UNIVERSIDADE FEDERAL DO TOCANTINS

CÂMPUS UNIVERSITÁRIO DE PALMAS

CURSO DE CIÊNCIA DA COMPUTAÇÃO

RICARDO HENRIQUE SOUZA MACHADO

ANÁLISE DE DADOS DAS QUEIMADAS OCORRIDAS NOS ÚLTIMOS ANOS NO ESTADO DO TOCANTINS

PALMAS (TO)

2022

RICARDO HENRIQUE SOUZA MACHADO

ANÁLISE DE DADOS DAS QUEIMADAS OCORRIDAS NOS ÚLTIMOS ANOS NO ESTADO DO TOCANTINS

Trabalho de Conclusão de Curso II apresentado à Universidade Federal do Tocantins para obtenção do título de Bacharel em Ciência da Computação, sob a orientação do(a) Prof.(a) Dra. Glenda Michele Botelho - UFT.

Orientador: Dra. Glenda Michele Botelho - UFT

PALMAS (TO)

2022

RICARDO HENRIQUE SOUZA MACHADO

ANÁLISE DE DADOS DAS QUEIMADAS OCORRIDAS NOS ÚLTIMOS ANOS NO ESTADO DO TOCANTINS

Trabalho de Conclusão de Curso II apresentado à UFT – Universidade Federal do Tocantins – Câmpus Universitário de Palmas, Curso de Ciência da Computação foi avaliado para a obtenção do título de Bacharel e aprovada em sua forma final pelo Orientador e pela Banca Examinadora.

Data de aprovação: 15 / 12 / 2022

Banca Examinadora:

Prof. Ma. Juliana Leitão Dutra - UFT

Profa. Dra. Anna Paula de S. P. Rodrigues - UFT

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

M149a Machado, Ricardo Henrique Souza.

Análise de dados das queimadas ocorridas nos últimos anos no estado do Tocantins. / Ricardo Henrique Souza Machado. – Palmas, TO, 2022.

45 f.

Monografia Graduação - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Ciências da Computação, 2022.

Orientadora : Glenda Michele Botelho

1. Análise de dados. 2. Descoberta de conhecimento em base de dados. 3. Queimadas. 4. Tocantins. I. Título

CDD 004

Elaborado pelo sistema de geração automática de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

Dedico esse trabalho de pesquisa
aos meus pais e aos meus avós, que
sempre me deram forças para
continuar avançando a cada
obstáculo

AGRADECIMENTOS

Gostaria de agradecer aos meus pais, que sempre acreditaram em mim e na minha educação, não pouparam esforços para que eu fosse mais longe, e me apoiaram nos momentos difíceis.

Aos meus amigos, que me deram apoio e conforto em momentos difíceis, assim como muita alegria.

À minha orientadora, Glenda Michele Botelho, por ter feito o possível me auxiliando a fazer um bom trabalho.

RESUMO

No Brasil, o uso de fogo para limpeza do terreno é algo muito presente culturalmente nas comunidades rurais, embora essa prática seja crime se executada sem autorização das autoridades competentes e possa ocasionar a destruição de áreas extensas quando o fogo foge do controle. Para combatê-la, existem sistemas de monitoramento capazes de detectar com relativa precisão a ocorrência de queimadas, o que auxilia autoridades como, por exemplo, o IBAMA (Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis), no combate às queimadas e aplicação das devidas punições aos responsáveis. Neste contexto, este trabalho propõe a realização de uma análise de dados coletados referentes às queimadas ocorridas nos últimos anos no Tocantins, estado pertencente à Amazônia Legal (área que compreende a os estados do Acre, Amapá, Amazonas, Mato Grosso, Pará, Rondônia, Roraima, Tocantins e parte do estado do Maranhão),, transformando-os em informações úteis por meio da realização de uma Análise Exploratória nos dados disponíveis. Com isso, este trabalho busca compreender a magnitude do avanço dos incêndios nos últimos anos nesse território, assim como prováveis causas associadas a esse fenômeno, e por fim ressaltar ações que possam ser tomadas para que haja diminuição nas queimadas, assim como fornecer conhecimento que possa ser utilizado em estudos posteriores. Para isso, serão utilizados dados que mostram a quantidade e a localização dos incêndios fornecidos pelo INPE (Instituto Nacional de Pesquisas Espaciais) e dados meteorológicos da região da Amazônia Legal fornecidos pelo INMET (Instituto Nacional de Meteorologia).

Palavra-chave: Análise de Dados. Descoberta de Conhecimento em Base de Dados. Análise Exploratória de Dados. Queimadas. Tocantins. Amazônia Legal.

ABSTRACT

In Brazil, using fire for ground cleaning is heavily present in the culture of the countryside communities. However, this practice is a crime if done without authorization from competent authorities and may cause large areas of destruction when fire runs out of control. In order to fight these fires, there are monitoring systems able to detect with relative precision the occurrence of fires that assists authorities, for example, IBAMA (Brazilian Institute of Environment and Renewable Natural Resources), helping to fight this practice and to punish those responsible for the fires. In this context, this work proposes the execution of computational analysis of data collected about forest fires in the last years, transforming them into useful information through an Exploratory Data Analysis. Through this, we hope to identify the advance of forest fires in the previous years in the State of Tocantins, which belong to the area designated as Amazônia Legal(area that comprehends the totality of the Acre, Amapá, Amazonas, Mato Grosso, Pará, Rondônia, Tocantins states and part of Maranhão state), on the way to turn it into useful information through the Exploratory Analysis technique. This work seeks, then, to comprehend the magnitude of the advance in fires in the last years on this territory, just like the causes of such, and to highlight possible actions to reduce the amount of fire and to make knowledge available for use in future studies. For doing this, there will be used data showing the amount and location of the fires provided by INPE (National Research Institute) and meteorological data about the Amazônia Legal territory provided by INMET (Meteorology National Institute).

Keywords: Data Analysis. Knowledge Discovery in Databases. Exploratory Data Analysis. Forest Fires. Tocantins. Amazônia Legal.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

As queimadas são uma prática extensivamente utilizada por agricultores na região amazônica, tendo como principais objetivos transformar florestas em pastagens e controlar a proliferação das plantas indesejadas com baixos custos. A alternativa ao fogo é o alto investimento em máquinas pesadas capazes de remover as árvores do terreno, assim como a limpeza manual de ervas daninhas menores presentes no terreno. Dessa maneira, o fogo parece a única opção viável, uma vez que, ao contrário de suas alternativas, apresenta custo mínimo (DIAZ et al., 2002)

Muitas vezes, o uso do fogo para limpar áreas específicas pode sair do controle e chegar a áreas adjacentes. Algumas técnicas são utilizadas para evitar essa "fuga", dentre elas o aceiro, que consiste em uma faixa desprovida de vegetação, que pode ser feito tanto manualmente quanto com maquinário e, é utilizado por até 98% dos agricultores. Outra estratégia utilizada é avisar os vizinhos antes de realizar a queimada, para que se atentem a qualquer evento inesperado e ajudem no controle do fogo (ALENCAR et al., 1997). Quando essas técnicas falham e o fogo atinge áreas indesejadas ocorrem tanto perdas ambientais quanto financeiras (ALENCAR et al., 1997).

Portanto, percebe-se que os maiores problemas relacionados às queimadas advém de seu descontrole e expansão desenfreada. No entanto, mesmo a queimada controlada, na Amazônia, expõe um solo pouco fértil para o cultivo, que obtém os nutrientes da área queimada através das cinzas da floresta, mas depois de aproximadamente três anos já não está mais adequado para cultivo, fazendo com que uma nova área seja desmatada. Essa área desmatada propicia também o aumento de pragas e plantas invasoras, aumentando o uso de herbicidas e novos incêndios (CABRAL; FILHO; BORGES, 2013).

Atualmente, a forma mais prática de monitorar alterações negativas em florestas é por meio do sensoriamento remoto. Dentre as várias vantagens que levam esse método a ser o mais utilizado quando se trata do cuidado com florestas destacam-se o baixo custo pela cobertura de grandes áreas, a ausência de contato direto com o objeto de estudo e os dados obtidos, geralmente, são mais informativos (ZHICHKINA et al., 2020). Nos últimos anos vários países têm adotado métodos de registrar, com a ajuda de satélites, as modificações em suas florestas como, por exemplo, a China, o Canadá e vários países da União Européia (CHENG; WANG, 2008). No Brasil, o sistema foi criado em 1998 e fornece recortes anuais da situação da floresta amazônica, orientando eventuais ações de proteção por parte do governo. Em 2004, um novo sistema, denominado DETER (Sistema de Detecção de Desmatamento em Tempo Quase Real) foi desenvolvido, com capacidade de análises diárias de desmatamentos na região amazônica. No entanto, esse sistema era limitado a detectar áreas desmatadas com mais de 25 hectares, o que foi corrigido em uma segunda versão lançada em 2015, o DETER-B, capaz de detectar

áreas desmatadas de até 1 hectare (DINIZ et al., 2015).

O panorama atual da região amazônica é resultado das mais diferentes situações advindas das atividades econômicas desenvolvidas na região nos últimos 50 anos (MELLO; ARTAXO, 2017). Apesar de ter muitos dados registrados que ajudariam a entender melhor as modificações do espaço nos últimos anos, os dados são tantos que tornam difícil qualquer análise e uso prático dos dados coletados com sistemas de monitoramento. Nesse trabalho será utilizada a Análise Exploratória, aliada a técnicas de pré-processamento nos dados de onde as queimadas ocorreram e os dados meteorológicos associados a esses locais, de forma que esses dados serão analisados e representados de forma que possam gerar conhecimento ao fim do processo. Essa sequência de extração segue a hierarquia exemplificada na Figura 5.

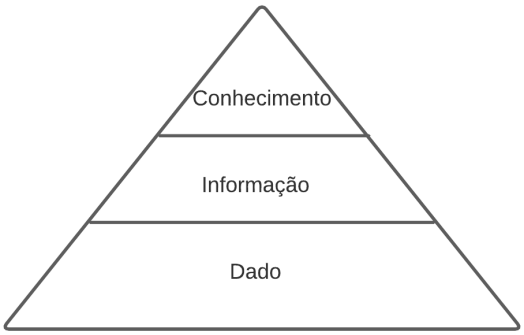

Figura 1 – Hierarquia entre dado, Informação e Conhecimento

1.1 Justificativa

As perdas ambientais decorrentes do descontrole das queimadas são gigantescas, indo desde a interrupção da sucessão ecológica, que consiste em alterações sobre a vegetação até que atinja uma situação de equilíbrio (PEREIRA-SILVA et al., 2011), até o agravamento do efeito estufa e o aumento da vulnerabilidade da floresta para futuros incêndios, a morte de animais lentos demais para escapar do fogo e a escassez de alimentos para animais que se alimentam dos frutos de árvores destruídas no processo (CABRAL; FILHO; BORGES, 2013).

Apesar da vasta quantidade de dados disponíveis referentes às queimadas na região da Amazônia Legal, eles são usados principalmente pelos órgãos fiscalizadores para punir queimadas ilegais, havendo na literatura escassez de análises que tentem entender melhor como os vários diferentes fatores se relacionam na ocorrência dessas queimadas e através do tempo. A aplicabilidade da análise proposta nesse trabalho é que o entendimento desse fenômeno tão recorrente pode servir tanto para o embasamento de políticas públicas no sentido de evitar queimadas no futuro quanto para outros estudos que tenham como ponto de partida a compreensão mais abrangente das queimadas no estado do Tocantins.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo principal deste trabalho é analisar os dados relacionados aos incêndios registrados nas últimas duas décadas no estado do Tocantins, por meio de técnicas estatísticas e análise exploratória, considerando a base de dados do Programa Queimadas do Instituto Nacional de Pesquisas Espaciais (INPE) (ESPACIAIS, 2020), assim como dados meteorológicos do Instituto Nacional de Meteorologia (INMET). Baseado nisso, espera-se encontrar possíveis relações existentes entre os vários aspectos dos focos de incêndio, principalmente os aspectos geográficos, meteorológicos e temporais atrelados a cada um.

1.2.2 Objetivos Específicos

Para atender o objetivo geral desta trabalho, os seguintes objetivos específicos devem ser alcançados:

- Utilizar métodos da área de estatística, como Média, Mediana, Moda, dentre outros, para quantificar e mostrar de forma discreta a evolução dos incêndios.
- Aplicar técnicas básicas para lidar com problemas comuns a dados brutos, como dados faltantes, inconsistentes, e advindos de diferentes fontes
- Possibilitar uma visão mais clara de como os incêndios se apresentaram no período analisado.
- Obter um panorama das áreas mais propensas a serem atingidas por incêndios
- Disponibilizar os resultados de modo a possibilitar um melhor entendimento dos incêndios no que se refere à aplicação de políticas públicas.

1.2.3 Organização do Trabalho

Esse trabalho está dividido em 5 capítulos, onde a análise exploratória de dados de queimadas e meteorológicos é abordada, sendo eles:

- Capítulo 1: apresenta a introdução do trabalho, apresentando as justificativas e objetivos do mesmo.
- Capítulo 2: discute alguns trabalhos recentes com temas pertinentes a esse estudo, como análise de dados de incêndios florestais, análise de dados geoespaciais e análise exploratória de dados.
- Capítulo 3: apresenta uma revisão bibliográfica dos temas envolvidos no estudo, como pré-processamento e técnicas comumente utilizadas para análise de dados.

- Capítulo 4: expõe a metodologia adotada na realização do trabalho, assim como os materiais utilizados para tal e as conclusões finais a respeito da execução do projeto.
- Capítulo 5: apresenta as conclusões referentes à execução do trabalho, de forma a sintetizar o conhecimento adquirido, assim como traçar ideias para trabalhos futuros.

2 ESTADO DA ARTE

As técnicas de sensoriamento remoto utilizadas nas últimas décadas apresentam muitas vantagens, sendo uma delas também um problema para o pós-processamento, ou seja, a grande quantidade e detalhamento dos dados. Para solucionar essa questão, tornam-se necessários métodos que reduzam a dimensionalidade. Sendo assim, neste capítulo serão apresentados trabalhos focados no uso do processo de descoberta de conhecimento em base de dados para lidar com essa explosão de dados. Também serão apresentados trabalhos focados na análise de dados advindos de sensoreamento remoto, especialmente os focados na detecção de queimadas.

2.1 Integrated Spatio-temporal Data Mining for Forest Fire Prediction

No trabalho de (CHENG; WANG, 2008), os autores explicam os enormes prejuízos trazidos pelas queimadas das florestas. Esses prejuízos são financeiros, pois ao atingir uma grande escala o custo para controlar uma queimada pode ser de milhões de dólares; ecológicos, pois pode causar a perda de centenas de quilômetros de vegetação, assim como pode colocar muitas vidas humanas em risco.

Diante disso, a maioria dos países vêm construindo bases de dados ou sistemas de informações com registros temporais da ocorrência de queimadas em seus territórios, de forma a gerenciar e controlar devidamente sua ocorrência. Registrando esses dados sistematicamente se torna possível entender, gerenciar e controlar as queimadas de forma mais eficiente. No processo de entender esses dados e torná-los úteis é necessário encontrar padrões profundamente ocultos, o que pode ser alcançado pela mineração de dados espaço-temporal (STDM - Spatio-temporal data mining), que é a extração de conhecimento desconhecido e implícito a partir de bases de dados espaço-temporais. O artigo busca usar a mineração de dados espaço-temporal de forma a prever a ocorrência de queimadas para então preveni-las e, enfatiza como os métodos existentes de previsão têm dificuldade em lidar com a dinâmica da mudança das queimadas em florestas através do espaço.

As técnicas de mineração espaço-temporal analisadas são a Previsão Espaço-Temporal e Análise de Tendências, Mineração de Regras de Associação Espaço-Temporais e Mineração de Padrões Sequenciais Espaço-Temporais, Mineração de Agrupamento de Caraterísticas e Regras Discriminatórias. Como o processo de previsão de queimadas é altamente dinâmico exigindo grande adaptação, o modo mais apropriado é utilizando uma Rede Neural Dinâmica Recorrente, que é uma rede neural com conexões de *feedback* em que a saída não depende apenas da entrada atual da rede, mas também de entradas e saídas anteriores, assim como do estado da rede.

Para trabalhar a previsão das queimadas usando redes neurais, os autores propõem o *Framework* de Previsão Integrada Espaço-Temporal (ISTFF) executando os seguintes passos: construção de modelos de tempo estocásticos para capturar características temporais de cada

subcomponente independente espacialmente e, posteriormente, construção de uma rede neural dinâmica recorrente para descobrir a correlação espacial entre o subcomponente alvo e os demais. Após isso, combina-se as previsões temporais e espaciais anteriores, baseadas em regressão estatística, para produzir a previsão final do subcomponente alvo.

Em suma, o artigo apresenta um *framework* de previsão espaço-temporal melhorado, introduzindo a rede neural Elman de previsão espacial, ilustrado na previsão de queimada florestal no Canadá. Chegou-se à conclusão de que o *framework* apresentado tem melhor performance de previsão se existe forte correlação entre o subcomponente alvo e os demais subcomponentes.

2.2 Statistics based predictive geo-spatial data mining: forest fire hazardous area mapping application

Os autores do artigo têm como foco principal a prevenção de incêndios florestais através da descoberta da distribuição espaço-temporal desses incêndios e previsão de áreas perigosas a partir de mapas florestais, mapas topográficos e dados do histórico dos incêndios. Um método de mineração geoespacial preditivo baseado em estatística é desenvolvido para prever a área de risco de incêndio na região de Youngdong da província de Kangwon, na Coreia do Sul. O artigo utiliza dois modelos de previsão baseados em estatística espacial, sendo eles probabilidade condicional e taxa de tendência. A probabilidade condicional é a probabilidade de que um dado pixel escolhido aleatoriamente contém uma área conhecida de início de incêndio florestal. Essa probabilidade encontrada pode ser usada, na falta de outros dados, como um indicador das chances de um incêndio florestal vir a acontecer naquela área no futuro. Já a taxa de tendência representa a razão de duas funções de distribuição espacial, a com os incêndios florestais e a sem nenhuma ocorrência. Após comparar os métodos, o método de taxa de tendência se saiu melhor nas avaliações.

2.3 A Data Mining Approach to Predict Forest Fires using Meteorological Data

O artigo apresenta uma solução com o propósito de prever a ocorrências de incêndios florestais, especificamente aplicada no parque natural Montesinho, na região Trás-os-Montes em Portugal. Existem três principais opções para analisar remotamente dados relacionados a incêndios florestais, sendo elas: baseado em satélites, *scanners* de infravermelho/fumaça e sensores locais meteorológicos. Devido aos problemas de confiabilidade dos dados de satélite disponíveis e ao alto custo dos *scanners* locais, os sensores meteorológicos acabam sendo a opção mais barata e confiável. Após várias experimentações, os autores utilizam na solução proposta apenas quatro variáveis obtidas dos dados meteorológicos: chuva, vento, temperatura e umidade, em conjunto com uma SVM (Máquina de Vetores de Suporte), capaz de prever pequenas áreas queimadas, que são a maioria.

2.4 Exploratory data analysis of activity diary data: a space–time GIS approach

Analisar os dados de movimentação diária tem se tornado um tópico importante para entender os padrões de movimentação diários dos cidadãos e poder assim alcançar um planejamento urbano melhor, em áreas como o transporte público e construção civil em geral, de forma a atender essas necessidades. Os pesquisadores (CHEN et al., 2011) analisaram os dados de pessoas em Pequim, China, já que esses dados estão cada vez mais fáceis de serem gravados e disponibilizados, com o avanço de tecnologias de monitoramento de localização geográfica em dispositivos móveis cada vez mais ampla.

O estudo é concluído com o desenvolvimento de uma extensão no software ArcGis, possibilitando a análise exploratória de dados geográficos associados ao tempo, podendo ser usado para entender movimento de pessoas, a partir de GPS de veículos, dados de localização de celulares e rastreamento de encomendas, etc.

3 REFERENCIAL TEÓRICO

Diante do aumento da capacidade de coleta de dados visto nos últimos anos, vê-se uma crescente lacuna entre essa capacidade de coleta e a capacidade dos cientistas de analisar esses dados. Isso tem tornado o processo de análise de dados tradicional, com cientistas observando-os e tirando conclusões a partir disso, cada vez mais defasado, uma vez que a quantidade de dados cresce em quantidade e dimensionalidade tão rapidamente que nem mesmo aumentar o tamanho das equipes é suficiente para manter as análises eficientes. Uma solução para esta questão é a utilização de técnicas para reduzir dados em abundância, transformando-os em quantidades menores e passíveis de serem efetivamente analisadas e interpretadas por um ser humano (FAYYAD; HAUSSLER; STOLORZ, 1996).

Quando a escala de manipulação, exploração e inferência extrapola a capacidade humana, busca-se tecnologia computacional para realizar a automação dos processos. Nesse caso, o problema da extração de conhecimento está atrelado a vários passos que envolvem desde manipulação e recuperação de dados a matemática básica e inferência estatística e, embora extrair conhecimento de dados não seja um tópico nada novo, a extração no contexto de grandes banco de dados ainda possui muito a se explorar (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

A necessidade de meios para efetivamente manipular essas grandes bases de dados computacionalmente deu origem a processos capazes de extrair tanto quanto possível de modo que o conhecimento adquirido seja útil para outras aplicações e tomada de decisões.

3.1 Análise Exploratória

Uma técnica a ser ressaltada é a Análise Exploratória, que busca olhar para os dados de tantos ângulos quanto possível em busca de recursos que sejam de algum interesse. Ela não busca um efeito em particular, e não possui um modelo matemático pronto, por isso é difícil delimitá-la e, muitas vezes, a única delimitação acaba sendo a imaginação do analista de dados. Os dados são considerados listas de números, e o analista é livre para escolher quaisquer procedimentos que queira para analisá-los, sendo o objetivo principal olhar os dados e pensar sobre eles de vários pontos de vista diferentes, o que leva às primeiras conclusões informais. Um dos primeiros passos a serem executados é transmitir as descobertas feitas com os dados para meios gráficos de visualização, com o tipo de disposição gráfica dos elementos e resultados dependendo do que foi aprendido. A partir daí conjecturas acerca dos dados podem surgir, e como a análise exploratória não está sujeita a um modelo específico para confirmação, é preciso buscar novos dados que confirmem ou neguem aquela conjectura inicial. Uma grande aliada da Análise Exploratória é a estatística, uma vez que essa pode ser entendida como a ciência dos dados. A finalidade é sempre obter dos dados a maior quantidade possível de informação, para

isso utilizando áreas como a Estatística Descritiva, que se apresenta a sintetização dos dados, utilizando gráficos, tabelas e medidas descritivas como ferramentas, e a Estatística Inferencial, que utiliza um conjunto de técnicas que permitem utilizar dados oriundos de uma amostra para generalizar ações sobre a população. Alguns conceitos estatísticos são importantes na abordagem da Análise Exploratória de Dados, como o de população, que é o conjunto de todos os indivíduos que representam pelo menos uma característica em comum cujo comportamento possa interessar, o de amostra, que pode ser definida como o subconjunto selecionado para se fazer inferência sobre as características da população.

Outro conjunto de procedimentos para a análise de dados, principalmente em grandes volumes, é o KDD (Descoberta de Conhecimento em Bases de Dados). Aliado às técnicas de *Data Mining*, ele pode levar ainda mais longe a descoberta de fatores desconhecidos intrínsecos a determinadas bases de dados. O KDD pode andar junto com a Análise Exploratória, assim como o faz com a Mineração de Dados, incorporando-as ao seu processo.

3.2 KDD

Ao utilizar a computação para extrair o conhecimento intrínseco de grandes bases de dados, é crucial definir o processo que será utilizado, uma vez que o termo *Data Mining* é usualmente utilizado em dois contextos diferentes. No primeiro contexto, *Data Mining* é o único processo responsável pela extração de conhecimento do banco de dados, o que pode levar à extração de conhecimento inútil ou mesmo inexistente, pois não há nenhuma triagem ou ajuste dos dados que serão analisados. Já no segundo caso o *Data Mining* é apenas um passo automatizado de um processo maior, o KDD.

Segundo Fayyad et al. (1996), KDD (*Knowledge Discovery in Databases*) é o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados. O processo de KDD em si consiste em usar o banco de dados junto com qualquer seleção necessária, pré-processamento, subamostragem e suas transformações, aplicar métodos de mineração de dados (algoritmos) para enumerar padrões, e avaliar os produtos de *Data Mining* para identificar o subconjunto dos padrões enumerados, considerados "conhecimento"(FAYYAD et al., 1996). O processo de KDD lembra, em alguns aspectos, o próprio desenvolvimento de *software*, pois possui uma considerável quantidade de níveis, fases e passos, além de várias técnicas alternativas poderem ser usadas, possuir muita iteração e a adição ou atualização dos dados interferir na execução do processo (ZHONG et al., 1997). Esse processo é tanto interativo, pois requer a interação com o usuário para atingir seus objetivos de forma mais acertada, quando iterativo, uma vez que requer vários passos, os quais são influenciados por decisões realizadas pelo usuário (FAYYAD et al., 1996).

3.2.1 Passos do KDD

Os passos do KDD são vários, embora alguns sejam muito semelhantes. Antes de tudo, é essencial o entendimento do domínio da aplicação e do pré-conhecimento relevante para atingir seu objetivo, assim como identificar qual exatamente é esse objetivo, podendo tal entendimento ser definido como o primeiro passo (FAYYAD et al., 1996). Em seguida, se inicia o trabalho de preparação dos dados, o pré-processamento. Antes de qualquer análise, uma etapa extremamente necessária é o pré-processamento, um dos pontos principais desse trabalho. Ele consiste de uma série de passos onde os dados brutos são transformados de forma a estarem "limpos"e organizados para serem apropriadamente utilizados posteriormente. O primeiro passo, de limpeza, se faz necessário pois no mundo real os dados geralmente estão "bagunçados", incompletos e com "ruído", no sentido de que podem existir valores incorretos e inconsistentes.

Para lidar com a situação onde há dados faltando, existem várias possibilidades. A primeira delas é simplesmente ignorar os dados faltantes, o que pode ser problemático se a porcentagem variar consideravelmente nos registros. Outra opção é preencher manualmente os valores faltantes com valores que façam sentido e não irão afetar a análise final, o que pode ser muito trabalhoso ou até inviável para uma quantidade muito grande de valores. A terceira opção é preencher com um valor esperado, seja esse valor a média dos demais valores ou o valor obtido por algum método preditivo, o que pode muitas vezes tornar os dados tendenciosos, dependendo do quão errados estão os valores inseridos.

Já para lidar com dados com "ruído", ou incorretos, alguns métodos básicos são: categorização, onde os dados são separados em categorias de acordo com os valores em sua volta, e no final os valores mais discrepantes não são considerados na análise, apenas as categorias em si; agrupamento, onde os valores são agrupados de acordo com sua similaridade, fazendo com que os valores discrepantes sejam eliminados; e *Machine Learning*, onde técnicas como análise regressiva, por exemplo, são utilizadas para normalizar os dados.

Para lidar com dados inconsistentes ou duplicatas, na maioria das vezes é possível utilizar correções manuais, uma vez que se conheça o tipo de dado que se está trabalhando. Os registros de temperatura, por exemplo, são fáceis de filtrar retirando temperaturas acima do que é possível de ser registrado.

Após essa limpeza inicial, há a parte de integrar dados advindos de inúmeras fontes diferentes de forma que fiquem juntos e façam sentido, o que pode gerar um grande número de desafios.

Após esses dados estarem agregados, no caso de grandes volumes de dados, como é o caso desse trabalho, é necessário também reduzi-los de forma a serem mais facilmente processados e analisados. Para isso pode ser que parte dos dados seja perdida ou não, tendo em vista o objetivo final de que a base de dados resultante seja mais efetiva para análise e ainda coerente com sua versão inicial (MALLEY; RAMAZZOTTI; WU, 2016).

A próxima etapa é a mineração de dados, essencial para a descoberta de conhecimento,

na qual métodos inteligentes são aplicados para extrair padrões de dados. Podem ser procurados tanto uma forma de representação específica quanto um conjunto dessas representações, como regras de classificação ou árvores, regressão, agrupamento, modelagem de sequência etc. A eficiência das técnicas de mineração de dados é melhorada quando combinadas com o usuário executando corretamente os passos anteriores, e deve sempre ser observado se determinada técnica de mineração de dados é a mais adequada para a aplicação pretendida, uma vez que é fato que não há uma técnica que seja adequada para todo e qualquer cenário (BENOMRANE; AYED; ALIMI, 2013).

Como os dados já foram previamente limpos, integrados em um banco de dados e transformados de forma a serem usados na mineração, executa-se o algoritmo de mineração de dados e, então, a representação do conhecimento encontrado por meio de técnicas de visualização. Esses dados minerados que são representados podem ser arquivos planos, *Data Warehouses* (repositório de dados coletados de várias fontes, muitas vezes heterogêneas), bancos de dados com registros de transações, bancos de dados multimídia, dentre outros.

A mineração de dados é um processo iterativo e, após ter o conhecimento apresentado para o usuário, pode ser regulado para se obter melhores resultados na próxima iteração (LAKSHMI; RAGHUNANDHAN, 2011). Depois dessas iterações e de ter sido apresentado para o usuário, é necessário consolidar o conhecimento, seja incorporando-o a outro sistema ou simplesmente documentando-o, sendo essa a etapa final do KDD (FAYYAD et al., 1996). A Figura 5 representa o processo de KDD.

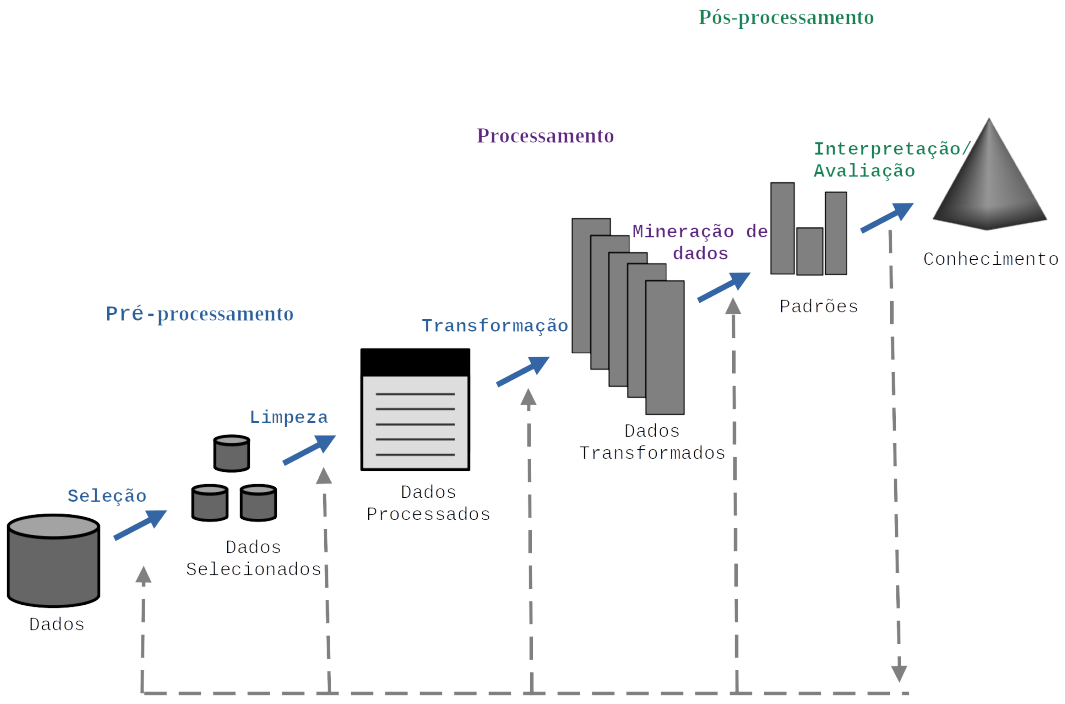

Figura 2 – Passos da descoberta de conhecimento. Figura adaptada de Fayyad, Piatetsky-Shapiro e Smyth (1996a).

A mineração de dados é um passo decisivo para alcançar sucesso na descoberta de conhecimento e, seu objetivo é definido ou pela criação de um modelo descritivo, que apresenta de forma concisa as principais características do conjunto de dados, ou pela criação de um modelo preditivo, que busca prever um valor desconhecido, geralmente do futuro, de uma variável específica. Se o valor buscado é uma classe (número predefinido discreto), a tarefa de mineração de dados é chamada de classificação, e se o valor buscado é um número real, é chamada regressão. (JAIN; SRIVASTAVA, 2013).

3.2.2 Algoritmos de mineração de dados

Na Mineração de Dados existem 3 tipos diferentes de associar objetos, conhecidos como classificação, regressão e clusterização. A **classificação**, como o próprio nome diz, separa a informação em classes, e a caracterização dessas classes pode ser utilizada para supor detalhes sobre a nova informação sem classificação. A classificação possui duas fases, que são a fase de aprendizado, em que há a análise dos dados de treinamento e a criação de regras e padrões, e a segunda fase, que usa o conhecimento obtido para classificar novos dados. A regressão mapeia dados em uma variável de grande valor preditivo, e pode ser utilizada para mostrar a conexão entre um ou mais fatores livres e dependentes. Por fim, a clusterização não tem classes e agrupa os dados em *clusters*, ou grupos, de itens de alguma forma semelhante (GURUVAYUR; SUCHITHRA, 2017).

Os algoritmos ou técnicas de classificação mais comuns são:

- **Árvore de decisão:** É uma estrutura recursiva usada com o objetivo de expressar regras de classificação (QUINLAN, 1990), sendo implementada como um grafo em que cada nó simboliza um teste no valor de um atributo, cada ramo simboliza um resultado do teste e as folhas representam as classes ou a distribuição de classes (JAIN; SRIVASTAVA, 2013). Essa árvore é uma árvore direcionada e possui um nó raiz, sem nenhum ramo de entrada nele, apenas de saída, ao passo que todos os outros nós tem um ramo de entrada. Os nós que possuem ramos de saída são chamados de nós internos, e os nós sem ramos de entrada são chamados de folhas, ou nós de decisão (ROKACH; MAIMON, 2005). O processo de classificação pela árvore de decisão começa com um universo de objetos, pertencentes a determinadas classes, sendo as propriedades desses objetos conhecidas através dos valores de seus atributos. Esses atributos podem ser discretos, ou seja, retirados de um conjunto limitado de valores possíveis, ou contínuos, com os valores sendo valores reais (QUINLAN, 1990).

- **Rede bayesiana:** Modelo gráfico para conexões entre um arranjo de diferentes componentes variáveis. Consiste em um grafo acíclico e cada um dos nós estão em correspondência coordenada com os componentes de um conjunto de dados. Esse método tem mostrado grande exatidão e velocidade quando conectado a vastos bancos de dados (GURUVAYUR; SUCHITHRA, 2017).

- **Máquina de Vetor de Suporte:** É um algoritmo de treinamento de classificação que prepara o classificador para pressupor a classe da nova amostra (GURUVAYUR; SUCHITHRA, 2017). Ele maximiza uma função matemática particular em relação a uma dada coleção de dados(NOBLE, 2006).

- **Rede neural:** É um modelo matemático ou computacional que busca simular as redes neurais formadas por inúmeros neurônios presentes no sistema biológico dos seres vivos. Um motivo para utilizar esse algoritmo na etapa de mineração de dados é que ele fornece um viés indutivo mais adequado do que outros algoritmos. Para entender o significado desse viés indutivo deve-se pensar que para cada algoritmo de aprendizado, infindáveis modelos podem resultar da análise dos dados, tendo cada algoritmo um viés indutivo relativo aos modelos que mais provavelmente serão retornados. Esse viés indutivo de um algoritmo pode ainda ser subdividido em dois aspectos: seu viés de espaço de hipóteses restrito e seu viés de preferência. O viés de espaço de hipóteses restrito se refere às limitações de um algoritmo de aprendizado ligadas às hipóteses que ele é capaz de construir, enquanto o viés de preferência se refere à ordem de preferência em que os modelos são postos dentro do espaço de hipóteses (CRAVEN; SHAVLIK, 1997).

A **clusterização** não tem classes e é um procedimento de agrupar conjuntos de dados em *clusters*, ou grupos, de forma que os objetos em cada um tenham alta similaridade, enquanto são extremamente diferentes dos presentes em outros grupos. Algoritmos de clusterização são usados para organizar dados, categorizá-los, comprimi-los e construir modelos, detectar valores atípicos, dentre outros (GURUVAYUR; SUCHITHRA, 2017).

Já os algoritmos mais de comuns de clusterização são:

- **K-means:** É a técnica de agrupamento mais famosa atualmente em operações lógicas e mecânicas. Seu nome se dá porque são separados um número k de *clusters* por média ou média ponderada de acordo com os focos de cada grupo, chamado centroide. Não funciona tão bem para todas as abordagens, mas funciona bem com dados numéricos. A função alvo é a distância total entre uma coleção de pontos e seu centroide (GURUVAYUR; SUCHITHRA, 2017).

- **Clusterização hierárquica:** Consolida objetos em *clusters* e agrupa esses *clusters* em grupos ainda maiores, criando uma hierarquia, que pode ser representada em uma árvore chamada dendrograma. Objetos isolados são as folhas da árvore e os nós internos são *clusters* não vazios (GURUVAYUR; SUCHITHRA, 2017).

Por fim, a **regressão** estuda o relacionamento entre uma variável chamada de dependente, Y, e uma ou mais variáveis independentes, X (X_1, X_2, X_p) (RODRIGUES, 2012). Ela é similar à classificação, com a diferença de ser utilizada para valores numéricos ao invés de categóricos, podendo assim estimar o valor de uma variável analisando os valores das demais (CAMILO; SILVA, 2009).

- **Regressão linear:** É a forma mais simples de regressão, e modela uma variável aleatória Y chamada variável de resposta, como uma função linear de outra variável X que é chamada variável preditiva (RATHI, 2010).
- **Regressão não-linear:** Semelhante à regressão linear, mas a relação entre uma variável e a outra é representada por uma variável não-linear.

3.2.3 Pré-processamento

4 MATERIAIS E MÉTODO

Esse capítulo tem como objetivo apresentar a metodologia proposta para o desenvolvimento deste trabalho. Seguindo o processo de Análise Exploratória, se torna importante olhar pelo máximo de ângulos possíveis para o objeto de estudo no qual será realizada a análise dos dados. A aplicação se trata das queimadas na Amazônia Legal, focando especialmente no estado do Tocantins. Para isso, obteve-se dados de registros de incêndios ocorridos na Amazônia Legal entre os anos 2000 e 2021, fornecidos pelo Programa Queimadas do Instituto Nacional de Pesquisas Espaciais (INPE), assim como os dados meteorológicos, também da Amazônia Legal, entre os anos de 2000 e 2021, fornecidos pelo Instituto Brasileiro de Meteorologia (INMET). A seção 4.1 trata das bases de dados e ferramentas utilizadas na realização do projeto, enquanto a seção 4.2 reúne os procedimentos a serem realizados nesses dados de modo a aplicar o processo de KDD e obter resultados a partir dele.

4.1 Materiais

4.1.1 Bases de dados

Dentre os dados históricos disponíveis para análise das queimadas na Amazônia Legal, duas bases de dados foram escolhidas, sendo uma delas referente aos pontos geográficos onde os focos de incêndio foram identificados e outra referente às condições climáticas durante o mesmo período na região objeto de estudo, conforme detalhado nas subseções abaixo.

4.1.1.1 Banco de Dados de Queimadas - BDQueimadas

O BDQueimadas é uma aplicação parte do programa Queimadas do INPE, focado em disponibilizar dados geoespaciais referentes às queimadas ocorridas no Brasil.O sistema também possui dados de risco de fogo, meteorologia e coleções de limites político-administrativos (Países, Estados, Municípios), Unidades de Conservação, Biomas Brasileiros, Áreas industriais, Terras Indígenas e Polígonos de Desmatamento (PRODES -INPE).

A aplicação permite a exportação dos dados nos formatos CSV (Valores Separados por Vírgula), GeoJSON (Formato de dados geoespaciais estruturados), KML (Formato de marcação de dados Geográficos do Google Earth) e Shapefile (arquivos de marcação de formas geográficas)(SETZER; MORELLI; SOUZA, 2019). O formato de exportação utilizado para essa pesquisa foi CSV. Como a exportação dos dados somente permite um intervalo máximo de 366 dias, cada ano foi exportado em um arquivo CSV/Tabela diferente, onde cada linha contém informações de um foco de incêndio. A Tabela 1 apresenta os atributos relacionados a cada registro de incêndio nos dados disponibilizados pelo INPE.

No entanto, apesar dos atributos supostamente estarem presentes nos dados, os dados

Tabela 1 – Atributos presentes nos dados de incêndios do INPE

Atributo	Descrição
Data/Hora	Data e hora em que foi registrado o foco de incêndio
Satélite	Nome do algoritmo utilizado e referencia ao satélite provedor da imagem
País	Pais em que o foco de incêndio foi identificado
Estado	Estado em que o foco de incêndio foi identificado
Município	Município em que o foco de incêndio foi identificado
Bioma	Bioma em que o foco de incêndio foi identificado
N. Dias Sem Chuva	Quantidades de dias sem chuva que o local do incêndio estava quando ocorreu
Precipitação	Valor da precipitação acumulada no dia até o momento da detecção do foco
Risco Fogo	O Risco de Fogo (RF) é um produto que apresenta a suscetibilidade da vegetação para sua queima, do ponto de vista meteorológico
Latitude	Latitude do centro do pixel de fogo ativo apresentada em unidade de graus decimais
Longitude	Longitude do centro do pixel de fogo ativo apresentada em unidade de graus decimais
FRP	FRP: "Fire Radiative Power" ou Potência Radiativa do Fogo, mede a energia radiante liberada por unidade de tempo, associada à taxa de queima da vegetação, em MW - megawatts.

de Precipitação, Número de Dias sem Chuva, Risco Fogo e FRP praticamente não possuem nenhum dado. Logo, foi necessário considerar outras fontes de dados em busca de mais detalhes sobre a ocorrência dos incêndios.

4.1.1.2 Banco de Dados Meteorológicos (BDMEP) do INMET

Como a base de dados de queimadas não possui as informações meteorológicas relacionadas aos incêndios florestais como esperado, foram buscados os dados do INMET (Instituto Nacional de Meteorologia) e obtidos os dados meteorológicos da região da Amazônia Legal no mesmo período dos dados obtidos no Programa Queimadas, entre 2000 e 2021.

Os dados meteorológicos fornecidos pelo INMET são divididos em estações meteorológicas, onde é possível obter tanto dados por hora, quanto por dia ou por mês. Para esse trabalho, optou-se pelos dados diários, uma vez que é mais simples associar o incêndio ao clima do dia em que ocorreu. As estações meteorológicas dividem-se em dois tipos: Estações Convencionais e Estações Automáticas (ou Telemétricas). Cada uma possui suas peculiaridades, vantagens, desvantagens, além de diferentes atributos presentes nos dados. A Figura 3 mostra a localização das Estações Convencionais (em azul) e das Estações Automáticas (em verde) no estado do Tocantins, e a Tabela 2 mostra a relação das estações e a data em que iniciaram suas operações.

Tanto a Figura 3 quanto todas as outras Figuras de mapas com pontos de marcação, foram geradas em Python utilizando as bibliotecas Geopandas e Matplotlib, que utilizando as marcações geográficas de Latitude e Longitude e os Arquivos de forma geográfica disponibilizados pelo IBGE, marcou no mapa os pontos de interesse e com as cores desejadas, seja pontos maiores para representar a localização das poucas estações meteorológicas, seja com pontos maiores para representar a grande quantidade de incêndios na 6, seja com alterações nos parâmetros do mapa para mostrar um mapa de calor como na 9.

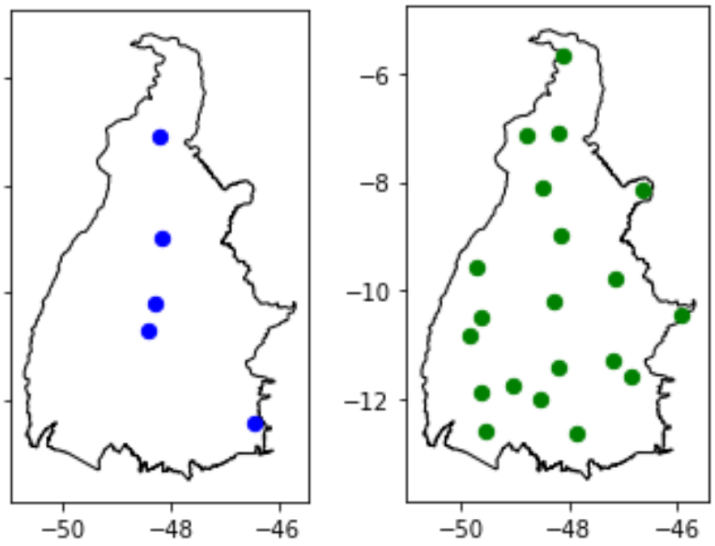

Figura 3 – Localização das Estações Convencionais (esquerda) e Automáticas (direita) no estado do Tocantins

As Estações Convencionais são instalações com uma diversidade de equipamentos, nos quais os profissionais realizam uma variedade de medições das condições meteorológicas. Como esses equipamentos existem desde anos atrás, essas estações estão presentes a muitas décadas e possuem uma série de dados históricos muito mais ampla do que a presente nas Estações Automáticas. Os dados medidos pelas Estações Convencionais possuem os atributos apresentados na Tabela 3.

Tabela 2 – Relação das Estações Convencionais e Automáticas no estado do Tocantins

ESTAÇÃO	TIPO	INÍCIO DE OPERAÇÃO
Palmas	Convencional	07/10/1993
Araguaína	Convencional	20/08/1984
Pedro Afonso	Convencional	03/03/1977
Porto Nacional	Convencional	31/12/1914
Taguatinga	Convencional	21/12/1915
Paranã	Automática	03/03/2005
Peixe	Automática	30/11/2006
Gurupi	Automática	19/12/2006
Pedro Afonso	Automática	22/01/2007
Araguaína	Automática	25/01/2007
Dianópolis	Automática	29/08/2008
Formoso do Araguaia	Automática	26/04/2008
Mateiros	Automática	24/10/2012
Marianópolis do Tocantins	Automática	23/11/2012
Campos Lindos	Automática	28/10/2012
Araguatins	Automática	11/03/2009
Santa Fé do Araguaia	Automática	29/07/2016
Colinas do Tocantins	Automática	10/11/2016
Rio Sono	Automática	02/08/2016
Santa Rosa do Tocantins	Automática	13/08/2015
Almas	Automática	31/03/2016
Araguaçu	Automática	18/08/2015
Lagoa da Confusão	Automática	14/11/2016

Já as Estações Automáticas consistem em uma torre equipada com sensores capazes de captar dados climáticos de forma contínua, sem interferência humana. Elas apresentam vantagens como, por exemplo, a consistência dos dados, já que não precisam de operação de profissionais para funcionar, o que possibilita também que haja a medida de dados durante a noite, enquanto nas estações convencionais é feita uma estimativa de certos dados que não podem ser medidos fora do horário de trabalho de um profissional na estação. Em contrapartida, as Estações Automáticas podem não ter a mesma precisão das estações convencionais, por utilizar sensores mais simples e acoplados e, como a tecnologia usada é relativamente recente, a maioria não tem mais de duas décadas de dados e, portanto, não pode ser usada para análise de dados de datas anteriores a isso. Os dados medidos pelas estações automáticas possuem os atributos apresentados na Tabela 4.

Os dados também foram obtidos em CSV e cada arquivo/tabela possui os dados de uma estação, de determinado município, onde cada linha possui as condições meteorológicas mais relevantes do dia. Além desses dados, foram obtidas também duas tabelas fornecidas pelo INMET, com o catálogo dos dois tipos de estações. Essas tabelas contém o Município da Estação, o Estado, a Situação (Operante ou em Pane), a Latitude, a Longitude, a Altitude, a

Tabela 3 – Atributos presentes nos dados de Estações Convencionais

ATRIBUTO	DESCRIÇÃO
Data Medicao	Data em que a medição foi realizada
EVAPORACAO DO PICHE, DIARIA(mm)	Medida diária do poder evaporante do ar à sombra
INSOLACAO TOTAL, DIARIO(h)	Intervalo total de tempo (entre o nascimento e o por do sol) em que o disco solar não esteve oculto por nuvens ou fenômenos atmosféricos de qualquer natureza
PRECIPITACAO TOTAL, DIARIO(mm)	Água proveniente do vapor de água da atmosfera depositada na superfície terrestre sob qualquer forma: chuva, granizo, neblina, neve, orvalho ou geada.
TEMPERATURA MAXIMA, DIARIA(^{o}C)	Temperatura máxima diária
TEMPERATURA MEDIA COMPENSADA, DIARIA(^{o}C)	Média entre 5 valores: temperatura das 9h, 15h, 21h e a máxima e a mínima
TEMPERATURA MINIMA, DIARIA(^{o}C)	Temperatura mínima diária
UMIDADE RELATIVA DO AR, MEDIA DIARIA(%)	Relação entre quantidade de água que existe no ar (umidade absoluta) e quantidade máxima de água que poderia existir na mesma temperatura (ponto de saturação)
VENTO, VELOCIDADE MEDIA DIARIA(m/s)	Velocidade média diária do vento

Data de início de Operação e o Código da Estação.

4.1.2 Softwares e ferramentas computacionais

Foram usadas ferramentas mistas para tratar e visualizar os dados. O Software *KNIME Analytics Platform* foi utilizado para etapas mais triviais da limpeza dos dados, evitando a necessidade de se fazer tudo através de programação. Como exemplos de sua utilização destaca-se a união de dados de vários anos (de todos os bancos de dados), a eliminação de linhas vazias nas tabelas, a ordenação dos dados por ordem cronológica e remoção de colunas desnecessárias ou vazias.

Também foi utilizada a plataforma *Jupyter Notebook*, em um ambiente Anaconda, para toda a parte envolvendo programação. Na linguagem Python, foram utilizadas várias bibliotecas voltadas à análise de dados, como Pandas (para processamento de dados em geral), Geopandas (para processamento de dados geográficos), Matplotlib (para exibição de gráficos em geral), Scipy (para análises espaciais) e Sklearn (para preenchimento de dados). Para plotagem de mapas foram utilizados os arquivos do tipo Shapefile fornecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

Tabela 4 – Atributos presentes nos dados de Estações Automáticas

ATRIBUTO	DESCRIÇÃO
Data Medicao	Data em que a medição foi realizada
PRECIPITACAO TOTAL, DIARIO(AUT)(mm)	Água proveniente do vapor de água da atmosfera depositada na superfície terrestre sob qualquer forma: chuva, granizo, neblina, neve, orvalho ou geada.
PRESSAO ATMOSFERICA MEDIA DIARIA (AUT)(mB)	Pressão exercida pela atmosfera sobre qualquer superfície, em virtude de seu peso.
TEMPERATURA DO PONTO DE ORVALHO MEDIA DIARIA (AUT)(C)	A temperatura na qual o ar deve ser esfriado a uma pressão constante para ser saturado.
TEMPERATURA MAXIMA, DIARIA (AUT)(C)	Temperatura máxima diária
TEMPERATURA MEDIA, DIARIA (AUT)(C)	Média entre 5 valores: temperatura das 9h, 15h, 21h e a máxima e a mínima
TEMPERATURA MINIMA, DIARIA (AUT)(C)	Temperatura mínima diária
UMIDADE RELATIVA DO AR, MEDIA DIARIA (AUT)(%)	Relação entre quantidade de água que existe no ar (umidade absoluta) e quantidade máxima de água que poderia existir na mesma temperatura (ponto de saturação)
UMIDADE RELATIVA DO AR, MINIMA DIARIA (AUT)(%)	Relação entre quantidade de água que existe no ar (umidade absoluta) e quantidade máxima de água que poderia existir na mesma temperatura (ponto de saturação)
VENTO, RAJADA MAXIMA DIARIA (AUT)(m/s)	Mudanças bruscas na velocidade do vento em um pequeno intervalo de tempo
VENTO, VELOCIDADE MEDIA DIARIA (AUT)(m/s)	Velocidade média diária do vento

4.2 Métodos

Seguindo a abordagem da Análise Exploratória, buscou-se a melhor maneira de selecionar os dados de forma que o resultado final fosse a melhor visualização possível. Optou-se por usar dados de 2000 à 2021, por ser um período extenso o suficiente para obter conclusões, mas não extenso demais, visando o foco nos acontecimentos da atualidade. Além disso, optou-se, em um primeiro momento, limitar a análise às queimadas do Estado do Tocantins, visto a quantidade e complexidade dos dados necessários para lidar com uma análise completa de toda a Amazônia Legal.

Após ler os arquivos no formato CSV com os dados de incêndios entre 2000 e 2021, foram filtrados os incêndios do estado do Tocantins, e as colunas de longitude e latitude fo-

ram unificadas em apenas uma, em um formato adequado para que a biblioteca Geopandas identificasse cada coordenada. O mesmo processo foi executado nas tabelas com os dados de localização das estações, obtendo uma coluna única com as coordenadas da localização das estações. Após isso, as duas tabelas (com a localização das estações automáticas e convencionais) foram unificadas em uma só, com a localização de todas elas.

Tendo os dados sobre as estações em si todos unificados em uma única tabela, foi executada uma operação onde as duas tabelas, com os dados dos incêndios e com os dados das estações, foram associadas em uma só, de modo que cada incêndio fosse associado à estação meteorológica mais próxima dele. Isso foi feito para obter os dados meteorológicos que mais representem as condições em que ocorreu o incêndio, mas essa etapa apenas adicionou uma coluna com o código da estação mais próxima a cada incêndio, sem os dados de fato.

Para fazer a junção de todos os dados meteorológicos, os atributos que ambos os tipos de estação possuem em comum foram mantidos e renomeados, ao passo que os que diferem, foram excluídos. Os atributos mantidos foram: Data Medição, Precipitação Total, Temperatura Máxima, Temperatura Média, Temperatura Mínima e Umidade Relativa do Ar.

Enquanto isso, os dados das estações meteorológicas foram lidos, tanto os advindos das Estações Convencionais quanto Automáticas, separadamente. Os dados das Estações Automáticas não apresentaram lacunas, ao contrário dos advindos das Estações Convencionais, onde a maioria das estações apresentava menos de 1% de dados faltantes, com algumas poucas apresentando por volta de 10%. Para preencher essas lacunas foi utilizada a função *Simple Imputer* da biblioteca *sklearn*, que preencheu as lacunas com a média dos valores. Além disso, todos os dados de todas as estações sendo analisadas foram unidos em uma única tabela, sendo adicionada uma coluna identificando o código daquela estação específica.

Com as tabelas dos dados de incêndios com a data de sua ocorrência e o código de sua estação meteorológica mais próxima, e os dados das estações meteorológicas com a data de cada medição e seu código, as duas tabelas foram unidas em uma só, obtendo finalmente tanto a data e localização de cada incêndio como também dados sobre o clima naquele mesmo dia e mesmo local, possibilitando análises a partir disso. O processo de associação de todos os dados em uma tabela final contendo todo o objeto de estudo está esquematizado na Figura 4.

Após a unificação de todas as bases de dados foram elaborados gráficos e tabelas para visualizar o processo de ocorrência dos incêndios nos últimos anos, buscando acompanhar a evolução e as relações entre os dados disponíveis, para um melhor entendimento desse processo. A metodologia completa adotada no trabalho é explicitada na Figura 5.

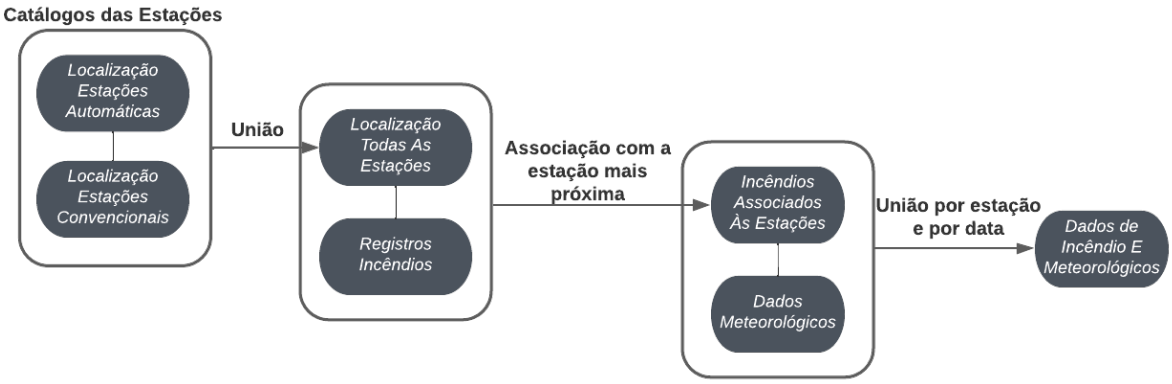

Figura 4 – Processo de unificação das bases de dados.

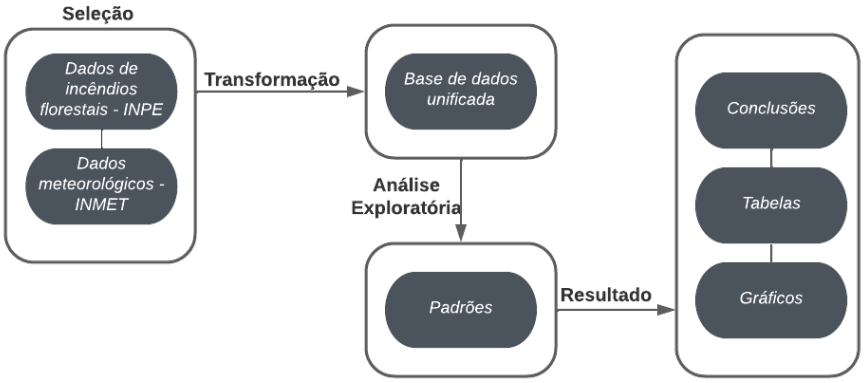

Figura 5 – Metodologia adotada no trabalho

4.3 Resultados

Este capítulo tem por objetivo apresentar os resultados das análises dos dados de queimadas do estado do Tocantins, considerando também dados meteorológicos. Uma das questões mais importantes a ser visualizada em posse de todos os dados é onde e quando houve a maior prevalência de incêndios nos últimos 21 anos. Os mapas na Figura 6 mostram os pontos onde esses incêndios ocorreram e dão uma noção de como foi essa evolução através dos anos. Percebe-se que na primeira década, a ocorrência de queimadas foi mais dispersa, podendo ser vista uma vasta área onde houve pouca ou nenhuma ocorrência de incêndios durante o ano. Já após o ano de 2010 é visível o aumento significativo nos registros, de forma que em alguns anos houveram pouquíssimas regiões do estado onde não houveram incêndios. No entanto, percebe-se uma pequena quebra nesse aumento desenfreado nos incêndios no ano de 2020, onde novamente há uma concentração mínima. Um dos fatores que ajudam a explicar essa redução em 2020 é o fato de que o Instituto Natureza do Tocantins (NATURATINS), responsável, dentre outras coi-

sas, por combater os incêndios no estado, ampliou entre 2019 e 2020 sua brigada de combate a incêndios em um experimento. Enquanto normalmente as brigadas de combate a incêndio são contratadas apenas nos meses de maior seca, entre 2019 e 2020 eles foram contratados durante os 12 meses, o que possibilitou que muitas ações preventivas e educativas fossem desenvolvidas quando as equipes não estavam atuando diretamente no combate ao fogo, o que levou a uma grande diminuição nas ocorrências (SEMARH, 2021) (TOCANTINS,).

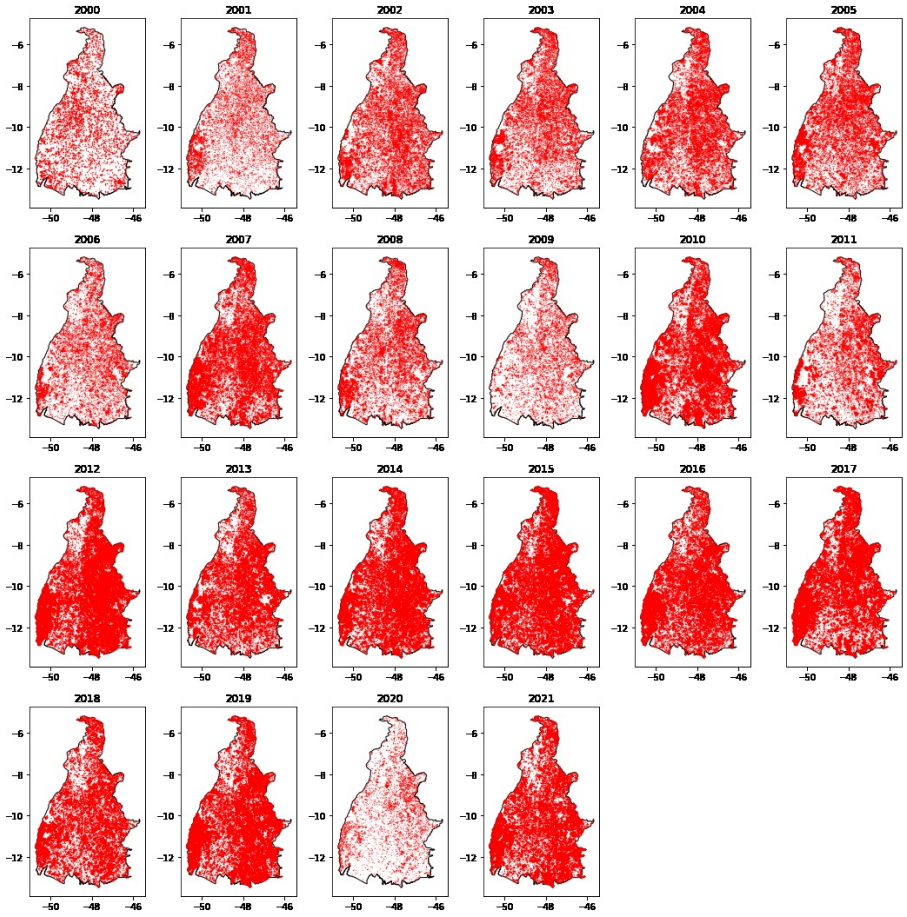

Figura 6 – Incêndios no Tocantins entre 2000 e 2021

Pode-se observar também que em todos os anos a ocorrência de incêndios na região noroeste do estado é inferior em relação ao restante do estado. Se compararmos com o mapa dos biomas do Tocantins, na Figura 7, percebe-se que essa área coincide com a parte do estado composta pelo bioma Amazônia, conhecido pela umidade e florestas mais fechadas, o que dificulta muito a propagação do fogo quando comparado ao Cerrado, predominante no restante do estado. Esse aspecto do bioma é corroborado se consideramos a umidade associada a cada incêndio que ocorreu entre 2000 e 2021, na Figura 8, onde é possível ver uma divisão clara na umidade registrada na região norte e sul do estado.

Associada a essa umidade há também a temperatura máxima em que os incêndios ocorrem, pois como pode-se observar na Figura 9, na mesma região onde há mais umidade a temperatura máxima de ocorrência de incêndios é menor. Isso não quer dizer necessariamente que

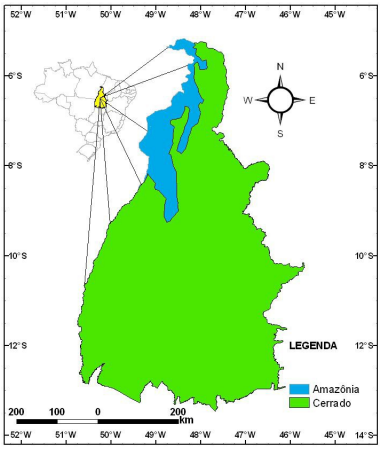

Figura 7 – Biomas presentes no Estado do Tocantins. Figura adaptada de (MARCUZZO; GOULARTE, 2013)

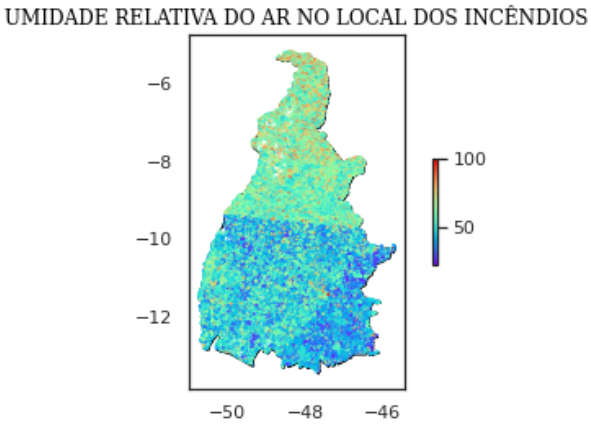

Figura 8 – Umidade relativa do ar no local dos incêndios, no período de 2000 e 2021, no estado do Tocantins.

os incêndios ocorrem com baixa temperatura local e, sim que, devido à alta umidade e baixa temperatura, há uma redução significativa na quantidade de incêndios, como pode ser visto na Figura 6, e os incêndios que ainda ocorrem são com a temperatura mais baixa.

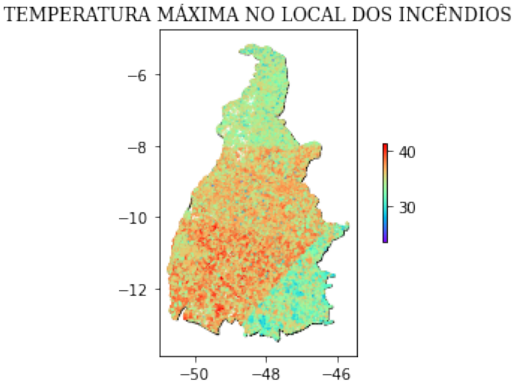

Figura 9 – Temperatura associada aos incêndios no estado do Tocantins.

Um aumento na quantidade de incêndios pôde ser visto na Figura 10, onde na primeira

década desse século os picos de incêndio em um único mês chegam no máximo a quase 50 mil, ao passo que na segunda década o pico é quase triplicado, chegando próximo a 140 mil.

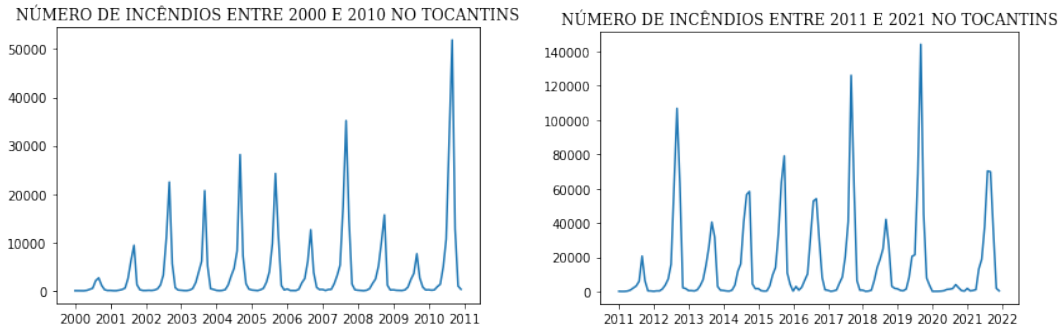

Figura 10 – Incêndios no Tocantins entre 2000 e 2021, por ano.

Além do aumento do número de incêndios de modo geral com o passar dos anos, outra forma de analisar é referente à sazonalidade com que eles ocorrem. Na Figura 11 pode-se observar a predominância dos registros de incêndios no mês de setembro, ano após ano, havendo leves variações de um mês antes ou depois. Com isso, pode-se perceber, além da relação geográfica, a relação temporal entre alguns fatores climáticos e o número de incêndios. Existem algumas relações claras entre os fatores climáticos e esse pico anual de incêndios, com alguns acontecimentos levando a outros. Começando pela precipitação, observa-se que desde maio há uma queda brusca no volume, chegando próximo a 0 e permanecendo assim todos os anos nos meses de Junho, Julho e Agosto. A Figura 12 apresenta a precipitação mensal no Estado do Tocantins, no período de 2000 a 2021.

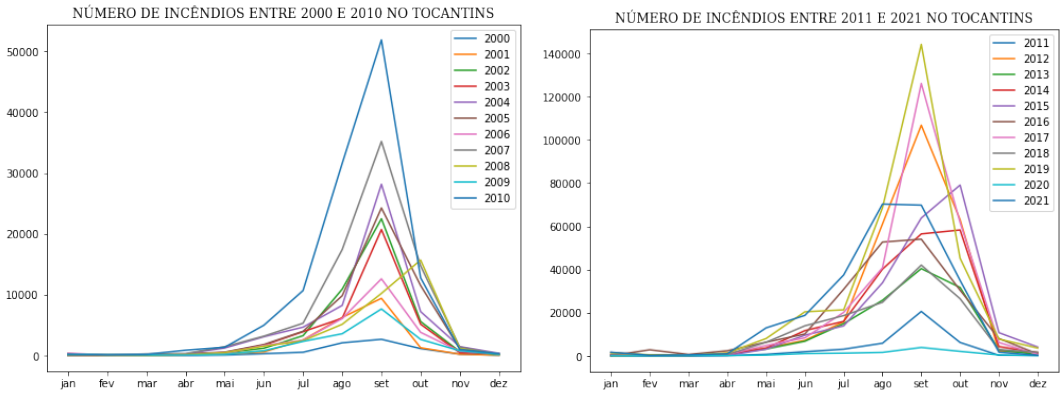

Figura 11 – Incêndios no Tocantins, entre 2000 e 2021, por mês.

Essa queda na precipitação naturalmente acarreta na diminuição da umidade relativa do ar, cuja queda, como vemos na Figura 13, ocorre um pouco após a queda na precipitação, uma vez que é consequência da mesma.

Com esse cenário montado, outro fator importante na propagação de incêndios é o vento que, embora com bem menos constância que os fatores vistos anteriormente, no geral apresenta aumento de velocidade no período entre Junho e Outubro, conforme pode ser observado na Figura 14.

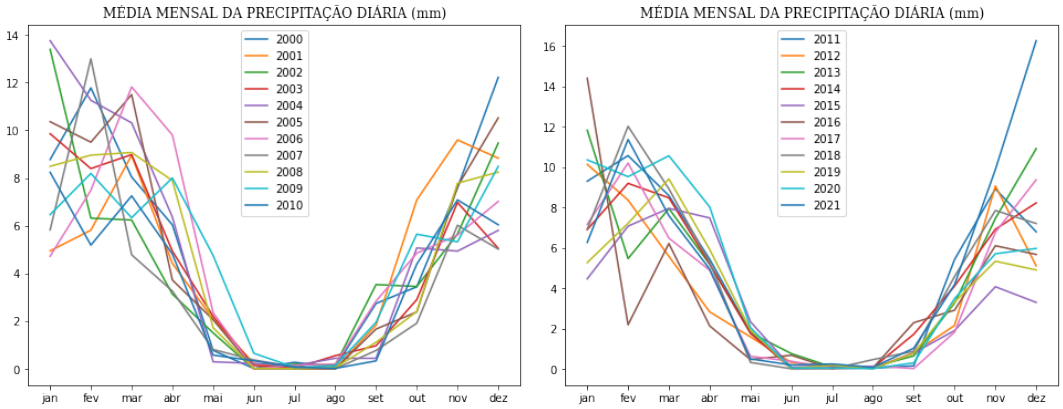

Figura 12 – Precipitação no Tocantins entre 2000 e 2021 por mês

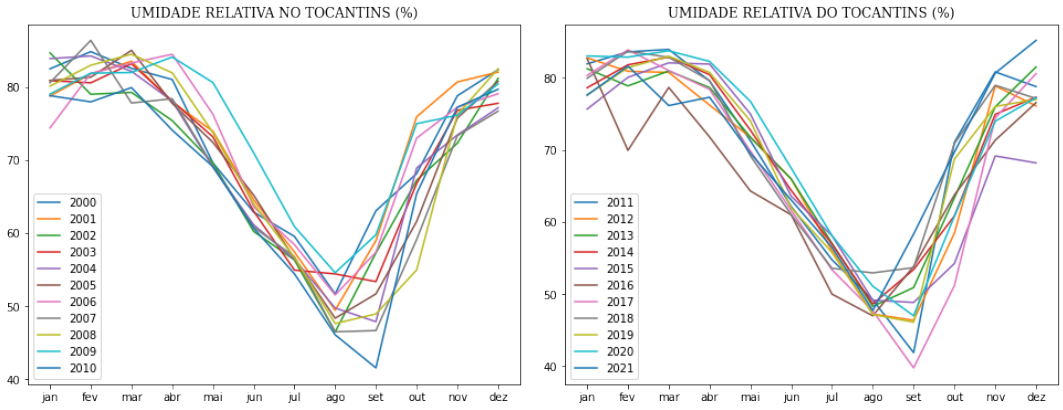

Figura 13 – Umidade Relativa do Ar no Tocantins entre 2000 e 2021 por mês

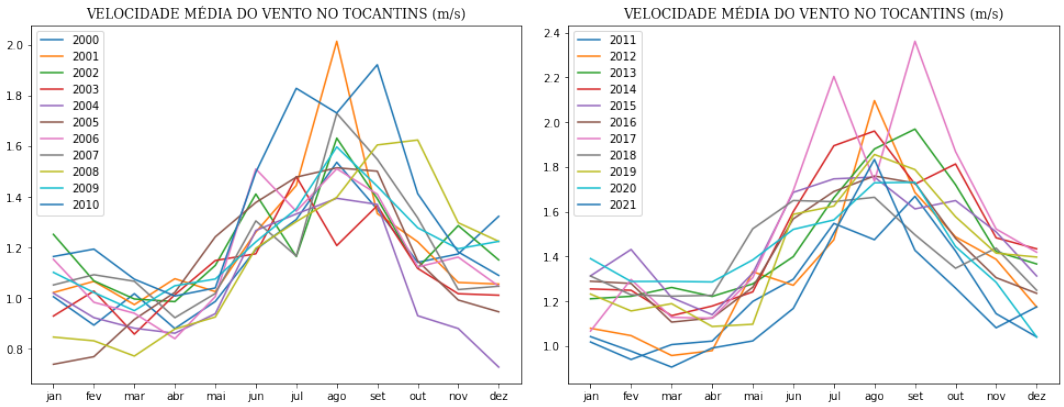

Figura 14 – Velocidade Media do Vento no Tocantins entre 2000 e 2021 por mês

4.4 Considerações Finais

Atualmente há uma consciência cada vez maior da necessidade de combate às queimadas, e conhecimento acerca dos impactos ambientais imensos tanto local afetado quanto para o planeta em geral, com o agravamento do efeito estufa. Há também a consciência de que a tecnologia é uma aliada no combate às mesmas, com dados sendo gerados e armazenados em escalas gigantescas o tempo todo. No entanto, esse excesso no volume de dados acaba por

tornar-se um problema a partir do momento que não encontra um destino e um uso, tornando-se mais inacessível adquirir conhecimento a partir dele conforme seu aumenta.

Adquirir o conhecimento intrínseco aos dados disponíveis passa por diversos obstáculos, enfrentados em maior ou menor medida no decorrer desse trabalho, como a disponibilidade, a integridade, a separação em partes incompatíveis à primeira vista e o volume gigantesco que torna quaisquer operações demoradas quando não inviáveis. No entanto, superados esses obstáculos, pôde-se converter esses dados acumulados e inacessíveis de forma a se analisar e entender as regiões mais afetadas, os períodos com mais e menos ocorrência, as tendências de alguns cenário se repetirem e outros novos que surgiram através dos anos.

5 CONCLUSÕES

Neste trabalho, um grande volume de dados, de duas fontes de dados distintas, sendo elas o BDQueimadas, com os dados de mais de 250 milhões de pontos de incêndio fornecidos pelo INPE (Instituto Nacional de Pesquisas Espaciais) e o BDMEP, banco de dados meteorológicos do INMET (Instituto Nacional de Meteorologia), foi estruturado e unido de forma adequada para permitir a realização de uma análise, gerando conhecimento útil por meio da criação de gráficos e tabelas. Com isso, percebe-se a importância do processo de Análise Exploratória, na busca de conhecimento nas bases de dados analisadas. Seguindo ela, buscou-se estruturar os dados e usar métodos estatísticos para transformar a totalidade dos registros em intervalos discretos, como a média mensal, conseguindo assim acompanhar a evolução das variáveis de incêndio, temperatura, umidade, vento, precipitação, entre outros, em gráficos de fácil entendimento, possibilitando visualizar a evolução e características de como os incêndios se apresentaram em cada período, como por exemplo quais áreas foram mais afetadas, quais foram menos afetadas, e as causas para ambos. O trabalho também atingiu o objetivo de aplicar técnicas ao tratar os dados brutos obtidos, possibilitando o procedimento de análise em si. Então, como planejado, várias conclusões foram obtidas a partir dessa análise, algumas já esperadas como, por exemplo, a correlação de variáveis climáticas entre si e com a ocorrência de incêndios. Outras eram imaginadas, como a relação geográfica e do bioma, mas seguindo procedimento da Análise Exploratória, foram buscados dados adicionais (localização dos biomas), que confirmaram a hipótese inicial. De forma semelhante, a queda na quantidade de incêndios no ano de 2020 precisou de dados adicionais para confirmar que havia uma relação com uma ação governamental de combate às queimadas. Apurou-se que, o fato do Governo do Tocantins ter realizado um teste entre 2019 e 2020 em que, ao invés de contratar as brigadas de incêndio apenas nos meses de seca, realizou a contratação durante todo o ano para trabalhar em ações preventivas e educativas, causou uma redução drástica na quantidade de incêndios registrada em 2020.

5.1 Trabalhos Futuros

Após a realização desta análise de dados de queimadas do estado do Tocantins, espera-se no futuro, expandir essa análise para os demais estados da Amazônia Legal, utilizando critérios semelhantes e as técnicas necessárias para lidar com uma quantidade de dados ainda maior. Com isso, pode ser possível utilizar os aprendizados das ações já tomadas em um estado, sendo aplicados em outro e vice-versa, possibilitando uma otimização das ações concretas, no sentido de proteger o território da Amazônia Legal, agora e no futuro. Além disso, em um próximo trabalho espera-se expandir além da Análise Exploratória e aplicando o KDD em sua totalidade, com um algoritmo de mineração de dados sendo executado para buscar padrões que não foram

descobertos.

REFERÊNCIAS

ALENCAR, A. et al. Uso do fogo na amazônia: Estudos de caso ao longo do arco de desmatamento. **World Bank, unpublished report**, 1997.

BENOMRANE, S.; AYED, M. B.; ALIMI, A. M. An agent-based knowledge discovery from databases applied in healthcare domain. In: IEEE. **2013 International Conference on Advanced Logistics and Transport**. [S.l.], 2013. p. 176–180.

CABRAL, A. L. A.; FILHO, L. O. M.; BORGES, L. A. C. Uso do fogo na agricultura: Legislação, impactos ambientais e realidade na amazônia. **Periódico Eletrônico Fórum Ambiental da Alta Paulista**, v. 9, n. 5, 2013.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, v. 1, n. 1, p. 1–29, 2009.

CHEN, J. et al. Exploratory data analysis of activity diary data: a space–time gis approach. **Journal of Transport Geography**, Elsevier, v. 19, n. 3, p. 394–404, 2011.

CHENG, T.; WANG, J. Integrated spatio-temporal data mining for forest fire prediction. **Transactions in GIS**, Wiley Online Library, v. 12, n. 5, p. 591–611, 2008.

CRAVEN, M. W.; SHAVLIK, J. W. Using neural networks for data mining. **Future generation computer systems**, Elsevier, v. 13, n. 2-3, p. 211–229, 1997.

DIAZ, M. d. C. V. et al. O preço oculto do fogo na amazônia: Os custos econômicos associados as queimadas e incêndios florestais. **Belém: Instituto de Pesquisa Ambiental da Amazônia**, 2002.

DINIZ, C. G. et al. Deter-b: The new amazon near real-time deforestation detection system. **Ieee journal of selected topics in applied earth observations and remote sensing**, IEEE, v. 8, n. 7, p. 3619–3628, 2015.

ESPACIAIS, I. I. N. de P. **Programa Queimadas**. 2020. Acessado em: 26/03/2020. Disponível em: <http://queimadas.dgi.inpe.br/queimadas/portal>.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, ACM New York, NY, USA, v. 39, n. 11, p. 27–34, 1996.

FAYYAD, U. M.; HAUSSLER, D.; STOLORZ, P. E. Kdd for science data analysis: Issues and examples. In: **KDD**. [S.l.: s.n.], 1996. p. 50–56.

FAYYAD, U. M. et al. Knowledge discovery and data mining: Towards a unifying framework. In: **KDD**. [S.l.: s.n.], 1996. v. 96, p. 82–88.

GURUVAYUR, S. R.; SUCHITHRA, R. A detailed study on machine learning techniques for data mining. In: IEEE. **2017 International Conference on Trends in Electronics and Informatics (ICEI)**. [S.l.], 2017. p. 1187–1192.

JAIN, N.; SRIVASTAVA, V. Data mining techniques: a survey paper. **IJRET: International Journal of Research in Engineering and Technology**, Citeseer, v. 2, n. 11, p. 2319–1163, 2013.

LAKSHMI, B.; RAGHUNANDHAN, G. A conceptual overview of data mining. In: IEEE. **2011 National Conference on Innovations in Emerging Technology**. [S.l.], 2011. p. 27–32.

MALLEY, B.; RAMAZZOTTI, D.; WU, J. T.-y. Data pre-processing. **Secondary analysis of electronic health records**, Springer, p. 115–141, 2016.

MARCUZZO, F.; GOULARTE, E. Caracterização do ano hidrológico e mapeamento espacial das chuvas nos períodos Úmido e seco do estado do tocantins (characterization of the hydrological year and spatial rains mapping of wet and dry periods in the state of tocantins) - karakterisering af det hydrologiske år og rumlig kortlægning af nedbør i de våde og tørre perioder af staten tocantins - - link: http://rigeo.cprm.gov.br/jspui/handle/doc/14837. v. 6, p. 091, 07 2013.

MELLO, N. A. G. A. R. d.; ARTAXO, P. EvoluÃ§Ãdo Plano de AÃ§Ãpara PrevenÃ§Ãe Controle do Desmatamento na AmazÃ´nia Legal. **Revista do Instituto de Estudos Brasileiros**, scielo, p. 108 – 129, 04 2017. ISSN 0020-3874. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0020-38742017000100108&nrm=iso>.

NOBLE, W. S. What is a support vector machine? **Nature biotechnology**, Nature Publishing Group, v. 24, n. 12, p. 1565–1567, 2006.

PEREIRA-SILVA, E. F. et al. Sucessão ecológica e o uso de nitrogênio em florestas tropicais. **Revista Interciência & Sociedade**, v. 1, p. 149–159, 2011.

QUINLAN, J. R. Decision trees and decision-making. **IEEE Transactions on Systems, Man, and Cybernetics**, IEEE, v. 20, n. 2, p. 339–346, 1990.

RATHI, M. Regression modeling technique on data mining for prediction of crm. In: SPRINGER. **International Conference on Advances in Information and Communication Technologies**. [S.l.], 2010. p. 195–200.

RODRIGUES, S. C. A. **Modelo de regressão linear e suas aplicações**. Tese (Doutorado) — Universidade da Beira Interior, 2012.

ROKACH, L.; MAIMON, O. Decision trees. In: **Data mining and knowledge discovery handbook**. [S.l.]: Springer, 2005. p. 165–192.

SEMARH, S. do Meio ambiente e R. H. Plano de ação para prevenção controle do desmatamento e queimadas do Tocantins. [S.l.], 2021.

SETZER, A.; MORELLI, F.; SOUZA, J. C. O banco de dados de queimadas do inpe. **Biodiversidade Brasileira-BioBrasil**, n. 1, p. 239–239, 2019.

TOCANTINS, R. C. do. **Governo do Tocantins apresenta redução nos focos de queimadas em 2020**. <https://www.to.gov.br/noticias/governo-do-tocantins-apresenta-reducao-nos-focos-de-queimadas-em-2020/4peikhqaqwl2>. Último acesso em 29/11/2022.

ZHICHKINA, L. et al. Satellite monitoring systems in forestry. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2020. v. 1515, n. 3, p. 032043.

ZHONG, N. et al. Kdd process planning. In: **KDD**. [S.l.: s.n.], 1997. p. 291–294.