetivos específicos desse trabalho são:

1. estudar e compreender o método de Análise Envoltória de Dados, bem como sua formulação;
2. identificar os fatores e parâmetros que devem ser considerados na formulação do problema;
3. criar fluxo de extração, transformação e carregamento dos dados para utilização modelos analíticos;
4. modelar o problema pelo método de Análise por Envoltória de Dados, utilizando o *software* Alteryx Designer.

1.4 Metodologia da pesquisa

O presente trabalho pode ser caracterizado como de natureza aplicada e quantitativa que, segundo Silva e Menezes (2005), busca gerar conhecimentos práticos e dirigidos à solução de problemas específicos além de utilizar-se de recursos estatísticos para traduzir em números opiniões e informações, que serão posteriormente analisadas.

De acordo com Silva e Menezes (2005), as pesquisas podem ser classificadas quanto aos objetivos e procedimentos técnicos, sendo a pesquisa descritiva uma das categorias de objetivos. Para este tipo de pesquisa os autores definem como aquela que visa descrever as características de determinada população ou fenômeno ou o estabelecimento de relações entre variáveis.

Tendo em vista a denominação anterior, este trabalho enquadra-se neste tipo de pesquisa, descritiva, pois envolve entendimento das variáveis que irão influenciar o sistema e suas interações

Conforme a categoria referente aos procedimentos técnicos apresentados por Silva e Menezes (2005), este trabalho caracteriza-se como, pesquisa bibliográfica e estudo de caso.

De acordo com o trabalho do autor, pesquisa experimental determina um objeto de estudo, variáveis que poderiam influenciá-lo, formas de controle e de observação dos

efeitos e estudo de caso consiste no estudo profundo e exaustivo de um ou poucos objetos, de maneira que permita seu amplo e detalhado conhecimento.

Para realizar o estudo proposto foram coletados dados a respeito de cada ponto de venda, de cada distribuidora de recarga, no nível de dia da transação. Desta forma, têm-se uma base com informações bem detalhadas, além das informações qualitativas a respeito de cada local, como o bairro onde está localizado, coordenadas geográficas, segmento de atuação, entre outras. Por último, buscou-se realizar o enriquecimento da base com bases externas, providas pelo IBGE e pela Prefeitura de Fortaleza, de forma a inserir mais informações a respeito de cada bairro, como população, área, renda por domicílio, quantidade de estabelecimentos comerciais, entre outras.

1.5 Estrutura do trabalho

Este trabalho foi dividido em quatro capítulos e em cada capítulo há subdivisões que objetivam uma organização adequada e de fácil acompanhamento por parte do leitor.

No primeiro capítulo, Introdução, é apresentada uma abordagem inicial da monografia, no qual são apresentadas a contextualização do desenvolvimento do estudo, a justificativa de sua aplicação, a definição dos objetivos geral e específicos do trabalho e a metodologia de trabalho do estudo.

No segundo capítulo, a fundamentação teórica, é apresentada primeiramente uma discussão a respeito do tema Teoria da Decisão, o qual serve de base para o desenvolvimento do estudo e abordagem da problemática. Em seguida, conceitos relacionados a *Big Data* e *Business Inteligence* são apresentados, assim como a teoria que irá fundamentar a implementação do modelo de Análise por Envoltória de Dados.

No terceiro capítulo, o estudo de caso, há a apresentação do atual cenário encontrado, além da caracterização da problemática e a forma como inicialmente foi abordado o projeto para entender como o tipo de tecnologia apresentada no trabalho poderia ajudar a otimizar as decisões diárias dos gestores.

O quarto capítulo corresponde às conclusões. São explicitadas as considerações finais sobre o modelo proposto e as ferramentas utilizadas, é abordado como se obteve os objetivos específicos e geral e sugerido trabalhos futuros neste assunto.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo encontra-se o referencial teórico com o intuito de melhorar o discernimento acerca do estudo de caso. Serão apresentados os conceitos de Teoria da Decisão, *Big Data*, processos ETL, *Business Inteligence* e Análise por Envoltória de Dados.

2.1 Teoria da decisão

De acordo com Levin e Milgron (2004), a ação de tomar decisão é uma atividade executada por cada indivíduo todos os dias e, para realização desta tarefa, diversas variáveis podem influenciar e impactar a decisão final, incluindo a opinião de pessoas, especialistas, empresas, consultorias, entre outros.

A teoria da decisão muitas vezes pode se enquadrar tanto como atividade científica como atividade profissional e a principal característica que leva a essa dinâmica de entendimento é a sua abordagem formal e abstrata. Pelo termo formal, devemos entender como a própria linguagem utilizada para descrever o tema, o qual busca-se reduzir ao máximo as ambiguidades existentes na comunicação humana. Já a questão da abstração, refere-se à utilização de uma linguagem que possa ser entendida independente do domínio de discussão. Dessa forma, a utilização de uma abordagem formal e abstrata irá contribuir para a adoção de um modelo de racionalidade, conceito esse de extrema importância dentro da teoria da decisão (TSOUKIAS, 2006).

Tsoukiàs (2006), afirma que a utilização de uma abordagem formal e abstrata apresenta uma série de vantagens e desvantagens. Entre as desvantagens pode-se citar:

- A ineficácia quando comparada a comunicação direta;
- A redução de ambiguidade na comunicação que nem sempre é algo desejado;
- O fato de conceber restrições a intuição e a criatividade inerente ao comportamento humano.

Já em relação as principais vantagens da abordagem, tem-se:

- Permitir a compressão mútua de todos participantes envolvidos no processo de decisão;
- Facilitar a transparência e participação dos envolvidos permite identificar estruturas que podem ser reutilizadas em outras situações ou problemáticas.

As diferentes estruturas relacionadas ao domínio da teoria de decisão podem ser separadas em quatro diferentes categorias: Abordagem normativa, descritiva, prescritiva e construtiva (TSOUKIAS, 2006).

Hansson (1994) define a teoria de decisão normativa como sendo a teoria sobre como as decisões deveriam ser tomadas. Estas decisões irão seguir normas de racionalidade que para o autor, são as únicas e mais importantes regras que o decisor deve optar para a tomada de decisão.

Para Tsoukiàs (2006), cada diagnóstico, representado por um estado natural, está associado com uma probabilidade e cada tratamento, consistindo de ações potenciais, está associado a consequências. Diante disso, deve-se construir uma função de utilidade do decisor (cliente), pela qual busca-se medir a preferência dos indivíduos, sendo esta essencial para aplicação da teoria da decisão. Diante desta função, busca-se sua maximização que permitirá identificar a melhor solução dentro do cenário proposto. Esta solução irá maximizar o valor esperado pelo decisor.

Segundo Savage (1972), a existência dessa função é garantida por uma série de axiomas que enunciam aquilo que na teoria iria constituir o comportamento racional do agente decisor. Sendo assim, no contexto normativo o decisor deve adaptar seu comportamento aos axiomas estabelecidos do contrário não será considerado racional.

Na abordagem descritiva, Hansson (1994) acredita que esta teoria representa como as decisões são de fato realizadas. Tsoukiàs (2006), aprofunda este assunto constatando que quando o decisor não está disposto a seguir os axiomas estabelecidos pela teoria clássica, ele irá procurar um modelo de racionalidade baseado não mais na teoria e sim em uma base empírica. A filosofia por trás dessa abordagem se resume ao questionamento: se há agentes decisores que adotaram estratégias de sucesso alternativas, dentro de condições similares, porque não replicar esta decisão? Dessa forma, para Tsoukias (2006) a abordagem descritiva se compromete em estabelecer modelos e estratégias baseados na observação do comportamento de reais agentes.

A teoria da decisão prescritiva procede de uma situação mais complexa que as anteriores. Essa abordagem provém de situações onde o decisor não se encaixa em nenhum dos modelos de racionalidade. No caso, o cliente ou decisor, pode ter preferências distintas, porém não possui os recursos, seja tempo ou financeiro, para ir adiante. Mesmo assim, é necessário propor alguma recomendação.

Para Roveri (2011), a abordagem prescritiva diz respeito a todos os métodos e técnicas relacionados tanto a auxiliar o decisor na tomada de decisão quanto a fornecer uma solução

para o problema da decisão através de *inputs* do decisor. Sendo assim, Tsoukiàs (2006) afirma que a legitimidade dessa abordagem vem do próprio cliente. Sendo assim, o modelo busca suas proposições e restrições a partir da situação problemática e do *modus operandis* dos agentes decisores, sem forçar o mesmo a aceitar as recomendações que não são convenientes para os seus problemas.

A abordagem construtiva trata do cenário mais próximo da realidade, no qual os clientes não possuem ideias suficientemente claras para fornecer algo semelhante a um modelo de racionalidade. Questões como: será que foram realizados todos os possíveis diagnósticos da situação? ou será que foram analisados todos os possíveis tratamentos?, irão caracterizar o contexto no qual insere-se essa abordagem. Desta forma, o que torna complexa a situação apresentada é o fato de normalmente as pessoas não saberem ou compreenderem do que realmente se trata o problema.

Tsoukiàs (2006) esclarece que essa abordagem trata de construir o problema e sua solução em paralelo. Diante disso, faz-se necessário, junto ao cliente, construir uma representação da situação problema e formalizar a formulação do problema com o aval do cliente para então construir o modelo mais apropriado de suporte a decisão. Cabe ressaltar aqui, que esta abordagem possui um fator aprendizagem intrínseco a sua forma de trabalho onde o cliente irá aprender como entender o seu problema de um ponto de vista formal e abstrato e a equipe responsável irá aprender a compreender o problema do ponto de vista do cliente.

Conforme Resnick (1987), os principais clientes de estudo da área de teoria da decisão e Pesquisa Operacional são empresas e organizações cujo foco envolva administração de redes. Exemplos são: distribuidoras de água, empresas do ramo de telecomunicações, ferroviárias, companhias de transporte e companhias aéreas.

Os estudos relacionados a complexidade das questões apresentadas por essas empresas torna-se um problema de extrema importância. De acordo com Tsoukias (2006), variados algoritmos desenvolvidos para solução desses problemas acabam por não ser escaláveis na presença de uma grande quantidade de dados, tornando inviável o tempo para encontrar uma solução ótima.

Novas tecnologias vêm sendo desenvolvidas para abordar a questão da viabilidade no tempo. Exemplo dessas técnicas são os programas de inteligência artificial que buscam a criação de máquinas pensantes, capazes de aprender padrões com dados passados. Essas abordagens também tratam de buscar soluções satisfatórias e não mais ótimas (TAURION, 2013).

Nas últimas décadas, muitos trabalhos foram desenvolvidos na problemática de multicritérios de decisão que eventualmente encontra-se em conflitos.

Para Tsoukias (2006), o conceito de eficiência torna-se extremamente importante para introduzir uma definição objetiva dos problemas. Segundo o autor, uma solução é dita eficiente caso não haja nenhuma outra solução que seja, no mínimo, tão boa quanto ela nos critérios analisados e melhor que ela em ao menos um critério.

A dificuldade que aparece nessa abordagem é a possibilidade de muitas soluções serem consideradas eficientes, tornando complexo o processo de selecionar uma alternativa possível. Desta forma, diferentes técnicas vêm sendo desenvolvidas para explorar o conjunto de soluções viáveis e encontrar aquela que seria mais conveniente.

2.1.1 Teoria da decisão especial

Desde a década de 60, variadas pesquisas nas áreas de ciência da informação geográfica e sistemas de apoio a decisão vêm sendo publicadas. Rafaeli Neto (2004), explica que estas duas áreas amplamente estudadas, possuem nuances e características que podem ser exploradas em conjunto. Exemplo de trabalhos citados por Rafaeli Neto (2004) é uma publicação de Densham em 1991, intitulado "*Spatial Decision Support Systems*" que relaciona essas duas áreas criando este novo campo de estudo denominado em português de Sistema de Apoio à Decisão Espacial (SADE). Desde então, este termo vem sendo utilizado para designar sistemas de informação com capacidade de auxiliar o ser humano a tomar decisões baseadas em dados referenciados geograficamente.

Atualmente, diversas ferramentas vêm sendo utilizadas para visualizar e manipular dados georeferenciados, seja em um computador local ou em um computador cliente (via Internet, com certa limitação e uso específicos), buscando auxiliar na tomada de decisões relacionadas, por exemplo, a problemas de roteamento de veículos e localização de facilidades. Tais ferramentas fazem parte do chamado SADE, nos quais técnicas para a solução de problemas específicos são implementadas sobre um ambiente de Sistemas de Informações Geográficas (SIGs).

De acordo com Rafaeli Neto (2004, p.02), um problema espacial é representado por "uma insatisfação gerada sobre o estado atual da posição, conformação ou atributos de entidades pertencentes a sistemas geográficos, onde a decisão espacial designa a decisão tomada com base em dados espaciais (posição ou conformação)".

Desta forma, as tecnologias concebidas para apoio à decisão espacial devem dar suporte ao tomador de decisões durante todo o processo decisório, auxiliando-o na exploração do espaço-problema para que possa buscar diferentes alternativas para a solução do problema específico chegando finalmente na escolha da melhor alternativa para o cenário demandado, resultando em sua decisão final.

Baseado nisso, Pomerol *et al.* (2004), citam a existência de 3 etapas fundamentais dentro deste processo, sendo elas: Etapa de inteligência, etapa de projeto e etapa de escolha.

Para Bruschi *et al.* (2004), a etapa de inteligência consiste em uma busca detalhada e precisa de informações analisando o ambiente e identificando as situações que exigem decisão. Sendo assim, esta fase demanda conhecimento especialista sobre as variáveis que influenciam ou não o problema especificado. O decisor será responsável por analisar o sistema geográfico onde o problema se encontra e delimitar os componentes deste sistema, assim como entender os elementos as relações entre estes elementos.

Segundo Rafaeli Neto (2004), o SADE pode atuar como repositório de informação sobre o sistema em análise, além de prover ferramentas para que o decisor explore, manipule e apresente os dados de interesse.

Hoppen *et al.* (1989), tratam a etapa de projeto como responsável pela modelagem das possíveis soluções. Rafaeli Neto (2004) explica que o decisor deve construir alternativas de solução para o problema, ponderando o papel de cada agente causador. Segundo o autor, o uso de ferramentas como modelos de otimização, simulação, previsão, análises heurísticas, entre outros, irá embasar essa etapa. Assim, o SADE irá representar uma simulação fidedigna do mundo real, onde o decisor poderá analisar os possíveis cenários existentes e suas consequências, tomando a decisão que for mais conveniente ao negócio.

A fase final é representada pela etapa de escolha. Para Bruschi *et al.* (2004), esta etapa consiste na definição da ação a ser seguida, isto é, escolha de uma das alternativas encontradas. Hoppen *et al.* (1989) definem esta fase de forma parecida, sendo esta responsável pela seleção da solução satisfatória.

Dentro desta fase o uso de técnicas de Pesquisa Operacional como modelos de decisão multicritério podem oferecer maior embasamento ao decisor, classificando e ordenando as alternativas segundo pesos e/ou critérios além de prover análises de sensibilidade para avaliar os diversos cenários possíveis. Vale ressaltar que o uso de interfaces visuais pode facilitar o entendimento do modelo considerado.

2.1.2 Sistemas de apoio a decisão espacial

Para Denshaw (1991), SADEs foram concebidos para prover ao usuário do sistema, um ambiente relacionado a tomada de decisão que permita realizar análises que envolvam informações geográficas de uma maneira flexível. Com isso, são sistemas de informação destinados a auxiliar decisões baseadas em dados geográficos (posição, geometria e atributos).

Para análise de problemas espaciais, há que se dar atenção aos modelos de simulação, por representarem o comportamento dinâmico do sistema do mundo real.

De acordo com Rafaeli Neto (2004), a maior utilidade do SADE está no suporte a problemas não estruturados e semi-estruturados. Segundo o autor, a estrutura de um problema se refere ao nível de conhecimento que existe sobre as causas, as consequências e o processo de solução.

Um problema é considerado estruturado se sua definição e fases de operação para chegar aos resultados desejados estão bem claros e sua execução repetida é sempre possível. Os problemas semiestruturados são problemas com operações bem conhecidas, mas que contêm algum fator ou critério variável que pode influenciar o resultado, como acontece com o problema de previsão. Nos problemas não estruturados, tanto os cenários, como os critérios de decisão, não estão fixados ou conhecidos *a priori* (TURBAN e ARONSON, 1998).

Rafaeli Neto (2004) estabelece o uso do paradigma diálogo-dado-modelo (DDM) como estrutura para entender os elementos incluídos no SADE, de forma que ele possa dar suporte a decisões. O DDM indica primeiramente a necessidade de informação que é representada pelos dados armazenados no Banco de Dados, já os modelos são representados por algoritmos computacionais responsáveis pela modelagem matemática das relações entre os diversos objetos georreferenciados e a comunicação é representada por interfaces gráficas computacionais, com as quais o decisor interage.

Atualmente, novas abordagens para estruturar o cenário SADE procuram explorar mecanismos da Inteligência Artificial. O uso de conceitos como aprendizado de máquina e agente inteligente estão sendo adaptados para o uso na problemática de decisões espaciais. No aprendizado de máquina, a análise de dados passados é essencial para o treinamento do modelo que a partir de novos *inputs* e *outputs*, será criado um novo conjunto de regras que irá contemplar a nova situação, buscando evoluir na forma como entende os padrões de informações passadas.

O conceito de agente inteligente, segundo Rafaeli Neto (2004), é representado por um conjunto de processos autocontidos que se executam em segundo plano. Desta forma os

agentes têm o papel de automatizar certos procedimentos, como manter o *software* atualizado automaticamente, responder questionamentos do usuário, processar dados, gerar relatórios, entre outros.

2.1.3 Construção de sistemas SADE

Neto e Rodrigues (2001) mostraram que, nas principais estratégias praticadas de desenvolvimento de sistemas SADE, o sistema de informação geográfica (SIG) é considerado o subsistema principal e a modelagem matemática científica, o subsistema secundário. De acordo com os autores, as diferentes estratégias variam na proximidade lógica e física entre SIG e *softwares* de modelagem científica, na forma de transação de dados entre si e na proximidade lógica e física do subsistema de integração com os dois subsistemas integrados por este.

O termo SIG, segundo Pesse *et al.* (2003), é utilizado para caracterizar sistemas computacionais que tratam dados com características espaciais, ou seja, manipulam banco de dados geograficamente referenciados.

Num contexto mais amplo, os SIG's permitem capturar, modelar, recuperar, manipular, consultar, apresentar e analisar bases de dados conectados a informações geográficas.

Para Neto e Rodrigues (2001), as estratégias de acoplamento de subsistemas de *software* se ramificam em 4 opções diferentes, sendo elas: acoplamento livre, acoplamento próximo, acoplamento rígido e acoplamento pleno.

Kirschbaum (2016), em uma definição generalizada, refere-se diretamente à noção de acoplamento fraco como a situação que permite flexibilidade ao sistema. Assim, a ação local é possível sem que uma mudança global ocorra.

Segundo Rafaeli Neto (2004), no acoplamento livre não há integração lógica nem física entre os subsistemas. Seu maior atrativo está no aproveitamento integral de subsistemas existentes. Os esforços de programação se concentram sobre o *software* que realiza o controle da integração, conferindo custo e tempo de desenvolvimento relativamente sintéticos. O desempenho do sistema tende a ser baixo devido à necessidade de operações de tradução e depuração dos dados transferidos entre os subsistemas. O SADE também tende a ser lento, quando se realizam simulações de comportamento do sistema real, que envolvam um número significativo de classes de entidades geográficas.

No acoplamento próximo, a sistemática continua muito semelhante ao acoplamento livre com a diferença que o *software* de controle da integração seria incorporado a um dos subsistemas integrados por ele. Diante disso, ainda não há integração lógica nem física entre os subsistemas (RAFAELI NETO, 2004).

No caso do acoplamento rígido, ainda existem subsistemas independentes. Sendo assim, para Rafaeli Neto (2004), as tecnologias de SIG e de sistema de modelagem científica (SMC) resultam de modelos conceituais distintos, desenvolvidos em separado, por empresas ou instituições independentes, em diferentes épocas. Porém, diferente dos outros casos, os dados são transferidos através de arquivos intermediários, sendo um processo originado diretamente nas estruturas de armazenamento dos subsistemas SIG e/ou SMC. Isto confere uma proximidade lógica maior entre os subsistemas o que provoca um melhor desempenho do SADE.

A integração plena se caracteriza quando o *software* de integração, o SIG e o SMC fazem parte de um modelo conceitual único. Para haver integração plena é necessária a integração das partes responsáveis pela concepção e manuseio do SIG e SMC. O modelo conceitual único é necessário, pois garantirá o nível de acoplamento necessário para que dados, modelos científicos, relatórios e interfaces com o usuário garantam apoio eficiente e eficaz aos processos decisórios envolvendo problemas semi-estruturados (RAFAELI NETO, 2004).

2.2 Definindo *Big Data*

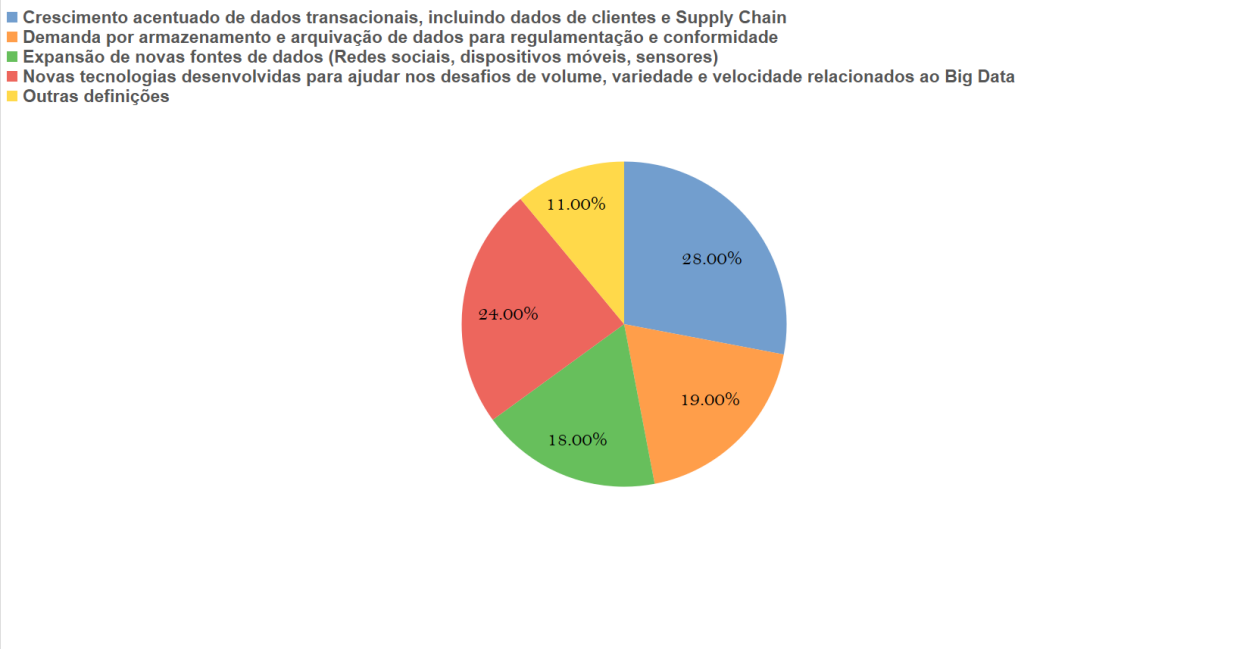
Atualmente, o conceito *Big Data* vem se difundindo dentro do mercado de trabalho e comunidade acadêmica, porém sua origem ainda é bastante recente. Para Gartner (2012), *Big Data* pode ser definido como sistemas informacionais que apresentam alto grau de volume, velocidade e variedade e necessitam de ferramentas e abordagens eficientes para processar os dados e extrair insights que irão auxiliar na toma de decisão organizacional.

Entre os grandes difusores deste conceito encontra-se a empresa IBM, que através de sua tecnologia voltada para *Question Answering* (QA), concebeu uma máquina, dentro de uma iniciativa de marketing, capaz de processar grande volume de dados e competir de igual para igual com especialistas humanos em um programa de TV, estilo pegunta-resposta (IBM, 2011).

De acordo com Gandomi *et al.* (2015), o termo *Big Data* vem evoluindo rapidamente, sendo assim sua definição acaba gerando questionamentos e dúvidas a respeito da

abrangência dessa área. Uma pesquisa *on-line* realizada pela empresa Harris *Interactive* ("*Small and midsize companies look to make big gains with big data*", 2012), consolidou as respostas de 154 executivos a respeito de como estas pessoas definiriam o termo *Big Data*. Na Figura 1 é ilustrada a divergência de respostas recebidas e como os participantes abordaram a pergunta.

Figura 1– Pesquisa on-line *Big Data*



Fonte: Autor

Com base na Figura 1 pode-se constatar que a consolidação das respostas aponta para "volume" ou "tamanho" como sendo a grande referência quando se pensa em *Big Data*.

Para Taurion (2013), volume é com certeza uma das fortes características que define esse fenômeno de *Big Data*, porém, para complementar a composição, ele sugere ainda a existência de mais duas características fortes: Variedade e Velocidade. Esses três V's formam a mais básica estrutura que irá compor a área de *Big Data*.

O primeiro "V" e talvez o principal diz respeito ao volume, ou seja, a magnitude relacionada a quantidade de dados a serem processados. Um exemplo desse tamanho segundo Wang et al. (2016), são as mídias sociais e sensores de localização que geram *terabytes* de dados a cada dia na internet.

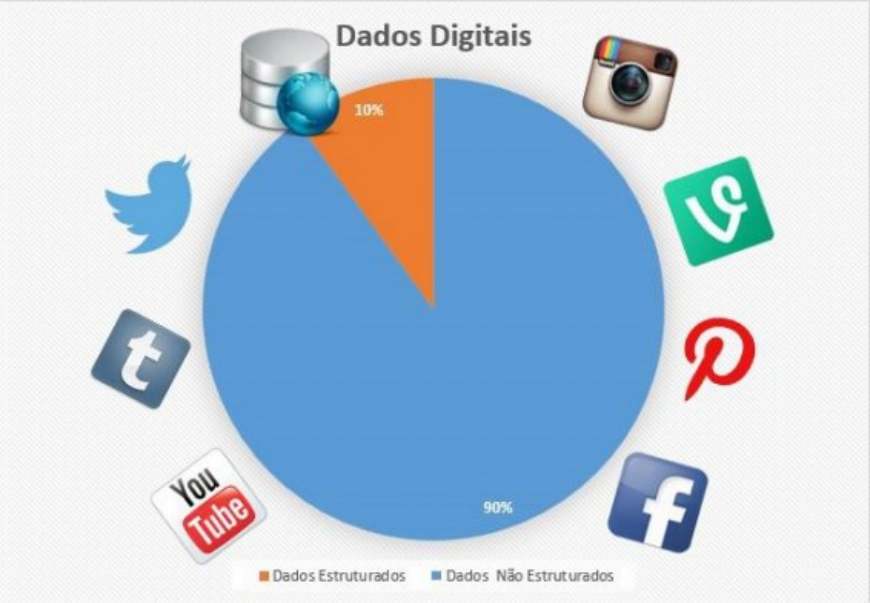
Segundo Taurion (2013), apenas a companhia Google sozinha processa mais de 24 *petabytes* de dados por dia e o Facebook faz *upload* de pelo menos 10 milhões de novas fotos a cada hora. Sendo assim, os dados de hoje vêm em todos os tipos de formato, sendo gerados milhões de dados por segundo e vindo de diversas fontes, implicando nas dimensões velocidade e variedade.

Para Gandomi *et al*. (2015), as definições relacionadas ao fator volume são relativas e podem variar tanto no tempo quanto no tipo dos dados. Para o autor, aquilo que hoje é considerado *Big Data*, pode vir a ser superado visto novas tecnologias e técnicas de processamento que podem evoluir com o tempo. Quanto ao tipo de dado, diferentes tecnologias também são escolhidas para tipos distintos. Exemplos de dados distintos referem-se ao processamento de textos, imagens, vídeos, entre outros.

O segundo "V" diz respeito a variedade intrínseca aos dados necessários para análises organizacionais. Segundo Gandomi *et al*. (2015), este fator refere-se a estrutura heterogênea encontrada em bases de dados.

De acordo com Gandomi *et al*. (2015), as estruturas dos dados podem ser classificadas em 3 categorias: estruturados, semi-estruturados e não-estruturados. Os dados estruturados referem-se as planilhas e bases de dados relacionais. Já os dados não-estruturados são representados por imagens, vídeos, áudios e textos, sendo as categorias de dados mais complexas para trabalhar apesar do recente desenvolvimento de técnicas computacionais para processar tais informações. Por último, os dados do tipo semi-estruturado como próprio nome já diz, apresentam certo grau de padronização, facilitando o acesso as informações contidas em seu meio. Como exemplo típico deste tipo de estrutura de dados, temos documentos XML (*Extensible Markup Language*). Na Figura 2, é apresentada, em percentual, a geração de dados divididos por tipos, considerando semi-estruturado como não-estruturado.

Figura 2– Divisão estrutura de dados



Fonte: Taurion (2013)

Para Taurion (2013), o alto grau de variedade intrínseco aos dados necessários para análises organizacionais não é um problema atual, porém, as tecnologias para manuseio e processamento destas informações possuem sua origem recente.

Como exemplo, tem-se empresas como a *Riminder*, francesa responsável por desenvolver algoritmos de inteligência artificial para processamento de currículos, relacionando os diferentes perfis de contratantes e contratados (LE MONDE, 2015). Dessa forma, novas técnicas relacionada a *Big Data* podem oferecer vantagem competitiva para empresas que passam a obter maior assertividade em suas decisões.

O último "V" do tripé básico diz respeito a velocidade que, como definido por Gandomi *et al*. (2015), refere-se a taxa na qual é gerado os dados assim como a rapidez que necessitam ser processados e analisados.

De acordo com Taurion (2013), a informação em tempo real, ou quase, permite a empresa ser muito mais ágil do que concorrentes. Uma das líderes mundiais em *Big Data Analytics*, a companhia SAS, afirma que *tags* de RFID, sensores, celulares e contadores inteligentes estão impulsionado a necessidade de lidar com imensas quantidades de dados em tempo real, ou quase real.

Os dados emitidos por dispositivos portáteis vêm ganhando importância dentro deste contexto e podem gerar diversas informações úteis para a tomada de decisão, como por exemplo análises comportamentais e de sentimento dos clientes. Desta forma, esta grande massa de dados necessita de ferramentas para armazenamento e processamento, gerando informações em tempo real que irão auxiliar gestores nas decisões diárias.

Para complementar este tripé, Wang *et al.* (2016) cita mais dois fatores intrínsecos a área de *Big Data*:

- Veracidade: Corresponde ao nível de confiança que pode ser atribuído aos dados recebidos direto da fonte. Um exemplo citado são os dados advindos de sensores, nos quais, os próprios dispositivos possuem margens de erro ou podem não estar funcionando corretamente, comprometendo os dados que serão transmitidos.
- Valor: Descreve o potencial financeiro que a organização pode conseguir através do uso de técnicas de *Big Data*. Dois aspectos são o grande valor potencial e a baixa densidade de valor. Estes conceitos representam oalto valor que pode ser obtido através do uso de *Big Data Analytics* versus o baixo valor que os dados originais têm (sem nenhum tipo de processamento).

Sendo assim, para Gandomi *et al*. (2015) as definições para cada um desses fatores devem adaptar-se as diferentes empresas, pois cada empresa possui, por exemplo, diferentes tamanhos, variedades e valores demandados de seus dados, de forma que sua realidade deve ser considerada dentro da criação de limites para a abrangência que terá a área de *Big Data* em sua organização.

O Quadro1 resume as definições citadas anteriormente.

Quadro 1 - 5 V's do Big Data

Atributos	**Definição**
Volume	Magnitude relacionada a quantidade de dados a serem processados.
Variedade	Estrutura heterogênea encontrada em bases de dados.
Velocidade	Taxa na qual é gerado os dados assim como a rapidez que necessitam ser processados e analisados.
Veracidade	Nível de confiança que pode ser atribuído aos dados recebidos direto da fonte.
Valor	Potencial financeiro que a organização pode conseguir através do uso de técnicas de *Big Data.*

Fonte: Autor

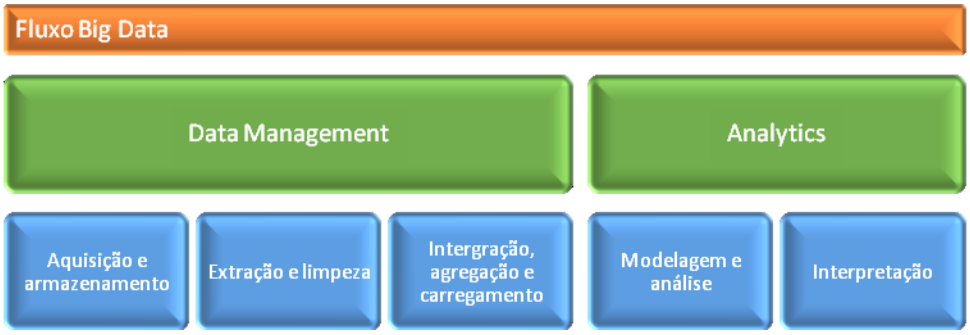
2.2.1 Big Data Analytics

Big Data Analytics representa um domínio ainda pouco explorado apesar da crescente difusão nos últimos anos. Porém, o valor da informação nunca esteve tão em cheque e as organizações podem obter vantagem competitiva através de boas práticas de processamento destes dados aliados a mão de obra especializada capaz de analisar e interpretar as informações processadas (LABRINIDIS e JAGADISH, 2012).

O potencial do *Big Data Analytics* é percebido quando o processo de tomada de decisão é alavancado através do seu uso. Cada vez mais as empresas estão buscando meios eficientes de transformar grandes e variados volumes de dados em poderosos *insights*. Desta forma, Labrinidis e Jajadish (2012) consideram cinco fases principais como base para uso do *Big Data* no processo de tomada de decisão, sendo estas fases subdivididas em 2 grupos: *data Management* e *Analytics*.

Na Figura 3 estas fases são apresentadas em seus respectivos grupos:

Figura 3 – As 5 fases principais do *Big Data*



Fonte: Adaptação Gandomi et al. (2015)

Para Gandomi *et al.* (2015), *Data Management* envolve as etapas de aquisição e armazenamento de informações que irão antecipar a transformação dos dados, removendo inconsistências e estruturando a base para ser utilizado na preparação de modelos e análises.

No caso de *Analytics*, o mesmo autor define como sendo técnicas usadas para analisar os dados de forma a extrair *insights* que possam ser utilizados para gerir os negócios de forma mais inteligente. Este último é onde encaixa-se o termo *Big Data Analytics*.

De acordo com a Gartner (2014), *Big Data Analytics* é uma prioridade para grandes negócios obterem vantagem competitiva, impelido pela necessidade de tornar mais acessível esses tipos de análises avançadas, assim como expandir o suporte a tomada de decisão. Segundo esta consultoria, o segmento de *Big Data Analytics* é um dos grandes mercados crescentes, superando a marca de 1 bilhão de dólares em 2013.

Para Gartner (2014), este segmento pode ser dividido basicamente em 4 tipos distintos de análises, sendo elas: Descritiva, diagnóstica, preditiva e prescritiva.

A análise descritiva inicializa o processo com a pergunta "O que aconteceu?". Segundo a IM *Advisor* (2016), essa análise é o ponto de partida da cadeia de valor do *Big Data Analytics*, porém pode vir a ser útil através da percepção de padrões que podem gerar *insights* interessantes ao modo como o negócio está sendo gerido.

Esta primeira análise se compromete essencialmente em buscar o que aconteceu no passado e no presente, para depois tentar entender o porquê das causas. Para isso, faz-se uso de técnicas gráficas para organizar os dados adquiridos. Exemplos de gráficos utilizados são: gráficos de barras, grafos, gráfico em pizza, mapas, gráficos de dispersão, entre outros. Todos estes procedimentos visuais facilitam o entendimento, provendo insights das informações contidas na base. Exemplos de aplicação dessa etapa, é o uso da performance financeira passada para entender tendências futuras de certos clientes (RAJARAMAN, 2016).

A análise diagnóstica procede a etapa de análise descritiva. A pergunta essencial que ela busca responder é “Por que aconteceu?”. Desta forma, segundo a empresa Hekima (2016), enquanto a análise descritiva busca detalhar uma base de dados, a análise diagnóstica tem como objetivo compreender de maneira causal (Quem, Quando, Como, Onde e Por quê) todas as suas possibilidades.

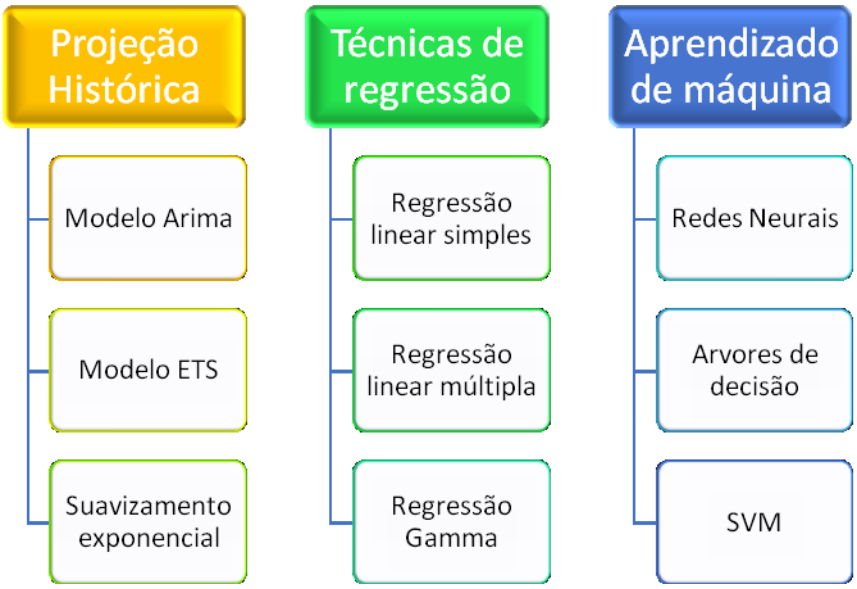
Para grande parte dos autores, uma aplicação básica diz respeito ao departamento de marketing e campanhas promovidas. De acordo com a IM *Advisor* (2016), a partir da análise descritiva você pode ver a quantidade de citações, postagens, seguidores, visualizações de páginas e então com a análise diagnóstica buscar uma visão geral dessas métricas e entender o que funcionou e/ou pode ser melhorado das campanhas passadas.

Sendo assim, esta análise irá funcionar como uma espécie de relatório expandido e quando feita em uma base de dados volumosa, permite entender a razão de cada um dos desdobramentos das ações adotadas e, a partir disso, mudar estratégias ineficazes ou reforçar as eficazes.

A análise preditiva tem seu funcionamento em grande parte baseado na análise descritiva e busca responder questões do tipo “O que irá acontecer?”. Para Gandomi *et al.* (2015), essa etapa do *Big Data Analytics*, compreende uma infinidade de técnicas que buscam prever os resultados futuros através de análises históricas e correlações entre variáveis. A IM Advisor (2016) entende esta fase como sendo um tipo de análise que busca entender padrões passados para prever o futuro.

Seguindo esse raciocínio, Rajaraman (2016), cita alguns dos métodos encontrados na literatura, como: Séries temporais, métodos estatísticos de regressão, redes neurais e variados algoritmos de aprendizado de máquina. Sendo assim, pode-se ver a concepção de 3 grandes grupos de técnicas utilizadas para análise preditiva, como mostrado na Figura 4, com alguns exemplos.

Figura 4– Métodos de análise preditiva



Fonte: Adaptação Rajaramam (2016)

Gandomi *et al.* (2015), vão além e propõem também uma divisão das técnicas baseado nas variáveis de saída, podendo ser modelos com variáveis de saída contínua ou discretas. Como exemplos de aplicações, Gandomi *et al.* (2015) cita a estimativa do preço de venda e aluguel de imóveis, assim como a previsão de bons pagadores ou de inadimplência dos inquilinos.

Fan *et al.* (2014), afirmam que as técnicas de análise preditiva são fundamentadas, quase que em sua totalidade, em métodos estatísticos; porém, existem diversos fatores que influenciam o desenvolvimento de novos métodos estatísticos para o processo de *Big Data Analytics*, entre eles:

- Significância estatística: Métodos convencionais são fundamentados no conceito de significância estatística. No entanto, visto a abrangência do *Big Data*, as grandes massas de dados representam quase que a maioria e em algumas vezes toda a população, não havendo necessidade de induzir resultados a partir de amostras. Desta forma, este conceito se torna irrelevante.
- Eficiência computacional: Métodos convencionais muitas vezes utilizados para amostras pequenas tornam-se ineficientes e não escaláveis para *Big Data*.
- Heterogeneidade: O grande volume de dados é altamente heterogêneo. Desta forma, pequenas amostras podem ser consideradas valores discrepantes devido a uma baixa frequência. No entanto, o tamanho de grandes conjuntos de dados cria a

oportunidade única de modelar a heterogeneidade decorrente de dados sub-populacionais, o que exigiria técnicas estatísticas sofisticadas.

- Acumulação de ruído: Algumas variáveis com significativo poder explicativo podem ser ignoradas como resultado do acúmulo de ruído.
- Correlação espúria:Refere-se a variáveis não correlacionadas sendo falsamente indicadas como correlacionadas, devido ao enorme tamanho do conjunto de dados. Dentro dessa questão, pode-se abordar a discussão entre causalidade versus correlação.
- Endogeneidade Incidental: Uma suposição comum na análise de regressão é o pressuposto de exogeneidade, que significa que as variáveis explicativas (preditores) são independentes do termo residual. A validade da maioria dos métodos estatísticos usados na análise de regressão depende dessa suposição. Em outras palavras, a existência de endogeneidade, ou seja, a dependência do termo residual em alguns dos preditores, prejudica a validade dos métodos estatísticos utilizados para a análise de regressão.

O último tipo de análise é a prescritiva. Esta análise se compromete em responder à pergunta “Como fazer acontecer?”. Segundo Rajaraman (2016), busca-se, através dos dados recebidos, identificar oportunidades de otimizar as soluções para os problemas existentes. Esse tipo de análise tem proximidade com técnicas de Pesquisa Operacional.

Para Hekima (2016), a análise prescritiva apresenta uma forma de definir qual escolha será mais efetiva em determinada situação, traçando as possíveis consequências de cada ação. No entanto, a análise prescritiva ainda é pouco utilizada, na maioria das vezes, por causa de desconhecimento e, segundo Gartner (2012), apenas 3% das empresas fazem uso dessa análise.

Um exemplo de aplicação é a precificação dos assentos em companhias aéreas, baseada no histórico dados, padrões de viagem, origens e destinos populares, grandes eventos, feriados, entre outros dados com o objetivo de maximizar o lucro da empresa.

2.2.1 Extração, transformação e carregamento

De acordo com Abreu (2010), a sigla ETL tem origem na língua inglesa, sendo melhor traduzida por "*Extract, Transform and Load*". Esta sigla representa um processo baseado em ferramentas que se destinam a extração, transformação e carregamento de dados.

Os dados trabalhados dentro deste processo podem vir das mais diversas bases, desde simples planilhas como grandes repositórios e sistemas. Quanto ao destino dado a estes dados, a carga pode ser direcionada a bancos de dados, sistemas de informação, *data warehouses* ou podem ser utilizados como carga para outros modelos e sistemas de inteligência de negócios.

No processo ETL, a etapa opcional diz respeito a manipulação e transformação dos dados, porém, sempre haverá o acontecimento das etapas de extração e carga, que irão justificar a utilização do processo. No entanto, normalmente os dados não se encontram em conformidade ou necessitam ser ajustados para atender as demandas em relação a carga que será realizada necessitando da etapa de transformação (CIELO, 2010).

O processo ETL pode ser desenvolvido através do auxílio de linguagens de programação com Python e R, porém, atualmente, existem diversos *softwares* que agilizam a construção e gerenciamento desse processo de forma segura e eficaz, além de facilitar a automatização e replicação dos processos caso seja necessário a reutilização dos mesmos procedimentos.

Segundo Date (2000) o ETL é utilizado como processo primordial em projetos de migração de dados para sistemas de informação, *business intelligence* e aplicações de *data warehouse*. Atualmente sua utilização vem crescendo na área de análise de dados, com a etapa de transformação sendo a principal para alimentação de modelos estatísticos e de mineração de dados no intuito de extrair *insights* que auxiliem os gerentes e tomadores de decisão no exercício de suas funções.

Sendo assim, o ETL pode ser utilizado em várias áreas de inteligência de negócios permitindo processamento dos dados para consumo em várias formas: a) Relatórios; b) Dashboards (painéis, *balanced scorecard*); c) Indicadores de Desempenho ("KPIs"); d) Multi-dimensional (OLAP).

2.2.1.1 Extração

Segundo Cielo (2010), o primeiro passo dentro do processo ETL é simplesmente a definição das fontes de dados e a extração delas. As origens podem ser várias e em diferentes

formatos, onde poderemos encontrar desde os sistemas transacionais das empresas até simples planilhas e arquivos textos.

Nesta fase, é necessário identificar o tipo, forma de armazenamento, estrutura e modelagem dos dados a serem extraídos, além da necessidade de viabilizar através da ferramenta de extração um meio de acesso a estes dados de origem.

2.2.1.2 Transformação

Este é o processo responsável pelo tratamento e transformação dos dados. Após o processo de extração, caso os dados não se encontrem em conformidade, deve-se realizar diversos procedimentos a fim de normalizar os dados e torna-los aptos a utilização na carga que será dada (CIELO, 2010).

Na obtenção dos dados em fontes muitas vezes desconhecidas ou que foram gerenciadas por sistemas de informação antigos, podendo existir falhas no projeto ou sem a utilização de um sistema gerenciador de banco de dados adequado, é comum encontrar problemas de integridade referencial ou inconsistências, como datas inválidas, atributos obrigatórios não preenchidos, somatórios numéricos inconsistentes, falta de normalização e diversos outros problemas.

Sendo assim, as inconsistências encontradas devem ser tratadas, buscando-se garantir confiabilidade as informações processadas. Segundo Cielo (2010), na maioria das vezes a transformação é necessária pois como os dados possuem frequentemente origens diferentes, eles deverão ser padronizados para que os dados que apresentam mesma informação possam ter seu valor comumente nomeado no banco de dados destino.

2.2.1.3 Carga

A etapa de carga ou carregamento deverá copiar os dados extraídos, tratados e manipulados nas etapas anteriores, em bancos de dados de destino ou em planilhas e outras extensões que possam vir a ser utilizadas pelos usuários.

De acordo com a necessidade dos usuários finais, a fase de carga poderá ser realizada uma única vez ou de forma periódica para atualização de dados, como por exemplo em projetos de inteligência de negócios (ABREU, 2010).

2.3 *Business Inteligence*

Business Intelligence (BI) é definida pela literatura e estudiosos de formas semelhantes. Noble (2006) define BI como uma ferramenta capaz de fornecer vantagens tecnológicas e de informação ao negócio, aumentando o ganho de produtividade e a eficiência com a qual as organizações realizam suas atividades.

Para Singer (2001), BI é uma ferramenta de agregação de valor que ajuda as organizações a obter informações auxiliares a tomada de decisões que relatórios regulares não são capazes de prover. Singer (2001) afirma que BI requer ferramentas, aplicativos e tecnologias focadas na melhora da tomada de decisão e é comumente usada na cadeia de suprimentos, vendas, finanças e marketing.

Os desafios na implementação de BI incluem a colaboração entre os setores de negócios e tecnologia da informação (TI) das organizações que resulta na transformação de dados em informações. A implementação de sistemas de BI é realizada através de uma metodologia que envolve um conjunto de processos, métodos e regras aplicadas dentro de uma disciplina.

No caso do BI, para termos uma metodologia bem-sucedida deve-se necessariamente concentrar na cadeia de valor da informação e menos no desenvolvimento de *software*, este último sendo normalmente o foco de desenvolvimento nos setores de TI tradicionais. Estudos mostraram que os ciclos de vida do tipo cascata e práticas de desenvolvimento de *software* tradicionais não são bem-sucedidos na aplicação a BI. *Software* e *hardware* não fornecem valor a organizações e sim o uso de informações (LARSON, 2009).

Segundo Larson (2009) o desenvolvimento de *software* é parte do processo de transformação de dados em informações úteis. No entanto, no caso do BI, o desenvolvimento consiste menos em criação de um programa e mais sobre a aplicação dos dados ao contexto de negócios.

Alguns *softwares* usados em BI incluem sistemas de gerenciamento de banco de dados, limpeza de dados, transformação de dados e sistemas analíticos. O escopo de desenvolvimento do BI foca mais na aplicação da lógica de negócios do que na programação.

Sendo assim, Larson (2009) explica que a fim de entender como aplicar a lógica necessária e configurar o *software*, os responsáveis pela implementação deverão primeiramente compreender o contexto de utilização e a relação dos dados com o negócio da organização.

No que diz respeito aos projetos na área de BI, existem obstáculos comumente encontrados pelas empresas, incluindo: requerimentos confusos; falta de entendimento sobre origem e uso dos dados; qualidade dos dados não é conhecida; restrições do sistema fonte impactam as possibilidades de serviços e utilização dos dados; divergências entre os departamentos de TI e as partes interessadas de negócios (LARSON, 2009).

A implementação de sistemas de BI tende a ser um processo em que as expectativas dos clientes seguem um ciclo de descoberta e refinamento, de forma que se torna importante a consolidação e ajustes de requerimentos. Transformar dados em informação não é um processo simples, mesmo com o uso de especialistas no assunto.

Um projeto de BI começa com algumas perguntas-chave: quais questões referentes ao negócio precisam ser respondidas? Que fontes de dados são necessárias para abordar estas questões? Como os dados serão utilizados? Estas questões são abordadas através de um processo de descoberta que examina como os dados são criados e como os dados serão convertidos em informação. Diante disso, a concepção de sistemas de BI inclui vários componentes, tais como ETL, bancos de dados e ferramentas de *front-end*. A infraestrutura de um sistema de BI é o ponto chave para extrair valor dos dados organizacionais (YEOH e KORONIOS, 2010).

2.3.1 Infra-estrutura de Business Inteligence

Yeoh e Koronios (2010) afirmam que um sistema de BI não é um sistema convencional de TI do tipo transacional, comumente conhecido. No entanto, os sistemas de BI têm características semelhantes a sistemas empresariais ou projetos de infraestrutura.

A implementação de sistemas de BI constitui uma atividade complexa que envolve *hardware*, *software* e recursos sobre a vida útil do sistema. A complexidade da infraestrutura do sistema BI aumenta com a abrangência de sua utilização. Um sistema de BI corporativo pode incluir um *data warehouse*, estruturas de dados integrados, sistemas de visualização e grandes volumes de dados.

Mungree, Rudra e Morien (2013) concluíram uma revisão de 15 anos do assunto que incluiu a pesquisa realizada por Yeoh e Koronios (2010). Nesta revisão, identifica-se fatores chave para o sucesso na implementação de sistemas de BI, que estão mostrados no Quadro 2:

Quadro 2 - Fatores chave em BI

Fatores chave para implementação de sistemas de BI
Gestão comprometida com resultados.
Estrutura técnica e recursos adequados.
Alinhamento do projeto com a estratégia de negócios.
Visão e requisitos extremamente claros e bem definidos.
Concepção orientada para o utilizador.
Gerenciamento de dados eficaz.
Gerenciamento do escopo do projeto.

Fonte: Adaptação Mungree, Rudra e Morien (2013)

Desta forma, o Quadro 2 reforça a necessidade principal dos projetos estarem alinhados com a estratégia de negócios, sendo feito em colaboração com os clientes e possuindo flexibilidade no que diz respeito ao escopo planejado.

2.3.2 Ciclo de vida de um projeto de Business Inteligence

Larson e Chang (2016) conceberam uma metodologia de implementação de sistemas de BI com base em pesquisas e experiência, que relaciona desenvolvimento *agile* com as práticas tradicionais de implementação BI.

O uso de boas práticas para o desenvolvimento de BI é justificado pelo surgimento do fenômeno de *Big Data*, onde se torna cada vez mais importantes conceitos como *fast analytics* no intuito de prover uma vantagem competitiva as empresas. Na Figura 5 são ilustradas as etapas da metodologia proposta por Larson e Chang (2016), as quais serão descritas posteriormente.

Figura 5 – Metodologia *agile* para *Business Intelligence*

Fonte: Adaptação Larson e Chang (2016)

2.3.2.1 Descoberta

Segundo Larson e Chang (2016), as expectativas dos projetos de BI nem sempre são claras para os interessados. Os usuários finais sabem que precisam de informação e habilidades analíticas, enquanto o departamento de TI precisa entregá-los sistemas que possam ser usados para melhorar a tomada de decisão.

O levantamento e *brainstorming* de demandas relacionadas a área servirá como ponto principal para coleta dos requisitos necessários. Estas perguntas serão o ponto de partida e servirão para fornecer *insights* sobre fontes de dados, dimensões e fatos necessários.

A qualidade e disponibilidade dos dados também são de extrema importância e irão determinar o que poderá ou não ser realizado.

Uma vez que as fontes de dados tenham sido mapeadas, o próximo passo é obter uma melhor compreensão dos dados e sua distribuição. Esta atividade é chamada de *data profiling* e é responsável por fornecer uma visão geral dos dados a partir de estatísticas descritivas, tais como: frequência de distribuição, valores máximo e mínimo, campos nulos ou em branco, exceções a valores de domínio, média, mediana, moda, e desvio padrão (LARSON e CHANG, 2016).

Para Larson e Chang (2016), o conhecimento adquirido a partir desta etapa irá fornecer a base para as métricas de qualidade dos dados e poderá ser usado mais tarde para a modelagem, desenvolvimento e testes. Este conhecimento também será útil na priorização das demandas que serão capazes de ser entregues.

Desse modo, com o uso de conceitos como *fast analytics* e *data science*, a análise exploratória de dados servirá como processo complementar para obtenção de relações entre as variáveis, assim como escolha dos registros e campos a serem considerados no uso de um dado modelo analítico, como, por exemplo, em técnicas de *data mining*.

2.3.2.2Arquitetura

No início de um programa de BI, a arquitetura tem de ser estabelecida. Criando uma arquitetura flexível e escalável, é essencial para dar suporte ao crescimento. Planejar a arquitetura é um passo essencial dentro de um ambiente *Agile* de desenvolvimento. Larson e Chang (2016) afirmam que uma boa forma de modelar a arquitetura de um projeto de BI é através do uso de técnicas de diagramação. Segundo eles, diagramas representam boas

práticas dentro do desenvolvimento *Agile*, visto sua facilidade e flexibilidade para sofrer alterações diferente de documentos texto. Diagramas incluem modelos de dados, fluxos de dados, fluxos de processo, e diagramas de infraestrutura. Quanto a arquitetura técnica do projeto, a entrega pode ser um diagrama descrevendo as diferentes tecnologias necessárias para seu desenvolvimento e como elas estão relacionadas.

Um modelo conceitual pode ser o início da arquitetura de dados. Diagramas são eficientes, porém, embora sirvam para facilitar o entendimento não irá servir como prova para implementação final. Cabe ressaltar que as decisões de arquitetura são normalmente difíceis de ser revertidas uma vez implementadas (LARSON e CHANG, 2016).

A abordagem de uma implementação de referência funciona bem no paradigma *Agile*. Como um miniprojeto, uma implementação de referência é um modelo de trabalho, mas se concentra em provar a funcionalidade da arquitetura desenhada. Implementações de referência para a arquitetura ETL, por exemplo, pode demonstrar se os níveis de serviço são possíveis e remover suposições sobre o potencial da tecnologia. A prova de conceito (POC) é também outra abordagem utilizada na validação de decisões de arquitetura (LARSON e CHANG, 2016).

Sendo assim, embora implementações de referência e POC's normalmente sejam utilizadas no desenvolvimento tradicional de software, para obter-se uma estrutura *Agile* em BI, esta metodologia é uma regra fundamental.

2.3.2.3 Concepção

As atividades que serão concluídas na fase de concepção da estrutura de BI são a modelagem e o mapeamento. Para realização desta fase irá se usar o produto gerado nas etapas anteriores. Como busca-se uma implementação baseada em desenvolvimento *Agile*, as atividades da fase de concepção também serão executadas de forma iterativa buscando flexibilidade no processo, ao longo de sua execução. Sendo assim, as atividades realizadas anteriormente de *data profiling* e construção dos diagramas de arquitetura servirão como base para a concepção (LARSON e CHANG, 2016).

Nesta fase de concepção, apesar de termos um desenvolvimento iterativo a modelagem irá focar em requisitos priorizados, levando-se em consideração um escopo mais estável, porém ainda permitirá modificações e incrementos.

A atividade de mapeamento irá validar a compreensão de dados e regras de negócios, além das atividades necessárias para manipulação e remoção de inconsistências.

Após isso poderá dar início o desenvolvimento dos processos de ETL assim como das funcionalidades necessárias ao usuário final do sistema. Vale ressaltar que todo esse processo é iterativo e tanto analistas como responsáveis da área de TI deverão trabalhar em conjunto para o aprimoramento do sistema ao longo da fase de concepção.

2.3.2.4 Desenvolvimento

Segundo Larson e Chang (2016), dentro de um ambiente de trabalho Agile a etapa de desenvolvimento é responsável por constantemente estar concebendo softwares e aplicações sob demanda. No caso do BI, possíveis entregáveis para esta fase incluem os próprios processos ETL, visualizações/*dashboards*, aplicações *data mining* e geradores de relatórios. Larson e Chang (2016), afirma que independente do sistema de BI desenvolvido sempre haverá um processo ETL resultante.

Dentro desta fase será melhor entendido a capacidade de entrega de requerimento baseado nas iterações sucessivas inerente ao desenvolvimento. As iterações envolvem, como nas fases anteriores, a colaboração entre área de negócios e TI para entendimento dos requerimentos, geração dos entregáveis, validação e interpretação dos resultados e arquivos gerados (LARSON e CHANG, 2016).

A partir do desenvolvimento deste sistema de BI, será mais fácil a utilização do conceito de *fast analytics*, onde os usuários finais terão maior autonomia para utilizar o sistema e produzir suas próprias análises como análises *ad-hoc*, geração de relatórios e modelos analíticos. Após conclusão desta fase, o sistema estará quase apto a entrar em produção necessitando apenas da fase de testes para validações finais (LARSON e CHANG, 2016).

2.3.2.5 Teste

A partir de uma abordagem *Agile*, os testes irão ocorrer de forma dinâmica pelos *stakeholders* e irão abranger diversas fases citadas anteriormente. Essa metodologia irá garantir confiabilidade nos resultados, pois ao longo do desenvolvimento vários colaboradores, incluindo usuários finais do produto poderão cooperar para verificar os resultados durante o ciclo de vida do projeto de BI, garantindo assim maior qualidade nos resultados.

Para Larson e Chang (2016), devido à natureza complexa dos sistemas de BI, é necessário um controle formalizado de modificações que venham a ser realizadas, isso facilitará a gestão do conhecimento ao longo do desenvolvimento do projeto. Desta forma, esta atividade de testes encontra-se incluída principalmente nas fases de arquitetura, concepção e desenvolvimento do projeto.

2.3.2.6 Produção

A etapa de produção normalmente denomina-se *go-live* do projeto. Nesta etapa, é necessário que exista um suporte pós-produção além de procedimento formalizados de manutenção caso o sistema venha a necessitar, garantindo assim o bom funcionamento do sistema. Cabe ressaltar, que durante essa etapa já não há mais a flexibilidade instaurada ao longo de todo o processo de concepção e desenvolvimento e desse modo, novos incrementos devem ser controlados e analisados cuidadosamente para que não impactem no funcionamento normal da ferramenta. Como atividade final necessária tem-se a necessidade de *feedbacks* por parte dos usuários dos sistemas que provavelmente irão identificar novas necessidades ou falhas a serem corrigidas. A partir desta fase, o sistema já se encontra normalizado e pronto para uso (LARSON e CHANG, 2016).

2.4 Análise por Envoltória de Dados

A Análise por envoltória de dados ou *Data Envelopment Analysis* (DEA) é vista como uma técnica de programação matemática, fazendo parte do conjunto de métodos encontrados na área de Pesquisa Operacional, que possibilita a análise do grau de eficiência de múltiplas unidades produtivas, chamadas de *Decision Making Units* (DMU's), nas quais avalia e compara os insumos com os produtos gerados dentro de um sistema.

Banker (1993) declara que a técnica DEA foi concebida como uma sistemática de programação matemática para análise da eficiência relativa de unidades comparativas com processos de produção similares.

O conceito básico por trás desta técnica está ligado à comparação de eficiências entre unidades semelhantes e suas efetivas operações, desta forma não faz uso de uma abordagem que busca um ideal ou produção ótima.

Segundo Thanassoulis (2001), as eficiências estimadas são comparativas ou relativas porque permitem alterações a uma unidade, tanto relacionadas aos recursos de entrada quanto

aos recursos de saída, quando comparada com outras unidades, em vez de comparar com algum senso absoluto. Da mesma forma, Ferreira Gomes (2009), alerta para o fato de que os modelos DEA referem-se às eficiências relativas do conjunto de dados analisados e não determinam eficiências absolutas.

Cooper *et al*. (2000) cita particularidades da DEA, além de suportarem sua utilização como ferramenta de apoio a decisão e avaliação de cenários *what-if*, sendo capaz de modelar situações diversas encontradas no mercado:

a) dados são valores reais que representam interesses dos analistas e gerentes e os valores devem ser positivos para cada DMU;

b) as unidades de medida das diferentes entradas e saídas não precisam ser congruentes. Exemplos podem envolver número de pessoas, espaços, dinheiro gasto, etc.

c) possui habilidade de identificar fontes e quantidades de ineficiência em cada *input* e *output* de cada entidade;

d) possui capacidade de identificar membros de referência no conjunto eficiente para assim avaliar e identificar fontes de ineficiências.

f) reconhece a probabilidade de que entidades *outliers* não representem somente desvio sem relação ao comportamento médio do conjunto, mas possíveis referências para *benchmark* a serem estudados pelas demais DMU's;

De acordo com Cooper *et al.* (2000), a programação matemática utilizada no modelo é composta basicamente de três elementos básicos:

1. A função-objetivo: Função linear de variáveis de decisão, que deve ser otimizada (maximizada ou minimizada).

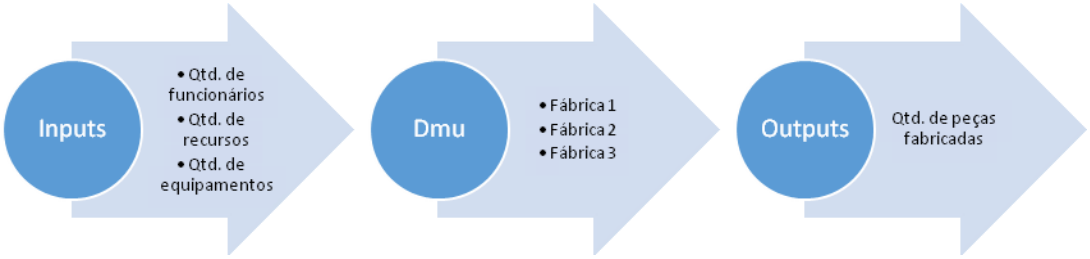
2. Funções Restrições: Tratam das relações de interdependência entre as variáveis de decisão, sendo expressas por um conjunto de equações e/ou inequações lineares.

3. Variáveis do modelo: Deverão assumir valores não-negativos.

É importante notar que os itens que irão compor as aplicações DEA devem ser detalhados, de modo que cada DMU possua os mesmos insumos, referentes aos recursos empregados na produção, e os mesmos produtos, referentes à produção gerada, diferenciando-se em suas quantidades, mas sendo similares em sua natureza (Ferreira Gomes, 2009).

A Figura 6 mostra um exemplo básico de sistema com seus respectivos inputs e outputs.

Figura 6– Exemplo geral de sistema



Fonte: Autor

A metodologia DEA permite, então, uma análise da eficiência de várias DMU's que possuem um ou mais *inputs* e/ou *outputs*, por meio da construção de uma fronteira de eficiência, linear por partes. As entidades que possuírem melhor taxa "*Output/Input*" serão as DMU's consideradas mais eficientes dentro do conjunto analisado e irão estar situadas na fronteira, enquanto as ineficientes estarão situadas em uma região abaixo desta fronteira, denominada de envoltória convexa (SEIFORD, 1999).

A característica desses modelos de poderem considerar ao mesmo tempo vários *inputs* e *outpus*, sem qualquer suposição sobre a distribuição dos dados, pode ser considerada como uma grande vantagem (JI e LEE, 2010).

Embora a técnica DEA tenha sido concebida em período relativamente recente, seu desenvolvimento e aplicabilidade vêm se desenvolvendo rapidamente, contando com uma ampla base teórica e variedade de aplicações práticas como: em economia (LOVELL e PASTOR, 1995), educação (SARRICO, 1997), dentre outras áreas.

2.4.1 Eficiência Técnica

COSTA (2012) afirma que fatores externos à organização devem ser considerados ao se definir o termo eficiência a partir de uma análise macroeconômica, conceituando-a como a propensão a gerar um arranjo institucional que maximize a produção, dado certo estoque de recursos e tecnologia.

No Quadro 3 são apresentados dois possíveis conceitos de eficiência segundo COSTA (2012).

Quadro 3 - Conceitos de eficiência

Tipo de eficiência	Definição

Produtiva	Utilização, com máximo rendimento, da planta produtiva instalada e respectivas tecnologias, de forma que os recursos estejam operando no limite máximo de seus potenciais.
Alocativa	Representa recursos escassos e necessidades ilimitáveis manifestadas pela sociedade que implicam em escolhas para satisfazer alguns desejos ou necessidades. As escolhas implicam em custos de oportunidade.

Fonte: Adaptação Pascal da Costa (2012).

Wilhelm (2000) descreve a eficiência técnica como uma comparação dos níveis de insumos e produtos observados com os níveis de insumos e produtos ideais, sendo assim, seria a taxa entre a produção realizada e a capacidade máxima de produção. Dessa forma, existem duas abordagens que podem ser dadas a eficiência técnica, na qual uma busca aumentar a geração de *outputs* e a outra reduzir os *inputs*.

Para Varian (1992), a eficiência técnica é definida pela relação entre input e output do mesmo sistema e o propósito principal pode ser produzir uma maior quantidade de produtos com a mesma quantidade de insumos ou produzir a mesma quantidade de produtos utilizando uma quantidade menor de insumos.

O critério de eficiência na produção está associado aos conceitos de racionalidade econômica e de produtividade material e revela a capacidade da organização de produzir um máximo de resultados com um mínimo de recursos (BELLONI, 2000).

Debreu (1951), ao determinar o coeficiente de utilização de recursos, concebeu a origem dos indicadores de eficiência produtiva. Orientado para a minimização do consumo de recursos, esse coeficiente consiste na redução proporcional máxima possível em todos os recursos, mantida a produção da mesma quantidade de um produto.

Koopmans (1990) *apud* Lovell (1993), propõe uma definição equivalente à noção de Ótimo de Pareto. Assim, uma unidade é tecnicamente eficiente no sentido Koopmans-Pareto se puder produzir os mesmos produtos reduzindo pelo menos um dos insumos ou se puder usar os mesmos insumos para produzir pelo menos mais um dos produtos. Raciocínio esse semelhante ao citado por Tsoukias (2016).

Farrell (1957) como uma continuação dos trabalhos desenvolvidos propôs uma metodologia para cálculo do indicador de eficiência produtiva de Debreu. No entanto, os cálculos foram restringidos à eficiência produtiva com um único resultado, embora tivesse formulado o problema para o caso com múltiplos resultados.

Farrel (1957) descreve o conceito de eficiência produtiva como sinônima de eficiência técnica e então apresenta um método de medição que se deve a uma tecnologia de produto único. Assumindo vários fatores de produção para um único *output* a rendimentos constante

de escala, Farrel (1957) utiliza como referência uma combinação eficiente de fatores para um dado nível de produto, classificando os desvios em relação à essa combinação como ineficiência.

Sendo assim, cabe ressaltar a forma mais utilizada para quantificar a eficiência, mediante a razão entre a quantidade gerada de produtos e a quantidade utilizada de insumos, conforme ilustra a Equação 1 e considerando os ambientes complexos em que as organizações estão inseridas.

$$\text{Eficiência} = \frac{\Sigma_r u_r y_r}{\Sigma_i v_i x_i} \qquad (1)$$

Na equação, u_r e v_i são pesos, ou seja, o grau de importância que a empresa atribui a quantidades y_r de output r e x_i de input i.

Charnes *et al.* (1978) analisaram os estudos de Farrel (1957) tanto no sentido de trabalhar com múltiplos recursos e múltiplos resultados, quanto na obtenção de um indicador que atendesse ao conceito de eficiência de Koopmans. Essa generalização deu origem a uma técnica de construção de fronteiras de produção e indicadores da eficiência produtiva conhecida como Análise por Envoltória de Dados.

2.4.2 Etapas de aplicação dos modelos DEA

Segundo Meza (1998), na modelagem DEA devem-se seguir 3 etapas para implementar o problema:

a) Definição e seleção de DMU's;

b) seleção das variáveis (inputs e outputs);

c) identificação da orientação do modelo e retornos de escala;

d) identificação e aplicação dos modelos.

2.4.2.1 Seleção de unidades

Para Cooper *et al.* (2000), a primeira observação a ser feita diz respeito à homogeneidade das DMU's. Por DMU's homogêneas entendem-se aquelas que possuem os mesmos insumos, referentes aos recursos empregados na produção, e os mesmos produtos,

referentes à produção gerada, que estejam trabalhando nas mesmas condições de mercado, diferenciando-se em suas quantidades, mas sendo similares em sua natureza.

As entidades escolhidas devem ser suficientemente semelhantes, de modo que faça sentido a comparação entre elas assim como também devem ser suficientemente diferentes, de forma que seja possível discriminá-las (FERREIRA e GOMES, 2009).

Dentro desta etapa, deve-se também definir o número de entidades que será incluído no modelo. Segundo Lins e Meza (2000) o número de entidades contempladas no modelo deve ser, no mínimo, o dobro do número de variáveis utilizadas no modelo, em se tratando de modelos DEA tradicionais. Desta forma, teremos uma quantidade suficientemente grande, de forma que a discriminação entre elas seja possível.

2.4.2.2 Seleção de variáveis

De acordo com Cooper *et al.* (2000), o método possui vantagens no que diz respeito ao uso de múltiplos *inputs* e *outputs*, não sendo necessário ter atenção as unidades de medidas utilizadas, que podem ser das mais variadas possíveis. No entanto, quando se trata da questão de escolha de variáveis que irão compor o modelo, esta seleção deve ser feita com extrema cautela.

Segundo Lins e Meza (2000), quanto maior o número de variáveis em relação ao número de DMU's, mais difícil será o processo de ordenação pelas eficiências, visto a tendência de várias DMU's acabarem sendo posicionadas na fronteira de eficiência. De acordo com Soares de Mello *et al.* (2004), para abordar este problema torna-se necessário buscar metodologias para restringir o número de variáveis usadas no modelo. Desta forma, deve-se definir quais serão as variáveis pertinentes para a análise e quais são dispensáveis, de modo que o modelo continue descrevendo fielmente a realidade.

Dentro deste contexto, Lins e Meza (2000) propõem um método baseado na relação causal entre insumos e produtos. O método I-O *Stepwise*, conduz a modelos com forte relação causal. Complementando esta metodologia, Senra *et al.* (2007) apresenta o método I-O *Stepwise* Exaustivo Completo que se baseia no fato de que existem variáveis que pouco influenciam a eficiência média do modelo, de forma que o modelo não será prejudicado caso estas variáveis venham a ser excluídas.

Para esta metodologia de seleção de variáveis temos o seguinte fluxo de etapas para sua implementação:

1. Calcular a eficiência média de cada par *input-output* possível. Dessa forma, teremos n x m pares. Para cada resultado calcula-se a eficiência média de todas as DMU's;

2. Escolher o par *input* e *output* inicial que gerou a maior eficiência média;

3. Uma vez de posse do par inicial, rodar modelo com mais uma variável, um para cada variável que ainda não foi incluída no modelo;

4. Calcular a eficiência média para cada variável acrescentada;

5. Escolher para entrar no modelo a variável que gerou a maior eficiência média;

6. Verificar se o aumento da eficiência foi significativo. Em caso afirmativo, repetir o passo três. Caso contrário, retirar a última variável incluída e finalizar o processo.

2.4.2.3 Identificação da orientação do modelo e retorno de escala

Para a aplicação da análise por envoltória de dados, os modelos são classificados de acordo com o tipo de superfície envoltória e a sua orientação. Tradicionalmente são possíveis duas orientações distintas: uma a input e outra a output.

Para Mello *et al*. (2004), o *benchmark* das unidades ineficientes é determinado pela projeção destas na fronteira de eficiência. A forma como é feita esta projeção determina a orientação do modelo: orientação a *inputs*, quando a eficiência é atingida por uma redução proporcional de entradas, mantidas as saídas constantes; e orientação a outputs, quando se deseja maximizar os resultados sem diminuir os recursos.

A relação entre *inputs* e *outputs* é denominada retorno de escala. Segundo Mello *et al.* (2004) existem dois tipos básicos de modelos, conhecidos como retorno de escala constante ou CCR (Iniciais de Charnes, Cooper e Rhodes) e retorno de escala variável ou BCC (Iniciais de Banker, Charnes e Cooper).

O modelo CCR tem como propriedade principal a proporcionalidade entre *inputs* e *outputs* na fronteira, ou seja, o aumento (decremento) na quantidade dos *inputs* provocará acréscimo (redução) proporcional no valor dos *outputs*. Já o modelo BCC é invariante a translações a *outputs* quando é orientado a *inputs* e vice-versa, além disso, a DMU que tiver o menor valor de um determinado *input* ou o maior valor de um certo *output* será eficiente (MELLO *et al.*, 2004).

2.4.2.4 Identificação e aplicação do modelo

Segundo Cooper *et al.* (2000), os modelos diferenciam-se em dois pontos principais:

a) Suposições sobre retornos de escala;
b) Projeção do plano ineficiente à fronteira de eficiência.

2.4.2.4.1 Modelo CCR

O modelo CCR representa a metodologia DEA inicialmente proposta por Charnes, Cooper e Rhodes em 1978. O modelo constrói um único *input* e *output* virtual a partir dos dados de entrada e busca ajustar pesos a cada DMU com ajuda de programação linear, buscando maximizar a relação entre *output* virtual e *input* virtual (COOPER *et al.*, 2000).

Esse problema de programação fracionária, mediante alguns artifícios matemáticos, pode ser linearizado e transformado no Problema de Programação Linear apresentado na figura 7. A primeira restrição pode ser definida como o resultado da empresa, pois nada mais é do que a subtração dos produtos (somatório das quantidades produzidas multiplicadas pelos pesos dos produtos) dos insumos (somatório dos insumos consumidos multiplicados pelos respectivos pesos). Ele está limitado a 0. Dessa forma, as empresas eficientes obterão resultado 0. A segunda restrição é o somatório da multiplicação das quantidades consumidas pelos pesos específicos para a DMU k, devendo ser igual a 1. Se a DMU k for eficiente, h_k será igual a 1. Se não for, obterá um indicador sempre inferior a 1 (MEZA *et al.*, 2007).

Os modelos CCR são representados conforme mostra Figura 7.

Figura 7 - Modelagem CCR

Modelo CCR – Orientação *input*	**Modelo CCR – Orientação *output***
Maximizar $h_k = \sum_{r=1}^{s} u_r y_{rk}$ Sujeito a: $\sum_{r=1}^{m} u_r y_{rj} - \sum_{i=1}^{n} v_i x_{ij} \leq 0$ $\sum_{i=1}^{n} v_i x_{ik} = 1$ $u_r, v_i \geq 0$ Considerando: y = *outputs*; x = *inputs*; u, v = pesos; r = 1,..., m; i = 1,..., n; e j = 1,..., N	Minimizar $h_k = \sum_{i=1}^{n} v_i x_{ik}$ Sujeito a: $\sum_{r=1}^{m} u_r y_{rj} - \sum_{i=1}^{n} v_i x_{ij} \leq 0$ $\sum_{r=1}^{m} u_r y_{rk} = 1$ $u_r, v_i \geq 0$ Considerando: y = *outputs*; x = *inputs*; u, v = pesos; r = 1,..., m; i = 1,..., n; e j = 1,..., N

Fonte: Périco *et al.*, (2008)

2.4.2.4.2 Modelo BCC

O modelo BCC, também chamado de VRS (V*ariable Returns to Scale*), considera situações de eficiência de produção com variação de escala e não assume proporcionalidade entre *inputs* e *outputs*.

A formulação do modelo BCC usa para cada DMU o problema de Programação Linear apresentado na Figura 8.

Figura 8– Modelo BCC

Modelo BCC – Orientação *input*	**Modelo BCC – Orientação *output***
Maximizar $\sum_{r=1}^{m} u_r y_{rk} - u_k$ Sujeito a: $\sum_{r=1}^{n} v_i x_{ik} = 1$ $\sum_{r=1}^{m} uryrj - \sum_{i=1}^{n} v_i x_{ij} - u_k \leq 0$ $u_r, v_i \geq 0$ Considerando: y = *outputs*; x = *inputs*; u, v = pesos; r = 1,..., m; i = 1,..., n; e j = 1,..., N	Minimizar $\sum_{i=1}^{n} v_i x_{ki} + v_k$ Sujeito a: $\sum_{r=1}^{m} u_r y_{rk} = 1$ $\sum_{r=1}^{m} u_r y_{jr} - \sum_{i=1}^{n} v_i x_{jr} - v_k \leq 0$ $u_r, v_i \geq 0$ Considerando: y = *outputs*; x = *inputs*; u, v = pesos; r = 1,..., m; i = 1,..., n; e j = 1,..., N

Fonte: PÉRICO et al., 2008

Neste modelo, para a DMU *o* em análise, a eficiência é dada por h_o; x_{ij} representa o input *i* da DMU *k*; y_{jr} representa o output *j* da DMU *k*; v_i e u_r representam os pesos dados aos inputs *i* e aos outputs *j*, respectivamente; u_k é um fator de escala (quando positivo, indica que

a DMU está em região de retornos decrescentes de escala; se negativo, os retornos de escala são crescentes) (MEZA *et al*., 2007).

Segundo Meza *et al*. (2007), em uma linguagem não matemática, para o modelo BCC, uma DMU será considerada eficiente se, na escala em que opera, ela for a mais hábil na utilização de seus insumos. No caso do modelo CCR, uma DMU é eficiente quando apresentar melhor quociente de outputs com relação aos inputs, ou seja, aproveitar melhor os inputs sem considerar a escala de operação da DMU.

2.4.2.5 Limitações da técnica DEA

Na concepção de modelos DEA, deve-se ressaltar possíveis desvantagens do método que podem influenciar na modelagem e interpretação dos resultados:

a) Geração de pesos nulos para variáveis que são consideradas fundamentais na construção do modelo, o que irá gerar um resultado do modelo incompatível com o cenário real.

b) Segundo Cooper *et al.* (2000), a modelagem DEA apresenta limitações com respeito a utilização de valores negativos, impossibilitando em certos cenários a construção de modelos com esses dados. Alguns autores propõem como forma de superar essa limitação avaliar a possibilidade de exclusão das unidades que tenham valores negativos em recursos e produtos, se o número de unidades sob avaliação for grande.

c) Enfatizando a afirmação de Lins e Meza (2000), quanto maior o número de variáveis em relação ao número de DMU's, mais difícil será o processo de ordenação pelas eficiências, visto a tendência de várias DMU's acabarem sendo posicionadas na fronteira de eficiência. Desse modo, deve-se encontrar uma boa relação entre o número de variáveis presentes no modelo e a quantidade de DMU's.

3. ESTUDO DE CASO

O presente capítulo apresenta o estudo de caso que analisa possíveis formas de implementar um sistema de apoio a decisão espacial, iniciando por toda a etapa de processo ETL, passando pela aplicação de modelos analíticos (DEA) até o carregamento dos dados em visualizações de mapas, concebendo uma interface de manuseio ao usuário.

O modelo será testado em um cenário de decisão sobre avaliação dos pontos de vendas passíveis de serem removidos, levando em consideração o nível de significância de seus faturamentos.

3.1 Caracterização da empresa

A empresa Redeinova Tecnologia é a renovação da JMR, empresa fundada em 1996. Atualmente atua no segmento de distribuição de recargas pré-pagas por meio da prestação de serviços em tecnologia da informação.

Possuindo mais de 600.000 mil clientes ativos, abrangendo todo o território nacional e participando em mais de dois bilhões de transações de recarga ao ano, a empresa desenvolve softwares que permitem o levantamento e o controle de dados em toda a cadeia de distribuição de recarga de celulares pré-pagos. Desta forma, ela busca oferecer as operadoras e distribuidoras, seus clientes diretos, serviços que as auxiliem a melhorar a gestão de seus negócios.

Sendo assim, esta empresa opera com um grande volume de dados gerados por cada transação de recarga feita, buscando formas de analisá-los e processá-lo e criando modelos de soluções que possam ser implementadas pelos seus clientes.

A problemática abordada neste projeto derivou-se de uma situação de engenharia reversa, na qual após visto que tipos de dados e objetivos a empresa têm, analisa-se o que poderia ser desenvolvido com estes mesmos dados.

3.2 Etapas do estudo

O estudo foi iniciado com a escolha do tema a ser abordado, o qual consiste em conceber um sistema de apoio a decisão espacial que possa ser utilizado pela companhia para propor serviços de análise de dados georreferenciados relacionado a gestão dos pontos de vendas de distribuidoras de recargas e operadoras.

Com o problema descrito, foi apresentado o referencial teórico sobre Sistemas de Apoio a Decisão, *Big Data*, Projetos de *Business Intelligence*, Análise por Envoltória de dados, mostrando seus principais conceitos e explicitando trabalhos desenvolvidos nas áreas.

É importante ressaltar que visto a necessidade de envolver paralelamente áreas de negócio e desenvolvimento, buscando uma metodologia eficiente de projeto, foi adotado uma estratégia de desenvolvimento *Agile*, metodologia essa explicitada na seção 2.3.2 da revisão de literatura.

Feito isso, foi dado início ao desenvolvimento do sistema de apoio a decisão para problemas georreferenciados, visto que no modelo atual esse tipo de problemática era tratado a partir de esforços manuais e sem muitos recursos para análise espacial.

Desta forma, um novo modelo abrangendo todas as etapas de extração, processamento, tratamento e carregamento dos dados com a utilização da análise por envoltórias de dados para priorização de ações em bairros, foi desenvolvido dentro de um mesmo ambiente. A partir deste, foi proposto uma aplicação para tomada de decisão relacionada a remoção de pontos de vendas, relacionando uma situação atual com a proposta pelo modelo.

Por fim, foram apresentadas as conclusões sobre os resultados obtidos e idéias para projetos futuros no assunto.

O estudo foi dividido, com base na metodologia *Agile* proposta por Larson e Chang (2016) na seção 2.3.2 da revisão de literatura, em 5 etapas descritas a seguir:

1. **Descoberta:** realização das atividades de levantamento e *brainstorming* de idéias, mapeamento das fontes de dados necessárias e o processo de *data profiling*, no qual busca-se uma compreensão da qualidade dos dados e sua estrutura.
2. **Arquitetura:** execução de três atividades principais, começando pelo mapeamento das ferramentas utilizadas, descrição de escopo base do projeto e a prova de conceito.
3. **Concepção/Desenvolvimento:** descrição do modelo criado, dividido pelo processo ETL, aplicação de análise por envoltória de dados e metodologia para remoção de pontos de vendas. Apresentado toda a sistemática utilizada para construção bem como as considerações a serem realizadas.
4. **Teste/Produção:** criação de métricas para avaliação de desempenho da solução e disposição dos *outputs* gerados pelo modelo.
5. **Aplicação Prática:** aplicação visando a redução do número de pontos de vendas. Análise do cenário atual de uma distribuidora de recarga e sugestão de solução a partir do nível de faturamento realizado. Por fim, é comparado os resultados.

Essas etapas serão desenvolvidas nos tópicos a seguir.

3.3 Descoberta

Esta etapa é composta basicamente por 3 atividades: O levantamento e *brainstorming* de idéias, já validado na abordagem inicial, o mapeamento das fontes de dados necessárias e o processo de *data profiling*, no qual busca-se uma melhor compreensão dos dados e sua distribuição através de estatísticas descritivas, tais como: frequência de distribuição, valores máximo e mínimo, campos nulos ou em branco, exceções a valores de domínio, média, mediana, moda, e desvio padrão.

O mapeamento foi realizado com o auxílio do gerente da área, o colaborador Fonseca, que visto o levantamento das demandas realizadas em etapa anterior, foi traçado as informações necessárias para a execução do projeto e então as bases nas quais encontravam-se estas informações.

Sendo assim, as bases utilizadas para o processo resumem-se basicamente à 3 fontes distintas: Interno, com dados relacionados aos pontos de vendas, suas coordenadas geográficas e faturamento diário; Instituto Brasileiro de Geografia e Estatística (IBGE), com dados relacionados a censo demográfico; Portal Fortaleza Dados Abertos, com dados relacionados aos bairros de Fortaleza e seus polígonos para visualização espacial. As bases consideradas são:

- Repositório Geral (Interna): Base da própria empresa que abriga informações das transações diárias realizadas por todos os PDV's, indicando o dia, mês e a receita gerada por cada transação.
- BI PDV (Interna): Esta base também se encontra no domínio da empresa em questão e abrange informações descritivas relacionadas a cada PDV.
- Censo Bairros Fortaleza (IBGE): Esta base concentra várias informações relacionadas a demografia dos bairros de Fortaleza, incluindo divisão por faixa etária.
- Bairros *Shape* (Portal Fortaleza Dados Abertos): A base inclui os polígonos referentes aos bairros de Fortaleza.
- Bairros IDH (Portal Fortaleza Dados Abertos): Base contém informações a respeito dos índices de desenvolvimento humano (IDH) para renda, longevidade e educação. Além do IDH consolidado.

Após mapear as bases que seriam utilizadas, foi realizada a atividade de *data profiling* necessária para entender melhor a qualidade das métricas que compunham as bases. É importante ressaltar que as métricas de análise serão diferentes para os casos onde os campos sejam de caracteres ou de dígitos. As Figuras 9 e 10 correspondem a base Repositório Geral (Interna), as Figuras, 11, 12 e 13 ao BI PDV (Interna).

Figura 9– Campos String em Repositório Geral (Interna)

Record #	FieldName	Average_Length	Count_Blank	Count_Non_Null	Count_Null	Count_Unique	Longest_Length
1	competencia	6.0	0	3258810	0	21	6
2	dia	1.7	0	3258810	0	31	2
3	idpdv	12.6	0	3258810	0	> 10000	13
4	idproduto	2.2	0	3258810	0	27	3

Fonte: Autor

Figura 10–Campos numéricos em Repositório Geral (Interna)

Record #	FieldName	Average	Count_Non_Null	Count_Null	Maximum	Minimum	Percentile25	Percentile50	Percentile75	Standard_Deviation	Variance
1	faturamento	52.6444226573504	3258810	0	20000	-70	14.776	29.085	58.978	85.6108540822072	7329.21833668497
2	quantidade	12.1654005603272	3258810	0	10000	-4	0.776	2.398	4.305	63.1539870166078	3988.42607609387

Fonte: Autor

Figura 11– Campos String em BI PDV (Interna)

Record #	FieldName	Average_Length	Count_Blank	Count_Non_Null	Count_Null	Count_Unique	Longest_Length
1	Bairro	11.1	0	119058	0	254	29
2	BairroGerente	6.3	12892	119058	0	21	18
3	BairroSupervisor	10.2	9183	119058	0	30	21
4	BairroVendedor	10.9	7	119058	0	132	29
5	CEP	8.0	0	119058	0	4059	8
6	CidadeGerente	8.1	12892	119058	0	14	23
7	CidadeSupervisor	8.3	9183	119058	0	16	17
8	CidadeVendedor	8.8	7	119058	0	28	14
9	Distribuidor	13.6	0	119058	0	8	18
10	Endereco	22.7	0	119058	0	> 10000	51
11	EnderecoGerente	13.3	12892	119058	0	22	32
12	EnderecoSupervisor	14.0	9183	119058	0	37	34
13	EnderecoVendedor	16.7	7	119058	0	230	34
14	Filial	16.9	0	119058	0	9	22
15	GeoLocalizacaoPreenchida	3.0	0	119058	0	2	3
16	Gerente	22.8	14	119058	0	27	29
17	IDPDV	12.4	0	119058	0	> 10000	13
18	Supervisor	22.5	14	119058	0	49	39
19	TipoDistribuidor	8.0	0	119058	0	2	8
20	TipoGerente	8.4	0	119058	0	3	10
21	TipoPessoa	1.0	0	119058	0	2	1
22	TipoSupervisor	8.3	0	119058	0	3	10
23	TipoVendedor	7.9	0	119058	0	3	10
24	UfGerente	2.0	0	119058	0	10	2
25	UfSupervisor	2.0	0	119058	0	9	2
26	UfVendedor	2.0	0	119058	0	9	2
27	Vendedor	24.8	0	119058	0	268	43
28	cidade	9.0	0	119058	0	1	9
29	cpfGerente	9.8	12892	119058	0	25	11
30	cpfSupervisor	10.2	9181	119058	0	45	11
31	cpfVendedor	10.7	2996	119058	0	249	11
32	cpf_cnpj	12.2	0	119058	0	> 10000	14
33	ddd	2.0	0	119058	0	2	2
34	email_pdv	5.1	92175	119058	0	1458	50
35	idDistribuidor	2.0	0	119058	0	9	2
36	idDistribuidorFilial	2.4	0	119058	0	9	3
37	idGerente	1.3	0	119058	0	22	6
38	idGerenteComposto	12.4	0	119058	0	26	13
39	idSupervisor	1.7	0	119058	0	33	6
40	idSupervisorComposto	12.4	0	119058	0	48	13
41	idVendedor	3.1	0	119047	11	257	7
42	idVendedorComposto	12.4	0	119058	0	273	13
43	nomefantasia	27.6	0	119058	0	> 10000	62
44	operacao	5.9	0	119058	0	2	6
45	razaosocial	28.3	0	119058	0	> 10000	70
46	regional	3.0	0	119058	0	1	3
47	segmento	55.4	0	119058	0	30	88
48	uf	2.0	0	119058	0	1	2

Fonte: Autor

Figura 12 - Campos numéricos em BI PDV (Interna)

Record #	FieldName	Average	Count_Non_Null	Count_Null	Maximum	Minimum	Percentile25	Percentile50	Percentile75	Standard_Deviation	Variance
1	Latitude	-3.2852404594877	98821	20237	0	-5.9039302	-3.796	-3.754	-3.724	1.2632917059295	1.59590593427027
2	Longitude	-33.5866658676384	98821	20237	0	-39.5276938	-38.577	-38.534	-38.492	12.9043068470443	166.521135202675
3	PDV_Pontuacao	19.0894270019654	119058	0	3502	0	25.044	3502.178	3502.178	84.3801321993822	7120.00670998522

Fonte: Autor

No caso da base em questão, os valores mais importantes a notar são a contagem de nulos (*Count_Null*) para os campos de latitude e longitude que são bastante expressivos, assim como seu valor máximo (*Maximum*) que é igual zero valor este que não faz sentido visto as fronteiras geográficas da cidade de Fortaleza. Outro ponto importante é a quantidade de categorias distintas (*Count_Unique*) para bairros que é bem maior do que a quantidade total de bairros em Fortaleza.

As figuras 13 e 14 estão relacionadas com a base Censo Bairros Fortaleza (IBGE).

Figura 13 - Campos String em Censo Bairros Fortaleza (IBGE)

Record #	FieldName	Average_Length	Count_Blank	Count_Non_Null	Count_Null	Count_Unique	Longest_Length
1	Bairros	11.8	0	119	0	119	24
2	Regional	10.0	0	117	2	7	10

Fonte: Autor

Figura 14 - Campos numéricos em Censo Bairros Fortaleza (IBGE)

Record #	FieldName	Average	Count_Non_Null	Count_Null	Maximum	Minimum	Percentile25	Percentile50	Percentile75	Standard_Deviation	Variance
1	% da população total 2000	0.43859649122807	114	5	3	0	0	0	1	0.652226787737553	0.425399782642447
2	Area 2015	2651976.43	119	0	14425594.9	335140.68	1080365.3	1739312.94	3277151.35	2588417.92390115	6699907348772.76
3	Area Regional 2015	58246719.2770938	117	2	122240258.009999	4856893.22	26633404.63	44504526.5	59408356.8299999	37836117.179914	1.43157176325219e+015
4	Dens. Demografica 2000	96.3157894736842	114	5	266	2	51	95	139	57.697765449511	3329.03213786679
5	Domicílios particulares permanentes 2010	25.7286722689076	119	0	960	1.05	3.275	5.338	8.321	113.501127744499	12882.505999273
6	EMP_COM	921.773109243697	119	0	29903	0	0	191	642	2980.74100156711	8884816.9184233
7	EMP_IND	879.084033613445	119	0	11769	0	0	157	827	1839.40180645512	3383399.00559037
8	EMP_PUB	719.647058823529	119	0	33466	0	0	0	0	3613.70730007953	13058880.4506481
9	EMP_SER	2168	119	0	43659	0	0	273	1165	6178.405231057	38172691.1991525
10	EMP_SET_PR	10.2352941176471	119	0	216	0	0	0	1	34.7454925456997	1207.24925224327
11	M2_COM	138863.537815126	119	0	2104132.4	3382.38	50609.02	91628.08	155787.41	222684.383880994	49588334824.4581
12	M2_IND	8176.69974789916	119	0	80588.26	0	505.85	1987.02	6992.86	15886.8047067105	252390563.78916
13	M2_RES	449898.188991597	119	0	6655010.1	4814	147217.47	288914.79	477596.76	751176.407010461	564265994449.145
14	M2_SER	11698.2276470588	119	0	201915.47	37.2	2477.09	4756.51	10181.52	26138.4639158865	683219295.882103
15	PESO	4837659.79563025	119	0	296953009.96	0	4467.05	107935.7	599858.68	28795948.5788959	829206654558420
16	Populacao 2000	18.7842280701754	114	5	80.303	1.576	8.646	14.952	24.698	14.0872226352763	198.449841575842
17	Populacao 2010	20.6065966386555	119	0	76.044	1.342	10.103	16.405	28.538	14.7731872238585	218.247060751175
18	Populacao 2015	20544.9495798319	119	0	75963	1342	10056	16399	28154	14749.3657449665	217543789.878792
19	Populacao Regional 2015	410833.170940171	117	2	540239	28154	334082	362796	539808	111124.943664102	12348753104.3498
20	VALOR	9740901.28731092	119	0	187689758.45	0	542150.12	2089666.3	7032190.41	25871601.4732387	669339762790086
21	VAL_M2_COM	2325.51907563025	119	0	4612.85	845.88	1780.7	2166.68	2674.41	759.35977772089	576627.272020319
22	VAL_M2_RES	2194.49420168067	119	0	4495.48	860.62	1756.44	2126.39	2504.84	605.0557032579	366092.404044912
23	VAL_M2_SER	2142.77025210084	119	0	5685.83	0	1599.36	2044.18	2715.7	1190.0082450266	1416119.62323129
24	VAL_M2_TER	278.88974789916	119	0	966.51	0	212.97	253.05	302.93	142.115314621387	20196.7626499359

Fonte: Autor

Visto as métricas analisadas, deve-se ter cuidado com os campos que serão utilizados vide a quantidade de nulos, na quinta coluna, que dependendo da variável pode reduzir a quantidade de bairros em 5. As Figuras 15 e 16 estão relacionadas com a base Bairros Shape (Portal Fortaleza Dados Abertos).

Figura 15 - Campos String em base Bairros Shape (Portal Fortaleza Dados Abertos)

Record #	FieldName	Average_Length	Count_Blank	Count_Non_Null	Count_Null	Count_Unique	Longest_Length
1	Bairros	11.8	0	119	0	119	24

Fonte: Autor

Figura 16 - Campos numéricos em Bairros Shape (Portal Fortaleza Dados Abertos)

Record #	FieldName	Average Area (Sq Miles)	Average Length (Miles)	Average Number of Points	Count_Non_Null	Count_Null	Count_Polygon	Largest Area (Sq Miles)	Largest Number of Points	Longest Length (Miles)
1	SpatialObjBairro	1.0	4.6	254.9	117	2	117	5.3326628076208	1368	13.8284754973297

Fonte: Autor

Nesta base, o campo mais importante é a quantidade de nulos, sétima coluna, que irá invalidar o uso daqueles bairros que não possuem polígono preenchido.

1. As Figuras 17 e 18 estão relacionadas com a base Bairros IDH (Portal Fortaleza Dados Abertos).

Figura 17 - Campos String da base Bairros IDH (Portal Fortaleza Dados Abertos)

Record #	FieldName	Average_Length	Count_Blank	Count_Non_Null	Count_Null	Count_Unique	Longest_Length
1	Bairros	11.8	0	119	0	119	24
2	Ranking IDH	3.1	0	117	2	117	4

Fonte: Autor

Figura 18 - Campos numéricos da base Bairros IDH (Portal Fortaleza Dados Abertos)

Record #	FieldName	Average	Count_Non_Null	Count_Null	Maximum	Minimum	Percentile25	Percentile50	Percentile75	Standard_Deviation	Variance
1	IDH	0.381018627333235	117	2	0.953077045299187	0.106724071719799	0.253841670573652	0.34718521159236	0.49977680784526	0.179698400652454	0.03229151519705
2	IDH-Educação	0.948477797006105	117	2	1	0.882591093117409	0.928137651821862	0.956477732793522	0.972672064777328	3.13982608821292e-002	9.85850786422244e-004
3	IDH-Longevidade	0.419113577920449	117	2	1	5.43657331136738e-002	0.233113673805601	0.419275123558484	0.573311367380056	0.228513626823634	5.22184776440908e-002
4	IDH-Renda	0.186251149500491	117	2	1	1.01437195760765e-002	6.88596580991 67e-002	0.109081393682992	0.227428083583556	0.204272864168834	4.17274030357391e-002

Fonte: Autor

Neste caso, é importante notar, assim como anteriormente, que existem alguns valores nulos para bairros, ou seja, alguns bairros não poderiam ser levados em consideração caso utilizada a variável de IDH. Outro fator que chama a atenção é o valor mínimo para o IDH-educação que se encontra bastante elevado, além de possuir baixa variação entre os registros vide o desvio padrão extremamente baixo.

Desta forma, pode-se notar que está etapa de descoberta é bastante útil como ponto de partida para entender como as fontes necessárias para o trabalho podem ser trabalhadas para adquirir um conjunto de dados mais limpo e com uma estrutura rigorosamente adequada para aplicação de modelos analíticos.

3.4 Arquitetura

Esta etapa contemplará a realização de três atividades principais: o mapeamento das ferramentas necessárias para o desenvolvimento do modelo que será proposto, a descrição do

escopo que servirá como base para desenvolvimento do projeto e a chamada prova de conceito que serve como referência para a arquitetura proposta.

3.4.1 Caracterização da arquitetura

No que diz respeito a arquitetura do projeto, o escopo do modelo proposto inicia com as fases relacionadas ao processo ETL. Irá ser realizada a conexão com banco interno da empresa para realizar as extrações necessárias ao desenvolvimento do projeto. Em paralelo, será obtido as bases externas relacionadas aos bairros para enriquecimento do projeto.

Desta forma, após adquiridas todas as bases, será iniciado o tratamento e manipulação das bases de forma a obter um *dataset* limpo para utilização em análises espaciais.

Neste *dataset* irá ser aplicado um modelo de análise por envoltória de dados, para avaliar a eficiência de cada bairro segundo variáveis demográficas e internas a empresa, no intuito de obter uma lista de priorizações de ações que seriam estabelecidas sobre os bairros considerados ineficientes, além de projeções de eficiência para melhorar a performance destes bairros.

Os resultados do modelo anterior serão novamente juntados a base principal, agora possuindo campo relacionado a eficiência dos bairros. A partir desta aplicação será avaliada a distribuição dos pontos por faixas de referência para a eficiência, visualizando as regiões de priorização de ações vide resultado do modelo DEA.

A aplicação prática em um problema de redução de quantidade de pontos de venda servirá como exemplo de ação sobre o produto gerado por esta arquitetura.

3.4.2 Mapeamento de ferramentas necessárias

Para execução do projeto a tecnologia necessária irá envolver primeiramente ferramentas para extração e manipulação e carregamento dos dados. Em momento posterior será necessário também ferramenta capaz de implementar o modelo de análise por envoltória de dados que será aplicado a nível de bairros. Como último passo, o processo irá envolver tecnologias de visualizações de dados para geração de painéis que ajudem a compreender as informações geradas pelo modelo, além de facilitar as análises e visualizações espaciais.

Desta forma, tendo rastreado as necessidades tecnológicas para desenvolvimento do projeto os softwares/ferramentas escolhidas para utilização serão:

- Alteryx Designer: *Software* que será responsável por todo o processo de extração, manipulação e carregamento dos dados que serão utilizados. Além disso, possui conexão com linguagens de programação como R, além de recursos para análise espacial de dados georreferenciados. Esta ferramenta foi escolhida também pela facilidade em ajustar, replicar e automatizar qualquer processo ETL criado, funcionando através de programação visual e tornando o processo intuitivo com a geração de fluxos de trabalho.
- Linguagem R: Esta linguagem foi escolhida para implementação do modelo de análise por envoltória de dados através da biblioteca intitulada "*Benchmarking*", concebida por Peter Bogetoft e Lars Otto. Esta linguagem foi priorizada diante de outras como Python e Java, devido a sua sinergia com outros programas, permitindo dentro do mesmo ambiente do software Alteryx Designer a construção de todo o processo ETL e carregamento direto para o modelo implementado através da linguagem.
- Tableau *Software*: Ferramenta escolhida para visualização de dados não georreferenciados. Apesar de possuir funcionalidades para criação de mapas a partir de coordenadas geográficas e objetos espaciais, apresenta ainda certas deficiências quando comparado com ferramentas próprias para análises espaciais.
- CartoDB: *Software* as a servisse (SaaS) escolhido para visualização de dados georreferenciados. Esta plataforma é utilizada através de processamento em nuvem e possui uma infinidade de recursos para análises de objetos espaciais assim como visualização em mapas extremamente atuais. Possui grande facilidade para implementação de diversas camadas sobrepostas, enriquecendo as visualizações criadas.

3.4.3 Prova de conceito

A aplicação destas ferramentas dentro do contexto de um mini-projeto, faz parte do processo de validação do que foi proposto na seção de arquitetura e servirá como referência para o desenvolvimento mais rigoroso e aprofundado do projeto proposto.

Para a prova de conceito, será avaliada a possibilidade de uso das ferramentas dentro de um processo ETL de menor escala com aplicação de Análise por Envoltória de dados simplificada.

A prova de conceito irá começar pela extração de uma amostra da base relacionada aos pontos de vendas de distribuidoras de recarga, ainda não havendo nesta informação geográfica.

Nessa primeira etapa foi utilizado um total de 20 competências para a primeira tabela consolidada de dados.

Ainda na extração, vale ressaltar que visto a grande quantidade de registros (+10 Milhões) no banco utilizado, foi necessário aplicar certos filtros para agilizar a conclusão da extração. Desta forma, aplicou-se um filtro para o estado apenas do Ceará, mais especificamente na cidade de Fortaleza.

Ao final da extração, gerou-se uma base com 315.000 mil registros e 72 campos distintos. Dentro desta base consolidada, possui dados a respeito de cada PDV, mostrando seu faturamento ao longo de cada competência existente, além de informações relacionadas a sua localização e ao seu proprietário.

Esta base serviu de ponto de partida para análise, visto todas as explicações e esclarecimentos a respeito da forma de negócio da empresa, suas fontes de receitas e seus custos. Como passo inicial, buscou-se formas de agregar estes dados, através de campos que tivessem grande importância dentro do contexto do negócio.

Para isso, buscou-se contato com próprio gerente da área para entender como a empresa julgava a importância dos campos. Por fim, verificou-se que os principais campos para agregação seriam “bairro”, “segmento”, “cep” e “codcliente original”, respectivamente.

As agregações geraram grupos com quantidades bem reduzidas de registros por competência, sendo estas quantidades no mês de agosto/2016: 5158 Cep’s distintos, 94 diferentes segmentos e 115 bairros.

A partir desta reduzida quantidade de categorias distintas, pensou-se em uma forma de poder comparar estes registros e analisar aquelas categorias ou grupos que estariam melhor posicionados diante dos outros, como um modelo de benchmarking.

A utilização da metodologia de Análise por Envoltória de Dados (DEA) é um procedimento que se encaixa de forma interessante nesta problemática. Esta técnica busca analisar a relação entre outputs e inputs, buscando aquelas entidades que conseguiram gerar maior quantidade de produtos com menores insumos.

A partir disso, pode-se ter uma pontuação de eficiência relativa, comparando as diferentes entidades existentes no grupo. Por fim, as entidades consideradas não eficientes, podem ser otimizadas, compreendendo em quanto elas devem aumentar seus produtos ou reduzir tais insumos para se alcançar a fronteira de eficiência.

Para melhor compreender a evolução dos dados, buscou-se primeiramente a aplicação nos grupos agregados de bairros da cidade de Fortaleza. Desta forma, todos os PDV’s foram agregados por bairros, obtendo a quantidade de PDV’s e o faturamento totalde cada bairro.

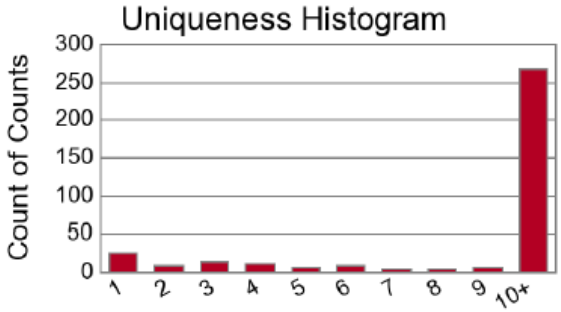
Embora o intuito seja fazer uma prova de conceito em tempo hábil para validar continuidade do projeto, após etapa de extração foram necessárias algumas transformações aplicadas na base com intuito de tratar os dados. Estas manipulações ocorreram ao analisar o campo “bairros” da base estudada.

A figura 19 apresenta resumo do campo “bairros” presente na base.

Figura 19 - Prova de conceito campo bairros sem tratamento

Name: **Bairro**
Type: V_String
Size: 8000
Source: File: C:\Users\leona\Desktop\RecargasCE.yxdb

Count Null: 0
Count Non Null: 149,013
Number of Different Values: 360
Count of Unique Values: 25
Unique Density: 0.0%



Top 10 Values:

Value	%	Count
CENTRO	6.59%	9,816
MESSEJANA	3.35%	4,989
BARRA DO CEARA	2.84%	4,225
PARANGABA	2.52%	3,757
MONDUBIM	2.49%	3,709
JANGURUSSU	2.19%	3,256
ALDEOTA	2.00%	2,977
CONJUNTO CEARA	1.96%	2,926
PASSARE	1.87%	2,781
BOM JARDIM	1.75%	2,615

Fonte: Autor

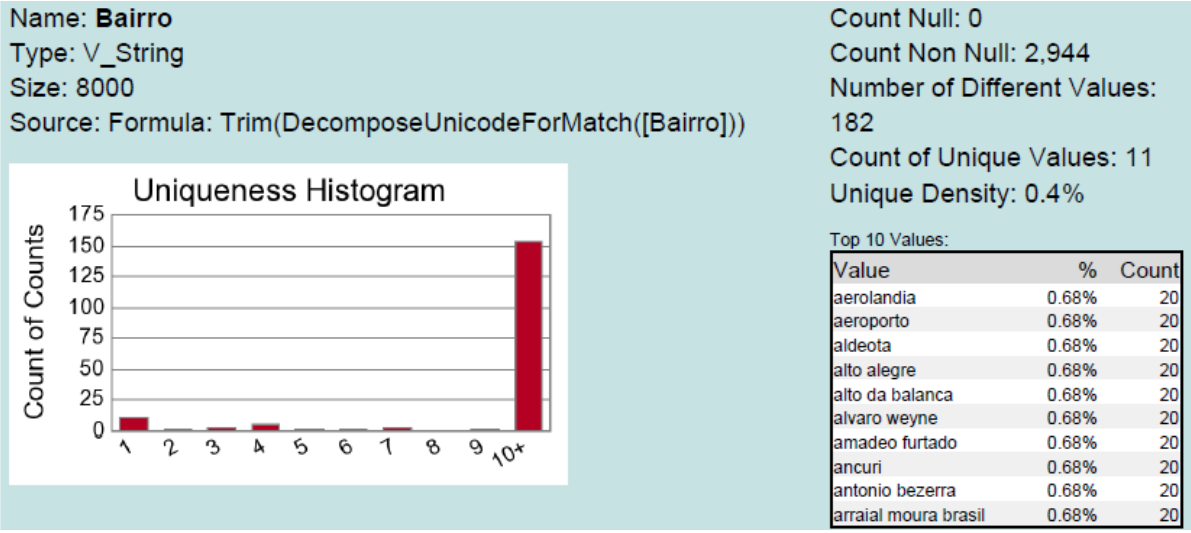
Sabendo que na cidade de Fortaleza, existem ao total 119 bairros cadastrados na prefeitura, observou-se pela figura 19 que a base possuía vários registros preenchidos de forma incorreta elevando a quantidade bairros distintos da base a 360.

Desta forma, para validar uma maior quantidade de registros para análise, a base necessitava de um tratamento neste campo com o intuito de padronizar as nomenclaturas utilizadas. Este tratamento foi realizado através de duas etapas: uma automatizada através de lógica *fuzzy* e a segunda feita de forma manual. Estes tratamentos serão melhores abordados nas etapas de concepção e desenvolvimento.

Depois de aplicados os devidos tratamentos, o número de categorias distintas foi reduzido a 182 e então se utilizou de tabela compondo todos os bairros de fortaleza para filtrar apenas as linhas da base que tivessem nomenclatura igual a um dos bairros existentes.

A Figura 20 apresenta resumo do campo “bairros”, após realizado os devidos tratamentos.

Figura 20 - Prova de conceito campo bairros com tratamento



Value	%	Count
aerolandia	0.68%	20
aeroporto	0.68%	20
aldeota	0.68%	20
alto alegre	0.68%	20
alto da balanca	0.68%	20
alvaro weyne	0.68%	20
amadeo furtado	0.68%	20
ancuri	0.68%	20
antonio bezerra	0.68%	20
arraial moura brasil	0.68%	20

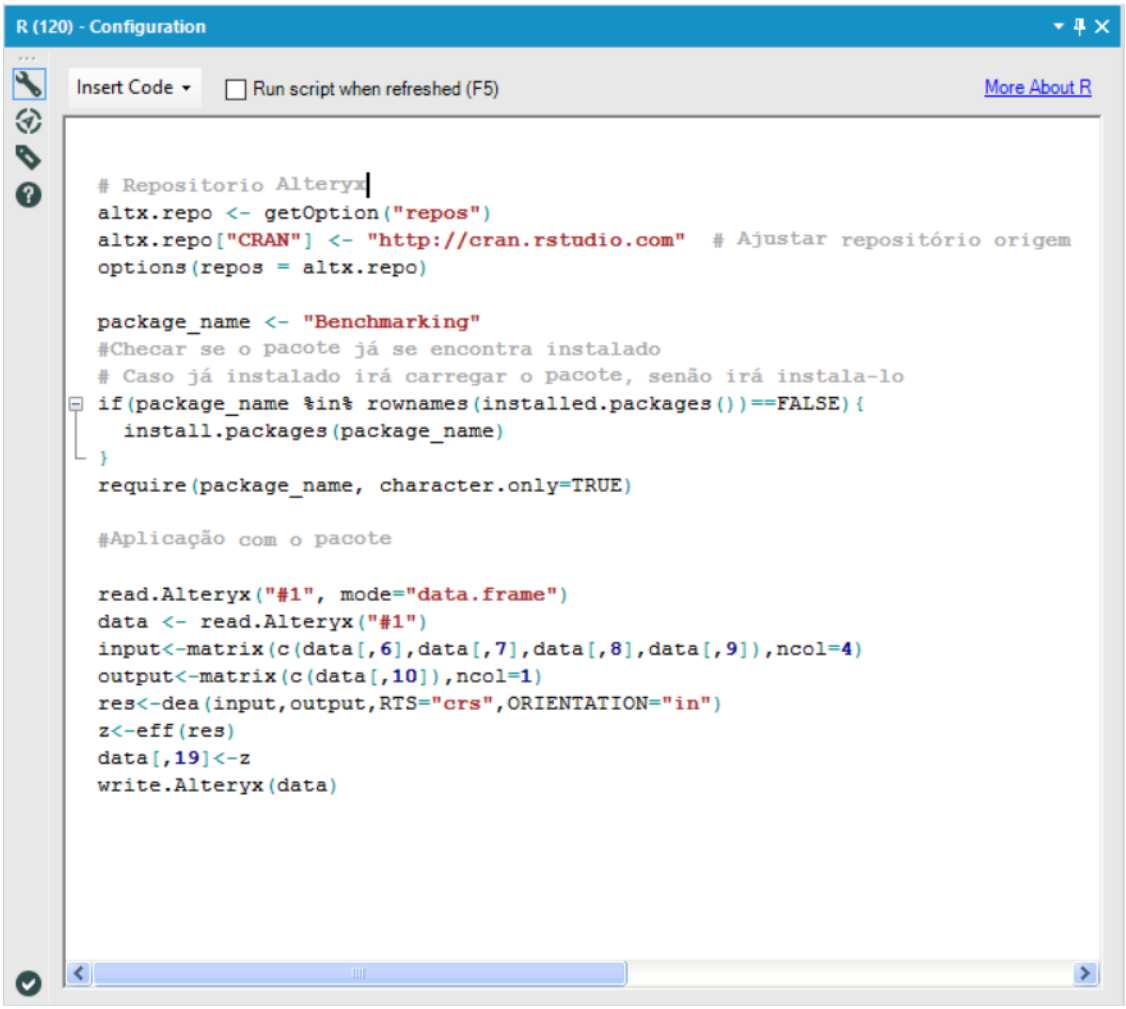
Fonte: Autor

Por fim, a quantidade de bairros reduziu para 116 categorias distintas que abrange basicamente toda a cidade de Fortaleza. Com as nomenclaturas corretas aplicadas aos bairros, tornou-se possível finalizar a etapa de carregamento para utilização da metodologia DEA.

A configuração utilizada foi para aplicação de modelo CCR, com a variável de entrada sendo apenas a quantidade de pontos de venda e a de saída sendo o faturamento agregado do bairro.

A Figura 21 apresenta o código aplicado em linguagem R para utilização do algoritmo de Análise por envoltória de dados aplicados a base em questão.

Figura 21 - Código para implementação simples de análise envoltória por dados



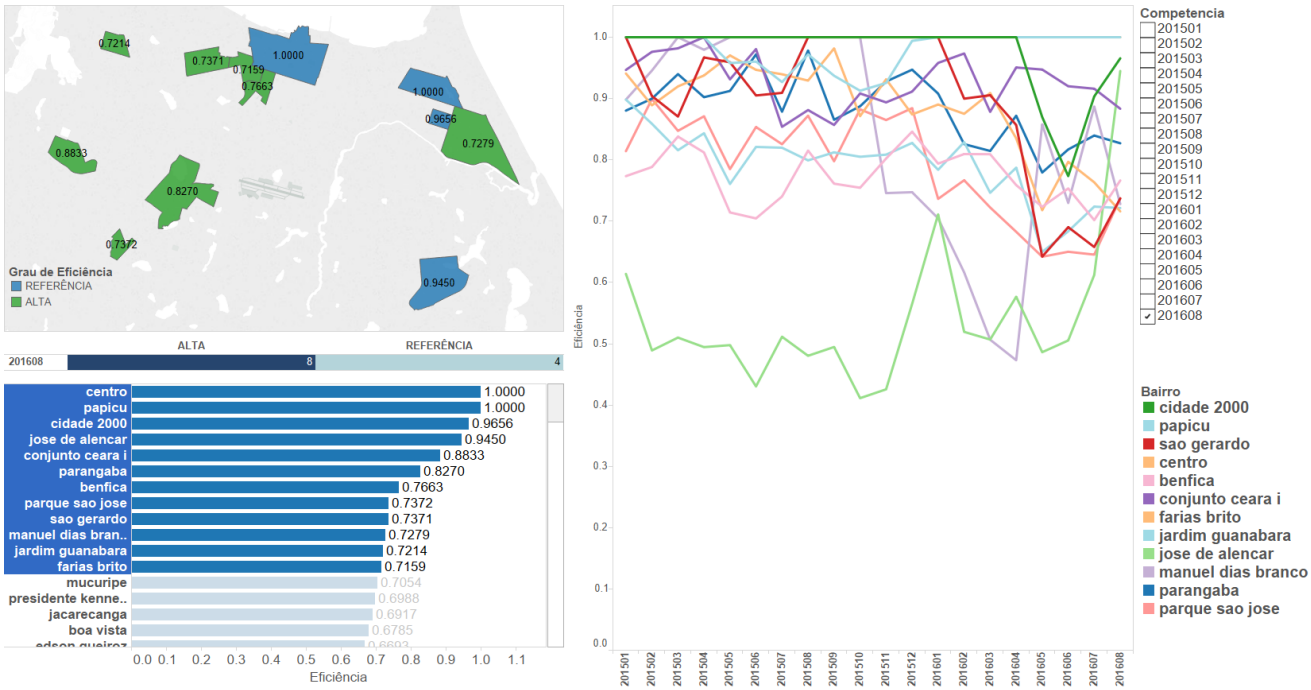

```
# Repositorio Alteryx
altx.repo <- getOption("repos")
altx.repo["CRAN"] <- "http://cran.rstudio.com"  # Ajustar repositório origem
options(repos = altx.repo)

package_name <- "Benchmarking"
#Checar se o pacote já se encontra instalado
# Caso já instalado irá carregar o pacote, senão irá instala-lo
if(package_name %in% rownames(installed.packages())==FALSE){
  install.packages(package_name)
}
require(package_name, character.only=TRUE)

#Aplicação com o pacote

read.Alteryx("#1", mode="data.frame")
data <- read.Alteryx("#1")
input<-matrix(c(data[,6],data[,7],data[,8],data[,9]),ncol=4)
output<-matrix(c(data[,10]),ncol=1)
res<-dea(input,output,RTS="crs",ORIENTATION="in")
z<-eff(res)
data[,19]<-z
write.Alteryx(data)
```

Fonte: Autor

Neste início buscou-se aplicar a técnica DEA separando a base competência a competência, ou seja, mensalmente. No final de todo o processo, foi feita a união de todas as bases processadas mês a mês pelo modelo DEA e gerado um novo arquivo base para análise.

A partir deste arquivo foi criado o *Dashboard* mostrado na figura 22 com o intuito de avaliar a evolução dos níveis de eficiência de cada bairro ao longo dos dois últimos anos.

Figura 22 - Dashboard histórico de eficiência I



Fonte: Autor

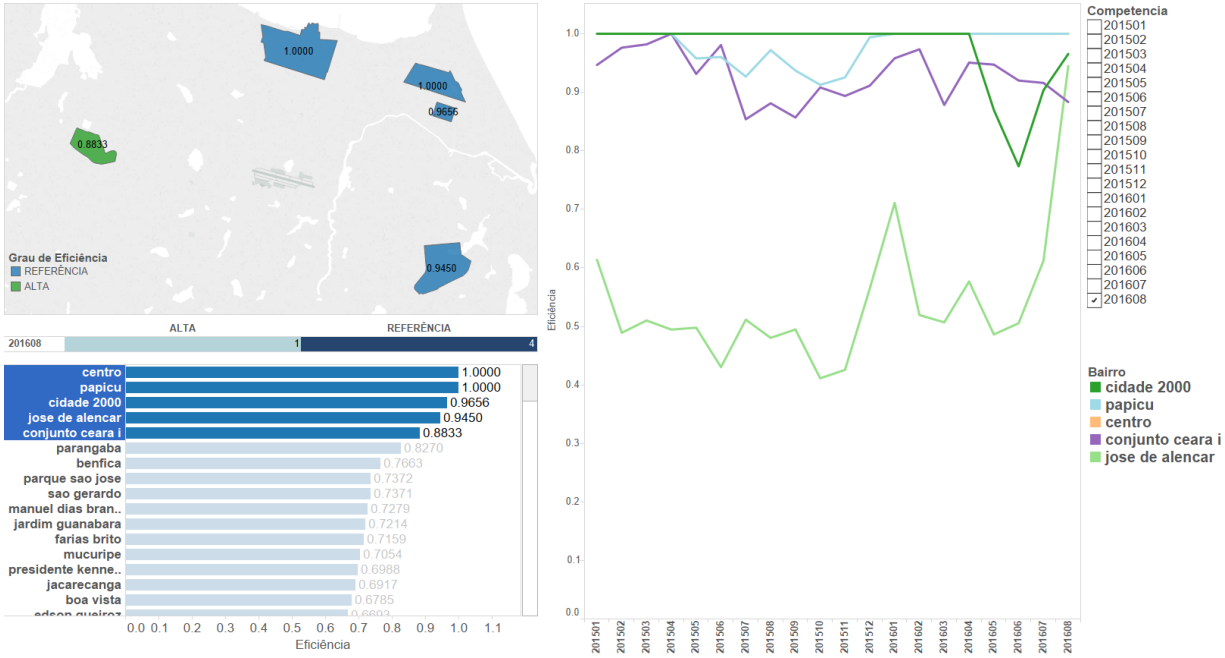
Pela Figura 22, selecionando apenas os 12 bairros mais eficientes no mês de agosto/2016, observa-se que alguns bairros permaneceram várias competências em torno do grau de referência 1.

Porém, nota-se também que vários bairros vêm decaindo em grau de eficiência ao longo dos últimos meses o que deve ser levado em consideração para analisar as causas e diagnosticar os problemas relacionados a esse decaimento.

Por último, têm-se também os bairros que se mantiveram em oscilação como o Conjunto Ceará I, a Parangaba e o Benfica.

A Figura 23 busca analisar se de fato existe aquelas entidades que podem ser consideradas como referência para um *benchmarking* comparativo, estas entidades seriam aquelas que continuam se mantendo com alto grau de eficiência relativa diante dos outros bairros.

Figura 23 - Dashboard histórico de eficiência II



Fonte: Autor

No gráfico em linhas, na figura XX, tem-se uma análise histórica dos 5 bairros mais eficientes no mês de Agosto/2016. Assim, é possível notar que existem dois bairros que conseguem se manter quase que constantemente com grau de eficiência relativa máxima, visto que na maioria das competências sua linha está indicando eficiência igual a 1 pela escala de eficiência no eixo Y do gráfico.Estes bairros são Papicu e Cidade 2000.

Fora estes, existe ainda o centro sendo o único bairro que se mantêm em todas as competências com eficiência relativa máxima igual a 1.

O bairro José de Alencar é um caso interessante, pois apesar de possuir sua eficiência variando entre 0.5-0.6 em grande parte das competências analisadas, nos últimos meses teve forte evolução.

Já o Conjunto Ceará, apesar de obter alto grau de eficiência, possui comportamento extremamente irregular ao longo de seu histórico.

A partir dos resultados obtidos, foi validado pela empresa e pelo autor do trabalho, a real possibilidade de desenvolver o projeto proposto com o auxílio das ferramentas mapeadas,

buscando conceber um sistema de apoio a decisão com uso de Análise por Envoltória de dados para priorização de ações. Dessa forma, desenvolveu-se o objeto de estudo que será tema da próxima seção.

3.5 Concepção/Desenvolvimento

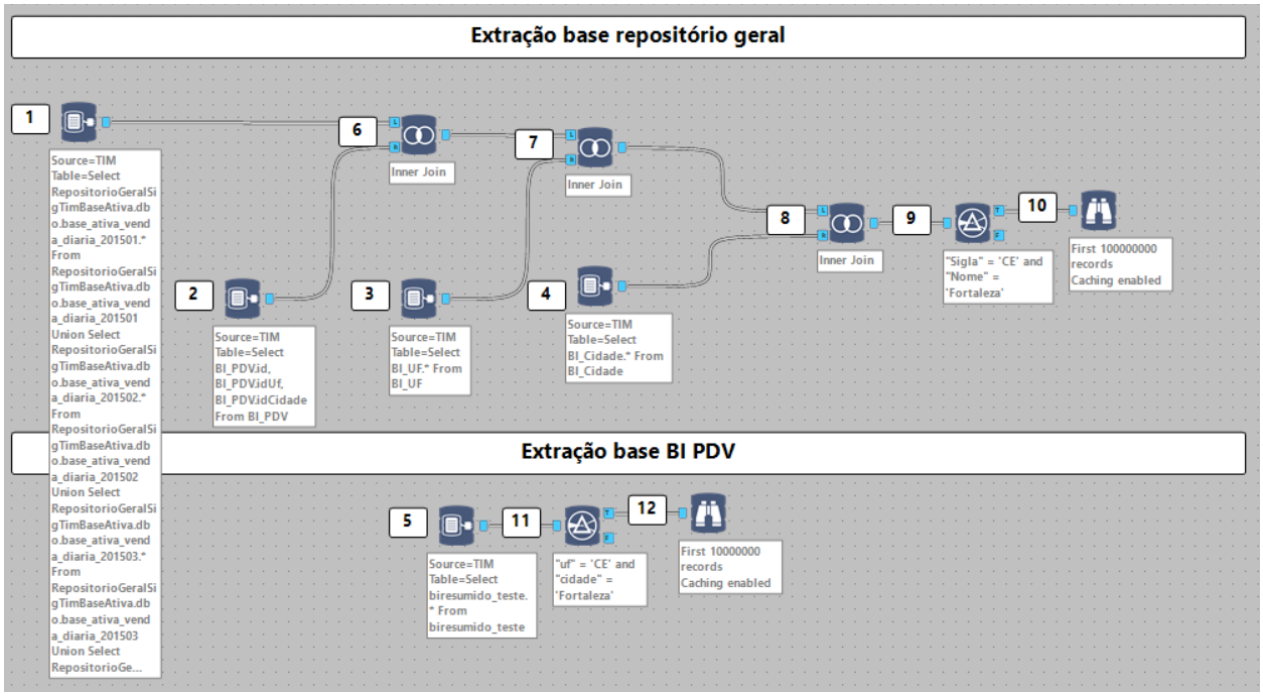
Para o projeto em questão as atividades que irão compor esta etapa são: Desenvolvimento de processo ETL, aplicação da metodologia DEA e criação de macro para análise dos pontos de vendas passíveis de remoção.

3.5.1 Processo ETL

Esta fase se inicia pela extração e aquisição de todas as bases necessárias para desenvolvimento do projeto. Desta forma, foi criado o fluxo abaixo para extração das bases internas da empresa.

A Figura 24 mostra o fluxo criado para aquisição das bases internas da empresa.

Figura 24 - Processo de extração das bases internas



Fonte: Autor

Na figura 24, os passos número 1, 2, 3, 4 e 5 representam os códigos utilizados para acessar as bases internas necessárias para extração.

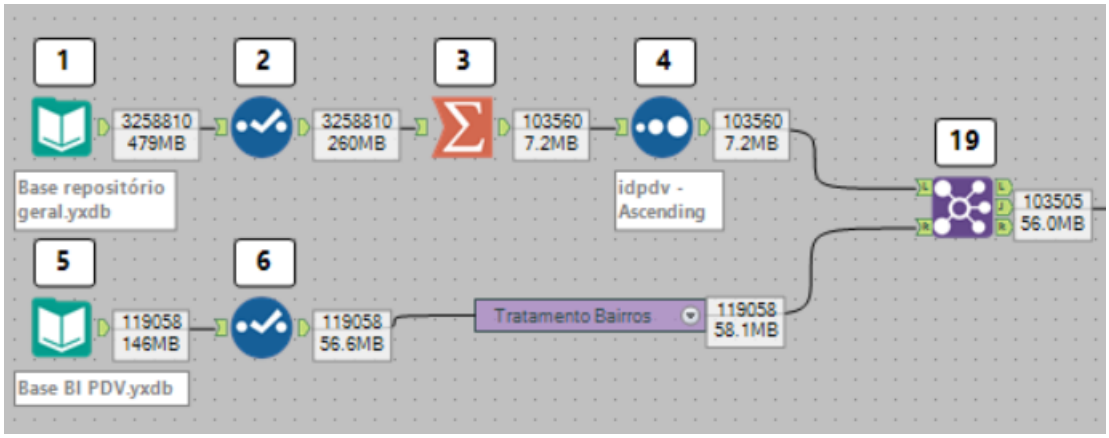
Os passos 6, 7 e 8 são junções e irão consolidar dentro da tabela repositório geral, informações relacionadas a unidade federal e cidade do PDV. Desta forma, será aplicado o passo 9 e 11 para filtrar as bases em questão apenas para a cidade de Fortaleza no estado do Ceará.

A base "repositório geral" inicia na competência de janeiro/2015 e possui um total de 3.258.810 linhas, sendo o output do passo 10, enquanto a base "BI PDV" possui um total de 119.058 linhas, sendo o output do passo 12.

Desta forma, vide a grande quantidade de linhas, assim como a dificuldade de acesso ao sistema que necessitava da rede local da empresa, optou-se por gerar as bases separadamente e trabalhar com elas off-line.

A próxima parte no processo ETL das bases internas é a preparação dos dados para posterior carregamento. A Figura 25 mostra a primeira etapa deste processo.

Figura 25 - Primeira etapa do processo ETL



Fonte: Autor

Na figura 25, os passos 1 e 5 representam a alimentação das bases geradas pelo fluxo mostrado na figura 24. Os passos 2 e 6 irão desmarcar aqueles campos considerados desnecessário para seguir no fluxo, ou seja, aqueles que não serão utilizados para análise.

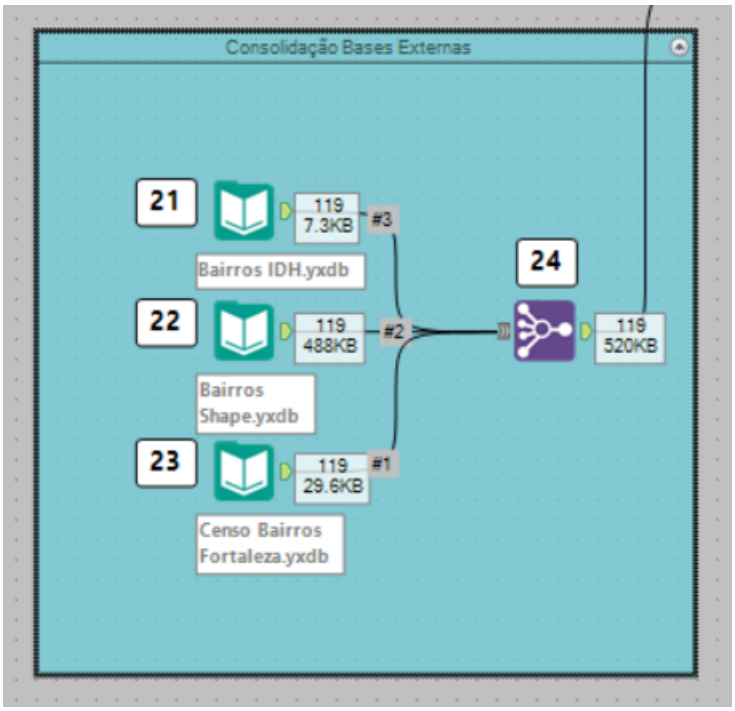
O passo 3 irá agregar os dados da base "repositório geral" por PDV e competência para obter o faturamento de cada PDV por mês e logo após no passo 4 será ordenado a base em ordem crescente de ID. Por fim, no passo 19 será feita uma junção com a base "BI PDV" através dos campos "IDPDV" e "Competência", desta forma teremos consolidado em uma

mesma base as informações descritivas de cada PDV assim como seus respectivos faturamentos mensais.

Assim como nas bases internas da empresa, deve-se realizar a junção das bases externas com informações relacionadas aos bairros que serão adicionadas a base gerada no fluxo anterior, para enriquecimento. Estas informações são necessárias para os modelos DEA que serão aplicados posteriormente.

Desta forma, buscou-se adquirir as bases relacionadas a bairros a partir dos próprios sites do IBGE e do portal de dados de Fortaleza. A Figura 26 apresenta o fluxo utilizado para consolidar estas informações.

Figura 26 - Consolidação bases externas



Fonte: Autor

Os passos 21,22 e 23 representam os inputs das três bases externas com informações relacionadas aos bairros. Na sequência, o passo 24 realiza a junção das três bases através do campo “bairros”.

Após consolidação, a base com informações relacionadas aos bairros está pronta para ser juntada a base interna consolidada, porém, antes de realizar este processo é necessário tratamento das categorias de bairros presentes na base interna, visto que a atividade de *data*

profiling, na etapa de descoberta, indicou 254 categorias distintas dentro da base, enquanto, para a cidade de Fortaleza, existem apenas 119 bairros.

Analisando este campo percebe-se que a base possui vários registros preenchidos de forma incorreta elevando a quantidade de bairros distintos, como mostrado na Figura 27.

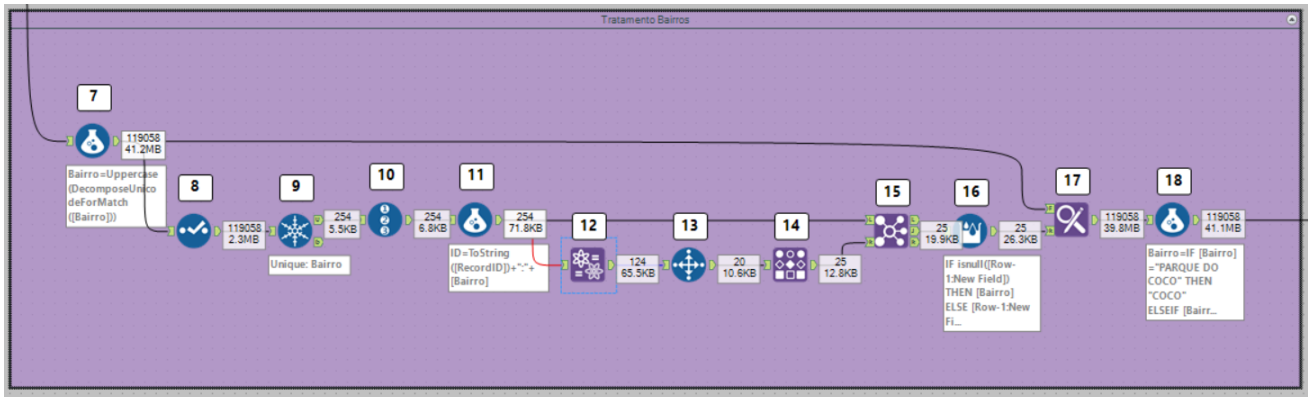
Figura 27 - Exemplo preenchimento incorretos

43	CID. FUNCIONARIOS
44	CIDADE 2000
45	CIDADE DOS FUNCIONAR
46	CIDADE DOS FUNCIONARIOS
47	CIDADE FUNCIONAR
48	CIDADE FUNCIONARIOS

Fonte: Autor

Desta forma, para validar uma maior quantidade de registros para análise, foi realizado um tratamento neste campo com o intuito de padronizar as nomenclaturas utilizadas. A Figura 28 mostra o tratamento realizado através de duas etapas: uma executada através de lógica *fuzzy* e a segunda feita de forma manual.

Figura 28 - Tratamento Bairros



Fonte: Autor

No passo 7, é aplicado uma formula ao campo “Bairro” para retirar qualquer tipo de caractere especial, como acentos, e colocar todas as categorias em letra maiúscula. Já no passo 8 e 9, respectivamente, será selecionado apenas o campo “Bairro” que é objeto de tratamento e obtido todas as categorias distintas existentes na base.

O passo 10 irá criar um campo ID para cada linha que será utilizado no passo 11 para criação de uma chave única de cada registro da base que será composta pelo “RecordID” + “Bairro”.

A lógica *fuzzy* utilizada no tratamento chama-se “Distância de Levenshtein” que mede a quantidade de operações necessárias para realizar a transformação de uma *string* em outra.Na figura 28, esta lógica é aplicada no passo 12.

No passo 13 e 14, são criados grupos de nomenclaturas similares como mostra a Figura 29.

Figura 29 - Exemplo de grupo gerado pela lógica fuzzy

Record #	Group	Key
1	43:CID. FUNCIONARIOS	43:CID. FUNCIONARIOS
2	43:CID. FUNCIONARIOS	46:CIDADE DOS FUNCIONARIOS
3	43:CID. FUNCIONARIOS	48:CIDADE FUNCIONARIOS
4	43:CID. FUNCIONARIOS	45:CIDADE DOS FUNCIONAR
5	43:CID. FUNCIONARIOS	47:CIDADE FUNCIONAR

Fonte: Autor

No exemplo da figura 29, o grupo 43 irá abranger todas as nomenclaturas mostradas pelo campo “*Key*”. No passo 15 será realizado a junção do resultado anterior na base de bairros distintos e no passo 16 será criado um campo com a nomenclatura padronizada do bairro, ou seja, será retirado os ID’s do campo “*Group*”, desta forma teremos na mesma base a nomenclatura antiga e a nomenclatura padronizada.

A partir disso, poderá ser realizado um “de - para”, na base principal, substituindo as categorias por seus valores padronizados. Este processo é executado no passo 17.

O passo 18 representa o tratamento manual. Neste buscou-se entre as categorias distintas substituir manualmente aquelas categorias que não foram processadas pela lógica *fuzzy*, além de padronizar todas as categorias utilizando a base da prefeitura como referência. A Figura 30 mostra exemplo dos tratamentos manuais realizados.

Figura 30 - Exemplo de tratamento manual bairros

```
Expression:
IF [Bairro]="PARQUE DO COCO" THEN "COCO"
ELSEIF [Bairro]="AGUA FRIA" THEN "EDSON QUEIROZ"
ELSEIF [Bairro]="MOURA BRASIL" THEN "ARRAIAL MOURA
BRASIL"
ELSEIF [Bairro]="BOA VISTA / CASTELÃO" THEN "BOA VISTA"
ELSEIF [Bairro]="BOM SUCESSO" THEN "BONSUCESSO"
ELSEIF [Bairro]="CONJUNTO CEARA" THEN "CONJUNTO CEARA I"
ELSEIF [Bairro]="BAIRRO DE FATIMA" THEN "FATIMA"
ELSEIF [Bairro]="MANUEL SATIRO" THEN "MANOEL SATIRO"
ELSEIF [Bairro]="NOSSA SENHORA DAS GRACAS" THEN
"PIRAMBU"
ELSEIF [Bairro]="PLANALTO AIRTON SENA" THEN "PLANALTO
AYRTON SENNA"
ELSEIF [Bairro]="PRAIA  IRACEMA" THEN "PRAIA DE IRACEMA"
ELSEIF [Bairro]="PRAIA DO FUTURO" THEN "PRAIA DO FUTURO
I"
ELSEIF [Bairro]="PREFEITO JOSE WALTER" THEN "PREFEITO
JOSE VALTER"
ELSEIF [Bairro]="SAO JOAO TAUAPE" THEN "SAO JOAO DO
TAUAPE"
ELSEIF [Bairro]="SAPIRANGA" THEN "SAPIRANGA/COITE"
ELSEIF [Bairro]="VILA PERI" THEN "VILA PERY"
ELSEIF [Bairro]="VICENTE PINZON" THEN "VINCENTE PINZON"
ELSEIF [Bairro]="MANOEL DIAS BRANCO" THEN "MANUEL DIAS
BRANCO"
ELSEIF [Bairro]="AMADEU FURTADO" THEN "AMADEO FURTADO"
ELSEIF [Bairro]="CONJUNTO GUARARAPES" THEN "GUARARAPES"
ELSEIF [Bairro]="CACHOEIRINHA" THEN "PADRE ANDRADE "
```

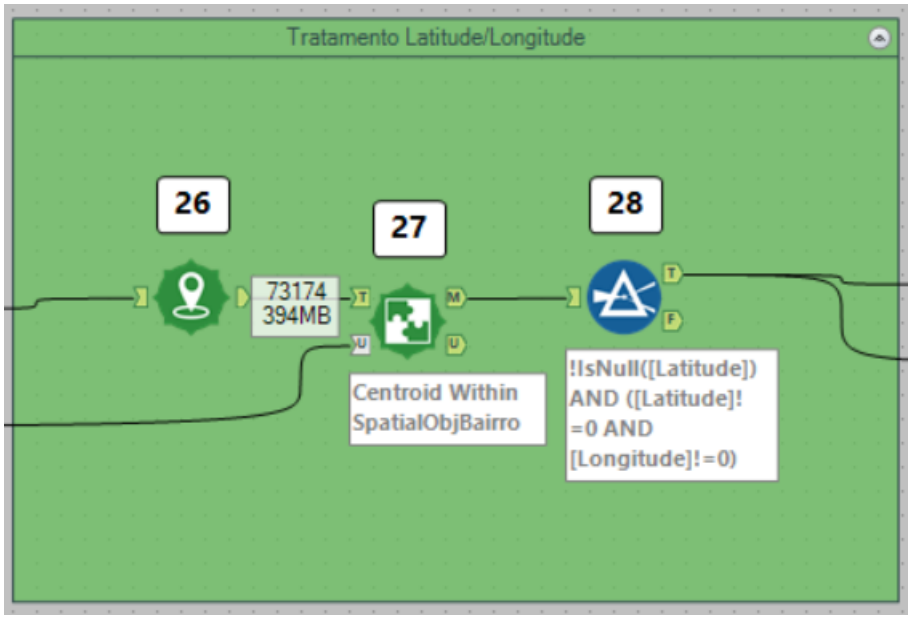
Fonte: Autor

Sendo assim, conseguiu-se chegar a um valor bastante razoável de 131 categorias distintas, sendo o excesso provocado por bairros que não pertencem a Fortaleza, porém foram preenchidos como sendo pertencentes a cidade de forma incorreta.

O último tratamento a ser realizado na base é referente ao campo coordenadas geográficas de cada PDV que, de acordo com a atividade de *data profiling* realizada anteriormente, possuem 20.237 linhas nulas além de coordenadas fora das fronteiras de referência para cidade de Fortaleza.

A Figura 31 mostra a parte responsável por realizar este processo.

Figura 31 - Tratamento campos Latitude/Longitude



Fonte: Autor

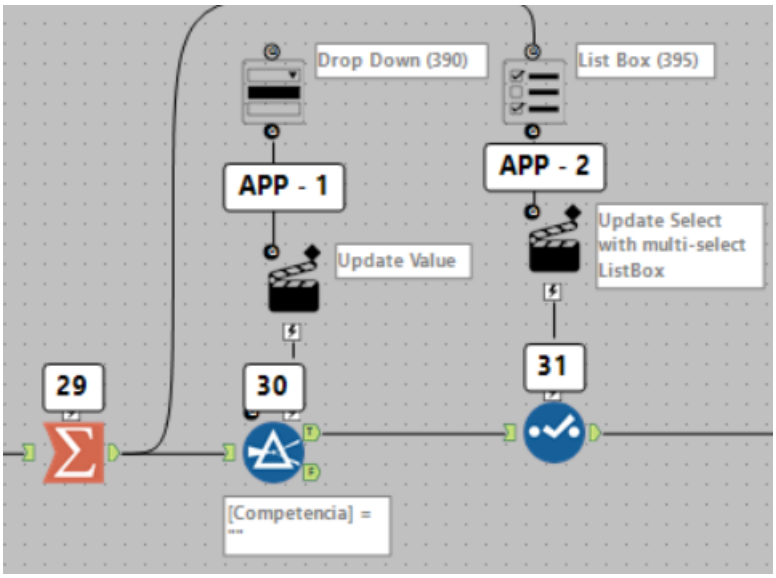
O passo 26 irá criar um objeto espacial a partir dos campos de latitude e longitude existentes na base. No passo 27, cada ponto gerado no passo 26, será conflitado com os objetos espaciais que representam os bairros, desta forma apenas os pontos que estiverem contidos no universo dos bairros serão escolhidos para continuar no fluxo.

Para os registros nulos, optou-se por remover da base estas linhas que não possuíam informações confiáveis, sendo necessário posterior análise juntos as distribuidoras. Este filtro é aplicado pelo passo 28.

A última etapa do processo ETL envolve a agregação dos dados para adquirir a quantidade de PDV's no par competência-bairro e seus respectivos faturamentos. Após isso os dados serão carregados para o modelo de acordo com a competência escolhida pelo usuário.

A Figura 32 mostra a parte final do fluxo ETL.

Figura 32 - Parte final Processo ETL



Fonte: Autor

O passo 29 agrupa os dados primeiramente por competência e depois por bairros. Após isso é realizado uma contagem do campo “idpdv” para adquirir a quantidade de pontos de vendas em cada par competência-bairro. Por fim, realiza-se a soma do campo de faturamento no agregado.

O passo 30 irá selecionar a competência a partir da seleção realizada pelo usuário. Esta seleção é permitida pelo “App - 1” que gera uma interface com as possíveis opções.

O passo 31 permite ao usuário a escolha das variáveis que irão entrar no modelo DEA, escolhendo entre os possíveis campos existentes na base.

3.5.2 Aplicação da metodologia DEA

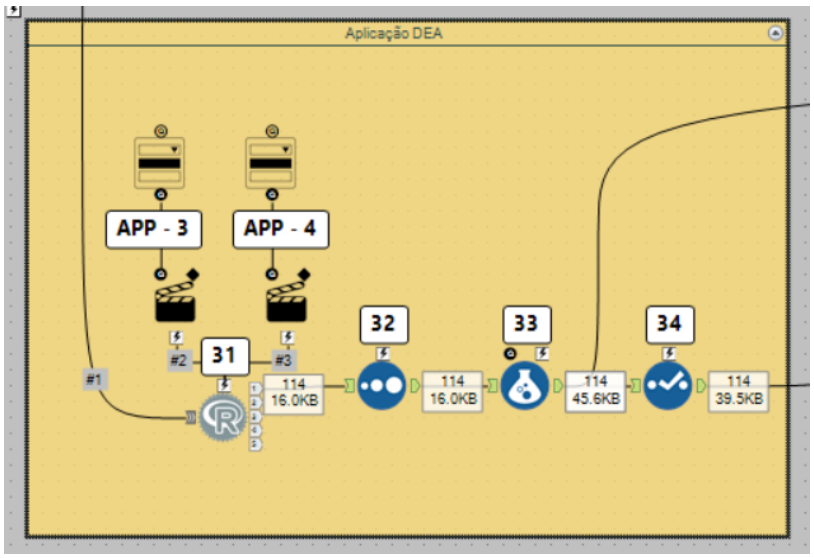
Nesta etapa, as DMU’s da simulação são os próprios bairros de Fortaleza. Espera-se que o modelo ordene os bairros que possuem maior faturamento diante das condições demográficas da localidade e da quantidade de PDV’s.

A implementação do modelo servirá para compreender a fronteira de eficiência e obtendo DMU’sde referência que servirão como benchmarking para analisar como os bairros ineficientes podem melhorar sua performance. Além disto, o ranking de eficiência gerada pode ser utilizado como referência para priorizar os bairros nos quais devem ser tomadas as principais decisões e realizado ações.

Outro valor importante gerado pelo modelo são as projeções de eficiência. Estes valores servirão de referência para analisar o resultado final do projeto, visto que eles representam a quantidade de recursos que se necessita reduzir para que a unidade analisada alcance o patamar de eficiente.

A seguir a Figura 33 mostra a parte final do fluxo relacionada a aplicação da metodologia DEA a base *output* do processo ETL.

Figura 33 - Metodologia DEA



Fonte: Autor

Na figura 33, o passo 31 representa a ferramenta na qual foi utilizado o código em R apresentado na figura 34. Os campos de “APP” permitem criar uma interface para que o usuário possa escolher o tipo de modelo que deseja aplicar (*Constant returns to scale, Variable returns to scale, Decreasing returns to scale, Increasing returns to scale*) e a orientação do modelo (*Input, Output*).

Figura 34 - Código para implementação de modelagem DEA

R (275) - Configuration

Insert Code | Run script when refreshed (F5) | More About R

```
# Alteryx repositorio
altx.repo <- getOption("repos")
altx.repo["CRAN"] <- "http://cran.rstudio.com"  # Setar repositório primário
options(repos = altx.repo)

package_name <- "Benchmarking"

if(package_name %in% rownames(installed.packages())==FALSE){ # Checar se o pacote já se encontra instalado
  install.packages(package_name) # Caso já instalado irá carregar o pacote, senão irá instala-lo
}
require(package_name, character.only=TRUE)
# Aplicação com o pacote
read.Alteryx("#1", mode="data.frame")
data <- read.Alteryx("#1")
# Alimentando matriz R com input e output
x<-matrix(c(data[,3],data[,4],data[,5],data[,6]),ncol=4)
y<-matrix(c(data[,7]),ncol=1)
# Escolha do modelo e orientação a ser utilizado na aplicação DEA
e<-dea(x,y,RTS="modelo",ORIENTATION="orientacao")
z<-eff(e) # Nível de eficiência gerado pelo modelo
ex<-excess(e,x,y) # Projeções de eficiência
# Criação de novos campos com os resultados do modelo DEA
data[,9]<-z
data[,10]<-ex[data[,8]]
data[,11]<-ex[data[,8]+114]
data[,12]<-ex[data[,8]+228]
data[,13]<-ex[data[,8]+342]

# Geração de base output do modelo DEA
write.Alteryx(data)
```

Fonte: Autor

O passo 32 irá ordenar a base em ordem decrescente de eficiência.

No Passo 33, será criado campos percentuais a partir das projeções de eficiência geradas pelo modelo, para entender o percentual necessário de redução para que tal entidade torne-se eficiente. Por fim, o passo 34 irá renomear os campos gerados pelo modelo, segundo a figura 48.

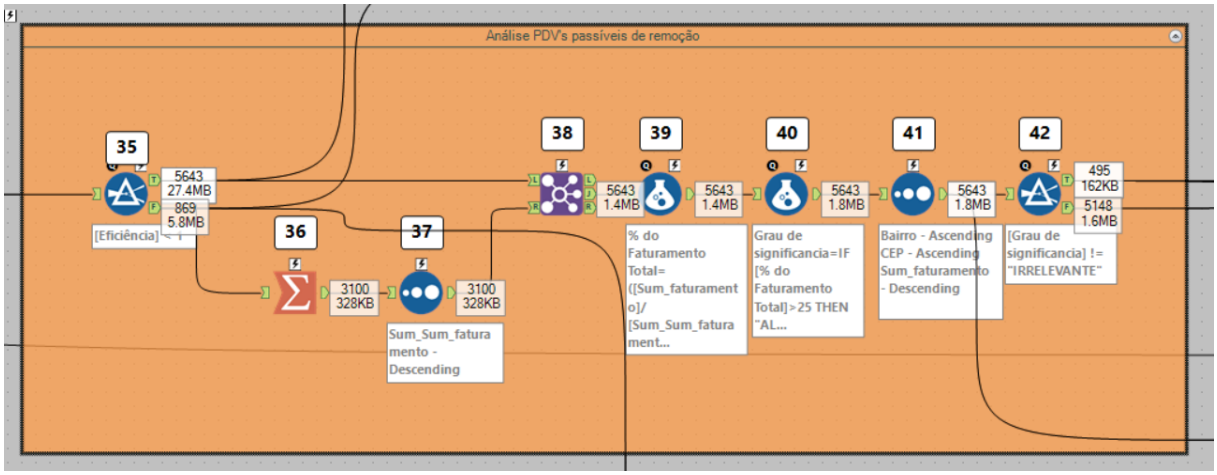
3.5.3 Análise dos pontos de vendas passíveis de remoção

Nesta etapa busca-se a criação de um fluxo para processar os registros existentes na base e avaliar quais pontos estariam passíveis de serem removidos vide o critério de percentual de significância no faturamento total do CEP analisado.

Cabe ressaltar que os bairros que foram considerados eficientes não serão selecionados para análise, pois como apresentado pelo modelo já possuem boa relação entre as variáveis de interesse.

A figura 35 apresenta o fluxo montado para realização desta etapa.

Figura 35 - Análise Pontos de Vendas



Fonte: Autor

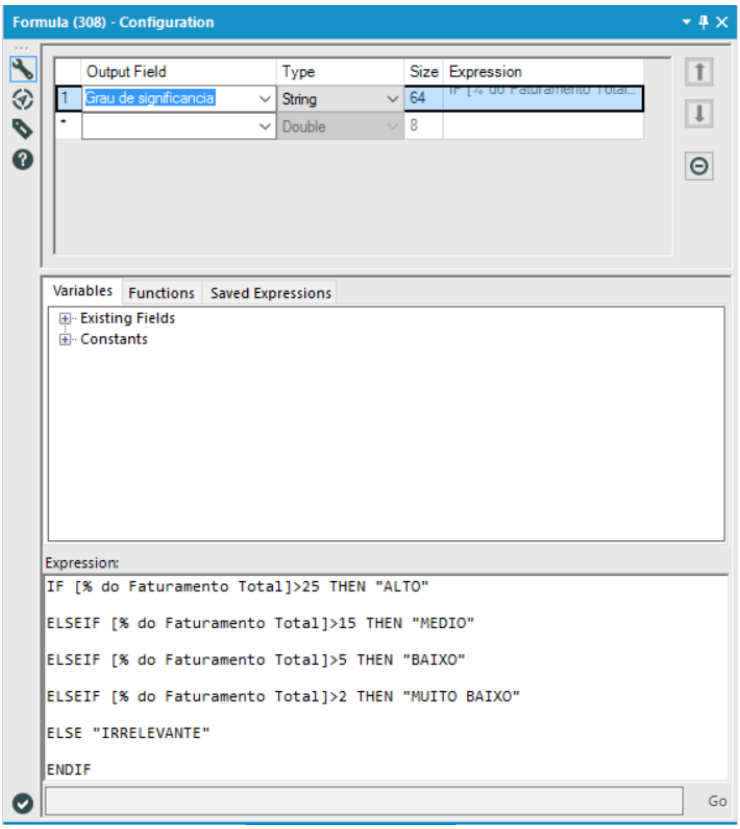
A partir da base consolidada e dos valores resultados da aplicação DEA, todos os PDV's representando bairros considerados eficientes foram filtrados da próxima etapa no passo 35.

No passo 36, será realizado primeiramente uma agregação por bairro e depois por CEP para obter o faturamento total por cada par Bairro-CEP. Logo após o passo 37 irá organizar a base em ordem decrescente de faturamento.

No passo 38 é feito a junção novamente das bases pelos campos Bairro e CEP. O intuito é após esse passo obter na mesma base os campos de faturamento do PDV e faturamento total do CEP em que o PDV se encontra.

O passo 39 irá calcular o campo "% do faturamento total" que servirá como base para classificação dos PDV's. Desta forma, o passo 40 irá categorizar os PDV's de acordo com as faixas apresentadas na figura 36.

Figura 36 - Categorias de faturamento



Fonte: Autor

O filtro aplicado pelo passo 42 irá eliminar todos os pontos de vendas encaixados na faixa abaixo de 2% do faturamento total, considerados de baixa significância vide faturamento total.

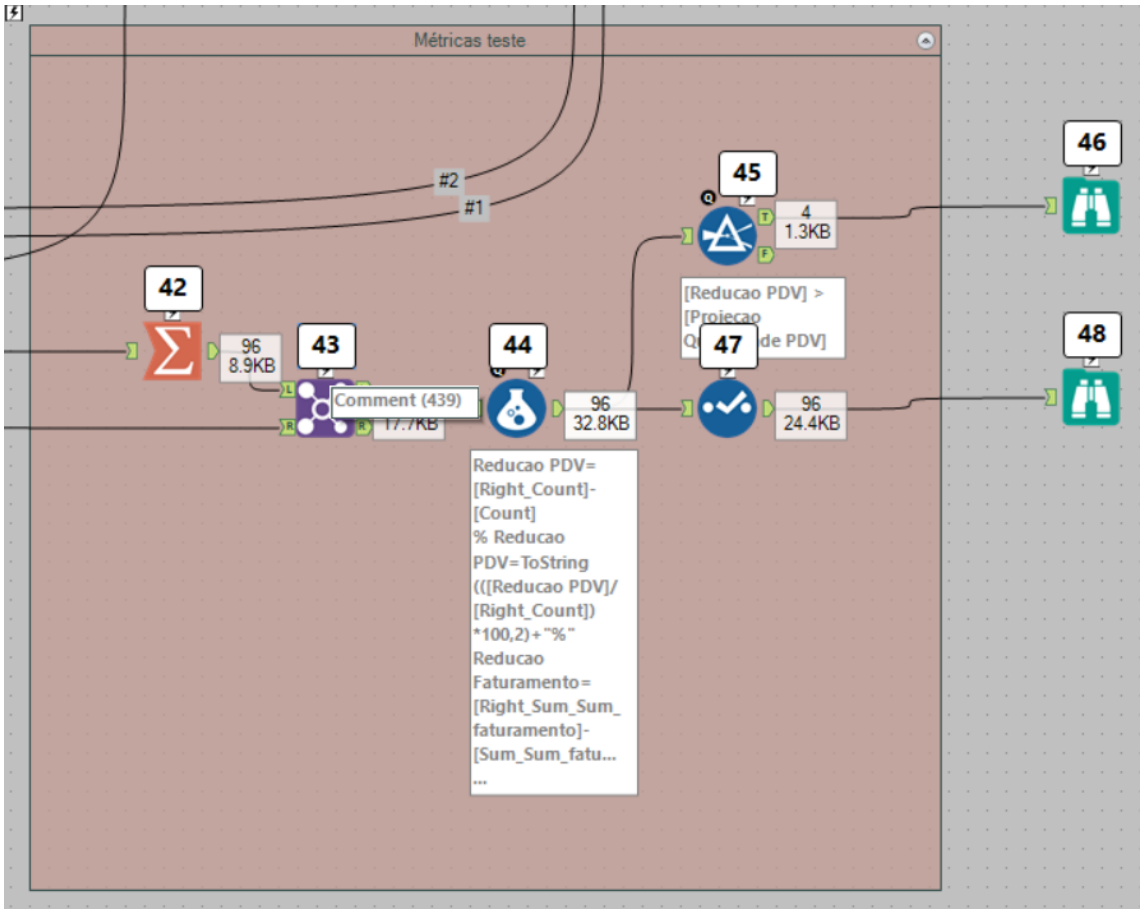
É importante ressaltar que todos estes valores são parâmetros que podem ser modificados de acordo com o cenário específico que uma empresa pode-se encontrar.

3.6 Teste/Produção

Para esta etapa foram criadas algumas métricas para avaliar a eficiência da solução proposta no caso analisado, visando entender as reduções resultantes tanto na quantidade de PDV's por bairro como nos seus respectivos faturamentos.

A figura 37 apresenta a metodologia para cálculo desses valores:

Figura 37 - Métricas de teste



Fonte: Autor

No passo 42, é novamente realizado agregações no nível de bairro para obter a soma do faturamento e a quantidade de PDV's após remoção feita pelo passo 41.

No passo 43, a base é anexada a base gerada após passo 33, com o intuito de buscar as informações de faturamento e quantidade de PDV's por bairro antes de remoção dos pontos, afins de comparação. Esta junção é feita pelo campo de "Bairro".

No caso do problema supracitado, algumas métricas foram calculadas para avaliar o desempenho da solução gerada. Estas métricas foram geradas pelo passo 44 e estão apresentadas no Quadro 4.

Quadro 4 - Definição de métricas teste

Métricas	Definição de cálculo

Redução PDV	Representa a redução da quantidade de PDV's em valor absoluto. Calculada reduzindo-se o valor inicial do valor obtido após aplicação do fluxo completo
% Redução PDV	Representa o percentual de redução da quantidade de PDV's. Calculada pela relação entre a métrica "Redução PDV" e o valor inicial de quantidade de PDV's.
Redução Faturamento	Representa a redução do faturamento em valor absoluto. Calculado reduzindo-se o valor inicial de faturamento do valor obtido após aplicação do fluxo completo.
% Redução Faturamento	Representa o percentual de redução do faturamento. Calculado pela relação entre a métrica "Redução faturamento" e o valor inicial de faturamento.
Gap PDV	Representa o valor faltante de PDV's a serem reduzidos para chegar ao valor de referência obtido através do resultado da aplicação DEA. Calculado pela subtração da projeção de eficiência PDV e a quantidade de PDV's removidos pelo fluxo.

Fonte: Autor

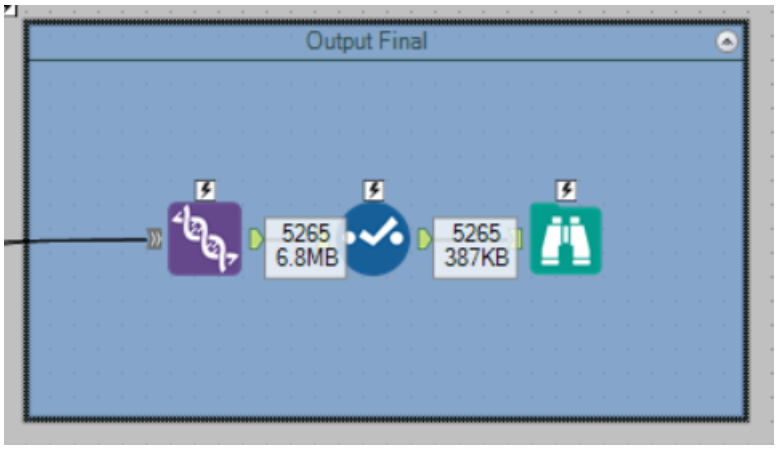
Por fim, os *outputs* 46 e 48 são respectivamente a lista dos PDV's que atenderam a quantidade referência gerada pelo modelo DEA e a tabela-resultado das métricas calculadas. O passo 46 é calculado através do filtro aplicado pelo passo 45 para bairros com quantidade reduzida maior do que a projeção de eficiência.

Como fase final foram adicionados importantes *outputs* para análise desta situação e tomada de decisão, de forma a visualizar os pontos de venda a partir de um mapa. Estes são descritos a seguir.

- Output Geral: mapa dos pontos de vendas antes de qualquer manipulação. Situação inicial.
- Output Eficientes: Apenas os pontos de vendas pertencentes aos bairros considerados eficientes pelo modelo.
- Output Ineficientes: Apenas os pontos de vendas pertencentes aos bairros considerados ineficientes pelo modelo.
- Output Removidos: Todos os pontos que foram considerados como irrelevante pelo modelo.
- Output Final: mapa dos pontos de vendas após aplicado o fluxo por completo. Situação final.

A Figura 38 apresenta exemplo para o caso do *output* final. Nos outros casos, os *outputs* irão seguir mesma estrutura e o último passo da Figura 38, gera o mapa para interação com o usuário.

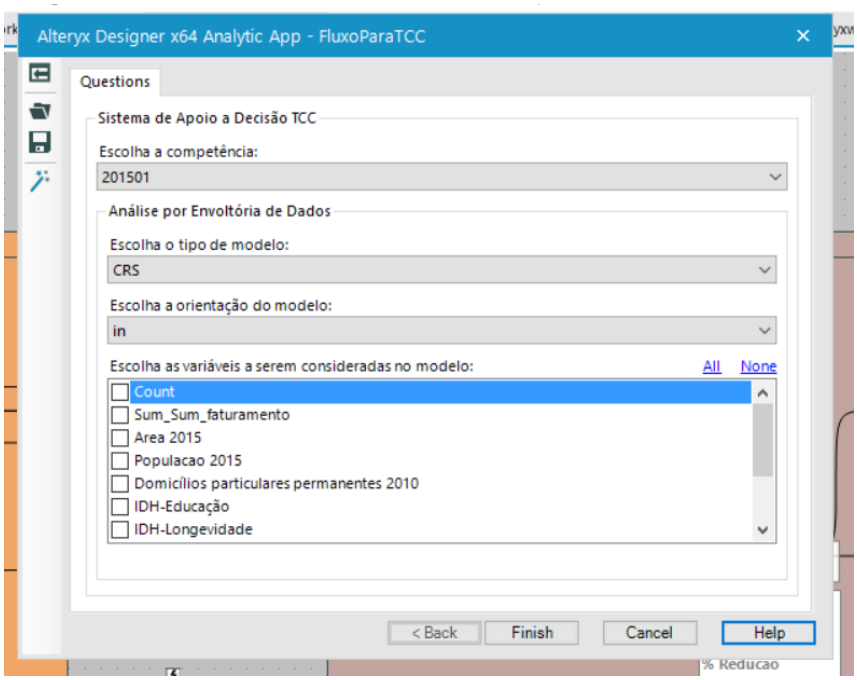
Figura 38 - Exemplo estrutura de output



Fonte: Autor

Por fim, é mostrado na Figura 39 a interface da aplicação para selecionar parâmetros relativos ao modelo.

Figura 39 - Interface usuário



Fonte: Autor

3.7 Aplicação prática

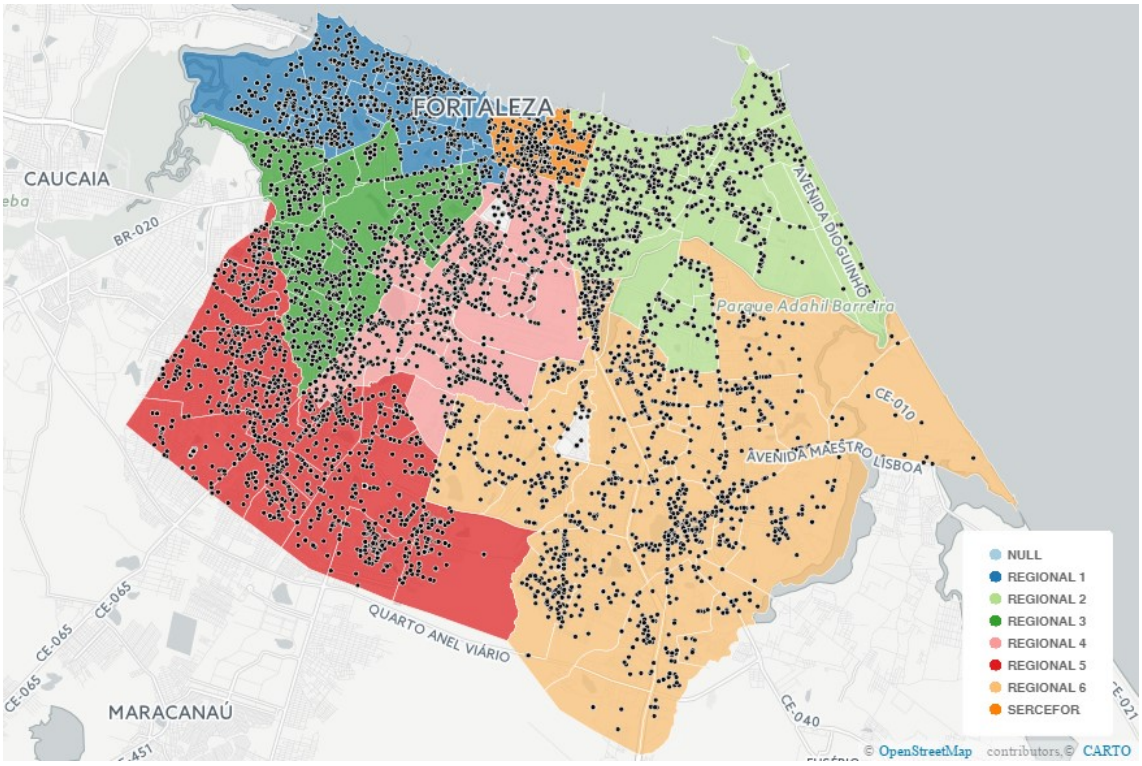
Para aplicação prática, escolheu a base mensal de Janeiro de 2015 pois apresentava maior quantidade de registros com dados georreferenciados. Também foi adotado apenas análise de uma das distribuidoras, sendo escolhida aquela com maior representatividade em ponto de vendas.

3.7.1 Situação atual

Atualmente, existe uma demanda de pontos de vendas que é exigida pela operadora. Essa demanda é relacionada com a quantidade de habitantes e no caso da operadora em questão, é exigido dos distribuidores que exista 1 ponto de recarga para cada grupo de 1.000 pessoas. Desta forma, passa-se as distribuidoras a tarefa de encontrar as melhores localizações para seus recursos.

A Figura 40, mostra como estão distribuídos os diversos pontos de vendas da base em questão analisada. O gráfico em questão foi gerado pelo *output* Geral.

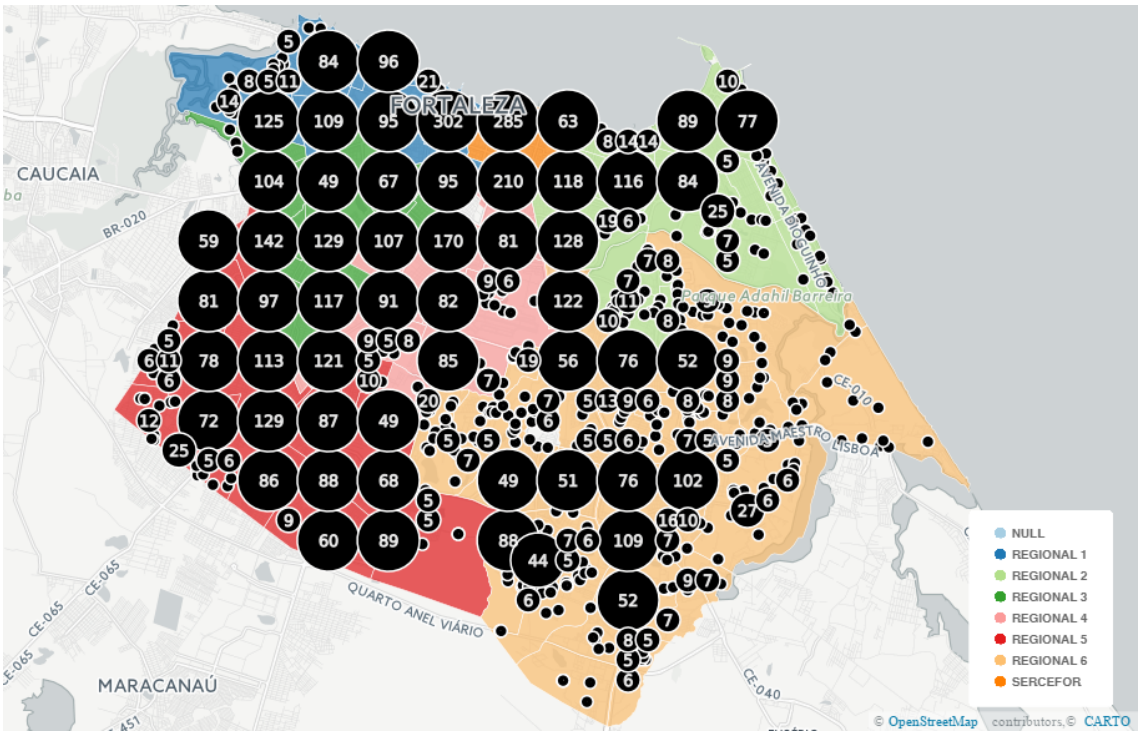
Figura 40 - Situação Janeiro/2015



Fonte: Autor

A criação de “clusters” pode facilitar o entendimento das concentrações de pontos de vendas. Esta análise é apresentada na figura 41.

Figura 41 - Clusters Janeiro/2015



Fonte: Autor

Desta forma, observa-se pela Figura 41 que existe uma imensa concentração de Ponto de Vendas (PDV’s) nas regiões 3, 5 e no centro da cidade de Fortaleza. Sendo os pontos de maiores densidades vistos o tamanho dos grupos agregados encontrados.

3.7.2 Situação proposta

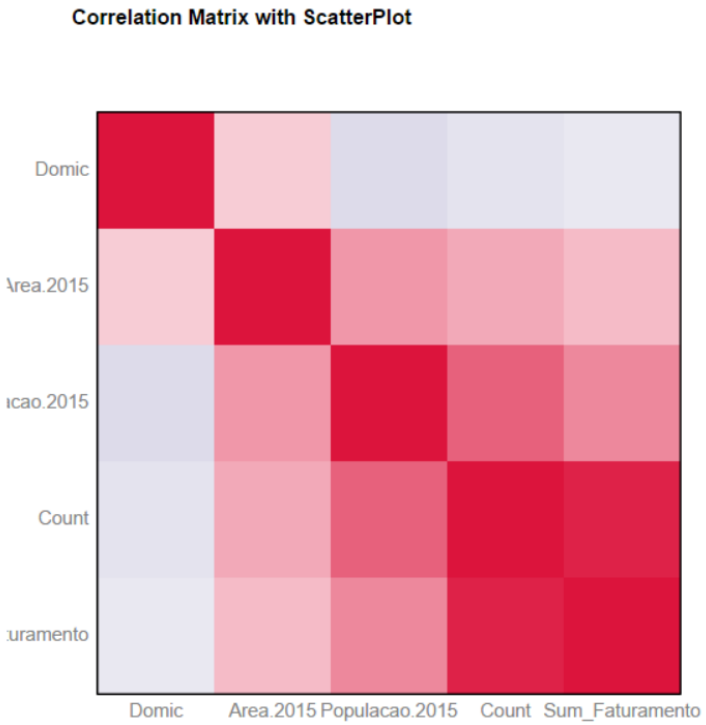
Para seleção das variáveis que iram compor o modelo, foi realizado discussão com o gerente da área que a partir dos dados já adquiridos em bases, procurou-se aquelas que mais tinham influência no faturamento agregado mensal.

Desta forma, entre as variáveis existentes no parâmetro anteriormente criado optou-se por 4 variáveis referentes aos bairros analisados, sendo elas: quantidade de pontos de vendas, população, área e quantidade de domicílios existentes.

A análise das 4 variáveis escolhidas indica claramente que um aumento dos seus valores vai de conflito com os objetivos das distribuidoras. Desta forma, estas quatro variáveis devem ser consideradas como input enquanto o faturamento será o único output.

Afim de validar a classificação sugerida com o auxílio do gestor da área, as Figuras 42 e 43 mostram os índices de correlação Pearson destas variáveis.

Figura 42 - Matriz correlação Pearson



Fonte: Autor

Figura 43 - Planilha correlação Pearson

Focused Analysis on Field Sum_Faturamento

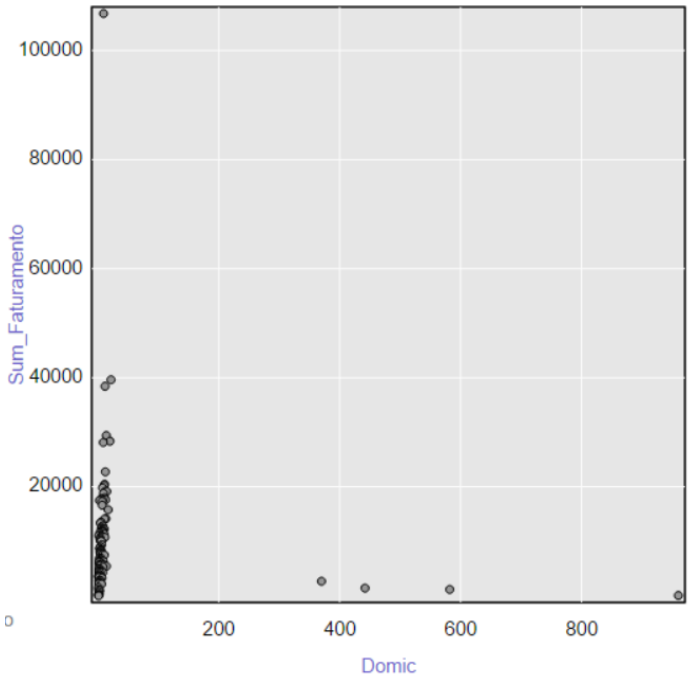
	Association Measure	p-value	
Count	0.93989	0.0000e+00	***
Populacao.2015	0.50818	5.7676e-09	***
Area.2015	0.28861	1.6786e-03	**
Domicílios.particulares.permanentes.2010	-0.12045	1.9775e-01	

Fonte: Autor

Pela análise acima, vemos que apenas a variável de domicílios particulares apresentou correlação negativa e baixo nível de significância com a variável faturamento, quando aplicada metodologia Pearson.

No entanto, o gráfico de dispersão da figura 44 mostra que a relação entre essas duas variáveis apresenta comportamento muito diferente do linear, que é a base de análise para esse tipo de correlação.

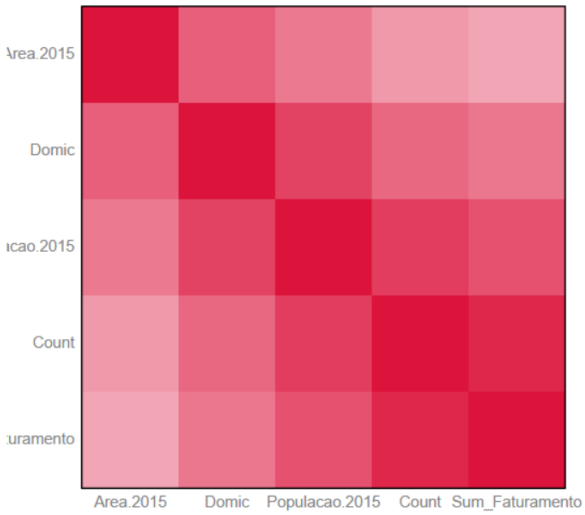
Figura 44 - Faturamento x Quantidade de domicílios



Fonte: Autor

Por este motivo, é interessante analisar os índices de correlação *Spearman* que representam melhor o comportamento sugerido pelo gráfico XX. Esta análise é mostrada nas figuras 45 e 46.

Figura 45 - Matriz correlação Spearman



Fonte: Autor

Figura 46 - Planilha correlação Spearman

Focused Analysis on Field Sum_Faturamento

	Association Measure	p-value	
Count	0.92195	0.0000e+00	***
Populacao.2015	0.74131	0.0000e+00	***
Domicílios.particulares.permanentes.2010	0.57764	1.1323e-11	***
Area.2015	0.37962	2.6367e-05	***

Fonte: Autor

Desta forma, observa-se que de fato todas estas variáveis apresentam forte nível de influência no faturamento mensal, o que também pode ser visto pelo *p-value* muito próximo de zero para cada uma delas.

Outro fator importante é a homogeneidade dos resultados. A Figura47 analisa o coeficiente de variação (CV), dado pela razão entre o desvio padrão e a média dos dados, próprio para comparações entre variáveis distintas.

Figura 47 - Coeficiente de variação

Record #	FieldName	Average	Percentile25	Percentile50	Percentile75	Standard_Deviation	CV
1	Area 2015	2691091.131897	1080365.3	1739312.94	3375244.24	2608780.383404	96.94%
2	Count	22.560345	10	16	30	21.320857	94.51%
3	Domicílios particulares permanentes 2010	26.30856	3.289	5.338	8.629	114.91284	436.79%
4	Populacao 2015	20787.775862	10148	16399	28717	14792.077369	71.16%
5	Sum_Faturamento	10254.258621	4371	7654	12332	11712.806321	114.22%

Fonte: Autor

Analisando a Figura 47, depreende-se que há preponderância de variáveis com alto valor de CV, ou seja, há preponderância de dados heterogêneos. Esta heterogeneidade é característica essencial para que faça sentido a comparação entre as DMU's escolhidas tornando possível a discriminação entre elas.

Para aplicação, o modelo BCC foi considerado o modelo mais propício visto que acréscimos nos *inputs*, podem promover ou não acréscimos no output, e esta relação não é proporcional. Da mesma forma, ele se adequa bem à heterogeneidade da amostra e ao porte relativamente não uniforme das unidades analisadas.

Foi adotada a orientação a inputs, uma vez que se busca a minimização das variáveis que impactam o custo da empresa, ressaltando que a única que pode ser de fato reduzida é a variável de quantidade de PDV's, pois faz parte das operações internas a empresa.

Após decisão sobre as variáveis e configurações a serem adotadas no modelo DEA, basta rodar o fluxo e analisar os resultados obtidos.

Na figura 48, temos o resultado do modelo DEA aplicado a competência de Janeiro/2015, para mostrar os campos gerados pelo modelo.

Figura 48 - Resultado aplicação DEA

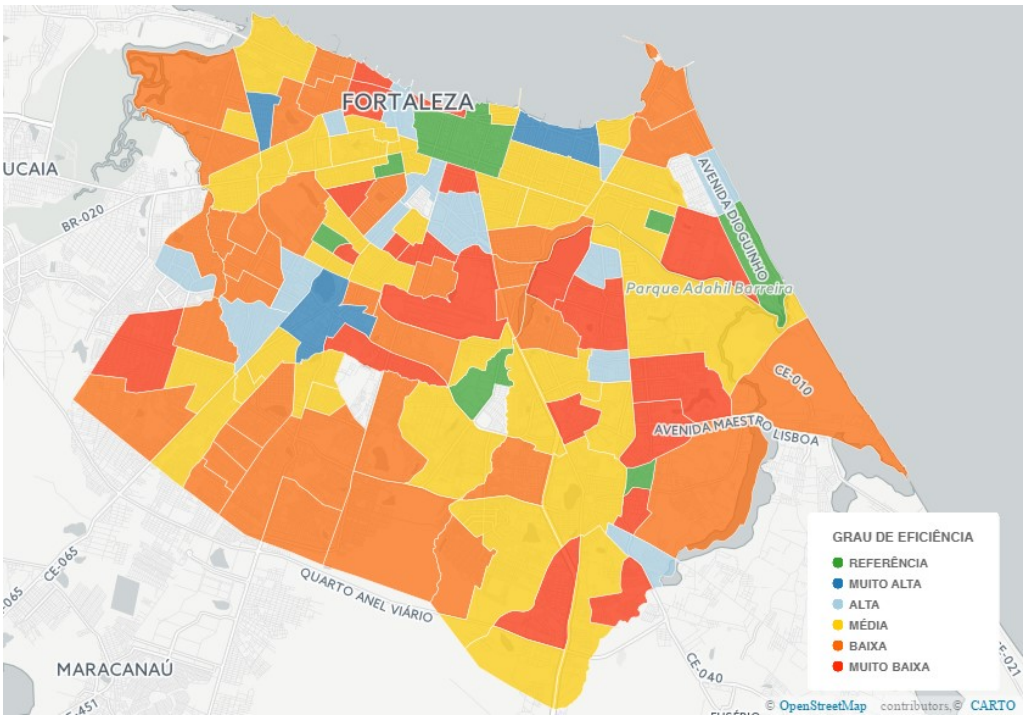
Bairro	Eficiência	Projeção Quantidade PDV	% Projecao PDV	Projeção Area	% Projecao Area	Projeção População	% Projecao Populacao	Projeção Domicilios	% Projecao Domicilios
boa vista	1	0	0.00%	0	0.00%	0	0.00%	0	0.00%
centro	1	0	0.00%	0	0.00%	0	0.00%	0	0.00%
cidade 2000	1	0	0.00%	0	0.00%	0	0.00%	0	0.00%
curio	1	0	0.00%	0	0.00%	0	0.00%	0	0.00%
pan americano	1	0	0.00%	0	0.00%	0	0.00%	0	0.00%
parque araxa	1	0	0.00%	0	0.00%	0	0.00%	0	0.00%
praia do futuro ii	1	0	0.00%	0	0.00%	0	0.00%	0	0.00%
meireles	0.964964	2.697737	3.50%	90430.739851	3.50%	1294.318043	3.50%	44.460103	3.50%
jardim iracema	0.877178	5.772623	12.28%	135231.355411	12.28%	2840.130671	12.28%	80.693905	12.28%
pedras	0.82564	1.917965	17.44%	747205.678637	17.44%	233.991701	17.44%	64.51336	17.44%
parangaba	0.822213	19.556561	17.78%	714994.763078	17.78%	5472.992572	17.78%	1640.084345	17.78%
benfica	0.814811	11.481726	18.52%	174832.722769	18.52%	1639.479372	18.52%	550.937663	18.52%
varjota	0.775902	7.843428	22.41%	118797.129721	22.41%	1885.335994	22.41%	625.681457	22.41%
bonsucesso	0.756893	25.769303	24.31%	612503.725579	24.31%	9996.301703	24.31%	285.407189	24.31%
praia do futuro i	0.744354	7.413726	25.56%	415949.272992	25.56%	1682.660178	25.56%	492.118025	25.56%
jacarecanga	0.672839	30.7531	32.72%	411467.840452	32.72%	4560.619329	32.72%	1363.278389	32.72%
damas	0.672588	8.840118	32.74%	285383.096359	32.74%	3501.341708	32.74%	1150.197627	32.74%
conjunto ceara i	0.626205	63.545143	37.38%	588678.284154	37.38%	7176.863216	37.38%	2045.779811	37.38%
guararapes	0.623506	1.882471	37.65%	508922.791067	37.65%	1982.618493	37.65%	582.060044	37.65%
vila ellery	0.618543	6.484775	38.15%	176014.862436	38.15%	2997.11054	38.15%	874.300287	38.15%
fatima	0.597323	33.019541	40.27%	1150622.348948	40.27%	9313.121398	40.27%	2919.813354	40.27%
messejana	0.580263	88.984146	41.97%	2549983.28939	41.97%	17442.991279	41.97%	5109.872605	41.97%
maraponga	0.576585	27.945369	42.34%	724834.552442	42.34%	4296.812258	42.34%	1297.342605	42.34%
monte castelo	0.575309	21.234535	42.47%	336567.176404	42.47%	5610.164155	42.47%	1629.962909	42.47%
cajazeiras	0.574538	8.509233	42.55%	1436036.991871	42.55%	6150.048187	42.55%	1877.136811	42.55%
passare	0.573055	20.066411	42.69%	3060045.094943	42.69%	21595.30075	42.69%	6385.815095	42.69%
parque manibura	0.568186	6.477214	43.18%	544501.716892	43.18%	3251.129529	43.18%	88.090108	43.18%
praia de iracema	0.558124	10.163141	44.19%	226432.065006	44.19%	1379.535928	44.19%	481.202635	44.19%
mucuripe	0.558097	22.53707	44.19%	386466.438023	44.19%	6068.658432	44.19%	1965.144109	44.19%
padre andrade	0.557981	9.724427	44.20%	539423.591663	44.20%	5668.899095	44.20%	1661.10899	44.20%
joquei clube	0.555259	12.897482	44.47%	756823.85563	44.47%	8581.717513	44.47%	252.168005	44.47%
jardim cearense	0.542966	5.027373	45.70%	393390.149074	45.70%	4595.933136	45.70%	1328.597616	45.70%
parque sao jose	0.52808	26.89943	47.19%	282936.514575	47.19%	4942.888231	47.19%	1423.782107	47.19%
sao gerardo	0.523903	11.426328	47.61%	696250.431706	47.61%	6881.02998	47.61%	2101.49217	47.61%
antonio bezerra	0.516791	63.30033	48.32%	1057616.515268	48.32%	12449.387028	48.32%	3613.434102	48.32%
joaquim tavora	0.51493	41.230977	48.51%	953340.032081	48.51%	11332.212834	48.51%	3598.73671	48.51%
cambeba	0.505375	9.8925	49.46%	1344717.338439	49.46%	3766.569506	49.46%	1065.422287	49.46%
parquelandia	0.504727	20.306201	49.53%	620852.269527	49.53%	7058.633466	49.53%	2193.069674	49.53%
canindezinho	0.494556	34.875606	50.54%	1905031.720305	50.54%	20818.209645	50.54%	5834.840539	50.54%
bom jardim	0.491264	99.203524	50.87%	1232748.807114	50.87%	19178.839176	50.87%	5322.396229	50.87%
dionisio torres	0.488939	23.508812	51.11%	880909.75003	51.11%	7983.796899	51.11%	2475.580091	51.11%
vila pery	0.464773	46.029508	53.52%	799389.447613	53.52%	11043.870669	53.52%	3261.137152	53.52%
democrito rocha	0.464548	23.559901	53.55%	429133.308064	53.55%	5886.227195	53.55%	1753.606301	53.55%
coacu	0.452854	1.641439	54.71%	913489.918925	54.71%	3932.887168	54.71%	1115.084036	54.71%
montese	0.452367	65.716019	54.76%	1046359.227315	54.76%	14202.327049	54.76%	4351.495748	54.76%
farias brito	0.447623	13.809437	55.24%	507710.72893	55.24%	6637.920032	55.24%	1991.320772	55.24%
papicu	0.445034	49.391992	55.50%	1160277.828586	55.50%	10179.19018	55.50%	3079.507486	55.50%
paupina	0.4423	37.365911	55.77%	3036086.154062	55.77%	8139.076167	55.77%	2350.148483	55.77%
dias macedo	0.44097	20.125093	55.90%	857099.761511	55.90%	6732.402524	55.90%	1945.984654	55.90%
jardim guanabara	0.435928	27.639504	56.41%	423200.132349	56.41%	8410.30631	56.41%	2418.174591	56.41%
bela vista	0.432339	21.571127	56.77%	553450.370262	56.77%	9507.758134	56.77%	2794.028631	56.77%
barra do ceara	0.421273	70.604735	57.87%	2453949.069812	57.87%	41826.939632	57.87%	11736.011689	57.87%
aldeota	0.416952	100.867236	58.30%	2262723.443298	58.30%	24645.422421	58.30%	8001.162335	58.30%
presidente ken...	0.415829	11.683429	58.42%	995095.38947	58.42%	13371.68418	58.42%	389.642348	58.42%
jardim america	0.413804	34.585545	58.62%	452255.616829	58.62%	7175.621207	58.62%	2124.373111	58.62%
cidade dos func...	0.411293	54.161044	58.87%	1652068.645466	58.87%	10704.45939	58.87%	3142.517969	58.87%
edson queiroz	0.409175	62.62742	59.08%	8154593.998582	59.08%	13108.628091	59.08%	3486.45668	59.08%
jangurussu	0.404618	104.191862	59.54%	5378354.860515	59.54%	30029.880668	59.54%	8485.385211	59.54%
coco	0.397203	11.453144	60.28%	1975457.122729	60.28%	12352.516877	60.28%	3881.41012	60.28%
cais do porto	0.394292	18.171228	60.57%	1552307.164405	60.57%	13545.439215	60.57%	3828.677784	60.57%
cristo redentor	0.390542	30.472915	60.95%	716367.754724	60.95%	16251.814853	60.95%	4410.64967	60.95%
prefeito jose va...	0.378482	76.446737	62.15%	6617014.523873	62.15%	20369.015587	62.15%	5962.223981	62.15%
conjunto esper...	0.373279	36.976521	62.67%	693080.385033	62.67%	10277.592617	62.67%	2978.803446	62.67%
parque santa ro...	0.372603	27.605459	62.74%	626239.071753	62.74%	8018.75835	62.74%	2337.053036	62.74%
joao xxiii	0.365165	33.646231	63.48%	744905.963461	63.48%	11643.500328	63.48%	3320.819487	63.48%
henrique jorge	0.361876	77.851144	63.81%	1240142.772098	63.81%	17207.017117	63.81%	4987.578186	63.81%
parque dois irm...	0.358435	37.210752	64.16%	2829565.359157	64.16%	17464.674116	64.16%	4798.262351	64.16%
domlustosa	0.353194	3.234031	64.68%	768757.318281	64.68%	8499.680352	64.68%	2477.914575	64.68%
parque preside...	0.35052	11.690638	64.95%	1068356.374695	64.95%	4659.368854	64.95%	1264.537379	64.95%
vila uniao	0.349907	22.753245	65.01%	943855.6543	65.01%	9936.667274	65.01%	2933.86846	65.01%
mondubim	0.340831	132.493045	65.92%	6241838.13978	65.92%	50072.483647	65.92%	14563.688292	65.92%
alvaro weyne	0.33183	40.758375	66.82%	953484.683045	66.82%	15764.805068	66.82%	4480.080443	66.82%
alto da balanca	0.327799	28.904651	67.22%	616038.608305	67.22%	8602.830865	67.22%	2535.542899	67.22%
vincente pinzon	0.321248	66.51774	67.88%	2088308.570559	67.88%	30679.610783	67.88%	8628.301156	67.88%
genibau	0.320169	58.465446	67.98%	1474766.808405	67.98%	27403.978092	67.98%	7711.320355	67.98%
quintino cunha	0.315151	40.406115	68.48%	1935112.955456	68.48%	32293.388802	68.48%	9003.030267	68.48%
aerolandia	0.309608	72.49121	69.04%	756785.853665	69.04%	7840.787391	69.04%	2270.700866	69.04%
vila velha	0.303622	43.175458	69.64%	4981997.103826	69.64%	42893.425027	69.64%	12065.451449	69.64%
rodolfo teofilo	0.301437	41.2152	69.86%	1215019.166932	69.86%	13307.619692	69.86%	3962.946273	69.86%
lagoaredonda	0.297655	70.234534	70.23%	8294422.558018	70.23%	19553.996715	70.23%	5582.240792	70.23%

serrinha	0.29488	60.640344	70.51%	2100100.355111	70.51%	20248.938976	70.51%	5834.165166	70.51%
jardim das olive...	0.280987	55.363968	71.90%	1709420.145673	71.90%	21243.945585	71.90%	5957.738209	71.90%
floresta	0.27873	5.04889	72.13%	1237109.36873	72.13%	20839.652846	72.13%	596.490254	72.13%
granja portugal	0.27847	55.557792	72.15%	1825495.612946	72.15%	28584.844528	72.15%	7786.027647	72.15%
sao joao do tau...	0.267361	53.482676	73.26%	1827568.700132	73.26%	20196.670206	73.26%	6081.639619	73.26%
barroso	0.26611	33.758931	73.39%	2518327.975985	73.39%	21900.005614	73.39%	6106.697051	73.39%
planalto ayrton...	0.262045	37.6357	73.80%	2961974.334692	73.80%	28989.820465	73.80%	8144.80828	73.80%
siqueira	0.256418	32.717615	74.36%	4603952.572681	74.36%	24993.283283	74.36%	6880.36565	74.36%
autran nunes	0.248048	17.294904	75.20%	748021.213922	75.20%	15939.886303	75.20%	4217.700834	75.20%
sabiaguaba	0.243209	11.351861	75.68%	10917156.624...	75.68%	1584.719807	75.68%	440.45221	75.68%
itaoca	0.235867	12.990261	76.41%	566618.413194	76.41%	9533.323198	76.41%	2853.272589	76.41%
conjunto palme...	0.231745	33.803241	76.83%	2970952.851899	76.83%	28117.382049	76.83%	7001.112124	76.83%
manoel satiro	0.23051	63.098195	76.95%	2365412.646336	76.95%	29165.986226	76.95%	8559.808743	76.95%
pici	0.211315	56.785348	78.87%	3000563.263572	78.87%	33504.932435	78.87%	9362.48418	78.87%
pirambu	0.18997	12.150443	81.00%	447400.781396	81.00%	14145.545366	81.00%	3867.890919	81.00%
sapiranga/coite	0.187941	32.482378	81.21%	3865554.820001	81.21%	26110.959275	81.21%	7007.260919	81.21%
salinas	0.177485	1.645031	82.25%	2110679.248412	82.25%	3535.170599	82.25%	1007.581197	82.25%
bom futuro	0.176771	16.464579	82.32%	315789.441035	82.32%	5269.488667	82.32%	162.176108	82.32%
parque iracema	0.172273	6.621814	82.77%	1308299.035631	82.77%	6953.732532	82.77%	2263.832695	82.77%
ancuri	0.17011	48.963525	82.99%	3438886.595861	82.99%	16543.032366	82.99%	4837.430303	82.99%
aeroporto	0.157163	3.371347	84.28%	5163764.272712	84.28%	7248.396202	84.28%	2007.637181	84.28%
josebonifacio	0.146244	16.221369	85.38%	758079.926873	85.38%	7522.446649	85.38%	2419.545319	85.38%
amadeo furtado	0.131604	6.947164	86.84%	803520.994268	86.84%	10149.806665	86.84%	2929.966435	86.84%
carlito pamplona	0.115802	22.989158	88.42%	1203286.873995	88.42%	25690.383919	88.42%	7353.877923	88.42%
luciano cavalca...	0.110501	31.132472	88.95%	3417614.687808	88.95%	13780.121764	88.95%	3977.840468	88.95%
itaperi	0.109349	40.969953	89.07%	2256342.402846	89.07%	20090.417918	89.07%	6283.54384	89.07%
arraial moura b...	0.107091	4.464543	89.29%	381274.908612	89.29%	3343.942853	89.29%	93.755407	89.29%
guajeru	0.097463	13.538053	90.25%	975069.519259	90.25%	6018.115867	90.25%	1674.205899	90.25%
granja lisboa	0.095772	27.126847	90.42%	4334180.195278	90.42%	47019.868013	90.42%	13043.492232	90.42%
jose de alencar	0.094201	19.021773	90.58%	2837289.942852	90.58%	14494.590885	90.58%	4136.782688	90.58%
couto fernandes	0.052263	2.843212	94.77%	317625.375241	94.77%	4984.151278	94.77%	1473.731743	94.77%
sao bento	0.035145	10.613401	96.49%	2142157.64426	96.49%	11482.735343	96.49%	3412.690942	96.49%
manuel dias br...	0.020029	1.959941	98.00%	4367432.739415	98.00%	1418.017414	98.00%	433.146992	98.00%
parreao	0.01159	1.97682	98.84%	1015069.958287	98.84%	10871.519794	98.84%	3171.807166	98.84%

Fonte: Autor

Desta forma, pelos resultados acima mostrados, vemos que sete bairros foram considerados como eficientes pelo modelo e irão servir como referência para os demais bairros considerados ineficientes. A partir deste resultado, o *output* geral foi usado para gerar a Figura 49.

Figura 49 - Visualização grau de eficiência

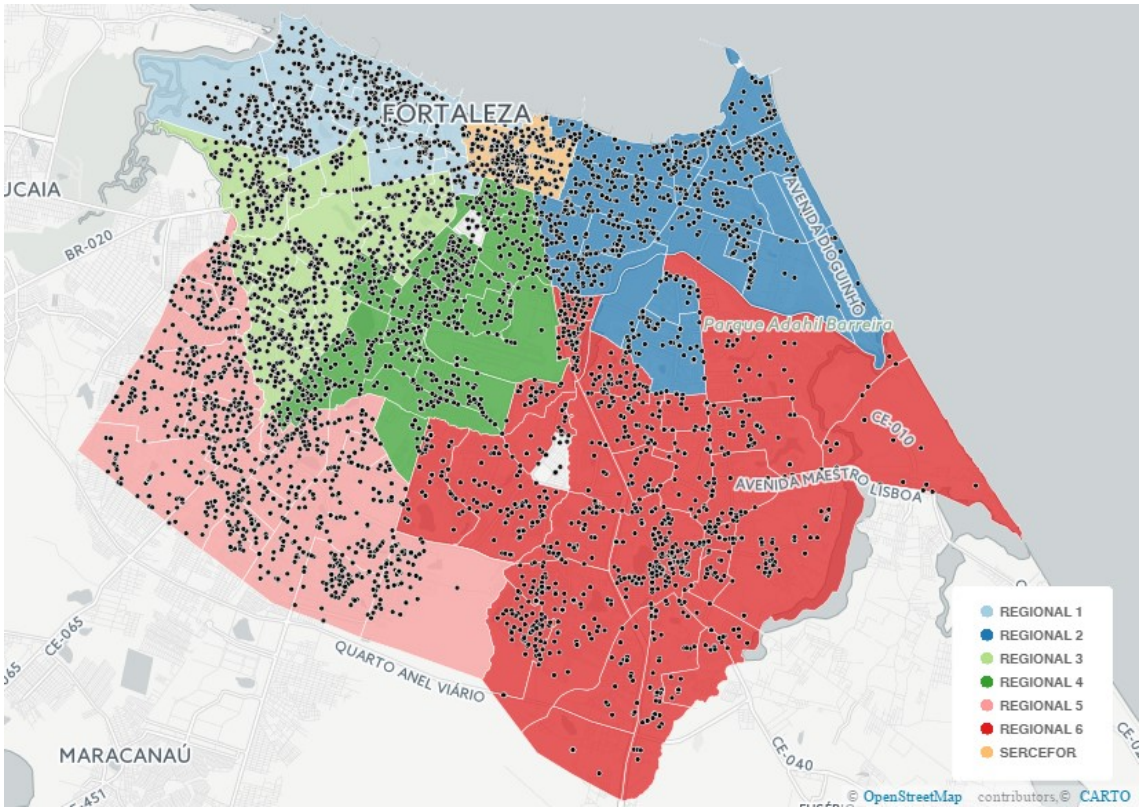


Fonte: Autor

É possível notar que há maior predominância de bairros com maior eficiência nas regiões 2,3 e 4. Porém, a faixa de eficiência dominante encontra-se entre média e baixa (0.2-0.6). Vale ressaltar que a regional 5 é a única que não possui nenhum bairro considerado acima da faixa média (0.4-0.6) de eficiência.

A partir desses resultados a figura 50 mostra o cenário obtido através do modelo, quando removido os PDV's que não possuem valor significativo de faturamento diante do agregado para seu CEP. Esta situação é representada pelo *output* final.

Figura 50 - Situação proposta pelo modelo

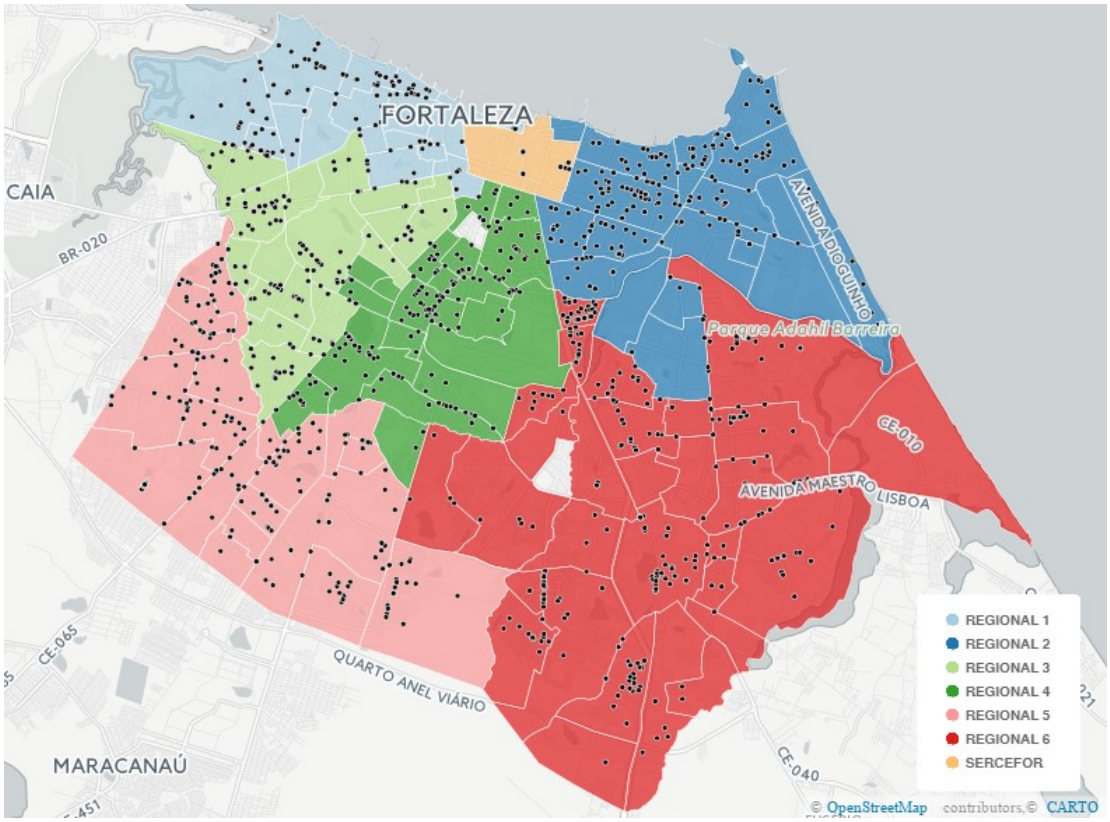


Fonte: Autor

Pela figura 50 vemos que apesar de retirar 1.247 pontos de vendas com o modelo, a distribuição continua bastante satisfatória, abrangendo a totalidade do território estudado.

A figura 51 mostra apenas aqueles pontos considerados como passíveis de remoção. Este gráfico foi gerado pelo *output* removidos.

Figura 51 - PDV's removidos



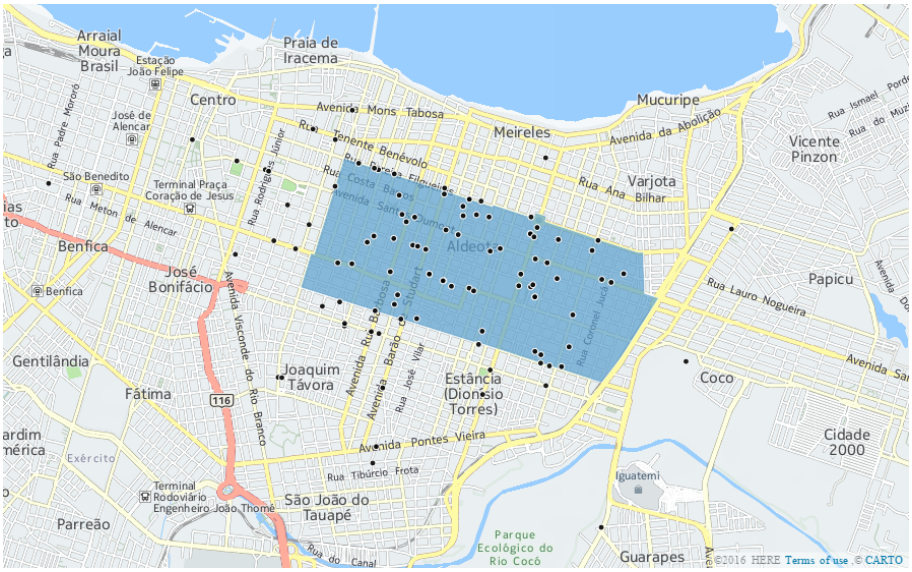
Fonte: Autor

Para finalizar as figuras 52 e 53 mostram uma comparação do bairro aldeota nas duas situações. Este bairro foi escolhido como exemplo por ter sido um dos bairros que mais reduziu em quantidade de PDV's. Vale ressaltar que com o modelo implementado, a aplicação de filtros para bairros se torna processo prático dentro do ambiente de visualização.

Figura 52 - Bairro Aldeota situação inicial

Fonte: Autor

Figura 53 - Bairro Aldeota situação final



Fonte: Autor

3.7.3 Comparação de resultados

A figura 54 mostra a relação das métricas criadas para avaliar a performance da solução aplicada após fluxo completo.

Figura 54 - Avaliação métricas performance

Bairro	Reducao PDV	% Reducao PDV	% Projecao PDV	Reducao Faturamento	% Reducao Faturamento	Gap PDV
aerolandia	49	46.67%	67.39%	5866	13.08%	21.758839
aeroporto	0	0.00%	83.28%	0	0.00%	3.331169
aldeota	67	38.73%	58.06%	6929	4.41%	33.439494
alto da balanca	7	16.28%	66.90%	575	1.75%	21.766661
alvaro weyne	8	13.11%	66.44%	858	1.62%	32.530711
amadeo furtado	0	0.00%	86.21%	0	0.00%	6.896539
ancuri	35	59.32%	81.83%	2295	8.63%	13.280143
antonio bezerra	34	25.95%	48.03%	4427	3.41%	28.918384
arraial moura brasil	0	0.00%	89.16%	0	0.00%	4.458076
autran nunes	3	13.04%	74.47%	302	1.76%	14.128268
barra do ceara	26	21.31%	57.31%	2818	1.85%	43.922421
barroso	0	0.00%	72.89%	0	0.00%	33.529742
bela vista	4	10.53%	56.05%	385	0.86%	17.299333
benfica	10	16.13%	18.91%	837	0.95%	1.721473
bom futuro	0	0.00%	80.85%	0	0.00%	16.169243
bom jardim	37	18.97%	49.51%	4480	2.84%	59.550064
bonsucesso	15	14.15%	24.46%	1920	2.00%	10.922422
cais do porto	8	26.67%	60.15%	791	2.25%	10.04608
cajazeiras	4	20.00%	41.83%	490	1.43%	4.366149
cambeba	3	15.00%	48.54%	243	0.90%	6.708976
canindezinho	11	15.94%	49.44%	1207	1.18%	23.116833
carlito pamplona	2	7.69%	87.80%	123	1.32%	20.827181
cidade dos funcionarios	27	29.35%	57.96%	2443	3.26%	26.319345
coacu	0	0.00%	53.98%	0	0.00%	1.61952
coco	0	0.00%	58.91%	0	0.00%	11.193828
conjunto ceara i	42	24.71%	37.27%	3738	2.99%	21.350903
conjunto esperanca	22	37.29%	62.43%	2127	4.34%	14.833504
conjunto palmeiras	10	22.73%	76.21%	1004	3.25%	23.533539
couto fernandes	0	0.00%	94.06%	0	0.00%	2.821743
cristo redentor	12	24.00%	59.94%	1368	2.62%	17.9723
damas	4	14.81%	32.74%	320	0.64%	4.839568
democrito rocha	8	18.18%	52.88%	734	1.66%	15.268145
dias macedo	12	33.33%	55.20%	1173	2.96%	7.87146
dionisio torres	9	19.57%	50.87%	737	1.29%	14.398207
domlustosa	0	0.00%	64.19%	0	0.00%	3.209588
edson queiroz	43	40.57%	58.27%	4510	5.26%	18.764172
farias brito	7	28.00%	54.71%	741	2.24%	6.676781
fatima	31	37.80%	38.92%	4135	3.58%	0.915317
floresta	0	0.00%	71.43%	0	0.00%	5.000325
genibau	16	18.60%	67.54%	1420	1.90%	42.084261

granja lisboa	0	0.00%	89.93%	0	0.00%	26.979832
granja portugal	10	12.99%	71.81%	997	1.57%	45.296063
guajeru	1	6.67%	89.51%	110	2.49%	12.426202
guararapes	0	0.00%	37.38%	0	0.00%	1.868942
henrique jorge	26	21.31%	63.66%	3551	4.03%	51.659767
itaoca	1	5.88%	76.02%	125	1.05%	11.923682
itaperi	3	6.52%	88.39%	195	1.25%	37.657682
jacarecanga	31	32.98%	32.12%	5798	5.38%	-0.811861
jangurussu	39	22.29%	59.03%	4861	2.99%	64.29575
jardim america	8	13.56%	57.77%	876	2.07%	26.083406
jardim cearense	0	0.00%	45.28%	0	0.00%	4.981118
jardim das oliveiras	15	19.48%	70.99%	1178	2.03%	39.66315
jardim guanabara	8	16.33%	55.85%	701	1.61%	19.364735
jardim iracema	3	6.38%	12.97%	325	0.61%	3.093747
joao xxiii	6	11.32%	63.20%	675	1.41%	27.497276
joaquim tavora	17	20.00%	48.37%	1935	1.93%	24.113561
joquei clube	2	6.90%	41.68%	242	0.87%	10.087299
josebonifacio	3	15.79%	84.99%	130	1.56%	13.14885
jose de alencar	2	9.52%	89.89%	20	0.32%	16.877795
lagoaredonda	27	27.00%	69.43%	2173	3.18%	42.427318
luciano cavalcante	0	0.00%	88.19%	0	0.00%	30.864792
manoel satiro	8	9.76%	76.39%	1071	1.89%	54.643222
manuel dias branco	0	0.00%	97.24%	0	0.00%	1.944847
maraponga	20	30.30%	42.29%	2588	3.83%	7.910647
meireles	27	35.06%	2.03%	3256	2.85%	-25.434821
messejana	46	21.70%	40.86%	5397	2.23%	40.628746
mondubim	34	16.92%	65.32%	4709	2.57%	97.302092
monte castelo	8	16.00%	42.46%	856	1.44%	13.230442
montese	47	39.17%	54.37%	4743	4.30%	18.243488
mucuripe	12	23.53%	43.49%	1257	1.95%	10.182032
padre andrade	0	0.00%	43.47%	0	0.00%	9.563177
papicu	32	35.96%	54.73%	3466	4.34%	16.706191
parangaba	41	37.27%	16.36%	4967	2.38%	-22.998797
parque dois irmaos	0	0.00%	63.84%	0	0.00%	37.02766
parque iracema	0	0.00%	81.82%	0	0.00%	6.545923
parque manibura	2	13.33%	37.81%	160	1.09%	3.671469
parque presidente var...	2	11.11%	64.50%	215	1.28%	9.61047
parque santa rosa	1	2.27%	62.53%	5	0.01%	26.514988
parque sao jose	6	10.53%	46.05%	908	2.15%	20.250045
parquelandia	2	4.88%	49.12%	387	0.72%	18.138504
parreao	0	0.00%	93.63%	0	0.00%	1.872535
passare	4	8.51%	42.12%	593	0.73%	15.798682
paupina	4	5.97%	54.82%	322	0.53%	32.726127
pedras	2	18.18%	17.02%	190	1.28%	-0.1273
pici	18	25.00%	78.38%	1719	3.77%	38.43459
pirambu	1	6.67%	80.16%	210	2.41%	11.023501
planalto ayrton senna	7	13.73%	73.15%	490	1.22%	30.308422
praia de iracema	1	4.35%	44.22%	30	0.14%	9.170027
praia do futuro i	3	10.34%	24.88%	347	0.78%	4.215067
prefeito jose valter	37	30.08%	61.45%	4185	4.01%	38.583023
presidente kennedy	2	10.00%	57.34%	93	0.61%	9.467654
quintino cunha	16	27.12%	68.08%	1787	3.24%	24.169018
rodolfo teofilo	10	16.95%	69.21%	1078	2.44%	30.834566
sabiaguaba	0	0.00%	74.86%	0	0.00%	11.228662
salinas	0	0.00%	82.30%	0	0.00%	1.646008
sao bento	0	0.00%	95.60%	0	0.00%	10.516025
sao gerardo	6	25.00%	47.03%	529	1.42%	5.286563
sao joao do tauape	11	15.07%	72.73%	1493	2.82%	42.095024
sapiranga/coite	6	15.00%	80.54%	409	1.78%	26.216261
serrinha	13	15.12%	69.86%	1122	1.76%	47.076653
siqueira	4	9.09%	73.54%	354	1.03%	28.358863
varjota	6	17.14%	21.43%	1563	2.87%	1.499191
vila ellery	1	5.88%	38.29%	115	0.40%	5.509597
vila pery	17	19.77%	52.88%	2142	2.54%	28.473009
vila uniao	4	11.43%	64.20%	325	0.90%	18.471514
vila velha	6	9.68%	69.07%	1020	1.80%	36.823288
vincente pinzon	23	23.47%	66.77%	1864	1.99%	42.434815

Fonte: Autor

Para a competência analisada, houve uma redução total de 19,15% dos PDV's contemplados na base, sem redução significativa no faturamento.

Sendo assim, pelos resultados mostrados na figura 54, vemos que apesar de apenas 4 bairros ter atingido o valor de referência para as projeções de eficiência (Jacarecanga,

Meireles, Parangaba e Pedras), muitos obtiveram valor satisfatório de redução (considerando a projeção como referência), sem impactar o nível de faturamento. Visto modelo desenvolvido e aplicação prática estabelecida, o Quadro 4, irá resumir os principais benefícios ganhos com a tecnologia desenvolvida no trabalho.

Quadro 5 - Benefícios gerados no trabalho

Benefícios Gerados
Integração dos ambientes relacionados ao processamento ETL, aplicação de modelo analítico e carregamento para visualizações.
A partir da base gerada pelo modelo DEA outras problemáticas podem ser abordadas utilizando como referência o ranking de eficiência para priorização de ações.
Visualizações espaciais, permitindo avaliar diversos ângulos do negócio, como bairros eficientes, ineficientes, diferentes cenários, faixas de referência para faturamento e eficiência.
Intereface permitindo controle sobre a configuração do modelo DEA e competência a ser analisada.
Tratanmento nos campos de latitude/longitude dando a possibilidade de realizar diagnóstico sobre os erros de localização e analisar possíveis ações a serem tomadas.
Tratamento realizado nas categorias de bairros, tornando possível o uso da base com maior grau de confiança e validando maior número de registros para análises posteriores.

Fonte: Autor

A maior base utilizada no estudo apresentava mais de 3 milhões e o tempo total de processamento do fluxo criado, levando em considerações todas as etapas avaliadas no projeto levou em média 10.5 segundos de execução, o que gera flexibilidade na parametrização para testes de possíveis outros cenários que possa vir a ser interessante.

Cabe enfatizar aqui, que apesar do uso de algumas ferramentas sofisticadas para visualizações, a empresa em questão não possuía ainda poucas competências no que diz respeito a processamento de dados e visualizações espaciais.

Desta forma a ferramenta, mostra-se como uma proposta de serviço da empresa a seus clientes (distribuidoras e operadoras), podendo fornecer insights valiosos para um bom gerenciamento dos recursos envolvidos.

4. CONCLUSÃO

Diante dos resultados obtidos no trabalho, pode-se afirmar que o objetivo geral do trabalho foi atingido mediante a concepção de um ambiente de apoio a decisão contemplando todo o processo ETL, aplicação de modelagem DEA para geração de escala de eficiência referente aos bairros, utilizada para priorização de ações com possibilidade de visualizar e analisar os resultados em mapas. Este sistema foi testado em uma aplicação prática de classificação de pontos de vendas para remoção, gerando resultados satisfatórios no que diz respeito a redução geral (19,15% dos PDV's) sem impactos significativos no faturamento.

O objetivo específico, estudar e compreender o método de Análise por Envoltória de Dados, bem como sua formulação, foi atingido com a seção 2.4 da revisão de literatura.

Quanto ao objetivo específico identificar os fatores e parâmetros que devem ser considerados na formulação do problema, este objetivo foi obtido vide colaboração com pessoas da área de negócio e com as ferramentas estatísticas apresentadas na seção 3.7.2 do estudo de caso.

O terceiro objetivo específico, criar fluxo de extração, transformação e carregamento dos dados para utilização de modelos analíticos, foi obtido no processamento realizado em seção 3.5.1 do estudo de caso. A partir deste obtém-se *dataset* limpo para aplicação em diversos modelos analíticos independentes.

Por fim, o objetivo específico, modelar o problema pelo método de Análise por Envoltória de Dados, utilizando o software Alteryx Designer, foi obtido com auxílio da comunidade referente ao software e está exposto na seção 3.5.2 do estudo de caso.

O estudo apresentou algumas limitações vide a qualidade dos dados georreferenciados presentes na base. Estes dados são obtidos pelos próprios supervisores responsáveis pelos pontos de venda, porém nem sempre são fieis a realidade. Limitação semelhante é a baixa quantidade de dados georreferenciados diante do total de registros da base, visto que apenas algumas distribuidoras mantêm esse tipo de informação em suas bases. Da mesma forma, necessita-se melhor preenchimento dos campos relacionados as categorias de bairros.

Para trabalhos futuros, recomenda-se: a utilização de tratamento nos casos de pontos de vendas que apresentem latitude e longitude fora das fronteiras do seu bairro; Análise de critérios que possam ser vinculados ao modelo para avaliar quão

significativo é um ponto de venda; Criação de uma interface acessível aos usuários deste modelo para facilitar a parametrização dos critérios utilizados como Nível de significância e outros que possam vir a ser considerados.

REFERÊNCIAS

ABREU, F. S. G. da G. Desmistificando o conceito de ETL. Disponível em: http://www.fsma.edu.br/si/Artigos/V2_Artigo1.pdf.

ANGULO MEZA, L. Data envelopment analysis na determinação da eficiência dos programas de pós-graduação da COPPE/UFRJ. 1998. Tese (Mestrado em Engenharia de Produção) – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

ANGULO MEZA, L.; SOARES DE MELLO, J.C.C.B.; GOMES, E.G.; FERNANDES, A. J. S. Seleção de variáveis em DEA aplicada a uma análise do mercado de energia eléctrica, 2007.

BANKER, R. D. Maximum likelihood, consistency and Data Envelopment Analysis: A statistical foundation. Management Science, Vol. 39, nº 10, pp. 1265-1273, 1993.

BRUSCHI, A. G.; BREVE, Fabrício Aparecido; GIORDANO, Luís Gustavo. Construindo Sistemas de Apoio à Decisão, 2003(tese).

CIELO, I. ETL – Extração, Transformação e Carga de Dados. Disponível em: http://www.datawarehouse.inf.br/etl.htm.

COOPER, W. W.; SEIFORD, L.M.; TONE, K. Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software, 2000. Kluwer Academic Publishers, Boston-USA.

DA COSTA, P. Cours d'introduction à l'analyse économique, 2012. Polycopie Ecole Centrale ParisSupelec.

DATE, C. J. Introdução a Sistemas de Banco de Dados. Rio de Janeiro: Campus, 2000. P. 803.

DEBREU G. The Coefficient of Resource Utilization, 1951. Econometrica, Vol. 19, No. 3, 1951, pp. 273-292. http://dx.doi.org/10.2307/1906814.

DENSHAW, P. Spatial decision support systems. In: Maguire, D. J.; Goodchild, M. F.; Rhind, D. W., Geographical Information Systems: principles and applications, New York, Longman, vol. 1, 1991, 403-412.

FAN, J.; HAN, F.; LIU, H. Challenges of big data analysis, 2014. National Science Review, pp. 293–314.

FARRELL, M. J. The Measurement of Productive Efficiency, 1957. Journal of the Royal Statistical Society. Series A (General), Vol. 120, No.3 (1957), 253-290.

FERREIRA, C. M. de C.; GOMES, A. P. Introdução à análise envoltória de dados: teoria, modelos e aplicações. Viçosa – MG: Editora UFV, 2009.

GANDOMI, A.; HAIDER, M. Beyond the hype: big data concepts, methods, and analytics, 2015. Int J Inf Manag 35 (2):137–144

Gartner IT Glossary. Definition of BIG DATA, 2012. Retrieved from: http://www.gartner.com/it-glossary/big-data/

Gartner IT Glossary. Definition of Business Inteligence, 2014. Retrieved from: http://www.gartner.com/it-glossary/business-intelligence-bi/

HANSSON, S. Decision Theory: A Brief Introduction, 1994.

HOPPEN, N.; ESPERANCA, L. G. Geradores de sistemas de apoio à decisão e seu uso num processo de gestão orçamentária. Rev. adm. empres., São Paulo , v. 29, n. 2, p. 33-45, June 1989.

JI, Y.; LEE, C. Data envelopment analysis. The Stata Journal. Vol. 10, nº 2, pp. 267-280, 2010.

KIRSCHBAUM, C. As Redes Intraorganizacionais são inclusivas? Utopia e Testes. Organ. Soc., Salvador, v.22, n.74, p.367-384, Sept. 2015.

LABRINIDIS, A.; JAGADHIS, H. V. Challenges and opportunities with big data, 2012. Proc. VLDB Endow. 5, 12 (August 2012), 2032-2033.

LARSON, D. BI principles for agile development: keeping focused, 2009. Business Intelligence Journal, 14(4), 36–41. Retrieved from Business Source Complete database.

LARSON, D.; CHANG, V. A review and future direction of agile, business intelligence, analytics and data science, 2016.

LE MONDE. Riminder, la start-up big data qui veut optimiser le recrutement de talents africains, 2015.
Retrieved from: http://www.lemonde.fr/afrique/article/2015/05/13/riminder-la-start-up-big-data-qui-veut-optimiser-le-recrutement-de-talents-africains_4632717_3212.html

LINS, M.P.E.; MEZA, L.A. Análise envoltória de dados e perspectivas de integração no meio ambiente de apoio à decisão. Rio de Janeiro: COPPE, 2000.

MILGRON, P.; LEVIN, J. Introduction to Choice Theory". web.stanford.edu. Stanford University.

NETO, S. L. R.; RODRIGUES, M. Um modelo conceitual para integração de modelos científicos e informação geográfica. In: III Workshop Brasileiro de Geoinformática – GEOINFO, 3., Rio de Janeiro, 2001. Anais. Rio de Janeiro, Sociedade Brasileira de Computação. 2001. 71-78.

NOBLE, J. The core of IT, 2006. pp. 15–17. CIO Insight.

NORTH, D. Institutions, institutional change and economic performance. Cambridge University Press, New York. 1990.

PÉRICO, A. E.; REBELATTO, D. A. N; SANTANA, N. B. Eficiência bancária: os maiores bancos são os mais eficientes? Uma análise por envoltória de dados. Revista Gestão & Produção, São Carlos, v. 15, n. 2, maio/ago. 2008;

PESSE, R.; GALVAO, R. D. Sistemas Georeferenciados de apoio à decisão espacial via internet, 2003.

POMEROL, J-Ch; ADAM F. Practical Decision Making – From the Legacy of Herbert Simon to Decision Support Systems, 2004

RAFAELI NETO, S. L.; SOUZA, A. P. de; MORAES, R. A. R. de. Potencial de sistemas de informação geográfica como sistemas de apoio à decisão espacial para gerenciamento de recursos hídricos, 2002.

RAFAELI NETO, S. L. Sistemas de Apoio à Decisão Espacial: uma contribuição à teoria em geoprocessamento, 2004.

RAJARAMAM, V. Big Data Analytics, 2016. Supercomputer Education and Research Centre Indian Institute of Science Bengaluru.

RESNIK, M.. Choices: An Introduction to Decision Theory. Univ of Minnesota Press. 1987.ROVERI, Guilherme De Oliveira. Aplicação da teoria de análise de decisão na avaliação de investimentos, 2011.

SAVAGE, L. J. The Foundations of Statistics. J. Wiley, New York, 1954. second revised edition, 1972.

SEIFORD, L.M.; ZHU, J. An investigation of returns to scale under data envelopment analysis. International Journal of Management Science, v. 27, p.1–11. 1999.

Silva, E. L. da Metodologia da pesquisa e elaboração de dissertação/Edna Lúcia da Silva, Estera Muszkat Menezes. – 4. ed. rev. atual. – Florianópolis: UFSC, 2005. 138p.

SINGER, T. Information engineering: the search for business intelligence, 2001. Plant Engineering, 34–36.

SOARES DE MELLO, J. C. C. B.; ANGULO MEZA, L.; GOMES, E.G.; BIONDI NETO, L. Curso de análise envoltória de dados, 2005.

TAURION, C. Big Data. Rio de Janeiro: Brasport, 2013. https://pt.scribd.com/doc/259741402/Big-Data-Cezar-Taurion.

THANASSOULIS, E. A comparison of regression analysis and data envelopment analysis as alternative methods of performance assessment, 1993. Journal of the Operational Research Society 44 (11), pp. 1129-1144.

THANASSOULIS, E. Introduction to the theory and aaplication of data envelopment analysis: A foundation text with integrated software. New York: Kluwe Academic, 2001.

TSOUKIÀS, A. “De la théorie de la décision à l'aide à la décision”, in D. Bouyssou, D. Dubois, M. Pirlot, H. Prade (eds.), Concepts et Méthodes pour l'Aide à la Décision, Hermés, Paris, 25 - 69, 2006.

TURBAN, E.; ARONSON, J. E. Decision Support Systems and Intelligent Systems, 1998.

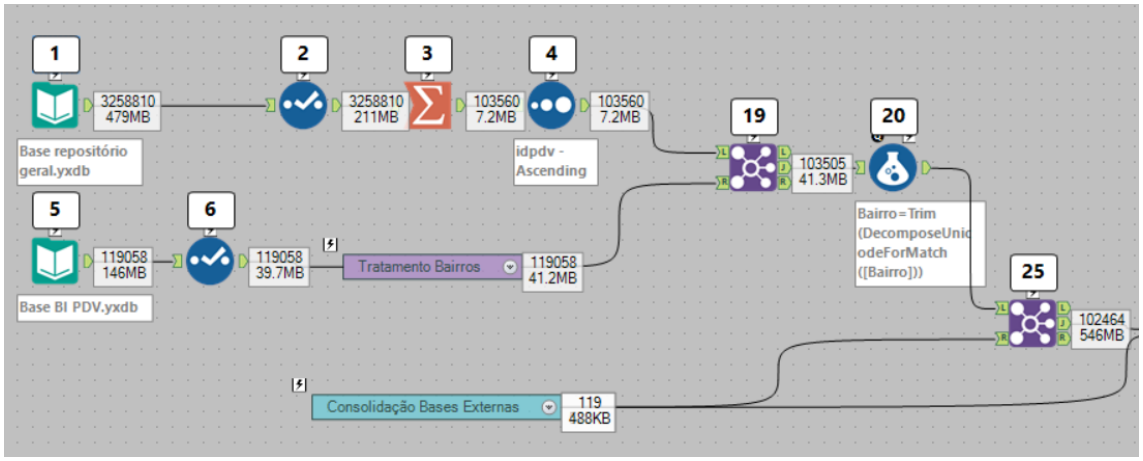
VARIAN, H.R. Microeconomic analysis. New York: W.W. Norton, 1992.

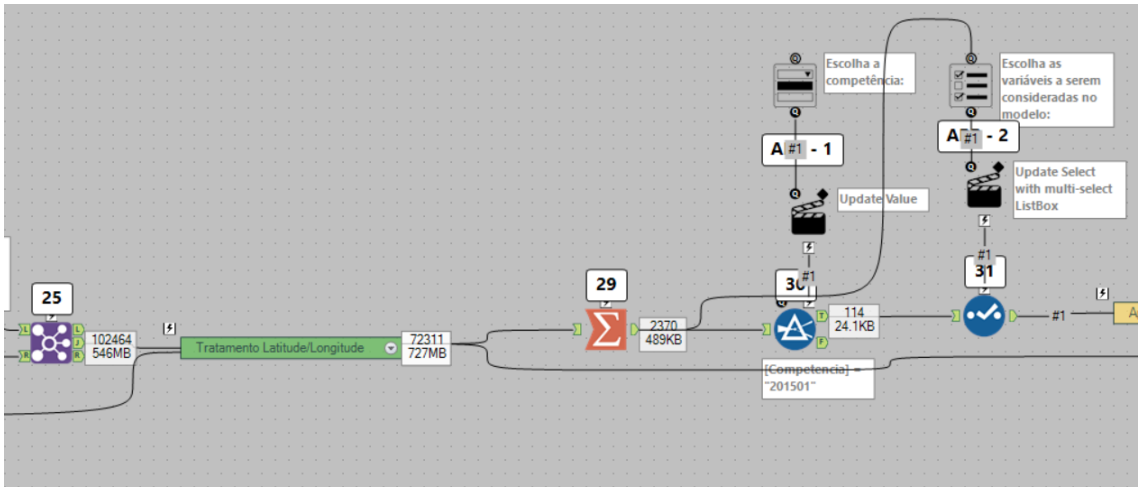
WANG, H.; XU, Z.; FUJITA, H.; LIU, S. Towards felicitous decision making: An overview on challenges and trends of Big Data, 2016.

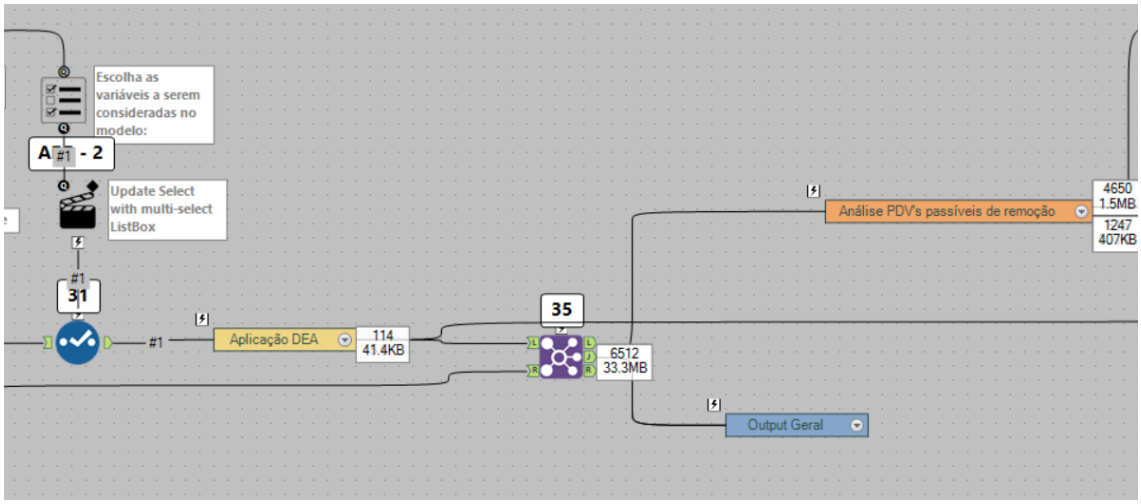
WILHELM, V. E. 2000. Análise da eficiência técnica em ambiente difuso. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2000.

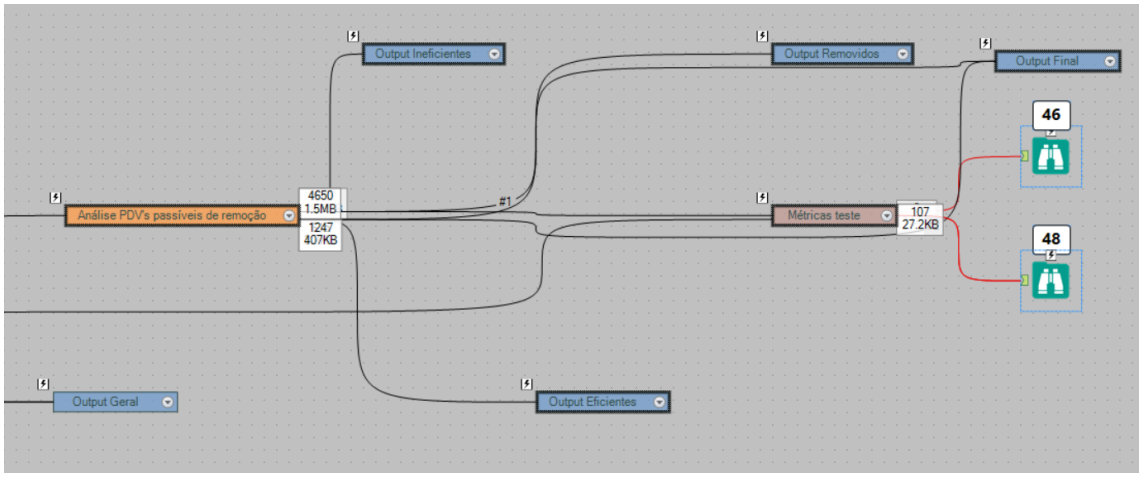
YEOH, W.; KORONIOS, A. Critical success factors for business intelligence systems, 2010. Journal of Computer Information Systems, 50(3), 23–32.

APÊNDICE A – FLUXO VERSÃO FINAL






Escolha as variáveis a serem consideradas no modelo:
A #1 - 2
Update Select with multi-select ListBox
#1
31
#1
Aplicação DEA
114
41.4KB
35
6512
33.3MB
Análise PDV's passíveis de remoção
4650
1.5MB
1247
407KB
Output Geral


Output Ineficientes
Output Removidos
Output Final
46
48
Análise PDV's passíveis de remoção
4650
1.5MB
1247
407KB
#1
Métricas teste
107
27.2KB
Output Geral
Output Eficientes