Alexandre Henrique Soares Dias

Uma Análise Exploratória de Dados sobre incêndios florestais no Brasil

Natal – RN

Novembro de 2019

Alexandre Henrique Soares Dias

Uma Análise Exploratória de Dados sobre incêndios florestais no Brasil

Trabalho de Conclusão de Curso de Engenharia de Computação da Universidade Federal do Rio Grande do Norte, apresentado como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Computação

Orientador: Ivanovitch Medeiros Dantas da Silva

Universidade Federal do Rio Grande do Norte – UFRN

Departamento de Engenharia de Computação e Automação – DCA

Curso de Engenharia de Computação

Natal – RN

Novembro de 2019

UFRN / Biblioteca Central Zila Mamede
Catalogação da publicação na Fonte.

DIAS, Alexandre Henrique Soares. **Uma Análise Exploratória de Dados sobre incêndios florestais no Brasil**. 2019. 52f.

Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) - Universidade Federal do Rio Grande do Norte, Centro de Tecnologia, Departamento de Engenharia de Computação e Automação, Natal, 2019.

Dedico este trabalho a Deus, Criador e Soberano sobre todas as coisas nos céus e na terra, visíveis e invisíveis. Porque Dele e por Ele, e para Ele, são todas as coisas; glória, pois, a Ele eternamente. (Romanos 11:36)

AGRADECIMENTOS

Agradeço a Deus por ter sido inspiração, baluarte e consolo nos momentos alegres e difíceis.

Agradeço à minha família, por todo o carinho, esforço e apoio provido desde o início da minha jornada acadêmica.

Agradeço a todos os amigos que me acompanharam nesta caminhada. Em especial, àqueles que sempre estiveram próximos e me ajudaram a enxergar o que muitas vezes eu não pude. Obrigado, Massao Henrique, Fábio Abreu, Lucas Lyon, Gabriel Coelho, Josué Araújo, Higor Felype, Andressa Jamar, Jayne Izabel, Maria Luiza Cartaxo e Élida Oliveira.

Agradeço aos professores Sérgio Luiz, Ivanovitch Medeiros e Luiz Affonso pela confiança, determinação e lições de vida transmitidas ao longo de todos esses anos.

Agradeço às pessoas que compõem a Cru Campus, NPC, e a IPN, por me ajudarem a compreender meu propósito dentro e fora da universidade.

Agradeço a Universidade Federal do Rio Grande do Norte, seu corpo docente e administrativo, por me proporcionarem a oportunidade de ter vivido os melhores anos da minha vida ao lado de pessoas tão excepcionais.

"Time has come to go
Pack your bags, hit the open road
Our hearts just won't die
It's the trip, keeps us alive."
- Still Corners

RESUMO

Incêndios florestais geram impactos sociais e ambientais nos locais onde ocorrem. As florestas brasileiras, especialmente nos meses mais secos do ano, estão mais suscetíveis a este fenômeno. Assim, este trabalho busca elucidar padrões geográficos e sazonais no contexto de focos de incêndios em todo o território brasileiro em seus 6 biomas naturais. Os dados utilizados nesta pesquisa foram coletados banco de queimadas do Instituto Nacional de Pesquisas Espaciais (INPE) e do órgão americano para Administração Nacional Aeronáutica e Espacial (*NASA*). Com essas informações, são aplicadas as técnicas mais comuns em análise de dados, desde o tratamento e limpeza dos dados, e da concepção de hipóteses até a visualização de gráficos e mapas. A pesquisa tem como foco o bioma amazônico e a região política da Amazônia Legal, dada a recente repercussão em volta do tema na grande mídia nacional e internacional.

Palavras-chaves: Ciência de Dados. Análise. Incêndios Florestais.

ABSTRACT

Forest wildfires cause social and environmental impacts where they occur. Brazilian forests, especially in the driest months of the year, are more susceptible to this phenomenon. Thus, this work seeks to elucidate geographic and seasonal patterns in the context of fire outbreaks throughout the Brazilian territory in its 6 natural biomes. The data used in this research were collected from the National Institute for Space Research (INPE) and the US National Aeronautical and Space Administration (NASA) database. It is shown some of the most common data analysis techniques and when to apply them, from data handling and data cleansing, to hypothesizing and graphs visualization. The research focuses on the Amazon biome and the political region of the Legal Amazon, given the recent repercussion around the topic of wildfires in the Amazon Rainforest in the national and international mainstream media.

Keywords: Data Science. Wildfires. Amazon Rainforest.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

INPE	*Instituto Nacional de Pesquisas Espaciais*
NASA	*National Aeronautics and Space Administration*
FIRMS	*Fire Information for Resource Management System*
MODIS	*Moderate Resolution Imaging Spectroradiometer*
IDE	*Integrated Development Environment*

SUMÁRIO

1 INTRODUÇÃO

A Ciência de Dados tem se tornado um tópico cada vez mais abordado e praticado no contexto das ciências sociais aplicadas (administração, logística, gestão pública, etc.) e ciências naturais como a física, a química, a astronomia, entre outras. Isso se deve ao fato de que suas aplicações abrangem escopos multi-disciplinares e polivalentes que se utilizam de algoritmos, sistemas, processos e métodos científicos para extrair informações e ideias de dados estruturados ou não-estruturados (DHAR, 2012).

Dessa maneira, a quantidade de técnicas e métodos desenvolvidos diariamente para aumentar a qualidade das aplicações em ciência de dados mantem um ritmo crescente. Em contrapartida, é um consenso na literatura de que não existe um processo padrão, linear e sistemático de como se deve fazer um projeto de Ciência de Dados, principalmente na fase inicial de exploração e análise de dados. Pelo contrário, como um fluxo cíclico, é necessário por diversas vezes repetir ações e reavaliar decisões que foram tomadas em etapas anteriores desde a obtenção dos dados até a elaboração de um modelo capaz de descrever com precisão os fenômenos que se queiram explicar ou analisar.

O processo supracitado dá-se segundo (WICKHAM; GROLEMUND, 2017) como na figura 1.

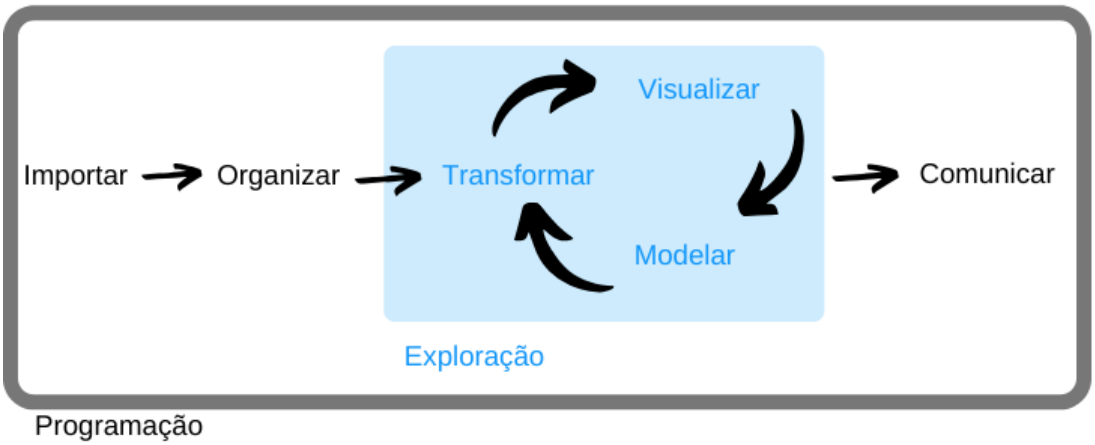

Figura 1 – Fluxo de um projeto de Ciência de Dados

Fonte – Livro R for Data Science - Adaptado

Isto é, inicialmente, os dados são armazenados em arquivos, banco de dados ou APIs *web* (conjunto de rotinas e padrões de programação para acesso a um aplicativo de software ou plataforma na internet) de onde devem ser **importados**. Uma vez que os dados são carregados em um ambiente de desenvolvimento, é importante adequá-los ao objetivo que se deseja atingir. Isto é, os dados devem ser **organizados** ou formatados

de modo que sua utilização seja viável para construção de gráficos e até mesmo modelos aprendizagem de máquina.

Em seguida, um passo natural é o de **transformar** os dados. A transformação de dados inclui desde a aplicação de filtros até a criação de novas variáveis que são funções de variáveis já existentes (como o índice de massa corporal que é a razão entre peso e o quadrado da altura de um indivíduo). Após essa etapa, são produzidas **visualizações** que ajudam a esclarecer e evidenciar causalidades e/ou correlações até então possivelmente não conhecidas ou incertas. Uma vez que as questões e discussões levantadas são suficientemente precisas para o objetivo buscado no projeto, é natural construir **modelos** que possam tornar todo o processo escalável e facilmente reproduzível. Ainda segundo (WICKHAM; GROLEMUND, 2017), tais modelos são fundamentalmente ferramentas matemáticas ou computacionais com grande poder de escalabilidade.

Todavia, como retratado através Figura 1, este processo ocorre ciclicamente pois à medida que novas descobertas e variáveis são adicionadas ao problema, é necessário voltar à fase de transformação dos dados e construir novos gráficos que incorporam as modificações introduzidas. Por último, tendo concluído o ciclo de exploração, uma parte crítica de um projeto de Ciência de Dados é a **comunicação**. Um projeto com modelos robustos e visualizações completas não tem relevância se não for bem comunicado.

Neste trabalho, é apresentada uma análise exploratória de dados sobre incêndios florestais no Brasil a partir de dados coletados do Instituto Nacional de Pesquisas Espaciais (INPE) e da Agência Espacial Norte Americana (*NASA*). Por meio desses dados, buscar-se-á percorrer todas as etapas anteriormente descritas que estão presentes em um projeto de Ciência de Dados.

1.1 Motivação

A tarefa de realizar uma análise exploratória de dados é normalmente árdua, no sentido de que muitas vezes chega-se a gastar até cerca de 80% do tempo no processo de limpeza e preparação dos dados que posteriormente serão utilizados em modelos e gráficos (DASU; JOHNSON, 2003).

Ainda, não há muitas fontes que relatam como fazer o processo de limpeza de dados de uma maneira fácil e eficiente, pois, além de requerer atitudes específicas em cada problema, o fluxo de uma análise de dados ocorre repetidamente enquanto surgem novos desafios, ideias, e novos dados são coletados (WICKHAM, 2014b). Corroborando com o pensamento de que a análise de dados é uma etapa crucial de qualquer projeto de Ciência de Dados, segundo (PENG, 2016):

> *[...] A análise exploratória de dados é importante porque permite que o investigador possa tomar decisões críticas sobre o que é interessante ser buscado e o que merece atenção (dentro do escopo do projeto) pois as relações contidas nos dados não são evidentes (e provavelmente nunca serão mesmo que se investigue bastante). É importante tomar esses tipos de decisões porque é a partir delas que se determina se um projeto vai avançar ou permanecer dentro do orçamento.*

Portanto, dada a relevância e grau de dificuldade na execução de uma análise de dados, a motivação desse trabalho é fornecer um estudo de caso sobre a análise exploratória de dados no contexto de incêndios florestais no Brasil. Esse por sua vez, tem sido um tópico amplamente discutido entre entidades governamentais, ONGs e ambientalistas, devido a sua relevância diante das transformações as quais nosso planeta está sendo submetido em virtude de ações naturais, e em grande parte, antrópicas.

1.2 Objetivos

Pelo que já foi dito até então, o principal objetivo é entender, descobrir e relatar padrões (geográficos, sazonais, etc.) de incêndios florestais nos biomas florestais brasileiros levantando hipóteses e até mesmo checando algumas suposições. Para isso é abordado um conjunto de técnicas e práticas comuns aplicáveis durante uma análise exploratória de dados.

Para atingir esse objetivo principal, surgem como necessários os seguintes objetivos específicos:

(a) Coleta de dados das bases do INPE e da *NASA*;

(b) Processamento, organização, limpeza e filtragem de dados;

(c) Analise crítica de resultados.

1.3 Estrutura do Trabalho

Este trabalho apresenta uma introdução sobre o tema, mostrando a dificuldade presente no processo da prática da análise de dados, além de uma justificativa da proposta e seus respectivos objetivos e finalidades. Em seguida, O Capítulo 2 fornece uma fundamentação teórica para contextualizar o leitor sobre as ferramentas utilizadas e a temática abordada. No Capítulo 3, são relatados os procedimentos de coleta e armazenamento dos dados. Posteriormente, no Capítulo 4 é relatado todo o processo da análise exploratória de dados por meio de questionamentos e motivações que permeiam o texto nessa seção. O Capítulo 5, por sua vez, fornece uma breve discussão em volta das descobertas feitas na

análise exploratória. E, por fim, o Capítulo 6 traz consigo as principais conclusões acerca dos resultados encontrados.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresentado o sumário as principais ferramentas utilizadas no percurso do desenvolvimento deste trabalho. Tal descrição também visa possibilitar ao leitor uma melhor compreensão sobre o tópico abordado.

2.1 Ciência de Dados

Brevemente introduzida no Capítulo 1, até aqui não foi detalhado o que de fato é a chamada Ciência de Dados. Segundo Frank Lo, um dos editores da plataforma online de busca de empregos *DataJobs*[1], a Ciência de dados é um campo multidisciplinar que mistura a inferência, o desenvolvimento de algoritmos, e a tecnologia para resolver problemas analiticamente complexos. Assim, utilizando dados obtidos de diversas fontes, esta nova forma de fazer ciência encontra-se na sobreposição entre as áreas de Ciência da Computação, Matemática e Estatística e Especialização Científica que é o termo usado para se referir ao conhecimento subjetivo sobre o problema que se quer resolver bem como suas nuances. Para tal relacionamento, Drew Conway, CEO e fundador da startup de tecnologia *Alluvium*[2], sugere o seguinte Diagrama de Venn[3] mostrado na Figura 2:

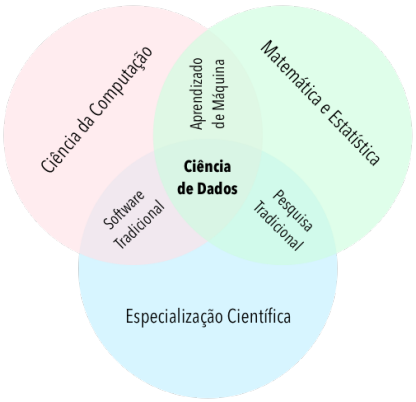

Figura 2 – Diagrama de Venn.

Fonte – Drew Conway

Dessa maneira, pela combinação e cooperação mútua entre todos estes campos do conhecimento, a ciência de dados se utiliza de técnicas advindas da interseção entre os conjuntos para por fim, atingir um objetivo maior. Neste mesmo contexto entra a análise

[1] https://datajobs.com/what-is-data-science, acesso em 01 de Novembro de 2019

[2] https://www.augury.com, acesso em 24 de Outubro de 2019

[3] http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, acesso em 24 de Outubro de 2019

exploratória de dados, que é utilizada como prática para extrair informações dos dados que estão sendo manipulados.

2.2 Análise Exploratória de Dados

Dentro do escopo das atividades em Ciências de Dados, existe a Análise Exploratória de Dados (AED). Em (PENG, 2016) é relatado que os objetivos de uma AED são muitos, mas incluem a identificação de relações entre variáveis que são particularmente interessantes ou inesperadas, verificando se há alguma evidência a favor ou contra uma hipótese declarada, verificando se há problemas com os dados coletados, como dados ausentes ou erro de medição), ou identificar determinadas áreas em que mais dados precisam ser coletados. Nesse ponto, detalhes mais precisos da apresentação dos dados e evidências, importantes para o produto final, não são necessariamente o foco.

Em outras palavras, em uma AED, o analista não preocupa-se necessariamente com a qualidade das visualizações produzidas, mas com as informações que podem ser obtidas dos dados. Desse modo, essa é uma atividade comum em projetos de Ciência de Dados pois permite que informações antes desconhecidas sejam obtidas por meio das observações e o relacionamento entre variáveis.

2.3 Amazônia Legal

A Amazônia Legal[4] trata-se de uma região delimitada através de intenções sociopolíticas dados os problemas comuns ambientais, políticos e de infraestrutura encontrados pelos Estados da Federação que a compõem. Sua região abrange uma área de 5 217 423 km^2, correspondendo a 61% do território brasileiro, contendo 20% do bioma Cerrado e parte do Pantanal no Mato Grosso. Em todos os 9 Estados que compõem essa região reside cerca de 55% da população indígena de todo o País.

Segundo o Ministério do Meio Ambiente por meio do Caderno Da Região Hidrográfica[5], A relevância desse território dá-se pela sua diversidade ambiental como também pela quantidade de recursos naturais disponíveis.

Nele está localizada a Bacia Amazônica a maior bacia hidrográfica do mundo, com cerca de um quinto do volume total de água doce do planeta. Com aproximadamente 40 mil espécies de plantas e mais de 400 de mamíferos. E ainda quase 1.300 espécies de pássaros, cerca de 3 mil espécies de peixes e milhões de insetos.

[4]https://www.oeco.org.br/dicionario-ambiental/28783-o-que-e-a-amazonia-legal/, acesso em 15 de Novembro de 2019

[5]https://www.mma.gov.br/estruturas/161/_publicacao/161_publicacao03032011024915.pdf, acesso em 20 de Novembro de 2019

2.4 Bioma

Bioma é uma unidade biológica ou espaço geográfico cujas características específicas são definidas pelo macroclima, a fitofisionomia, o solo e a altitude, dentre outros critérios. São tipos de ecossistemas, habitats ou comunidades biológicas com certo nível de homogeneidade (FARIA, 2019).

Segundo o Ministério do Meio Ambiente[6], o Brasil é formado por seis biomas de características distintas: Amazônia, Caatinga, Cerrado, Mata Atlântica, Pampa e Pantanal. Cada um desses ambientes abriga diferentes tipos de vegetação e de fauna.

Nos conjuntos de dados utilizados neste trabalho, existem informações sobre todos os 6 biomas brasileiros. Todavia, o enfoque da AED é voltado para o bioma Amazônico e Cerrado, dado que esses biomas ocupam mais da metade da área de todo território brasileiro.

2.5 Espectrorradiômetro de Imagem com Resolução Moderada (*MODIS*)

O Sistema de Informação de fogo para gerenciamento de recursos (*FIRMS*) da *NASA* fornece dados de incêndio ativos quase em tempo real dentro de 3 horas da observação por satélite. O equipamento que captura os dados utilizados neste trabalho e a principal fonte de detecção de focos incêndios pela *NASA* desde 2002 é o chamado Espectrorradiômetro de imagem com resolução moderada (*MODIS*, do inglês, *Moderate Resolution Imaging Spectroradiometer*).

Além de dados tabulares sobre variáveis atmosféricas e naturais, *MODIS* (Coleção 6) também disponibiliza imagens de alta resolução da superfície da terra que auxiliam na detecção de focos de incêndios. Cada local de sinalização do *MODIS* representa o centro de um pixel de 1 que representa uma área em volta de 1 km^2, podendo ser sinalizado pelo algoritmo de detecção como contendo um ou mais incêndios.

MODIS[7] é um instrumento essencial a bordo dos satélites *Terra* (EOS AM-1) e *Aqua* (originalmente conhecido como EOS PM-1). O satélite *Terra* orbita ao redor da Terra cronometradamente para que passe de norte a sul através do equador pela manhã, enquanto o *Aqua* passa de sul a norte sobre o equador à tarde.

O *Terra MODIS* e o *Aqua MODIS* visualizam toda a superfície da Terra a cada 1 a 2 dias, adquirindo dados em 36 bandas espectrais ou grupos de comprimentos de onda. Com essas informações, é possível melhorar a compreensão das dinâmicas e processos globais que

[6] https://www.mma.gov.br/biomas.html, acesso em 29 de Novembro de 2019
[7] https://modis.gsfc.nasa.gov/about/, acesso em 20 de Outubro de 2019

ocorrem na terra, nos oceanos e na atmosfera mais baixa. O *MODIS* está desempenhando um papel vital no desenvolvimento de modelos validados, globais e interativos do sistema *Earth* da *NASA*, capazes de prever mudanças globais com precisão suficiente para ajudar os formuladores de políticas públicas a tomar decisões sólidas em relação à proteção do meio ambiente.

O INPE também se utiliza deste dispositivo para capturar os dados que são disponibilizados em sua plataforma. O que torna este aparelho um importante contribuidor para o desenvolvimento e a própria existência deste trabalho.

2.6 Trabalhos Relacionados

Como já foi dito, a tarefa de realizar uma Análise Exploratória de Dados não é simples. Não existe um consenso comum sobre como exatamente se executa tal atividade. No entanto, o que existe é uma gama de técnicas criadas para contornar a maioria dos problemas enfrentados por cientistas de dados, desde a manipulação dos dados e sua limpeza até as maneiras de exibir as informações da forma mais clara possível. Nesta seção são apresentados alguns trabalhos relacionados com a prática da análise exploratória de dados no contexto de incêndios florestais.

Em (GAITHER et al., 2011), os autores propõem um estudo de caso sobre como riscos de incêndios florestais e a vulnerabilidade social se relacionam no sudeste dos Estados Unidos. A abordagem deste trabalho está voltada para uma análise qualitativa espacial do objeto de pesquisa. Os autores mostram que as áreas mais pobres na região analisada estão mais suscetíveis a ocorrências e desastres causados por incêndios florestais.

Já (HOLMES; HUGGETT; WESTERLING, 2008), propõe uma análise estatística sobre grandes incêndios florestais nos Estados Unidos, focando nos aspectos econômicos para combatê-los e no gerenciamento de riscos e recursos. Considerando que eventos extremos de incêndios florestais podem levar a mudanças súbitas e massivas de ecossistemas naturais e na economia local, é proposta uma análise sobre a probabilidade de haver alguma mudança em virtude de catástrofes. Também é mostrado que 94% do dinheiro investido entre 1980 e 2002 nos EUA para combate de incêndios, é decorrente de apenas 1,4% do total de ocorrências no período.

Em outra vertente, (ALVES; NÓBREGA, 2011) apresenta uma análise espacial sobre o risco de incêndio florestal através de uma análise de dados climáticos de precipitação e umidade. A área de atuação da análise se restringe ao Parque Nacional do Catimbau, uma unidade de conservação do estado de Pernambuco. Os resultados deste trabalho foram importantíssimos para o planejamento e gestão do meio ambiente do Parque Nacional do Catimbau.

Semelhantemente, (WHITE; RIBEIRO, 2011) apresenta uma análise de dados relacionando a precipitação de água da chuva na ocorrência de incêndios florestais dentro do Parque Nacional Serra de Itabaiana, em Sergipe. No estudo, é mostrado que para a região há uma correlação negativa entre a precipitação e a quantidade de incêndios.

Por outro lado, em (ARAUJO et al., 2007) é demonstrada uma análise exploratória do número de ocorrência de incêndios na região chamada de "Arco do Desflorestamento", que se estende de Rondônia ao Maranhão. Essa região tem como uma de suas principais características a intensa exploração madeireira e a utilização do fogo para fins de agricultura e agropecuária. Os resultados mostraram que cerca de 80% das ocorrências de incêndios florestais na região Amazônia Legal estão na área do "Arco do Desflorestamento".

Dentre os trabalhos apresentados, nenhum aborda com tanto grau de generalidade como o proposto por este trabalho sobre a problemática de incêndios florestais no Brasil nos últimos anos. (GAITHER et al., 2011) e (HOLMES; HUGGETT; WESTERLING, 2008), além de serem estudos aplicados em outro país, focam em incêndios florestais sobre os prismas econômicos e sociais. Em contrapartida, (ALVES; NÓBREGA, 2011) e (WHITE; RIBEIRO, 2011) realizam análises de dados sobre incêndios florestais mas apenas em parques ecológicos de Pernambuco e Sergipe, e considerando apenas variáveis climáticas, o que se torna uma análise muito limitada tendo em vistas as dimensões continentais do Brasil. Por fim, (ARAUJO et al., 2007) propõe um estudo simplista sobre a ocorrência de incêndios apenas na região do "Arco do Desflorestamento", que apesar de ser uma região com extrema relevância ambiental, abrange apenas parte do território da Amazônia Legal.

3 METODOLOGIA PARA OBTENÇÃO E RECONHECIMENTO DOS DADOS

Neste Capítulo, é abordada a metodologia seguida para o desenvolvimento deste trabalho. Na Seção 3.1, é apresentada a classificação desta pesquisa quanto a sua natureza, abordagem e objetivo. Na Seção 3.2, são descritas as ferramentas utilizadas no desenvolvimento deste trabalho. Por último, na Seção 3.3 são mostrados os procedimentos para captura e leitura dos dados, bem como seus tipos e algumas considerações para a análise.

3.1 Classificação da Pesquisa

Uma pesquisa pode ser classificada a partir de três perspectivas distintas, a saber, sua natureza, abordagem de dados, objetivos e procedimentos técnicos.

Quanto à natureza de uma pesquisa, (SILVA; MENEZES, 2005) propõe as classificações de básica e aplicada. A pesquisa básica objetiva a geração de novos conhecimentos que visam o avanço do estado da arte científica de uma determinada área sem necessariamente uma aplicação imediata. Por outro lado, a pesquisa aplicada objetiva a construção do conhecimento através de aplicações práticas, abordando verdades e interesses locais. Portanto, este trabalho de caráter exploratório pode ser classificado como uma pesquisa aplicada.

Ainda segundo (SILVA; MENEZES, 2005), quanto aos tipos de abordagem aplicadas aos dados, uma pesquisa classifica-se entre quantitativa ou qualitativa. Uma pesquisa é quantitativa se ela pode ser traduzida em termos puramente numéricos através da utilização de modelos matemáticos, estatísticos e classificatórios. Em contrapartida, uma pesquisa é qualitativa se ela está buscando descobrir ou transmitir aspectos subjetivos sobre o tema pesquisado, sem atentar com representatividade numérica de resultados e procedimentos. Dessa maneira, o conteúdo de pesquisa exposto aqui recai sobre as duas categorias, o que implica que esta é uma pesquisa com aspectos qualitativos e quantitativos.

A depender dos objetivos de uma pesquisa, ela pode ser classificada como exploratória descritiva ou explicativa. Para (GIL, 2007), uma pesquisa que busca identificar padrões e fatores que determinam ou contribuem para a ocorrência de um fenômeno, tem caráter explicativo. Dessa maneira, esta pesquisa tem aspectos descritivos e explicativos, pois busca elucidar padrões e características sobre incêndios florestais no Brasil e promover um aprofundamento acerca do entendimento sobre o tema.

3.2 Instrumentação

Nesta seção são apresentadas as ferramentas utilizadas para o desenvolvimento deste trabalho, sendo elas, a linguagem de programação R e alguns de seus pacotes de código-fonte aberto, e o ambiente de desenvolvimento integrado, *RStudio*.

3.2.1 Linguagem de Programação R

R foi a linguagem de programação escolhida para o desenvolvimento deste trabalho por ser livre, ter código fonte aberto e ser disponível nos principais ecossistemas de computadores. Consequentemente, se uma análise é feita em *R*, qualquer um pode facilmente reproduzi-la. Dentre outros motivos, também destacam-se os listados por (WICKHAM, 2014a) em seu livro, *Advanced R*:

- *R* concebe um conjunto massivo de pacotes para modelagem estatísticas de dados, aprendizagem de máquina, visualização, importação e manipulação de dados;
- Ferramentas poderosas para comunicação de resultados;
- Um Ambiente de Desenvolvimento Integrado (sigla em inglês, *Integrated Develpment Environment*) especialmente destinada a análise de dados e programação voltada à prática da estatística;
- Facilidades de metaprogramação. Os recursos de meta-programação em *R* permitem escrever funções magicamente sucintas e concisas.

Corroborando, (VENABLES et al., 1997) caracteriza esta linguagem de programação como um "ecossistema" com a finalidade de rotular *R* como um sistema totalmente planejado e coerente, em vez de um acréscimo incremental de ferramentas muito específicas e inflexíveis, como é frequentemente o caso de outros softwares de análise de dados.

3.2.2 Ambiente de Desenvolvimento Integrado: *RStudio*

O *RStudio*[1] é uma *IDE* para *R* que conta com um *console*, um editor de texto com realce de sintaxe que suporta execução direta de código, e ferramentas para gráficos, histórico, *debug* e gerenciamento do espaço de trabalho (RStudio Team, 2015).

Todo o desenvolvimento deste trabalho foi realizado na plataforma do *RStudio* desde a importação dos dados até a geração dos resultados. Este *software* foi escolhido devido ao suporte de diversos meios que possibilitam a programação em *R* ser mais fluida, bem como a fácil instalação de pacotes, escrita e manutenção de código.

[1] http://www.rstudio.com, acessado em 01 de Setembro de 2019

3.2.3 Pacotes e bibliotecas

Nesta seção são apresentados os principais pacotes e bibliotecas utilizadas no desenvolvimento deste trabalho. Vale ressaltar, que a lista não se limita aos elementos citados abaixo, menções honrosas vão para as bibliotecas *stringr*, *TSA*, *sf*, *rgdal* e *zoo*.

3.2.3.1 *dplyr*

Segundo (WICKHAM et al., 2019), *dplyr* é uma gramática para manipulação de dados que provê um conjunto consistente de verbos que ajudam a resolver os desafios mais comuns em manipulação de dados. A maior vantagem deste pacote é a forma bem definida e estruturada com a qual é possível manipular dados por meio de funções concisas e intuitivas. Como consequência de ser baseado em uma gramática, suas propriedades tornam fácil a tarefa de encadear múltiplas chamadas de funções simples para atingir um objetivo complexo.

3.2.3.2 *tidyr*

Criado por (WICKHAM; HENRY, 2019), o principal objetivo da biblioteca *tidyr* é de criar dados em um formato ideal para sua manipulação, o que os autores chamam de "dados organizados" (do inglês *tidy data*). Por definição, em um conjunto de dados "organizados" toda coluna é uma variável, toda linha é uma observação e toda célula corresponde a um único valor. Para (WICKHAM, 2014b), conjuntos de dados ditos "organizados" possibilitam uma maneira padronizada de vincular a estrutura de um conjunto de dados (sua disposição física) com sua semântica (seu significado). Este pacote desempenhou o papel principal na etapa de organização dos dados neste projeto.

3.2.3.3 *ggplot2*

Segundo (WICKHAM, 2016), *ggplot2* é um sistema para criação de gráficos baseado nos princípios da gramática dos gráficos introduzida por (WILKINSON, 2012), que por sua vez, é um sistema coerente para descrição e construção de gráficos. Assim como o pacote *dplyr* descrito na seção 3.2.3.1, uma grande vantagem de utilizar o pacote *ggplot2* é a forma concisa para construir e manipular elementos gráficos, incorporando o melhor de uma programação simples e um modelo sistemático de chamadas de funções que culminam em gráficos robustos e informativos.

3.2.3.4 *lubridate*

Trabalhar com estruturas de dados com marcação de tempo em R pode ser uma tarefa desafiadora e até mesmo frustrante. Além do mais, os métodos e funções utilizadas para manipular tais estruturas devem prever diferenças entre fuso-horários, anos bissextos,

horário de verão e etc. Desenvolvido por (GROLEMUND; WICKHAM, 2011), esta biblioteca possibilitou a manipulação de dados em formatos de data neste projeto, tornando esta uma tarefa tão simples quanto invocar algumas funções no ambiente de desenvolvimento.

3.3 Fontes dos dados

Os dados deste trabalho advém de duas fontes distintas, a saber, do *FIRMS* e do banco de queimadas do INPE, disponibilizado pelo seu Programa Queimadas. Embora utilizem-se do mesmo dispositivo (*MODIS*) para capturar os dados, devido ao tratamento intermediário que ambos os órgãos realizam antes de disponibilizá-los nas suas plataformas, esses dados não podem ser diretamente cruzados. Em contrapartida, cada um deles serve de auxílio ao outro no que concerne à validação de hipóteses acerca da discussão proposta.

Os dados obtidos diretamente da plataforma[2] da *NASA* pelo *FIRMS* podem ser baixados por meio de uma requisição ao preencher algumas informações que serão utilizadas para envio posterior dos dados ao requisitante no formato especificado (*Shapefile*, *CSV*, *JSON*, *KML*, etc.). Por outro lado, o INPE também disponibiliza informações para *download* no seu ambiente de monitoramento de focos de incêndio[3] (*Shapefile*, *CSV*, *GeoJSON*, *KML*).

E ainda, deve-se notar que, no caso dos dados do *FIRMS*, há uma distinção entre os dados de antes e depois de 01 de Agosto de 2019. Os dados anteriores a 01 de Agosto de 2019 são classificados como "Dados com Padrão de Qualidade Científica" enquanto os dados do período posterior são chamados de *NRT* (do inglês, *Near-Real Time*).

A diferença entre essas duas classes está basicamente na qualidade da precisão dos locais de incêndio (posições ou geolocalização). O *FIRMS* distribui dados de incêndios ativos quase em tempo real (*NRT*) após 3 horas da observação por satélite. O *FIRMS* leva entre 2 e 3 meses para, a partir dos dados *NRT*, disponibilizar os dados com Padrão de Qualidade Científica. Em virtude do período em que este trabalho foi desenvolvido, não foi possível obter dados com qualidade científica para os últimos 2 meses de informação analisada. Mais informações sobre a diferença entre os dois tipos de dados citados podem ser consultadas na plataforma do *FIRMS*[4].

Para fins de transparência, os dados obtidos diretamente da plataforma do *FIRMS* com qualidade *NRT* correspondem a apenas 11.77% da totalidade das observações utilizadas neste trabalho. E ainda, as informações obtidas em ambas as fontes abrangem dados no intervalo de 01 de Janeiro de 2015 até 30 de Setembro de 2019.

[2] https://firms.modaps.eosdis.nasa.gov/download/, acesso em 10 de Outubro de 2019
[3] http://queimadas.dgi.inpe.br/queimadas/bdqueimadas/, acesso em 10 de Outubro de 2019
[4] https://earthdata.nasa.gov/faq/firms-faq#ed-nrt-standard/, acesso em 01 de Novembro de 2019

3.3.1 Dados do INPE

A obtenção dos dados do banco de queimadas do INPE é direta. Como só é permitido baixar dados em um período de até 364 dias, foram baixados vários arquivos no formato *CSV* com amostras de intervalos diferentes de tempo que se complementam entre 01 de Janeiro de 2015 a 30 de Setembro de 2019. Só então todas observações foram concatenadas em um só *dataframe*. O procedimento é mostrado em seguida, na Figura 3:

```
```{r}
temp <- list.files(path = "./database/modis_15-19", pattern = "*.csv")
myfiles = lapply(paste0("./database/modis_15-19/", temp), read.csv)

data <- bind_rows(myfiles)
```
```

Figura 3 – Carregamento de dados na *IDE*

Fonte – Elaborado pelo Autor

Inicialmente, é comum construir uma visualização sobre as primeiras observações com o *dataframe* do qual partirá a análise como na Figura 4:

	datahora	satelite	pais	estado	municipio	bioma	diasemchuva	precipitacao	riscofogo	latitude	longitude	frp
1	2015/08/11 17:18:00	AQUA_M-T	Brasil	PARA	ITAITUBA	Amazonia	NA	0	0.96	-6.514	-56.060	NA
2	2015/08/14 17:48:00	AQUA_M-T	Brasil	AMAZONAS	APUI	Amazonia	NA	0	0.95	-7.010	-59.305	NA
3	2015/08/15 16:53:00	AQUA_M-T	Brasil	PARA	ALTAMIRA	Amazonia	NA	0	1.00	-6.124	-53.357	NA
4	2015/08/03 16:28:00	AQUA_M-T	Brasil	PARA	SANTA MARIA DAS BARREIRAS	Amazonia	NA	0	NaN	-8.673	-49.987	NA
5	2015/08/03 16:28:00	AQUA_M-T	Brasil	MARANHAO	MIRADOR	Cerrado	NA	0	1.00	-6.508	-44.766	NA

Figura 4 – Dados do Banco de Queimadas do INPE

Fonte – Elaborado pelo Autor

Nessa imagem, é possível visualizar todas as variáveis disponibilizadas pelo INPE. As variáveis apresentam observações espaciais (latitude, longitude, Estado e Município de detecção do foco), ecológica (Bioma), meteorológicas (quantidade de dias sem chuva e precipitação) e dados sobre a intensidade do foco de incêndio. Mais informações e a descrição detalhada sobre cada variável são fornecidas na Seção A.0.1 do apêndice A.

3.3.2 Dados do *FIRMS*

O *FIRMS* disponibiliza os dados de maneira semelhante, porém com mais possibilidades de formatos para *download* do que o INPE. A obtenção das observações, assim como no caso anterior, é direto. Exceto que neste caso, é possível baixar as informações para o período completo da análise de uma só vez, em dois arquivos diferentes de acordo com a qualidade dos dados (Padrão Científico ou *NRT*):

```r
# Carregando dados com qualidade científica padrão (jan/15 a jul/19)
nasaModis1 <-  read.csv("./database/DL_FIRE_M6_79475/fire_archive_M6_79475.csv")

# Carregando dados com qualidade NRT (01/08 a 30/09/19)
nasaModis2 <- read.csv("./database/DL_FIRE_M6_79475/fire_nrt_M6_79475.csv")
```

Figura 5 – Carregamento de dados do *FIRMS*

Fonte – Elaborado pelo Autor

Após essa primeira etapa, é importante explicitar individualmente as características de cada tipo de informação. Na Figura 6, são mostradas as primeiras linhas dos dados com Qualidade Científica Padrão:

	latitude	longitude	brightness	scan	track	acq_date	acq_time	satellite	instrument	confidence	version	bright_t31	frp	daynight	type
1	-20.2440	-40.2393	308.6	1.7	1.3	2015-01-01	143	Terra	MODIS	50	6.2	293.3	13.1	N	2
2	-20.3650	-40.9498	305.0	1.5	1.2	2015-01-01	143	Terra	MODIS	63	6.2	288.6	13.1	N	0
3	-20.2327	-40.2416	306.2	1.7	1.3	2015-01-01	143	Terra	MODIS	28	6.2	293.2	10.3	N	2
4	-18.0882	-42.7508	305.1	1.3	1.1	2015-01-01	144	Terra	MODIS	63	6.2	290.9	8.9	N	0
5	-19.9724	-41.9862	302.0	1.3	1.1	2015-01-01	144	Terra	MODIS	46	6.2	288.5	7.0	N	0
6	-14.2319	-51.6894	306.4	1.7	1.3	2015-01-01	145	Terra	MODIS	56	6.2	290.6	10.9	N	0
7	-14.2300	-51.6740	305.2	1.6	1.3	2015-01-01	145	Terra	MODIS	42	6.2	290.6	9.3	N	0

Figura 6 – Dados de Qualidade Científica Padrão do *FIRMS*

Fonte – Elaborado pelo Autor

Por outro lado, os dados de qualidade *NRT*, diferem um pouco da versão de Qualidade Científica Padrão, e são mostrados na Figura 7:

	latitude	longitude	brightness	scan	track	acq_date	acq_time	satellite	instrument	confidence	version	bright_t31	frp	daynight
1	-22.906	-43.737	308.5	1.0	1.0	2019-08-01	135	Terra	MODIS	75	6.0NRT	291.6	8.9	N
2	-22.668	-42.717	303.6	1.0	1.0	2019-08-01	135	Terra	MODIS	56	6.0NRT	287.7	7.1	N
3	-21.972	-42.926	303.4	1.0	1.0	2019-08-01	135	Terra	MODIS	55	6.0NRT	288.1	6.8	N
4	-20.789	-41.144	314.7	1.2	1.1	2019-08-01	135	Terra	MODIS	90	6.0NRT	288.5	17.5	N
5	-20.216	-41.790	308.4	1.1	1.0	2019-08-01	135	Terra	MODIS	74	6.0NRT	288.8	10.2	N
6	-14.400	-48.690	313.1	1.3	1.1	2019-08-01	135	Terra	MODIS	62	6.0NRT	294.7	14.9	N

Figura 7 – Dados de qualidade *NRT* do *FIRMS*

Fonte – Elaborado pelo Autor

Em um primeiro momento, a única diferença entre as duas bases parece ser apenas que a primeira tem a variável **type**, enquanto a segunda não. Como o período de amostragem é o mesmo, isso sugere que para concatenar os dois *dataframes*, deve-se remover a coluna excedente (**type**).

A Seção A.0.2 do apêndice A apresenta a descrição detalhada de cada variável. Nela, é possível ver que a variável **type** pode assumir 4 valores categóricos, dos quais, apenas o valor 0 representa incêndios recorrentes de áreas florestais. Com uma simples conta, viu-se dados deste tipo representam 99.2% de todo o conjunto obtido.

Portanto, ao filtrar o conjunto de dados considerando apenas as observações para as quais **type** é igual a zero, se perde apenas uma quantidade irrisória de observações. Dessa

maneira, aplicando-se tal filtragem, a coluna **type** pode ser removida. Posteriormente os dois *dataframes* foram concatenados, tornando-se apenas um.

Uma outra observação é apontada em relação às duas variáveis **bright_t31** e **brightness**, que aparentam representar a mesma quantidade, porém, são substancialmente diferentes e existem para propósitos distintos. Essas características representam a temperatura estimada no pixel que ativou o algoritmo de detecção de focos de incêndio do *MODIS*. A temperatura do brilho é na verdade uma medida dos fótons em um comprimento de onda específico recebido pela sonda, mas apresentado em unidades de temperatura.

Segundo (GIGLIO; SCHROEDER; JUSTICE, 2016), o canal 31 (**bright_t31**), recebe ondas com comprimento de cerca de 11 µm e seu propósito é a detecção ativa de incêndio, com mascaramento de nuvens, e rejeição de limpeza de florestas. Por outro lado, os canais 21/22 (**brightness**) recebem ondas com comprimento de aproximadamente 4 µm e seu propósito é ser um canal de baixo e longo alcance para detecção ativa de incêndios.

4 ANÁLISE DE DADOS

O objetivo de uma análise exploratória de dados é permitir ao analista obter respostas para perguntas ou suposições iniciais do problema. Assim, ao longo do processo de análise, as perguntas são refinadas e/ou novos dados são coletados para validar as respostas encontradas. Tudo isso ocorre iterativamente.

Nesta seção são mostradas algumas abordagens no processo de questionamento e conhecimento dos dados e, a medida que novas questões são levantadas, busca-se respondê-las através de métodos comumente utilizados na prática da análise de dados.

Primeiro, analisaremos apenas os dados da base de queimadas do INPE. Em um segundo momento, faremos o mesmo para os dados do *FIRMS*, a fim de obter novas descobertas e validar algumas conclusões feitas no primeiro momento.

4.1 Uma análise exploratória sobre os dados do banco de queimadas do INPE

Nesta seção, será descrito o processo de reconhecimento e exploração dos dados do banco de queimadas do INPE.

4.1.1 Estruturas e tipos de dados

O *R* possui tipagem dinâmica de variáveis. Em outros palavras, ele atribui os tipos das variáveis em tempo de execução do programa. Contudo, é importante ter uma primeira noção sobre quais estruturas de dados estão sendo utilizadas para cada variável a fim de que se tenha conhecimento de quais operações são possíveis em cada uma e até mesmo entre elas. A seguir, são mostradas as estruturas de dados iniciais de cada variável do *dataframe*, que é uma estrutura de dados comum e tabular para armazenar variáveis e manipular dados:

```
'data.frame':   885117 obs. of  12 variables:
 $ datahora    : chr  "2015/08/11 17:18:00" "2015/08/14 17:48:00" "2015/08/15 16:53:00" "2015/08/03 16:28:00" ...
 $ satelite    : Factor w/ 1 level "AQUA_M-T": 1 1 1 1 1 1 1 1 1 1 ...
 $ pais        : Factor w/ 1 level "Brasil": 1 1 1 1 1 1 1 1 1 1 ...
 $ estado      : chr  "PARA" "AMAZONAS" "PARA" "PARA" ...
 $ municipio   : chr  "ITAITUBA" "APUI" "ALTAMIRA" "SANTA MARIA DAS BARREIRAS" ...
 $ bioma       : Factor w/ 6 levels "Amazonia","Caatinga",..: 1 1 1 1 3 3 3 3 3 3 ...
 $ diasemchuva : int  NA NA NA NA NA NA NA NA NA NA ...
 $ precipitacao: num  0 0 0 0 0 0 0 0 0 0 ...
 $ riscofogo   : num  0.96 0.95 1 NaN 1 1 1 1 1 1 ...
 $ latitude    : num  -6.51 -7.01 -6.12 -8.67 -6.51 ...
 $ longitude   : num  -56.1 -59.3 -53.4 -50 -44.8 ...
 $ frp         : num  NA NA NA NA NA NA NA NA NA NA ...
```

Figura 8 – Estruturas de dados do INPE

Fonte – Elaborado pelo Autor

Instantaneamente, é possível notar algumas anormalidades de acordo com o que se espera pela natureza das variáveis, por exemplo:

1. A coluna de **datahora** é do tipo *character*, quando na verdade, o mais natural seria o tipo *datetime*;
2. As colunas **satelite** e **pais**, são do tipo *factor* e possuem apenas um nível, indicando que assumem apenas um valor. Em outras palavras, só é possível tirar uma única informação de cada uma delas, isto implica que pode-se dispensá-las tendo em vista que essa operação nos ajudará a reduzir a complexidade do processo de análise;
3. As variáveis **diasemchuva** e **frp** aparentemente contam com diversas entradas indisponíveis (NA). É necessário que se verifique essa suposição com mais profundidade.

Dadas estas considerações, o tipo de **datahora** foi convertido para *datetime*, e as colunas **satelite** e **pais** foram removidas do *dataframe*.

4.1.2 Completude de dados

A fim de se validar a última hipótese levantada na seção 4.1.1 sobre a completude das variáveis **diasemchuva** e **frp**, foi desenvolvido um código simples para aferir a proporção de entradas nulas nessas colunas. O código e o resultado são mostrados na Figura 9.

```r
n_linhas <- nrow(data)
noWeatherInfo <- is.na(data$diasemchuva)
noFrpInfo <- is.na(data$frp)

paste("Porcentagem de entradas nulas em `FRP` =",
      round(100 * sum(noFrpInfo) / n_linhas, 2), "%")
paste("Porcentagem de entradas nulas em `diasemchuva` =",
      round(100 * (sum(noWeatherInfo)) / n_linhas, 2), "%")
```

```
[1] "Porcentagem de entradas nulas em `FRP` = 68.75 %"
[1] "Porcentagem de entradas nulas em `diasemchuva` = 45.36 %"
```

Figura 9 – Porcentagem de entradas nulas em **frp** e **diasemchuva**

Fonte – Elaborado pelo Autor

Como demonstrado, **frp** tem quase 70% de ausência de dados, enquanto **diasemchuva** tem pouco mais de 45%. A fim de afunilar o escopo da análise, a coluna de **frp** foi desconsiderada do conjunto de dados.

Vale salientar que as demais variáveis também demostraram uma completude próxima ou igual a 100%.

4.1.3 Estatísticas descritivas

Uma outra operação comum a se fazer quando em contato com os dados em um primeiro momento, é calcular algumas estatísticas sumárias sobre os mesmos. Assim, é possível identificar propriedades como dispersão, assimetria, escala, completude e distribuição.

A Tabela 1, contém algumas estatísticas sumárias para as variáveis numéricas. É então possível perceber que o número total de observações é de 885 117 e a média de dias sem chuvas até a detecção de um foco incêndio é de 9,85 dias, e que a precipitação máxima diária registrada em um local de incêndio foi de 309,8 mm de água. Além disso, note que a variável **riscofogo** varia entre 0 e 1, indicando a sua escala, em que, quanto maior o índice, maior era o risco de incêndio na área detectada.

Estatísticas Sumárias ($N = 885,117$)			
	diasemchuva	**precipitacao**	**riscofogo**
min	0	0	0
mediana (I. Interq.)	0 (0,00; 8,00)	0,00 (0,00; 0,00)	1,00 (0,37; 1,00)
média (desv. padr.)	9,85 ± 21,61	1,08 ± 6,14	0,71 ± 0,38
máx	120	309,8	1

Tabela 1 – Tabela de estatísticas sumárias da base de dados de queimadas do INPE

Adicionalmente, vale salientar que as estatísticas mostradas na tabela 1 são para todas as observações, ou melhor, para todos os biomas. Algo natural a se perguntar a

partir dessas informações é **qual a participação de cada bioma no conjunto total das observações**.

A Figura 10 destaca a quantidade relativa de observações por bioma. Nela é possível ver que o bioma Amazônico e Cerrado correspondem juntos a mais de 80% da ocorrência de incêndios florestais em todo o período analisado. Em quantidades, isso representa 727 394 registros em um total de 885 117.

Decorrente dessas informações, pode-se perceber que essas estatísticas são coerentes pelo fato de que a Amazônia ocupa 49,5% de todo o território brasileiro, seguido pelo Cerrado (23,3%), Mata Atlântica (13%), Caatinga (10,1%), Pampa (2,3%) e Pantanal (1,8%). É natural que se chegue a estes números.

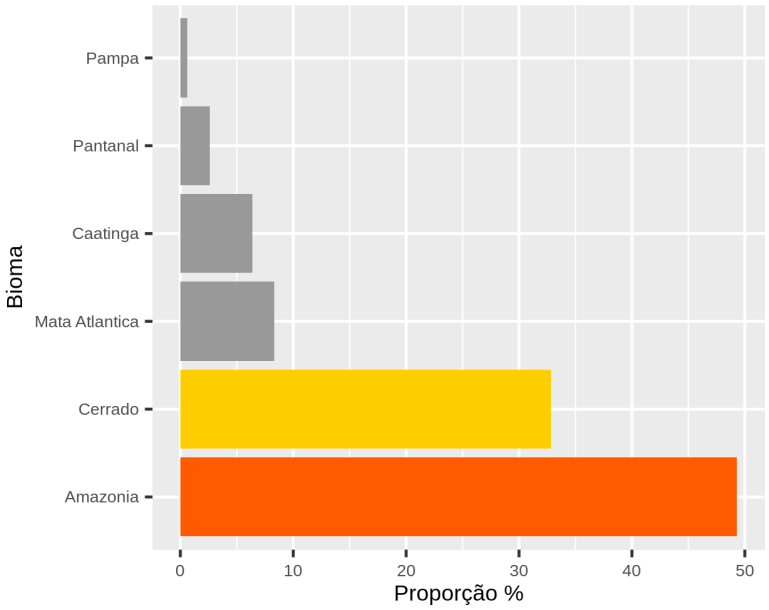

Figura 10 – Proporção de observações por Bioma

Fonte – Elaborado pelo Autor

A discussão sobre estatísticas sumárias como um meio de fornecer um conhecimento geral sobre as propriedades matemáticas mais básicas não se esgota apenas aos dois tipos de ações exemplificadas aqui. Ainda há diversas abordagens e métricas diferentes que podem ser exploradas para prover ideias e *insights* durante esta fase do projeto.

4.1.4 Sazonalidade temporal

No campo de ideias para a proposta deste trabalho, a pergunta inicial levantada sobre os dados foi acerca da periodicidade de ocorrências de incêndios florestais. **Afinal, existe algo como uma estação com maior ocorrência de incêndios?**

Nesta seção, será elucidada a existência ou não de tal fenômeno. Numa primeira

tentativa para encontrar padrões de periodicidade, foi construído o gráfico da figura 11 com a quantidade diária de incêndios florestais no período analisado.

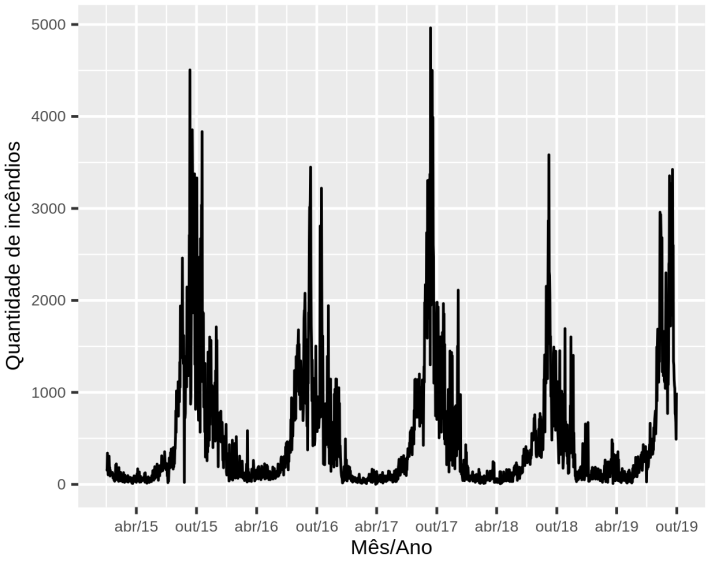

Figura 11 – Quantidade de incêndios por dia

Fonte – Elaborado pelo Autor

Nitidamente, os dados apresentam padrões que indicam periodicidade, isto é, que as ocorrências de incêndios são sazonais levando em conta todos os biomas. No entanto, não pode-se repousar apenas capacidade visual humana, que por vezes, pode ser enganosa.

Uma abordagem alternativa é a análise no domínio da frequência através da transformada de Fourier. A Transformada de Fourier decompõe um sinal em todas as possíveis frequências contidas nele. Assim, tratando a série temporal da Figura 11 como um sinal, pode-se, por meio da função de densidade espectral de potência do sinal, a informação que indica qual a periodicidade das ocorrências de incêndio.

```r
sinal_INPE <- data %>%
  mutate(date = date(datahora)) %>%
  group_by(date) %>%
  summarise(n = n())

p <- periodogram(sinal_INPE$n, plot = F)

dd = data.frame(freq=p$freq, spec=p$spec)
order = dd[order(-dd$spec),]
top3 = head(order, 3)

# frequências com 3 maiores potências
top3
```

	freq <dbl>	spec <dbl>
5	0.002777778	413000315
10	0.005555556	99518858
15	0.008333333	25265475

3 rows

Figura 12 – Código para obtenção das frequências de maior relevância do sinal

Fonte – Elaborado pelo Autor

Da Figura 12, conclui-se que as 3 frequências de maiores potências são 0.002 77 Hz, 0.005 55 Hz e 0.008 33 Hz. O que, pela equação,

$$T = \frac{1}{f} \tag{4.1}$$

Implica que os três períodos mais preponderantes no sinal são aproximadamente 360 dias, 180 dias, e 120 dias, respectivamente. A função de densidade espectral do sinal é apresentada na Figura 13. Pela imagem, a frequência de 0.002 77 Hz expressivamente domina o cenário. Assim, pode-se afirmar categoricamente que **as observações para todos os incêndios florestais no período de tempo analisado ocorrem com periodicidade anual**.

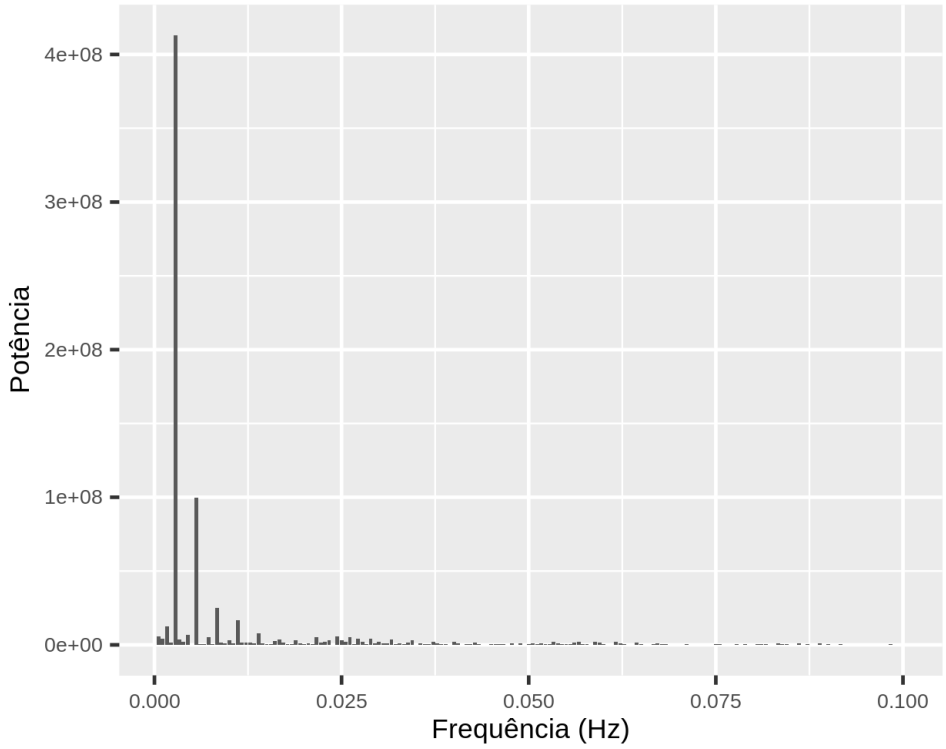

Figura 13 – Densidade Espectral de Potência da série temporal de incêndios por dia

Fonte – Elaborado pelo Autor

Uma pergunta natural que surge após a conclusão anterior é sobre **quais meses apresentam maior incidência de focos de incêndios?**

A Figura 14 mostra a quantidade de incêndios a cada mês de todos os anos analisados. Através dela, fica claro que a alta estação de incêndios acontece entre Junho e Novembro.

Sendo assim, pode-se cruzar as informações das Figuras 10 e 14, com isso, obtem-se o gráfico da Figura 15, que mostra que a sazonalidade temporal existe, e que ela também é consistente entre cada bioma.

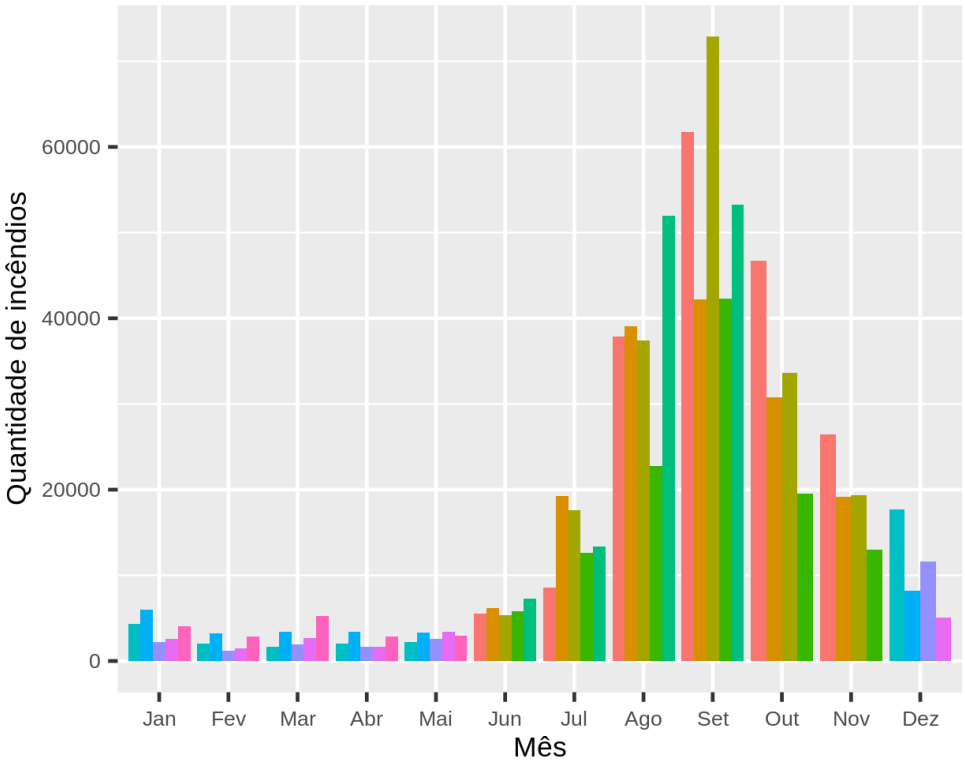

Figura 14 – Quantidade de incêndios por mês

Fonte – Elaborado pelo Autor

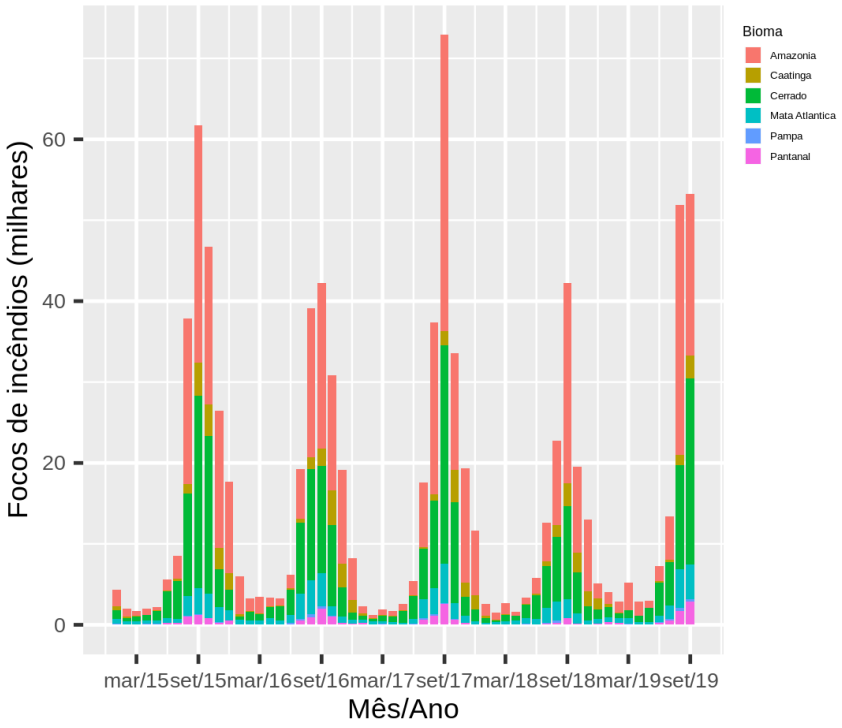

Figura 15 – Participação de cada bioma na quantidade mensal de incêndios

4.1.5 Uma visão por Estado

O gráfico da Figura 15 sugere uma sazonalidade temporal e espacial sobre a incidência de incêndios florestais. Informações que o banco de queimadas do INPE fornece adicionalmente àquelas do *FIRMS* são a respeito do Estado e Município onde ocorreu o registro de incêndio.

Aproveitando-se dessas variáveis adicionais, podemos também buscar um entendimento sobre como são distribuídos o número de incêndios por Estado. Ações como essas possibilitam a entidades governamentais agir sobre as áreas mais suscetíveis ao fogo e a encontrar potenciais causas de incêndios bem como compreender o fluxo rural local que, por vezes, também é responsável por grandes incêndios para lavoura.

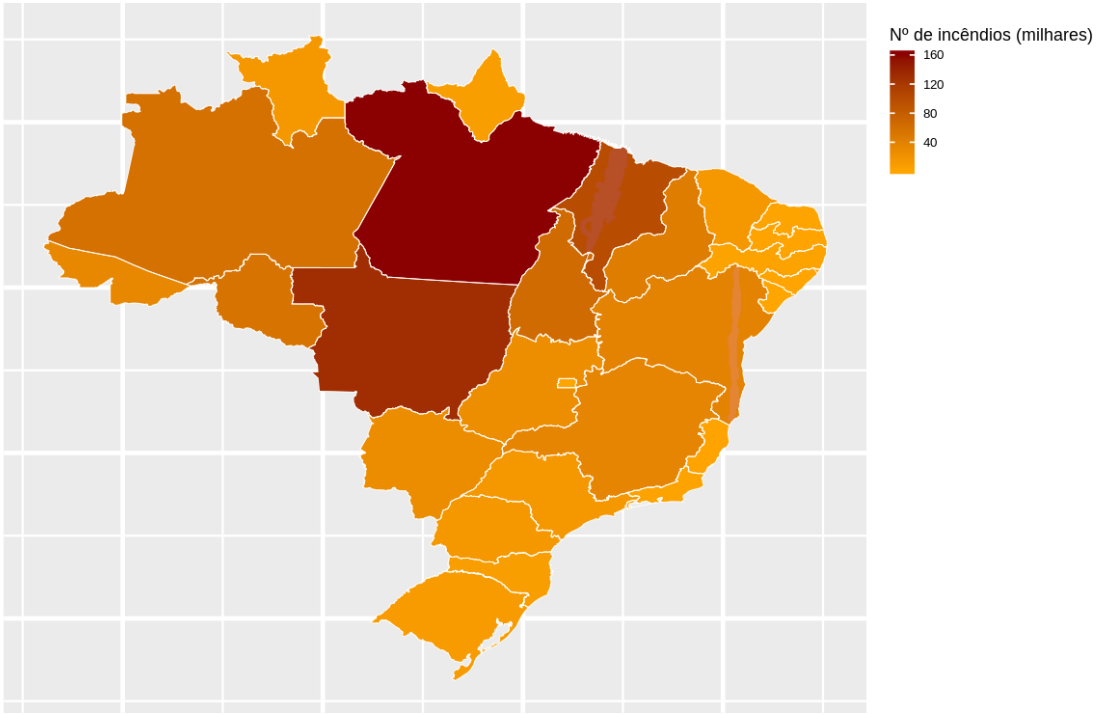

Figura 16 – Número de incêndios por Estado

Fonte – Elaborado pelo Autor

Pela Figura 16, vê-se que os Estados com maior frequência de queimadas são Pará e Mato Grosso. Isto é condizente com o esperado já que se sabe de antemão que o bioma amazônico é o mais atingido, e ele está presente nesses Estados. Um outro ponto a notar é que estes Estados tem grande atividade agropecuária, o que também justifica uma porção da quantidade de incêndios detectados nesses regiões. Algo que pode-se verificar é se correto afirmar que durante todos os anos **o padrão de distribuição de incêndios por estado é o mesmo**.

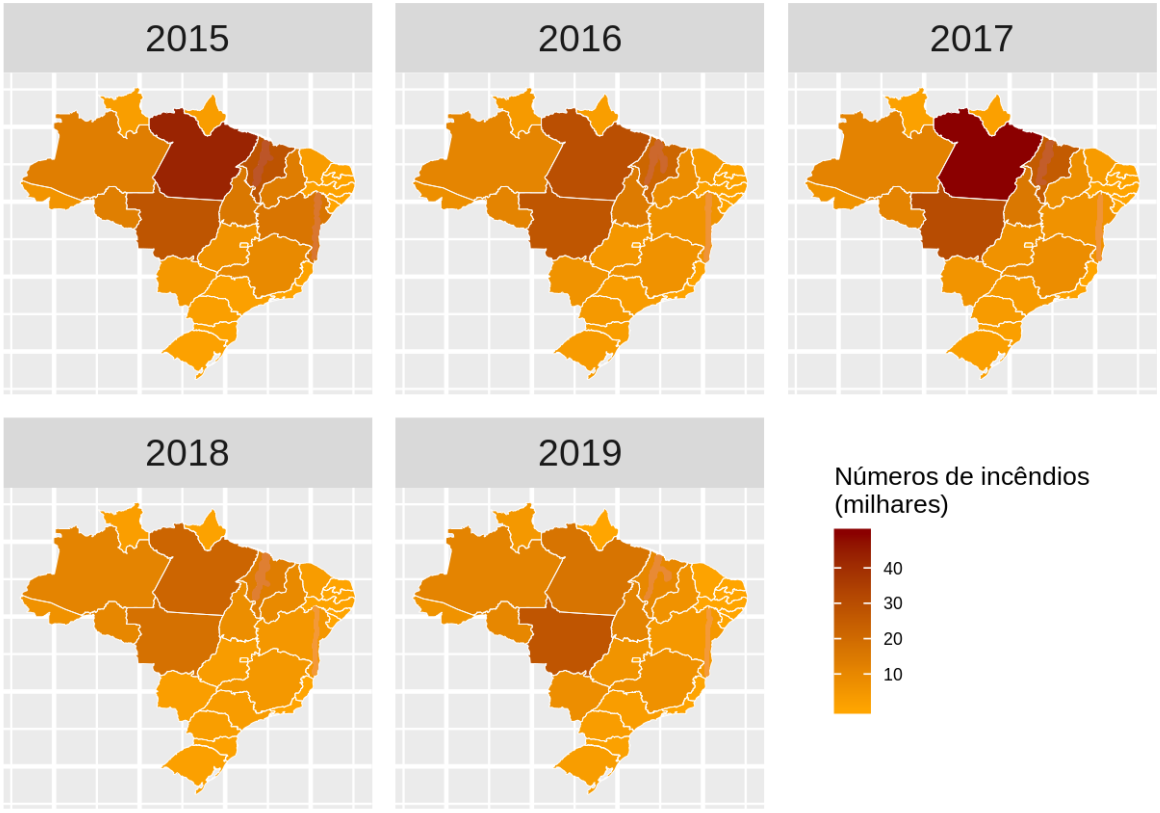

Figura 17 – Quantidade de incêndios em cada estado de 2015 a Setembro de 2019

Fonte – Elaborado pelo Autor

Pela Figura 17 percebe-se que a distribuição de incêndios por estado ao longo dos anos é aproximadamente o mesmo. Apesar de uma redução relativa no ano de 2016 e 2018, o Pará é o Estado com maior número de ocorrências de fogo.

Considerando que o Pará é o Estado com maior registro de focos, um questionamento comum a se fazer em uma análise de dados com esse caráter é sobre **quais municípios são os mais atingidos**.

Na Figura 18 é mostrado o ranking dos 10 municípios mais atingidos. Dentre os 10, 3 estão no estado do Pará, do qual o segundo, São Félix do Xingu, que fica localizado na região sudeste do Estado, abriga em seu território a reserva ambiental Triunfo do Xingu, o que, por si só, é uma informação preocupante.

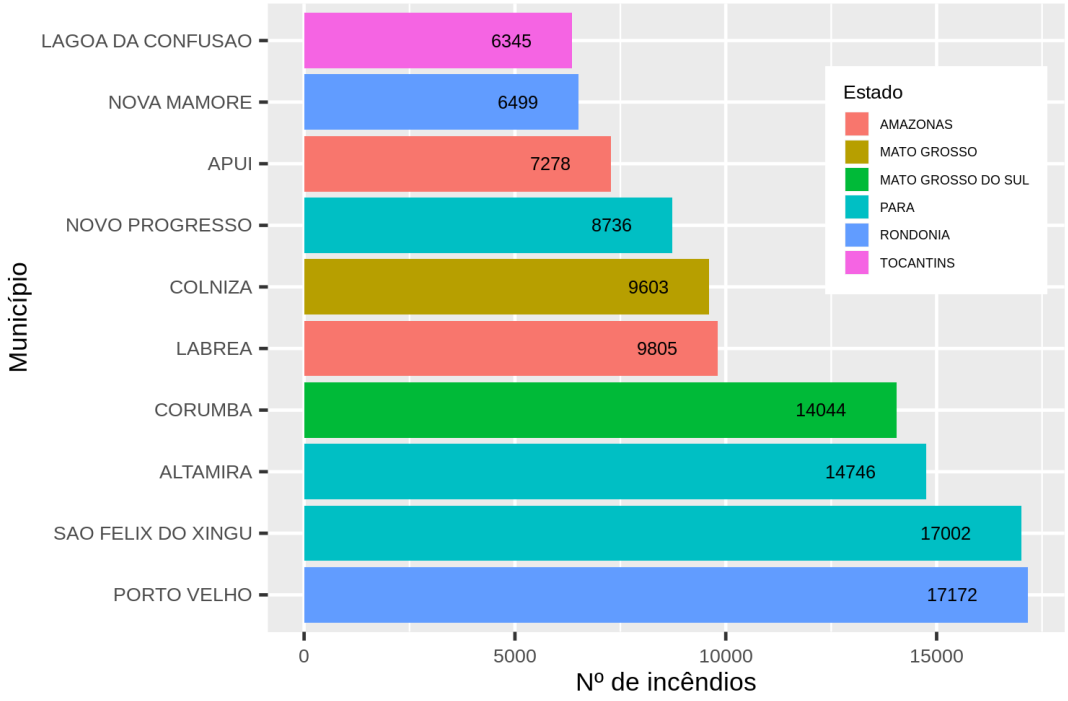

Figura 18 – 10 municípios com maior ocorrência de focos de incêndio entre Jan/2015 e Set/2019

Fonte – Elaborado pelo Autor

4.1.6 Sazonalidade Espacial: Amazônia Legal

Até aqui foi mostrado que a incidência de focos de incêndio aumenta em um determinado período do ano e também foi fornecida uma visão geral por Estado com relação a contagem bruta de detecções em todo o período analisado. No entanto, pode-se ser mais preciso nesse sentido.

Identificar exatamente quais as regiões mais suscetíveis pode nos fornecer ainda mais ideias de intervenção e prosseguimento com a análise. Desta vez, o escopo de exploração tem enfoque na região da Amazônia Legal, a qual foi introduzida introduzida na Seção 2.3.

Delimitando essa região pela área contida dentro do polígono de bordas amarelas, a Figura 19 mostra a distribuição espacial de ocorrências de incêndio na vegetação dentro e fora da Amazônia Legal para todo o período de Janeiro a Setembro de 2019.

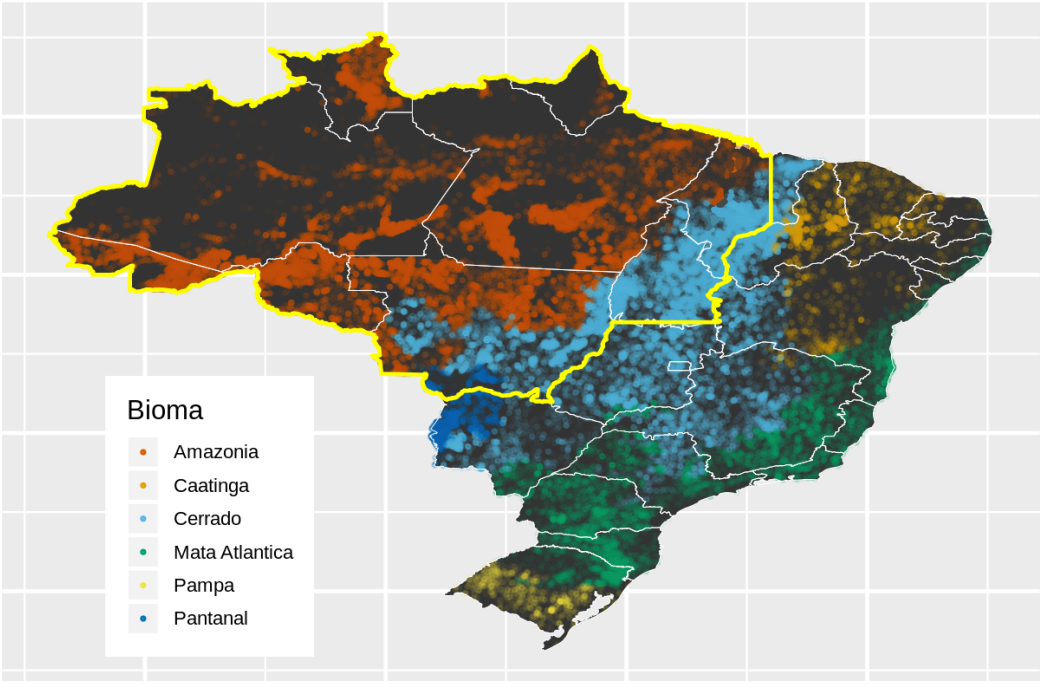

Figura 19 – Distribuição de focos de incêndios no Brasil e na Amazônia Legal entre Janeiro e Setembro de 2019

Fonte – Elaborado pelo Autor

A partir da Figura 19, podem surgir questionamentos como: **Qual a proporção total de incêndios que ocorreram na Amazônia Legal nos primeiros 9 meses de 2019?** 68,96% das detecções (99 116 de 143 734) fornecidas pelo INPE estão contidas dentro dessa região.

Da Figura 19, pode-se perceber que a grande maioria das observações estão contidas na região dos Estados do Mato Grosso, Tocantins e Pará, sendo predominante na região Sul-Sudeste do Pará e no corredor entre o Nordeste de Mato Grosso e Leste do Maranhão, que é a região chamada de "Arco do Desflorestamento".

Em contrapartida, para termos um referencial do que essa proporção representa, a Figura 20 concebe a mesma estatística para os anos de 2015 a 2019. Nela é possível perceber que a tendência espacial também se mantém, apontando assim, potenciais regiões onde projetos de intervenção poderiam causar o maior impacto para a redução de queimadas ilegais.

A contagem acumulada relativa do número de observações contidas dentro da região da Amazônia Legal em Janeiro e Setembro foi de 70% em 2015, seguido de 71% em 2016 e 2018, 75% em 2017 e 69% em 2019.

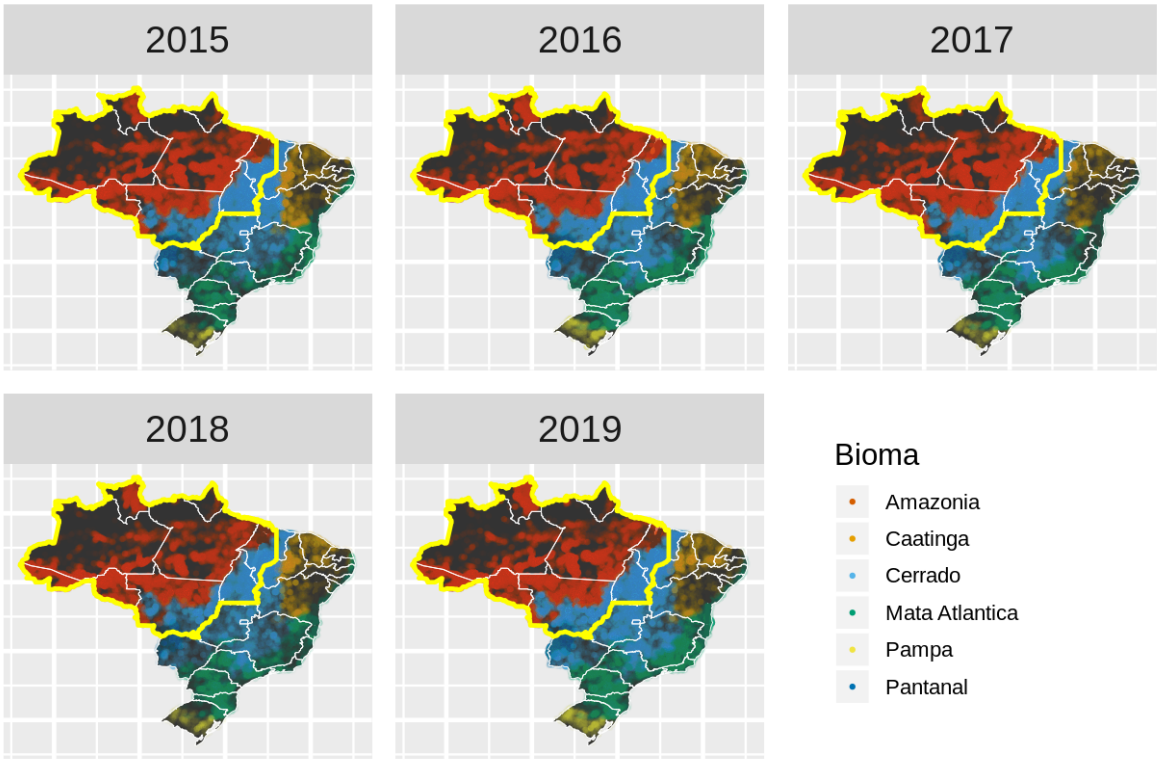

Figura 20 – Distribuição de focos de incêndios no Brasil e na Amazônia Legal entre Janeiro e Setembro de 2015 a 2019

Fonte – Elaborado pelo Autor

4.1.7 Relação entre chuva e fogo

Algumas variáveis ainda não foram profundamente exploradas até aqui, como **riscofogo** e **precipitacao**. Embora a investigação seja voltada ao caráter periódico das detecções de focos de incêndios, nessa Seção será abordado um pouco do que pode-se aprender a partir dessas variáveis acerca da temática de ocorrência de focos de incêndios.

A seguir, é mostrado na Figura 21, um gráfico de dispersão que objetiva esclarecer a relação entre essas duas variáveis.

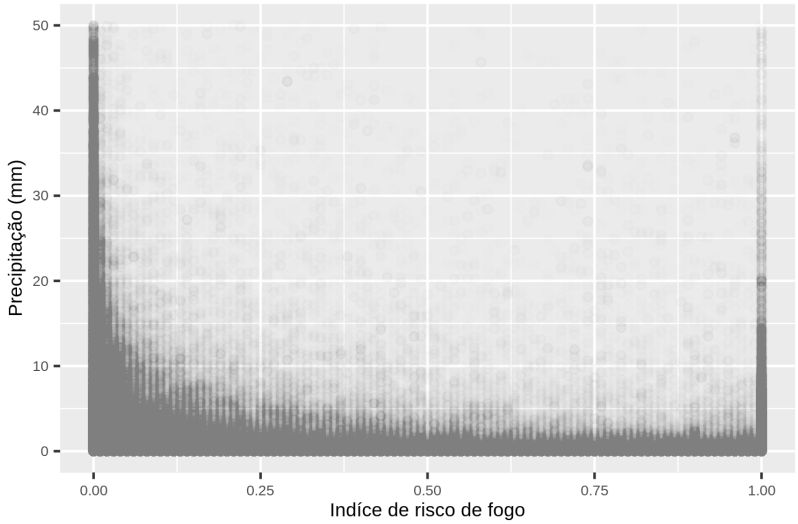

Figura 21 – Gráfico de dispersão risco de fogo vs precipitação (mm)

Fonte – Elaborado pelo Autor

Do ponto de vista da análise, na Figura 21 é possível perceber uma situação interessante. O conjunto de dados contêm algumas centenas de milhares de observações, para poder mostrar todos os pontos, como há muita sobreposição, utilizou-se o recurso de adicionar transparência aos pontos, assim, as regiões de maior sobreposição se destacam. No entanto, para este caso, o melhor a se fazer é construir um gráfico bidimensional de densidades. Como é mostrado na Figura 22.

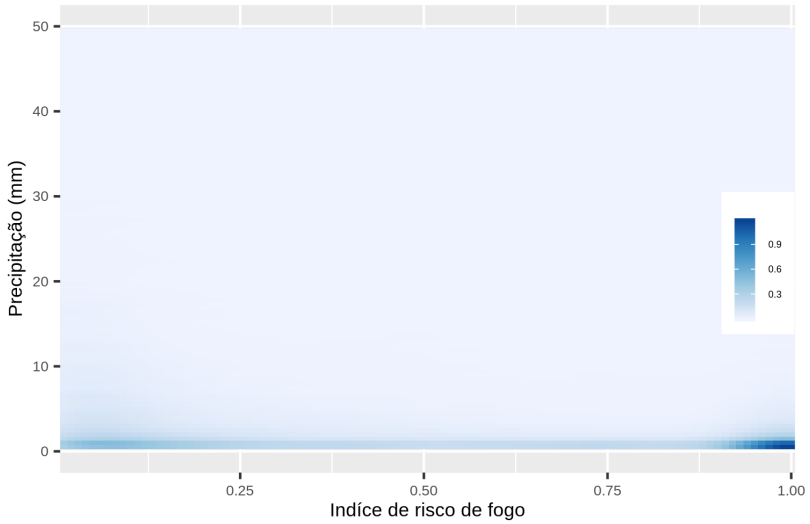

Figura 22 – Gráfico de densidade: risco de fogo vs precipitação (mm)

Fonte – Elaborado pelo Autor

A partir do gráfico de densidade da Figura 22, fica ainda mais claro que quanto menor a previsão de precipitação maior é o risco de fogo no local do incidente. Apesar

disso parecer óbvio, tal relação não poderia ser notada se olhada apenas pelo espectro de visão da Figura 21.

4.2 Uma análise exploratória sobre os dados do *FIRMS* (*NASA*)

Nesta seção, vamos explorar o conjunto de dados do *FIRMS* da *NASA*. Aqui, nos dedicaremos principalmente a responder as perguntas que não poderiam ser respondidas (ou teriam pouco grau de significância estatística) com os dados do banco de queimadas do INPE.

Um apontamento inicial vai para a variável **frp** que, na Seção anterior, não pôde ser explorada devido ao seu alto índice de ausência dentre as observações. No entanto, nenhuma das variáveis do conjunto de dados do *FIRMS* apresentam valores nulos.

4.2.1 Restringindo o escopo de observações e variáveis

A grosso modo, os dados do INPE provêm da mesma fonte, no entanto, não é revelado exatamente quais os procedimentos aplicados até a disponibilização dos mesmos em sua plataforma online.

O que chega a ser intrigante neste caso, é que o número de entradas dos dados do *FIRMS* é 1 637 180, quase o dobro do número de dados do INPE.

Como se tratam de muitos dados, considerando também a resolução de 1 km^2 do produto *MODIS*, buscou-se maneiras de aplicar filtros para reduzir a complexidade de análise:

1. A variável **confidence** denota o nível de confiança de que o disparo é realmente um incêndio. Para considerar apenas os casos mais precisos, foi aplicado um filtro para manter apenas as observações tais que **confidence** $> 50\%$.

2. A variável **type** descreve o tipo inferido de incêndio (se em solo, ou água, ou detecção de atividade vulcânica, por exemplo). Os dados foram filtrados para manter apenas aqueles para os quais **type** $= 0$, como comentado na seção 3.3.2. Isto é, incêndio de vegetação (ver Tabela 3, apêndice A).

3. As variáveis **scan** e **track**, servem apenas como parâmetros para descrever o tamanho real dos píxeis de detecção, que aumentam em direção à borda da captura. Isto é, os píxeis nas bordas "Leste"e "Oeste"da captura são maiores que 1 km^2. É 1 km^2 apenas ao longo do nadir (vertical exato do satélite). Considerando que na média, esse tamanho é o padrão (1 km^2 quadrado) e ele não varia para muito mais que essa medida, estas variáveis serão desconsideradas.

4. As variáveis, **satellite**, **instrument** e **version**, não são relevantes para a análise proposta pois se tratam apenas de informações técnicas sobre os equipamentos de detecção. Portanto, serão desconsideradas.

Após todas as considerações, restaram 79,15% do total de observações iniciais, o que em valores representam um montante de 1 295 786 observações (esta quantidade ainda corresponde a mais de 46% da quantidade lida do banco de queimadas do INPE), e um conjunto de dados com as seguintes características:

	latitude	longitude	brightness	acq_date	acq_time	confidence	bright_t31	frp	daynight
1	-20.3650	-40.9498	305.0	2015-01-01	143	63	288.6	13.1	N
2	-18.0882	-42.7508	305.1	2015-01-01	144	63	290.9	8.9	N
3	-14.2319	-51.6894	306.4	2015-01-01	145	56	290.6	10.9	N
4	-14.2206	-51.6810	311.4	2015-01-01	145	67	290.5	17.2	N
5	-11.5154	-49.6146	309.4	2015-01-01	146	77	292.3	9.7	N
6	-11.5167	-49.6250	307.7	2015-01-01	146	72	292.2	8.5	N

Figura 23 – Conjunto de dados do *FIRMS* após restrição de escopo

Fonte – Elaborado pelo Autor

4.2.2 Intensidade e ocorrência de incêndios ao longo dos anos

Devido a ausência de observações, na Seção 4.1.4 não foi possível explorar uma variável extremamente importante à análise da problemática proposta neste trabalho, o **frp** (poder radiativo de fogo). Essa variável literalmente descreve qual a intensidade dos focos de incêndios detectados. A Figura 24 mostra a quantidade de incêndios acumulada ao logo de todos os anos disponíveis nos dados.

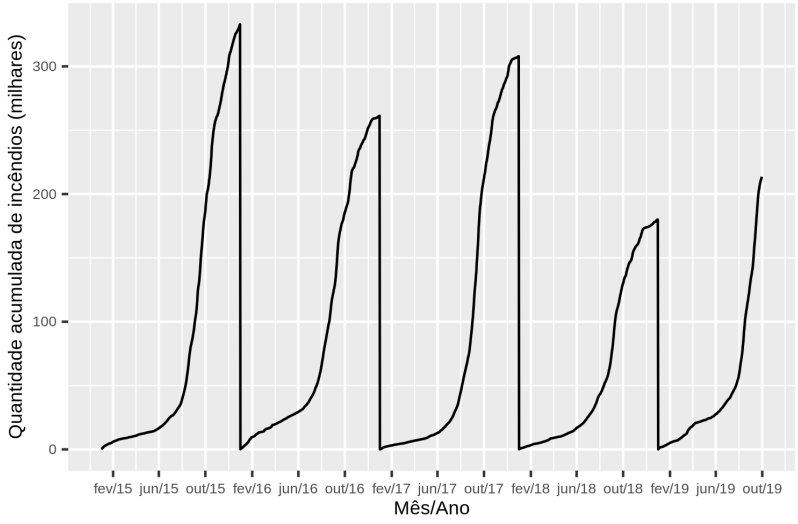

Figura 24 – Curva cumulativa de incêndios por dia

Fonte – Elaborado pelo Autor

Pela Figura 24, é possível ver que o ano com maior número bruto de detecções foi 2015, com um crescimento bastante acentuado na segunda metade do ano, quando comparado com o segundo ano com maior número de ocorrências, 2017. Por outro lado, percebe-se que embora os dados de 2019 correspondam apenas ao período de 01 de Janeiro a 30 de Setembro, a quantidade de focos detectados já ultrapassou o total registrado no ano anterior. Isto é, para todo o ano de 2018 e de Janeiro a Setembro de 2019, o total de incêndios passou de 180 045 para 213 576, o que representa um aumento de 16,62%.

Todavia, nessa situação a quantidade bruta de detecções por si só pode ser enganosa pois não considera o poder destrutivo de cada incêndio. Por exemplo, 1 incêndio grande sozinho pode ter mais impacto destrutivo do que vários menores.

A fim de contornar essa situação, uma curva acumulada do poder radiativo ao longo dos dias de cada ano mostraria exatamente em qual ano houve mais destruição de vegetação causada por fogo, proposital ou não. Tal visualização é fornecida pela Figura 25.

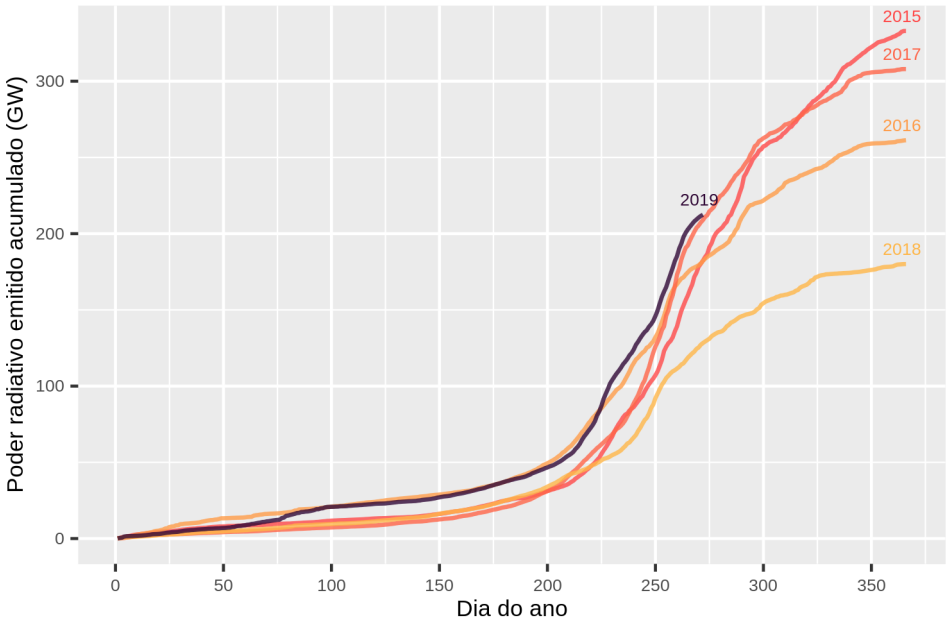

Figura 25 – Curva cumulativa de poder radiativo de fodo emitido por dia

Fonte – Elaborado pelo Autor

Pela Figura 25, até 30 de Setembro de 2019, o número de queimadas detectadas já havia superado todos os anos anteriores em poder radiativo emitido.

4.2.3 Correlação entre poder radiativo de fogo e temperatura estimada de detecção (considerações sobre *outliers*)

Até o momento, discutiu-se sobre o papel da análise da variável de poder radiativo de fogo. No entanto, esta variável e seu significado pode parecer um pouco abstrato do

ponto de vista físico. Algo mais palpável seria relacioná-la diretamente com algo que vive-se no dia a dia, por exemplo, medidas de temperatura. A variável **brightness** parece ótima para nos contextualizar a esse respeito, uma vez que se trata de uma medida de temperatura em uma unidade conhecida, Kelvin.

Primeiro, vale notar que **frp** tem diversas entradas que podem ser consideradas *outliers*, como é possível observar pela Figura 26. Devido alta taxa de dispersão das medidas, a variável foi transformada para uma escala logarítmica. Esse método é comumente aplicado nas análises exploratórias para aproximar as observações de variáveis que contém um alto valor de variância/desvio padrão.

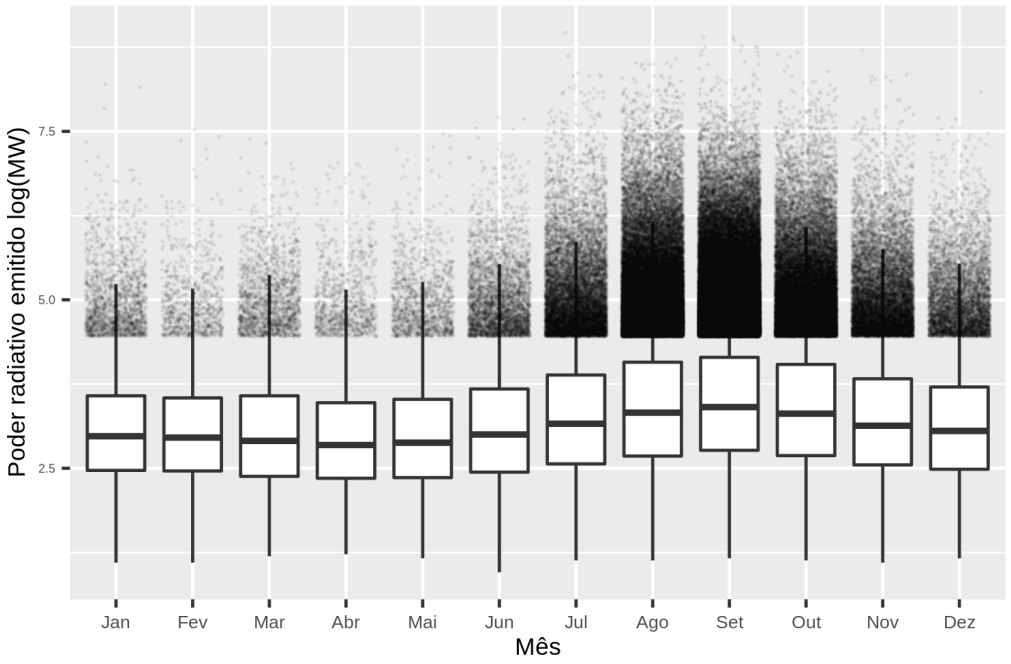

Figura 26 – Boxplot de Poder Radiativo de Fogo para cada mês do ano

Fonte – Elaborado pelo Autor

A Figura 26 transmite a informação de que incêndios de grande magnitude ocorrem justamente na alta estação. Para a análise de correlação entre as variáveis **frp** e **brightness** fazer sentido, é necessário remover os *outliers* do conjunto de dados nas duas variáveis, principalmente na primeira. *Outliers* são consideradas aquelas observações as quais seus valores estão fora do Intervalo interquartil multiplicado 1,5 vezes.

Assim, o processo de remoção de *outliers* para as duas variáveis foi análogo, primeiro foi calculado o intervalo interquartil ($IQR = Q_3 - Q_1$). Depois, com a informação dos quartis Q_1 e Q_3, foram removidas todas as observações fora do intervalo $[Q_1 - 1,5IQR; Q_3 + 1,5IQR]$.

Como é possível observar nas Figuras 27 e 28, as duas variáveis estão intrinsecamente ligadas, de tal modo que apresentam um relacionamento linear quase perfeito, com um

coeficiente de correlação de 0,904.

Este é um exemplo de como pode-se explorar a relação entre variáveis para entender um fenômeno, ou o que uma delas representa em contextos físicos abstratos ou palpáveis.

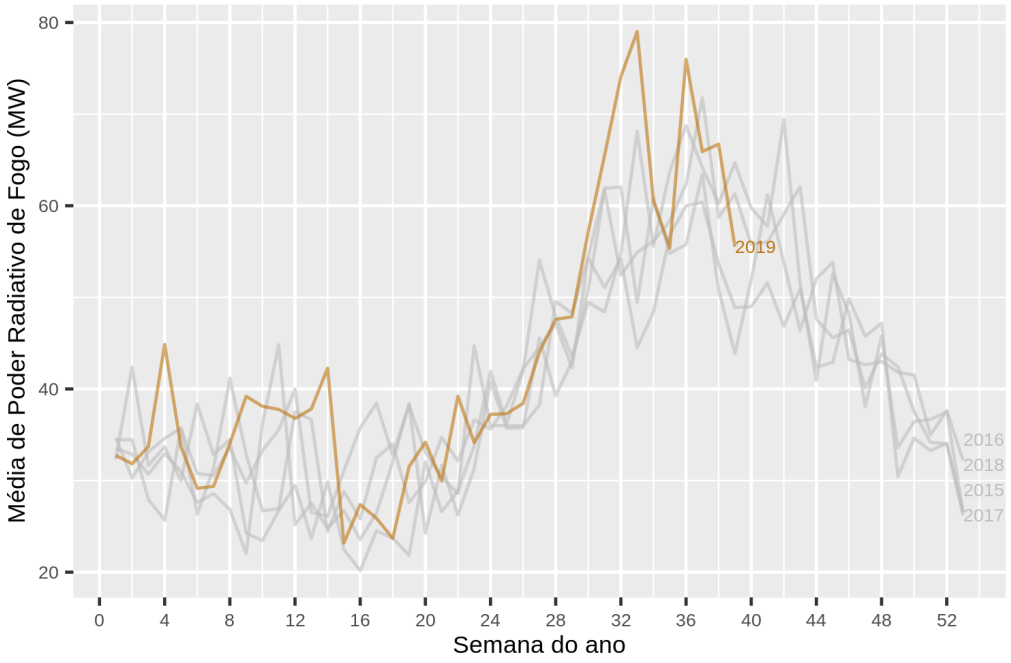

Figura 27 – Poder Radiativo de Fogo Médio por semana

Fonte – Elaborado pelo Autor

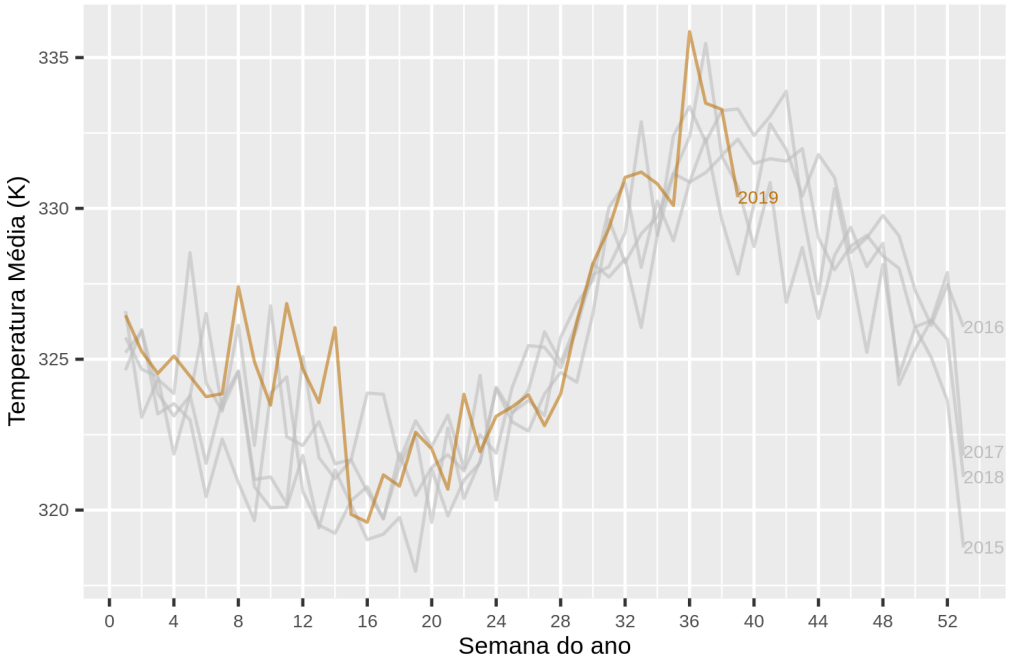

Figura 28 – Temperatura Média estimada nos píxeis de detecção por semana

Fonte – Elaborado pelo Autor

5 DISCUSSÃO

Apesar dos resultados expostos até então, o objetivo deste trabalho não é esgotar a discussão sobre o tema abordado pela análise de dados. De fato, a própria finalidade de uma análise não é responder a todos os questionamentos levantados, mas promover um entendimento amplo acerca do objeto de pesquisa. Esta seção propõe uma discussão acerca dos produtos e ideias repercutidas no Capítulo 4.

Diante da difusão midiática acerca da agressão ambiental praticada à floresta Amazônica nos últimos meses. A questão motivadora em torno da temática abordada propaga-se em volta da frequência com que a floresta tropical Amazônica sofre com incêndios. Nas últimas décadas, devido ao manejo humano (rural ou indígena) da terra utilizando o fogo como ferramenta e a secas de grandes magnitudes, as florestas tropicais se tornaram mais suscetíveis a incêndios florestais recorrentes, proporcionando consequências negativas para a biodiversidade (SILVEIRA et al., 2013).

(SILVEIRA et al., 2016) mostra que incêndios florestais recorrentes causam impactos bastante negativos na estrutura de florestas, reduzindo a biomassa de árvores e mudas vivas. Tais efeitos não se estendem apenas à flora, mas também à fauna desses ecossistemas, promovendo a redução da riqueza, biodiversidade e composição de espécies animais.

No Capítulo 4, utilizando dados de dois órgãos distintos, mostrou-se que há um forte componente sazonal (temporal e geográfico) na ocorrência de incêndios florestais (Figuras 11-14 e 17). Adicionalmente, também foi apontado pela Figura 20 que mais de 2/3 do total de incêndios nos primeiros 9 meses de cada ano desde 2015 ocorrem na região da Amazônia Legal. Esta informação, agregada ao que foi dito no parágrafo anterior, convida não só entidades governamentais como também a nós, cidadãos, a fomentarmos posturas mais responsáveis em direção à conservação e proteção de nossas florestas.

Na direção oposta a tal comportamento, as Figuras 24 e 25 reforçam que entre Janeiro e Setembro de 2019, incêndios florestais registrados nos 5 biomas em todas as regiões brasileiras já haviam causado mais destruição ambiental do que em todo período do ano anterior.

Por meio de um estudo examinando regiões da floresta Amazônica atingidas por fogo há mais de 10 anos antes, (SILVEIRA et al., 2013) mostra que os efeitos desse fenômeno no passado ainda são aparentes na biodiversidade local, destacando a importância de políticas efetivas para o gerenciamento de incêndios na vegetação, especialmente da Amazônia e Cerrado. Biomas estes que, somadas as ocorrências, abrangem mais que 80% dos registros de focos disponibilizados pelo INPE no intervalo analisado (Figura 10).

6 CONCLUSÃO

A Análise Exploratória de Dados é uma atividade metódica ou ferramenta que utiliza-se de conhecimentos multidisciplinares para atingir um objetivo comum relacionado com a descoberta e obtenção de ideais e informações sobre um determinando tópico. No entanto, não existe ainda uma metodologia linear aplicável em todo o processo de uma análise. Este trabalho buscou elucidar alguns procedimentos comuns em situações recorrentes no cotidiano de um cientista de dados.

Aplicando essa motivação no contexto de incêndios florestais no Brasil, questionamentos pertinentes e recorrentes em torno da temática foram respondidos e padrões de ocorrências foram identificados ao longo do período de tempo analisado.

O maior grau de dificuldade encontrado no desenvolvimento deste trabalho está relacionado com a complexidade de analisar tamanha quantidade de observações sobre incêndios florestais. Muitas vezes, a magnitude dos dados em si pode se tornar um fator impeditivo do ponto de vista da viabilidade do desenvolvimento de um projeto. Por esta razão, foram analisados dados apenas no período de 2015 a Setembro de 2019. Abordagens adicionais como redução do escopo de variáveis dos conjuntos de dados também tiveram que ser aplicadas para não retardarem o fluxo de ideias desenvolvidas aqui.

Adicionalmente, uma possível melhoria para este trabalho seria aumentar o intervalo de tempo amostral dos dados utilizados na análise exploratória, o que agregaria mais significância estatística às informações transmitidas. Também, promover uma análise mais profunda sobre as demais variáveis encontradas em ambos os conjuntos de dados utilizados ou até mesmo buscar por formas de cruzar diretamente as informações do INPE e do *FIRMS*.

Uma outra ação pertinente seria a utilização de algoritmos de aprendizagem de máquina para revelar informações ainda mais ocultas acerca do tema abordado. A própria *NASA* busca maneiras de aplicar inteligência artificial na previsão de propagação de incêndios florestais[1]. Todavia, a quantidade de variáveis envolvidas e o curto espaço de tempo disponível para a execução de um modelo com tal finalidade e a aplicabilidade de seus resultados são fatores ainda bastante impeditivos. Porém, muito progresso está sendo feito nesta direção.

[1] https://www.nasa.gov/feature/goddard/2019/fire-forecasting-from-smart-phone-in-wilderness, acesso em 10 de Novembro de 2019

A APÊNDICE

Neste apêndice são apresentadas as descrições de todas as variáveis dos dois conjuntos de dados analisados neste trabalho. A Seção A.0.1 fornece a descrição detalhada das características do conjunto de dados do INPE. Por outro lado, a Seção A.0.2 detalha o que as variáveis do conjunto de dados da *NASA* representam.

A.0.1 Descrição de variáveis do conjunto de Dados do INPE

A Tabela 2, descreve com detalhes o que cada variável do conjunto de dados do INPE representa:

Descrição de variáveis do Banco de Queimadas do INPE	
Variável	Significado
datahora	Horário de referência da passagem do satélite segundo o fuso horário de Greenwich (GMT);
satelite	Nome do algorítmo utilizado e referência ao satélite provedor da imagem;
pais/estado/municipio	Entidades político-geográficas onde foram detectadas os focos de incêndio. Neste caso, em especial, o país se refere apenas ao Brasil;
bioma	Nome do Bioma segundo referência do Instituto Brasileiro de Geografia e Estatística (IBGE) 2004[1];
diasemchuva	Número de dias sem chuva até a detecção do foco;
precipitacao	Valor da precipitação acumulada no dia até o momento da detecção do foco;
riscofogo	Valor do Risco de Fogo previsto para o dia da detecção do foco;
latitude/longitude	coordenadas geográficas do centro do pixel que contém um possível foco de incêndio;
frp	Poder radiativo de fogo em MegaWatts (MW).

Tabela 2 – Tabela de descrição de variáveis do Banco de Queimadas do INPE

A.0.2 Descrição de variáveis do conjunto de Dados do *NASA*

A Tabela 3, descreve com detalhes o que cada variável do conjunto de dados da *NASA* representa:

[1] http://www.ibge.gov.br/home/presidencia/noticias/21052004biomashtml.shtm, acesso em 20 de Agosto de 2019

Descrição de variáveis do produto *MODIS/Aqua+Terra* do FIRMS	
Variável	Significado
latitude/longitude	Centro de 1 km de pixel de incêndio, mas não necessariamente o local real do incêndio, pois um ou mais incêndios podem ser detectados dentro do pixel de 1 km;
brightness	Canal 21/22 de temperatura de brilho do pixel de fogo medido em Kelvin (K);
scan/track	O algoritmo produz pixels de disparo de 1 km, mas os pixels MODIS aumentam em direção à borda da digitalização. o *scan* e *track* refletem o tamanho real do pixel;
acq_date/acq_time	Data e hora de aquisição do dado (UTC);
satellite	Satélite de aquisição (A = Aqua and T = Terra);
confidence	Este valor é baseado em uma coleção de quantidades intermediárias de algoritmos usadas no processo de detecção. Destina-se a ajudar os usuários a avaliar a qualidade de pontos de acesso / pixels individuais de incêndio. As estimativas de confiança variam entre 0 e 100% e são atribuídas a uma das três classes de incêndio (incêndio de baixa confiança, fogo de confiança nominal ou fogo de alta confiança);
version	A versão identifica a coleção (por exemplo, coleção 6 do *MODIS*) e a fonte do processamento de dados: quase em tempo real (sufixo *NRT* adicionado à coleção) ou processamento padrão (apenas coleção);
bright_t31	Temperatura de brilho do canal 31 do pixel de fogo medido em Kelvin;
frp	Descreve a potência radiativa de fogo integrada em pixels em MW (MegaWatts);
type	tipo de incêndio inferido (0 = vegetação, 1 = vulcão ativo, 2 = outras fontes estáticas terrestres, 3 = oceano);
daynight	detecção de fogo durante o dia ou noite.

Tabela 3 – Tabela de descrição de variáveis do produto *MODIS/Aqua+Terra* do *FIRMS*

REFERÊNCIAS

ALVES, K. M. A. da S.; NÓBREGA, R. S. Uso de dados climáticos para análise espacial de risco de incêndio florestal. *Mercator-Revista de Geografia da UFC*, Universidade Federal do Ceará, v. 10, n. 22, p. 209–219, 2011. 20, 21

ARAUJO, L. M. A. de et al. Análise dos focos de calor em áreas florestais ao longo do arco do desflorestamento. 2007. 21

DASU, T.; JOHNSON, T. *Exploratory data mining and data cleaning*. [S.l.]: John Wiley & Sons, 2003. v. 479. 14

DHAR, V. Data science and prediction. NYU Working Paper, 2012. 13

FARIA, C. C. *Bioma*. [S.l.], 2019. InfoEscola. Acesso em: 29 de Nov. de 2019. Disponível em: <https://www.infoescola.com/geografia/bioma>. 19

GAITHER, C. J. et al. Wildland fire risk and social vulnerability in the southeastern united states: An exploratory spatial data analysis approach. *Forest Policy and Economics*, Elsevier, v. 13, n. 1, p. 24–36, 2011. 20, 21

GIGLIO, L.; SCHROEDER, W.; JUSTICE, C. O. The collection 6 modis active fire detection algorithm and fire products. *Remote Sensing of Environment*, Elsevier, v. 178, p. 31–41, 2016. 28

GIL, A. C. Métodos e técnicas de pesquisa social. 8 reimpr. *São Paulo: Atlas*, v. 201, 2007. 22

GROLEMUND, G.; WICKHAM, H. Dates and times made easy with lubridate. *Journal of Statistical Software*, v. 40, n. 3, p. 1–25, 2011. Disponível em: <http://www.jstatsoft-.org/v40/i03/>. 25

HOLMES, T. P.; HUGGETT, R. J.; WESTERLING, A. L. Statistical analysis of large wildfires. In: *The Economics of Forest Disturbances*. [S.l.]: Springer, 2008. p. 59–77. 20, 21

PENG, R. *Exploratory Data Analysis with R*. [S.l.]: Lulu.com, 2016. ISBN 1365060063, 9781365060069. 14, 18

RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA, 2015. Disponível em: <http://www.rstudio.com/>. 23

SILVA, E. L. d.; MENEZES, E. M. A pesquisa e suas classificações. _____. *Metodologia da pesquisa e elaboração de dissertação*, v. 3, p. 19–23, 2005. 22

SILVEIRA, J. M. et al. The responses of leaf litter ant communities to wildfires in the brazilian amazon: a multi-region assessment. *Biodiversity and conservation*, Springer, v. 22, n. 2, p. 513–529, 2013. 47

SILVEIRA, J. M. et al. A multi-taxa assessment of biodiversity change after single and recurrent wildfires in a brazilian amazon forest. *Biotropica*, Wiley Online Library, v. 48, n. 2, p. 170–180, 2016. 47

VENABLES, B. et al. Notes on r: a programming environment for data analysis and graphics. *Dept. of Statistics, University of Adelaide and University of Auckland*, 1997. 23

WHITE, B. L. A.; RIBEIRO, A. de S. Análise da precipitação e sua influência na ocorrência de incêndios florestais no parque nacional serra de itabaiana, sergipe, brasil. *Ambiente & Água-An Interdisciplinary Journal of Applied Science*, Universidade de Taubaté, v. 6, n. 1, p. 148–156, 2011. 21

WICKHAM, H. *Advanced r*. [S.l.]: Chapman and Hall/CRC, 2014. 23

WICKHAM, H. Tidy data. *The American Statistician*, v. 14, 09 2014. 14, 24

WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: <https://ggplot2.tidyverse.org>. 24

WICKHAM, H. et al. *dplyr: A Grammar of Data Manipulation*. [S.l.], 2019. R package version 0.8.3. Disponível em: <https://CRAN.R-project.org/package=dplyr>. 24

WICKHAM, H.; GROLEMUND, G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2017. ISBN 1491910399, 9781491910399. 13, 14

WICKHAM, H.; HENRY, L. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. [S.l.], 2019. R package version 0.8.3. Disponível em: <https://CRAN.R-project-.org/package=tidyr>. 24

WILKINSON, L. The grammar of graphics. In: *Handbook of Computational Statistics*. [S.l.]: Springer, 2012. p. 375–414. 24