UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Allan do Amaral Alves

Implementação de WebGIS para análise de mercado e processo de compra e venda

Recife

2020

Allan do Amaral Alves

Implementação WebGIS para análise de mercado e processo de compra e venda

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco
UFRPE Departamento de Estatística e Informática Curso de Bacharelado em Sistemas de Informação

Orientadora: Roberta Macêdo Marques Gouveia
Coorientadora: Maria da Conceição Moraes

Recife

2020

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

A474i Alves, Allan
Implementação de WebGIS para análise de mercado e processo de compra e venda / Allan Alves. - 2020.
63 f. : il.

Orientadora: Roberta Macedo Marques Gouveia.
Coorientadora: Maria da Conceicao Moraes.
Inclui referências e apêndice(s).

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Recife, 2022.

1. GIS. 2. Análise de compra e venda. 3. K-means. 4. Clustering. 5. Base de dados geográficos. I. Gouveia, Roberta Macedo Marques, orient. II. Moraes, Maria da Conceicao, coorient. III. Título

CDD 004

ALLAN DO AMARAL ALVES

IMPLEMENTAÇÃO DE WEBGIS PARA ANÁLISE DE MERCADO E PROCESSO DE COMPRA E VENDA

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em: 05 de Novembro de 2020.

BANCA EXAMINADORA

Roberta Macêdo Marques Gouveia (Orientadora)
Maria da Conceição Moraes (Coorientadora)
Departamento de Estatísitca e Informática
Universidade Federal Rural de Pernambuco

Rodrigo Gabriel Ferreira Soares
Departamento de Estatísitca e Informática
Universidade Federal Rural de Pernambuco

Agradecimentos

Dedico este trabalho à minha família pelo amor, apoio e preocupação mesmo que à distância. À Deus por sempre me levar ao caminho de aprendizado e solução. Aos meus irmãos Alexsandra e Júnior pela cobrança e preocupação. Às minhas orientadoras Roberta Macêdo e Conceição Moraes pelo suporte, paciência e por acreditar no desenvolvimento deste trabalho. À Rodolpho pelo apoio constante e por sempre tentar me manter calmo. Aos meus gatinhos Alejandro e Fernando por atrapalhar mas também me acalmar nos momentos de ansiedade e estresse. À Ariane e Olyntho pela torcida na conclusão deste projeto. À Marcela, Isabella e Wagner pela parceria durante a graduação até os dias atuais. À minha tia Inalva, que sempre esteve na cobrança e na torcida pelas minhas conquistas e certamente continua torcendo por mim lá em cima.

Por último e mais importante, agradeço à minha mãe, a minha maior protetora. Quem se preocupa infinitamente comigo e estará sempre na torcida por novas vitórias pessoais. Este trabalho é dedicado principalmente à ela.

Resumo

Com a crescente utilização de plataformas para comércio eletrônico no país e as diversas crises econômicas afetando o número de vendas dos estabelecimentos desde 2014, pequenas e grandes companhias do varejo se vêem com a necessidade de realizar uma análise cada vez mais cuidadosa do ambiente em que estão inseridas, a fim de identificar os potenciais compradores de seus produtos em perfil, localização geográfica e outros atributos para otimizar o direcionamento dos seus serviços, prevendo as possíveis alterações de demanda e obtendo um menor risco perante os investimentos realizados. Com o acompanhamento de softwares mais avançados e a tecnologia atual, sistemas de informação geográfica se tornaram aliados para o estudo de grandes bases de dados, gerando resultados que auxiliam a tomada de decisão destas empresas. Este trabalho tem por objetivo implementar uma aplicação WEBGIS para análise de dados e resgate de significativas informações geográficas, utilizando algoritmo clustering para calcular e simular cenários de melhoria, identificar regiões com mais compradores e indicar as melhores localizações para venda de produtos classificados em diversos setores da região metropolitana do Recife através de dados existentes em notas fiscais eletrônicas.

Palavras-chave: webgis, notas fiscais, sistema de informações geograficas, k-means, banco de dados, informação.

Abstract

With the growing use of e-commerce platforms in the country and the various economic crises affecting the number of establishments sales since 2014, small and large retail companies are faced with the need to carry out an increasingly careful analysis of the environment in which they are located, in order to identify potential buyers of products in profile, geographic location and other attributes to optimize the direction of your services, anticipating possible changes in demand and obtaining a lower risk in association to the investments made. With the current technology, geographic information systems have become allies for the study of large databases, generating results that help the decision making of these companies. This work aims to implement a WEBGIS application for data analysis and rescue of significant geographical information, using a clustering algorithm to calculate and simulate improvement scenarios, identify regions with more buyers and indicate the best locations for selling classified products in different sectors of the market. metropolitan region of Recife using data from electronic invoices.

Keywords: webgis, invoices, geographic information system, k-means, database, information.

Sumário

Lista de ilustrações

Lista de abreviaturas e siglas

API - Interface de programação de aplicações

B2B - *Business-to-business*

BD - Base de dados

CNPJ - Cadastro Nacional da Pessoa Jurídica

PHP - *PHP: Hypertext Preprocessor*

NF-e - Nota fiscal eletrônica

SIG - Sistema de Informação Geográfica

SQL - Linguagem de Consulta Estruturada

XML - Linguagem de Marcação Extensível

UF - Unidade da Federação

1 Introdução

Neste capítulo tem-se a introdução ao tópico abordado no trabalho e a apresentação de motivações e justificativas para realização deste estudo de caso.

1.1 Apresentação

Foi possível acompanhar, nas últimas décadas, uma expansão gradativa no uso da internet pela população brasileira. Tal informação pôde ser confirmada através de uma recente pesquisa realizada pela TIC Domicílios [1], com dados indicadores da população brasileira em 2019. A partir da pesquisa, obteve-se a constatação de que 74% da população acima dos 10 anos de idade já utiliza a internet através de dispositivos móveis ou computadores. Tal crescimento deve prosseguir nos próximos anos e aumentar a utilização de conteúdos digitais também nas zonas rurais (onde, atualmente, apenas 53% da população possui internet). Estes indicadores de acesso acabam por apontar uma mudança comportamental da população, que inclui uma consequente mudança no processo de contratação de serviços e do comércio em geral, com indicação de que 39% dos usuários da internet realizaram alguma transação de compra online no ano indicado. As demonstrações da Figura 1 e da Figura 2 trazem as principais atividades realizadas na internet e os principais serviços consumidos pela população brasileira através da internet em 2019.

As informações citadas demonstram que a sociedade passou, gradativamente, a confiar em transações online e optar pelos serviços de ecommerce pela possibilidade de ter maior comodidade durante a pesquisa, realizar a compra de produtos a qualquer horário do dia e ter uma cobertura de estoque em maior abrangência, evitando situações onde a rede comerciante detém o produto almejado mas a loja física mais próxima do comprador está com as unidades esgotadas. Além disso, o e-commerce trouxe a praticidade de receber as aquisições em domicílio.

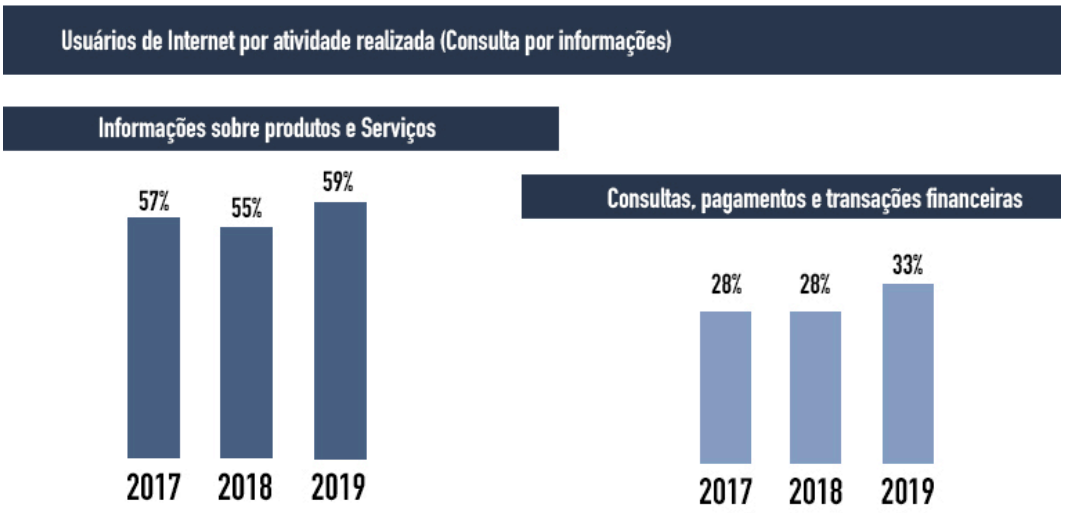

Figura 1: Quadro com as atividades online mais realizadas pelos brasileiros entre os anos 2017 e 2019 [1].

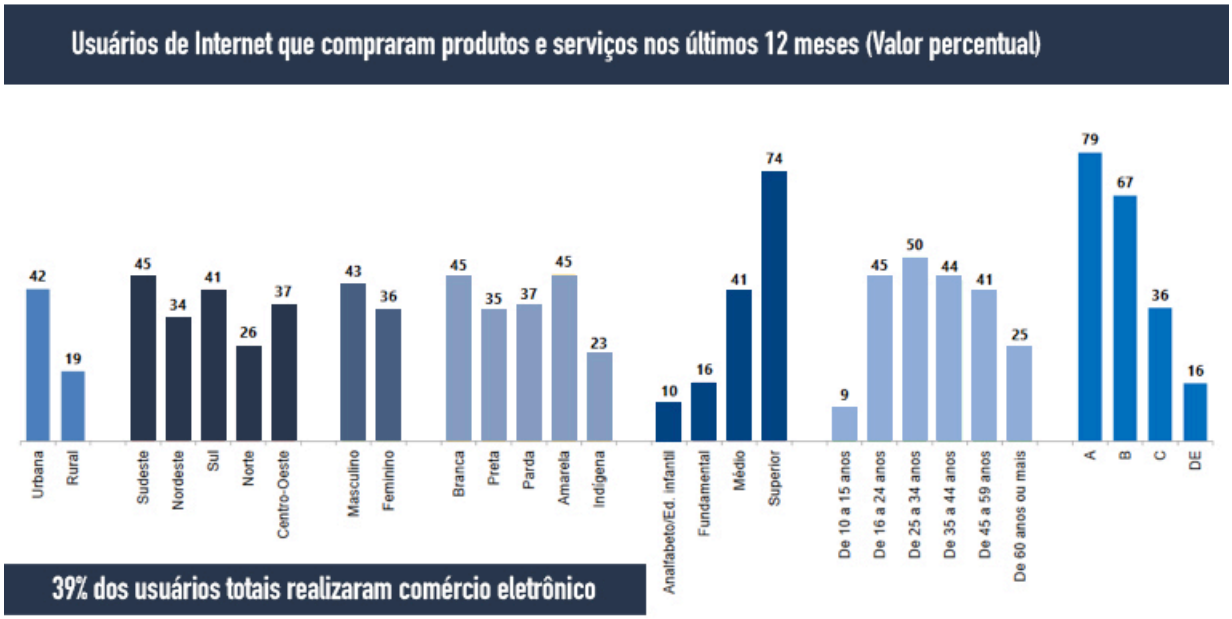

Figura 2: Quadro com as categorias de usuários que compraram produtos e serviços pela internet em 2019 [1].

Segundo a Associação Brasileira de Comércio Eletrônico ABComm em janeiro deste ano [2], o crescimento do setor de comércio eletrônico em 2019 superou as expectativas e as previsões para o ano seguinte seriam de contínuo crescimento, com um aumento de 18% e uma movimentação acima dos 105 bilhões em vendas. Ao analisar anos anteriores, é possível encarar uma realidade expressivamente menor. Em 2011, mesmo com a informática já incluída no cotidiano da população, o

faturamento final do comércio eletrônico brasileiro foi de R$ 18,5 bilhões. Tais informações acabam por demonstrar um crescimento constante e significativo mesmo em momentos de crise econômica no país, que ocorre desde o ano de 2014 [3], enquanto as vendas no varejo sofrem um impacto muito maior. A Figura 3 exibe a variação nas vendas do varejo no Brasil, que apontou em 2010 um crescimento de 10% mas reduziu a uma variação negativa nos anos de 2015 e 2016. Enquanto isso, a Figura 4 exibe a variação das vendas online no Brasil, onde o crescimento se manteve nos anos de 2015 e 2016 ainda que em menores proporções.

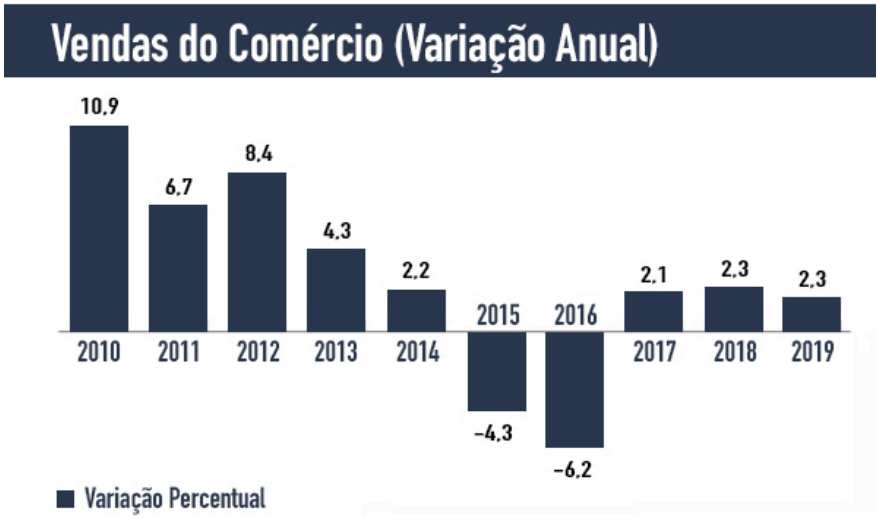

Figura 3: Variação percentual de vendas no mercado brasileiro de varejo entre 2010 e 2019 [4].

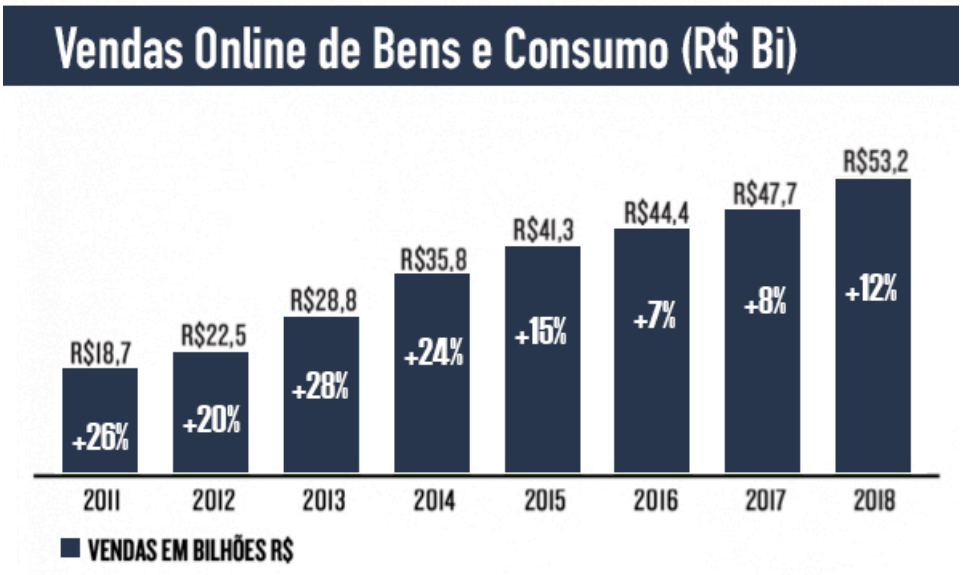

Figura 4: Variação percentual de vendas online no mercado brasileiro entre 2011 e 2018 [5].

O ano de 2020 acabou por traçar direções imprevistas a partir de uma situação inédita à nível global causada pela pandemia do vírus Covid-19, originado na China [6]. Tal problemática exigiu mudanças comportamentais para a população de diversos países do mundo, incluindo o Brasil, com a adesão de uma reclusão geral pela sociedade a fim de reduzir ao máximo o contágio deste novo vírus. O encerramento de centros comerciais e a redução no número de visitantes à lojas físicas foi algo extremamente indicado pela Organização Mundial da Saúde durante a fase de quarentena. Com esta mudança, a população em geral acabou sendo impulsionada a realizar diversos processos online, incluindo compras. Este cenário trouxe um impulsionamento para a camada que, em teoria, levaria mais alguns anos para incluir as transações online em seu cotidiano.

A aceleração deste aumento na utilização de redes online para o varejo nacional passou a render altos índices para o comércio eletrônico mesmo em meio a uma crise econômica alastrada pela pandemia. Segundo BOWLES [7], foi possível identificar, através de um estudo da consultoria de gestão estratégica Kearney, os impactos da Covid-19 no comportamento de consumo dos brasileiros. Ele indica que as compras online devem registrar R$ 111 bilhões em 2020 — 49% a mais do que em 2019. Para os quatro anos seguintes, a análise aponta que o mercado deve continuar com crescimentos à uma taxa de 17,3% ao ano.

Este novo hábito de compras virtuais, facilitado com a popularização dos smartphones nos interiores do país [1], trouxe para as companhias de varejo a necessidade de realizar uma mudança estratégica, que consiga suprir as dificuldades de rendimento mediante a redução de visitas a estabelecimentos físicos e, consequentemente, atualizar o processo de vendas existentes para o mercado em constante crescimento.

A década de 2000 apresentou diversos projetos de plataformas comerciais para vendas online por lojas que já existiam no varejo. O processo eletrônico que serviria como complemento comercial para as companhias já renomadas acabou por não receber grande enfoque de investimento e, consequentemente, atrapalhar o crescimento destas com o desenvolvimento posterior do setor. De exemplos mal sucedidos, temos o do primeiro portal de compras Carrefour, que precisou encerrar as atividades online em 2012 [9] e retornou apenas cinco anos depois, e a falência da maior loja de brinquedos americana "Toy 'r' us" [10]. Para uma grande parte das redes físicas do mercado, no entanto, não ocorreram grandes problemas na disponibilização de uma plataforma digital própria.

O comércio eletrônico, como anteriormente citado, traz ao consumidor uma variedade de vantagens. Já para os vendedores, existe a expansão na amplitude de consumidores, com estoques consultados em qualquer região do país através de uma plataforma digital e aumento de visibilidade dos produtos. Tal fator, ao mesmo tempo que aumenta a exibição dos produtos de uma loja para diversos clientes em potencial, pode também contrastar a diferença de preços existente para um mesmo produto dentre as diversas lojas do setor e prejudicar vendas de uma companhia que conseguia utilizar preços mais elevados a partir da não existência de lojas físicas com estoque semelhante.

A necessidade de maior cuidado estratégico mediante uma competição comercial global, unida à baixos resultados de vendas desde 2014, serviu de alerta para pequenas e médias empresas estarem dispostas a realizar um estudo mais aprofundado dos compradores em potencial e das diferentes formas de otimizar seus serviços no objetivo de reduzir custos, atender a um maior número de clientes e prever valores de demanda. Para uma empresa de varejo com loja física e e-commerce, torna-se importante conhecer o ambiente no qual atua e quais os locais com maior densidade de usuários dos seus serviços na hora de optar por uma mudança de localização física ou mesmo do centro de distribuição online, permitindo uma redução no custo de entrega e possibilitando um menor preço aos compradores ou um atrativo de entregas mais ágeis para eles.

Com esta necessidade de análises mais precisas do poder de compra da sociedade, classificando, dentre outras categorias, geograficamente, surgiu a possibilidade de utilizar dados resgatados da receita federal através dos projetos de pesquisa realizados por LIRA (2016) [11] e NETO (2018) [12]. Tais projetos foram responsáveis por retornar e estruturar uma gama de dados de notas fiscais eletrônicas pertencentes a compras de diversos setores do mercado na Região Metropolitana do Recife. Estas notas podem ter seus atributos analisados, transformando em informação significativa no apoio a tomada de decisão empresarial.

Nos trabalhos citados, foram resgatados dados referentes a mais de 3 milhões de itens de compra da região metropolitana do Recife durante o período de 2010 a 2015. O objetivo definido no trabalho atual é que, com a utilização destes dados, seja possível identificar pontos onde setores específicos do varejo possuam um maior número de compradores em potencial, além de traçar possíveis melhorias no rendimento da companhia a partir de mudanças geográficas do seu negócio. O uso de dados de cinco anos atrás unido ao fato de que tal tipo de informação é facilmente

mutável com novos padrões de compra da sociedade, os resultados apresentados por uma análise podem não ser eficazes ou coerentes com a realidade atual. A partir desta possível defasagem dos dados, este trabalho se propõe a implementar uma aplicação que realize a análise dos dados de compra, por meio de notas fiscais, em uma base de dados dinâmica, possibilitando a atualização do banco de dados sempre que necessário para um retorno de informações mais recentes. Tais objetivos definiram a necessidade de implementar um Sistema de Informação Geográfica (SIG) que seja capaz de realizar agrupamentos e cálculos geográficos dinamicamente para contemplar simulações e resultados baseados em diferentes contextos.

Um Sistema de Informação Geográfica (SIG) permite a visualização, consulta e análises de dados associados a pontos geográficos para compreender relações, padrões e tendências entre eles. Os SIGs auxiliam organizações em decisões estratégicas de qualquer dimensão e setor, o que desperta um interesse crescente por tais aplicações. Com o ponto de atuação definido e a base de dados a ser trabalhada já disponível, a implementação do SIG neste trabalho de conclusão se tornou viável.

1.2 Motivação e Justificativa

O avanço tecnológico e a necessidade do setor de varejo em aprimorar os métodos de pesquisa de mercado e seu relacionamento com o cliente geraram uma motivação para análises com maior eficácia e dinamismo. É possível afirmar que a análise de dados reais de compra a partir das notas fiscais eletrônicas pode ser um fator contributivo nos resultados e informações geradas, visto que os procedimentos mais comuns para pesquisa de mercado se baseiam em questionários públicos. Tais pesquisas de opinião nem sempre conseguem atingir respostas próximas da realidade, por não serem realizados com o público alvo da companhia, por serem efetuados com uma amostra pequena de consumidores cuja opinião difere da opinião geral ou pelas respostas dos entrevistados não corresponderem com seus reais comportamentos como consumidores. Além disso, tais métodos de pesquisa são custosos, levam um tempo considerável para retornar os resultados e dificilmente conseguem ter seus objetivos adaptados durante o processo.

1.3 Objetivos

A partir dos pontos segmentados acima, pode-se formar o seguinte objetivo principal do trabalho: *Avaliação espacial dos dados de compra e venda da região Metropolitana do Recife com informações que auxiliem na identificação de melhorias no processo e no reconhecimento de tendências do mercado, através da identificação geográfica dos compradores e suas distâncias com a empresa distribuidora, proporcionando um sistema de apoio à decisão para empreendedores;*

O trabalho é dividido nos seguintes objetivos específicos:

- Realizar o pré-processamento dos dados de notas fiscais eletrônicas já existentes e documentar as condições para análise desta base de dados;
- Agrupar os dados de notas fiscais por localização com o objetivo de extrair conhecimento e identificar pontos de convergência na base de dados que sirvam de auxílio para o estudo do mercado online.
- Implementar SIG para análise dos dados válidos da base refatorada, gerando novas informações geográficas e estatísticas;
- Possibilitar a análise dos dados em diferentes cenários, contemplando possíveis atualizações na base de dados do sistema;

1.4 Estrutura do Trabalho

Além da introdução, o presente trabalho de conclusão de curso (TCC) apresenta os seguintes tópicos:

- Capítulo 2 - Referencial Teórico, apresentando definições e exemplos para os principais temas abordados;
- Capítulo 3 - Trabalhos relacionados, realizando uma breve descrição dos artigos científicos, teses e/ou dissertações associadas ao tema do TCC e o que foi relevante para o desenvolvimento;
- Capítulo 4 - Metodologia, com a definição das ferramentas e processos aplicados para a implementação do trabalho e o pré-processamento dos dados utilizados;
- Capítulo 5 – Desenvolvimento, com a descrição dos passos de implementação do projeto e informações sobre sua programação;

- Capítulo 6 - Resultados e discussões, com a conclusão do trabalho e quais informações relevantes foram retornadas a partir do que foi implementado;

2 Referencial Teórico

Este capítulo tem como objetivo explanar os principais conceitos associados ao trabalho, aprofundando sua fundamentação teórica através de definições encontradas em livros, artigos e documentações oficiais.

2.1 Nota Fiscal Eletrônica

A Nota Fiscal Eletrônica (NF-e) é um documento eletrônico que contém dados do contribuinte remetente, do destinatário e da operação a ser realizada [13]. A NF-e enquadra-se na convergência dos objetivos do projeto de modernização da administração tributária e aduaneira da receita federal do Brasil e dos correspondentes programas de modernização da administração 35 tributária dos estados e municípios [14]. Sua finalidade é definir e implementar as regras de negócio, requisitos e funcionalidades de um processo informatizado para a emissão e controle de NFs em formato eletrônico.

Segundo o portal de notas fiscais da receita federal, o Projeto NF-e tem como objetivo a implantação de um modelo nacional de documento fiscal eletrônico que venha substituir a sistemática atual de emissão do documento fiscal em papel, com validade jurídica garantida pela assinatura digital do remetente, simplificando as obrigações acessórias dos contribuintes e permitindo, ao mesmo tempo, o acompanhamento em tempo real das operações comerciais pelo Fisco.

O portal ainda ressalta as principais vantagens com a conversão de emissões de notas fiscais do modo convencional para o eletrônico:

- Para as Administrações Tributárias: Aumento na confiabilidade da Nota Fiscal; Melhoria no processo de controle fiscal, possibilitando um melhor intercâmbio e compartilhamento de informações entre os fiscos; Redução de custos no processo de controle das notas fiscais capturadas pela fiscalização de mercadorias em trânsito; Diminuição da sonegação e aumento da arrecadação; Suporte aos projetos de escrituração eletrônica contábil e fiscal da Secretaria da Receita Federal do Brasil (RFB).

- Para a Sociedade: Redução do consumo de papel, com impacto positivo no meio ambiente; Incentivo ao comércio eletrônico e ao uso de novas tecnologias; Padronização dos relacionamentos eletrônicos entre empresas; Surgimento de oportunidades de negócios e empregos na prestação de serviços ligados à Nota Fiscal Eletrônica.
- Para o Contribuinte Comprador (Receptor da NF-e): Eliminação de digitação de notas fiscais na recepção de mercadorias; Planejamento de logística de entrega pela recepção antecipada da informação da NF-e; Redução de erros de escrituração devido a erros de digitação de notas fiscais; Incentivo ao uso de relacionamentos eletrônicos com fornecedores (B2B);
- Para o Contribuinte Vendedor (Emissor de NF-e): Redução de custos de impressão; Redução de custos de aquisição de papel; Redução de custos de envio do documento fiscal; Redução de custos de armazenagem de documentos fiscais; Simplificação de obrigações acessórias, como dispensa de AIDF; Redução de tempo de parada de caminhões em Postos Fiscais de Fronteira; Incentivo a uso de relacionamentos eletrônicos com clientes (B2B);

Definidos como conjuntos de dados inter-relacionados, os bancos de dados são organizadas de forma a permitir que sistemas de aplicação armazenem novos dados, encontrem dados armazenados, alterem seu conteúdo ou excluam dados indesejáveis por meio de métodos precisos de manipulação e localização [15].

Um banco de dados com seus valores tratados e filtrados, de acordo com as necessidades de um contexto, torna-se imprescindível para retornar informações que auxiliem em uma tomada de decisão.

2.2 Agrupamento de dados e K-Médias

Ao trabalhar com uma grande quantidade de dados e com o objetivo de obter informações significativas, é importante que eles estejam formatados e agrupados da forma mais eficaz possível. Segundo ORTEGA [16], Clustering é um dos principais procedimentos para obter informações sobre a natureza e a estrutura subjacentes dos dados. O processo decorre da organização de um conjunto de dados em grupos menores, chamados clusters, de forma que os elementos agrupados em cada cluster sejam semelhantes entre si e diferentes dos que estão em outros clusters a partir de aspectos determinados.

O Clustering é uma técnica existente no aprendizado de máquina (*Machine Learning*) que, por sua vez, pode ser definido como um conjunto de métodos capazes de detectar automaticamente padrões em dados e, em seguida, utilizá-los para prever dados futuros ou para realizar outros tipos de tomada de decisão sob critérios de incerteza [17]. Baseando-se no princípio de aprendizado por indução, onde a inferência lógica permite que conclusões gerais sejam obtidas de exemplos particulares e possa partir de um plano específico para o geral, o aprendizado de máquina em seu modelo de indução por exemplos reais pode ser classificado em dois tipos principais, Supervisionado e Não Supervisionado. O aprendizado supervisionado recebe em seu algoritmo um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido. Segundo MONARD [18], para o aprendizado supervisionado, “Cada exemplo é descrito por um vetor de valores de características, ou atributos, e o rótulo da classe associada. O objetivo do algoritmo de indução é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados, ou seja, exemplos que não tenham o rótulo da classe”. Enquanto isso, para o aprendizado não supervisionado, “o indutor analisa exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou clusters. Após a determinação dos agrupamentos, normalmente, é necessária uma análise para determinar o que cada agrupamento significa no contexto do problema que está sendo analisado.”

Este trabalho aplica o aprendizado não supervisionado para seu processo de análise de dados, utilizando-se de métodos de *clustering* conforme a necessidade do contexto existente. Dentre os diversos métodos de agrupamento, o K-means é amplamente utilizado para agrupamento de dados geográficos e se destaca pela variedade de exemplos e informações sobre suas implementações existentes dentre as plataformas de programação. A partir da descrição por ADAMS [19], K-means é um algoritmo iterativo que consiste em realizar ciclos (*loops*) de processamento de dados até convergir para uma solução (supostamente ideal). Em cada ciclo de processamento, são feitos dois tipos de atualizações: uma sobre os vetores do conjunto com o objetivo de identificar o ponto aleatório mais próximo, e outra para os pontos aleatórios inicialmente traçados, para ajustar suas posições para um posição média perante o agrupamento que o cerca. Um pseudocódigo de exemplo para o K-means pode ser demonstrado na Figura 5 a seguir:

```
k-means(CD, k, AG)

  input:
  CD= {D1, D2, ..., DN}  //conjunto de instâncias de dados a serem agrupadas.
  k                      // número de grupos a ser criado.

  output:
  AG ={G1,G2,...Gk}    //agrupamento formado por k grupos de instâncias de dados.

begin

  (1) escolher arbitrariamente k instâncias CD, cada um como centroidedos
  grupos G1,G2,...Gk       //após (1) cada um dos k grupos contém apenas o centroide
  (2) repeat
    (3) (re)atribuir cada instância Di CD ao grupo associado ao centroide que lhe
    seja mais próximo;
    (4) atualizar os centroides de cada um dos k grupos, como a média dos valores dos
    atributos entre as instâncias a ele associados;
    (5) until nenhuma alteração aconteça no agrupamento;
  end

  return AG = {G1,G2,...Gk}

end
```

Figura 5: Pseudocódigo de exemplo para o processo de agrupamento via K-means [20].

A idéia básica é conseguir dividir o conjunto de dados em grupos formados a partir de determinado critério e identificar qual o ponto médio do conjunto para cada grupo formado. O processo de implementação do algoritmo no projeto é descrito com mais detalhes no tópico de metodologia deste trabalho.

2.3 Sistema de Informação Geográfica (SIG)

O Sistema de Informação Geográfica, também referenciado como GIS (*Geographic Information Systems*) está associado a diversas áreas da ciência e, com isso, não possui uma definição única.

Uma das definições mais antigas encontradas para um Sistema de Informação Geográfica é de 1981 e, de acordo com OZEMOY, SMITH e SICHERMAN [21], GIS pode ser descrito como um conjunto de funções automatizadas, que fornecem aos profissionais capacidades avançadas de armazenamento, acesso, manipulação e visualização de informação georreferenciada.

Enquanto isso, DAVIS [22] afirma que existem diversas definições para GIS e que nenhuma delas pode descrever ou explicar suficientemente o termo. Às vezes é dito "mapeamento computacional avançado" mas uma melhor definição é "Uma tecnologia e metodologia computacional para coletar, gerenciar, analisar, modelar e apresentar dados geográficos para uma grande escala de aplicações

Para CASANOVA [23], a principal diferença de um SIG para um sistema de informação convencional é sua capacidade de armazenar tanto os atributos descritivos como as geometrias dos diferentes tipos de dados geográficos. Arquitetura de sistemas de informação geográfica. Do ponto de vista da aplicação, o uso de sistemas de informação geográfica (SIG) implica em escolher as representações computacionais mais adequadas para capturar a semântica de seu domínio de aplicação. Do ponto de vista da tecnologia, desenvolver um SIG significa oferecer o conjunto mais amplo possível de estruturas de dados e algoritmos capazes de representar a grande diversidade de concepções do espaço. A Figura 6 apresenta os componentes de arquitetura existentes em um banco de dados geográfico. A interface se encontra como componente mais externo, onde se define como o sistema deve ser operado visualmente pelo usuário. No intermédio, o sistema deve prever mecanismos de entrada, consulta, análise e visualização dos dados espaciais. Como camada interna, o SIG deve dispôr de um processo de gerência de bancos de dados geográficos que ofereça o armazenamento e recuperação dos dados espaciais trabalhados e seus atributos.

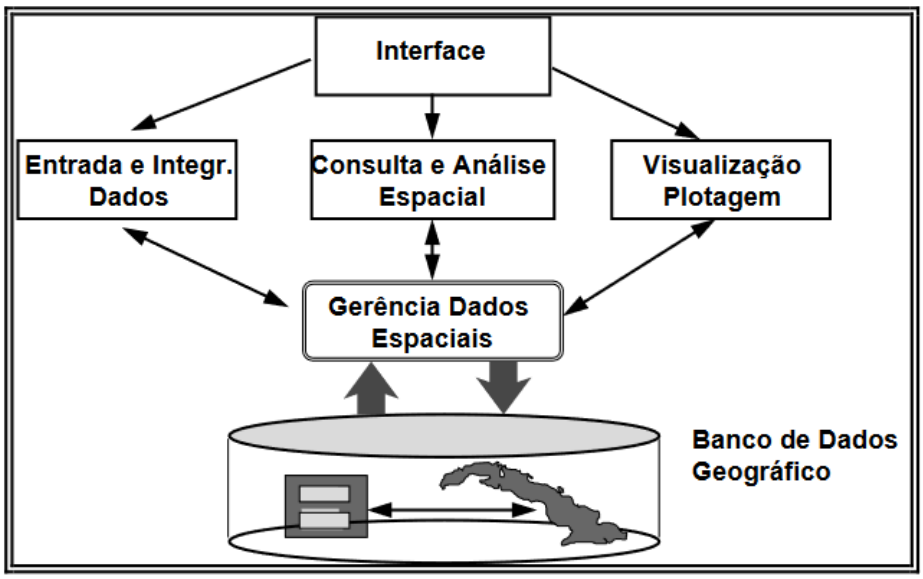

Figura 6: Arquitetura de sistemas de informação geográfica [23].

Com isso temos definições e referências básicas para a utilidade de um SIG. Hoje, mais do que nunca, temos uma vasta quantidade de dados disponíveis para os mais diversos serviços. Para uma análise eficaz de grandes conjuntos de dados que estejam, de alguma forma, associados a pontos e localizações geográficas, faz-se necessária a implementação de um SIG, reconhecendo padrões e os possíveis

conhecimentos gerados a partir do estudo dos atributos destes pontos geográficos em relação ao tema abordado.

2.3.1 Mapas de Kernel

Os mapas de Kernel são também conhecidos como mapas de calor e equivalem a exibição geográfica de intensidade da ocorrência de eventos a partir de dados relacionados ao contexto específico. A palavra Kernel significa “núcleo” e está associada a nomenclatura do mapa a partir da sua associação com o processo de Estimativa de Densidade Kernel (EDK) [24]. Neste processo estatístico para estimar curvas de densidades, cada observação é ponderada através da distância perante o núcleo de estudo.

Medeiros [25] complementa: Dito de forma simples, o Mapa de Kernel é uma alternativa para análise geográfica do comportamento de padrões. No mapa é plotado, por meio de métodos de interpolação (estimativa do valor de um atributo em locais não amostrados, a partir de pontos amostrados na mesma área ou região [26]), a intensidade pontual de determinado fenômeno em toda a região de estudo. Assim, temos uma visão geral da intensidade do processo em todas as regiões do mapa, conforme demonstrado na Figura 7. As regiões são preenchidas com as cores verde, amarelo e vermelho conforme sua concentração, onde verde representa uma densidade menor e vermelho uma densidade maior.

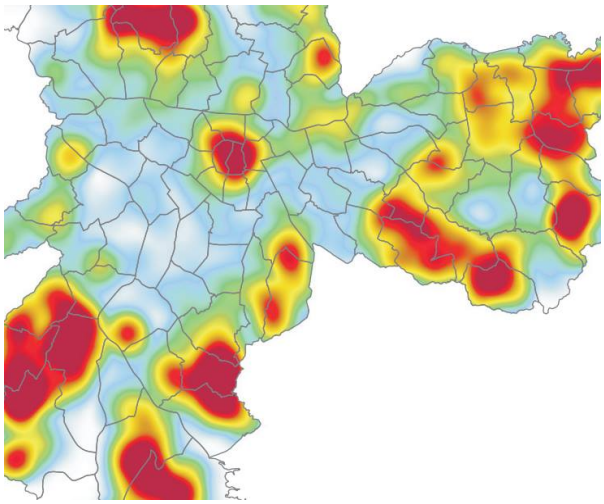

Figura 7: Exemplo de mapa de Kernell exibido em estudo da criminalidade do estado de São Paulo [27].

2.3.2 Framework e Padrão MVC

Segundo GAMMA [28], um framework equivale a um conjunto de classes cooperativas que formam um projeto reutilizável para uma categoria específica de software, fornecendo direcionamento arquitetural pelo particionamento do projeto em classes abstratas, definindo suas responsabilidades e colaborações. Assim, um desenvolvedor customiza um framework para uma aplicação específica, especializando e compondo instâncias destas classes.

Frameworks tornam-se essenciais para manter um padrão no processo de desenvolvimento e facilitar a criação e manutenção de código, além de fornecer bibliotecas que tornem diversas etapas da implementação de um projeto mais ágeis.

O padrão MVC (modelo-visão-controlador) foi desenvolvido por Trygve Reenskaug em 1979 e equivale a um padrão de arquitetura de aplicações que divide uma aplicação em três camadas: a visão (*view*), o modelo (*model*), e o controlador (controller) [29]. O padrão atribui responsabilidades distintas a cada camada e, segundo MACORATTI [30], estas podem ser descritas da seguinte forma:

- Model que representa os dados e não deve incluir detalhes de implementação podendo ter muitas Views associadas;
- A View que representa um componente de interface de usuário que está vinculado a um Model. Ela pode exibir os dados e permitir que a modificação dos dados pelo usuário. A View deve sempre refletir o estado do Model.
- Controller que fornece um mecanismo para o usuário interagir com o sistema definindo como a interface do usuário vai reagir à ação do usuário. Ele é responsável por trocar e interpretar mensagens entre a View e o Model.

2.3.3 Google Maps API

O Google Maps Application Programming Interface é um dos componentes do conjunto de serviços Google chamado ‘Google API’, um serviço público e gratuito executado através de API’s da organização que podem ser incorporadas a uma aplicação web de acordo com a necessidade existente. O site oficial do Google API[1] descreve seus serviços como:

- Google Maps JavaScript API - Incorpora um mapa do Google na página da web usando JavaScript. Manipula mapas e gera conteúdo com a ajuda de vários serviços. Através desta API, é possível implementar a exibição de

1 Site oficial Google API - https://console.cloud.google.com/apis/library

mapas de calor como no exemplo exibido na Figura 8.

- Serviços da web - Utiliza solicitações de URL para acessar informações de geocodificação, rotas, elevação e lugares dos aplicativos cliente e manipula os resultados em JSON ou XML.
- Google Maps Data API – Visualiza, armazena e atualiza dados de mapa por meio de feeds da Google Data API, usado um modelo de elementos (marcadores, linhas e formas) e coleções de elementos. Um exemplo demonstrativo de mapa com marcadores pode ser analisado na Figura 9.

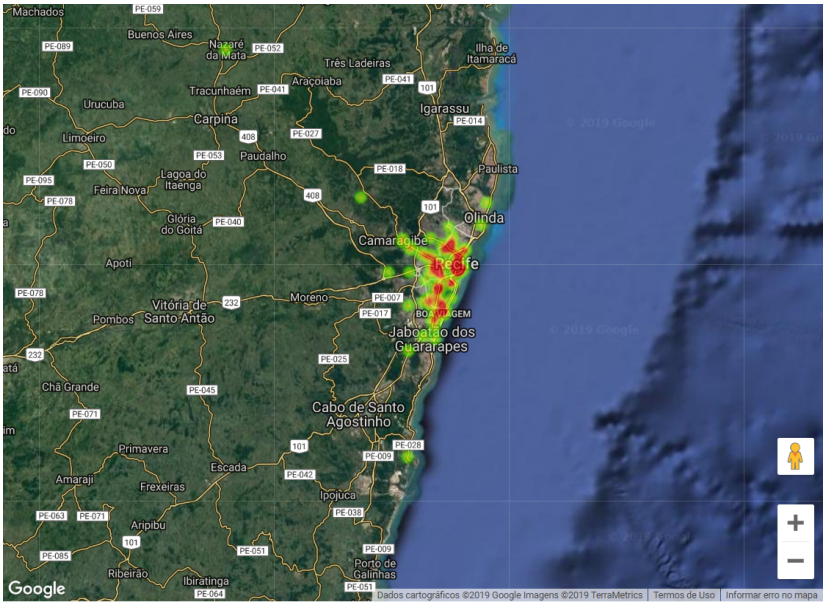

Figura 8: Implementação de mapas de calor com a utilização da ferramenta Google.

Fonte: Autoria própria.

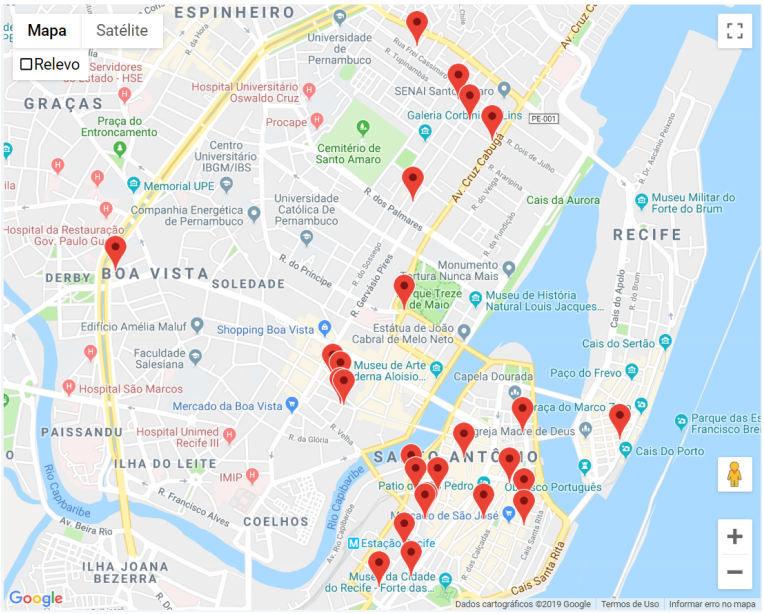

Figura 9: Implementação de mapas com marcadores específicos a partir da utilização da ferramenta Google. Fonte: Autoria própria.

2.3.4 High Maps

Enquanto o Google Maps API disponibiliza os seus serviços geográficos em API, o HighCharts é uma biblioteca de gráficos escritos em JavaScript, ofertando a implementação de gráficos complexos e mapas interativos em uma página web, chamados de *HighCharts Maps* ou simplesmente *HighMaps*. O serviço é executado apenas com a interpretação de seus arquivos javascript pelo *browser*, não precisando de *plugins* como Flash ou Java. A Figura 10 demonstra a exibição de um conjunto de dados a partir do componente *HighMaps* existente na biblioteca.

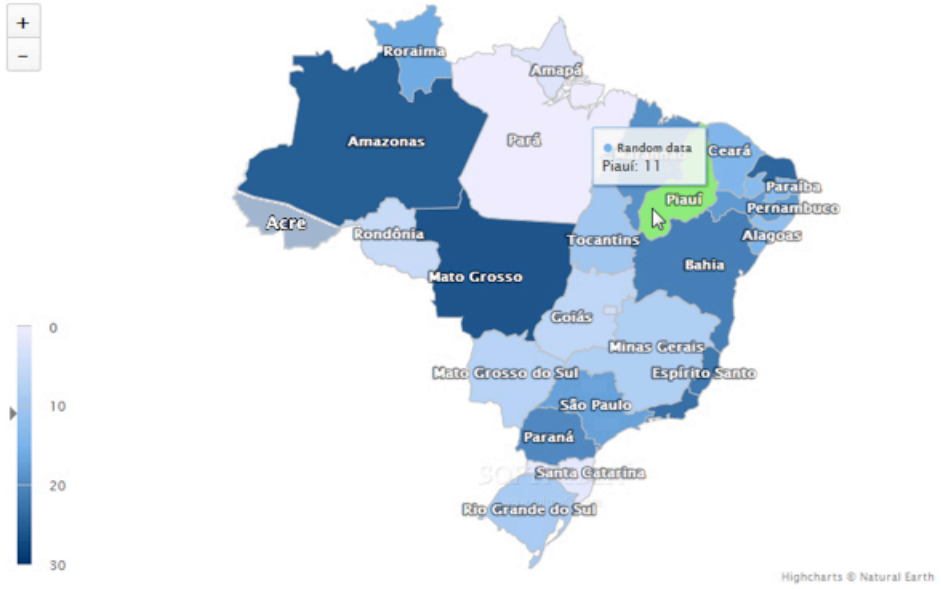

Figura 10: Simulação *HighMaps* do mapa Brasileiro com quantitativo de registros.
Fonte: Autoria própria.

3 Trabalhos Relacionados

O trabalho tem como escopo o tratamento, mineração e análise dinâmica dos registros existentes em uma grande base de dados de notas fiscais eletrônicas, a fim de retornar informações significativas do mercado consumidor que auxiliem na tomada de decisão de empresas do varejo e do comércio eletrônico.

Apesar do processo de notas fiscais ter sido convertido para o modo eletrônico há apenas uma década, foi possível encontrar importantes referências para o estudo do tema através de ferramentas de pesquisa, livros e artigos científicos. Neste tópico, são descritos os artigos que serviram de auxílio durante o processo de desenvolvimento deste trabalho de conclusão de curso.

O artigo realizado por LIRA [11] traz em detalhes um estudo do processo de implementação de modelos de dados multidimensional e não relacional para uma grande massa de dados. Para popular o banco criado, foram utilizados arquivos XML de notas fiscais eletrônicas. Já para cada um dos dois modelos de implementação estudados, a autora utilizou abordagens específicas como data warehouse para o modelo multidimensional e o banco de dados NoSQL HBase para implementar o modelo orientado a coluna. O projeto de LIRA está diretamente relacionado com este trabalho por fazer uso da mesma base de dados de NF-e. Apesar do foco deste não ser a comparação de performance de modelos de dados não relacionais e multidimensionais, o estudo do trabalho citado foi fundamental para conhecimento da semântica dos dados.

Após a implementação de LIRA [11] de bases de dados relacionais a partir de arquivos de notas fiscais eletrônicas, NETO [12] retoma a utilização deste *dataset* com o objetivo de obter a relação de demanda reprimida de produtos existentes nas NFe's, identificando os clientes interessados nas mercadorias do varejo que se encontram a uma distância elevada deles. A prova de conceito do estudo se baseou em três produtos específicos: gasolina de aviação (AVGAS), televisão de LED e diclorvol (inseticida). Foi possível identificar com as análises realizadas quais as cidades mais indicadas para uma iniciativa de ampliação de unidades das respectivas empresas distribuidoras. O trabalho de NETO [12] se torna importante por se tratar do processo de mineração de dados referentes ao mesmo contexto de notas fiscais eletrônicas. O atual trabalho se difere ao abordar uma solução dinâmica para a

análise dos dados desta base, com a utilização do processo de agrupamento automático K-means, a possibilidade de atualizações e, consequentemente, a indicação de informações atualizadas.

SOUZA [31] aborda um estudo para definição de propostas de melhoria na acessibilidade de pessoas com mobilidade reduzida na Universidade Federal Rural de Pernambuco. A pesquisa utilizou a implementação de um Sistema de Informação Geográfica para armazenar e identificar através de consultas específicas em Mapas de Kernel as áreas da universidade nas quais devem ser feitas ações de melhoria da acessibilidade ofertada. Com um contexto diferenciado, a autora utilizou ferramentas como GRASS e Quantum GIS para desenvolver um sistema geográfico com base na coleta de informações espaciais e avaliações específicas do local em análise. O trabalho torna-se importante para o presente TCC no processo de implementação de SIG's e uso de dados geográficos armazenados em um banco específico.

Para o processo de mineração de dados com a utilização de algoritmos de clustering como o K-means, foram analisados e considerados como base para implementação alguns artigos internacionais, como é o caso do projeto da Universidade de Harvard pelo mestrando SUD [32], envolvendo a população indiana e os preconceitos sociais existentes. Ao considerar que respostas indicadas em questionários nem sempre retratam a realidade de uma sociedade preconceituosa, o estudante resgatou uma gama de dados existente em um site de relacionamento e procura de matrimônio online da Índia a fim de dentificar, a partir da utilização do algoritmo de agrupamento K-means, o cenário mais realista acerca dos preconceitos sociais existentes e a disparidade entre o que se relata em pesquisas do país e o que se vivencia nas escolhas pessoais da população na procura por um relacionamento afetivo. O trabalho de SUD serviu como um exemplo de implementação do processo de agrupamento clustering a ser implementado neste projeto.

4 Metodologia

Este capítulo tem por objetivo descrever as metodologias definidas para o trabalho, a limpeza da base de dados utilizada e as principais decisões tomadas para o início de implementação do WebGIS.

4.1 Definição de requisitos

Como etapa inicial, foram levantados os objetivos principais citados anteriormente neste trabalho para, em seguida, realizar uma análise detalhada dos requisitos necessários para a implementação do SIG, além dos pontos de atenção a serem considerados durante o desenvolvimento.

O objetivo geral da aplicação descrito no tópico 1.3 pôde ser resumido da seguinte forma técnica: “Implementação de um SIG para análise dinâmica de uma grande base de dados de notas fiscais, com a utilização do algoritmo de agrupamento K-means para devolver informações significativas na tomada de decisão por empresas do setor de varejo e de comércio eletrônico”. A partir desta definição, foram elaborados os requisitos do projeto com as seguintes funcionalidades:

- Relatório geográfico de pontos de envio - Exibição de todos os emitentes de mercadorias existentes na base de dados estudada. A demonstração deve ser feita com exibição em três tipos de mapa: múndi, indicando a quantidade de registros existentes no *dataset* para o país e estados específicos, em mapas de calor com pontos de intensidade de registros e mapas com marcadores individuais por item existente;
- Relatório geográfico de compradores - Exibição de todos os destinatários de compras existentes na base de dados estudada. Equivalente ao requisito um, a demonstração deve ser feita com mapa-mundi, mapas de calor e mapas com marcadores individuais;
- Elaboração de relatórios de vendas - A aplicação deve possibilitar a partir de consultas automatizadas a base de dados trabalhada, a elaboração de relatórios ao usuário com gráficos e informações significativas sobre o processo de compra e venda da empresa ou região determinada como custo de distribuição e previsões de lucro;

- Importação de dados - O sistema deve permitir a inclusão de novos dados para a base de trabalho a partir de importações de registros em SQL enviados pelo usuário;
- Simulação de resultados - Também deve ser possível realizar uma simulação de resultados dos relatórios em caso de trocas de endereços de empresa, entre outras características que possam afetar seu cálculo de custos e lucros;
- Dashboard - Painel ilustrado com o resumo em gráfico do histórico de dados existentes no *dataset* e prévia de informações geradas pela aplicação;

Com as principais funcionalidades da aplicação enumeradas e a base de dados de estudo já disponível para utilização, o desenvolvimento pôde seguir adiante.

4.2 Base de Dados: Notas Fiscais Eletrônicas

A base de dados utilizada na aplicação foi recuperada a partir do projeto de pesquisa da Universidade Federal Rural de Pernambuco para resgate de NFE's da região metropolitana do Recife em arquivos XML [11]. Este conteúdo foi convertido para uma base de dados relacional através do trabalho *Análise de modelos de dados não relacionais e multidimensionais no contexto de big data* citado no capítulo 3 e analisada em seguida pelo trabalho *Extração de Informações Mercadológicas a partir de Notas Fiscais Eletrônicas*.

Tal conjunto de dados em uma base relacional contou inicialmente com um espaço de 1,9 Gigabytes composto por tabelas e registros correspondentes às notas fiscais, os seus emitentes, os destinatários, produtos relacionados, impostos, transporte, entre outras informações. Apesar da importância destes valores para os registros a nível de controle do governo federal em seu sistema origem de notas fiscais e de consulta pelos consumidores e entidades de envio, alguns dados pertencentes a esta base não têm utilidade para o estudo proposto neste trabalho, consumindo um tamanho exagerado de armazenamento e aumentando significativamente o tempo de processamento das consultas pela aplicação. A necessidade de realizar uma limpeza nesta BD iniciou a etapa seguinte de implementação do projeto.

4.3 Pré-Processamento

Com a grande quantidade de dados existente no banco de dados inicialmente estruturado no trabalho de LIRA [11], foi identificada a necessidade da realização de uma formatação para facilitar o desenvolvimento da aplicação e aprimorar sua performance. O seguinte processo de pré-processamento e remoção de dados foi definido:

- Os atributos de cada tabela existente na base foram analisados a fim de remover aqueles que já estavam sem utilidade para a aplicação. Exemplo: Na tabela Destinatario, os campos 'numero', 'logradouro', 'complemento', 'insc_suframa' (inscrição da empresa na Superintendência da Zona Franca de Manaus), 'ie', 'indicador_ie' (indicativo de contribuinte, não contribuinte ou isento), 'im' (Inscrição Municipal), 'data_insercao', 'fone' e 'email' não seriam necessários para qualquer processo da aplicação SIG a ser implementada. Além deles, o campo 'uf' existente poderia ser removido tendo em vista que a partir da tabela municipio é possível retornar o valor de UF correspondente a linha do destinatário.
- Não apenas atributos, mas tabelas inteiras foram identificadas como desnecessárias para a aplicação. Foram removidas a 'item_imposto_nf', 'imposto' e 'totais'.
- A tabela 'volume' foi identificada com registros onde todos os valores estavam vazios. Após uma consulta por registros da tabela com os campos 'quantidade', 'especie', 'marca_volume', 'numeracao', 'peso_liquido' e 'peso_bruto', foram retornados e removidos 261 mil registros vazios na tabela citada.
- Foi identificado que o registro da tabela 'transporte' com identificador 1 não possuía nenhum conteúdo relacionado, tornando os registros de volume associados a esse transporte inválidos e desnecessários para a aplicação. A partir desta análise, foi realizada a remoção dos itens de 'volume' com valor de transporte inválido com um consulta por registros associados ao identificador de transporte '1'. Esta consulta retornou uma remoção de 254 mil registros vazios na tabela citada.

Ao categorizar as informações inválidas e valores desnecessários, foi possível reduzir em quase 70% o tamanho em disco da base de dados inicial, com uma redução de três tabelas de dados completas e mais de 500 mil registros. Esta redução

possibilitou um significativo ganho na performance de consulta aos dados e, consequentemente, nas funcionalidades da aplicação. A Figura 11 exibe o diagrama com as tabelas e campos resultantes da limpeza dos dados na base:

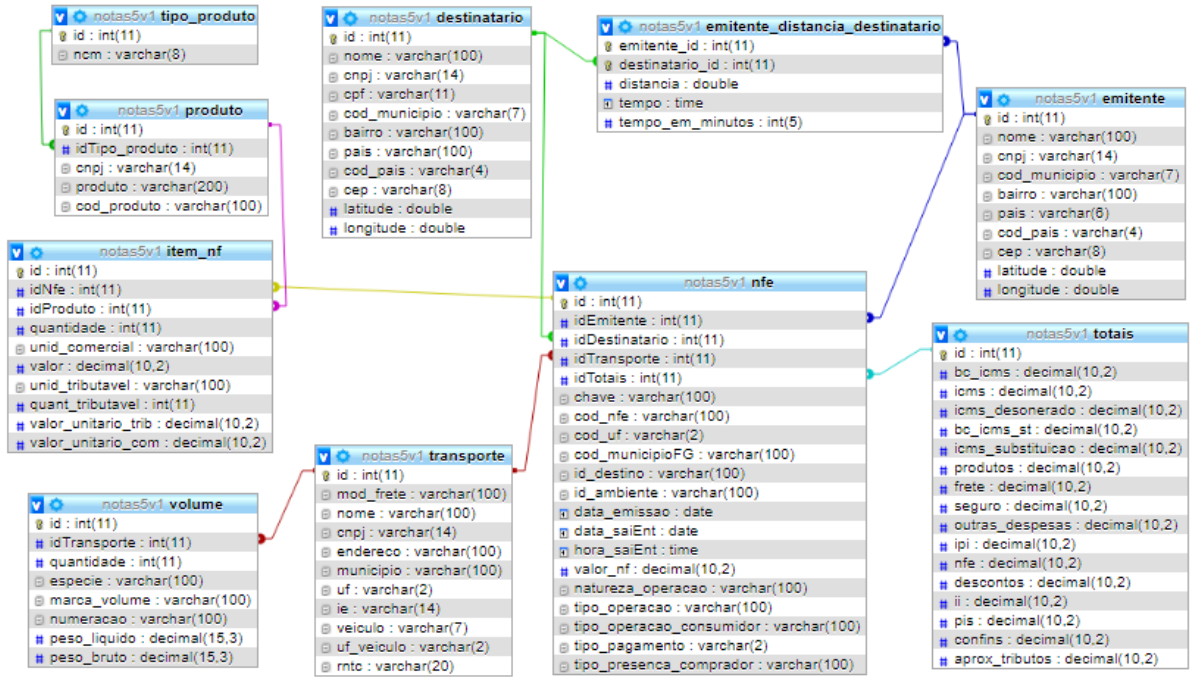

Figura 11: Diagrama de tabelas da base de dados de notas fiscais eletrônicas após o processo de limpeza. Fonte: Autoria própria.

4.4 Definição de Plataforma de Desenvolvimento

O projeto de implementação de Sistema de Informação Geográfico em processo de compra e venda teve, inicialmente, como um de seus principais objetivos, o de tornar dinâmica a análise dos dados existentes. Isso significa que além da base de dados recuperada e limpa na etapa anterior, deve ser prevista a inclusão de novos dados atualizados que possam gerar diferentes resultados de análise.

Além disso, mesmo após a redução considerável no tamanho da base de dados com sua limpeza e formatação, uma consulta completa executada nas análises desejadas necessitaria de um longo tempo para execução. Com isso, a escolha da plataforma de desenvolvimento utilizada para o processo de implementação do projeto precisaria contemplar o tamanho da base, a necessidade de executar grandes consultas de maneira ágil e de facilitar o processo de inclusão e resgate de novos dados nesta mesma base de trabalho.

Foram consideradas plataformas SIG especializadas como possíveis ferramentas de desenvolvimento da aplicação. Apesar de trazerem interessantes

funcionalidades no âmbito de análise geográfica, as plataformas analisadas não são *open source* ou não tornavam possível o desenvolvimento dos requisitos desejados no trabalho, o que acabou por retornar as buscas por uma plataforma de programação não apenas geográfica. A plataforma low coding Outsystems [39] também foi cogitada para o desenvolvimento das funcionalidades do trabalho, contando com extensões nativas geográficas e fácil manipulação de dados e mapas na plataforma. No entanto, grandes aplicações na plataforma também requerem pagamento de licença o que nos levou a uma linguagem de programação mais habitual: PHP com o gerenciamento da base de dados através da ferramenta phpMyAdmin.

4.5 Arquitetura do Projeto

Após o resgate de dados e a limpeza destes para utilização no projeto final, foi possível iniciar o processo de análise e definição da arquitetura e componentes necessários para o desenvolvimento da aplicação GIS.

Para o tema da aplicação, foram pesquisadas opções que pudessem possibilitar a exibição dos diversos dados e estatísticas de uma forma mais clara para o usuário, auxiliando uma absorção rápida dos valores e suas representações. A maior agilidade na apresentação dos dados e navegação se tornam fundamentais ao considerar a quantidade de dados que o projeto precisa executar. Tendo em vista a adoção da linguagem PHP, o mais apropriado para o layout das páginas existentes e para o desenvolvimento da aplicação seria a implementação de código a partir de um Framework PHP robusto.

Com uma grande quantidade de opções de Framework disponíveis para implementação na linguagem PHP, tornou-se necessário identificar quais os aspectos do projeto que deveriam ser priorizados para esta decisão. O sistema de informações geográficas proposto deveria priorizar a agilidade do desenvolvimento sem comprometer a qualidade e desempenho do código, com um padrão que consiga abstrair processos genéricos da implementação que possam levar tempos consideráveis para desenvolver e bibliotecas que permitam otimizar o tempo de execução de consultas e carregamento das funções propostas.

O CodeIgniter [38] apresentou características vantajosas e de maior importância para o contexto do projeto a ser desenvolvido. O framework disponibiliza uma arquitetura MVC, ferramentas de segurança integradas e uma rica

documentação para guiar durante sua instalação. O produto também oferece desempenho sólido, tornando-se uma boa opção para desenvolver aplicativos executados em servidores modestos. Por fim, o CodeIgniter possui licença pelo MIT, gratuita para utilização em qualquer aplicativo.

Com a implementação iniciada a partir da instalação do framework, o projeto seguiu a seguinte divisão de código e responsabilidades:

- Diretório Model – Com a responsabilidade de representar o negócio e realizar acessos e manipulações dos dados da aplicação.
- Diretório View - Com a responsabilidade de persistir e organizar as apresentações visuais da aplicação, exibindo as informações e funcionalidades ao usuário através de conteúdo hipertexto executável por navegadores web.
- Diretório Controller – Com a responsabilidade de ser a camada de controle da aplicação e realizar a ligação entre as duas camadas anteriores.
- Diretório Core – Com a responsabilidade de gerenciar as funções desenvolvidas para os processos de formatação e exibição de mapas, cálculo de distância entre localizações geográficas, funções de agrupamento, entre outros processos desenvolvidos pela aplicação e executados pela camada de View.

5 Desenvolvimento

Este capítulo tem por objetivo descrever os processos de implementação do WebGIS com foco em questões de programação e código aplicado.

5.1 Desenvolvimento da aplicação em PHP

A aplicação Web teve seu desenvolvimento iniciado após a conclusão dos testes de ferramentas e análises de viabilidade descritos no tópico 4.5. As consultas e mapas inicialmente desenvolvidos como testes das ferramentas no contexto do projeto trouxeram as exibições macro dos emitentes e destinatários existentes na base de dados de notas fiscais através da visão em mapa quantitativo, kernel e por marcadores. A Figura 12 exibe uma demonstração dos registros de compradores no sistema pela visão do mapa brasileiro via HighMaps. Enquanto isso, a Figura 13 demonstra a consulta destes dados com visão de mapas do Google API por marcadores e de calor.

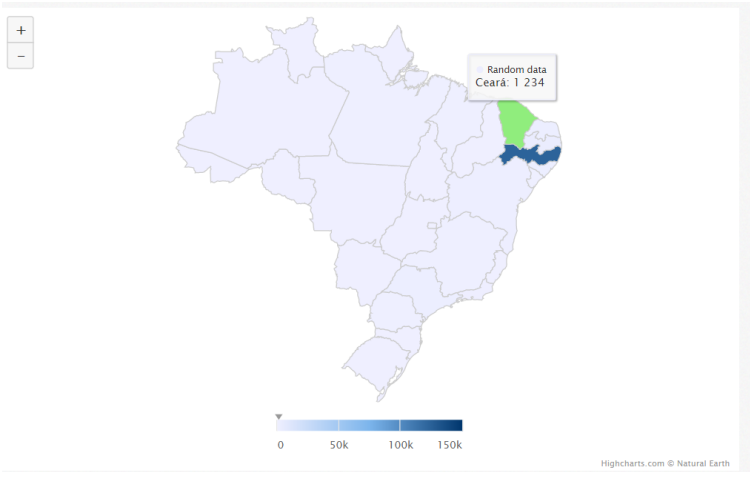

Figura 12: Demonstração do mapa Brasileiro com quantitativo de registros na aplicação.

Fonte: Autoria própria.

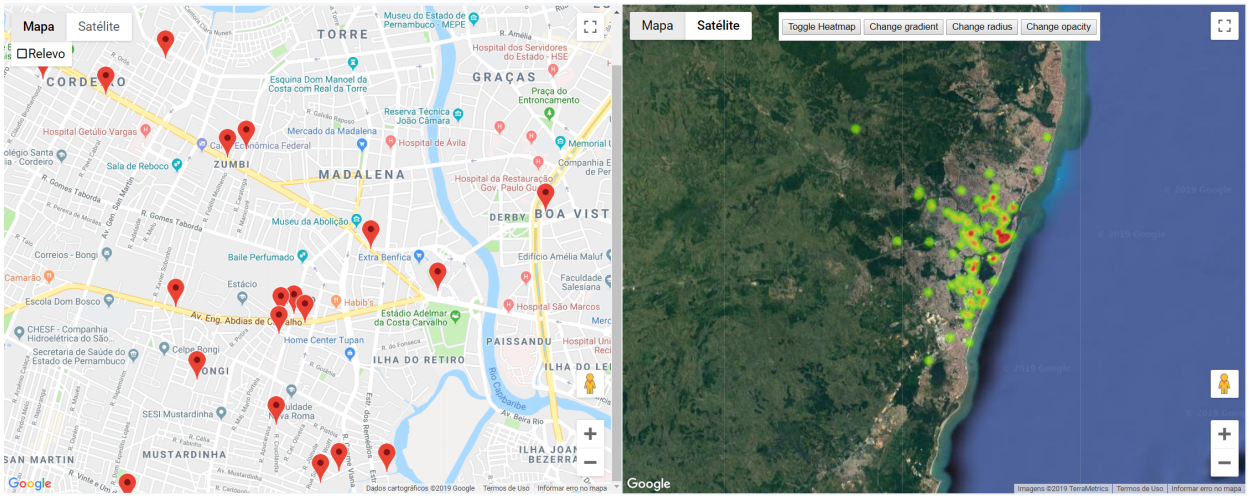

Figura 13: Demonstração de consulta de mapa
com marcadores de registros geográficos e mapa de kernel na aplicação. Fonte: Autoria própria.

A partir dos requisitos definidos no tópico 4.1, há a necessidade de gerar relatórios de vendas e clientes potenciais de emitentes específicos através do CNPJ ou outro critério informado. Os relatórios retornados pela aplicação devem ter a capacidade de exibir informações úteis a este usuário/empresa e, a partir delas, indicar alterações de contexto e simulações que auxiliem em uma tomada de decisão. Uma das propostas implementadas é a de trazer, neste relatório de vendas, pontos de localização geográfica que indiquem quais os locais mais centralizados em relação aos compradores existentes em histórico de um emitente ou de um tipo de produto específico.

A aplicação do projeto, durante seu desenvolvimento, foi batizada de GDash, ou Geographic Dashboard, considerando que os resultados obtidos estão em um modelo de métricas e análises via gráficos e mapas, semelhante a um dashboard. A página inicial exibe uma breve descrição sobre o projeto e direciona o usuário para a página de acesso aos relatórios de informação, conforme as demonstrações nas Figuras 14 e 15. A página ‘Relatórios’, por sua vez, possibilita o direcionamento para as exibições gerais dos emitentes e destinatários existentes em base e para a exibição de relatório de análise especializada para emitentes.

O sistema analisa, através da base de dados carregada, todas as compras efetuadas pelo critério definido, tipo de produto ou emitente, e gera uma quantidade de pontos médios P que indicam quais os locais mais interessantes para se estabelecer um negócio a partir da proximidade com as regiões de maiores

compradores. Estes pontos médios são gerados através do agrupamento via método clustering k-means, descrito com maior detalhe no tópico 5.2.

Figura 14: Demonstração de página inicial da aplicação. Fonte: Autoria própria.

Figura 15: Demonstração de página ‘Relatórios’ da aplicação. Fonte: Autoria própria.

O terceiro item do menu principal nomeado “Atualização” corresponde à uma página onde o usuário pode incluir novos arquivos sql que devem ser carregados na base de dados interna do sistema e considerada no processo de análise. Os dados carregados devem, no entanto, seguir o mesmo formato e padrão existente na base de dados previamente utilizada pela aplicação e formatada neste trabalho.

5.2 Implementação Algoritmo Clustering

Com a decisão de utilizar o método de clustering de dados via K-means para realizar os cálculos de agrupamento automático necessários para o SIG, iniciou-se uma pesquisa com o objetivo de implementar o algoritmo em um projeto PHP desenvolvido neste trabalho. Tornou-se possível incluir a implementação do processo

de agrupamento na linguagem de programação escolhida e chegar a um processo eficaz de utilização do método K-means para a aplicação em desenvolvimento.

Seguindo o mesmo raciocínio descrito por ADAMS [37] e citado no tópico 2.3 deste trabalho, a implementação do processo em PHP consiste em retornar uma lista de dados com duas dimensões (latitude e longitude) e um valor K para quantidade de pontos de agrupamento dinamicamente determinados. O código é responsável por traçar, inicialmente, um número K de pontos, onde a quantidade K é determinada pela aplicação, com valores de latitude e longitude aleatórios, e calcular a distância de cada item da lista para identificar qual ponto dinâmico existente em K está mais próximo. Tal combinação resulta em um conjunto de K grupos, conforme a aproximação dos valores de localização geográfica dos elementos da lista em comparação com os pontos dinâmicos. Após o agrupamento inicial realizado, cada ponto agrupador, chamados protótipo, deve ter seu valor atualizado para a média de todos os pontos associados ao seu grupo. O processo de validação do ponto K mais próximo para cada item da lista deve então ser repetido e os itens devem devem ter seu ponto agrupador atualizado, caso o mais próximo deixe de ser o anteriormente associado. A repetição deve ocorrer até que os agrupamentos estejam estáveis, ou seja, sem alterações nos agrupadores por dois ciclos de repetição e os pontos gerados para o conjunto K estejam organizados entre si, como demonstrado na Figura 16.

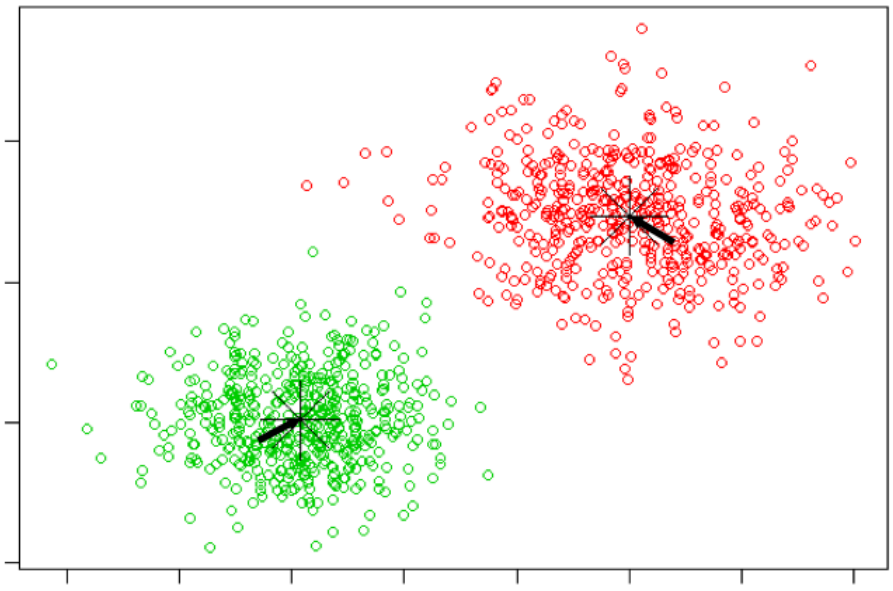

Figura 16: Demonstração visual de processo de agrupamento via K-Means. [38]

A implementação do K-means para o projeto levou em consideração implementações anteriormente existentes do algoritmo [37] para chegar a uma solução eficaz para o desenvolvido em PHP. A inclusão do processo de agrupamentos na aplicação desenvolvida seguiu a partir da implementação do código demonstrado na Figura 17 no projeto.

```php
function calculoKmeans($obj){
    //Carrega lista vazia
    $table = array();
    $centroid = array();

    //Recebe e Formata atributos do objeto $obj
    if(isset( $obj->k )){
        $k = $obj->k;
    }

    foreach($obj->set as $row){
        $table[] = new DataSet($row->x, $row->y);
    }

    //Cria lista inicial para itens médios a partir de valor K
    for($i=0; $i<$k; $i++)
        $centroid[] = new DataSet($table[$i]->x, $table[$i]->y);

    //Define limite de iterações caso haja uma definição prévia
    if(isset( $obj->limit ))  {
        $limitInteration = $obj->limit;
    } else {
        $limitInteration = 10;
    }

    //Iterações em clusters
    for ($iteration = 0; $iteration < $limitInteration; $iteration++){
        $cluster = dump($table, $centroid, $k);
        $group = array();
        for($i=0; $i<$k; $i++){
            $group[] = array();
        }
        $i = 0;
        foreach($table as $row){
            $group[ $cluster[$i] ][] = new DataSet( $row->x, $row->y );
            $i++;
        }

        // REALIZA NOVO CICLO CENTROID
        $new_centroid = dump_group($centroid, $group, $k);

        // ANALISA MUDANÇA NO CENTROID E REALIZA BREAK CASO ESTEJAM IGUAIS
        $i = 0; $flag = true;
        foreach($new_centroid as $g){
            if( $centroid[$i]->x != $new_centroid[$i]->x ||
                $centroid[$i]->y != $new_centroid[$i]->y) {
                $flag = false; break;
            }
            $i++;
        }
        if($flag) {
            break;
        }

        // COPIA NOVO CENTROID
        $i = 0;
        foreach ($new_centroid as $g) {
            $centroid[$i] = new DataSet( $g->x, $g->y );
            $i++;
        }
}
```

Figura 17: Função PHP “calculoKmeans” aplicada no projeto para processo de agrupamento.
Fonte: Autoria própria.

Utilizar o algoritmo no projeto trouxe uma forma mais eficaz de agrupar os itens de notas fiscais existentes e gerar resultados de pontos médios geográficos significativos a partir dos critérios que foram determinados.

5.3 Consulta à Base de dados

Para a obtenção dos dados necessários para análise de compras realizadas e elaboração de relatório dinâmico no sistema, foram implementadas as seguintes consultas SQL na base de dados trabalhada:

- Listar localização geográfica de todos os destinatários de notas fiscais com compras realizadas a um emitente, para exibição de mapas com marcadores que representam os destinatários existentes e para gerar mapas de calor com visão macro das regiões com mais compradores ativos;

- Retornar os quantitativos de vendas e faturamentos agrupados por mês e ano. Estes dados são importantes para montar gráficos de desempenho de um setor de mercado ou CNPJ específico que considerem as mudanças na quantidade de vendas de acordo com a sazonalidade;
- Retornar quantitativo de vendas dos 10 bairros que mais compraram itens de acordo com o emitente informado, traçando um possível padrão de regiões com compradores ativos no mercado;
- Número de vendas por mês, ano e emitente;

O relatórios de vendas de um CNPJ específico tem como campo de estudo todas as transações de clientes já realizadas pela empresa. Estas transações estão registradas na tabela de notas fiscais existente na base de dados da aplicação e com tais dados é possível a realização do cálculo médio de vendas a cada mês registrado na base de dados, a média de distanciamento entre a empresa e seus destinatários de mercadoria, além de um gráfico de vendas por região do país, conforme demonstrado na Figura 18.

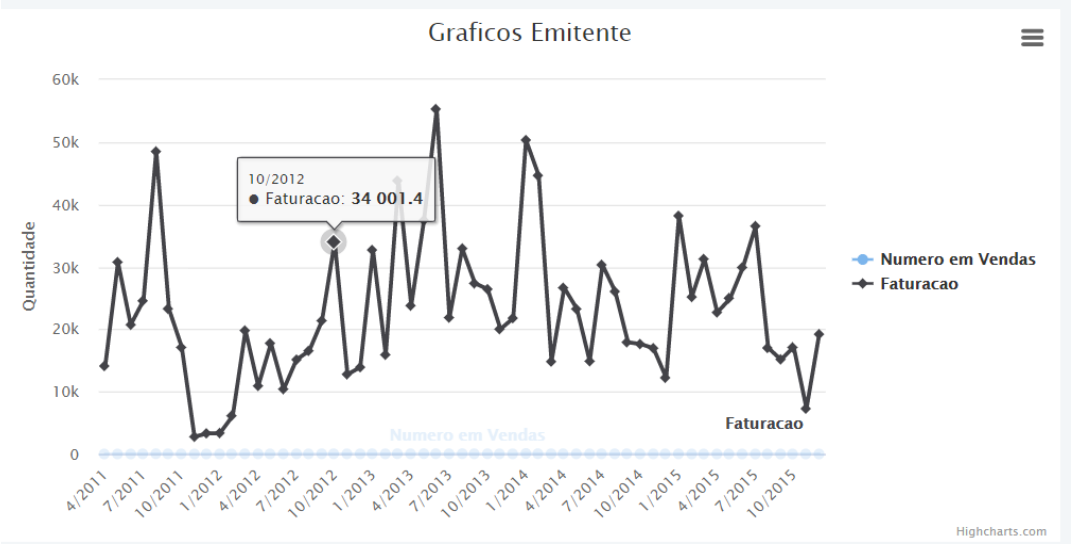

Figura 18: Gráfico gerado pela aplicação através da consulta de Vendas e Faturamento de um dos emitentes da base de dados. Fonte: Autoria própria.

O relatório também exibe gráficos e mapas indicativos relacionados ao volume das vendas realizadas e as localidades com maior número de consumidores ativos.

5.4 Implementação para Cálculo de Distância

As informações que devem ser contempladas nos relatórios de emitente estão diretamente associadas à identificação de compradores e regiões com maior densidade de consumidores ativos em um mercado. Apesar da visão global geográfica implementada, apresentando uma rápida dimensão e identificação dos pontos de compra, torna-se necessário também realizar levantamentos aproximados de distância média existente entre o emitente e seus compradores. Tal informação cria uma noção mais assertiva sobre os impactos que uma mudança de localização de um estabelecimento pode acarretar em seus consumidores de modo geral.

A implementação deste cálculo pôde ser implementada de duas maneiras: através da API Google, com solicitações aos serviços google de modo online para retornar a distância entre dois pontos geográficos, e através do cálculo em PHP puro, levando em consideração apenas o estudo das dimensões geográficas. A segunda forma de implementação é indicada na Figura 19.

```
// Diferença entre coordenadas
$raioTerrestre = 6371000;

//Converte de graus para raio
$latitudeOrigem = deg2rad($latitudeFrom);
$longitudeOrigem = deg2rad($longitudeFrom);
$latitudeDestino = deg2rad($latitudeTo);
$longitudeDestino = deg2rad($longitudeTo);

//Realiza diferença de valores em raio
$latitudeDelta = $latitudeDestino - $latitudeOrigem;
$longitudeDelta = $longitudeDestino - $longitudeOrigem;

//Cálculo de angulo
$angulo = 2 * asin(sqrt(pow(sin($latitudeDelta / 2), 2) +
  cos($latitudeOrigem) * cos($latitudeDestino) * pow(sin($longitudeDelta / 2), 2)));

//Resultado
$resultado = $angulo * $raioTerrestre;
```

Figura 19: Demonstração de cálculo de distância geográfica aplicada em PHP para o projeto.
Fonte: Autoria própria.

Apesar de haver pequenas divergências entre os resultados das duas implementações, a segunda foi a mais plausível para o contexto, levando em consideração que a realização de solicitações web aos serviços Google para cada linha de nota fiscal retornada do processo, somado ao já pesado carregamento dos

dados de notas fiscais na formação de relatório, acarreta em uma longa execução de código, levando à prováveis timeouts na aplicação.

5.5 Implementação para Cálculo de Frete

Uma outra funcionalidade desenvolvida na aplicação foi o cálculo médio de custo de envio e transporte dos produtos. Esta implementação teve como objetivo ilustrar um novo contexto que possa indicar futuros ganhos na escolha de uma localização geográfica de um emitente e consiste em calcular, a partir dos dados de compras registrados para aquela categoria de produto, ou distribuidor específico, o custo de envio de todos os produtos no contexto existente para, em seguida, simular qual seria o valor pago de transporte pelas mesmas compras caso o emitente estivesse em uma localização diferente.

De acordo com especialistas e empresas especializadas em e-commerce como a BERTHOLDO [33], o preço do frete pode ser um fator significativo para o aumento ou queda de vendas de um comércio eletrônico. AUGUSTO [34] complementa ao exemplificar que o consumidor sente muito mais ter que pagar R$ 15,00 de frete, do que R$ 15,00 a mais em um produto, mas com frete gratuito. Muitos clientes desistem da compra ao ver o valor do frete no momento do checkout. Torna-se importante para o gestor de uma loja virtual (ou mesmo física) estudar opções que contornem esta problemática, seja com táticas de aumento do preço do produto, frete grátis para compras acima de uma determinada quantia ou negociações com empresas de transporte.

A funcionalidade de simulação de resultados foi idealizada a partir deste problema, objetivando a possibilidade de retornar ao usuário da aplicação um comparativo entre a soma dos custos de entrega de mercadorias do emitente baseado na sua localização atual e a soma dos custos de entrega baseados em uma nova localização definida para a simulação. Com isso, é possível demonstrar ao usuário/gestor de uma loja virtual as possíveis reduções nos custos de transporte de produtos a partir de uma mudança física da sua distribuidora. Tal mudança pode ser estudada como facilitador para o aumento de vendas.

Os custos de entrega são demonstrados a partir de estimativas de frete disponibilizados pelos Correios [35]. Através de uma API oficial que retorna um documento XML com informações de frete e tempo de entrega de um produto previstos pelo sistema dos Correios, a funcionalidade de simulação de custos informa

os valores de CEP de origem, CEP de destino, dimensões, peso, valor do produto e tipo de entrega para realizar o cálculo de frete de uma compra através das informações obtidas nas notas fiscais da base de dados. Como resultado, a aplicação deveria retornar a soma de todos os custos previstos para o mesmo tipo de entrega com a origem inicial do emitente e o seu novo endereço simulado.

A função de cálculo de frete prevê os seguintes valores de código de serviço:

- 41106 PAC sem contrato
- 40010 SEDEX sem contrato
- 40045 SEDEX a Cobrar, sem contrato
- 40215 SEDEX 10, sem contrato

Para os cálculos automáticos da aplicação, é considerado apenas o serviço "41106 PAC sem contrato". Com os valores de frete existentes na base de dados original de notas fiscais, é provável que existam diferenças significativas entre os custos de envio de mercadorias via endereço original do emitente e os custos simulados de um novo endereço deste. Tais valores são demonstrados como referência e estimativa a nível proporcional, não sendo possível obter um resultado monetário assertivo.

5.6 Considerações Finais

Apesar do escopo inicial do projeto ter sua implementação bem sucedida de um modo geral, alguns pontos impactaram os resultados e objetivos almejados. A base de dados, por conter um grande volume de registros, tornou o processo de consulta e levantamento de relatórios bastante demorado em alguns cenários traçados, executando *timeouts* e a necessidade de uma reconstrução de código baseado em um modelo de implementação via múltiplas threads. Tal implementação foi cogitada porém não bem sucedida durante o tempo de realização do projeto, que acabou mantendo a estrutura de execução síncrona inicialmente planejada, o que, em alguns casos, limitou as consultas a cenários mais simples do negócio por depender do desempenho da aplicação.

Outro parcela impactante da implementação do projeto ocorreu durante a implantação de cálculo pelos correios. Primeiramente, o cálculo implementado é realizado através de uma API previamente existente dos correios brasileiros, que requer informações, além dos ceps origem e destino, sobre atributos do produto a ser enviado, como o volume do material e peso. Foi considerado que a identificação

destas informações necessárias para o produto estariam disponíveis através da tabela “Volume”, existente na base de dados. No entanto, no decorrer do desenvolvimento foi feita a identificação tardia da não existência de um relacionamento da tabela com as demais na base pré-carregada. Sem chaves estrangeiras para referenciar e associar registros de produtos e compras aos seus respectivos volumes, o cálculo de frete de envio se tornou bastante genérico, sem considerar as reais dimensões e pesos dos produtos em contexto na realidade estudada. Uma outra limitação em relação à implementação do processo foi o tempo de execução da API dos correios, visto que em diversos testes apresentou lentidão, impedindo o término do cálculo, que realiza a soma do frete de todas as compras realizadas pelo emitente.

Por fim, pretendia-se disponibilizar em servidores online o WebGIS GDash, a nível de demonstração da sua portabilidade em execução. No entanto, mesmo tendo sido implementado em PHP, com a utilização de API’s e uma base de dados flexível, os servidores encontrados tiveram restrições quanto à execução das queries existentes e o carregamento da base de dados, devido à grande quantidade de registros existentes.

6 Resultados e discussões

Através da aplicação GDash desenvolvida no presente trabalho, foi possível implementar um sistema de análise e demonstrações geográficas automáticas com informações existentes em uma extensa base de dados de notas fiscais eletrônicas. Por possibilitar um estudo dinâmico dos dados, os resultados gerados neste tópico do trabalho se resumem apenas ao conteúdo existente no banco originalmente resgatado e trabalhado.

A seguir, tem-se alguns exemplos dos resultados conquistados a partir da base de dados estudada. As informações que identificam as empresas dos resultados da aplicação, como o CNPJ, nome e endereço, foram omitidas das imagens e textos de demonstração, com o objetivo de preservar a privacidade dos dados destas.

Os relatórios para emitentes específicos são realizados através do menu topo “Relatórios” > “Relatório para emitentes”. Na página seguinte, o usuário deve informar o CNPJ associado à empresa na qual deseja visualizar o relatório de vendas pela aplicação, conforme exibido na Figura 21. É necessário que este CNPJ exista na base de dados, assim como dados de notas fiscais associadas à ele. Devem ser exibidas as opções “Resgatar Relatório” e “Simular Resultados” abaixo do campo de preenchimento. Enquanto o item “Resgatar Relatório” corresponde ao levantamento de informações já existentes em base para o emitente indicado, o botão “Simular Resultados” direciona para o conteúdo de cálculo de pontos médios via K-means e simulação de novas localizações, distâncias e custos de envio para o emitente.

Figura 21: Demonstração de preenchimento de CNPJ para retornar relatório de vendas.

Fonte: Autoria própria.

Exemplo 1: Relatório por emitente de CNPJ 04********1:**

O primeiro exemplo para demonstração de resultados considera o CNPJ de número 04**********1 para exibir o relatório de informações levantadas pelo WebGIS GDash na opção de resgate de relatório. A aplicação implementada retornou as informações sobre o emitente conforme a exibição da Figura 22:

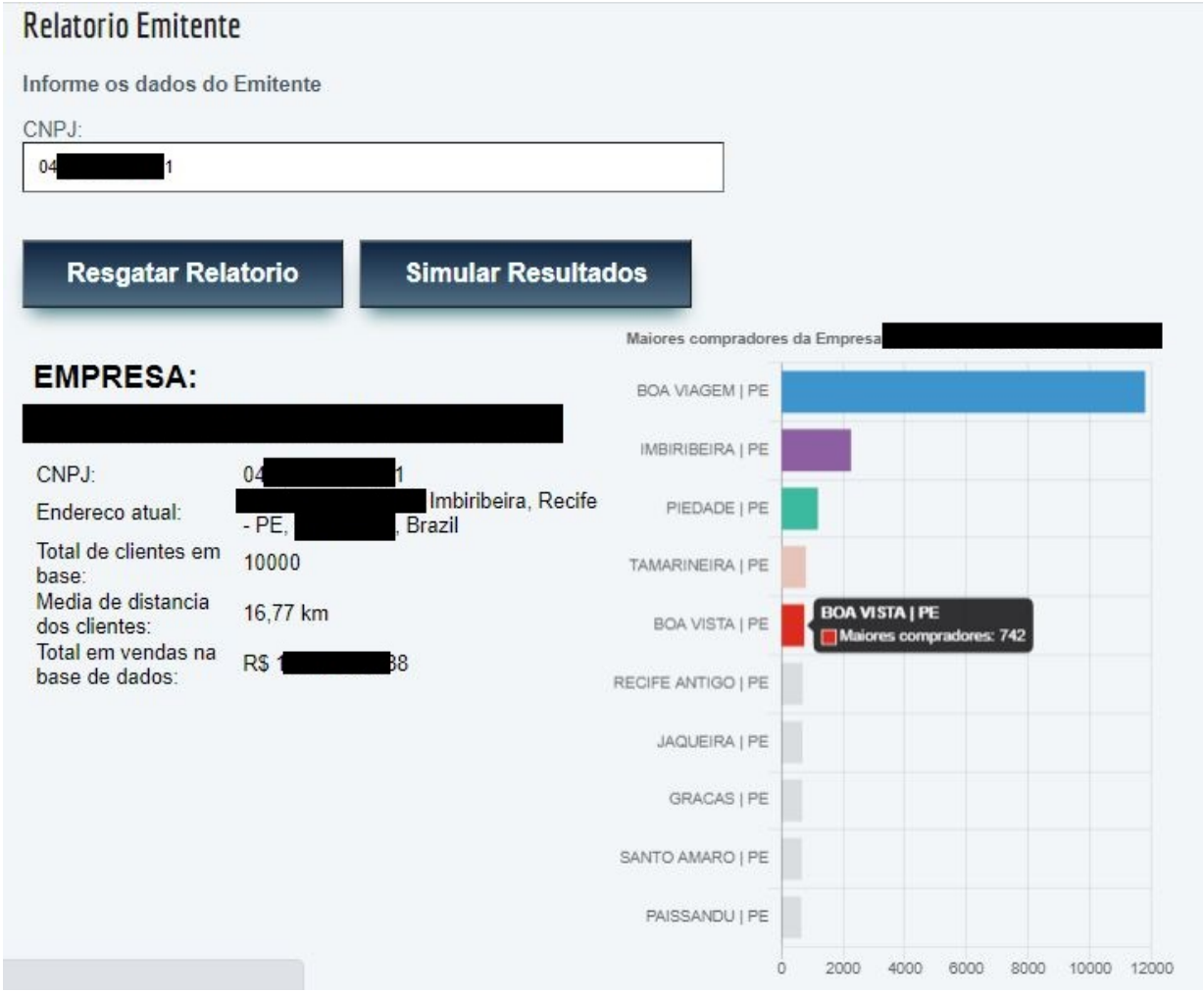

Figura 22: Relatório inicial gerado pela aplicação através da consulta pelo CNPJ de um dos emitentes da base de dados. Fonte: Autoria própria.

O primeiro bloco de conteúdo exibe o número total de clientes da empresa, a partir dos registros de compras efetuados, e a soma de todas as compras registradas em base para o CNPJ. Além disso, é devolvido o cálculo referente à média de distância da empresa para os seus compradores, entre as informações do bloco esquerdo. Considerando que compradores de outros estados acabam por afetar

consideravelmente uma média de distância geral, o resultado trazido para o CNPJ é bastante razoável, com 16,77 kms de distância média.

O segundo bloco de conteúdo da aplicação realiza uma demonstração em mapa de calor dos compradores ativos da empresa de acordo com os dados existentes na base carregada. Através da coloração esverdeada e avermelhada, é possível identificar os locais com maior número de compradores no mapa da região metropolitana do Recife (e de outras regiões, conforme navegação pelo componente). As regiões marcadas em tons vermelhos demonstram maior concentração. Abaixo, são exibidos dois gráficos evolutivos em linha do tempo, representando o faturamento e o número de vendas mensais conforme as datas de compras resgatadas da base de dados. O gráfico deve demonstrar os cinco tipos de produto mais vendidos pela empresa através dos registros existentes e traçar a mudança das vendas à nível temporal para cada um dos cinco produtos. Através dessa visão, é possível traçar a evolução de compra ou identificar um padrão de aumento ou redução de faturamento conforme o mês e ano. Estes resultados podem ser demonstrados nas Figuras 23 e 24.

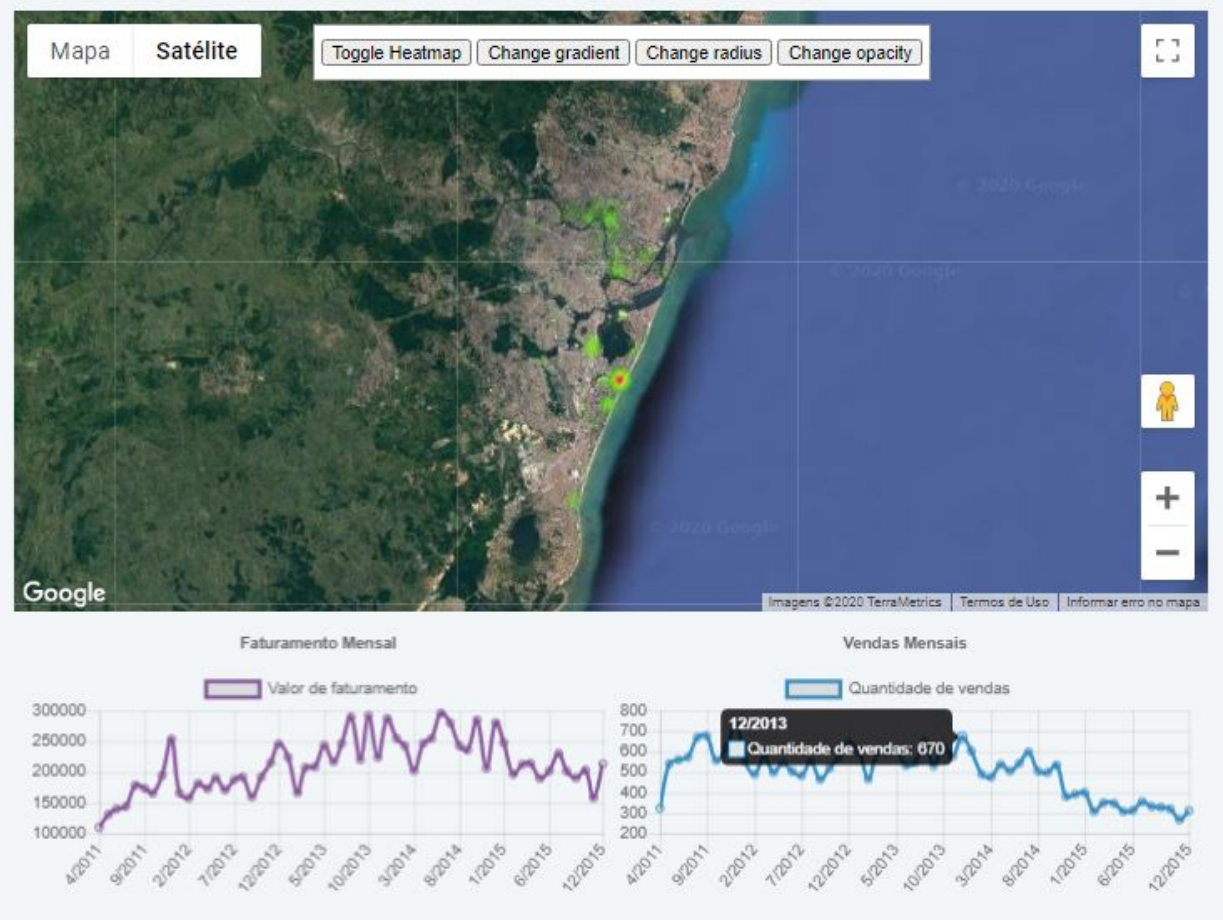

Figura 23: Mapa de calor, representando a intensidade de vendas de CNPJ por região no grande Recife. Abaixo, os gráficos de faturamento e vendas mensais da empresa. Fonte: Autoria própria

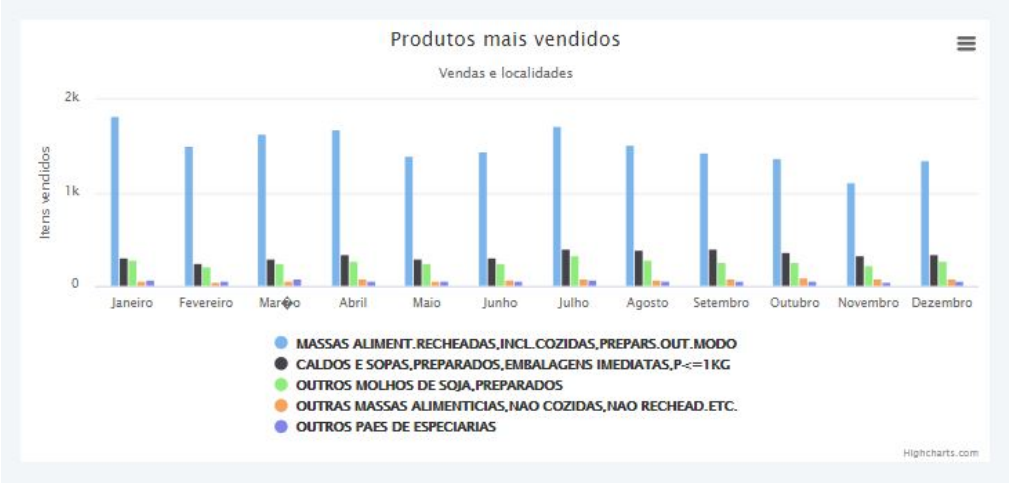

Figura 24: Levantamento realizado pela aplicação GDash de produtos mais vendidos por emitente indicado. Fonte: Autoria própria.

Exemplo 2: Relatório por emitente de CNPJ 03********2:**

Este segundo exemplo traz uma nova demonstração de relatório de emitente a partir de consultas pela base de dados existente. Esta empresa indicada tem seu endereço na cidade de Jaboatão dos Guararapes, em Pernambuco, e conta com 1.395 consumidores identificados com as notas fiscais eletrônicas. O resultado é demonstrado na Figura 25, a seguir:

Figura 25: Relatório inicial gerado pela aplicação através da consulta pelo CNPJ de um dos emitentes da base de dados. Fonte: Autoria própria.

Apesar do gráfico de maiores compradores da empresa indicar que o Distrito Industrial de Pernambuco, é possível identificar uma grande quantidade de consumidores em estados como o Ceará, Paraíba e Rio Grande do Norte. O emitente acaba por ter uma média de distância geral dos clientes consideravelmente maior que o exemplo anterior, com 405 quilômetros de distância média aos seus compradores.

A seguir, pode ser demonstrado o mapa de calor correspondente às vendas da empresa que tem uma concentração além da região metropolitana do Recife. Da mesma forma que para o primeiro exemplo, devem ser demonstrados os gráficos de faturamento e de produtos mais vendidos. Estas exibições podem ser demonstradas nas Figuras 26 e 27.

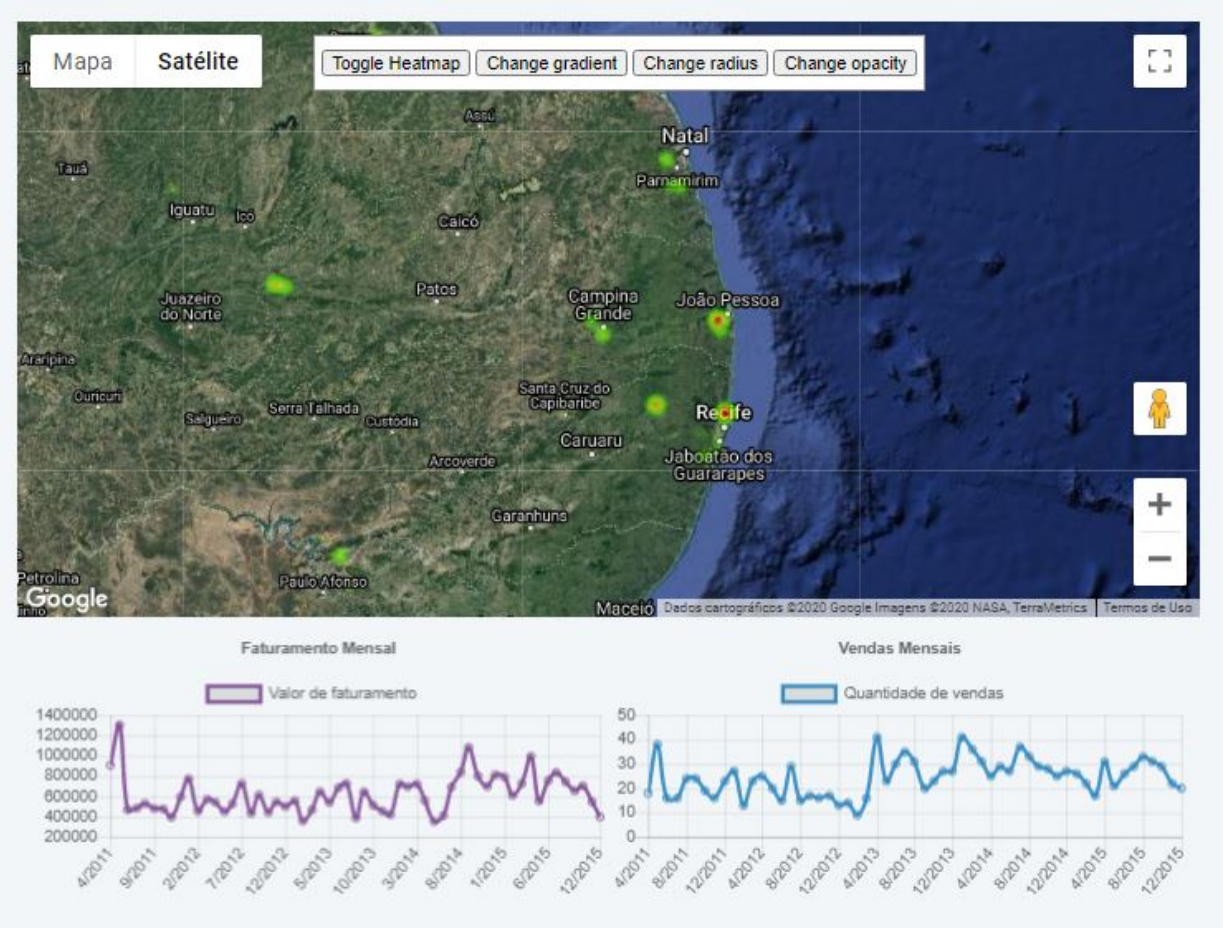

Figura 26: Mapa de calor, com marcadores, representando a intensidade de vendas de CNPJ por região. Abaixo, os gráficos de faturamento e vendas mensais da empresa. Fonte: Autoria própria.

Figura 27: Levantamento realizado pela aplicação GDash de produtos mais vendidos por emitente indicado. Fonte: Autoria própria.

É possível, através da Figura 27, perceber que as vendas da empresa se concentram principalmente nos produtos de mobiliário e armações de madeira. Também é possível notar um acréscimo expressivo nas vendas durante o mês de dezembro. É importante para a empresa perceber o que motiva esse aumento no mês indicado e se é um padrão frequente para as vendas dos anos anteriores para, talvez, aprimorar a venda nos demais meses ou se manter preparado, à nível de estoque e equipe, para este padrão de aumento.

Exemplo 3: Simulação e modificação de endereço para CNPJ 05********2.**

O terceiro exemplo é utilizado para demonstrar os resultados na segunda opção dos relatórios de emitentes, de “Simulação de resultados”. O usuário que está executando o sistema deve preencher o valor de CNPJ a ser consultado, da mesma maneira que para os exemplos anteriores, e, em seguida, selecionar o botão “Simular Resultados” existente, conforme demonstrado na Figura 28.

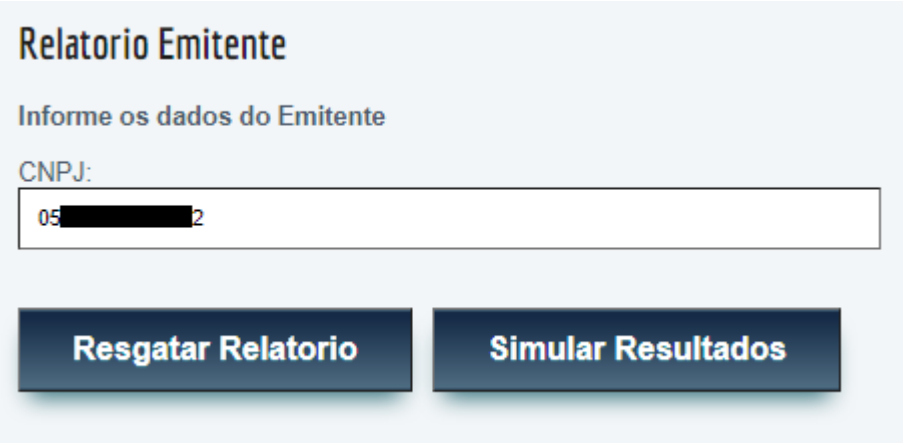

Figura 28: Demonstração da pesquisa inicial de relatórios para simular resultados em emitente. Fonte: Autoria própria.

O sistema apresenta, inicialmente, as informações básicas da empresa, semelhante ao relatório resgatado na primeira opção. No entanto, não são exibidos os gráficos de vendas e bairros. Este é substituído pelo painel de simulação/mudança de endereço da empresa. A partir desta opção, o usuário pode indicar um novo endereço da empresa (seja por rua e número ou por coordenadas gegráficas) para realizar o cálculo de distância perante os clientes registrados e os custos de envios para eles. Tal comportamento é demonstrado na Figura 29.

Figura 29: Relatório inicial de simulação gerado pela aplicação através da consulta pelo CNPJ de um dos emitentes da base de dados. Fonte: Autoria própria.

A empresa demonstrada no relatório tem como destaque sua distância média de 207 quilômetros dos seus compradores. Através do campo de simulação para nova localização, foi indicado um possível endereço para a empresa. Resultando no comportamento demonstrado na Figura 30.

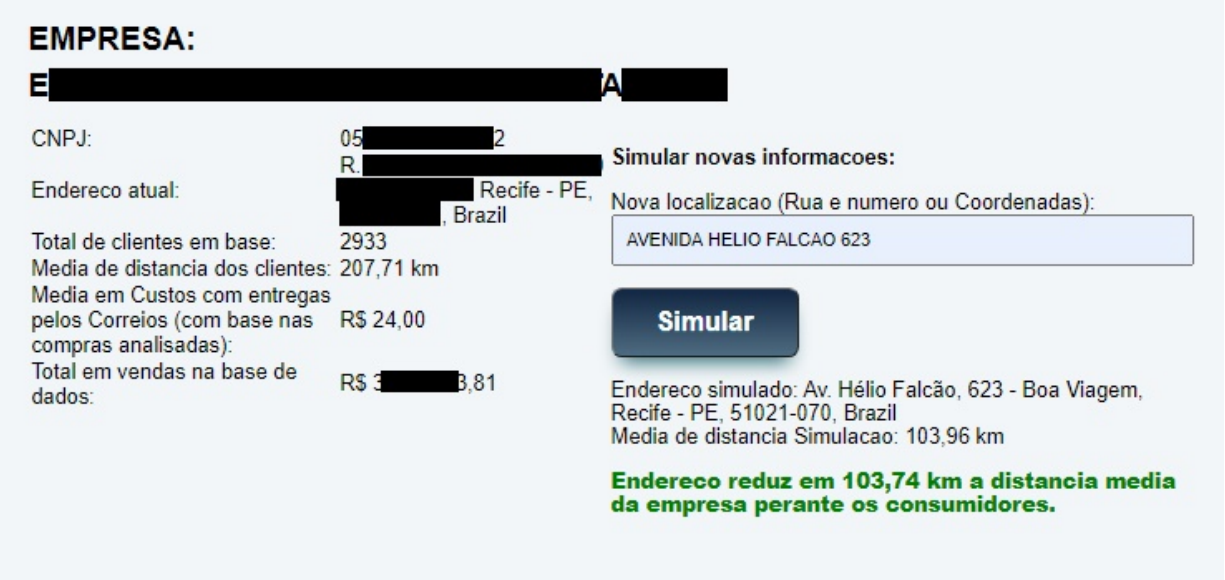

Figura 30: Relatório inicial de simulação com alteração positiva de endereço do emitente.

Fonte: Autoria própria.

O novo endereço definido resultou em um cálculo positivo de distanciamento dos clientes, indicando que, se a empresa mudar sua localização para o novo endereço descrito, a sua distância média perante os clientes sofrerá uma redução em torno dos 103 quilômetros.

Para demonstrar um resultado negativo, foi preenchido um endereço correspondente ao estado de São Paulo, conforme a Figura 31 exibe.

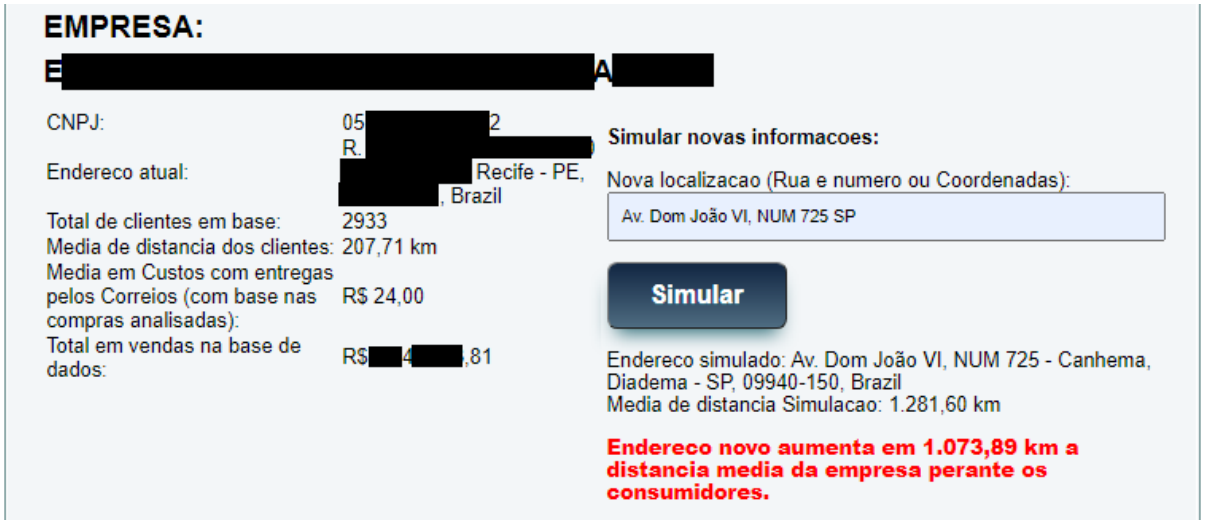

Figura 31: Relatório inicial de simulação com uma alteração negativa de endereço do emitente.

Fonte: Autoria própria.

Esta segunda simulação, por sua vez, trouxe a informação de que o endereço novo aumenta em 1073 quilômetros a distância média dos clientes registrados da empresa, tornando a opção completamente negativa para o negócio.

Exemplo 4: Pontos médios sugeridos.

Com a opção de simulação de resultados demonstrada no exemplo 3, tem-se, para o CNPJ indicado, o resultado visual do cálculo de pontos médios indicativos para a empresa, via k-means. O cálculo corresponde a um processo automático de sugestão para quais seriam os locais mais indicados para se estabelecer um ponto de comércio da empresa ou novo armazém de envio, assumindo o fator distância dos compradores favorecido. A seguir, na Figura 32, há um exemplo demonstrativo onde um dos dois pontos indicados para transferência da empresa se situa no bairro do Ibura, em Recife, Pernambuco.

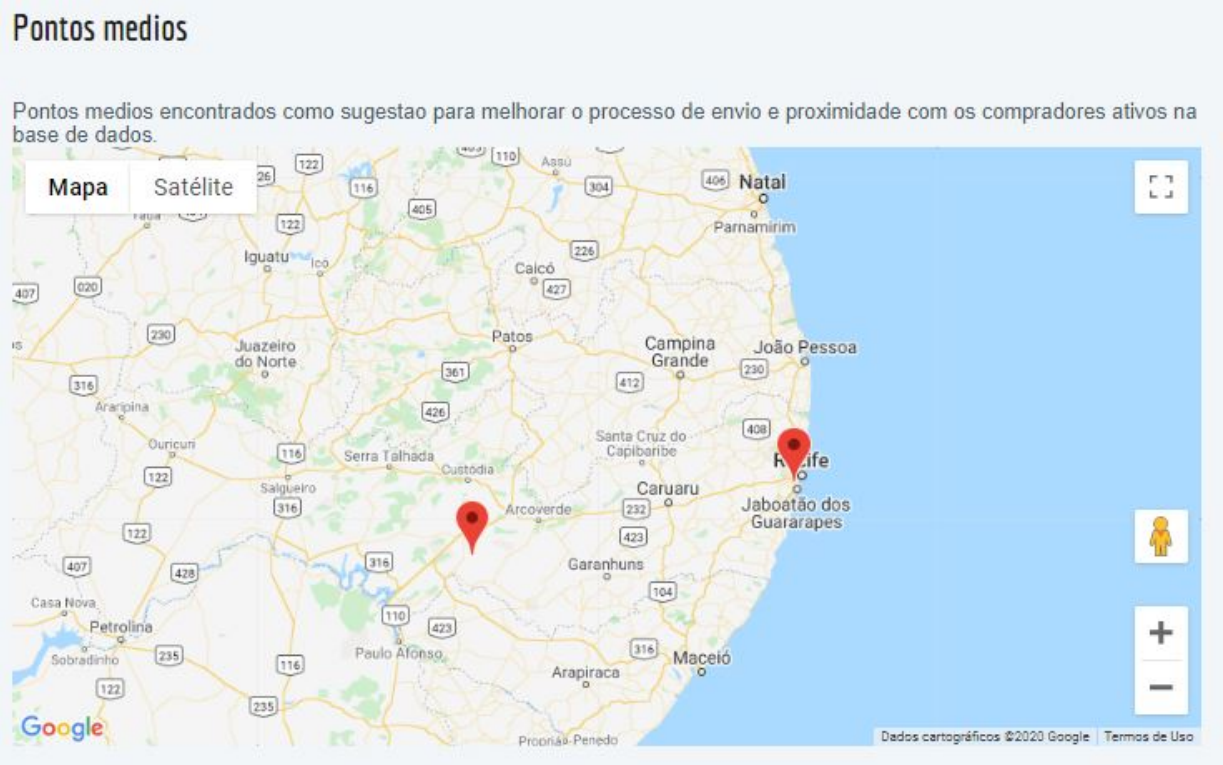

Figura 32: Exibição de relatório de simulação com o cálculo de pontos médios sugeridos.
Fonte: Autoria própria.

É importante ressaltar que o cálculo de pontos médios é feito com base em localizações geográficas puras, ou seja, coordenadas geográficas, e que os resultados exibidos podem não se encontrar em regiões habitáveis, como florestas,

estradas ou rios. Em um caso de análise real, é necessário identificar um local próximo que seja coerente para mudança da empresa e todos os demais fatores que podem afetar a decisão de uma instalação física. Para o exemplo da Figura 32, CNPJ 009**********7, foi realizada a simulação de mudança de endereço para o endereço existente em Camaragibe que foi indicado. O resultado de uma simulação para este endereço resultou numa redução de mais de 37 quilômetros de distância média da empresa perante seus principais consumidores, conforme demonstração da Figura 33.

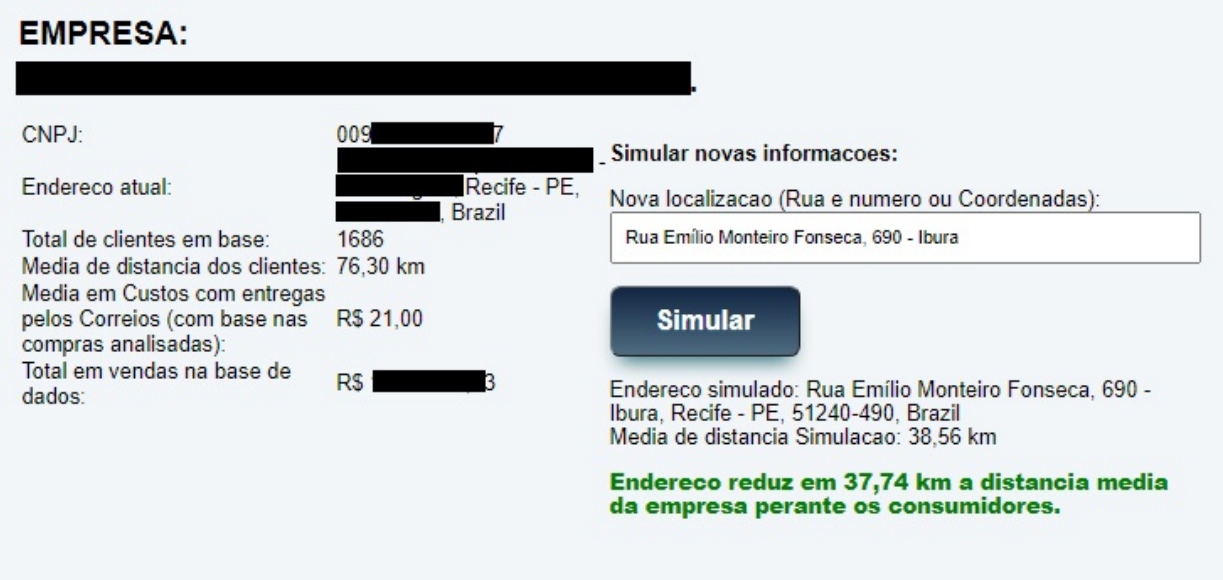

Figura 33: Simulação de novo endereço para o emitente, a partir de uma das indicações de cálculo médio via K-means. Fonte: Autoria própria.

Com a visão de pontos médios utilizando a mesma API google maps implementada para os mapas de calor e marcadores da aplicação, é possível customizar a exibição dos resultados com o modelo via mapa ou satélite e exibição opcional das nomenclaturas de ruas e regiões do mapa. Na figura 34, pode ser demonstrada uma diferente versão do resultado de pontos médios, em comparação com o da Figura 32. Os marcadores geográficos no mapa exibido devem ter uma caixa informativa com o endereço por extenso da localização indicada mais indicada para o negócio em contexto. Para os casos onde o ponto indicado não tenha um endereço definido, recomenda-se identificar os localidades significativas na proximidade.

Figura 34: Exibição de relatório de simulação com o cálculo de pontos médios sugeridos.
Fonte: Autoria própria.

Exemplo 5: Ponto médio sugerido para distribuidor de diferentes regiões.

Para este último exemplo, foi considerada uma empresa com vendas em grande quantidade ao redor do país. Através do relatório inicial demonstrado na Figura 35, foi possível identificar a existência de vendas da empresa para estados das regiões Nordeste e alguns do Sudeste.

Figura 35: Exibição de mapa de calor associado às vendas do emitente selecionado.
Fonte: Autoria própria.

Coincidentemente, o ponto médio indicado na Figura 36, através do cálculo de agrupamento de localizações por K-means para o emitente, é identificado na cidade de Caruaru, em Pernambuco.

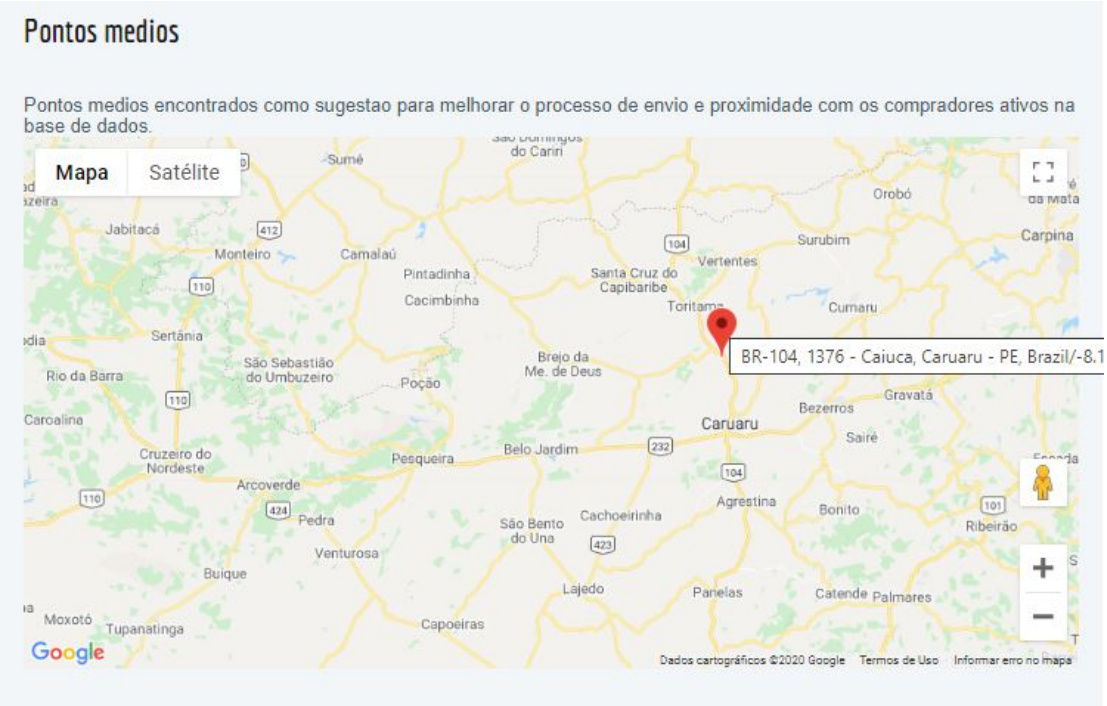

Figura 36: Indicação de ponto médio sugerido através do cálculo de agrupamento K-means.
Fonte: Autoria própria.

Apesar de não se tratar de um endereço completo, com rua e número, a simulação de distância perante os clientes da empresa para o município de Caiuca, em Caruaru, retornou uma melhoria de mais de 900 quilômetros na distância média aos compradores da distribuidora, conforme Figura 37, o que pode, unido à uma análise de contexto do negócio, representar um ganho financeiro através de uma proximidade com os destinatários dos produtos.

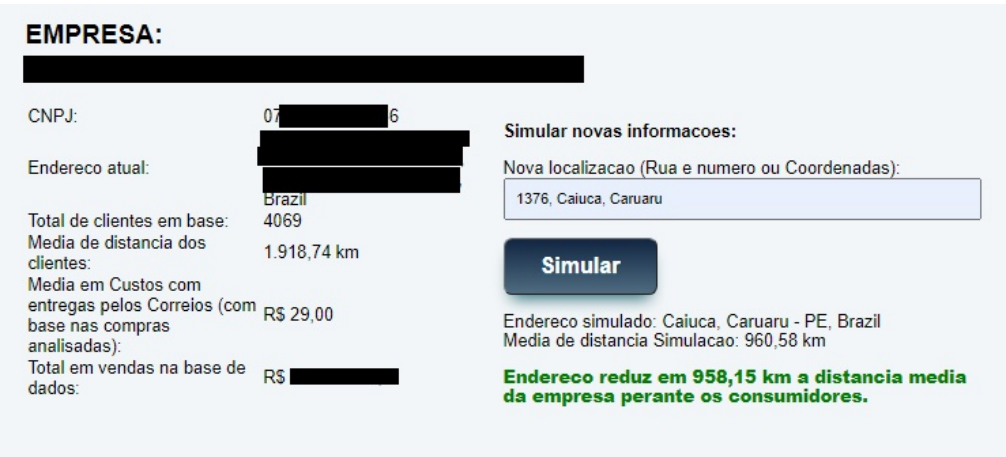

Figura 37: Simulação de novo endereço para empresa através de indicação do sistema GDash.
Fonte: Autoria própria.

7 Conclusão

O trabalho apresentado pôde demonstrar a implementação e utilização de uma plataforma dinâmica para análise de dados, aplicando sistemas de sugestão e simulação a partir de cenários reais do mercado de compra e venda da região metropolitana do Recife ou, pelo que pôde ser notado durante o desenvolvimento, em qualquer região do país.

A aplicação se fundamentou em um sistema de informação geográfica híbrido, concentrando-se em diferentes formas de exibição de mapas e gráficos demonstrativos para os relatórios formados. Através dele, tornou-se possível acrescentar novos exemplos e métodos de análise de dados, como o processo de sugestão automática via agrupamento de dados utilizando o algoritmo de aprendizado de máquina K-means, somado ao mecanismo de implementação de um sistema geográfico em uma plataforma web (WebGIS), com código fonte em PHP e integrações com variadas API's.

Também foi possível enfatizar a importância dos SIG's no estudo de comportamentos de compra, auxílio na gestão de comércio e a possibilidade de utilização das técnicas de informática e geolocalização nas mais variadas áreas de atuação. Tais aplicações podem ser ampliadas e aperfeiçoadas a fim de trazer os melhores resultados e auxílios para tomada de decisão no mercado de compra e vendas.

7.1 Trabalhos futuros

Com base nos resultados recebidos com este projeto, obteve-se algumas ideias para trabalhos futuros relacionados:

- Implementar uma aplicação de análise dinâmica de dados estatísticos para um segmento de mercado ou produto/serviço específico;
- Reutilizar a aplicação gerada no trabalho, aumentando seu escopo a nível de pontos analisados e funcionalidades previstas;
- Implementar um estudo semelhante que consiga abranger a sua funcionalidade para negócios sem uma base de histórico pré-carregada, oferecendo ferramentas de estudo, suporte e simulação de custos mesmo para empresas onde não existam registros de vendas na aplicação;

- Análises comparativas de estatísticas acerca dos preços de produtos e serviços por região do país;
- Integração com diferentes fontes de dados, como redes sociais, para análise estatística e pesquisa de mercado;
- Análises que permitam o desenvolvimento de serviços para a orientação de consumo inteligente a um consumidor final, indicando as melhores relações de custo/benefício para a aquisição de produtos e/ou serviços a partir das distribuições georreferencias de ofertas e demandas (assim como preço, disponibilidade e acesso).

Referências

[1] CETIC.BR - O Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação. Indicadores - TIC Domicílios 2019. 2020. Disponível em: <https://cetic.br/pesquisa/domicilios/indicadores/>. Acesso em: 2 jun. 2020.

[2] ABCOMM. Comércio eletrônico deve crescer 18% em 2020 e movimentar R$ 106 bilhões. 2019. Disponível em: <https://abcomm.org/noticias/comercio-eletronico-deve-crescer-18-em-2020-e-movimentar-r-106-bilhoes/>. Acesso em: 10 jul. 2020.

[3] BARBOSA FILHO, F.; A crise econômica de 2014/2017. Instituto Brasileiro de Economia. Disponível em: <https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142017000100051>. Acesso em: 15 jul. 2020.

[4] IBGE. Pesquisa mensal de comércio. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/comercio/9227-pesquisa-mensal-de-comercio.html>. Acesso em: 14 jul. 2020.

[5] MOREIRA, Paulo. Comércio eletrônico: antes e depois da pandemia do coronavírus. Disponível em: <https://www.ecommercebrasil.com.br/artigos/comercio-eletronico-antes-e-depois-dapandemia-do-coronavirus/>. Acesso em: 15 jul. 2020.

[6] World Health Organization, Coronavirus disease (COVID-19) pandemic. Disponivel em: < https://www.who.int/news/item/29-06-2020-covidtimeline>. Acesso em: 20 jun. 2020.

[7] BOWLES, E.; A Covid-19 e a transformação do comércio eletrônico no Brasil. Disponível em: <https://www.ecommercebrasil.com.br/artigos/a-covid-19-e-a-transformacao-do-comercio-eletronico-no-brasil/>. Acesso em: 9 ago. 2020.

[9] ALMEIDA, D.; CAMASMIE, A. *Carrefour fecha loja online e demite cerca de 50 pessoas.* 2012. Disponível em: <https://epocanegocios.globo.com/Informacao/Acao/

noticia/2012/12/carrefour-anuncia-suspensao-de-seu-e-commerce.html>. Acesso em: 5 mar. 2019.

[10] HIRSCH, L.; *Toys R Us stores closed on Friday, leaving behind nostalgia, anger and maybe a chance of revival.* 2018. Disponível em: <https://www.cnbc.com/2018/06/29/toys-r-us-closes-its-doors-on-friday-leaving-beind-nostalgia-anger-a.html>. Acesso em: 13 fev. 2019.

[11] LIRA, M. C. S.; *Análise de modelos de dados não relacionais e multidimensionais no contexto de big data,* 2016, Universidade Federal Rural de Pernambuco

[12] NETO, A. B. S.; *Extração de Informações Mercadológicas a partir de Notas Fiscais Eletrônicas.* Universidade Federal Rural de Pernambuco. 2018.

[13] DUARTE, R. D.; *Big Brother Fiscal - III, O Brasil na Era do Conhecimento,* 2009. p. 74.

[14] BARROS, M.; FARIA, M. C. P. G.; ALVES, P. H. B. P.; DOS SANTOS SOUZA, R. *Nota fiscal eletrônica*, 2008. p. 34, 35.

[15] FEITOSA, M. P.; *Fundamentos de Banco de Dados – Uma abordagem prático-didática*, 2013. p.14.

[16] PÉREZ-ORTEGA, J.; *The K-Means Algorithm Evolution, Introduction to Data Science and Machine Learning*, 2019. Disponível em: <https://www.intechopen.com/books/introduction-to-data-science-and-machine-learning/the-em-k-em-means-algorithm-evolution>.

[17] MURPHY, K. P.; *Machine Learning: A Probabilistic Perspective*, 2012. p. 2, 3.

[18] MONARD, M. C.; BARANAUSKAS, J. A.; Paradigmas de Aprendizado: Conceitos sobre Aprendizado de Máquina, 2003. p. 40. Universidade de São Paulo. Disponível em: <http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf>

[19] ADAMS, R. P.; K-Means Clustering and Related Algorithms. Princeton University <https://www.cs.princeton.edu/courses/archive/fall18/cos324/files/kmeans.pdf>

[20] OLIVEIRA, A. F.; Favorecendo o Desempenho do k-Means via Métodos de Inicialização de Centróides. 2018. Centro Universitário Campo Limpo Paulista. Disponível em: <www.cc.faccamp.br/Dissertacoes/AndersonFranciscoOliveira.pdf>

[21] OZEMOY, V. M.; SMITH, D. R.; SICHERMAN, A. Evaluating Computerized Geographic Information Systems Using Decision Analysis Interfaces, 11, 1981. p.92.

[22] DAVIS, B. E.; GIS: A visual approach, 2, 2001. p.13.

[23] CASANOVA, M. A.; CAMARA, G.; DAVIS JR, C. A.; VINHAS, L; Queiroz, GR. Bancos de dados geográficos. MundoGEO Curitiba, 2005. p.13.

[24] ROSENBLATT, M.; Remarks on Some Nonparametric Estimates of a Density Function, 1956.

[25] MEDEIROS, A.; Introdução aos Mapas de Kernel <http://www.clickgeo.com.br/mapas-de-kernel-parte-1/>. Acesso em: 26 de mai. 2019.

[26] JAKOB, A. A. E.; YOUNG, A. F.; O uso de métodos de interpolação espacial de dados nas análises sociodemográficas, 2006.

[27] MAGALHÃES JÚNIOR, M.; Olhar São Paulo, Violência e Criminalidade <http://smul.prefeitura.sp.gov.br/criminalidade>. Acesso em: 10 mai. 2019.

[28] GAMMA, Erich; et al., 1995. Tradução de Luiz A. Meireles Salgado. Padrões de Projeto: Soluções Reutilizáveis de Software Orientado a Objetos.

[29] MASSARI, J.; Padrão MVC | Arquitetura Model-View-Controller <https://www.portalgsti.com.br/2017/08/padrao-mvc-arquitetura-model-view-controller.html>. Acesso em: 5 jun. 2019.

[30] MACORATTI, J. C.; Entendendo o padrão MVC - Model - View - Controller <http://www.macoratti.net/14/05/net_mvc.htm>. Acesso em: 5 jun. 2019.

[31] SOUZA, F. C. M.; Implementação de SIG e Mapas de Kernel visando Acessibilidade na Educação Superior. Universidade Federal Rural de Pernambuco. 2015.

[32] SUD, N.; Indian Online Matrimony Data Exploration. Harvard University. 2013 <https://projects.iq.harvard.edu/matrimony_data_exploration>

[33] BERTHOLDO.; Saiba como o preço do frete prejudica as vendas Disponível em: <https://www.bertholdo.com.br/blog/preco-frete-prejudica-as-vendas/>. Acesso em: 02 jun. 2020.

[34] ANTONIO, A.; Você sabia? O preço do frete pode estar impedindo 26% das suas vendas! Disponível em: <https://blog.melhorenvio.com.br/o-impacto-do-preco-do-frete-nas-vendas/>. Acesso em: 02 jun. 2020.

[35] CORREIOS. Como calcular preços e prazos de entrega em sua loja on-line. Disponível em: <https://correios.com.br/solucoes-empresariais/comercio-eletronico/como-calcular-precos-e-prazos-de-entrega-em-sua-loja-on-line>. Acesso em: 15 abr. 2020.

[36] SANTANA, W. Como Calcular o frete dos Correios. Disponível em: <http://sooho.com.br/2017/03/26/como-calcular-o-frete-dos-correios/>. Acesso em: 15 abr. 2019.

[37] PHP kmeans Examples – Hot Examples <https://hotexamples.com/examples/-/-/kmeans/php-kmeans-function-examples.html>. Acesso em: 10 de agosto de 2020.

[38] CODEIGNITER < https://codeigniter.com/docs>. Acesso em: 8 de novembro de 2020.

[39] OUTSYSTEMS <https://success.outsystems.com/Documentation>. Acesso em: 8 de novembro de 2020.