Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática

Programa de Pós Graduação em Informática Aplicada

Extração de Informações Mercadológicas a partir de Notas Fiscais Eletrônicas

Ademir Batista dos Santos Neto

Dissertação de Mestrado

Rua Manoel de Medeiros, s/n - Dois Irmãos, Recife - PE
Fevereiro 2018

Universidade Federal Rural de Pernambuco
Departamento de Estatística e Informática

Ademir Batista dos Santos Neto

Extração de Informações Mercadológicas a partir de Notas Fiscais Eletrônicas

Trabalho apresentado ao Programa de Programa de Pós Graduação em Informática Aplicada do Departamento de Estatística e Informática da Universidade Federal Rural de Pernambuco como requisito parcial para obtenção do grau de Mestre em Informática Aplicada.

Orientador: *Prof. Tiago Alessandro Espínola Ferreira*
Coorientadora: *Profa. Maria da Conceição Moraes Batista*

Rua Manoel de Medeiros, s/n - Dois Irmãos, Recife - PE
Fevereiro 2018

Dedicado a Deus que sempre me dá forças para evoluir em minha vida.

Agradecimentos

Agradeço primeiramente a Deus por ter me concedido todas as condições necessárias para que eu pudesse concluir esse trabalho e me iluminar em todo o caminho que trilhei durante o meu mestrado.

Agradeço muito aos meu pais. A minha mãe Elisângela por representar tudo para mim e sempre ter me dado amor e carinho por toda a minha vida. Ao meu pai Ademir Junior que constantemente me mostrou o caminho certo a se percorrer e me orientou com muita sabedoria nas minhas decisões.

Agradeço a toda minha família pelo amor e carinho incondicionais. Em especial a meus irmãos: Eduarda, Thaís, Hugo e Marcella. Agradeço muito ao meu padrasto Mauricio que sempre me guiou no caminho dos estudos. Agradeço também aos meus avós: Maria José, Ademir e Maria Bernadete que sempre me deram muito carinho e afeto.

Agradeço de forma especial a três pessoas que sem elas não poderia ter concluído esse trabalho. Agradeço muito ao meu orientador, professor Tiago, que me orientou em diversos aspectos, acadêmicos ou não, com toda paciência e dedicação necessária nesse processo. A minha co-orientadora a professora Ceça que sempre esteve muito solícita em suas orientações e contribuiu bastante com esse trabalho. E também ao professor Paulo Renato que apesar de não estar envolvido diretamente nesse trabalho contribuiu muito para a minha formação acadêmica.

Agradeço aos meus professores que sempre me transmitiram conhecimentos para que eu pudesse me tornar uma pessoa melhor.

Agradeço aos meus amigos e colegas por serem essas pessoas tão especiais em minha vida, que sempre colaboram para o meu desenvolvimento.

Por fim, agradeço especialmente a uma pessoa muito importante na minha vida, a minha noiva Rhayane Vaz. Obrigado por todo carinho, apoio e dedicação em que teve comigo nesse período que estamos juntos.

"We can see only a short distance ahead, but we can see plenty there that needs to be done."

—ALAN TURING

Resumo

A maior parcela das transações comerciais realizadas no Brasil são informadas ao governo federal através de notas fiscais eletrônicas (NFe). Nessas notas existem informações a respeito de uma transação comercial realizada. Neste sentido, uma vez que a informação está disponível ao acesso público através dos sites na internet da Secretaria da Fazenda nos municípios brasileiros, em posse das chaves de acesso das NFe, é possível acessar e minerar dados das notas fiscais eletrônicas geradas no Brasil, podendo assim se ter um grande volume de dados contendo informações a respeito da maioria dos segmentos do mercado brasileiro. Com a modelagem e geração de uma base de dados contendo as informações das notas fiscais eletrônicas é possível aplicar técnicas de mineração de dados para extrair diversas informações, inclusive georreferenciadas e no tempo, acerca de diversos aspectos e da dinâmica do mercado brasileiro. Um dos principais problemas abordados nesse trabalho foi o da demanda reprimida, que consiste na existência de algum tipo de restrição que impede o consumo de um determinado produto por um grupo de clientes específico. A partir de informações coletadas das notas fiscais eletrônicas é possível avaliar a existência ou não de uma demanda reprimida por determinado produto em uma região. É avaliada de maneira quantitativa a viabilidade de uma empresa abrir uma filial em uma determinada região. Para essas análises, alguns mecanismos como os de reconhecimento de padrões, clusterização e modelos de séries temporais são empregados com o intuito de consolidar melhor os resultados obtidos. A presente dissertação demonstra um conjunto de informações sobre o comércio de vários produtos de diversos setores do mercado brasileiro. Dentre os resultados obtidos tem-se uma análise sobre informações da dinâmica das leis que governam o mercado pernambucano, mais especificamente o mercado da Região Metropolitana do Recife.

Palavras-chave: Mineração de dados, Demanda reprimida, Análises mercadológicas, Clusterização, Séries temporais.

Abstract

Most of the commercial transactions in Brazil are informed to the govenmet by eletronic invoice (NFe). In these invoices, there are information about the commercial transaction carried out. This information is avaliable to the general public throught the websites from the threasuty office in the Brazilians cities, with the right keys it is possible acces and mining data from the eletroninc invoices generated in Brazil, being able to have a huge volume of data containing information about all brazillian market segments. With the modelling and the generation of the database containing the data from the eletronical invoices it is possible apply techniques of data mining and extract information, includding georefenciated and in the time, about many aspects of the Brazilian market. One of the main problems addressed in this dissertation was the Pent-up demand. From the information colleted in the eletronical invoices it is possible evaluate the existency of the Pent-up demmand about some product in a specific region. It is quantitatively evaluated, for exemple, a company open a branch in a specific region. For this analysis, some mecahnisms as: pattern recognition, clusterization and time series are employeed with the intent of consolidate the results. This dissertation presents a set of information about the comerce of many products in many fields of the brazillian market. It is expected have an relevant analysis about the dynamc of the laws that govern the Brazilian market, more specifically in the Metropolitan Region of Recife.

Keywords: Data mining, Pent-up demand, Market analysis, Clusterization, Time series

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Com o objetivo de modernizar a escrituração contábil realizada no Brasil o governo criou o Sistema Público de Escrituração Digital (SPED) através da Medida Provisória 2.200-2, de 17/08/2001, que tem como uma das suas principais vertentes a Nota Fiscal Eletrônica (NFe) [1]. A NFe é um modelo nacional, com validade jurídica garantida pela assinatura digital do emitente, reduzindo gastos da recepção, digitação e armazenamento. As primeiras notas fiscais eletrônicas foram emitidas no Brasil em 2006 em um projeto piloto e em 2017, de acordo com as últimas atualizações do Portal da Secretária da Fazenda, já são mais de 17 bilhões de notas fiscais eletrônicas autorizadas[1]. Cada nota fiscal possui diversas informações a respeito da transação comercial realizada, como por exemplo: endereço do emitente, endereço do destinatário, valor unitário do produto, quantidade vendida, valor tributado, custos de transporte, entre outras [2]. A utilização do modelo de nota fiscal eletrônica trouxe vários benefícios para a sociedade como: padronização e melhor qualidade das informações, maior eficácia no controle fiscal, diminuição da sonegação, aumento da arrecadação e redução do tempo de emissão de documento fiscal [3]. Em relação ao consumidor, uma das grandes vantagens é a consulta do documento através dos endereços eletrônicos estaduais e federal da Secretaria da Fazenda, como por exemplo, o Sefaz PE[2]. As chaves de acesso são geradas através do algoritmo *Message-Digest algorithm 5*, ou simplesmente MD5 [4]. Esse é um algoritmo de *hash*[3] unidirecional, e sendo assim o *hash* não pode ser transformado novamente no texto que lhe deu origem. Esse mecanismo garante a segurança da geração das chaves pelos órgãos emissores. As notas fiscais eletrônicas são dados públicos e com a chave da nota, ela pode ser descarregada por qualquer indivíduo [6]. Sendo assim, as informações podem ser mineradas dos portais da Secretaria da Fazenda uma vez que se tenha as chaves corretas das NFe.

Mineração de dados, de acordo com Hand *et al* [7], é uma análise de grandes conjuntos de dados a fim de encontrar relacionamentos úteis, também é entendida como resumir os dados de uma maneira que eles possam expressar alguma informação relevante. Segundo Larose [8], a mineração de dados é uma das áreas mais promissoras da atualidade por conta da grande geração de dados por companhias públicas e privadas. A captação, organização, análise e mineração de dados massivos vêm se estabelecendo cada vez com maior intensidade, tanto na academia como no mercado. Na verdade, habilidades de manuseio de grandes massas de dados são a base para a competitividade, aumento da produtividade, inovação e aumento da gama de clientes em

[1]Essa informação foi obtida no portal da Fazenda no endereço eletrônico: https://www.nfe.fazenda.gov.br/portal/infoEstatisticas.aspx Acessado em: 05-09-2017.

[2]Endereço eletrônico da Secretaria da Fazenda em Pernambuco: https://www.sefaz.pe.gov.br/SitePages/Home.aspx Acessado em: 05-09-2017.

[3]*Hash* é um algoritmo que mapeia dados de comprimento variável para dados de comprimento fixo [5].

empresas ([9], [10]). A importância da mineração de dados é embasada por diversas pesquisas. Segundo Chen [11], em uma pesquisa corporativa envolvendo mais de 4000 profissionais de tecnologia da informação de 93 países, estes declararam que a área de análise de grandes massas de dados é umas das 4 áreas mais promissoras para a segunda década do século XXI. Ainda no sentido da importância da mineração de dados, uma pesquisa realizada pelo instituto Mckinskey Global [9] prevê que vai haver uma necessidade entre 140.000 a 190.000 de profissionais habilitados a trabalhar em análise de dados nos Estados Unidos, da mesma forma que vão haver 1,5 milhões de gestores sem conhecimento suficiente para conseguir tomar decisões frente a análise de informações provenientes de grandes massas de dados.

Com os *hashs* corretos é possível minerar dados e obter muitas informações relevantes sobre a dinâmica do mercado de uma região específica. Algumas informações que podem ser obtidas através da mineração de dados de notas fiscais são: o volume de venda de um determinado produto por uma empresa, variação de preço de um conjunto de produtos, demandas reprimidas e previsão de preços de produtos. Os portais da Fazenda geralmente possuem mecanismos de segurança para dificultarem o acesso automatizado à informação mesmo em posse de chaves válidas. O mecanismo de segurança utilizado pelos portais é o CAPTCHA [6]. Trata-se de um teste de Turing público completamente automatizado para diferenciação entre computadores e humanos. A ideia central desses mecanismos é propor um teste que um humano possa facilmente resolver, mas um programa vai falhar na maioria das tentativas [12]. Existem diversos tipos de CAPTCHAs, porém o que é mais utilizado é o CAPTCHA textual [13], que também é utilizado pelos portais da Secretaria da Fazenda. Através de técnicas de reconhecimento de padrões, como o *k nearest neighbor* (*k*-NN) [14], é possível automaticamente descobrir o CAPTCHA dando possibilidade de descarregar notas fiscais para se minerar dados e gerar informações relevantes para análise de mercado.

Uma das principais propostas do presente trabalho é a análise quantitativa da demanda reprimida por determinado produto. De acordo com o dicionário financeiro Farlex [15], a demanda reprimida é uma situação em que a demanda por um determinado produto cresce de modo precipitado. O consumidor deseja adquirir um determinado produto, mas algum fator como, por exemplo, a grande distância entre ele e a loja o impede. Nesse trabalho, a demanda reprimida é definida como a população que tem um deslocamento além de três vezes o valor da média dos consumidores para a compra de um produto. Como as informações das NFe são georreferenciadas é possível apresentar em um mapa as compras de um determinado produto. Com esse mapa, indivíduos que estejam localizados a uma distância três vezes maior que a distância média percorrida pelos compradores são considerados *outliers* do processo. O valor três tem base no teorema da distribuição normal, onde um valor que esteja em até três vezes o valor médio da distribuição deve contemplar mais do que 98% dos elementos dessa distribuição [16], atingindo assim um espalhamento suficientemente longe para representar o conjunto de dados. Se o número de *outliers* for maior que 5% do total de vendas de um produto, de acordo com a metodologia desenvolvida nesse trabalho, existe um indicativo de demanda reprimida sobre esse produto naquela região e que talvez seja interessante tomar medidas para que o consumo desse produto aumente, como por exemplo, abrir uma filial da empresa nessa região. Com o mapa das vendas dos produtos é possível através de algoritmos de clusterização, como *k-means* [14], identificar onde seriam as potenciais localizações para uma possível filial que

atenderiam a maior parte da demanda reprimida de uma região específica. Além da demanda reprimida, outras informações como a relação entre compras e população de um município também é uma informação relevante que se pode obter através das notas fiscais para se traçar o perfil do consumidor de uma região. Outra informação interessante é a série histórica de preços de um determinado produto. Os preços de um produto podem variar conforme vários fatores tais como: sazonalidade, demanda, especificidade do produto, dentre outros [17]. Outro aspecto que merece destaque nesse trabalho é a questão da privacidade das NFe. Não existe até o momento do desenvolvimento desse trabalho nenhuma legislação específica que impeça o uso das informações contidas nas NFe, respeitando a identidade dos envolvidos nos processos contidos nas NFe. Dessa forma qualquer indivíduo que estiver em posse das chaves corretas de acesso das NFe pode obter as notas através do portal da SEFAZ e assim utilizar as informações nelas contidas.

Por fim, o presente trabalho visa apresentar um sistema que automaticamente gere as chaves das notas fiscais e as descarregue do portal da Secretaria da Fazenda em um banco de dados. Através desse banco de dados é possível minerar dados e descobrir informações importantes para o entendimento da dinâmica mercantil de uma região. Para esse trabalho, as análises são direcionadas para a comercialização de produtos de um conjunto de empresas que em sua maioria estão localizadas na Região Metropolitana do Recife - Pernambuco.

1.1 Motivação

O trabalho de mineração de notas fiscais eletrônicas é muito importante por diversos motivos. Um dos principais motivos é a geração de informações a respeito da dinâmica do mercado local. A ferramenta gerada pode contribuir para gestores de empresas de diversas formas. A partir de uma mineração de dados sobre uma empresa é possível acompanhar suas vendas realizadas, inclusive em relação as vendas que foram canceladas. Existe ainda a possibilidade de se realizar uma mineração de notas fiscais de empresas concorrentes sendo possível avaliar como está o mercado e como estão sendo praticados os preços dos produtos dos concorrentes. Ainda sobre o auxílio em tomada de decisão, uma importante contribuição seria a possibilidade de avaliação de maneira quantitativa e eficiente da demanda reprimida para um determinado produto. Com essas informações os gestores podem se posicionar melhor sobre suas decisões e direcionar investimentos da empresa para um produto ou serviço que seja promissor em relação às análises realizadas.

Outro motivo pelo qual o trabalho é importante é a possibilidade de acompanhar o balanço fiscal de uma determinada empresa, uma vez que os documentos minerados possuem toda a descriminação tributária das transações realizadas. Sendo assim, empresas que declararem valores equivocados para a Receita Federal podem ser autuadas com base nos documentos fiscais emitidos e avaliados na ferramenta desenvolvida neste trabalho. Além disso, o trabalho apresenta um estudo sobre o comportamento das séries históricas da variação de preços. Uma vez que se sabe sobre essa variação de preço de um produto praticado por uma empresa em um intervalo de 5 anos (quantidade de anos em que a nota fiscal fica disponível), é viável realizar análises sobre as tendências desse preço em um futuro próximo, auxiliando gestores na administração de recursos de suas empresas.

1.2 Problema e Hipótese

O principal problema abordado pela pesquisa é consolidar uma base de dados de notas fiscais eletrônicas e assim obter de maneira quantitativa e dinâmica a relação da demanda reprimida de um determinado produto. Essa demanda reprimida se caracteriza aqui nesse trabalho pelos compradores que se deslocam além da média para adquirir um produto.

A hipótese é de que através da mineração de informações das notas fiscais eletrônicas pelos portais da Secretaria da Fazenda pode-se obter os endereços dos clientes e das empresas podendo assim, apresentar um mapa com os clientes e inferir sobre os seus deslocamentos e caracterizando através de modelos estatísticos se existe ou não uma demanda reprimida para aquele produto.

1.3 Objetivos

Este trabalho tem como objetivo principal modelar e gerar uma base de dados histórica e consolidada a partir de dados obtidos de notas fiscais eletrônicas para atender a consultas e análises da indústria a respeito das transações realizadas em diversos segmentos de negócios. Com estas consultas, a indústria pode, de forma dinâmica, com o uso das dimensões espacial (georreferenciada) e temporal, definir estratégias de negócios. Um exemplo de informação consultada através dessa base histórica de dados é verificar a existência de demanda reprimida por um determinado produto de maneira dinâmica.

Além desse objetivo principal, a presente dissertação tem como objetivos específicos:

- Desenvolvimento de ferramenta computacional para a automatização do download das NFe com reconhecimento automático de CAPTCHA;
- Análises de demandas reprimidas;
- Análise das séries de preços de produtos com base nas informações históricas promovidas pela base de dados.

1.4 Contribuições Esperadas

O trabalho possui contribuições para diversas áreas do mercado local analisado. Especificamente, as contribuições desse trabalho são:

- Análise eficiente e quantitativa da dinâmica mercadológica Pernambucana, mais especificamente de um grupo de empresas da Região Metropolitana do Recife;
- Um modelo empírico para determinação de demanda reprimida para um determinado produto;
- Geração e publicação de conhecimento científico sobre análise de grandes massas de dados. Dentre as contribuição científicas um artigo já publicado no periódico *International*

Journal of Computer Applications (IJCA) sobre a metodologia de decifrar CAPTCHAs baseados em texto;

- Por fim, com as consultas ao banco de dados gerado esperasse um incentivo real ao crescimento da produção da indústria, gerando um maior rendimento da sua produção e fazendo com que as políticas mercadológicas aplicadas na produção industrial sejam ajustadas com maior verossimilhança às reais necessidades do mercado.

1.5 Estrutura da Dissertação

A presente dissertação está estruturada da seguinte forma:

- O Capítulo 1 apresenta a introdução do trabalho. Esse capítulo contém a motivação, objetivos e contribuições esperadas para o trabalho;
- O Capítulo 2 aborda a fundamentação teórica da dissertação. Trazendo os principais temas trabalhados, como: nota fiscal eletrônica, algoritmos de reconhecimento de padrão, demanda reprimida, clusterização e séries temporais;
- O Capítulo 3 traz alguns trabalhos relacionados ao tema pesquisado e também apresenta uma tabela com os pontos positivos e limitações dos trabalhos apresentados;
- O Capítulo 4 expõe a metodologia científica e prática do trabalho desenvolvido. Além disso, apresenta as ferramentas utilizadas, funcionalidades do sistema desenvolvido e comparação da dissertação com os trabalhos relacionados;
- O Capítulo 5 apresenta os resultados finais do trabalho e as análises realizadas sobre eles;
- O Capítulo 6 contém a conclusão do trabalho e também levanta alguns possíveis trabalhos futuros.

CAPÍTULO 2

Fundamentação Teórica

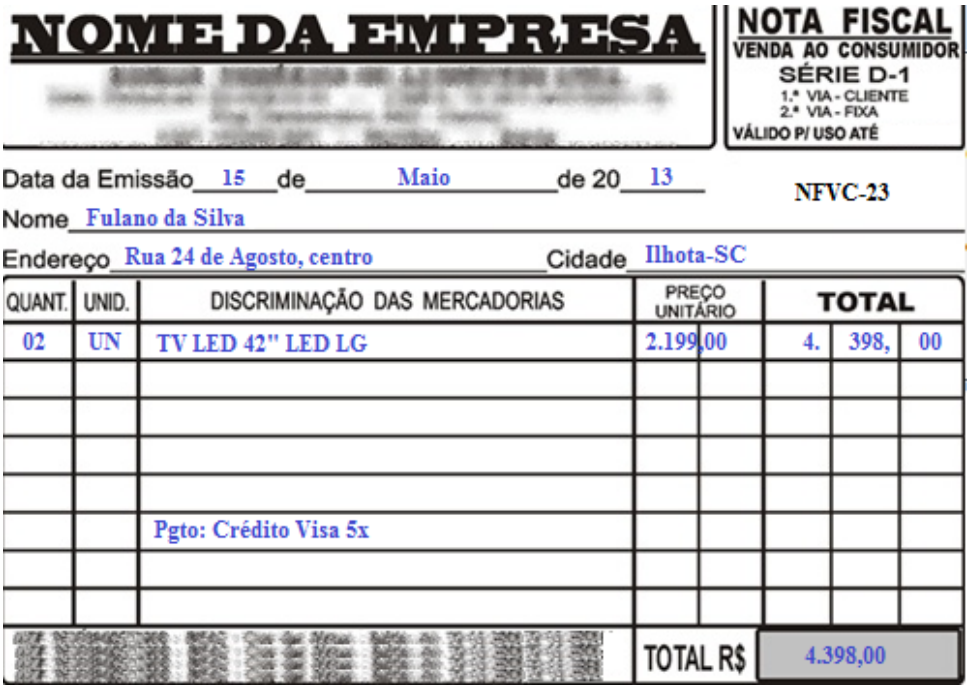
O Capítulo 2 apresenta o fundamentação teórica dos principais componentes utilizados no desenvolvimento do trabalho. Esse capítulo consiste do referencial teórico com base na literatura, relacionado ao desenvolvimento desse trabalho.

2.1 Nota Fiscal Eletrônica

A nota fiscal, de acordo com Calderelli [18] é a relação detalhada das mercadorias fornecidas por um vendedor a um comprador. Esse documento expressa determinados requisitos legais previstos em leis federais e estaduais. Esse registro comprova a existência de um ato comercial, sendo emitida a cada circulação de mercadoria, bem ou prestação de serviços. A nota fiscal é composta por diversos itens e deve conter informações a respeito do destinatário, inscrição, localização, tipo de mercadoria, quantidade, valor, imposto sobre circulação de mercadorias e serviços (ICMS) destacado e outras mais [19]. O objetivo de uma nota fiscal é o registro de uma transferência de propriedade sobre um bem ou uma prestação de serviços por uma empresa à uma pessoa física ou jurídica. A emissão da nota fiscal é uma obrigação de todos os comerciantes, pois proporciona aos fornecedores e consumidores garantias e deveres, o direito de recorrer de alguma inconsistência, e além disso é um componente importante do sistema tributário [20].

Durante muito tempo se emitiu notas fiscais manualmente em blocos de papel, como exemplifica a Figura 2.1 que expõe o modelo de nota fiscal predecessor da Nota Fiscal Eletrônica. Em 1987, foram realizados os primeiros esforços em direção da automação da emissão de notas fiscais, com a adoção do cupom fiscal em máquinas registradoras que automatizavam mecanicamente o controle de caixa [19]. Com o avanço da tecnologia surgiram novas necessidades de se utilizar o meio digital e adotar um modelo que fosse mais eficiente, nesse caso a nota fiscal eletrônica.

A informatização do projeto SPED foi instituída pelo Decreto n° 6.022, de 22 de janeiro de 2007, constituiu-se em um avanço para a informatização na relação entre o fisco e os contribuintes [19]. A Emenda Constitucional número 42 introduziu o Inciso XXII ao Artigo 37 onde determinou que as administrações tributárias da União dos Estados, do Distrito Federal e dos municípios atuassem de maneira integrada, inclusive compartilhando cadastros e informações fiscais, gerando um grande avanço para a modernização da Administração Tributária Brasileira. Para deferir essa Emenda foram realizados diversos encontros denominados Encontro Nacional dos Administradores e Coordenadores Tributários (ENCAT) tendo como participantes pessoas das administrações tributárias das esferas governamentais [19]. Desses encontros


NOME DA EMPRESA

NOTA FISCAL
VENDA AO CONSUMIDOR
SÉRIE D-1
1.ª VIA - CLIENTE
2.ª VIA - FIXA
VÁLIDO P/ USO ATÉ

Data da Emissão 15 de Maio de 20 13 NFVC-23

Nome Fulano da Silva

Endereço Rua 24 de Agosto, centro Cidade Ilhota-SC

QUANT.	UNID.	DISCRIMINAÇÃO DAS MERCADORIAS	PREÇO UNITÁRIO	TOTAL
02	UN	TV LED 42" LED LG	2.199,00	4.398,00
		Pgto: Crédito Visa 5x		
			TOTAL R$	4.398,00

Figura 2.1 Exemplo de nota fiscal no modelo 1A (papel)

surgiu o modelo da NFe e foram definidos as regras de sua implantação [1].

O Projeto NFe teve como objetivo a implantação de um modelo nacional de documento fiscal eletrônico, visando substituir a emissão do documento fiscal em papel, com validade jurídica garantida pela assinatura digital do emitente. A nota fiscal eletrônica é um documento que existe exclusivamente no meio digital, emitido e armazenado eletronicamente, com o intuito de registrar uma operação de circulação de mercadorias ou prestação de serviços, no campo de incidência do ICMS, cuja validade jurídica é garantida por duas condições necessárias: a assinatura digital do emitente e a autorização de uso fornecida pela administração tributária da esfera do contribuinte [6]. De acordo com Pereira [21] as características da nota fiscal são:

- Documento digital, que atende aos padrões definidos na MP 2.200/01, no formato *Extended Markup Language* (xml);
- Garantia de autoria, integridade e irrefutabilidade, certificada através de assinatura digital do emitente, definido pela infraestrutura de Chaves Públicas Brasileiras (ICP Brasil)[1];
- O arquivo da NFe deve seguir o *layout* de campos definidos em legislação específica como definido nos ENCAT [1];
- A NFe deve conter um código numérico, obtido por meio de algoritmo fornecido pela administração tributária, que compõe a chave de acesso de identificação da NFe, jun-

[1] A Infraestrutura de Chaves Públicas Brasileira – ICP-Brasil é uma cadeia hierárquica de confiança que viabiliza a emissão de certificados digitais. Disponível em: http://www.iti.gov.br/icp-brasil Acessado em: 12-02-2018.

tamente com o cadastro nacional de pessoa jurídica (CNPJ) do emitente e número da NFe;

- As NFe devem ser emitidas em ordem consecutiva, crescente, e sem intervalos a partir do primeiro número sequencial, sendo vedada a duplicidade ou reaproveitamento dos números inutilizados ou cancelados;
- A transmissão da NFe é efetivada, via internet, por meio de protocolo de segurança ou criptografia.

A utilização do modelo digital trouxe diversos benefícios para várias áreas da sociedade. Dentre os benefícios, de acordo com Pereira [21], pode-se citar:

- Benefícios para o emissor
  - Redução de custos de impressão;
  - Redução de custos de envio do documento fiscal;
  - Redução de custos de armazenagem de documentos fiscais.
- Benefícios para o destinatário
  - Eliminação de digitação de notas fiscais na recepção de mercadorias;
  - Planejamento de logística de entrega pela recepção antecipada da informação da NFe;
  - Redução de erros de escrituração devido a erros de digitação de notas fiscais.
- Benefícios para a sociedade
  - Redução do consumo de papel, com impacto em termos ecológicos;
  - Incentivo ao comércio eletrônico e ao uso de novas tecnologias;
  - Surgimento de oportunidades de negócios e empregos na prestação de serviços ligados à NFe.
- Benefícios para administrações tributárias
  - Aumento da confiabilidade da nota fiscal;
  - Diminuição da sonegação e aumento da arrecadação;
  - Melhoria no processo de controle fiscal, possibilitando um melhor intercâmbio e compartilhamento de informações ente os fiscos.

O processo de transição da nota fiscal para a NFe envolveu o uso de várias tecnologias tanto para a comunicação entre as partes integrantes do sistema, quanto para geração e validação do documento fiscal. Algumas das principais tecnologias envolvias no projeto da NFe são: certificação digital, padrão xml, criptografia e chaves públicas e privadas [1].

A certificação digital é o conjunto de operações utilizado para identificar e autenticar as transações eletrônicas, ela garante que algo disponível no meio digital seja de fato de quem a reivindica [22]. Portanto, através dessa tecnologia é possível garantir que o documento possa ser enviado pela internet e chegará seguro ao destinatário, e a veracidade e privacidade das informações nele contidas. Existem também as Autoridades Certificadoras que têm o poder de emitir os certificados digitais. Apenas o emitente possui o certificado digital que pode assegurar a veracidade das informações contidas no documento assinado, dessa maneira apenas o destinatário correto pode abrir o documento em questão, sem riscos de interceptação ou adulteração por terceiros. A NFe tem sua validade jurídica assegurada pela assinatura digital do emitente.

A linguagem de marcação xml é utilizada para criar documentos cujos dados precisem ser organizados hierarquicamente, e se concentra na estrutura da informação [1]. A justificativa principal para a escolha do xml como padrão para a NFe é que o formato "é livre de licenças" [23], o que permite que seja utilizado um *software* de leitura e emissão próprio.

Com a finalidade de evitar que um arquivo seja lido ou interceptado durante o envio, é utilizado o processo de criptografia. A criptografia é o ato de codificar as informações garantindo assim sigilo e sua autenticidade [24]. Uma vez criptografadas, as informações contidas no arquivo só poderão ser lidas com a correta chave e algoritmo de criptografia. Com relação à NFe, a criptografia é utilizada quando o emissor da nota valida as informações contidas através do seu certificado digital.

As chaves públicas e privadas são instrumentos utilizados na des/criptografia dos documentos. Também conhecida como criptografia assimétrica, é uma classe de criptografia baseada em algoritmos que requerem duas chaves, uma delas pública e outra privada. A chave privada é usada para encriptar texto e a pública é usada no processo inverso. No caso da NFe, a chave privada é a assinatura digital do emissor, pois somente ele pode atestar a propriedade da assinatura. O receptor em posse da chave pública (chave de acesso) pode descriptografar as informações da NFe e ter acesso livre a elas [24].

2.1.1 Descrição do Modelo de Comunicação

As Secretarias da Fazenda Estaduais disponibilizam os serviços de recepção e consultas de NFe. Cada um desses serviços é realizado por um *webservice*[2] específico. O fluxo de comunicação para acesso a NFe é iniciado por um aplicativo disponibilizado nos portais da Secretaria da Fazenda onde o cliente insere a chave de acesso da NFe e um código de segurança (CAPTCHA). Com isso o aplicativo envia uma mensagem para o *webservice* que processa a solicitação, manda uma resposta para o cliente confirmando a solicitação e enviando a resposta caso as informações informadas pelo cliente estejam corretas [6]. A Figura 2.2 apresenta o fluxo de comunicação entre cliente e servidor para acesso a NFe.

Seguindo o fluxo da Figura 2.2, (1) O aplicativo do contribuinte inicia a conexão enviando uma mensagem de solicitação de serviço para o *webservice*. (2) O *webservice* recebe a mensagem de solicitação de serviço e encaminha ao aplicativo da NFe que irá processar o serviço solicitado. (3) O aplicativo da NFe recebe a mensagem de solicitação de serviço e realiza o

[2] *Webservice* é uma solução utilizada na integração de sistemas e na comunicação entre aplicações diferentes.

Figura 2.2 Comunicação entre contribuinte e portal SEFAZ para transações envolvendo NFe

processamento, devolvendo uma mensagem de resultado do processamento ao *webservice*. (4) O *webservice* recebe a mensagem do resultado do processamento e o encaminha ao aplicativo do contribuinte.

2.1.2 Composição da Nota Fiscal Eletrônica

A chave de acesso da nota fiscal eletrônica é o *hash* para acesso do documento nos portais da Secretaria da Fazenda. Essa chave contém algumas informações sobre a NFe e é composta por uma sequência de 44 caracteres numéricos, representados na Tabela 2.1.

Tabela 2.1 Composição da chave de acesso da NFe

	cUF	AAMM	CNPJ	Mod	Série	nNF	tpEmis	cNF	cDV
Quantidade de Caracteres	02	04	14	02	03	09	01	08	01

As siglas apresentadas na Tabela 2.1 representam, de acordo com o Manual do Contribuinte [6], as seguintes informações:

- cUF - Código da UF do emitente do Documento Fiscal. Cada Unidade Federativa possui um código de dois dígitos que o representa de acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE);
- AAMM – Ano e mês da emissão da NFe;
- CNPJ – CNPJ do emitente;
- Mod – Modelo do Documento Fiscal que geralmente é representado pelo valor 55 indicando modelo de NFe emitido em substituição ao modelo 1 ou 1A que representa notas fiscais emitidas em papel;
- Série - Série do Documento Fiscal;
- nNF - Número do Documento Fiscal;
- tpEmis - Tipo da emissão;
- cNF - Código numérico que compõe a chave de acesso;

- cDV - Dígito verificador da chave de acesso. O dígito verificador irá garantir a integridade da chave de acesso, protegendo-a principalmente contra digitações erradas. O dígito verificador da chave de acesso da NFe é baseado em um cálculo do módulo 11[3] dos 43 primeiros dígitos da chave da nota fiscal.

O cliente pode ver as informações da NFe através do documento auxiliar da nota fiscal eletrônica (DANFE) que pode ser impresso a fim de facilitar o acesso as informações, colher firma do destinatário no ato da entrega e acompanhar o trânsito de mercadorias. A Figura 2.3 demonstra um exemplo de DANFE contendo os principais campos de uma nota fiscal.

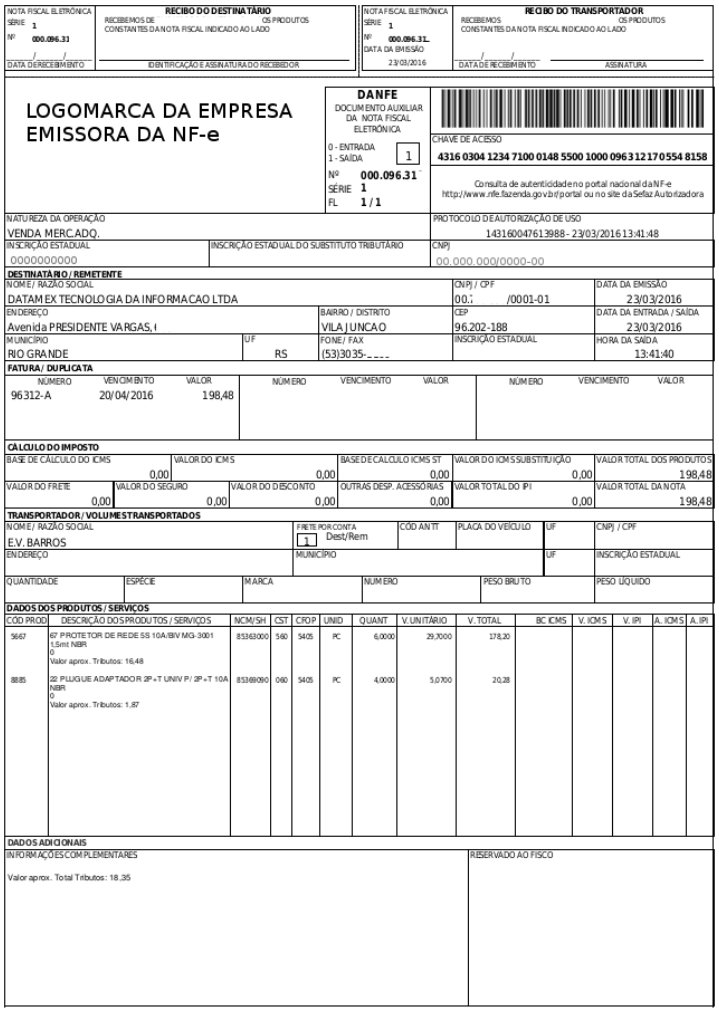

NOTA FISCAL ELETRÔNICA
SÉRIE 1
Nº 000.096.31
DATA DE RECEBIMENTO
RECIBO DO DESTINATÁRIO
RECEBEMOS DE OS PRODUTOS CONSTANTES DA NOTA FISCAL INDICADO AO LADO
IDENTIFICAÇÃO E ASSINATURA DO RECEBEDOR
NOTA FISCAL ELETRÔNICA
SÉRIE 1
Nº 000.096.31.
DATA DA EMISSÃO 23/03/2016
RECIBO DO TRANSPORTADOR
RECEBEMOS OS PRODUTOS CONSTANTES DA NOTA FISCAL INDICADO AO LADO
DATA DE RECEBIMENTO
ASSINATURA
LOGOMARCA DA EMPRESA EMISSORA DA NF-e
DANFE
DOCUMENTO AUXILIAR DA NOTA FISCAL ELETRÔNICA
0 - ENTRADA
1 - SAÍDA 1
Nº 000.096.31
SÉRIE 1
FL 1 / 1
CHAVE DE ACESSO
4316 0304 1234 7100 0148 5500 1000 0963 1217 0554 8158
Consulta de autenticidade no portal nacional da NF-e http://www.nfe.fazenda.gov.br/portal ou no site da Sefaz Autorizadora
NATUREZA DA OPERAÇÃO
VENDA MERC.ADQ.
PROTOCOLO DE AUTORIZAÇÃO DE USO
143160047613988 - 23/03/2016 13:41:48
INSCRIÇÃO ESTADUAL
0000000000
INSCRIÇÃO ESTADUAL DO SUBSTITUTO TRIBUTÁRIO
CNPJ
00.000.000/0000-00
DESTINATÁRIO / REMETENTE
NOME / RAZÃO SOCIAL
DATAMEX TECNOLOGIA DA INFORMACAO LTDA
CNPJ / CPF
00. /0001-01
DATA DA EMISSÃO
23/03/2016
ENDEREÇO
Avenida PRESIDENTE VARGAS, (
BAIRRO / DISTRITO
VILA JUNCAO
CEP
96.202-188
DATA DA ENTRADA / SAÍDA
23/03/2016
MUNICÍPIO
RIO GRANDE
UF
RS
FONE / FAX
(53)3035-____
INSCRIÇÃO ESTADUAL
HORA DA SAÍDA
13:41:40
FATURA / DUPLICATA
NÚMERO VENCIMENTO VALOR
96312-A 20/04/2016 198,48
CÁLCULO DO IMPOSTO
BASE DE CÁLCULO DO ICMS 0,00
VALOR DO ICMS 0,00
BASE DE CÁLCULO ICMS ST 0,00
VALOR DO ICMS SUBSTITUIÇÃO 0,00
VALOR TOTAL DOS PRODUTOS 198,48
VALOR DO FRETE 0,00
VALOR DO SEGURO 0,00
VALOR DO DESCONTO 0,00
OUTRAS DESP. ACESSÓRIAS 0,00
VALOR TOTAL DO IPI 0,00
VALOR TOTAL DA NOTA 198,48
TRANSPORTADOR / VOLUMES TRANSPORTADOS
NOME / RAZÃO SOCIAL
E.V. BARROS
FRETE POR CONTA 1 Dest/Rem
CÓD ANTT
PLACA DO VEÍCULO
UF
CNPJ / CPF
ENDEREÇO
MUNICÍPIO
UF
INSCRIÇÃO ESTADUAL
QUANTIDADE ESPÉCIE MARCA NÚMERO PESO BRUTO PESO LÍQUIDO
DADOS DOS PRODUTOS / SERVIÇOS
COD PROD DESCRIÇÃO DOS PRODUTOS / SERVIÇOS NCM/SH CST CFOP UNID QUANT V. UNITÁRIO V. TOTAL BC ICMS V. ICMS V. IPI A. ICMS A. IPI
5667 67 PROTETOR DE REDE SS 10A/BIV MG-3001 1,5mt NBR 0 Valor aprox. Tributos: 16,48 85363000 560 5405 PC 6,0000 29,7000 178,20
8885 22 PLUGUE ADAPTADOR 2P+T UNIV P/ 2P+T 10A NBR 0 Valor aprox. Tributos: 1,87 85369090 060 5405 PC 4,0000 5,0700 20,28
DADOS ADICIONAIS
INFORMAÇÕES COMPLEMENTARES
Valor aprox. Total Tributos: 18,35
RESERVADO AO FISCO

Figura 2.3 Exemplo de documento auxiliar de nota fiscal

O DANFE é uma visualização do arquivo xml que representa a NFe. Porém o arquivo em formato xml é onde se encontram todas as informações das notas fiscais. A Figura 2.4 apresenta

[3]O módulo 11 de um número é calculado multiplicando-se cada algarismo pela sequência de multiplicadores 2,3,4,5,6,7,8,9,2,3, ..., posicionados da direita para a esquerda. O somatória dos resultados das ponderações dos algarismos é dividida por 11 e o DV (dígito verificador) será a diferença entre o divisor (11) e o resto da divisão.

um trecho de um arquivo xml de NFe. Nesse arquivo a hierarquia das informações são divididas em *tags* e *subtags* onde os valores são destacados dentro dessas estruturas hierarquicamente distribuídas.

Figura 2.4 Trecho de arquivo xml de uma NFe

A NFe possui informações sobre: emitente, destinatário, produto, impostos, transporte e outras mais. A Tabela 2.2 apresenta os principais componentes e suas respectivas informações que estão presentes na NFe [6].

Tabela 2.2 Campos presentes na NFe

Grupo NFe	**Informações**
Identificação NFe	Informações da chave de acesso, data emissão e data saída.
Identificação do Emitente	CNPJ, CPF, nome, endereço, telefone, inscrição estadual e Classificação Nacional de Atividade Econômica (CNAE).
Identificação do Destinatário	CNPJ, CPF, nome e endereço.
Produto e Serviços	Código Produto, nome, Nomeclatura Comum do Mercosul (NCM), unidade comercial, quantidade comercializada, valor unitário, valor total bruto e valor unitário de tributação.
Grupo de Tributação	ICMS, Programa de Integração Social (PIS) e COFINS.
Valores totais	Valor total da NFe, valor frete e valor total tributado.
Informações de transporte	Modalidade de frete, CNPJ ou CPF do transportador, endereço, valor do serviço, placa do veículo e UF do veículo.

Após ser explicado o modelo da nota fiscal eletrônica, agora é abordado a classe e algoritmo que foi utilizado para o reconhecimento de caracteres na comunicação com o servidor da

SEFAZ.

2.2 Algoritmos de Reconhecimento de Padrões

De acordo com Duda [14], reconhecimento de padrões é um campo da ciência voltado a utilização de máquinas para o reconhecimento de regularidades em ambientes ruidosos e complexos. Uma das vertentes da área de reconhecimento de padrões lida com a classificação de um objeto dentro de uma categoria (classe) de maneira automática [25]. Esse classificação pode ser realizada de maneira: (i) supervisionada (*eg.* análise discriminante) onde o padrão é definido como um membro da classe predefinida; (ii) não-supervisionada (*eg*. *k-means*) onde o padrão não é atribuído para nenhuma classe previamente definida [26]. Um sistema de reconhecimento de padrões envolve essencialmente três aspectos: aquisição e pré-processamento dos dados, representação do conjunto de dados e tomada de decisão. Usualmente a classificação consiste em: definir o espaço amostral dos elementos, utilizar uma técnica de preprocessamento, representar o esquema dos elementos e então realizar o modelo de decisão [27]. A Figura 2.5 representa uma sintetização do processo de reconhecimento de padrão supervisionado.

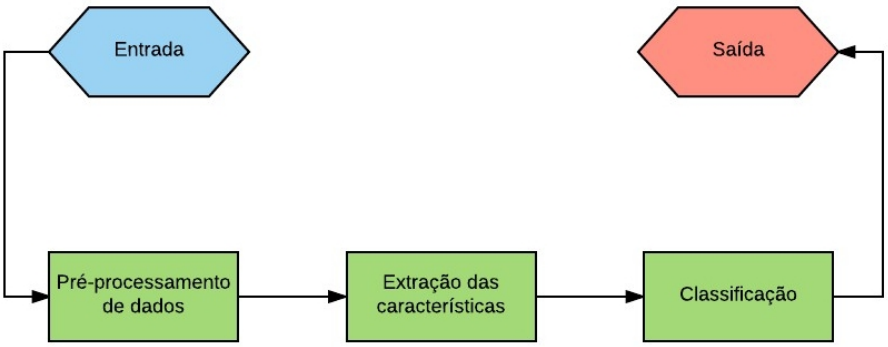


Figura 2.5 Síntese de um processo de reconhecimento de padrão

Existem diversos métodos de reconhecimento de padrões, dentre eles pode-se citar: reconhecimento de padrões estatísticos [27], clusterização [14], regras *fuzzy* [28], redes neurais [25], reconhecimento de padrões estrutural [29], SVM (*supported vector machine*) [30] e outros mais [31]. Dentre esses métodos, uma das técnicas supervisionadas mais simples e difundidas para reconhecimento de padrões em imagens é denominada *k*-NN. Devido a sua simplicidade em termos de implementação, aliada a eficiência da execução do algoritmo, essa técnica foi selecionada para integrar o sistema desenvolvido nesse trabalho no sentido de preencher automaticamente os CAPTCHAs das requisições ao servidor de notas fiscais.

2.2.1 K-NN

Dentre os algoritmos de classificação, existe a classe denominada não paramétrica, que significa que eles não utilizam parâmetros nas suas classificações. Alguns exemplos desses algoritmos

são: Estimação de Densidade, *Parzen Windows* e k-NN [14]. O k-NN (*k nearest neighbor*) é um algoritmo para classificar elementos em uma classe dependendo da distância desse elemento para os seus vizinhos em relação as suas características. O k-NN é classificado como um algoritmo de aprendizado preguiçoso porque toda a computação é deferida no momento da classificação do elemento. Esse é um dos mais simples algoritmos de aprendizado de máquina. O k-NN é uma técnica bastante difundida e vários trabalhos propõem modificações no algoritmo para melhorar a eficiência e acurácia em uma classificação específica ([32], [33]).

Seja n um arranjo de todos os elementos do universo U onde cada elemento pertence a uma classe C_i ($i = 1,2,\ldots,W$, onde existam W classes). As classes W são definidas *a priori* e contém os elementos que têm características similares. Suponha que existe um elemento x e ele pertence ao universo U. O algoritmo k-NN classifica o elemento x em uma das classes em W comparando as suas características com as dos outros elementos do universo U (n elementos). Para isso, o k-NN mede a distância entre x e todos os elementos do universo U e verifica a qual classe seus k vizinhos pertencem. Em caso de todos os elementos pertencerem ao uma mesma classe, o k-NN classifica o elemento x para a mesma classe que os k vizinhos pertençam. Caso os k vizinhos pertençam a duas ou mais classes, o algoritmo k-NN classifica o elemento x para a classe que tem a maior quantidade de vizinhos k. A Figura 2.6 apresenta uma distribuição de n elementos em um eixo cartesiano representando o universo U de três categorias distintas e o elemento x com os seus k vizinhos circulados.

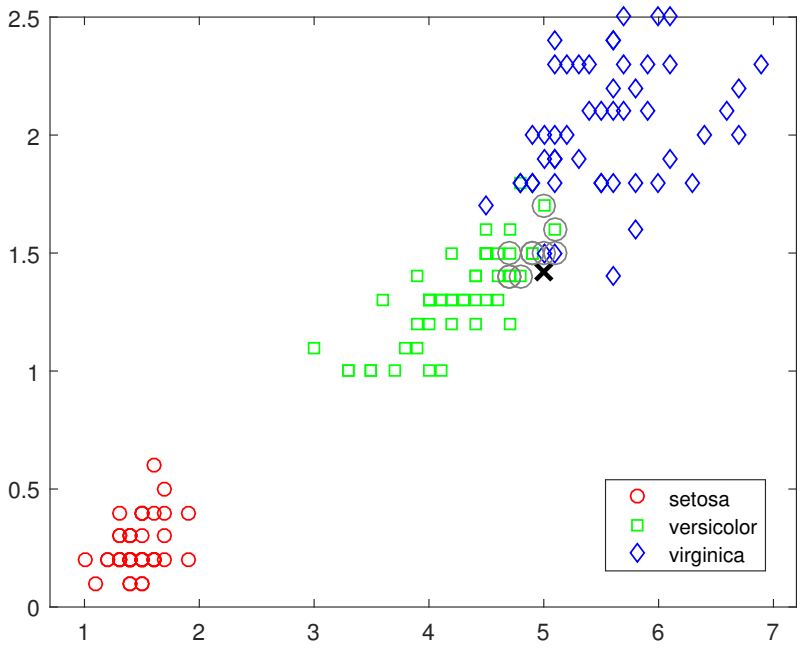


Figura 2.6 Exemplo de delimitação da vizinhança na classificação por k-NN

Essa Figura 2.6 apresenta um conhecido conjunto de dados das características de diferentes tipos de plantas íris [34]. A Figura representa a distribuição dos elementos conforme suas

categorias. O elemento x que está sendo classificado foi atribuído a classe 'versicolor' uma vez que a maior parte da sua vizinhança pertence a essa categoria.

Quanto maior o número de elementos no universo U, melhor é a estimação do algoritmo. Quando a quantidade de elementos em n tende ao infinito ($\lim_{n\to\infty}$), a classificação do k-NN converge para a classe a qual o elemento que está sendo classificado pertence [14]. O problema de um grande valor de n é o poder computacional requerido, porque quanto maior o número de elementos em n mais recursos computacionais são necessários para executar o algoritmo. O desafio é encontrar o maior número de n elementos que o computador possa executar em um tempo aceitável, com bons resultados. Nesse trabalho, como o universo possui 19 categorias foram utilizados 100 elementos de cada categoria, sendo assim um universo de 1900 elementos. Para estimar o número de k vizinhos, uma boa estimativa é a raiz quadrada do número de elementos no vetor n, então $k \approx \sqrt{n}$ [14]. Essa é apenas uma estimação, o número que minimiza o erro na classificação varia de acordo com as especificações do problema, tais como: tamanho do vetor n, o estudo de caso, a medida utilizada para checar a distância entre os elementos. Um valor pequeno para k pode não ser suficiente para caracterizar corretamente a qual classe W o elemento x pertence. Por outro lado, um valor alto de k talvez tenha muito mais elementos do que uma classe possui, prejudicando assim a classificação.

Cada elemento x possui suas próprias características (ou traços) e para medir a distância D entre esse elemento e os outros elementos incidentes do universo U é possível usar diferentes métricas. Em particular, a métrica de Minkowski é bastante utilizada [35]:

$$D(x, C_i) = \left(\sum_{v=1}^{n_i} |x - x_v|^p \right)^{1/p} \tag{2.1}$$

Onde n_i é o número de elementos pertencentes à classe C_i. A variável x representa o elemento que se está classificando e x_v um elemento pertencente a classe C_i. Portanto a Equação 2.1 é a medida da distância entre os elementos x e todos os elementos da classe C_i.

A Equação 2.1 é uma métrica geral para medir distância entre dois pontos. A métrica mais usada para medir distância entre os pontos é a distância Euclidiana (quando $p = 2$) e a Manhatann ou *city block* ($p = 1$) [35]. Contudo, existem outras métricas propostas, como a métrica de Tanimoto, comumente utilizada em taxonomia [14]. De fato, não é trivial a escolha de uma métrica mais adequada a um problema. A seleção da métrica é geralmente direcionada a limitações computacionais, requerendo experimentações para avaliar qual métrica é mais adequada. Nesse trabalho, após uma série de testes para ver quais seriam os melhores parâmetros para a classificação estudada, foi definido a distância euclidiana e o valor de k igual a 1.

Depois de definido a métrica que vai ser usada e o número de vizinhos k que serão aplicados através de testes estatísticos, o algoritmo vai comparar o elemento x em todo o vetor n e verificar quais são os k vizinhos mais próximos. Portanto, a classe que o elemento x vai pertencer vai ser a classe que mais for presente entre os seus k vizinhos. O Algoritmo 1 apresenta o pseudocódigo do funcionamento geral do k-NN.

Classificar (U,W,u) // U: dados de treinamento, W: rótulo das classes, u: elemento a ser classificado;
i=1;
while *i< U lenght* **do**
    Calcule a distância D(U_i,u);
    i = i+1;
end
Calcule o conjunto I que contém os índices para as k menores distâncias D(U_i, u);
Result: Rótulo que tiver maior representatividade em [W_i onde $i \in I$]

Algoritmo 1: K Nearest Neighbor

Apresentado o algoritmo que foi utilizado para a classificação das imagens de CAPTCHA nos servidores da SEFAZ, agora é demonstrada a metodologia proposta pelos autores de definição da demanda reprimida. Essa demanda reprimida pode ser calculada com as informações que foram obtidas pela mineração de notas fiscais eletrônicas.

2.3 Demanda Reprimida

A demanda natural é o interesse dos consumidores por determinado produto. Já a demanda reprimida ocorre quando um público tem o desejo, mas não as condições de adquirir um produto por algum motivo, como por exemplo: alto custo, acesso reduzido ao crédito, oferta reduzida, ação protecionista do governo e distância entre os compradores e o produto [36]. Para identificar uma demanda reprimida, a melhor saída é recorrer à análise de mercado. Desse modo, é possível entender o perfil e comportamento do público que se busca atingir, encontrando possíveis demandas reprimidas [36]. Usualmente a análise de mercado busca avaliar perfis, registrar informações e comparar dados. A análise de mercado é uma das maneiras para encontrar oportunidades de negócio. Morrison [37] afirma que a demanda reprimida acontece quando as vendas de um produto em lançamento são substanciais e não há estoque suficiente para atender a demanda. Uma outra forma de encontrar demandas reprimidas é acompanhar grandes companhias no seu mercado. Por exemplo, o crescimento de vendas de *smartphones* trouxe consigo a demanda por acessórios [36]. Muitas empresas também investem em pesquisa de mercado para saberem quais são as necessidades dos seus consumidores. Contudo, é muito difícil realizar uma avaliação quantitativa e dinâmica da demanda reprimida.

De posse de dados georreferenciados de clientes e vendedores, além dos produtos comercializados, é possível plotar as compras realizadas em um mapa e se ter a distribuição das vendas realizadas por uma empresa. Se a quantidade de dados analisados for relativamente considerável ao volume real de vendas da empresa, é possível traçar o perfil geográfico do consumidor. Dentro desses grupos, é possível destacar os consumidores que se deslocam além da média para obterem um determinado produto. Com informações das distâncias de todos os clientes de um produto e com base na teoria da distribuição normal [16] é possível afirmar que um cliente que esteja 3 vezes mais distante do que a distância média é considerado um *outlier* do processo. Sendo x o vetor das distâncias percorridas por clientes para comprar um produto, o raio da demanda reprimida (ϕ) pode ser calculada de acordo com a Equação 2.2. Nessa equação N representa o número de vendas e x um conjunto das distâncias percorridas.

$$\phi = 3.\frac{\sum_{x=1}^{x=N} x_i}{N} \tag{2.2}$$

Ao se identificar o grupo de pessoas que se deslocam além da média, pode-se inferir que o local não dispõe de um determinado produto, demandando que eles tenham que realizar esse deslocamento além do usual para realizar a compra. Se a quantidade de *outliers* do processo for considerável, conforme a Tabela 2.3 proposta por esse trabalho, é possível que seja uma boa oportunidade de negócio inserir um ponto de vendas do produto nesse local.

Tabela 2.3 Tabela indicando a categoria a qual pertence o produto conforme análise do percentual de vendas fora do raio da demanda reprimida

Classificação	**Percentual**
Não há demanda reprimida	>5%
Há indício de demanda reprimida	<5% >15%
Forte indício de demanda reprimida	<15%

A Figura 2.7 mostra um mapa com uma exemplificação de vendas de um produto por uma loja, na qual, o raio da demanda reprimida com a linha tracejada e o grupo de clientes que ficaram fora do raio da demanda reprimida.

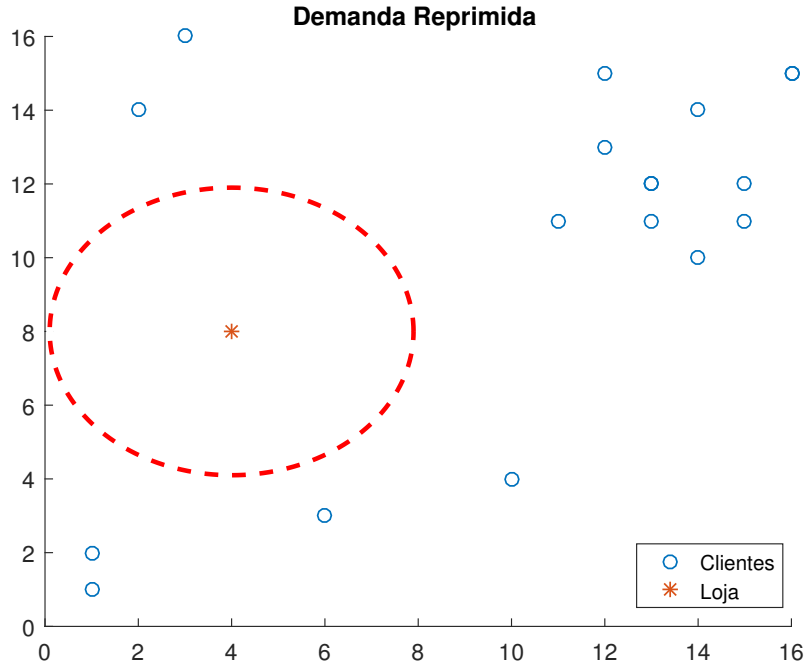


Figura 2.7 Demonstração da demanda reprimida no eixo cartesiano

Na Figura 2.7 a loja é destacada no centro da linha tracejada que representa o raio da demanda reprimida. Esse linha foi calculada com base na distância média percorrida pelos clien-

tes. A imagem apresenta alguns pontos fora desse raio que representam os *outliers* do processo. As áreas onde se concentram os *outliers* são os locais que existem demanda reprimida.

É muito difícil obter-se de maneira dinâmica e quantitativa a demanda reprimida por um produto. Em geral, o que existe são pesquisas de mercado que usam critérios subjetivos para análises de aceitação de produtos e acompanhamento de consumo. Uma análise tão acurada de demanda reprimida juntamente com a experiência dos gestores podem ser grandes aliadas na tomada estratégica de decisão. Através dessa metodologia proposta para se definir clientes além do raio de abrangência da demanda reprimida espera-se conseguir auxiliar empresas de vendas de produtos a descobrirem boas oportunidades de negócios.

Após visto o conceito definido para a demanda reprimida, é definido o conceito de clusterização (agrupamento) e o funcionamento do algoritmo *k-means* (k-médias) que é utilizado no trabalho. Esse conceito é importante nesse trabalho pois após identificados quais são os pontos onde existem *outliers* o algoritmo de agrupamento vai apontar onde eles se concentram e qual ou quais pontos são os mais indicados para representar um determinado grupo de *outliers*.

2.4 Clusterização

Clusterização consiste em agrupar os elementos de um conjunto de dados de acordo com suas características de modo que elementos similares estejam em um mesmo *cluster*[4] e elementos mais distintos em *clusters* diferentes [38]. O objetivo de uma clusterização é descobrir os agrupamentos naturais dos pontos ou objetos de um conjunto de dados a partir de suas características [39]. A clusterização pode ser realizada com um número k de *clusters* predefinido ou de maneira que o próprio algoritmo reconheça o número de *clusters* necessários [40]. Quando o número k de *clusters* é definido, essa clusterização é denominada "k-clusterização" [41], exemplos de algoritmos de k-clusterização são: *k-means* [42], *expectation maximization* [43] e *k-medoids* [44]. Um problema de clusterização demanda um alto processamento computacional uma vez que as possíveis combinações entre os elementos em um número predefinido de *clusters* é alta. Em uma k-clusterização, o número total de diferentes formas de agrupamento de n elementos de um conjunto em k *clusters*, equivale à função $N(n,k)$ apresentada na Equação 2.3 [38]. Nessa Equação n representa o número de indivíduos a serem agrupados em *clusters* e k representa a quantidade de *cluster*.

$$N(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^n \binom{k}{i}(k-i)^n \tag{2.3}$$

Essa função tem um crescimento exponencial. Para exemplificar, as possíveis combinações para essa função é apresentado esses dois exemplos: $N(10,2) = 511, N(100,2) = 6,3382510^{29}$. Outro aspecto da clusterização é como medir o quanto um elemento é similar ao outro. Para se realizar a separação dos elementos é utilizado a distância que existe entre eles. Sendo assim os elementos que possuírem menor distância entre eles devem estar agrupados em um mesmo *cluster*. Uma das medidas de distância muito utilizada é a distância Euclidiana [38].

[4] *Cluster* é um agrupamento de indivíduos similares em determinados aspectos para um objetivo comum.

Não é computacionalmente viável testar todas as possibilidades de *clusters* que existem em um conjunto de dados, sendo assim se faz necessário a adoção de alguma heurística para reduzir a complexidade do problema de clusterização. As heurísticas podem ser classificadas em hierárquica ou particionada [27]. Em uma clusterização hierárquica os *clusters* vão se formando gradativamente através das divisões dos *clusters*, gerando uma hierarquia de *clusters*, que pode ser representada em estrutura de árvore. Já em uma clusterização particionada cada elemento do conjunto é associado a um *cluster* distinto, e novos *clusters* vão se formando a partir de *cluster* existentes, essa união ocorre de acordo com a proximidade em que um *cluster* está do outro. Para técnicas de k-clusterização com heurísticas de aglomeração, como no *k-means*, o ponto no espaço que melhor representa o *cluster* é definido como o centróide do *cluster*, e tem os valores médios dos atributos considerados para aquele *cluster*. Esses centróides são utilizados para definir a qual *cluster* o elemento melhor se adequa [38].

Como já demonstrado na Equação 2.3 um algoritmo de clusterização lida com um problema difícil de análise combinatória. Para o raciocínio humano é difícil pensar em um procedimento de clusterização com um vetor de mais de duas dimensões, contudo a maioria dos problemas reais envolvem mais do que esse número de dimensões [45]. Não existe uma técnica universal que abranja todos os problemas de agrupamento, cada problema exige uma determinada particularidade no tratamento dos dados. Contudo, existe um conjunto de etapas estruturadas comuns a algoritmos de clusterização convencionais. De acordo com Jain e Dubes [46], uma clusterização típica contém as seguintes etapas: Seleção de características, identificar similaridades e agrupamento. A Figura 2.8 apresenta uma organização das etapas de um processo de clusterização.

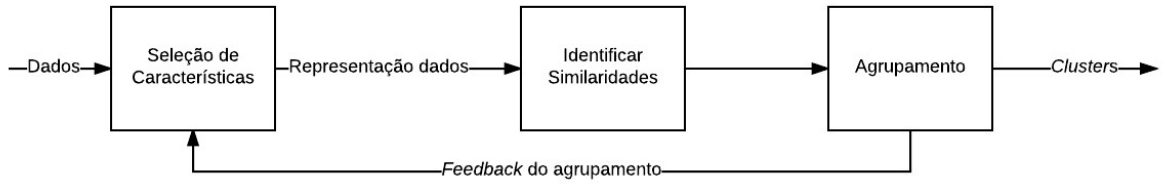


Figura 2.8 Etapas de um processo de clusterização

Seguindo o fluxograma da Figura 2.8 a primeira etapa é a *seleção de características* que identifica de maneira efetiva as características dos dados que serão usadas para a clusterização e as transcreve de modo que a representação dos dados esteja de maneira compacta. A etapa de *representação dos dados*, define-se o número de: classes, padrões disponíveis no contexto, tipo e as escalas que são acessíveis ao algoritmo de clusterização. A *identificação de similaridades* usualmente utiliza alguma função que calcula a similaridade entre os elementos. E por fim, *o agrupamento* posiciona os elementos que tem características mais similares, essa etapa pode ser executada várias vezes de maneira supervisionada ou não. O resultado do processo são elementos dos conjuntos de dados organizados em *clusters* [45].

2.4.1 *K-means*

Um algoritmo muito usado para clusterização é o *k-means*. Mesmo sendo proposto em 1956 por Steinhaus *et al* [42] ele continua sendo utilizado para diversos tipos de problemas de clusterização. Facilidade de implementação, simplicidade, eficiência e considerável sucesso nos resultados são algumas razões do porquê da popularidade dessa algoritmo [39].

Seja $X = \{x_i\}, i = 1,..,n$ um conjunto de dimensão de d pontos para serem clusterizados em k *clusters*, $C = \{c_k, k = 1,...,k\}$. O algoritmo *k-means* procura uma partição cujo o erro quadrático e a média empírica do *cluster* entre os pontos sejam minimizados. Seja μ_k a média do *cluster* c_k. O erro quadrático médio entre μ_k e os pontos no *clusters* c_k são calculados pela distância euclidiana entre esses pontos. O objetivo do algoritmo é reduzir ao mínimo possível o erro quadrático médio sobre todos os k *clusters* como mostra a Equação 2.4, onde o erro quadrático médio é representado por $J(C)$ [39].

$$J(C) = \sum_{i=1}^{k} \sum_{x_i \in c_k} \| x_i - \mu_k \|^2 \tag{2.4}$$

O algoritmo *k-means* começa com um valor inicial de k *clusters*, os vetores de dados e a métrica utilizada para calcular as distâncias. O erro quadrático reduz a medida que o número de *clusters* aumenta ($J(C) = 0$ quando $k = n$). Os passos principais do *k-means* são: seleção de uma partição inicial com k *clusters*, gerar uma nova partição colocando cada elemento pertencente a seu centróide mais próximo e computar novos centróides até que o erro se estabilize [46]. A escolha mais importante dos parâmetros é o valor de k. Não existe nenhum critério matemático que vai indicar o melhor valor para k, o que existem são heurísticas [47] que visam melhorar a estimação inicial para esse valor. Diferente valores de k podem gerar resultados do erro quadráticos diferentes. A Figura 2.9 mostra um exemplo de clusterização dividindo um conjunto de dados de dois grupos distintos. Dado esse vetor de pontos, o algoritmo ajusta os centróides dos k *clusters* (2 nesse caso) e atribui cada elemento para um dos *clusters* conforme as suas distâncias dos centróides de cada *clusters*. Esses valores dos centróides são ajustados de maneira que o erro quadrático entre os elementos seja o menor possível.

Ao longo do tempo, várias propostas foram formalizadas para melhorar o desempenho do *k-means* através de diversas técnicas ([48], [49]). Em especial, o *k-means* global, proposto por Likas *et al* [50], é uma adaptação do algoritmo original que não depende de nenhum parâmetro inicial, otimizando o tempo de execução do algoritmo e reduzindo os erros dos *clusters*. A ideia consiste em ao invés de distribuir os elementos nos *clusters*, o algoritmo adiciona os novos *clusters* de maneira incremental, otimizando assim o processo. Devido a sua eficiência e simplicidade em termos de implementação o algoritmo *k-means* foi utilizado nesse trabalho para a clusterização dos clientes que ficaram além do raio da demanda reprimida.

O Algoritmo 2 apresenta um pseudocódigo com funcionamento do algoritmo *k-means* descrito nessa seção e que foi utilizado nesse trabalho para o agrupamento dos *outilers* desse processo.

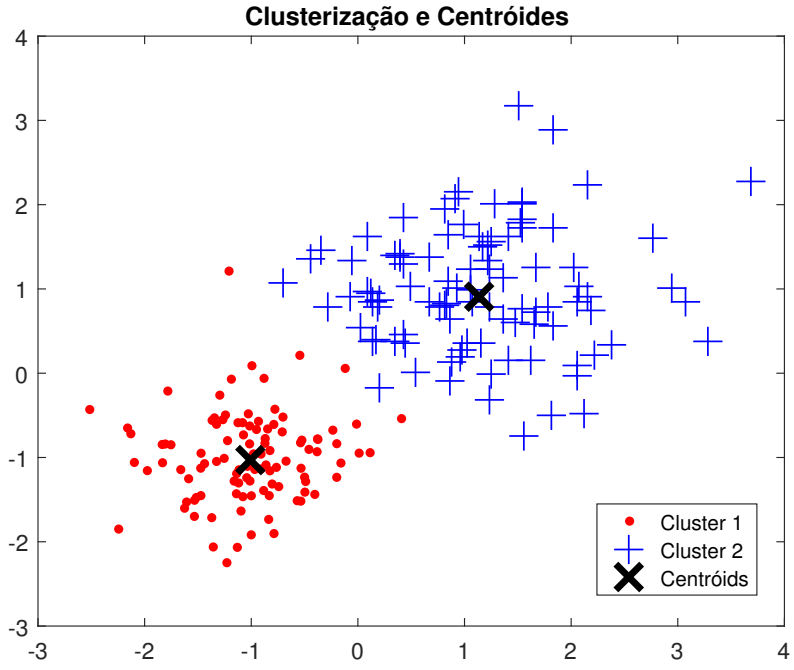


Figura 2.9 Exemplo de clusterização através do algoritmo *k-means*

K-means (X,K) // X: conjunto a ser clusterizado ($\{x_1, x_2, ..., x_n\}$), K: total de clusters;
cria_centroides(X,K)-> C // C: conjunto de centróides $\{c_1, c_2, ..., c_n\}$;
i=0;
while *i< k lenght* **do**
  u_k -> c_k // atribui cada centróide para um cluster, U: estrutura de dados com os elementos dos clusters $\{u_1, u_2, ..., u_n\}$;
  i = i+1;
  while *erro quadrático médio* > *x* **do**
    cluster[k] = {};
    while *k=1 to n* **do**
      atribui_elemento_cluster(X,cluster)
    end
  end
end
Result: U

Algoritmo 2: K-Means

Por fim, um último conceito abordado nessa seção é o de séries temporais e das funções de autocorrelação e autocorrelação parcial. Esses conceitos são importantes para entender as análises das séries temporais que foram obtidas através da mineração de NFe.

2.5 Séries Temporais

Séries temporais são definidas como um conjunto de observações organizadas em uma ordem cronológica, onde o tempo geralmente é uma variável discreta [51]. Uma das características mais importantes desse tipo de dados, que possibilita sua modelagem e previsão, é a dependência entre as observações vizinhas [52]. Essa característica permite que seja possível elaborar modelos para prever, por exemplo, qual seria o próximo valor assumido por determinada série temporal. Um dos principais objetivos no estudo de uma série é a sua previsão. Para isso devem ser consideradas muitas variáveis intrínsecas a série e também fenômenos do ambiente externo que possam estar relacionados a ela. Considere Y_i uma série temporal onde: $Y_i = \{y_i \in R | i = 1,2,3,...,N\}$, y_i é uma observação ordenada cronologicamente dada a ocorrência dos eventos e N é o número de observações. Para exemplificar séries temporais é apresentado a série *sunspots*. Essa série representa as manchas solares na superfície do sol. A Figura 2.10 apresenta essa série medida em anos onde o eixo horizontal representa os anos das amostras e o eixo vertical representa a quantidade de manchas solares observadas [53].

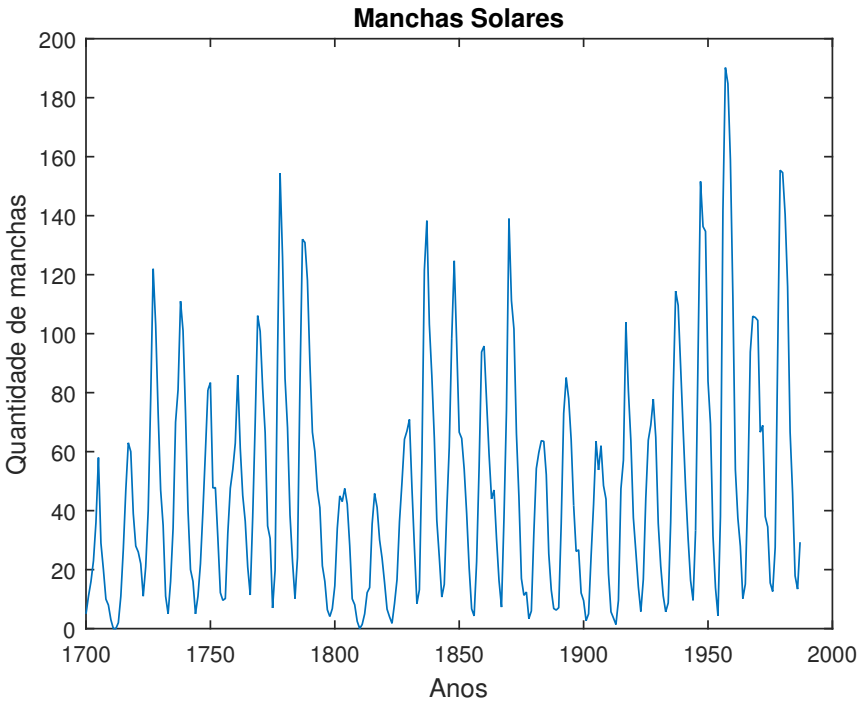


Figura 2.10 Quantidade de manchas solares anuais de 1700 a 1986

Algumas propriedades de séries temporais são muito importantes especialmente no que se diz respeito à sua predição. A sazonalidade de uma série é um certo padrão que tende a se repetir a cada determinado período de tempo [54]. Na série das manchas solares é possível observar que existe um comportamento cíclico, com um crescimento precedido de uma queda no número de manchas solares durante os anos. Outra característica de uma série temporal é a tendência. Uma série pode exibir uma tendência de crescimento ou decrescimento de seus

valores. Observando a série temporal das manchas solares é possível inferir que a quantidade de manchas solares apresenta uma tendência crescente ao longo dos anos. Uma outra característica muito importante de uma série temporal é a sua estacionariedade. Esse é um atributo elementar no que se diz respeito a previsão da série. Um processo é dito estacionário se a sua média e variância não variam ao longo do tempo. Logo, sendo y_t um ponto da série, μ a média, σ a variância e γ a covariância do processo, ele é estacionário se: $E[y_t] = \mu_t = \mu$; média constante para qualquer tempo t, sua variância é constante representada por $E[(y_t - \mu)^2] = E[(y_{t-s} - \mu)^2] = \sigma_y^2$ e a sua covariância também é constante descrita por $E[y_t, y_s] = [(y_t - \mu_t)(y_s - \mu_s)] = \gamma_{(|x-s|)}$ [51]. Para uma série que apresente tendências ou sazonalidades não há estacionariedade porque essas sazonalidades ou tendências vão afetar o valor da média da série temporal em diferentes períodos [55].

2.5.1 Funções das Séries Temporais

Uma função muito importante para estudar o comportamento de uma série temporal é a função de autocorrelação. A função de auto correlação utiliza os coeficientes de autocovariância para gerar uma função que revela a correlação entre quaisquer valores de uma série [56]. Essa função dá a ideia de como as observações são regidas ao longo do tempo e evidenciam a lei que governa o processo. Essa função é definida pela Equação 2.5.

$$\rho(\tau) = \frac{\gamma_\tau}{\gamma_0} = \frac{E[(y_t - \mu)(y_{t+\tau} - \mu)]}{E[(y_t - \mu)^2]} \tag{2.5}$$

A Figura 2.11 apresenta a função de autocorrelação para a série temporal das manchas solares obtida através do MatLab [57]. Cada ponto demostra o percentual de correlação que existe entre o ponto e o Lag (janela temporal) correspondente.

Outra função utilizada para estudar o comportamento de uma série é a função de autocorrelação parcial. A função de autocorrelação parcial, entre quaisquer dois pontos da série, é a correlação que permanece se o impacto de todas os outros pontos da série fosse eliminado. Da mesma forma que a função de autocorrelação a função de autocorrelação parcial também evidencia as leis que governam o modelo com a diferença que essas evidências são observadas par a para. Sendo ϕ_{kj} o *j-ésimo* coeficiente em uma função auto regressiva de ordem k e ϕ_{kk} o último elemento, então de acordo com [56] essa função é representada pela Equação 2.6

$$\rho_j = \phi_{k1}\rho_{j-1} + ... + \phi_{k(k-1)}\rho_{j-k+1} + \phi_{kk}\rho_{j-k}; com \quad j = 1, 2, 3, ..., k \tag{2.6}$$

As equações de Yule-Walker levam a expressão encontradas na Equação 2.6, que depois de resolvidas para um caso específico de um modelo AR(2)[5] por exemplo resulta na Equação 2.7.

$$\phi 22 = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \tag{2.7}$$

[5]AR(2) seguindo a metodologia de Box e Jenkins [23] representa um modelo auto regressivo de parâmetro 2 para a previsão da série temporal.

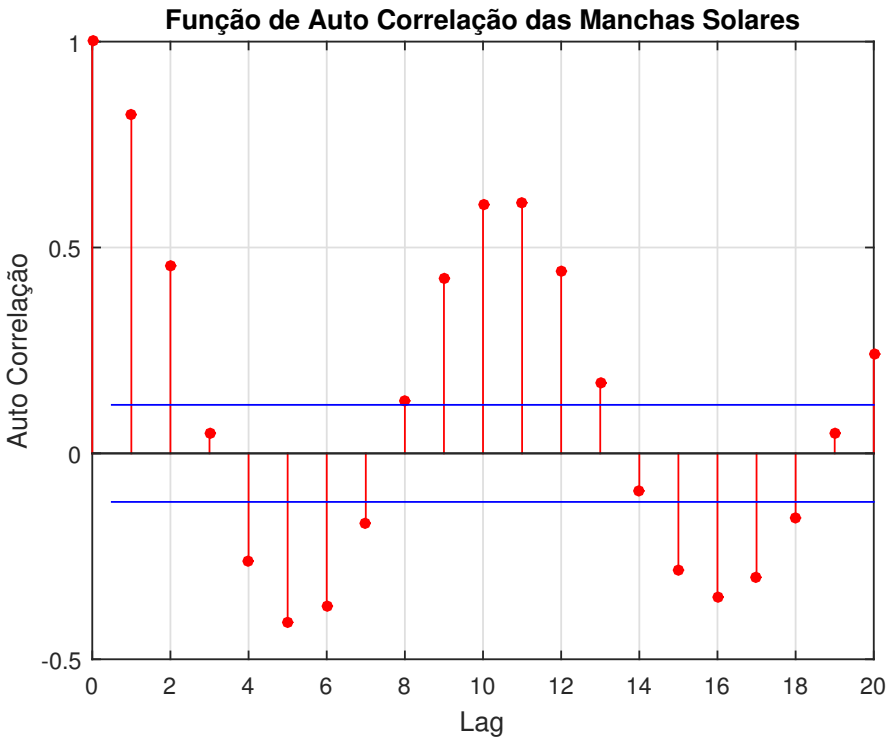


Figura 2.11 Função de autocorrelação da série das manchas solares

A Figura 2.12 apresenta a função de autocorrelação parcial da série temporal das manchas solares obtida através do MatLab [57].

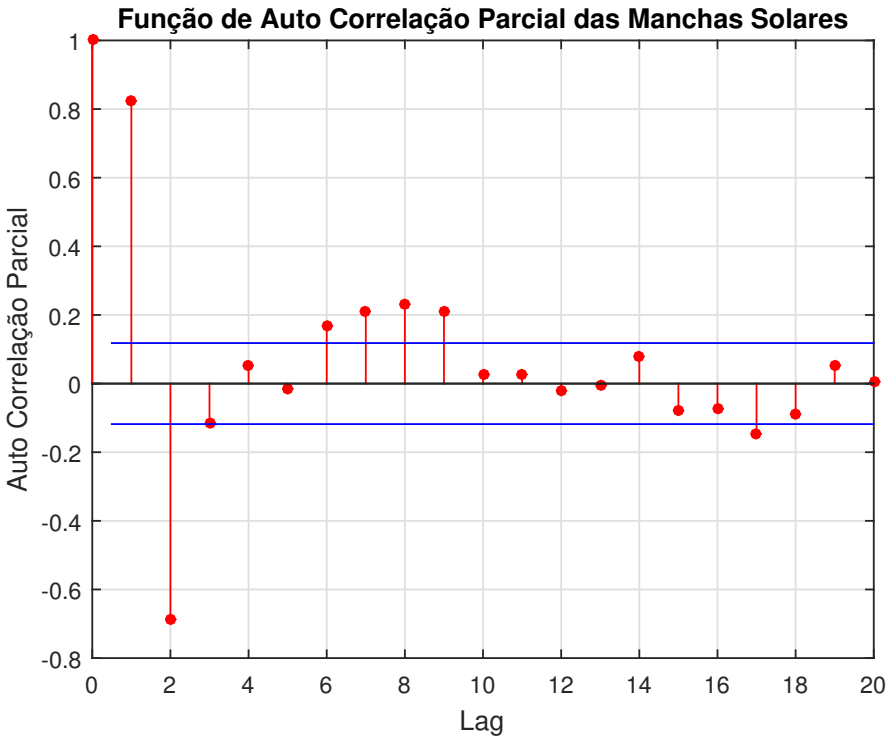


Figura 2.12 Função de autocorrelação parcial da série das manchas solares

CAPÍTULO 3

Trabalhos Relacionados

O escopo desse trabalho é realizar uma análise de mercado através de informações mineradas de notas fiscais eletrônicas. Com as informações contidas nas notas a presente dissertação tem como objetivo identificar clientes que estejam com um deslocamento acima da média para comprar determinados produtos, definindo assim uma demanda reprimida para esse produto. A definição de demanda reprimida nesse trabalho passa por diversas etapas, desde uma análise georreferenciada dos clientes de uma empresa até o agrupamento de clientes que estão fora do raio da demanda reprimida.

O tema análise de dados a partir de notas fiscais eletrônicas, apesar de bastante abrangente ainda é muito insipiente por se tratar de um processo que começou a ser implantado há pouco mais de uma década [1]. Com isso, poucos trabalhos no âmbito acadêmico tiveram semelhança com o objetivo proposto nessa dissertação. Um dos trabalhos semelhantes foi apresentado por Daozheng [58], onde algumas empresas cederam os seus dados comerciais e os autores criaram um mecanismo que analisava as compras no varejo através do fluxo de clientes. Fora do âmbito acadêmico, a operação serenata de amor[1] utiliza as notas fiscais eletrônicas para acompanhamento dos gastos dos parlamentares.

Para a seguinte revisão de literatura foi realizado uma pesquisa com a *string* de busca "demanda reprimida", que é um dos principais temas abordados neste trabalho, no engenho de busca Google acadêmico. Foram obtidos 10500 resultados que contém a palavra no corpo do trabalho, mas não abordam diretamente o tema. Além disso, a maioria dos trabalhos com essa *string* de busca resultou por trabalhos na área da saúde ressaltando a falta de médicos, hospitais e medicamentos. Por causa disso, a estratégia foi comparar alguns trabalhos relacionados às diversas etapas da dissertação buscando trabalhos nos temas de: reconhecimento de caracteres em CAPTCHA, clusterização de vendas, previsão de preços e análise de mercado. Os trabalhos foram selecionados pela aproximação com o trabalho aqui desenvolvido e cada um dos trabalhos são descritos nas próximas seções. Além da descrição dos trabalhos esse capítulo também aponta uma tabela avaliando os pontos positivos e limitações dos trabalhos relacionados que foram selecionados com relação a avaliação das suas metodologias e resultados.

3.1 Uma Simples Abordagem Genérica para Decifrar Textos Baseados em CAPTCHA

O trabalho de Gao *et al.* [59] apresenta um sistema genérico para decifrar textos baseados em CAPTCHAs por meio de algoritmos de tratamento de imagens e reconhecimento de pa-

[1]Informações do projeto disponível em: https://serenatadeamor.org/ Acessado em: 21-09-2017.

drões. Os resultados da metodologia proposta nesse trabalho para aplicações em CAPTCHAs de empresas como: Google, Microsoft e Amazon tiveram uma eficiência de acerto entre 5% e 77%. Os CAPTCHAs foram resolvidos em aproximadamente 15 segundos em um computador regular (2.8GHz Intel Core i3 CPU e 2 GB RAM).

O algoritmo proposto funciona utilizando filtros Log-Gabor. Filtros Gabor são algoritmos de processamento de sinal, eles informam a localização no espaço e a frequência simultaneamente. Um filtro Gabor é definido por um produto do seu núcleo Gaussiano e a sua sinuosidade complexa [60]. O filtro Log-Gabor tem como função de transferência uma Gaussiana em frequência logarítmica. Essa alteração melhora a representação dos filtros independente da largura da banda analisada [61]. A estratégia utilizada pelos autores nesse trabalho é utilizar filtro Log-Gabor para preprocessar a imagem e identificar os caracteres e o algoritmo k-NN para reconhecer os padrões desses caracteres e classificá-los.

A técnica proposta usa o filtro Log-Gabor para extrair as informações dos caracteres em diversos sentidos, assim apenas as letras e números que compõe o CAPTCHA são obtidas e os ruídos são eliminados. Esses caracteres são postos em tons de cinza. E a partição é realizada através de um grafo que é criado a partir da frequência de aparição de pixel nas imagens. Com o grafo da distribuição da imagem é possível distinguir os caracteres individualmente. Com os caracteres separados é possível classificá-los utilizando k-NN. O percentual de acertos juntamente com os tempos médios para quebrar o CAPTCHA estão presentes na Tabela 3.1. Em conclusão, Bursztein *et al.* [12] consideram que um esquema de CAPTCHA foi quebrado quando alguma técnica consegue um percentual de acerto acima de 1%. Sendo assim, devido aos resultados apresentados, os desfechos do trabalho são consideráveis.

Tabela 3.1 Resultados das tentativas de quebra de CAPTCHA

Esquema	**Taxa de acerto**	**Velocidade(s)**
Yahoo!	5.0%	28,56
Wikipedia	23.8%	3,74
Amazon	25.8%	13,8
Ebay	58.8%	5,8

3.2 Clusterizando Dados do Mercado de Ações Indiano

O trabalho de Nanda *et al.* [62] utiliza o algoritmo de clusterização *k-means* para gerar portfólios de ações baseados em *clusters*. Os autores utilizam também algumas outras técnicas de clusterização como *self organizing maps* (SOM) e Fuzzy C-means. Eles demonstram que através das análises dos resultados o algoritmo *k-means* conseguiu formar os *clusters* mais compactos e assim obter melhores resultados.

O modelo de Marlowitiz para portfólio de gerenciamento de ações [63] afirma que um portfólio eficiente demonstra o retorno de uma ação e o retorno médio entre o risco dessa ação. O portfólio deve conter os riscos das ações ponderados com os seus possíveis retornos.

A proposta desse trabalho visa minimizar os problemas de combinação entre os riscos e o retorno das ações através do seu agrupamento em *clusters* gerando assim um bom portfólio

para a melhor análise dos investidores. Os dados utilizados no trabalho foram da *Bombay Stock Exchange* da Índia entre os anos de 2007 e 2008. A ideia da metodologia é coletar os dados das ações e depois as clusterizar pelas características de ganho e retorno. Cada um desses *clusters* geram novos grupos de ações e então um portfólio com esses grupos de ações é gerado. Ações de maior risco e retorno são agrupadas em um *cluster*, por outro lado ações de menor risco e retorno são agrupadas em um *cluster* diferente. Um número de *k clusters* entre 2 e 12 foram testados e os indicativos apontam que a associação gerou melhor resultado para clusterizações com valores de 5 ou 6 *clusters*.

3.3 Análise de modelos de dados não relacionais e multidimensionais

O trabalho de Lira [64] é uma análise comparativa de dois modelos de dados (multidimensional e não relacional) no armazenamento e consulta no contexto de grandes massas de dados. Do mesmo modo que a presente dissertação, o trabalho de Lira utiliza arquivos em xml de NFe para realizar os dois tipos de modelagem de dados. Primeiramente, ela constrói um modelo relacional para conter as informações presentes nos xml. Após popular essa base relacional a autora utiliza duas ferramentas para criar os modelos de dados. Para o modelo multidimensional a autora usa uma abordagem com *data warehouse* (DW) através da ferramenta Pentaho [2]. A outra abordagem é a não relacional com o uso da ferramenta Hbase [65] que é um banco de dados não relacional orientado a colunas.

Nos resultados do seu trabalho, Lira demonstra um poco da dificuldade em se trabalhar com grandes massas de dados. Ela afirma que a limitação de recursos computacionais é um problema crucial na modelagem e tem que ser levado em consideração na escolha do modelo de dados adotado para projetos que trabalhem com grandes volumes de dados. No trabalho a autora mostra execuções de preenchimento de tabelas que duraram mais de 126 horas e mesmo assim não conseguiram ser concluídas com êxito. Da mesma forma a autora cita problemas para conseguir instalar corretamente o ambiente para utilização do modelo não relacional. Por fim, a autora conclui que o modelo relacional tem resultados superiores aos modelos não relacionais para o estudo de caso avaliado.

3.4 Método de Pesquisa de Mercado e Sistema para Coleta de Dados de Mercado

Daozheng *et. al* [58] desenvolveram em 1994 um mecanismo de análise de mercado através de cupons fiscais de empresas parceiras. Eles patentearam sua técnica de gerar uma cópia de cada cupom fiscal emitido pela empresa e depois os analisam em um sistema que é alimentado pelas informações desses cupons fiscais. O sistema então cruza as informações observadas dos cupons fiscais e traça um perfil para o consumidor dessa loja parceira para que ela consiga melhor atender as necessidades e demandas do seu cliente.

[2]Ferramenta para modelagem, criação e consulta em DW. Disponível em http://www.pentaho.com/ Acessado em: 27-02-2018.

O objetivo da patente é coletar de forma automatizada dados de mercado que incluem dados de venda junto com as transações de venda. Esses dados assim são inseridos em um sistema para que cada loja obtenha o perfil do seu cliente como: produtos mais vendidos, fluxo de vendas em diferentes horários e sazonalidades de determinados produtos. Aparelhos de vídeo monitoramento também foram instalados nas lojas parceiras para auxiliar os relatórios do fluxo de pessoal dentro dessas lojas.

3.5 Previsão do Preço do Milho através de Séries Temporais

O trabalho de Tibulo e Carli [66] utiliza modelos estatísticos de previsão para predição do valor do milho pelas séries históricas da varição de preços no Rio Grande do Sul. O milho é umas das culturas mais importantes para o estado do Rio Grande do Sul [67]. O milho produzido no Brasil é exportado e consumido mundialmente. Essa cultura é de alta produtividade, logo está sujeita a uma variação de preços por diversos fatores. Devido à variação de preço os comerciantes têm fechado contratos de compra da produção do milho, muitas vezes prejudicando o agricultor. O trabalho aponta uma forma de prever o preço do milho através de modelos estatísticos de séries temporais.

As observações realizadas no trabalho foram o preço médio mensal do milho entre os anos de 2004 e 2014 gerando uma média de 120 observações. Dessas observações as 5 últimas foram utilizadas para checar o modelo de previsão através das métricas MAD (*mean absolute deviation*)[3] e SSE (*sum of square for forecast error*)[4]. O modelo utilizado para realizar as previsões foi o modelo Autorregressivo Integrado e de Médias Móveis (ARIMA). Primeiramente, os autores avaliam se a série apresenta algum comportamento não estacionário e se for o caso aplicam um número de diferenciações até que essa série se torne estacionária [68]. Depois estimam os modelos através de funções de autocorrelação e autocorrelação parcial. E por fim, realizam uma validação desses modelos através dos critérios *Akaike Information Criteria* AIC e Bayesian Information Criteria BIC [69].

As métricas MAD e SSE apresentaram uma boa avaliação das predições do modelo estimado. O trabalho também demonstrou que é viável a utilização de modelos ARIMA ajustados para a previsão de preços de *commodities* através de suas séries históricas.

3.6 Operação Serenata de Amor

Um pouco fora do âmbito acadêmico, mas com grande aderência a dissertação existe a operação "Serenata de amor"[70]. A operação serenata de amor foi idealizada pelo cientista de dados Irio Musskopf com o intuito de monitorar os gastos públicos no congresso federal brasileiro. Mais especificamente, o *software* analisa os dados dos valores requisitados para reembolso da cota para exercício da atividade parlamentar que é uma cota destinada para custear gastos com

[3] O Desvio Médio Absoluto é o somatório do produto dos desvios (em módulo, ou valor absoluto) de cada valor observado em relação à média, pelas respectivas frequências, dividido pela frequência total.

[4] A Soma Quadrática dos Erros é uma métrica que expressa o somatório do valor absoluto dos erros das previsões.

viagens, hospedagens e alimentação dos parlamentares no exercício da profissão.

A operação serenata de amor utiliza algoritmos de inteligência artificial para monitorar as notas fiscais utilizadas pelos deputados com fins de fiscalizar os seus gastos. O projeto teve grande suporte da sociedade fornecendo o financiamento necessário para o seu desenvolvimento, sendo um dos projetos com maior número de colaboradores na plataforma de *crowfunding*[5], de acordo com Araújo [71]. Dentre algumas irregularidades identificadas pela operação estão:

- Uma refeição no valor de R$ 6.205,00 ;
- 30 tanques de gasolina completos em um mês, com custo em torno de R$ 6.000,00 ;
- 13 refeições realizadas no mesmo dia;
- Reembolso de bebida alcoólica na cidade de Las Vegas, EUA.

3.7 Avaliação dos Trabalhos Relacionados

A Tabela 3.2 apresenta um resumo dos trabalhos relacionados destacando seus pontos positivos e limitações.

[5] *Crowfunding* consiste na obtenção de capital para iniciativas de interesse coletivo através da agregação de múltiplas fontes de financiamento

Tabela 3.2 Resumo da avaliação dos trabalhos relacionados

Trabalho	Área	Pontos Positivos	Limitações
Uma simples abordagem genérica para decifrar textos baseados em CAPTCHA	Reconhecimento de padrões	Metodologia bem definida	Alta taxa de erro em algumas classificações
Clusterizando dados do mercado de ações indiano	Clusterização	Comparação de métodos de clusterização	Falta de critérios na comparação dos algoritmos
Análise de modelos de dados não relacionais e multimensionais	Mineração de dados	Apresentação dos reveses de se trabalhar com grande massas de dados	Falta de comparação dos modelos em ambientes equivalentes
Método de pesquisa de mercado e sistema para coleta de dados de mercado e análise de mercado	Análise de mercado	Método de análise de dados de vendas a varejo e análise de mercado	Limitação de análise as lojas parceiras do trabalho
Previsão do preço do milho, através de séries temporais	Previsão de séries temporais	Utilização de uma metodologia bem estruturada para previsão de séries temporais	Necessidade de melhor avaliação dos coeficientes AIC e BIC para encontrar o modelo que melhor se ajuste a série
Operação serenata de amor	Análise de notas fiscais	Utilização de notas fiscais para avaliação de gastos públicos	Falta de difusão acadêmica para os resultados desse trabalho

CAPÍTULO 4

Metodologia

A metodologia do trabalho se divide em duas partes: a metodologia científica e a metodologia do desenvolvimento dos resultados desse trabalho. Também nesse capítulo são apresentadas as tecnologias adotadas nesse trabalho, as funcionalidades do sistema desenvolvido e uma tabela comparativa com a dissertação aqui proposta e os trabalhos relacionados do Capítulo 3.

4.1 Metodologia Científica

A metodologia científica do presente trabalho envolve uma pesquisa: experimental, aplicada, descritiva, quantitativa e que tem método de coleta de dados através de documentos.

De acordo com Gil [72] uma pesquisa experimental se caracteriza por definir um objeto de estudo específico, selecionar as variáveis que podem influenciar esse objeto de estudo e definir os efeitos que as variáveis causam no objeto de estudo. O presente trabalho se caracteriza como uma pesquisa experimental pois se baseia em um objeto de estudo que são as massas de dados das notas fiscais e as variáveis que são as observações que podem ser retiradas dessa base de dados com técnicas de mineração.

De acordo com Gil [72] uma pesquisa aplicada tem como objetivo principal gerar conhecimento para aplicação prática, dirigida a solução de problemas específicos. A natureza da pesquisa desse trabalho é aplicada por conta das características de desenvolvimento de um sistema para analisar dinamicamente comportamento de mercados através de NFe. Isso é ressaltado nos objetivos da pesquisa que é, dentre outros mais, a aplicação de técnicas e ferramentas para obtenção dos resultados desejados.

Triviños [73] afirma que os estudos descritivos tem como objetivo descrever fatos e fenômenos de determinada realidade além de uma descrição detalhada desses fatos e fenômenos. Com base nisso, em relação aos objetivos dessa pesquisa, se caracteriza como uma pesquisa descritiva, pois o trabalho propõe descrever as características da experiência da manipulação de grandes massas de dados a fim de descobrir comportamentos de mercado.

A abordagem da pesquisa é quantitativa, uma vez que sua natureza envolve análise de dados numéricos, estatística e uma base matemática mais complexa. De acordo com Terrance [74] a pesquisa quantitativa caracteriza-se por: obedecer um plano pré estabelecido, examinar relações entre variáveis, confirmar hipóteses, dentre outras. Essas afirmações reforçam que o objetivo da pesquisa proposta nesse trabalho refere-se a uma pesquisa de natureza quantitativa.

A estratégia adotada para a coleta de dados da pesquisa é a utilização de documentos. Cada xml é um documento digital que contém as informações que serão utilizadas para se obter os resultados finais desse trabalho. Portanto, utilizar esses documentos é a técnica adotada para

coletar os dados da pesquisa. Gil [72] afirma que essas fontes de dados são muito ricas para a pesquisa.

4.2 Mineração de Dados de Mercado

Essa seção descreve os procedimentos metodológicos utilizados para obter as notas fiscais, analisar os dados, calcular a demanda reprimida, clusterizar os pontos de vendas e analisar as séries dos preços dos produtos.

A extração das notas fiscais eletrônicas se dá pela seleção de um CNPJ de uma das empresas cadastradas nos diretórios criados a partir das primeiras notas fiscais baixadas manualmente. Após selecionada a empresa, a chave de acesso é então gerada. Uma vez com a chave, o sistema vai obter a imagem de CAPTCHA do site da Secretaria da Fazenda do município e então extrair o texto da imagem. Com o texto da imagem do CAPTCHA e a chave é iniciado uma sessão com o servidor para se obter o xml da nota fiscal. Esse xml é então adicionado à base de dados e está pronto para análise. Existem diversas informações contidas no arquivo xml da nota fiscal, por isso se faz necessária uma limpeza dos dados. Após a limpeza dos dados, as informações relevantes para esse trabalho – a exemplo de itens vendidos, localização dos vendedores e compradores, preços e impostos – são carregadas no banco de dados relacional. Com essas informações é possível se calcular várias métricas inclusive a demanda reprimida que pode existir por algum produto. Com o mapa da demanda reprimida é possível definir parâmetros para algoritmos de clusterização e indicar possíveis pontos de concentração de vendas fora do raio de alcance da loja. Além dessa análise, com as informações é possível também avaliar a partir das séries da variação de preço de determinado produto qual será o seu comportamento no futuro. Uma visão geral do procedimento metodológico envolvido nesse trabalho está na Figura 4.1.

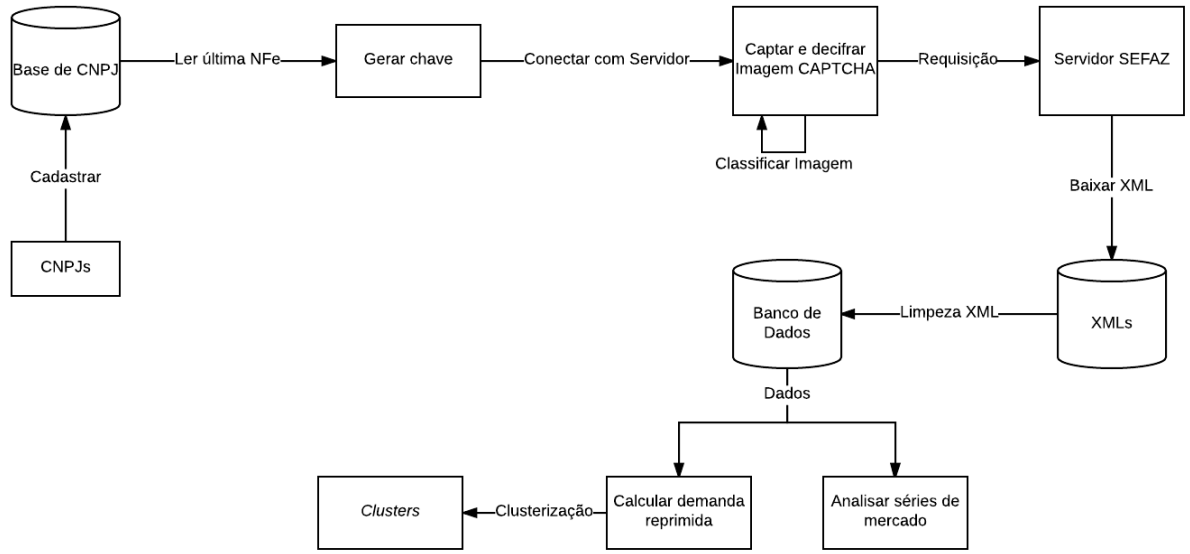


Figura 4.1 Visão geral do funcionamento do sistema de baixar NFe

Seguindo o fluxograma da Figura 4.1, inicialmente é cadastrado um banco de CNPJs coletando manualmente as notas fiscais que se têm em posse e separando os CNPJs em diretórios. Após a criação dos diretórios é baixado manualmente a primeira nota fiscal daquele CNPJ. O sistema vai então minerar as notas a partir dessa primeira nota fiscal adicionada ao sistema. Com a chave da primeira nota o sistema gera uma próxima chave e se conecta ao servidor da SEFAZ captura e decifra a imagem do CAPTCHA, e envia uma requisição da NFe com as informações do código do CAPTCHA e a chave da nota. O servidor da SEFAZ retorna o xml que o sistema baixa e o coloca no diretório criado para os CNPJ. Após baixar os xml, um outro módulo do sistema faz uma limpeza no xml e adiciona as informações no banco de dados. Com esses dados, pode-se calcular a demanda reprimida por produtos e também clusterizar os *outliers* do processo. Além do cálculo da demanda reprimida é possível também analisar séries de mercado.

4.2.1 Extração de Notas Fiscais Eletrônicas

O primeiro esforço do projeto foi a coleta de várias notas fiscais dos mais diversos segmentos. Cada nota fiscal está relacionada a uma empresa e essa empresa possui um CNPJ. Ao todo foram coletados notas de 120 empresas de diversos segmentos de mercado. Cada uma dessas empresas apresenta um classificação nacional de atividades econômicas (CNAE) específico. CNAE é um código de padronização das atividades econômicas e os critérios de enquadramento usados pelos mais diversos órgãos da administração tributária do Brasil [75]. Dentre as empresas estudadas existem uma diversidade de 57 CNAEs. A Tabela 4.1 apresenta a quantidade de empresas e CNAEs mais representativos da massa de dados avaliadas. As atividades relacionadas aos códigos foram obtidas através do site do IBGE[1].

Com o banco de CNPJs formado, cada CNPJ deve conter um arquivo xml de uma nota inicial que vai ser gerada a chave de acesso para então o sistema se comunicar com o servidor da Secretaria da Fazenda. Cada chave de acesso possui 44 dígitos, dentre esses 44 dígitos 9 são o número da nota fiscal (nNF). Uma vez que se tem uma nota fiscal é possível gerar uma provável chave na nota subsequente a que se tem na base. Além do número da nota fiscal existe um conjunto de 8 dígitos que é o código da nota (cNF). Muitas empresas utilizam o algoritmo MD5 [4] para gerar o código da nota fiscal utilizando como semente 35 primeiros dígitos da chave de acesso. A Figura 4.2 apresenta um exemplo de uma chave típica.

Contudo, uma boa parte dos emitentes de NFe não se preocupam em utilizar o algoritmo MD5 para geração do código, repetindo os 8 últimos dígitos do número da nota (nNF) no código (cNF) da chave. Com isso se tem maiores possibilidade de minerar notas fiscais das empresas. Assim para se criar uma chave de acesso com grandes chances de ser válida e tendo um exemplo de nota da empresa basta incrementar o número da nota (nNF), alterar o código (cNF) com o novo número gerado e por fim calcular novamente o dígito verificador (cDV) que é basedao no cálculo do módulo 11. Essa estratégia abrange um grande número de empresa e consegue-se obeter um grande número de notas fiscais.

Apesar dessa estratégia ter bons resultados, existem algumas situações em que essa lógica

[1]Informações relativas ao CNAE disponível em: https://concla.ibge.gov.br/busca-online-cnae.html. Acesssado em: 24-11-2017.

Tabela 4.1 CNAE de algumas empresas da base de dados

CNAE	№CNPJ	Atividade
4530703	8	Comércio varejista de peças para automóveis (alarme, radiador, capotas, som)
4744001	7	Comércio atacadista de equipamentos industriais (rolamentos para máquinas, correias industriais, furadeiras de bancada)
4713001	6	Loja de departamentos
4751201	5	Comércio varejista de equipamentos para informática (impressoras, computadores, monitores, teclados)
4754701	4	Comércio varejista de assentos (mesas e cadeiras, sofás, poltronas)
4511101	4	Comércio varejista e atacadista de automóveis, reboques e/ou caminhonetes
4753900	4	Comércio varejista de aparelhos eletrônicos (fogão, geladeira, refrigerador, máquina de costura, câmeras, aparelhos de som)
4741500	4	Comércio varejista de fluídos de construção (loja de tintas, solventes, pincéis, massa para pintura)

não funciona. Uma das situações acontece quando a nota subsequente da que se tem na base foi cancelada. A estratégia que se pode utilizar é incrementar em até duas vezes a nota fiscal para se confirmar a possibilidade de até as duas próximas notas terem sido canceladas por algum motivo. A outra possibilidade é a de que a data esteja errada, uma vez que a nota subsequente pode estar presente no mês ou até mesmo no ano posterior ao da última nota registrada na base. Caso alterando a data e a nota continue inválida é necessário trocar o CNPJ que se esteja obtendo a NFe pois provavelmente a nota da base foi a última nota válida emitida pela empresa. Então, a estratégia utilizada nesse trabalho para checar se uma chave é valida ou não consiste em:

(i) incrementar o número da nota, em até duas vezes (notas canceladas);

(ii) incrementar a data;

(iii) minerar outro CNPJ (provavelmente última nota emitida foi baixada).

Uma chave válida de acesso para esse projeto tipicamente tem a seguinte padronização:

- Código do estado (cUF) - Para esse trabalho o código utilizado foi o 26 que representa o estado de Pernambuco;
- Data (AAMM) - O mês e ano da nota fiscal. Exemplo 1701, que significa janeiro do ano de 2017;
- CNPJ - CNPJ do emitente exemplo 00118694000166;

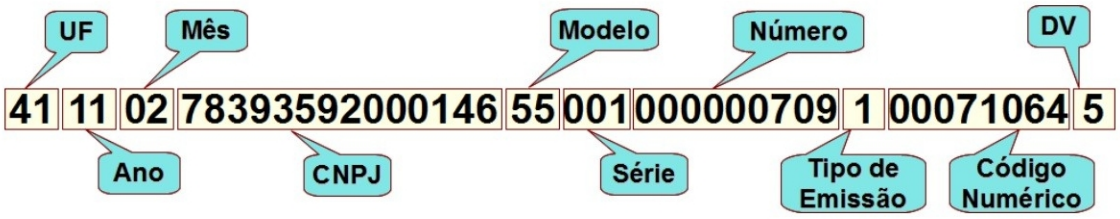


Figura 4.2 Exemplo de uma chave típica de nota fiscal para acesso ao documento no servidor da Secretaria da Fazenda

- Modelo do documento fiscal (Mod) - Utiliza-se o modelo 55 que significa modelo em substituição do modelo de papel para o eletrônico;
- Série - A serialização do documento fiscal. Usualmente se utiliza a primeira série de notas com o valor 001;
- Número do documento Fiscal (nNF) - O número da nota fiscal gerada. Que pode ser algo como 000000001;
- Tipo da emissão (TpEmis) - Tipo de emissão do modelo. O algarismo 1 representa modelos emitidos de forma eletrônica;
- Código da nota fiscal (cNF) - Código gerado pelo algoritmo. Esse código em muitos casos é uma repetição dos 8 últimos dígitos do número da nota. Um exemplo para esse número pode ser 00000001;
- Dígito verificador (cDV) - Um dígito entre 0 e 9 calculado a partir dos primeiros 43 números da nota fiscal.

Nesse sentido uma possível chave válida seguindo os itens citados seria 26170100118694000166550010000000011000000013. Com a chave de acesso se pode conectar o servidor da SEFAZ para se baixar as NFe através de requisições *http* nesse servidor. A seguir a Figura 4.3 apresenta um trecho do código desenvolvido para geração das chaves das NFe.

Nessa imagem a função "*generateKey*"tem como parâmetro uma variável do tipo booleano que representa a resposta do servidor se a chave foi ou não válida. Além desse parâmetro outras variáveis como, por exemplo, "*dataIncremented*"controlam o fluxo da execução para se gerar a chave da próxima NFe a ser baixada.

4.2.2 Classificação de Imagens

O mecanismo de segurança contra iteração automática dos portais da SEFAZ é um CAPTCHA baseado em texto. Para decifrar o texto contido na imagem é usado um algoritmo de

```
public String generateKey(Boolean InvalidNote) throws IOException{

    if(InvalidNote==false){// Simple download of notes
        this.dataIncremented = false;
        this.canceled=0;
        updateNote();
    }

    else if(InvalidNote==true && dataIncremented==false && this.canceled==0){//Note might be canceled
        this.dataIncremented = false;
        this.canceled +=1;
        updateNote();
        key.setNumberNfe(key.numberNfe);//add other number of note because the last number might be canceled
        key.setCode();//in order do do two notes add other else if and put the function setNumberNfe twice
        key.setModel11();
    }

    else if(InvalidNote==true && dataIncremented==false && this.canceled==1){//Note might be canceled
        this.dataIncremented = false;
        this.canceled +=1;
        updateNote();
        key.setNumberNfe(key.numberNfe);//add other number of note because the last number might be canceled
        key.setCode();//in order do do two notes add other else if and put the function setNumberNfe twice
        key.setModel11();
        System.out.println("Duas canceladas");
    }

    else if(InvalidNote==true && dataIncremented==false && this.canceled==2){//Note might be canceled
        this.dataIncremented = false;
        this.canceled +=1;
        updateNote();
        key.setNumberNfe(key.numberNfe);//add other number of note because the last number might be canceled
        key.setCode();//in order do do two notes add other else if and put the function setNumberNfe twice
        key.setModel11();
        System.out.println("Tres canceladas");
```

Figura 4.3 Trecho do código para gerar as chaves das NFe

classificação de imagens denominado *k*-NN. Para demonstrar como funciona a metodologia de classificação das imagens é utilizado um exemplo.

Inicialmente é necessário baixar muitas imagens de CAPTCHA para se definir o alfabeto, essas imagens descarregadas estão da maneira em que o *website*[2] mostra para o usuário. A Figura 4.4 apresenta a interface do *webservice* da SEFAZ PE para consulta completa de notas fiscais eletrônicas homologadas.

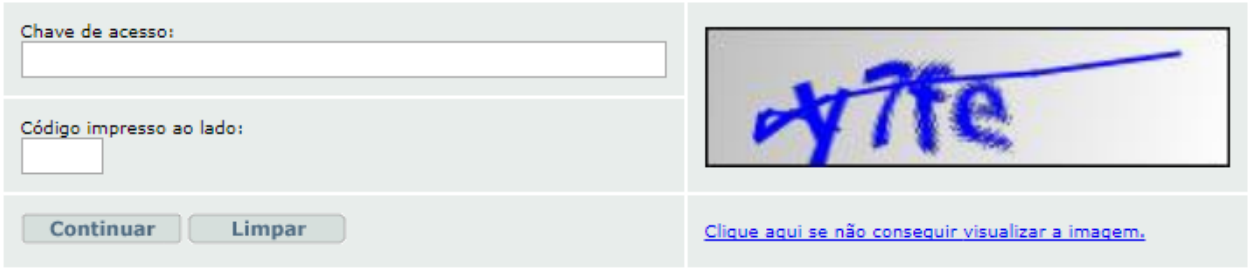


Figura 4.4 Inferface de consulta de NFe pelo portal da SEFAZ PE

Para se descobrir a quantidade de caracteres que compõe o CAPTCHA bem como os elementos que fazem parte do seu alfabeto foram descarregadas cerca de 2000 imagens de CAPTCHAs do *webservice* da SEFAZ a exemplo das imagem presente na Figura 4.4. Através da análise das imagens descarregadas, identificou-se que o CAPTCHA analisado sempre possui 4 caracteres pertencentes a um alfabeto de 19 caracteres, eles são: *2*, *3*, *4*, *5*, *6*, *7*, *8*, *b*, *c*, *d*, *e*, *f*, *g*, *m*, *n*, *p*, *w*, *x* e *y*. Para cada um desses caracteres é criado um repositório e adicionado 100 diferente exemplos de imagens do carácter específico.

O processamento da imagem é dividido em 4 etapas: baixar imagem, converter em tons de

[2]http://www.sefaz.pe.gov.br/nfe-web/consNfe?tp=C

cinza, remover as bordas em branco e dividir a imagem. A Figura 4.5 apresenta os resultados da etapa de processamento da imagem.

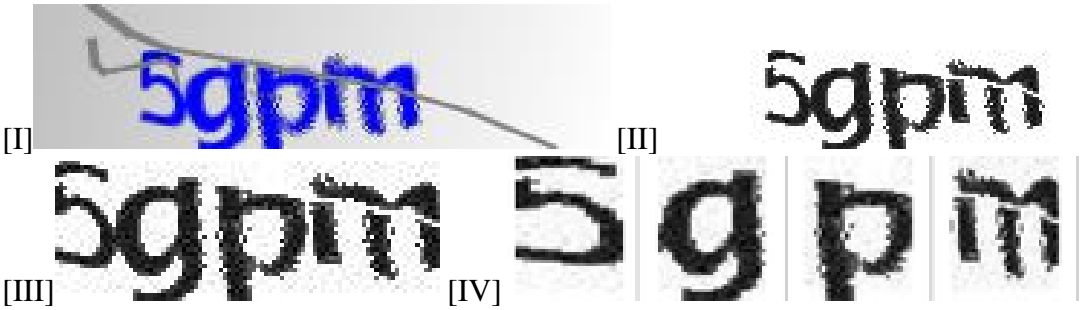


Figura 4.5 Fases do processamento da imagem

O primeiro processamento é realizado na imagem do modo em que está apresentado na Figura 4.5 (I), nesse sentido a figura é colocada em tons de cinza. Inicialmente o algoritmo converte a imagem para o formato *jpg* onde cada pixel da imagem tem um valor entre 0 e 255 para os tons de vermelho, verde e azul (RGB). Depois de convertida a imagem em *jpg* o algoritmo muda todos os pixels da imagem seguindo a regra: caso qualquer valor dos tons de vermelho, verde ou azul seja maior que 60 esse pixel é convertido em branco com RGB (255,255,255); Caso contrário esse pixel é convertido em preto com RGB (0,0,0). Essa primeira transformação pode ser notada na Figura 4.5 (II) onde é possível analisar que o fundo cinza e as linhas que dificultavam a visualização da imagem foram removidas.

Depois da imagem ser convertida em escalas binária (preto e branco), o algoritmo de processamento da imagem remove os espaços em branco das bordas da imagem para se avaliar apenas a parte da imagem que contém o texto. Para se retirar as bordas é realizada buscas horizontais e verticais procurando o primeiro pixel diferente de (255,255,255), então é definido os pontos de início e fim da imagem. A Figura 4.5 (III) apresenta o resultado desse processamento.

O último estágio é o de divisão da imagem. Essa fase é a mais complexa do processamento. Usualmente em CAPTCHAs baseados em texto, os caracteres não estão espaçados uniformemente, e algumas vezes eles estão sobrepondo outros caracteres. Essas técnicas são empregadas para dificultar um programa de computador descobrir os caracteres no CAPTCHA [12]. Então a estratégia é verificar a concentração de pixels brancos em cada linha vertical da imagem. Cada concentração dessa é armazenada em um vetor e assim é possível montar um histograma de quantidade de pontos brancos em cada linha vertical. Com isso, analisando os pontos máximos do histograma e a distância entre os pontos adjacentes através da heurística desenvolvida nesse trabalho e detalhada no próximo paragrafo, é possível inferir sobre os locais da divisão da imagem. No exemplo da Figura 4.5, a imagem do CAPTCHA foi dividida nos 4 elementos por três linhas. Os três maiores pontos no histograma indicam uma grande concentração de pontos brancos, o que demonstra uma grande possibilidade de ser um espaço em branco entre dois caracteres. Com esses três maiores pontos foi possível dividir a imagem presente na Figura 4.5 (III) em 4 caracteres apresentados na Figura 4.5 (IV). A distribuição dos pixeis brancos na Figura 4.5 (III) é:

vetor = [40, 34, 35, 32, 28, 31, 33, 32, 32, 33, 32, 32, 33, 34, 34, 25, 18, 18, 18,

*19, 30, 40, **42**, 42, 42, 42, 40, 36, 30, 31, 33, 34, 33, 33, 34, 33, 35, 34, 33, 31, 34, **38**, 38, 37, 34, 32, 29, 33, 33, 34, 33, 34, 34, 33, 34, 33, 32, 33, 32, 31, 34, 30, 31, 32, 31, 30, 32, 38, 41, **42**, 42, 42, 42, 42, 42, 42, 42, 41, 40, 37, 35, 34, 35, 36, 32, 32, 32, 31, 29, 29, 25, 26, 35, 37]*

Onde os três pontos sublinhados são as referências para dividir a imagem. O histograma do vetor pode ser visto na Figura 4.6, onde a linha tracejada indica o ponto de divisão da imagem, definindo assim cada um dos elementos do CAPTCHA.

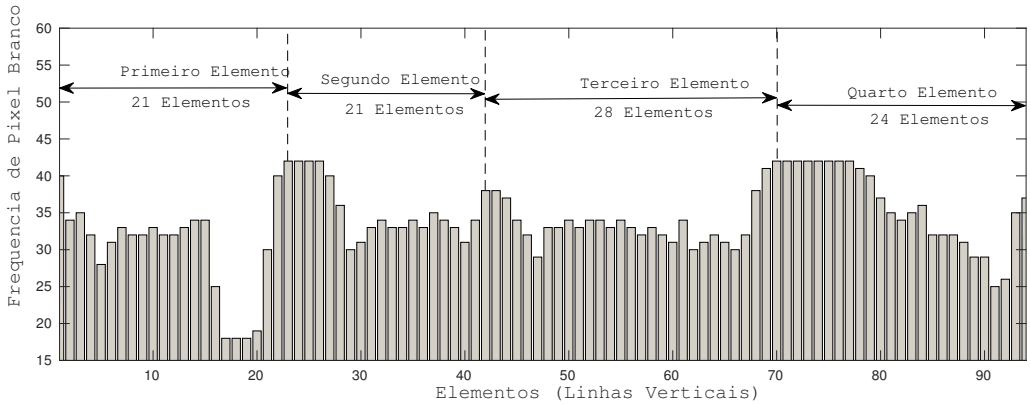


Figura 4.6 Histograma dos pontos brancos em cada linha vertical da imagem do CAPTCHA

Contudo, devido a variância do número de linhas brancas separando dois caracteres, essa estratégia algumas vezes não funciona. Então, a heurística adotada é de que após definido o primeiro ponto de divisão da imagem o segundo deve estar no mínimo de 10% da quantidade de linhas da imagem mais para frente do ponto anterior no vetor. Essa heurística evita que dois pontos de corte estejam muito próximos no vetor. Após as imagens serem separadas elas são processadas em um algoritmo de redimensionamento para que todas as 4 partes tenham as mesmas dimensões. A Figura 4.7 apresenta um trecho do código da função de fatiamento da imagem

Nesse trecho de código a função "*SliceFigure*"carrega a imagem e atribui as variáveis que são utilizadas para o seu fatiamento, como a frequência de pontos brancos nas linhas verticais na variável "*frequencyWhite*", depois de definido esse vetor o algoritmo procura os pontos de corte e os armazena no vetor *cutPoints*.

A outra etapa para preencher o CAPTCHA é classificar os caracteres usando *k*-NN. Esse algoritmo foi escolhido pela facilidade de implementação aliado com a rápida execução do algoritmo. Primeiramente, os diretórios dos caracteres são preenchidos com aproximadamente 100 imagens de cada um dos caracteres do alfabeto. Após isso, é definido o número de *k* vizinhos. Considerando que existem 19 classes diferentes no problema, o número *k* de vizinhos é definido seguindo as recomendações de Duda [14] sendo $k = \sqrt{(100 * 19)} \simeq 44$. Subsequentemente o carácter analisado é classificado com todos os elementos dos *clusters* através da distância Euclidiana, $p = 2$ na Equação 2.1. Os 44 elementos que tiverem a menor distância do carácter são considerados a sua vizinhança. Para a classificação do carácter *g* na Figura 4.5.(IV) usando os parâmetros da distância Euclidiana, a vizinhança é definida por: *vizinhança*(*g*) = [*4*,

```
public  void SliceFigure() throws IOException {
    try {
        File input = new File(this.path);
        BufferedImage image = ImageIO.read(input);
        int width = image.getWidth();
        int height = image.getHeight();
        int frequencyWhite[]= new int[width];
        int cutPoints[] = new int[3];
        int max;
        int index=0;

        for (int i = 0; i < width; i++) {//taking the occurrence of white points in a column
            int aux = 0;
            for (int j = 0; j < height; j++) {
                Color c = new Color(image.getRGB(i, j));
                if(c.getBlue() > 60 && c.getGreen() > 60 && c.getRed() > 60){
                    aux+=1;
                }
            }
            frequencyWhite[i]=aux;
        }

        max = 0;//taking the first cut point
        for (int i = 2; i < (int)(frequencyWhite.length/4+frequencyWhite.length * 0.1); i++) {
            if(frequencyWhite[i] > max){
                max = frequencyWhite[i];
                index=i;
            }
        }
        cutPoints[0]=index;
        frequencyWhite[index]=0;

        max = 0;//taking the second cut point
```

Figura 4.7 Trecho de código da fatiamento da imagem

4, *4*, *6*, *6*, *6*, *6*, *c*, *c*, *d*, *d*, *d*, *d*, *d*, *d*, *d*, *d*, *d*, *e*, *e*, *e*, *f*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *g*, *n*, *n*, *n*, *n*, *n*, *p*]

Esse vetor demonstra que a classe *g* é a dominante com cerca de 36% da ocorrência do vetor. Portanto, esse carácter foi corretamente classificado nessa execução.

O erro médio na classificação do algoritmo é de aproximadamente 6.8% conforme os métodos de amostragem estatística utilizados para se avaliar o erro da classificação [76]. Como o CAPTCHA analisado possui 4 caracteres, a probabilidade de classificar corretamente uma imagem é de $(1-(6.8/100))^4 * 100 = 75,31 \approx 75\%$ em média. Para cada 4 tentativas o algoritmo resolve corretamente 3 CAPTCHAs. A Tabela 4.2 apresenta a classificação dos caracteres para 10 tentativas para cada elemento. Essa tabela confusão [77] mostra os dados reais nas linhas e os resultados da classificação dos algoritmos nas colunas. Para avaliar o erro da classificação é possível agrupar o erro pelas classificação nas linhas que não correspondam o mesmo carácter nas colunas. Esta metodologia desenvolvida e aplicada aqui neste trabalho foi publicada na revista *International Journal of Computer Applications*, cuja a publicação pode ser vista na referência [76].

4.2.3 Análise Mercadológica dos Dados

As notas fiscais eletrônicas são descarregadas em formato xml. Essas NFe possuem diversos campos com informações relevantes, que podem ser muito úteis para análises mercadológicas. De acordo com o manual do contribuinte [6], alguns dos itens presente em uma nota fiscal são informações sobre: emitente, destinatários, produtos, valores, vendas e impostos. Com essas informações, é possível se traçar diversas análises como por exemplo "Qual a distância média percorrida entre clientes de um determinado produto e uma loja, filtrando os clientes por seus respectivos municípios?"ou "Quais foram os segmentos de mercado que mais emitiram NFe em um determinado estado, dentro de um período de tempo específico?". Para isso é necessário se desenvolver uma estrutura para armazenar os dados que são descarregados em

Tabela 4.2 Tabela confusão sobre o erro do algoritmo

	2	**3**	**4**	**5**	**6**	**7**	**8**	**b**	**c**	**d**	**e**	**f**	**g**	**m**	**n**	**p**	**w**	**x**	**y**
2	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	8	1	0	0	0	0	1	0	0	0	0	0
c	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	0	2	0	8	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
g	0	0	0	0	0	2	0	0	0	1	0	0	7	0	0	0	0	0	0
m	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
p	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0
w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0
x	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0
y	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	8

forma de xml. Para armazenar as notas, foi criado uma estrutura de banco de dados relacional em MySQL [78]. Esse modelo foi escolhido pois, apesar da grande quantidade de dados, sua estruturação facilita as pesquisas pelas junções de tabelas e tem um tempo de retorno aceitável. O modelo lógico da base de dados utilizada nesse trabalho está presente no Anexo A desse trabalho. A quantidade de notas fiscais utilizadas nesse trabalho foi de aproximadamente 5,5 milhões de notas fiscais baixadas. Para se escalar esse problema para mais empresas e outros estados é necessário a utilização de modelos multidimensionais ou não-relacionais. Lira [64] utiliza os mesmos tipos de dados e apresenta um comparativo entre modelos não-relacional e multidimensional, demonstrando que existem vantagens nas duas abordagens e demonstrando um pouco da dificuldade de se trabalhar com grandes volumes de dados. Um dos esforços desse trabalho foi a implementação de um modelo não relacional distribuído dos dados utilizando Hbase [65]. Essa implementação visou a escalabilidade das informações. Contudo, como a base de dados relacional apresentou resultados viáveis para as análises presentes nesse trabalho, foi decidido usar esse modelo.

Essa modelagem possui 17 tabelas, com vários relacionamentos entre elas. Algumas dessas tabelas possuem vários atributos como é o caso da tabela *nfe* que possui 33 colunas. Outras tabelas possuem muitos registros como é o caso da tabela imposto_item_nf com 14.509.220 registros. Essa quantidade de atributos e registros no banco aliado ao grande número de relacionamentos fazem a performance do banco diminuir [64]. Porém mesmo assim essa modelagem apresenta tempos consideravelmente bons para as análises desse trabalho. A seguir cada uma das tabelas presentes na figura do Anexo A é brevemente descrita e a quantidade de atributos delas é apresentada entre parênteses:

- cnae (11) - Apresenta as seções ou ramo de atuação da empresa, seja ela emitente da nota ou trasportadora;

- destinatário (23) - Cliente da nota fiscal, que pode ser uma empresa ou uma pessoa física. Essa tabela possui informações do destinatário como, por exemplo, o endereço do cliente;

- emitente (24) - A empresa que emitiu a NFe. Contém informações como cnae, endereço e outras mais;
- emitente_distancia_destinatario (5) - Apresenta informações sobre a relação da distância que existe entre os destinatários e emitentes registrados;
- imposto (13) - Informações sobre as alíquotas e valores de impostos operados em uma transação;
- imposto_item_nf (4) - Tabela de relacionamento entre os itens das NFe e os impostos incidentes nesses itens;
- item_nf (28) - Apresenta o detalhamento dos itens das NFe como: produto, valor, quantidade, descontos, frete e outros mais;
- municipio (5) - Tabela com os municípios e códigos cadastrados no sistema;
- ncm (5) - Variação do grupo de cnae de uma empresa;
- NFe (33) - Informações sobre o documento fiscal como: data, valor, chave de acesso e transporte;
- produto (7) - Produtos e códigos cadastrados no sistema;
- similaridade (8) - Tabela de relação entre produtos;
- tipo_imposto (2) - Tabela de relacionamento com os tipos de imposto e as NFe;
- tipo_produto (2) - Tabela de relacionamento com os produtos e as NFe;
- totais (17) - Valores totais de: imposto, produto e nota;
- transporte (11) - Informações sobre o transporte da mercadoria: empresa responsável, endereço da empresa, veiculo, etc;
- volume (8) - Especificações do volume transportado.

Com essa base de dados é possível realizar diversas análises mercadológicas, inclusive de demanda reprimida. Como os dados são georreferenciados, análises mais complexas também são viáveis utilizando essa modelagem de dados. Para exemplificar as consultas que podem ser realizadas no banco de dados, a Tabela 4.3 apresenta as 10 vendas mais caras de um determinado tipo de poltrona por uma loja de departamento que possui 4 filiais na cidade do Recife. Nessa consulta é apresentado o valor da poltrona, a data em que foi realizada a venda e o bairro onde a loja fica localizada.

Outra análise que se pode realizar com esses dados é a série histórica da variação do preço de um determinado produto. Para exemplificar foi selecionado um veneno de formiga vendido por uma fábrica de controle de pragas situada no agreste pernambucano. Esses dados são a variação do preço unitário do produto no ano de 2015 e estão presentes na Tabela 4.4. Com

Tabela 4.3 Variação de preço de uma poltrona na cidade do Recife

Valor	Data	Local
2200.00	2015-02-02	BOA VIAGEM
2200.00	2015-02-06	MADALENA
2200.00	2015-01-05	MADALENA
2090.00	2015-01-18	MADALENA
2090.00	2014-09-16	MADALENA
1898.10	2014-11-24	MADALENA
1399.00	2015-12-26	SANTO AMARO
1399.00	2015-12-28	BOA VIAGEM
1399.00	2015-12-21	MADALENA
1399.00	2015-12-17	MADALENA

Tabela 4.4 Variação no preço do produto formitol no ano de 2015

Período	Valor (R$)
01/2015	5,163
02/2015	5,100
03/2015	5,138
04/2015	5,115
05/2015	5,448
06/2015	5,459
07/2015	5,355
08/2015	5,463
09/2015	5,480
10/2015	5,390
11/2015	5,358
12/2015	5,466

essa série histórica é possível utilizar modelos de previsão para se estimar possíveis valores futuros de preço de um determinado produto.

Um outro tipo de comparativo pode ser a relação de vendas de um inseticida para baratas pela população do município para se ver onde o público consumidor de determinado produto está localizado. Em uma análise confrontando o número de habitantes estimado pelo censo do IBGE com o total de vendas de um veneno de barata, filtrando por município, é possível destacar que ao menos: 0.09% Recife, 0.06% Jaboatão dos Guararapes, 0.05% Olinda e 0.04% Paulista da população desses municípios consomem o produto. Também com os mesmos dados é possível afirmar que o município do Recife é o maior consumidor do veneno de barata no estado de Pernambuco, correspondendo a um total de 12,5% de todas as vendas registradas na base de dados desse produto. Aumentando a especificidade da análise é possível definir que o bairro na cidade do Recife que mais consome o produto é o de Boa Viagem com 123 diferentes clientes. Outros bairros como Santo Antônio (96), Imbiribeira (95) e Afogados (95) também possuem um quantitativo considerável de clientes registrados na base para esse produto.

4.2.4 Demanda Reprimida

A demanda reprimida ocorre quando por algum motivo o consumidor deseja, mas não consegue, adquirir um produto. Como exposto anteriormente existem diversos motivos para a ocorrência da demanda reprimida. No entanto, a demanda reprimida abordada nesse trabalho está relacionada a distância entre comprador e fornecedor, o que naturalmente dificulta o acesso do cliente ao produto. Mesmo com a análise de mercado bem elaborada, é muito difícil se afirmar que um determinado local possui realmente potenciais clientes para um determinado produto. Contudo, analisando as notas fiscais é possível inferir sobre a distância percorrida pelos clientes para adquirir um produto. Com essa análise, é possível traçar uma métrica do deslocamento médio e identificar alguns nichos de consumidores que se deslocam além do habitual para conseguir um produto. A proposta desse trabalho, além de indicar onde existe demanda reprimida por um produto de maneira eficiente, também gera informação sobre a microeconomia de uma região.

Nesta visão apresentada, a demanda reprimida pode ser estimada pela média das distâncias de todos os compradores de um determinado produto. Aqueles consumidores que estiverem além de um perímetro de 3 vezes o valor médio é considerado um *outlier* do processo. Quando a densidade de *outliers* é considerável, conforme a metodologia aqui proposta, em relação ao número de vendas totais do produto então pode-se afirmar que existe uma demanda reprimida por esse determinado produto. Quando existe uma concentração desses *outliers* em uma determinada região existe um indicativo que esse local seria uma boa localização para que o emitente das notas fiscais coloque algum ponto de facilitação de vendas desse produto como, por exemplo, uma filial. Logicamente existem diversas questões envolvendo tomada de decisões estratégicas desse nível. Contudo, o objetivo desse trabalho é elucidar a demanda reprimida que existe em uma região.

Por vezes, é necessário se realizar uma limpeza nos dados para que o valor na média não fique tão alto, prejudicando as análises de demanda reprimida. Essas limpezas consistem em desconsiderar elementos que estejam demasiadamente longe do local de venda como é o caso de compradores de outros países.

Para exemplificar o cálculo da demanda reprimida de um determinado produto foi selecionado um produto da área de controle de pragas onde sua origem é a cidade de Bezerros no interior do estado de Pernambuco. Foram contabilizados 11232 vendas do produto onde 607 estiveram fora do raio da demanda reprimida que é de 312,9 km do emitente das NFe. Esse número de clientes representa 5,4% do total de vendas desse produto. Alguns desses clientes estão localizados a 2748 km de distância. Com essas informações é possível avaliar que existem diversos compradores de diferentes regiões do país. Duas das vendas para um mesmo comprador localizado em Portugal não foram contabilizadas nessa avaliação.

4.2.5 Clusterização

Com a distribuição dos pontos de venda fora do raio que delimita a demanda reprimida é possível agrupá-los em *clusters* pela sua similaridade, que nesse caso é a distância entre si. Avaliando as informações do exemplo da demanda reprimida pelo produto de controle de pragas localizado em Bezerros (PE) é possível avaliar que existem ao menos três *clusters* espalhados

entre as regiões centro-oeste, norte e sudeste do país com base nas localizações dos *outliers* do processo. Aplicando o algoritmo *k-means* através do Matlab [57] e supondo que a empresa disponha de recursos para aportar três lojas é possível identificar os pontos que melhor representam a distribuição (centróides) da demanda reprimida pelo produto de controle de pragas que estão localizados nas cidades: Sinop (MT), Campinas (SP) e Belém (PA).

4.2.6 Avaliação das Séries Temporais de Preços dos Produtos

As séries do presente trabalho são montadas com os dados de diversas informações presentes na base de dados. Por exemplo, uma série que pode ser avaliada é a média mensal do preço do veneno de formiga que variou entre R$ 4,55 e R$ 5,46 no período de abril de 2012 até dezembro de 2015. A Figura 4.8 apresenta a série das observações dos preços médios mensais praticados sobre o produto.

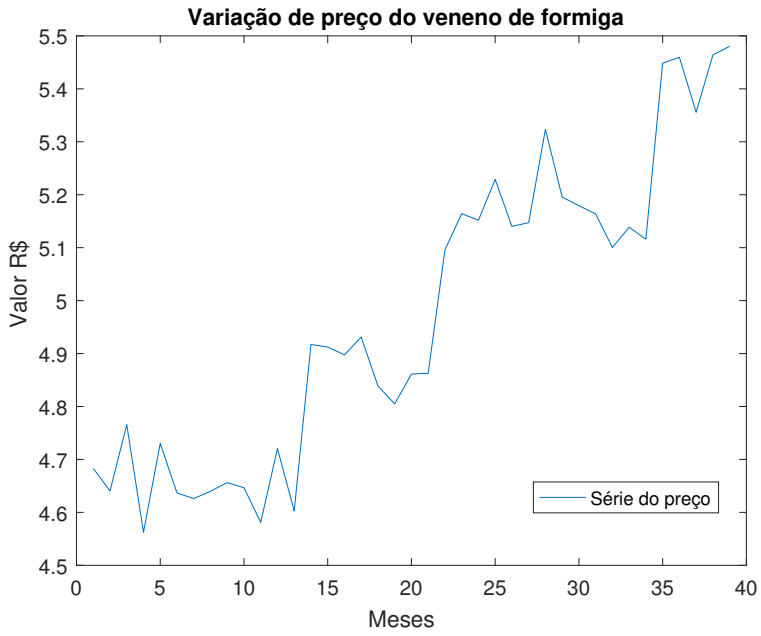


Figura 4.8 Série temporal da variação de preço do formitol

A partir dessas séries temporais é possível calcular as funções de autocorrelação e autocorrelação parcial dessa série. Essas funções são grandes indicativos da previsão de valores futuros das séries em modelos matemáticos de previsão. Outras informações podem ser obsercadas como, por exemplo, a natural tendência de elevação do preço médio mensal ao longo dos anos e uma característica de sazonalidade apresentada por essa série. A estacionariedade da série indica que os pontos dessa série são correlacionados e que existe a possibilidade de se usar preditores para prever os pontos futuros dessa série.

4.3 Tecnologias Adotadas

Dentre as tecnologias adotadas pode-se citar:

- A linguagem de programação JAVA [79]. A utilização dessa linguagem foi devido a experiência de utilização dessa linguagem juntamente com a sua portabilidade. Alguns módulos do sistema foram desenvolvidos em linguagem de *script* e a linguagem JAVA facilitou muito a integração com esses *scripts*. Da mesma forma a linguagem facilitou a persistência dos dados tanto em banco de dados relacionais quanto em tentativas de implementação em bancos NoSQL.
- A persistência dos dados foi feita utilizando MySQL [78]. Foi realizada uma modelagem dos dados com base nas *tags* dos arquivos xml. Algumas tabelas têm uma quantidade de atributos muito grande, outras possuem muitos registros. Porém, as consultas realizadas conseguiram retornar em tempo hábil os resultados e não foi necessário uma abordagem através de outra modelagem de dados.
- Outra ferramenta para análise de algoritmos de clusterização e geração das análises de séries temporais foi o MatLab [57]. O MatLab é uma ótima ferramenta que possui várias funcionalidades para calcular diversas funções, possui também diversos algoritmos implementados o que facilitou muito as análises realizadas nesse trabalho.
- A ferramenta de desenvolvimento utilizada foi o ambiente Eclipse [80]. Esse ambiente facilitou diversos aspectos do desenvolvimento do sistema como versionamento de código, *debug* e utilização de bibliotecas externas.
- Uma ferramenta relacionada diretamente ao sistema foi a biblioteca SELENIUM[3]. Essa biblioteca, usada geralmente para testes de sistemas *web*, manipula o *browse* e realiza comandos programados.
- O sistema operacional Debian [81] que facilitou vários aspectos do desenvolvimento do trabalho desde comunicações com o servidor da aplicação através de SSH até envio de e-mail informando que o sistema parou de funcionar por algum motivo
- Por fim, a ferramenta Batchgeo[4] foi importante para inserir os dados no mapa.

4.4 Comparativo com Trabalhos Relacionados

Essa seção apresenta uma comparativa dos trabalhos relacionados do Capítulo 3 com a presente dissertação, destacando o diferencial da dissertação proposta com relação aos trabalhos relacionados. A Tabela 4.5 apresenta esse comparativo.

[3]Disponível em: http://www.seleniumhq.org/ Acessado em: 11-10-2017.
[4]Disponível em: https://batchgeo.com/ Acessado em: 13-10-2017.

Tabela 4.5 Comparação dos trabalhos relacionados com a dissertação

Trabalho	Área	Comparação
Uma simples abordagem genérica para decifrar textos baseados em CAPTCHA	Reconhecimento de padrões	A metodologia utilizada nessa dissertação para o reconhecimento de caracteres em CAPTCHA é bem mais simplificada do que a proposta pelo trabalho relacionado e apresenta resultados similares.
Clusterizando dados do mercado de ações indiano	Clusterização	Apesar de usar o mesmo algoritmo de clusterização que o trabalho relacionado a presente dissertação faz uso de dados mais ricos. Uma vez que o trabalho relacionado aborda o mercado de ações, a dissertação trata de informações dinâmicas sobre o mercado, promovendo assim uma análise mais complexa e resultados finais mais elaborados.
Análise de modelos de dados não relacionais e multimensionais	Mineração de dados	Na dissertação foi utilizado o modelo de banco de dados relacional. O trabalho relacionado fornece uma descrição abrangente das dificuldades da utilização de outros modelos. Existe um grande indício da necessidade do uso de modelos não-relacionais num futuro próximo do trabalho dessa dissertação por conta da grande massa de dados que continua a crescer.
Método de Pesquisa de Mercado e Sistema para Coleta de Dados de Mercado e Análise de Mercado	Análise de mercado	Esse trabalho relacionado tem propostas similares com relação a análise de mercado. Contudo, a dissertação é bem mais abrangente nas suas análises não ficando limitada apenas a lojas parceiras como é o caso do trabalho relacionado.
Previsão do preço do milho, através de séries temporais	Previsão de séries temporais	A presente dissertação não tratou diretamente de previsão de séries temporais. Porém, em relação a análise de séries temporais se assemelhou bastante a metodologia proposta por esse trabalho relacionado.
Operação serenata de amor	Análise de notas fiscais	Esse é o trabalho relacionado que apresenta maior similaridade com a dissertação. Principalmente, no que se diz respeito a proposta. Contudo, o trabalho relacionado carece de um embasamento acadêmico e também de trabalhos divulgando para a comunidade científica suas abordagens com relação ao tema de mineração de notas fiscais.

CAPÍTULO 5

Resultados

Os resultados finais dessa dissertação serão apresentados nesta seção. É abordado aqui as análises das funcionalidades do sistema com as consultas no banco de dados, demanda reprimida e análise de séries temporais com suas respectivas discussões.

5.1 Consolidação dos Dados

A base de dados possui 120 empresas de 57 ramos diferentes. As 10 empresas que mais tiveram notas fiscais obtidads do portal da SEFAZ estão presentes na Tabela 5.1. Essa tabela possui a quantidade de notas obtidas da empresa, CNAE, município e bairro desses emitentes. As informações estão ordenadas pela quantidade de notas mineradas. A variedade de CNAE, bairros e até município indicam a grande variedade de emitentes de notas fiscais captados para a análise.

Tabela 5.1 Dez maiores emitentes de notas fiscais da base de dados

quantidade	CNAE	município	bairro
61565	4651601	RECIFE	VARZEA
49015	4753900	RECIFE	PINA
38947	4753900	RECIFE	BOA VIAGEM
36600	4742300	RECIFE	SANTO AMARO
28658	1094500	RECIFE	IMBIRIBEIRA
25032	2052500	BEZERROS	DISTRITO INDUSTRIAL
16970	4744001	RECIFE	IMBIRIBEIRA
15996	4635401	RECIFE	MACAXEIRA
15933	4713001	RECIFE	MADALENA
15298	4637199	RECIFE	BONGI

A base de dados possui um total de 139468 destinatários no geral. Esses clientes são de diversos estados e municípios, inclusive clientes de outros países. Os clientes da base de dados são na maioria do Brasil, mas a base possui também clientes de outros 10 países diferentes. A grande maioria das vendas se concentra no município do Recife e na região metropolitana dessa cidade. A cidade do Recife representa cerca de 55,8% de vendas da base analisada e o estado de Pernambuco tem um total de 88,8% das vendas da base de dados. Dentre os bairros aquele que apresenta maior representatividade é o bairro de Boa Viagem. A Tabela 5.2 apresenta a quantidade de vendas dos 10 bairros que mais possuem vendas no município do Recife. Cerca

de 24,4% dos emitentes são pessoas jurídicas representando comercializações do tipo *Business to Business (B2B)*[1].

Tabela 5.2 Tabela com os 10 bairros que mais registraram vendas

bairro	quantidade
BOA VIAGEM	66090
CENTRO	55255
IMBIRIBEIRA	22248
BOA VISTA	16992
MADALENA	15366
GRACAS	15251
SANTO AMARO	15179
ESPINHEIRO	12964
PINA	11432
SANTO ANTÔNIO	10798

Existe dentro da base uma grande diversidade de produtos das mais variadas linhas que vão de automóveis até água mineral. Existem ao todo na base de dados 171105 produtos registrados de 12498 categorias. Por exemplo, o produto *lâmpada incandescente 100W 220V PHILIPS* faz parte da categoria *máquinas, aparelhos e matériais elétricos*. Os 20 produtos mais vendidos estão presentes na Tabela 5.3 juntamento com suas respectivas quantidade de unidades vendidas. O produto mais caro vendido foi uma *BMW X6 XDRIVE 50I SPORT* no valor de R$ 462.000,00 o mais barato foi um *rebite de ferro duplo 10MM* no valor de R$ 0,11 .

A modelagem da base de dados permite consultas muito específicas. Por exemplo, o segmento de negócio que mais faturou no município do Recife foi o comércio varejistas de automóveis e utilitários novos com um faturamento total de R$ 472.921.840,43 no ano de 2015. A Tabela 5.4 apresenta a quantidade de emissões de notas fiscais dos 10 segmentos mais representativos do banco de dados. Através dessa tabela é possível avaliar que o segmento do comércio varejista de aparelhos eletrônicos como: aparelhos de som, máquinas de lavar, arcondicionado, televisão e outros eletrodomésticos apresenta muito faturamento de notas fiscais da base avaliada.

Existem notas de diversos valores na base de dados. Inclusive notas com valor de R$ 580.000,00 correspondente a 100.000 litros de gasolina para uso de veículos aéreos. Cada uma dessas notas possui discriminação de impostos dos tipo: ICMS, imposto sobre operação financeira (IOF), contribuição para o financiamento da seguridade social (COFINS), etc. Os dez produtos comercializados dentro do Recife que tiveram maior tributação até um teto de R$ 1000,00 são apresentados na Tabela 5.5. Essa tabela apresenta o valor total tributado, valor da nota fiscal, produto e quantidade. É importante afirmar que todas as análises foram realizadas com base nos dados minerados dentro de um período específico de tempo entre 2012 e 2017. Não podendo assim representar todo o mercado pernambucano, mas contendo um bom referencial para o que acontece nessa dinâmica mercadológica.

[1] *B2B* ou empresa para empresa é a modalidade de comércio realizada entre empresas.

Tabela 5.3 Vinte produtos mais vendidos e suas respectivas quantidades de unidades vendidas

quantidade	produto
1243650	OVOS BRANCO MEDIO
224040	COXA C/SOBRECOXA CONG.
75000	PORTA-ANILHA REF PA1 - PRETA
71999	VERNIZ PU 923-155 HS - GLASURIT 18L
51450	CAPA DE CADERNO CAPA DURA COM IMPRESSAO -
30000	ORINGBUNA-2118
30000	CONJUNTO MO100/4 NUTRICAO 1200ML P/ UND - UND
29250	PEIXE CORVINA
28880	SUCATA PLASTICA
27000	QUEIJO MUSSARELA PECA
25780	SUCATA DE ONIBUS
23000	PRESTACAO DE SERV. DE ARMAZENAGEM
22000	PAPEL CREPADO 30 X 30 CX C/2000 - AMCOR FLEXI. -
21636	MERLUZA DO ALASKA HGT IMPORTADA
20402	LIXA DAGUA NORTON 360
20000	DICLORVOS - DDVP 98%
19589	ESTOPA COR ROSA
18000	PREGO ANELADO PBL 50MM
18000	EMB TERM DESC 500ML S/D C/TAMPA UNID
17450	VERNIZ PU HS 255 5L

5.2 Demanda Reprimida

Como prova de conceito a análise de demanda reprimida aqui se baseou na escolha de três produtos diferentes: gasolina de aviação, televisão de led e diclorvol. A gasolina de aviação (AVGAS) é um combustível de alta octanagem utilizado em aeronaves. Esse produto é um exemplo do onde não ocorre demanda reprimida observável, onde a grande maioria dos seus produtos é vendido dentro de um determinado perímetro próximo ao ponto de comércio. Televisões de led englobam monitores e televisões entre 23 e 42 polegadas das marcas *AOC*, *Sansumg* e *Philips*. Esse é um exemplo de produto que existe demanda reprimida. Diclorvol é um inseticida líquido eficaz no combate de moscas, baratas e pernilongos. Esse é um produto onde existe um forte indicativo de demanda reprimida.

A primeira análise é a do produto gasolina de aviação. Foram analisados vendas de galões de 100 litros de AVGAS. Ao todo foram computadas 3650 vendas. O comércio desse produto fica localizado próximo ao aeroporto internacional dos Guararapes na cidade do Recife. O raio médio das vendas desse produto foi de 188 kilômetros. Ao todo 40 compras ficaram fora do raio da demanda reprimida. A Figura 5.1 apresenta um mapa indicando os clientes que ficaram fora do raio da demanda reprimida. É possível considerar que os clientes que ficaram fora do raio são inexpressivos em comparação ao total de vendas sendo que eles representam apenas 1,09% das vendas e ainda sim não concentrados em local específico.

Tabela 5.4 Dez segmentos com maior quantidade de notas emitidas no município do Recife

Quantidade NFe	CNAE	Descrição
90396	4753900	Comércio varejista de aparelhos eletrônicos
67576	4651601	Comércio atacadista de produtos de Informática
50003	4744001	Materiais metálicos de construção
39797	4742300	Materiais elétricos de construção
33252	4713001	Loja de departamento
28658	1094500	Massas alimentícias
26754	4663000	Máquinas e aparelhos de uso industrial
25032	2052500	Produtos de controle biológico de pragas
24856	4530701	Acessórios automobilísticos
19369	4511101	Comércio de automóveis

Tabela 5.5 Maiores tributação dos produtos comercializados em Recife

Tributação	Valor Nota	Produto	Quant
999,97	13.157,48	PERFIL LAMPVC VINC 7,00M BRAN 200	1.372
999,60	5.880,00	PERFIL LAMPVC VINC 6,00M 200	600
997,83	5.772,00	MINI SYSTEM 500W MX SAMSUNG	6
997,68	8.314,00	SMARTPH MOTO E DTV PTO	17
997,50	1.995,00	PERFIL LAMPVC FRIS200 6,00M	2.100
997,02	3.692,68	AVGAS - 100 LL	879
996,83	5.863,68	BARATOL GEL 10G	1.152
996,63	5.862,54	PCHAMEX 75G 210X297 MULTI,CHAMEX	491
996,63	5.885,00	PAPEL A4 75G. 500F. HP	500
996,45	8.303,72	1.1/38 SO T110 FIBRA POLIESTER	1.845

Uma outra análise foi a das vendas da televisão e monitores de led de vários tamanhos e marcas. Foi selecionado uma loja de departamento eletrônico localizada na cidade do Recife no bairro da Várzea. O total de vendas desses televisores contabilizado foi de 3909 produtos. O raio médio de venda desses produtos foi de 172 km. E ocorreram 248 vendas fora do raio da demanda reprimida. Isso indica que cerca de 6,3% das vendas são de demanda reprimida. A Figura 5.2 apresenta um mapa dos endereços dos clientes que estão fora do raio da demanda reprimida.

Como indicado na Figura 5.2 existe uma grande concentração de vendas fora do raio da demanda reprimida (mais de 5% do total de vendas) e supodo-se que os gestores querem ampliar sua loja em três pontos, aplicando o algoritmo *k-means* as cidades mais indicadas para comportarem pontos de distribuições para esse produto seriam: Rio Branco-AC, Teresina-PI e Campinas-SP. A Figura 5.3 indica os pontos no mapa de onde poderiam se localizar pontos de vendas para suprir a demanda reprimida pelo produto analisado.

O último produto analisado possui um grande indicativo de presença de demanda reprimida. As vendas contabilizadas do diclorvol CE de 1 litro foram de 4553 produtos. A distribuidora desse produto fica localizado no município de Sairé no agreste pernambucano. A distância

Figura 5.1 Mapa das vendas de gasolina de aviação fora do raio da demanda reprimida

média das vendas foram de 485 kilômetros. Ao todo 749 vendas estiveram fora do raio da demanda reprimida. A Figura 5.4 indica os pontos de venda que ficaram fora do raio da demanda reprimida.

Mais uma vez supondo que os gestores indiquem a possibilidade de expansão em três pontos, aplicando o algoritmo *k-means*, o resultado é apresentado na Figura 5.5 que apresenta a clusterização dos pontos de vendas do produto diclorvol. Nessa imagem fica evidenciado que o algoritmo indicou as cidades: Belém-PA, Feira de Santana-BA e Campinas-SP como possíveis pontos de vendas para o produto.

É importante ressaltar que todas as sugestões de pontos de comércio são baseadas em algoritmos de clusterização que não levam em consideração as viabilidades econômicas de cada local.

Figura 5.2 Mapa dos vendas de monitores e televisões de led fora do raio da demanda reprimida

5.3 Séries de Mercado

Com base em sua representatividade nas consultas realizadas na base de dados foram escolhidos dois produtos para a análise das séries temporais desse trabalho: um veneno contra cupins e um tubo de esgoto de 100 mm de uma marca específica. Para o veneno contra cupins foi analisada a flutuação média de preço em um período de 3 anos e 5 meses. Já para o tubo de esgoto de 100 mm foi realizada tanto uma análise da flutuação média de preço quanto uma análise sobre a quantidade de vendas no mesmo período de 3 anos e 5 meses. Essas vendas são a análise das lojas específicas onde se comercializam os produtos. O veneno de cupim fica localizado no município de Bezerros e a loja de tubulação se localiza no município do Recife. A seguir serão apresentados as séries temporais para as análises dos dois produtos juntamente com as funções de autocorrelação e autocorrelação parcial dessas séries. Tanto as representações das séries quanto o cálculo das funções de autocorrelação e autocorrelação parcial foram realizadas no MatLab versão R2016a. Os valores das séries avaliadas nesse trabalho estão presentes no Apêndice B desse trabalho.

Figura 5.3 Mapa dos *clusters* dos pontos de vendas para monitores e televisões de led fora do raio da demanda reprimida

A primeira análise é da variação de preço médio do veneno de cupim. Foram coletados dados do período de julho de 2012 até dezembro de 2015 totalizando 42 observações que variam de R$ 10,51 até R$ 14,51. A série dessa variação está presente na Figura 5.6. Como já esperado, a valor médio praticado pela empresa para o veneno de cupim teve um crescimento ao longo dos meses observados, tendo destaque para a variação entre os meses de outubro e novembro de 2013 onde é possível observar uma variação de R$ 0,90 o que é elevada se for comparado com a variação média dos preços que foi de R$ 0,17. Outra característica dessa série são padrões de sazonalidade. Apesar disso, testes de estacionariedade[2] indicaram que essa série tem comportamento estacionário, com as suas variáveis constantes ao longo do tempo.

Outra análise é o comportamento das funções de autocorrelação e autocorrelação parcial. A função de autocorrelação dessa série indica um decaimento suave dos valores da correlação entre os elementos da série o que indica uma possível característica de um modelo auto-regressivo. Essas funções estão representadas na Figura 5.7.

O outro produto analisado foi o tubo de esgoto de 100 mm. Primeiramente foi analisado a série histórica da variação média mensal de preço desse produto no intervalo de julho de 2012 até dezembro de 2015. Essa série tem 42 observações e possui valores que variam de R$ 38,22 até R$ 50,28. Mais uma vez, como já era esperado, o valor médio mensal do produto sofreu um aumento gradual no seu preço com algumas variações acima da média. Outra semelhança com a série anterior são os aparentes indícios de sazonalidade apesar de testes de estacionariedade também confirmarem que a série é estacionária. Essa série é apresentada na Figura 5.8.

[2]O teste de estacionariedade utilizado nesse trabalho foi o teste de Dickey-Fuller que verifica se o modelo autoregressivo da série possui ou não raiz unitária. Caso possua raiz unitária essa série é estacionária [82].

Figura 5.4 Mapa das vendas do diclorvol fora do raio da demanda reprimida

Na Figura 5.9 são apresentadas as funções de autocorrelação e autocorrelação parcial dessa série. A função de autocorrelação com decaimento suave indica que essa série tem um possível comportamento de um modelo auto-regressivo. A partir da análise da função de autocorrelação parcial esse modelo tem características de ser um modelo auto-regressivo de ordem dois).

Por fim a análise da quantidade de vendas do tubo de esgoto no mesmo período que foi realizada a análise anterior. A série da variação de vendas mensal está presente na Figura 5.10. Essa série apresenta valores que variam de 252 até 1155 unidades vendidas. Essa série, diferentemente das outas apresenta um comportamento de declínio, o que indica queda nas vendas do produto por alguma razão. Essa série falhou no teste de estacionariedade o que indica que ela tem ao menos uma propriedade que não está sendo constante no tempo.

A Figura 5.11 apresenta os comportamentos das funções de autocorrelação e autocorrelação parcial da série da variação de quantidade de unidades vendidas desse produto. Apesar da análise de queda não gradual da função de autocorrelação indicar um possível modelo de médias móveis para essa série ela falhou no teste de estacionariedade sendo necessário uma determinada quantidade de diferenciação para que seja possível a modelagem estatística dessa

Figura 5.5 Mapa dos *clusters* dos pontos de vendas do diclorvol

série.

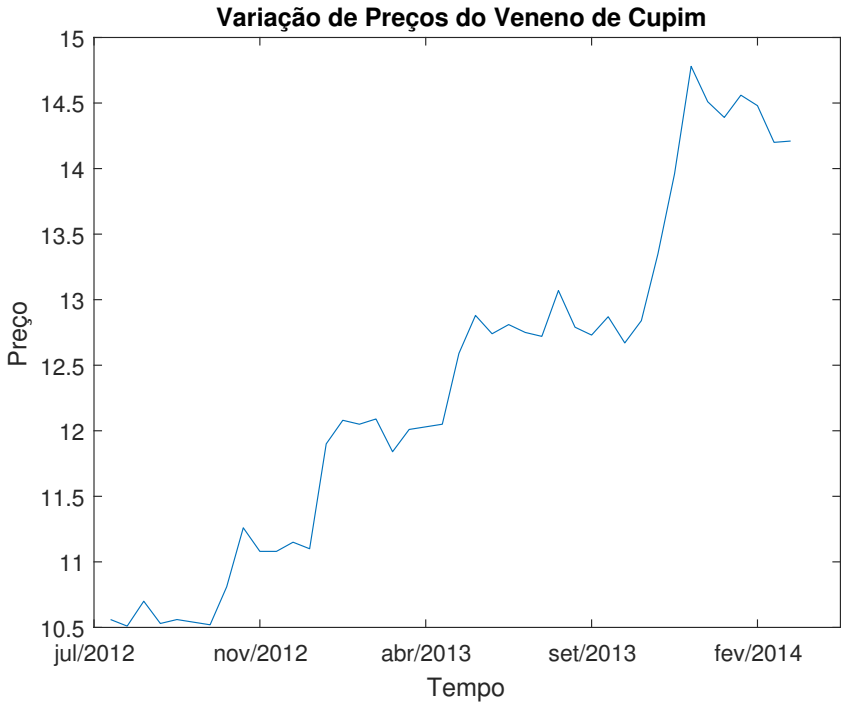


Figura 5.6 Série histórica da variação de preço do veneno de cupim num período de 3 anos e meio

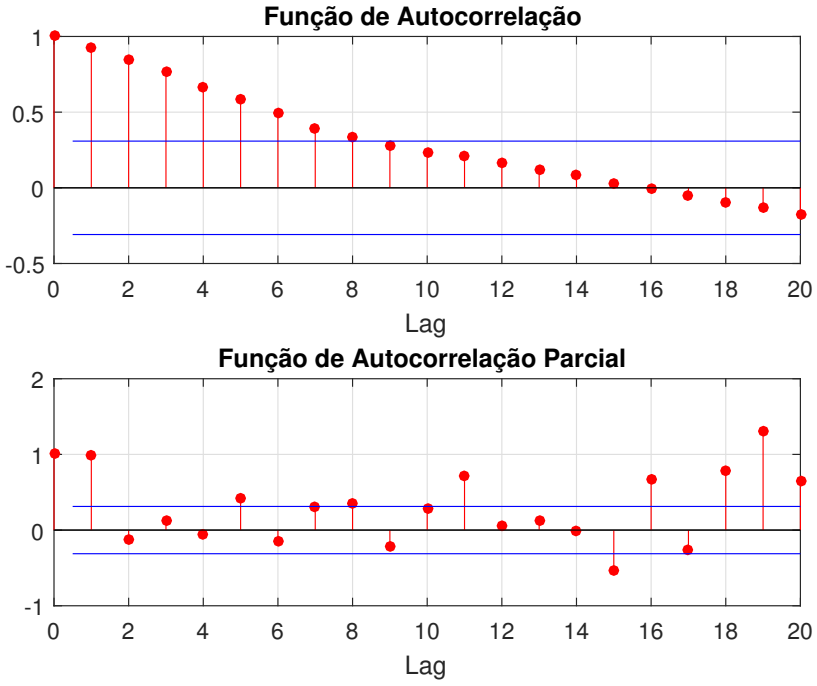


Figura 5.7 Função de autocorrelação e autocorrelação parcial da variação de preço do veneno de cupim num período de 3 anos e meio

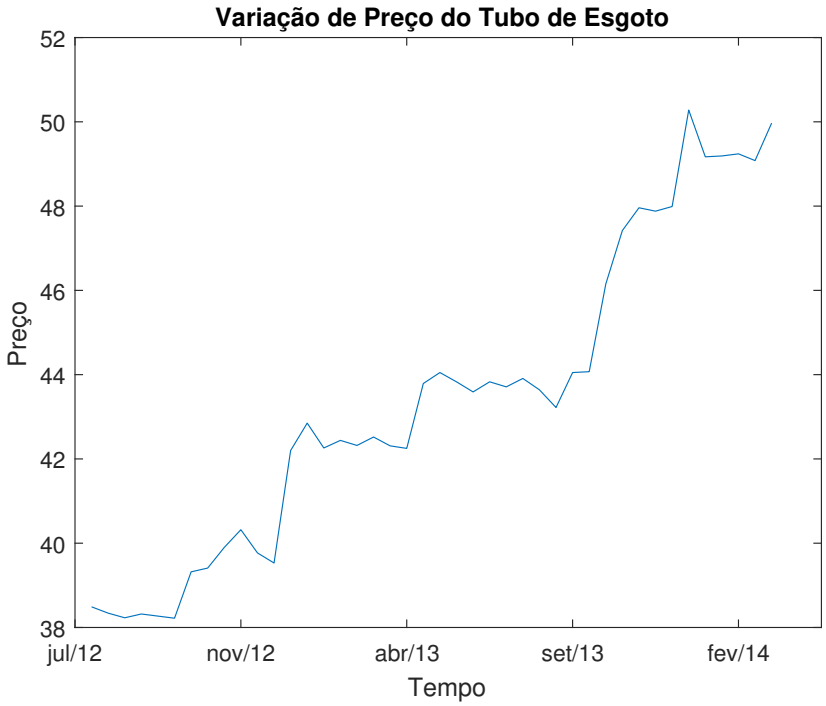


Figura 5.8 Série histórica da variação de preço do tubo de esgoto de 100mm num período de 3 anos e meio

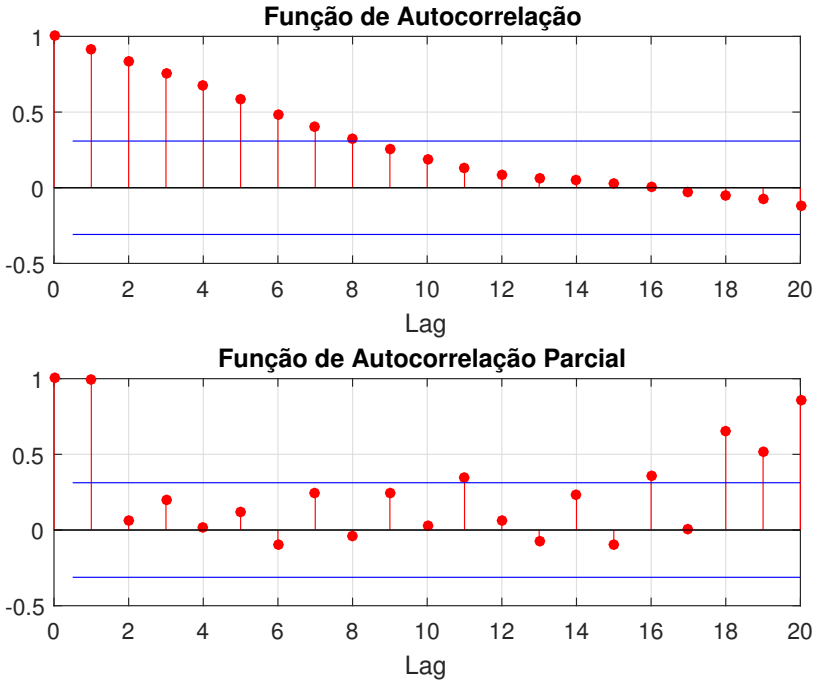


Figura 5.9 Função de autocorrelação e autocorrelação parcial da variação de preço do tubo de esgoto de 100mm num período de 3 anos e meio

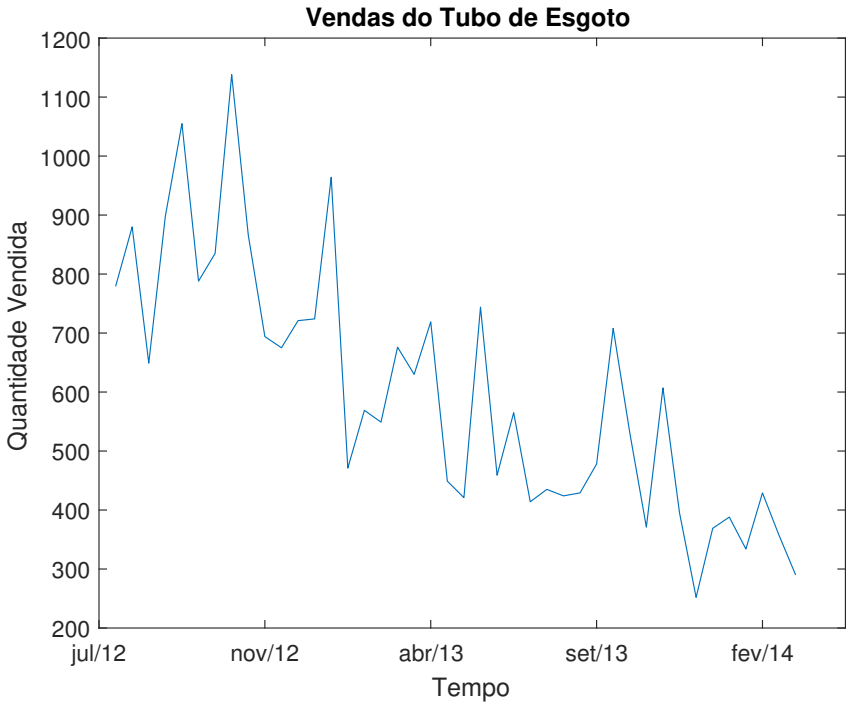


Figura 5.10 Série da quantidade de tubos de esgoto de 100mm vendidos num período de 3 anos e meio

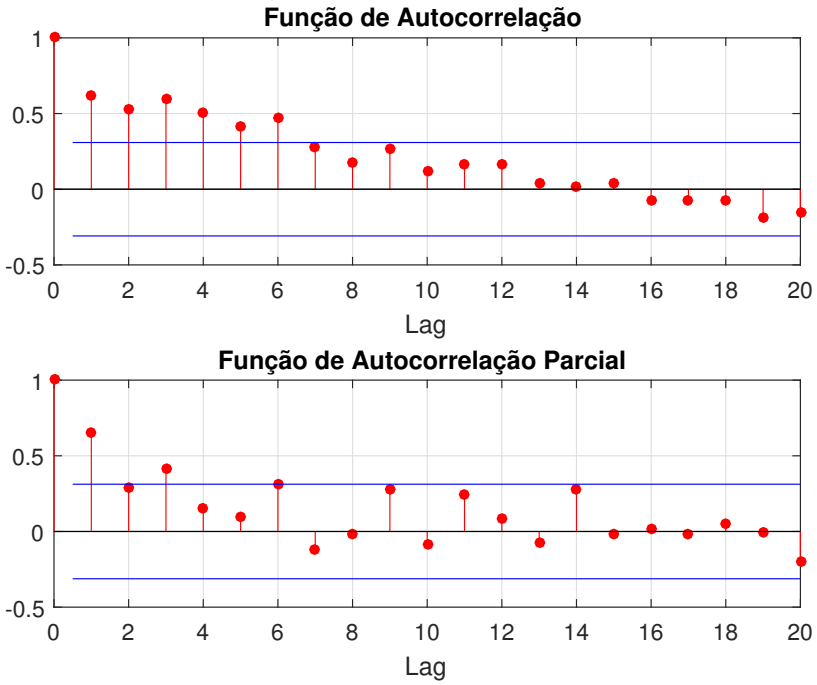


Figura 5.11 Função de autocorrelação e autocorrelação parcial da série da quantidade de tubos de esgoto de 100mm vendidos

CAPÍTULO 6

Conclusão e Trabalhos Futuros

O desenvolvimento desse trabalho apresentou vários aspectos da dinâmica mercadológica de algumas empresas de Pernambuco, boa parte situada na Região Metropolitana do Recife. Foram empregadas diversas técnicas de computação para se obter informações sobre o mercado local. De modo geral, as informações contidas nas notas fiscais eletrônicas podem ser usadas para várias análises sobre o comportamento do mercado. Um dos aspectos mais relevantes desse estudo foi a determinação da demanda reprimida por certo produto de uma maneira quantitativa e dinâmica. Pesquisa de aceitação ou intenção de compra de um determinado produto podem ser realizados através desse sistema. É possível avaliar também as séries históricas dos movimentos comerciais e como elas podem influenciar na tomada de decisão para gestores. Além disso, utilizando modelos estatísticos para modelagem de séries pode-se até mesmo prever o comportamento da demanda e do preço de um determinado produto. Em relação aos objetivos do trabalho pode-se dizer que foi construída uma base de dados consolidada com as informações das notas fiscais. Em relação aos objetivos específicos da pesquisa os resultados obtidos foram:

- Análises em tempo e espaço de produtos, como o veneno de cupim e a gasolina de avião;
- Análises de demandas reprimidas, como no caso do diclorvol;
- Análise das séries históricas, como no caso do tubo de esgoto;

A maioria dos objetivos específicos do trabalho foi concluída. Alguns pontos como análises do perfil geográfico de compras de um determinado grupo de consumidores e a previsão das séries temporais foram objetivos que ficaram como trabalhos futuros dessa dissertação por conta de limitações de tempo e poder computacional disponíveis para execução desse trabalho.

As ferramentas utilizadas nesse trabalho apresentaram grande eficiência na criação do sistema. Dentre elas é possível destacar os acervos de periódicos tanto da CAPES quanto do Google Acadêmico que contribuíram no material bibliográfico que deu suporte a essa pesquisa. As ferramentas e bibliotecas de desenvolvimento tanto da linguagem Java quanto o base de dados MySQL também merecem destaque. Outra importante ferramenta para executar algoritmos de clusterização e as funções relacionadas as séries temporais foi o MatLab. Por fim, outra ferramenta adotada que trouxe grandes benefícios para apresentar os pontos nos mapas foi a ferramenta Batchgeo.

6.1 Limitações

As principais dificuldades relacionadas ao trabalho foram a inexperiência com programação *web* e a dificuldade de comunicação com o servidor das NFe. Essa dificuldade foi contornada com o uso da biblioteca SELENIUM que possibilitou a mineração das notas fiscais. Além disso a própria estabilidade do servidor da SEFAZ foi outro problema enfrentado durante o desenvolvimento desse trabalho, uma vez que esse servidor muitas vezes se encontrava inoperante. Outro ponto que pode ser ressaltado na questão de limitações, é o foco de um segmento de mercado específico, procurando realizar as análises com o maior número possível de empresas de um mesmo segmento dentro de uma limitação geográfica. Com isso as análises perdem um pouco da generalidade mas contribuem mais pontualmente para o entendimento do mercado que se esteja estudando. Além dessa especificidade nas análises a falta de tempo e de recursos computacionais favoreceu a não realização de alguns objetivos iniciais do trabalho como: identificação de perfil geográfico de consumidor pro produto ou segmento e previsão das séries temporais de mercado.

6.2 Contribuições

As principais contribuições desse trabalho foram:

- Um sistema de mineração de notas fiscais que alimenta uma base de dados que é utilizada para fazer uma análise dinâmica e quantitativa do mercado pernambucano com base nas empresas cadastradas. O sistema desenvolvida tem alto potencial comercial e pode contribuir para tomada de decisão de muitos gestores tanto da área pública quanto da área privada;

- Um modelo empírico de determinação de demanda reprimida por um produto em relação ao deslocamento dos clientes;

- Análise de séries temporais de mercado com base nas informações das NFe mineradas;

- Publicação de conhecimento científico com artigos nas áreas de reconhecimentos de padrões e análise de séries temporais, submissões de artigos sobre modelos de dados e a conclusão de uma dissertação de mestrado.

6.3 Trabalhos Futuros

Com base nos resultados obtidos do presente trabalho têm-se ideias para alguns possíveis trabalhos relacionados. Algumas ideias para trabalhos futuros são:

- Análise de perfil geográfico dos consumidores. Uma análise mais profunda dos consumidores juntamente com dados de organizações e empresas como o IBGE, pode ter grande impacto nas análises de mercado;

- Uma comparação de uma pesquisa de mercado utilizando o sistema desenvolvida e os métodos tradicionais de questionários;
- Previsão das séries históricas das demandas por produtos e da variação de preços desses produtos.

APÊNDICE A

Modelo Lógico da Base de Dados

Esse apêndice apresenta o modelo Entidade Relacionamento (ER) do banco de dados do sistema desenvolvido nesse trabalho.

Figura A.1 Modelo ER da base de dados das NFe

APÊNDICE B

Séries Avaliadas no Trabalho

Nesse apêndice são apresentadas os dados das séries utilizadas na Seção 5.3.

Tabela B.1 Série da variação de preço do veneno de cupim

Data	**Preço**	**Data**	**Preço**
jul/12	10,56	abr/14	12,59
ago/12	10,51	mai/14	12,88
set/12	10,70	jun/14	12,74
out/12	10,53	jul/14	12,81
nov/12	10,56	ago/14	12,75
dez/12	10,54	set/14	12,72
jan/13	10,52	out/14	13,07
fev/13	10,81	nov/14	12,79
mar/13	11,26	dez/14	12,73
abr/13	11,08	jan/15	12,87
mai/13	11,08	fev/15	12,67
jun/13	11,15	mar/15	12,84
jul/13	11,10	abr/15	13,35
ago/13	11,90	mai/15	13,96
set/13	12,08	jun/15	14,78
out/13	12,05	jul/15	14,51
nov/13	12,09	ago/15	14,39
dez/13	11,84	set/15	14,56
jan/14	12,01	out/15	14,48
fev/14	12,03	nov/15	14,20
mar/14	12,05	dez/15	14,21

Tabela B.2 Série da variação de preço do tubo de esgoto de 100 mm

Data	**Preço**	**Data**	**Preço**
jul/12	38,49	abr/14	44,05
ago/12	38,34	mai/14	43,83
set/12	38,23	jun/14	43,59
out/12	38,32	jul/14	43,83
nov/12	38,27	ago/14	43,71
dez/12	38,22	set/14	43,91
jan/13	39,32	out/14	43,64
fev/13	39,41	nov/14	43,22
mar/13	39,90	dez/14	44,05
abr/13	40,32	jan/15	44,07
mai/13	39,77	fev/15	46,15
jun/13	39,53	mar/15	47,42
jul/13	42,20	abr/15	47,96
ago/13	42,85	mai/15	47,88
set/13	42,26	jun/15	47,99
out/13	42,44	jul/15	50,28
nov/13	42,32	ago/15	49,17
dez/13	42,52	set/15	49,19
jan/14	42,31	out/15	49,24
fev/14	42,25	nov/15	49,08
mar/14	43,79	dez/15	49,97

Tabela B.3 Série da quantidade de tubos de esgoto de 100 mm vendidos mensalmente

Data	**Quantidade**	**Data**	**Quantidade**
jul/12	779	abr/14	421
ago/12	880	mai/14	744
set/12	649	jun/14	459
out/12	898	jul/14	565
nov/12	1055	ago/14	414
dez/12	788	set/14	435
jan/13	835	out/14	424
fev/13	1138	nov/14	429
mar/13	866	dez/14	478
abr/13	694	jan/15	708
mai/13	675	fev/15	531
jun/13	721	mar/15	371
jul/13	724	abr/15	607
ago/13	964	mai/15	395
set/13	471	jun/15	252
out/13	569	jul/15	369
nov/13	549	ago/15	388
dez/13	676	set/15	334
jan/14	630	out/15	429
fev/14	719	nov/15	357
mar/14	449	dez/15	290

Referências Bibliográficas

[1] W. R. d. Oliveira and D. F. Maia, "Nota fiscal eletrônica: projeto nacional e a iniciativa municipal de são paulo - uma análise comparativa," in *Congresso Brasileiro de Contabilidade*, vol. 18, 2008, pp. 1–15.

[2] A. N. Müller, R. do Pilar, and V. M. Kido, *Manual da nota fiscal eletrônica*. Juruá, 2007.

[3] C. V. B. Ferreira and D. O. de Morais, "Nota fiscal eletrônica," *Congresso Brasileiro de Contabilidade*, p. 55, 2012.

[4] R. Rivest, "The md5 message-digest algorithm," *Internet Engeneering Task Force*, 1992.

[5] X. Yi, "Hash function based on chaotic tent maps," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 52, no. 6, pp. 354–357, 2005.

[6] SeFaz, "Manual de orientação do contribuinte," *Secretária da Fazenda do Governo Federal do Brasil*, 2015.

[7] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*. MIT press, 2001.

[8] D. T. Larose, *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

[9] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Global Institute*, 2011.

[10] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive datasets*. Cambridge university press, 2014.

[11] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact." *MIS quarterly*, vol. 36, no. 4, 2012.

[12] E. Bursztein, M. Martin, and J. Mitchell, "Text-based captcha strengths and weaknesses," in *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 2011, pp. 125–138.

[13] K. Chellapilla, K. Larson, P. Y. Simard, and M. Czerwinski, "Building segmentation based human-friendly human interaction proofs (hips)," in *2nd International Workshop on Human Interactive Proofs*. Springer, 2005, pp. 1–26.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[15] T. F. Dictionary, "Farlex financial dictionary," 2009.

[16] T. W. Anderson, *An introduction to multivariate statistical analysis*. Wiley New York, 1958, vol. 2.

[17] E. Rodrigues de Faria, M. A. M. Ferreira, L. Maia dos Santos, and S. d. F. R. Silveira, "Fatores determinantes na variação dos preços dos produtos contratados por pregão eletrônico," *Revista de Administração Pública-RAP*, vol. 44, no. 6, 2010.

[18] A. Calderelli, *Enciclopédia contábil e comercial brasileira*. CETEC, 1997.

[19] M. Barros, M. C. P. G. Faria, P. H. B. P. Alves, and R. dos Santos Souza, "Nota fiacal eletrônica," Ph.D. dissertation, Centro Universitário Católico Salesiano Auxilium, 2008.

[20] D. DIAS, "Função da nota fiscal na produção de informações contábeis, demonstrações e apuração dos impostos," *Revista Científica da Faminas*, vol. 2, p. 124, 2006.

[21] S. A. Pereira, R. Locks, D. S. Matos, and G. B. da Costa, "Governança eletrônica na administração pública: estudo de caso sobre a nota fiscal eletrônica - nf-e," in *XVIII Congresso Brasileiro de Contabilidade*, 2008.

[22] S. M. d. Silva Júnior, "Certificação digital a importância para órgão público," *Centro Universitário de Brasília*, 2016.

[23] B. Bos *et al.*, "Xml in 10 points," *W3C*, vol. 200, 1999.

[24] S. TRAIN, "Identidade digital: Como os certificados digitais estão facilitando a vida das pessoas," *Câmara Brasileira do Livro - Certisign Certificadora Digital*, vol. 2, 2007.

[25] S. Theodoridis and K. Koutroumbas, "Pattern recognition and neural networks," *Machine Learning and Its Applications*, pp. 169–195, 2001.

[26] S. Watanabe, *Pattern recognition: human and mechanical*. John Wiley & Sons, Inc., 1985.

[27] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[28] R. Bellman, R. Kalaba, and L. Zadeh, "Abstraction and pattern classification," *Journal of Mathematical Analysis and Applications*, vol. 13, no. 1, pp. 1–7, 1966.

[29] T. Pavlidis, *Structural pattern recognition*. Springer, 2013, vol. 1.

[30] H. Byun and S.-W. Lee, "Applications of support vector machines for pattern recognition: A survey," *Pattern recognition with support vector machines*, pp. 571–591, 2002.

[31] J. Liu, J. Sun, and S. Wang, "Pattern recognition: An overview," *IJCSNS International Journal of Computer Science and Network Security*, vol. 6, no. 6, pp. 57–61, 2006.

[32] R. Muralidharan and C. Chandrasekar, "Object recognition using svm-knn based on geometric moment invariant," *International Journal of Computer Trends and Technology*, vol. 1, no. 1, pp. 215–220, 2011.

[33] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2126–2136.

[34] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of human genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[35] W. Zimmermann, "The power counting theorem for minkowski metric," *Communications in Mathematical Physics*, vol. 11, no. 1, pp. 1–8, 1968.

[36] Online, "Demanda reprimida x demanda natural: o que você precisa saber," http://destinonegocio.com/br/negocios-online/demanda-reprimida-x-demanda-natural-o-que-voce-precisa-saber/, 2015, acessado em: 03-10-2017.

[37] J. Morrison, "How to use diffusion models in new product forecasting," *The Journal of Business Forecasting*, vol. 15, no. 2, p. 6, 1996.

[38] E. P. Pimentel, V. F. de França, and N. Omar, "A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização," in *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, vol. 1, no. 1, 2003, pp. 495–504.

[39] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[40] P. Berkhin *et al.*, "A survey of clustering data mining techniques." *Grouping multidimensional data*, vol. 25, p. 71, 2006.

[41] D. Fasulo, "An analysis of recent work on clustering algorithms," Technical report, Tech. Rep., 1999.

[42] H. Steinhaus, "Sur la division des corp materiels en parties," *Bull. Acad. Polon. Sci*, vol. 1, no. 804, p. 801, 1956.

[43] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[44] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[45] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[46] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[47] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

[48] J. C. Dunn, *A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters*. Taylor & Francis, 1973.

[49] M. Steinbach, G. Karypis, V. Kumar *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400, no. 1. Boston, 2000, pp. 525–526.

[50] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[51] G. Kirchgässner, J. Wolters, and U. Hassler, *Introduction to modern time series analysis*. Springer Science & Business Media, 2012.

[52] R. S. Ehlers, "Análise de séries temporais," *Laboratório de Estatística e Geoinformação. Universidade Federal do Paraná*, 2007.

[53] R. Bracewell, "The sunspot number series," *Nature*, vol. 171, no. 4354, pp. 649–650, 1953.

[54] S. Hylleberg, R. F. Engle, C. W. Granger, and B. S. Yoo, "Seasonal integration and cointegration," *Journal of econometrics*, vol. 44, no. 1, pp. 215–238, 1990.

[55] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2014.

[56] G. Box, G. M. Jenkins, and G. Reinsel, "Time series analysis: Forecasting & control," 1994.

[57] M. MatLab, "The language of technical computing," *The MathWorks, Inc. http://www. mathworks. com*, 2012.

[58] D. Lu, D. A. Kiewit, and J. Zhang, "Market research method and system for collecting retail store and shopper market research data," Jul. 19 1994, uS Patent 5,331,544.

[59] H. Gao, J. Yan, F. Cao, Z. Zhang, L. Lei, M. Tang, P. Zhang, X. Zhou, X. Wang, and J. Li, "A simple generic attack on text captchas," *Network and Distributed System Security Symposium (NDSS)*, 2016.

[60] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *JOSA A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[61] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Josa a*, vol. 4, no. 12, pp. 2379–2394, 1987.

[62] S. Nanda, B. Mahanty, and M. Tiwari, "Clustering indian stock market data for portfolio management," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8793–8798, 2010.

[63] H. Markowitz, "Portfolio selection," *The journal of finance*, vol. 7, no. 1, pp. 77–91, 1952.

[64] M. C. S. de Lira, "Análise de modelos de dados não relacionais e multidimensionais no contexto de big data," 2016, monografia (Bacharel em Sistemas de Informação), UFRPE (Universidade Federal Rural de Pernambuco), Recife, Brasil.

[65] A. H. Team, "Apache hbase reference guide," *Apache, version*, vol. 2, no. 10, 2015.

[66] C. Tibulo and V. D. C. Tibulo, "Previsão do preço do milho, através de séries temporais," *Scientia Plena*, vol. 10, no. 10, 2014.

[67] R. Battisti, P. C. Sentelhas, and F. G. Pilau, "Eficiência agrícola da produção de soja, milho e trigo no estado do rio grande do sul entre 1980 e 2008," *Ciência Rural*, vol. 42, no. 1, 2012.

[68] P. A. Morettin and C. Toloi, *Análise de séries temporais*. Blucher, 2006.

[69] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[70] I. Musskopf. (2016) Operação serenata de amor. [Online]. Available: https://serenatadeamor.org/

[71] M. D. M. d. Araújo, "Crowdfunding: o que as campanhas de sucesso fazem diferente? uma análise comparativa com uso de conjuntos fuzzy set," *Universidade do Vale do Rio dos Sinos*, 2017.

[72] A. C. Gil, "Como elaborar projetos de pesquisa," *São Paulo*, vol. 5, p. 61, 2002.

[73] A. N. S. Triviños, "Introdução à pesquisa em ciências sociais: a pesquisa qualitativa em educação. o positivismo; a fenomenologia; o marxismo," in *Introdução à pesquisa em ciências sociais: a pesquisa qualitativa em educação. O positivismo; a fenomenologia; o marxismo*. Atlas, 2015.

[74] A. C. F. Terence and E. Escrivão Filho, "Abordagem quantitativa, qualitativa e a utilização da pesquisa-ação nos estudos organizacionais," *Encontro Nacional de Engenharia de Produção*, vol. 26, 2006.

[75] I. CNAE, "Instituto brasileiro de geografia e estatística," *Classificação nacional das atividades econômicas.*, vol. 18, 2014.

[76] A. B. S. Neto, M. D. C. M. Batista, and T. A. E. Ferreira, "A simple approach to automatic filling captcha using pattern recognition," *International Journal of Computer Applications*, vol. 170, no. 2, pp. 1–7, Jul 2017. [Online]. Available: http://www.ijcaonline.org/archives/volume170/number2/28039-2017914669

[77] K. Doherty, R. Adams, and N. Davey, "Non-euclidean norms and data normalisation," in *Proceedings of European Symposium on Artificial Neural Networks*, 2004.

[78] A. MySQL, "Mysql documentation," *Disponível em http://www. mysql.com/doc*, 2000.

[79] K. Sierra and B. Bates, *Use a cabeça!: java*. Alta Books, 2007.

[80] E. Burnette, *Eclipse Ide Guia de Bolso*. Bookman, 2006.

[81] O. Aoki and P. R. Ormenese, "Referência debian," 2009, disponível em: http://www.debian.org/doc/manuals/reference/refere nce.pt-br.pdf acessado em: 03-10-2017.

[82] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American statistical association*, vol. 74, no. 366a, pp. 427–431, 1979.

[83] D. Barbin, *Componentes de variância: teoria e aplicações*. Fealq, 1993.

[84] A. Bifet, "Mining big data in real time," *Informatica*, vol. 37, no. 1, 2013.

[85] J. Gama, *Knowledge discovery from data streams*. CRC Press, 2010.