UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
ESCOLA DE INFORMÁTICA APLICADA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

CIÊNCIA DE DADOS E APRENDIZADO ESTATÍSTICO: UM ESTUDO DE CASO DA CAUSALIDADE ENTRE FATORES SOCIAIS E O DESMATAMENTO NA AMAZÔNIA LEGAL

GABRIEL BUZAK ZAMPIERI DE AZEVEDO
LUIZ FELIPPE BARBOSA NASCIMENTO

Orientador
BRUNO FRANCISCO TEIXEIRA SIMÕES

RIO DE JANEIRO, RJ
FEVEREIRO DE 2022

CIÊNCIA DE DADOS E APRENDIZADO ESTATÍSTICO: UM ESTUDO DE CASO DA CAUSALIDADE ENTRE FATORES SOCIAIS E O DESMATAMENTO NA AMAZÔNIA LEGAL

Projeto de Graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para obtenção do título de Bacharel em Sistemas de Informação

GABRIEL BUZAK ZAMPIERI DE AZEVEDO
LUIZ FELIPPE BARBOSA NASCIMENTO

Orientador
BRUNO FRANCISCO TEIXEIRA SIMÕES

Catalogação informatizada pelo(a) autor(a)

A994

Azevedo, Gabriel Buzak Zampieri de
CIÊNCIA DE DADOS E APRENDIZADO ESTATÍSTICO: UM ESTUDO DE CASO DA CAUSALIDADE ENTRE FATORES SOCIAIS E O DESMATAMENTO NA AMAZÔNIA LEGAL / Gabriel Buzak Zampieri de Azevedo. -- Rio de Janeiro, 2022.
85 f

Orientador: Bruno Francisco Teixeira Simões.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Estado do Rio de Janeiro, Graduação em Sistemas de Informação, 2022.

1. Aprendizado Estatístico. 2. Amazônia Legal. 3. Desmatamento. 4. PRODES. 5. IPS. I. Teixeira Simões, Bruno Francisco, orient. II. Título.

N244

Nascimento, Luiz Felippe Barbosa
CIÊNCIA DE DADOS E APRENDIZADO ESTATÍSTICO: UM ESTUDO DE CASO DA CAUSALIDADE ENTRE FATORES SOCIAIS E O DESMATAMENTO NA AMAZÔNIA LEGAL / Luiz Felippe Barbosa Nascimento. -- Rio de Janeiro, 2022.
85 f

Orientador: Bruno Francisco Teixeira Simões.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Estado do Rio de Janeiro, Graduação em Sistemas de Informação, 2022.

1. Aprendizado Estatístico. 2. Amazônia Legal. 3. Desmatamento. 4. PRODES. 5. IPS. I. Teixeira Simões, Bruno Francisco, orient. II. Título.

CIÊNCIA DE DADOS E APRENDIZADO ESTATÍSTICO: UM ESTUDO DE CASO DA CAUSALIDADE ENTRE FATORES SOCIAIS E O DESMATAMENTO NA AMAZÔNIA LEGAL

Aprovado em _____ / _____ / _____

_____

PROF. BRUNO FRANCISCO TEIXEIRA SIMÕES D. SC (UNIRIO)

_____

PROF.ª LETÍCIA MARTINS RAPOSO, D. SC (UNIRIO)

_____

PROF. REINALDO VIANA ALVARES D. SC (UNIRIO)

Rio de Janeiro, _____ de _____ de _____.

_____

_____

RIO DE JANEIRO, RJ – BRASIL.

FEVEREIRO DE 2022

AGRADECIMENTOS

Ao professor Bruno Francisco Teixeira Simões pela orientação e lapidação de nosso trabalho.

Aos meus familiares que me apoiam hoje e sempre.

Aos meus amigos e colegas, sempre presentes.

À UNIRIO, há 13 anos na minha vida e especialmente aos professores de BSI.

Gabriel Buzak Zampieri de Azevedo.

À minha família, por todo apoio durante minha trajetória acadêmica.

Aos meus amigos e colegas, companheiros de jornada.

À Bruno Francisco Teixeira Simões, pela orientação e paciência durante todo este trabalho.

Aos professores da UNIRIO pelos ensinamentos, em especial para os do BSI.

Luiz Felippe Barbosa Nascimento.

“O que as vitórias têm de mau é que não são definitivas. O que as derrotas têm de bom é que também não são definitivas”. (José Saramago)

RESUMO

Este trabalho visa, através do Aprendizado Estatístico, entender a relação entre o incremento do desmatamento na Amazônia Legal e indicadores sociais dos municípios em análise. A causalidade descrita não é necessariamente o fator causal principal. Utilizando a base de dados do PRODES (Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite) para dados sobre desmatamento e do IPS Amazônia (Índice de Progresso Social) para dados socioambientais, foi possível realizar as análises nos anos de 2014 e 2018. Dessa maneira, constatou-se que há uma relação de causalidade entre os dados socioambientais utilizados e o incremento do desmatamento. Fatores como o aumento de segurança, do acesso ao conhecimento e necessidades humanas básicas e de oportunidades (moradia, liberdade individual e de escolha, acesso ao conhecimento básico e educação superior) podem contribuir para a redução do incremento do desmatamento na região da Amazônia Legal.

Palavras-chave: Aprendizado Estatístico; Amazônia Legal; Desmatamento; PRODES; IPS.

ABSTRACT

This thesis intends to understand the relation between the increasing deforestation at the Legal Amazonia and the social indicators of the municipalities under analysis through the Statistical Learning. The causality described is not necessarily the main causal factor. By utilizing PRODES (Deforestation Monitoring Project in the Legal Amazon by Satellite) database in order to obtain data about deforestation and the IPS Amazonia (Social Progress Index) for socio-environmental data, it was possible to find the years of 2014 and 2018 analysis. Therefore, it was determined that there is a causality relation between the used socio-environmental data and the increasing deforestation. Some factors like the increasing inspection, knowledge access as well as the basic human necessities and equal opportunities (habitation, individual and choice freedom, basic knowledge access and higher education) can contribute to reduce the deforestation at the Legal Amazonia region.

Keywords: Statistical Learning; Legal Amazonia; Deforestation; PRODES; IPS.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ACP - Análise de Componentes Principais

AE - Aprendizado Estatístico

ARDS - Amazon Relational Database Service

AWS - Amazon Web Services

CVS - *Comma Separated Values*

DATASUS - Departamento de Informática do Sistema Único de Saúde

DETER - Detecção de Áreas Desflorestadas em Tempo Real

IBAMA - Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis

IBGE - Instituto Brasileiro de Geografia e Estatística

IDAM - Incremento do Desmatamento por Área de Município

IDE - *Integrated Development Environment*

IDH - Índice de Desenvolvimento Humano

INPE - Instituto Nacional de Pesquisas Espaciais

IPS - Índice de Progresso Social

MCTIC - Ministério de Ciência, Tecnologia, Inovações e Comunicações

MMA - Ministério do Meio Ambiente

MSE - *Mean Squared Error*

OMS - Organização Mundial de Saúde

PIB - Produto Interno Bruto

PRODES - Projeto de Estimativa de Desflorestamento da Amazônia

RMSE - *Root Mean Squared Error*

SGBD - Sistema de Gerenciamento de Banco de Dados

SUDAM - Superintendência de Desenvolvimento da Amazônia

SUMÁRIO

1. INTRODUÇÃO

1.1 Motivação

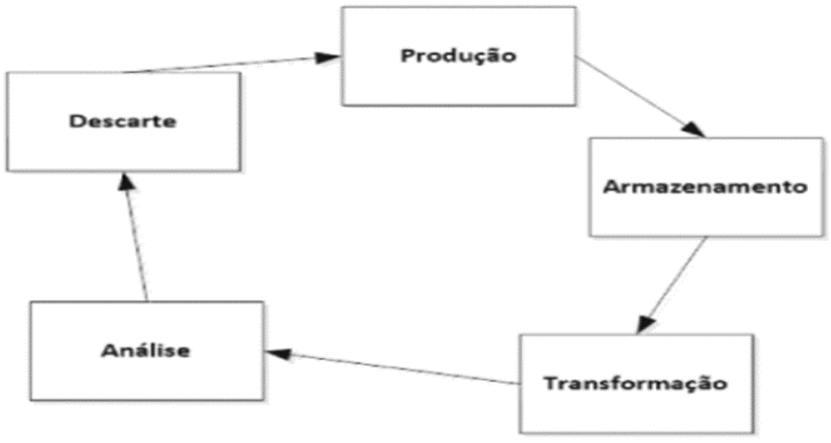
Para MOREIRA, DE CARVALHO e HORVÁTH (2018), dados são um conjunto de bits representando números, textos, imagens, entre outros. Os dados por si só são inúteis. Porém, quando adicionamos informação, os dados ganham significado e podem se tornar conhecimento.

A fim de extrair significado e conhecimento útil dos dados, utilizando o suporte de tecnologias, temos a Ciência de Dados. Ela tem o foco na criação de modelos capazes de extrair padrões dos dados, a fim de serem aplicados em problemas reais (MOREIRA; DE CARVALHO; HORVÁTH, 2018).

A Ciência de Dados é uma área relacionada à coleta, processamento, limpeza, análise e visualização de dados. Tem como finalidade tomar decisões baseadas em dados, podendo englobar várias atividades, como: análise e modelagem de dados, aprendizado de máquina, dentre outras, podendo utilizar ferramentas da Matemática e Estatística (SCHIAVON, 2019).

De forma sucinta, pode-se definir a Ciência de Dados como um conjunto de modelos, processos e tecnologias que estudam os dados durante seu ciclo de vida por completo (Figura 1), desde a produção até o seu descarte (AMARAL, 2016).

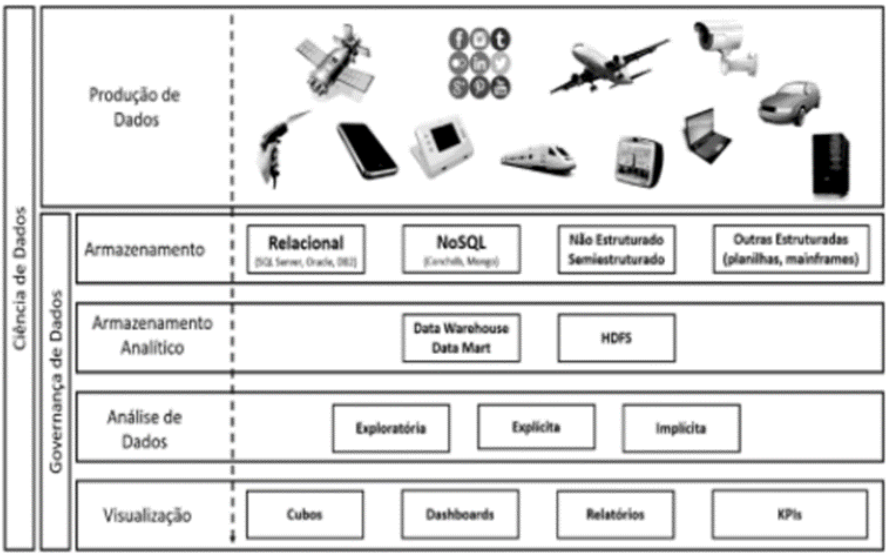
Figura 1 - Ciclo de vida de um dado.



Fonte: (AMARAL, 2016)

Por ter uma aplicação essencialmente interdisciplinar, a Ciência de Dados (Figura 2) vem ganhando cada vez mais importância em diversas áreas, como saúde, biodiversidade, bioinformática, esporte, energia, entre outras (PORTO; ZIVIANI, 2014).

Figura 2 - Visão panorâmica da Ciência de Dados.



Fonte: (AMARAL, 2016)

Aprendizado Estatístico (AE) é a habilidade de se extrair padrões de comportamento de variáveis de um certo escopo em uma escala temporal (SCHAPIRO; TURK-BROWNE, 2015). O AE tornou-se um alicerce no que se refere à construção de conhecimento de grande parte das atuais teorias de processamento de informações (BOGAERTS *et al*., 2022).

O AE pode ser determinado como um conjunto de abordagens para realizar a estimação de um desejado valor. Estas abordagens são classificadas em supervisionadas ou não supervisionadas (JAMES *et al*., 2021).

O aprendizado supervisionado envolve a construção de um modelo estatístico para prever ou estimar uma saída com base em uma ou mais entradas. Problemas de tal natureza ocorrem em campos tão diversos como negócios, medicina, políticas públicas, dentre outros. Com o aprendizado não supervisionado, também há a presença de entradas, porém nenhuma saída supervisionada, esperada. No entanto, é possível aprender relacionamentos e estrutura de tais dados, bem como a descoberta de padrões (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

1.2 Problema

De acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE), a Amazônia Legal é a área na qual atua a Superintendência de Desenvolvimento da Amazônia (SUDAM). Seus limites estão estabelecidos no Art. 2o da Lei Complementar n. 124, de 03.01.2007. Possui como finalidade promover o desenvolvimento includente e sustentável de sua área de atuação e a integração competitiva da base produtiva regional na economia nacional e internacional (IBGE).

A Amazônia é a maior floresta tropical do mundo, estando presente em nove países, sendo eles: Brasil, Bolívia, Colômbia, Equador, Peru, Venezuela, Guiana, Suriname e Guiana Francesa (SCHWERTNER, 2020).

Em território brasileiro, a Amazônia Legal ocupa uma área de 5.015.067,749 km². Essa porção equivale a 58,9% do território nacional, com sua presença em nove Estados, os quais: Amazonas, Acre, Rondônia, Roraima, Pará, Maranhão, Amapá, Tocantins e Mato Grosso; e 772 municípios, distribuídos da seguinte forma: 52 municípios de Rondônia, 22 municípios do Acre, 62 do Amazonas, 15 de Roraima, 144 do Pará, 16 do Amapá, 139 do Tocantins, 141 do Mato Grosso, bem como, por 181 municípios do Estado do Maranhão situados ao oeste do Meridiano 44º, dos quais, 21 deles, estão parcialmente integrados na Amazônia Legal (MIRANDA *et al*., 2021).

Com o intuito de realizar atividades de agricultura, pecuária, mineração, dentre outras, a Amazônia é vítima constante de desmatamento e queimadas, sendo as duas maiores questões ambientais enfrentadas pelo Brasil (GONÇALVES; DE CASTRO; HACON, 2012). Como essa prática é frequente, todos os anos, grandes quantidades de fumaça podem ser vistas nas grandes áreas desmatadas por meio de imagens de sensoriamento remoto. Essa nuvem de fumaça se espalha além do território amazônico, invadindo também outras regiões brasileiras (FIOCRUZ, 2019). O desmatamento coloca em risco características importantes do território amazônico. A Amazônia faz parte do patrimônio genético brasileiro, com variedade de fauna e flora, além de ser lugar de diversos povos indígenas (RAMOS, 2014). Mais um ponto importante é o ciclo das águas, em que a Amazônia Legal possui grande parte da reserva de água doce mundial, responsável por cerca de 20% do volume de água lançado nos oceanos. Tal ciclo é importante para a regulação do clima no território nacional e nos países vizinhos (RAMOS, 2014).

Uma fonte importante para a obtenção de dados referentes a desmatamento é o Instituto Nacional de Pesquisas Espaciais (INPE), por meio de dois projetos: o Projeto de Estimativa de

Desflorestamento da Amazônia (PRODES) e o Detecção de Áreas Desflorestadas em Tempo Real (DETER). Em ambos os projetos, são utilizados satélites para realizar cálculos da área desmatada (SOUZA *et al*., 2019).

A diminuição da área de florestas junto às práticas de queimadas causa efeitos colaterais principalmente na população local, visto que há mudanças climáticas, deterioração do solo, impacto na fauna e na flora e aumento da concentração de gases nocivos (PRATES, 2008).

Em questão de desenvolvimento local, a agropecuária é responsável por uma parcela do PIB da região amazônica, que, por sua vez, como depende de terras, tem relação com o desmatamento (PRATES, 2008). Portanto, do ponto de vista econômico, o desmatamento afeta a população local.

Para entender o progresso social na Amazônia Legal, o Índice de Progresso Social na Amazônia (IPS) mede o desempenho social e ambiental, utilizando indicadores sociais e ambientais. O IPS enfatiza os resultados e não os investimentos na área social. Nele, os indicadores econômicos não são suficientes, visto que um crescimento econômico sem desenvolvimento social resulta em degradação ambiental e problemas sociais (SANTOS *et al*., 2019). Sendo assim, é importante entender a relação dos indicadores sociais e ambientais com o desmatamento da Amazônia Legal.

O desmatamento na Amazônia Legal é uma problemática que afeta diversos aspectos ambientais, econômicos e sociais. Por isso, é necessário verificar o impacto dessa prática em relação à população local, indo além de aspectos econômicos e entender como os indicadores sociais se comportam diante a tal realidade.

1.3 Objetivos

O objetivo geral deste trabalho foi usar as ferramentas de Aprendizado Estatístico para a análise do desmatamento na Amazônia Legal e a relação deste com indicadores sociais, a fim de entender o impacto social no desmatamento.

Como objetivos específicos, temos:

- Analisar a relação entre variáveis sociais e o desmatamento.
- Extrair indicadores sociais que tenham maior relação com o desmatamento.
- Mapear possíveis cenários com as variáveis obtidas.

1.4 Organização do Texto

O presente trabalho está estruturado em capítulos e, além desta introdução, será desenvolvido da seguinte forma:

Capítulo II: Revisão bibliográfica - Este capítulo aborda estudos sobre Ciência de Dados; Aprendizado Estatístico; a relação entre sociedade e desmatamento na Amazônia Legal; explica o Monitoramento do Desmatamento da Floresta Amazônica Brasileira por Satélite e também a estrutura do Índice de Progresso Social da Amazônia.

Capítulo III: Metodologia – Discorre as ferramentas e tecnologias utilizadas no presente trabalho, além da metodologia aplicada.

Capítulo IV: Resultados e discussão - São expostos gráficos e tabelas e, posteriormente, feita uma discussão sobre os resultados obtidos.

Capítulo V: Conclusões – Agrupa as considerações finais, aponta as contribuições do trabalho e sugere possibilidades de trabalhos futuros.

2. REVISÃO BIBLIOGRÁFICA

Neste capítulo é abordado o que alguns estudos dizem sobre Ciência de Dados; Aprendizado Estatístico; a relação entre sociedade e desmatamento na Amazônia Legal; entender o que é o Monitoramento do Desmatamento da Floresta Amazônica Brasileira por Satélite e, por fim, explicar como é a estrutura principal do Índice de Progresso Social (IPS) Amazônia.

2.1 Ciência de Dados

Ciência de Dados foi uma área que alcançou maior visibilidade nas últimas décadas, sendo, em geral, associada aos desafios de análise de grandes bases de dados, juntamente à temática denominada *Big Data*. Muitos documentos externos à área acadêmica e artigos científicos foram desenvolvidos sobre este tema, procurando conceituar estes termos e avaliar os possíveis impactos na sociedade e no desenvolvimento tecnológico (KHOURY; IOANNIDIS, 2014).

Esta é uma área multidisciplinar que utiliza, como fonte, dados das mais diferentes naturezas, tais como: estruturados, não-estruturados, pequenos ou grandes volumes, constantes ou modificados em tempo real. As incumbências de Ciência de Dados incluem a extração, preparação, exploração, transformação, armazenamento e recuperação dos dados, além de manutenções computacionais, aplicações de técnicas de mineração de dados e aplicação de métodos de aprendizado de máquina, como a predição de dados (BORYCKI, 2019).

Uma extensa gama de técnicas e finalidades compõem o potencial da área de Ciência de Dados. De forma mais específica, domínios como análise de regressão, métodos de classificação, análise de agrupamento de dados, regras de associação, análise de séries temporais e sazonalidade, análise de sentimentos, padrões de comportamentos e detecção de anomalias, são ligados a esta grande área de conhecimento (SARKER, 2021).

Há uma grande diversidade de aplicações da Ciência de Dados no mundo real. Algumas áreas beneficiadas por estes avanços de conhecimentos são: Finanças e Negócios, Indústria, Medicina e Saúde Pública, Segurança de Computadores, Psicologia, Direito e Agronegócio (SARKER, 2021).

2.2 Aprendizado Estatístico

Aprendizado Estatístico refere-se a um conjunto de algoritmos e métodos pelos quais os computadores podem descobrir características e padrões importantes de conjuntos de dados de entrada que, geralmente, são muito grandes em tamanho. O termo aprendizado refere-se à tarefa de se descobrir informações e recursos a partir de dados brutos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Nos dias atuais, o AE pode ser aplicado de forma *online*, com o algoritmo recebendo dados de forma sequencial e constante. O modelo treinado em um conjunto de dados inicial é usado para prever um fluxo de observações de entrada, uma amostra por vez, e seus parâmetros são constantemente e automaticamente atualizados de acordo com o padrão do *input* observado. Essa ideia é semelhante à análise sequencial estudada em Estatística. O fato de haver a necessidade de armazenar uma grande quantidade de dados no modelo que não é em tempo real favorece a escolha pelo modelo *online* da aplicação do algoritmo (SAMBASIVAN; DAS; SAHU, 2020).

Como exemplos práticos da utilização de AE pode-se citar: predição do preço de venda ou aluguel de imóveis levando-se em conta uma série de covariáveis como taxa de criminalidade, número de quartos por residência, presença de comércio nos arredores; classificação de imagens aéreas, podendo ser aplicado no mapeamento de áreas desmatadas e com queimadas; predição do comportamento de vendas de uma locação comercial levando-se em conta sazonalidade, tendência e ciclos (SAMBASIVAN; DAS; SAHU, 2020).

2.3 Desmatamentos e Queimadas

As práticas de desmatamentos e queimadas são distintas, porém associadas. Isto porque é frequente que logo após a derrubada da vegetação ocorra a queimada. Essas atividades são comumente usadas para preparar o terreno para a realização de atividades como a agricultura e pecuária (GONÇALVES; DE CASTRO; HACON, 2012).

Essas práticas são tão frequentes na Amazônia, que existe uma região ao sul da floresta que foi nomeada como “Arco do desmatamento”. Essa região se estende do município de Paragominas – PA até Rio Branco – AC (FIOCRUZ, 2019).

Além da perda vegetal, um dos resultantes das queimadas é a emissão de poluentes. Por meio do transporte pelo fluxo de ar, os gases tóxicos, além de afetarem o local da queimada,

alcançam outras regiões (GONÇALVES; DE CASTRO; HACON, 2012). Tais poluentes atingem a população, afetando a qualidade de vida local e podendo ter reflexos na saúde, por exemplo.

Outro ponto importante é a geração de renda que acontece com o uso da terra desmatada, já que muitas famílias dependem da agropecuária como fonte de renda. Porém, em uma análise da relação entre o desmatamento da floresta amazônica e o bem-estar da população local, PRATES e BACHA (2010) concluíram que gerar mais desmatamento não necessariamente soluciona a problemática da renda, apresentando como uma possível solução a otimização do uso das áreas que já foram desmatadas a fim de aumentar a produtividade.

Já em um estudo feito por ARRAES, MARIANO e SIMONASSI (2012), em que se testou a eficácia da ação de órgãos públicos fiscalizadores e os efeitos de fatores socioeconômicos sobre as causas do desflorestamento na Amazônia Legal, foi notado que a redução do desmatamento local pode ser vista quando se tem um órgão ambiental oficial em cada município e ações como a redução de desigualdade de renda, cumprimento de leis regulatórias a fim de delimitar a expansão da fronteira agropecuária e o aumento do nível educacional.

2.4 Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite

O PRODES, ou Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite, é um projeto do INPE em que se realiza o monitoramento por satélite do desmatamento por corte raso na Amazônia Legal. Este produz as taxas anuais de desmatamento na região desde 1988, onde são usadas pelo governo brasileiro para o auxílio no planejamento de políticas públicas. A estimação das taxas anuais é feita a partir dos incrementos de desmatamento identificados em cada imagem de satélite que cobre a Amazônia Legal (INPE, 2022).

Além disso, o PRODES tem a colaboração do Ministério do Meio Ambiente (MMA) e do Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (IBAMA) e é financiado pelo Ministério de Ciência, Tecnologia, Inovações e Comunicações, o MCTIC (INPE, 2022).

A base de dados do PRODES é aberta e sua estrutura conta com as seguintes variáveis: ID do munícipio (código munícipio IBGE); ano; área (km^2); área desmatada total (km^2);

incremento em área desmatada (km^2); área de floresta (km^2); área encoberta por nuvens (km^2); área não-observada (km^2); área de não-floresta (km^2); área de hidrografia (km^2).

2.5 Aplicação de Aprendizado Estatístico sobre dados de desmatamento

Técnicas de Aprendizado Estatístico têm sido utilizadas com mais frequência com o desenvolvimento de computadores com maior capacidade de processamento. A partir da utilização de um algoritmo de árvore de decisão, foi possível realizar um processo de classificação supervisionada sobre os dados de desmatamento na Amazônia Legal divulgados anualmente pelo PRODES (MAURANO; ESCADA; RENNO, 2019).

Este estudo demonstrou que a metodologia de Aprendizado Estatístico, utilizada pela equipe, possibilitaria uma melhoria na qualidade dos dados de desmatamento obtidos pelo PRODES, com um índice de 98,5% de exatidão no mapeamento global das áreas desmatadas.

2.6 Índice de Progresso Social na Amazônia

Uma forma de entender mais sobre a população da Amazônia é por meio do IPS da Amazônia. Nele é medido o desempenho social dos municípios com o uso de indicadores sociais e ambientais. Ele é estruturado em três dimensões (Necessidades Humanas Básicas, Fundamentos para o Bem-Estar e Oportunidades) e doze componentes, nos quais cada componente possui de duas a cinco variáveis, sendo no total quarenta e três variáveis socioambientais (Figura 3). O valor da variável Índice de Progresso Social na Amazônia corresponde à média simples das três dimensões e o seu valor é uma variável entre 0 (pior) a 100 (melhor). Similarmente, as dimensões são obtidas pela média simples dos valores dos componentes que as formam. Já os componentes são resultados da Análise de Componentes Principais (ACP) entre as variáveis originais (SANTOS *et al*., 2019).

Figura 3 - Estrutura do IPS na Amazônia.

Fonte: (SANTOS *et al*., 2019).

3. METODOLOGIA

No presente capítulo, em primeiro momento, serão abordadas as principais ferramentas e tecnologias utilizadas no trabalho e, em seguida, a metodologia e a aplicação das ferramentas abordadas.

3.1 Ferramentas e Tecnologias

3.1.1 Python e bibliotecas

O Python (PYTHON, 2022) é uma linguagem de programação conhecida pela sua facilidade de aprendizado e pela sua enorme quantidade de bibliotecas disponíveis para uso público. Grande parte dessas bibliotecas podem ser encontradas no PyPI (PYPI, 2022), onde estão indexadas.

Devido às características citadas anteriormente, o Python tem se tornado popular e é comumente utilizado na área de Ciência de Dados.

Neste trabalho, o Python foi utilizado no pré-processamento dos dados (principalmente na limpeza), assim como para plotar mapas. As bibliotecas utilizadas foram: Pandas (PANDAS, 2022), Numpy (NUMPY, 2022), SqlAlchemy (SQLALCHEMY, 2022) e Plotly (PLOTLY, 2022). Sobre as bibliotecas, tem-se que:

- Pandas é uma biblioteca para análise de dados e ferramentas de manipulação. Neste trabalho, foi utilizado para manipulação dos dados juntamente com o Numpy, em que é uma biblioteca que disponibiliza funções matemáticas.
- O SqlAlchemy disponibiliza um pacote de ferramentas SQL e, dessa maneira, possibilitou a conexão com o banco de dados SQL.
- Plotly foi aplicado para a criação de mapas interativos.

3.1.2 PostgreSQL

Com origem em 1986, o PostgreSQL (POSTGRESQL, 2022) é um SGBD (Sistema de Gerenciamento de Banco de Dados) relacional de código aberto conhecido por sua confiabilidade, integridade de dados, pelo conjunto de recursos robustos e capacidade para executar em todos os sistemas operacionais. Este foi utilizado no presente trabalho por ser um

SGBD relacional com fácil uso e por ser gratuito, possuindo a plataforma *pgAdmin* como meio de se usar as funcionalidades da linguagem SQL no PostgreSQL.

Para facilitar o processo de conexão e acesso ao banco de dados, foi utilizado o serviço de nuvem da Amazon (AMAZON, 2022), o Amazon Web Services (AWS). O mesmo possibilita a hospedagem do banco de dados PostgreSQL na nuvem por meio do Amazon Relational Database Service (ARDS), sendo gratuito até certo nível de uso.

3.1.3 Google Colaboratory

O Google Colaboratory (GOOGLE, 2022), ou Google Colab, é um ambiente integrado de desenvolvimento (IDE) baseado no Jupyter notebook (JUPYTER, 2022). Ele funciona totalmente na nuvem e não precisa de configuração. Nele é possível escrever, executar, salvar e compartilhar códigos em Python e fazer anotações, sendo ideal para Ciência de Dados e áreas correlatas. O Google Colab permite programar em tempo real de forma colaborativa, é gratuito e disponibilizado via navegador.

O uso do Google Colab neste trabalho foi feito principalmente na fase inicial de limpeza de dados e em outro momento na plotagem de mapas.

3.1.4 R software e técnicas estatísticas

O R (THE R FOUNDATION, 2021) é uma linguagem de programação focada em computação estatística e gráficos. Além de ser um software gratuito e de código aberto, o R é capaz de disponibilizar um ambiente para diversos tipos de análises estatísticas graças ao grande número de funcionalidades implementadas pela comunidade de pesquisadores. Sendo assim, uma característica fundamental para a utilização do mesmo neste trabalho.

Essas funcionalidades são disponibilizadas por meio de pacotes que podem ser encontrados em repositórios públicos ou privados, sendo o Comprehensive R Archive Network (CRAN, 2021) um dos repositórios públicos mais conhecidos.

Os principais pacotes utilizados neste trabalho foram: EFAutilities (ZHANG *et al*., 2020), DBI (R SPECIAL INTEREST GROUP ON DATABASES; WICKHAM; MÜLLER, 2021), FactoMineR (HUSSON *et al*., 2020), factoextra (KASSAMBARA; MUNDT, 2020), corrplot (WEI; SIMKO, 2021), EFAtools (STEINER; GRIEDER, 2021), ggcorrplot

(KASSAMBARA, 2019), quantreg (KOENKER, 2022), Rcmdr (FOX; BOUCHET-VALAT, 2022), missMDA (JOSSE; HUSSON, 2020), dplyr (WICKHAM *et al*., 2022).

No contexto do uso do R software, foram aplicadas algumas técnicas estatísticas para efetuar a análise dos dados, sendo as principais: Análise dos Componentes Principais (ACP), o Modelo Linear de Regressão Quantílica e a Métrica do Erro Quadrático Médio (MSE).

I. Análise dos Componentes Principais

A Análise dos Componentes Principais (ACP), ou do inglês *Principal Component Analysis* (PCA), é uma técnica de análise de Estatística Multivariada em que transforma um conjunto de variáveis originais, por meio da informação contida na matriz de covariâncias ou correlação, em outro conjunto de variáveis de dimensionalidade menor, chamados de componentes principais. Estes componentes são teoricamente independentes entre si, conservam o máximo de informação das variáveis originais, em termos do percentual de inércia (percentual de explicação da variabilidade total dos dados), onde cada nova dimensão obtida tem um percentual de contribuição de cada variável. O poder de explicação da variabilidade total (contido na matriz de covariâncias) decai conforme o número de componentes, de forma que, a primeira é a que possui maior poder de explicação, a segunda é a que possui o seguindo maior poder e assim sucessivamente. Para maiores informações, veja JOHNSON e WICHERN (2014).

A ACP tem a ideia de reduzir a massa de dados com a menor perda possível de informação, sendo adequada para resumir um conjunto de dados de alta dimensionalidade (KENT, 1979).

Neste estudo, a ACP foi utilizada para reduzir a dimensionalidade das variáveis, ter indicadores mais relacionados ao desmatamento e reduzir o problema de multicolinearidade (variáveis redundantes em um mesmo modelo).

II. Modelo Linear de Regressão Quantílica

A análise por modelo de regressão é utilizada para que sejam estudados os relacionamentos entre variáveis. A reta de regressão fornece uma visão não completa de um dado conjunto de distribuições, informando apenas o valor médio previsto. Por sua vez, o modelo de regressão quantílica permite uma visão mais completa do conjunto estudado, à

medida que se estima o efeito da covariável para cada percentil. O modelo é robusto, sendo aplicado para os casos da violação da hipótese de normalidade da variável resposta e de presença de *outlier*, uma vez que não sofre influência destes pressupostos nos cálculos dos estimadores dos coeficientes dos efeitos das covárias. Para maiores detalhes, veja SANTOS (2012).

A regressão quantílica estima várias retas para diferentes quantis do mesmo conjunto estudado, ao invés de estimar apenas a esperança de Y dado um valor X, como ocorre numa regressão linear (BRAGA, 2019). A regressão usual limita-se em apresentar a relação de Y com suas covariáveis usando-se médias condicionais. Já a regressão quantílica permite que sejam observadas tais relações em qualquer quantil desejado, num intervalo de 0 a 1 (RASTEIRO, 2017). De forma concisa, a regressão quantílica é dada por:

$$Q\tau(y|x) = xT\beta(\tau) = \beta 0(\tau) + x1\beta 1(\tau) + x2\beta 2(\tau) + \cdots + xk\beta k(\tau) \quad (1)$$

em que $\beta(\tau)$ é o efeito marginal das variáveis explicativas X no τ-*ésimo* quantil de Y, efeito este que pode ser variante a depender do quantil escolhido.

O Modelo de Regressão Quantílica foi aplicado neste estudo para avaliar o impacto dos indicadores obtidos via ACP no Desmatamento.

Intervalos de Confiança de 95% foram construídos para os efeitos (estimativas dos coeficientes) das covariáveis do modelo**.**

$$E(Estimativa) \pm Z \frac{DP\,(Estimativa)}{\sqrt{n}} \quad (2)$$

III. Erro Quadrático Médio e Erro Quadrático Médio da Raiz

O valor esperado do desvio ao quadrado de um estimador em relação ao valor verdadeiro do parâmetro por ele estimado é chamado de erro quadrático médio (*Mean Squared Error*, MSE), podendo ser decomposto na variância do estimador somado ao quadrado do valor da tendenciosidade (MONTGOMERY,2016). Tendência é a diferença entre o valor esperado do estimador e o verdadeiro valor do parâmetro estimado.

Para a avaliação de uma regressão quantílica, realiza-se o cálculo dos valores do MSE. O MSE também é conhecido como perda quadrática, visto que a penalidade se dá pelo quadrado do valor do erro, e não o valor do erro em si. Um modelo é considerado bom quando o valor do MSE está mais próximo do zero. A raiz do MSE (*Root Mean Squared Error*, RMSE) mede a

magnitude média dos erros e se preocupa com os desvios do valor real. Um RMSE mais alto indica que há um grande desvio do residual para o valor esperado (BARROSO *et al*., 2015).

O MSE é dado pela fórmula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i\ )^2 \quad (3)$$

em que y é o valor real, ŷ é o valor estimado e n representa o tamanho da amostra. Já o RMSE é obtido calculando-se a raiz quadrada do MSE, definido em (3).

A métrica do MSE foi utilizada neste estudo para avaliar o desempenho do modelo de regressão em cada percentil avaliado.

3.2 Etapas de Pesquisa

I. Pré-processamento dos Dados

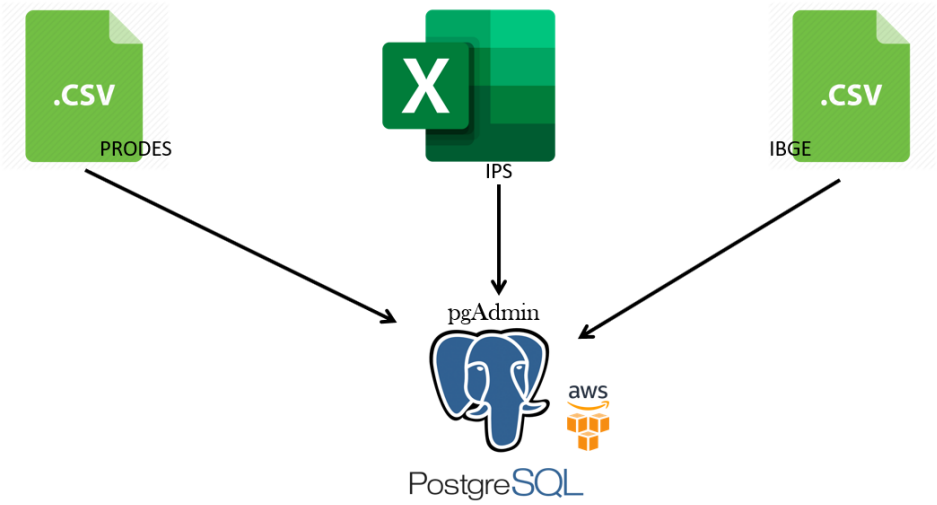
Inicialmente, o escopo deste trabalho trataria de doenças consideradas pela Organização Mundial de Saúde (OMS) como ligadas à qualidade do ar. Para isto, foram coletados dados do Departamento de Informática do Sistema Único de Saúde (DATASUS) em relação aos municípios da Amazônia Legal. Entretanto, a grande quantidade de dados faltantes inviabilizou a sequência do estudo com o tema citado, sendo posteriormente escolhido o tratado no presente trabalho.

Para que fossem utilizados dados relativos a desmatamento, foram extraídas informações do portal PRODES, disponível em formato *comma separated values* (csv). Os indicadores socioeconômicos foram obtidos do site Índice de Progresso Social da Amazônia (IPS AMAZONIA, 2022), que podem ser consumidos no formato de *Planilha do Microsoft Excel* (xls). As informações geográficas dos municípios da Amazônia Legal foram obtidas a partir do portal do IBGE, também no formato csv.

Pelo fato deste trabalho ter sido realizado por duas pessoas, decidiu-se usar o Google Colab para o desenvolvimento das análises realizadas, sendo utilizada a linguagem de programação Python. Em um primeiro momento, o Google Drive foi utilizado como um repositório online dos arquivos supracitados, devido à possibilidade de usar diretórios compartilhados.

Após análises exploratórias iniciais, foi possível detectar a viabilidade do estudo em questão. Neste momento, decidiu-se utilizar o SGBD PostgreSQL hospedado na AWS como a base de dados (Figura 4). A AWS é um serviço online e, portanto, permitiu o acesso a ambos os membros do grupo.

Figura 4 – Organização das bases utilizadas.



Fonte: Produzido pelos próprios autores.

Os três arquivos iniciais que serviram como fonte de dados foram carregados diretamente à base de dados *db_tcc* utilizando-se a ferramenta *pgAdmin*, que permite, de forma intuitiva, tal carregamento de forma facilitada (Apêndice 5).

II. Processamento dos Dados

Para realizar a conexão ao banco de dados a partir do Google Colab foi necessária a importação da biblioteca *SqlAlchemy* e o uso de seu método *create_engine*, usando como parâmetros o endereço do servidor, a porta usada, o usuário, a senha e o nome da base de dados.

A partir do uso da biblioteca pandas e seu método *read_sql_query*, foi possível realizar a leitura das tabelas carregadas na base de dados do PostgreSQL, utilizando-se simples *queries* de SQL.

A tabela referente aos dados do IBGE nomeada como *df_municipios* no Google Colab (Figura 5), e a partir do uso da biblioteca p*andas* e *unidecode*, os dados foram padronizados em letras maiúsculas e sem caracteres especiais. Estas ações foram necessárias para confirmar a

identidade dos valores no momento em que fossem cruzadas as tabelas, garantindo que os valores usados como chave estivessem iguais.

Figura 5 - Tabela *df_municipios* contendo dados do IBGE.

nm_regiao	cd_uf	nm_uf	sigla	cd_mun	nm_mun	area_tot	area_int	perc_int	lat_sede	lng_sede	sede_al	cd_mun_sus
NORTE	11	RONDONIA	RO	1100015	ALTA FLORESTA D'OESTE	7067.025	7067.025	100.0	-11.9342	-62.0041	True	110001
NORTE	11	RONDONIA	RO	1100023	ARIQUEMES	4426.571	4426.571	100.0	-9.9120	-63.0338	True	110002
NORTE	11	RONDONIA	RO	1100031	CABIXI	1314.352	1314.352	100.0	-13.4945	-60.5429	True	110003
NORTE	11	RONDONIA	RO	1100049	CACOAL	3792.892	3792.892	100.0	-11.4356	-61.4512	True	110004
NORTE	11	RONDONIA	RO	1100056	CEREJEIRAS	2783.300	2783.300	100.0	-13.1886	-60.8203	True	110005

Fonte: Produzido pelos próprios autores.

Os dados referentes ao desmatamento foram lidos como uma tabela nomeada *df_desmatamento_prodes* no Google Colab (Figura 6).

Figura 6 - Tabela *df_desmatamento_prodes* contendo dados do IBGE.

ano	cd_mun	area	desmatado	incremento	floresta	nuvem	nao_observado	nao_floresta	hidrografia	incremento_por_area
2000	1100015	7137.0	1761.1	NaN	3639.6	0.0	6.1	1708.0	22.2	NaN
2001	1100015	7137.0	1834.2	73.1	3566.5	0.0	6.1	1708.0	22.2	1.024240
2002	1100015	7137.0	1948.9	114.7	3451.8	0.0	6.1	1708.0	22.2	1.607118
2003	1100015	7137.0	2014.9	66.0	3385.8	0.0	6.1	1708.0	22.2	0.924758
2004	1100015	7137.0	2092.0	77.1	3308.7	0.0	6.1	1708.0	22.2	1.080286

Fonte: Produzido pelos próprios autores.

Para facilitar posterior cruzamento de dados, a coluna *id_municipio* foi renomeada como *cd_mun*. Em seguida, a variável resposta, que representa o desmatamento e é objeto de análise deste estudo foi criada, sendo nomeada como *incremento_por_area*, denotando o incremento do desmatamento por área de município (IDAM). Tal variável foi calculada a partir da divisão do valor da coluna *incremento* pela variável *area*, seguido pela multiplicação por 100, explicitada pela fórmula:

$$IDAM = \frac{incremento\ do\ desmatamento}{área} \times 100 \quad (4)$$

A coluna *incremento* denota o incremento de área desmatada (km²). Os incrementos são mapeados por meio de fotointerpretação por especialistas. O PRODES adota uma metodologia de mapeamento incremental, ou seja, para cada imagem são mapeados os incrementos de desmatamento que ocorreram no intervalo entre a data da imagem de um ano e a data da imagem

no ano subsequente. A coluna *área*, presente no quociente da Equação 4, representa a área, em km², dos municípios estudados. A criação desta variável se deu para que os dados fossem normalizados, dispostos em uma mesma métrica.

Na sequência, dividiu-se a tabela oriunda do PRODES em *df_desmatamento_2014* e *df_desmatamento_2018.*

Finalmente, os dados do IPS foram salvos na variável *df_ips* (Figura 7), contendo dados dos anos 2014 e 2018. Em seguida, tal variável foi dividida em *df_ips_2014* e *df_ips_2018*, de acordo com o ano observado.

Figura 7 - Visualização parcial das colunas da tabela *df_ips.*

IBGEDados	Município	Estado	Ano	Índice de Progresso Social	Necessidades Humanas Básicas	Fundamentos para o Bem-Estar	Oportunidades	Nutrição e cuidados médicos básicos	Água e saneamento	Moradia	Segurança pessoal	Acesso ao conhecimento básico	Acesso à informação e comunicação	Saúde e bem-estar	Qualidade do meio ambiente	Direitos individuais	Liberdade individual e de escolha	Tolerância e inclusão	Acesso à educação superior
1100015	Alta Floresta D'Oeste	RO	2018	60.81	58.84	68.53	55.08	81.19	13.98	78.92	61.25	67.08	74.89	58.10	74.04	49.79	78.90	69.18	22.44
1100023	Ariquemes	RO	2018	58.65	57.70	66.58	51.67	82.33	23.34	91.65	33.50	68.04	66.67	64.42	67.19	39.76	79.46	57.53	29.93
1100031	Cabixi	RO	2018	61.32	59.22	69.74	55.00	76.99	23.93	80.85	55.11	68.98	83.33	61.89	64.74	47.71	78.06	73.36	20.86
1100049	Cacoal	RO	2018	60.69	62.19	64.80	55.09	83.54	63.27	90.80	11.13	72.75	66.67	50.18	69.60	44.94	82.82	60.77	31.81
1100056	Cerejeiras	RO	2018	59.97	61.26	68.77	49.87	77.63	26.94	91.76	48.73	68.10	66.67	59.23	81.09	46.28	74.32	53.11	25.76

Fonte: Produzido pelos próprios autores.

Tratamentos de caracteres especiais foram realizados nas colunas das tabelas resultantes do IPS, visando evitar divergências entre as diferentes plataformas utilizadas. Foram, então, criadas as tabelas *df_desmatamento_ips_2014* e *df_desmatamento_ips_2018* a partir da junção das tabelas de IPS, sendo usadas como chaves as colunas *IBGEDados* (código do IBGE em que identifica cada município) e *Ano*, e as tabelas de desmatamento, tendo as chaves *cd_mun* e *ano* utilizadas. Estas novas tabelas foram inseridas no banco de dados relacional com o uso do método *to_sql* da biblioteca *pandas* (Apêndice 2).

A seguir, o software R, por meio da interface RStudio, passou a ser utilizado para a realização de análises estatísticas (Figura 8), aplicação da técnica de Análise de Componentes Principais e do Modelo de Regressão Quantílica, sendo usada a linguagem de programação *R*. A conexão ao banco de dados relacionais foi realizada a partir do método *dbConnect* da biblioteca *DBI*. As duas tabelas referentes ao IPS foram subdivididas em novas tabelas de acordo com a classificação existente dentro da origem: Dimensão, Componente e Indicador. No presente trabalho, apenas a classificação Componente foi utilizada posteriormente. Variáveis com muitas informações ausentes foram retiradas do banco de dados resultante.

Figura 8 – Fluxo das ferramentas utilizadas.

Fonte: Produzido pelos próprios autores.

Foram realizados testes de correlação entre as variáveis quantitativas para que fossem mantidas apenas as que possuíssem correlações maiores que 0,5 e significativas com o índice de desmatamento IDAM (Apêndice 3).

Em seguida, foi utilizada a biblioteca *scale*, nativa da linguagem *R*, para a normalização dos dados provenientes do IPS, resultando nas tabelas *norm_componente_df_desmatamento_ips_2014* e *norm_componente_df_desmatamento_ips_2018*. É importante frisar, que para aplicação da técnica de ACP, a variável resposta encontra-se separada desta tabela neste momento, e, portanto, não sofreu normalização.

Na sequência, a ACP foi aplicada a fim de promover a redução de dimensionalidade dos dados. Para tal, foi utilizado a biblioteca *FactoMineR*. Em seguida, com a aplicação do método *princomp* pertencente à biblioteca nativa *stats*, as variáveis foram agrupadas num conjunto de quatro variáveis, nomeadas como: *Seguranca_2014*, *Acesso_ao_conhecimento_e_tolerancia_e_Necessidades_Humanas_Basicas_2014*, *Bem_estar_2014* e *Oportunidade_2014* para o ano de 2014 e as mesmas, contendo o ano de 2018 no final, para o mesmo ano (Apêndice 4). Após esta estimação, as novas variáveis foram adicionadas às tabelas referentes. Estas foram as variáveis resultantes da ACP.

As colunas *inc_area_2014* e *inc_area_2018*, anteriormente separadas do banco de dados, foram novamente acopladas às suas respectivas tabelas. Com o auxílio da biblioteca *ggplot2* e da função *cor*, pertencente à biblioteca *stats*, foram calculadas e impressas as matrizes de correlação entre as variáveis quantitativas de cada tabela.

Ao se usar os métodos *dbCreateTable* e *dbWriteTable*, também da biblioteca *DBI*, foram adicionadas as tabelas *norm_componente_df_desmatamento_ips_2014_1* e *norm_componente_df_desmatamento_ips_2018_1* ao banco de dados relacional (Apêndice 5).

III. Modelagem dos Dados

A partir do uso das bibliotecas *EFAutilities* e *EFAtools*, foi realizada a regressão quantílica. Para o ano de 2014, a variável *inc_area_2014* foi utilizada como a variável resposta e *Bem_estar_2014*, *Seguranca_2014*, *Oportunidade_2014*, *Acesso_ao_conhecimento_e_tolerancia_e_Necessidades_Humanas_Basicas_2014* como as variáveis explicativas. Para o ano de 2018, foram usadas as mesmas variáveis, com as referências ao ano correto. O argumento *tau* foi definido em um intervalo de 0,05 a 0,95, com incremento de 0,05 para cada percentil.

Após a modelagem, foi possível observar os valores estimados dos efeitos das cováriáveis para cada percentil da distribuição de IDAM, o *p-valor* para a significância das variáveis explicativas em cada percentil avaliado, assim como o valor de MSE e, consequentemente, de RMSE (Apêndice 6), sendo os gráficos confeccionados no *software Microsoft Excel*.

IV. Mapas Interativos

Mais uma vez utilizando-se o Google Colab, a partir da biblioteca plotly foi possível construir mapas interativos, exportados e salvos no formato html, exibindo os valores de IDAM para os dois anos estudados e cada uma das quatro variáveis explicativas. Para o correto mapeamento dos municípios desejados e a construção do mapa, foram utilizados dados no formato GeoJSON, contendo informações geográficas de todos os municípios do Brasil (Apêndice 7).

4. RESULTADOS E DISCUSSÃO

Este capítulo apresenta o resultado das análises estatísticas feitas utilizando os métodos citados anteriormente no trabalho. Logo em seguida, é feita uma discussão sobre os resultados obtidos.

I. Resumos da variável-resposta

Tratando-se da variável-resposta, o IDAM, é possível perceber nas medidas-resumo (Tabelas 1 e 2) que ela possui médias de aproximadamente 0,10 e 0,12, em 2014 e 2018 respectivamente, bem como os valores de desvio-padrão iguais a 0,199949 e 0,259785. Desta forma, sinaliza que, em média, houve um incremento de aproximadamente 0,10% de área desmatada na Amazônia Legal em 2014 e aproximadamente 0,12% em 2018. Outro ponto de destaque é o incremento mais alto registrado, que em 2014 foi de aproximadamente 2,28% e em 2018 foi de 2,49%.

Tabela 1 - Medidas-resumo do indicador de incremento do desmatamento sobre área em 2014.

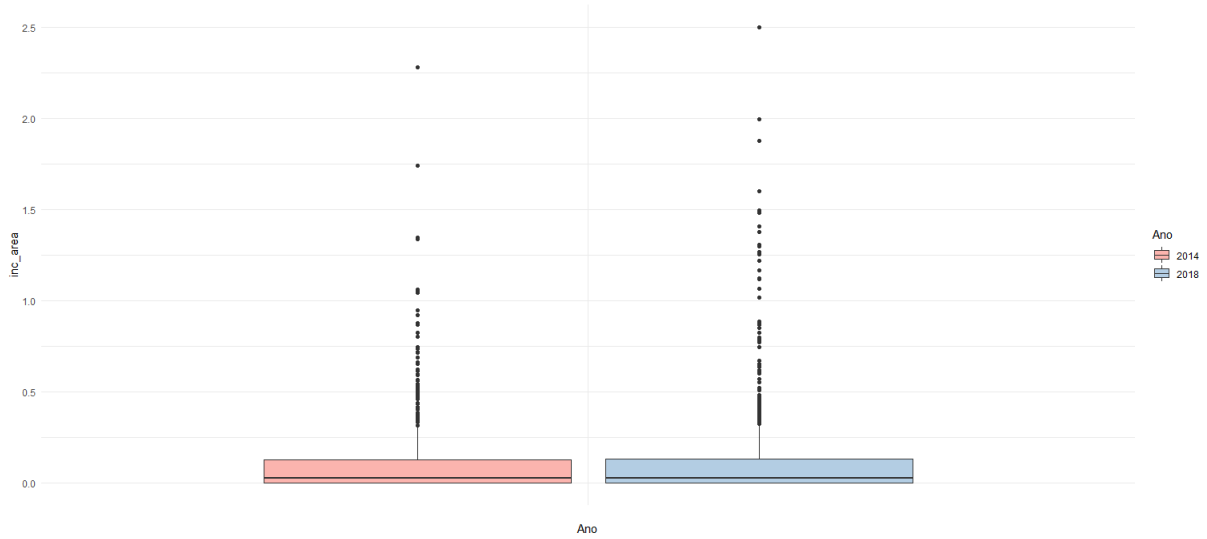
	IDAM 2014
Contagem	760
Média	0,103264
Desvio-padrão	0,199949
25%	0
50%	0
75%	0,02519
Máximo	2,28116

Tabela 2 - Medidas-resumo do indicador de incremento do desmatamento sobre área em 2018.

	IDAM 2018
Contagem	760
Média	0,122901
Desvio-padrão	0,259785
25%	0
50%	0
75%	0,025264
Máximo	2,497205

Nos boxplots (Figura 9) é notória a presença de valores discrepantes que acentuam a assimetria dos dados e demonstram que há municípios que apresentam altos valores de IDAM em comparação com os demais, já que é evidente que a maioria das ocorrências estão beirando até aproximadamente 0,025% de incremento em ambos os anos. Em 2014, podemos citar como exemplos de municípios que possuem valores discrepantes: Governador Luiz Rocha – MA, Buritis – RO e Fortuna – MA. Já em 2018, os municípios Alto Paraíso – RO, Alto Boa Vista – MT, Cujubim – RO.

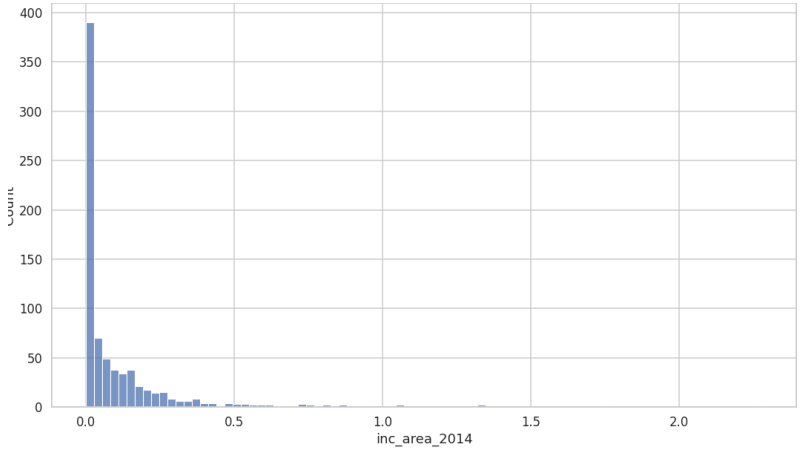
Figura 9 – Boxplot do indicador de incremento do desmatamento sobre área em 2014 e 2018.



Fonte: Produzido pelos próprios autores.

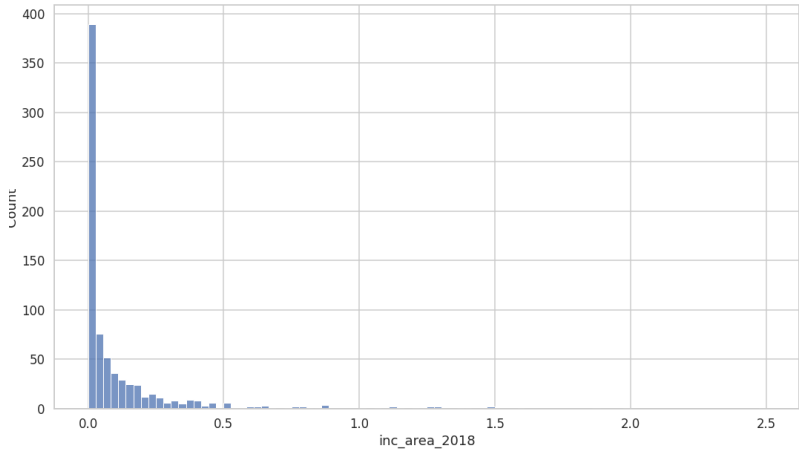
Tais assimetrias à direita também podem ser observadas nos gráficos de frequência (Figuras 10 e 11), em que se percebe que a grande maioria da frequência dos valores de IDAM se deu nos menores valores da distribuição.

Figura 10 – Histograma do indicador de incremento do desmatamento sobre área (IDAM) em 2014.



Fonte: Produzido pelos próprios autores.

Figura 11 – Histograma do indicador de incremento do desmatamento sobre área (IDAM) em 2018.



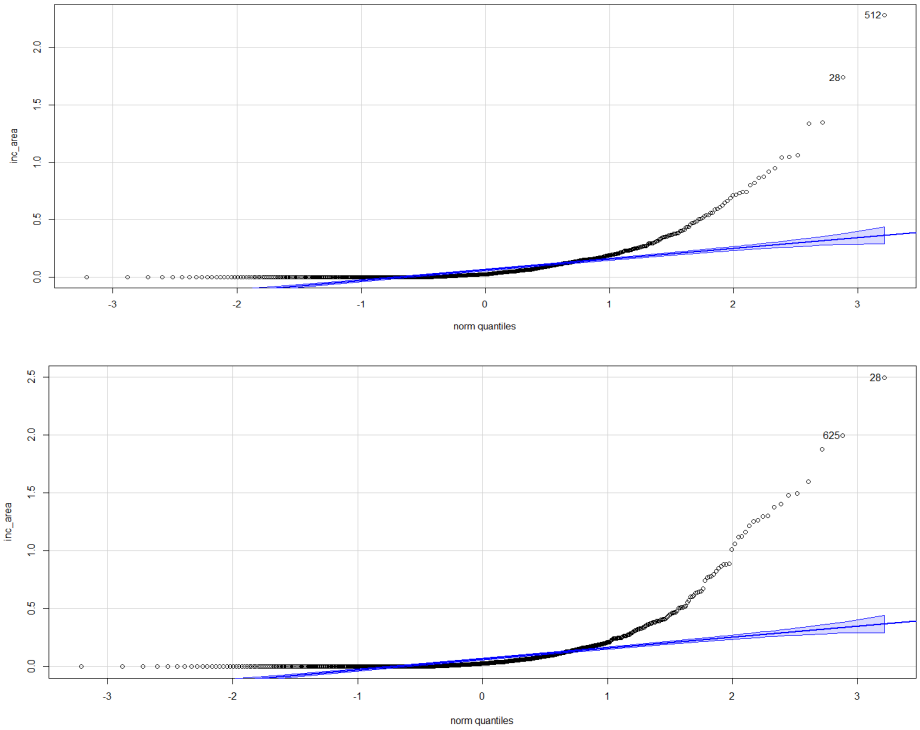
Fonte: Produzido pelos próprios autores.

Foi aplicado o teste de normalidade pelo método *Shapiro-Wilk* nos anos de 2014 e 2018 e obtido o *p-valor* < 0,001 em ambos, ou seja, descarta-se a hipótese principal que identifica a distribuição normal. Desta forma, assumimos que a distribuição não segue normalidade. A Figura 12 representa graficamente o resultado de não normalidade obtido. Caso apresentasse

distribuição normal, a sequência de pontos seguiria a reta azul dentro de seus limites estabelecidos.

O teste de normalidade foi importante para orientar o tipo de modelo de regressão a ser utilizado, e, por não seguir a distribuição normal, a regressão quantílica foi a escolhida.

Figura 12 - Gráficos de comparação de quantis em 2014 e 2018.

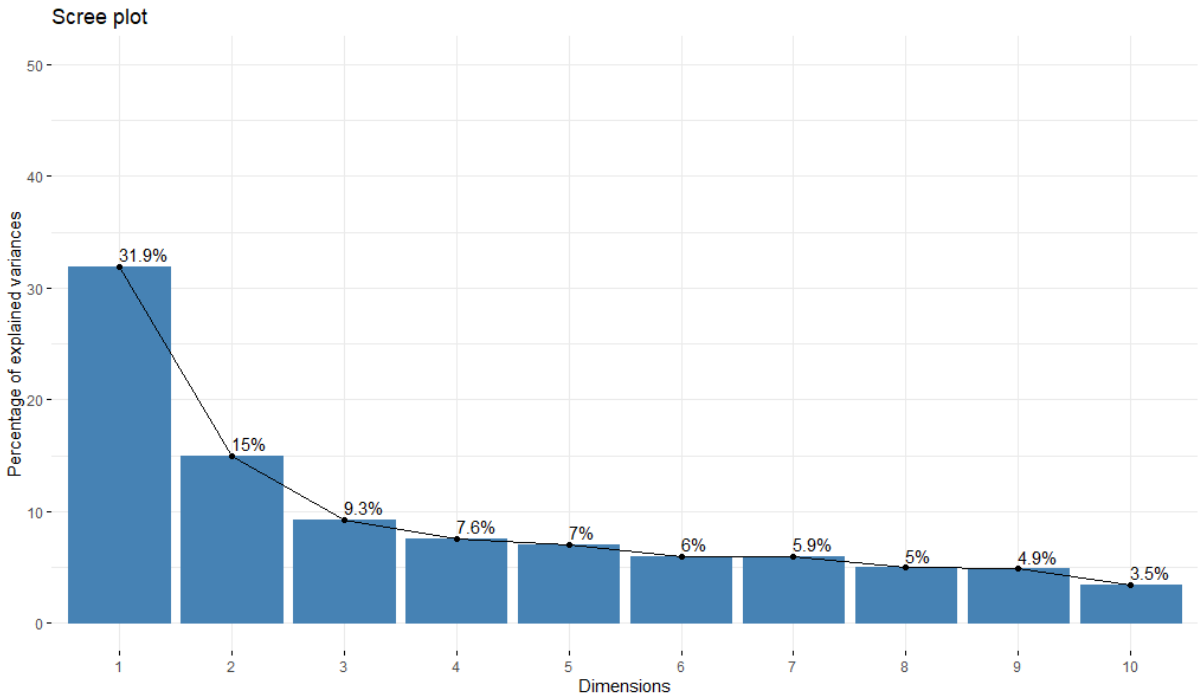


Fonte: Produzido pelos próprios autores.

II. PCA

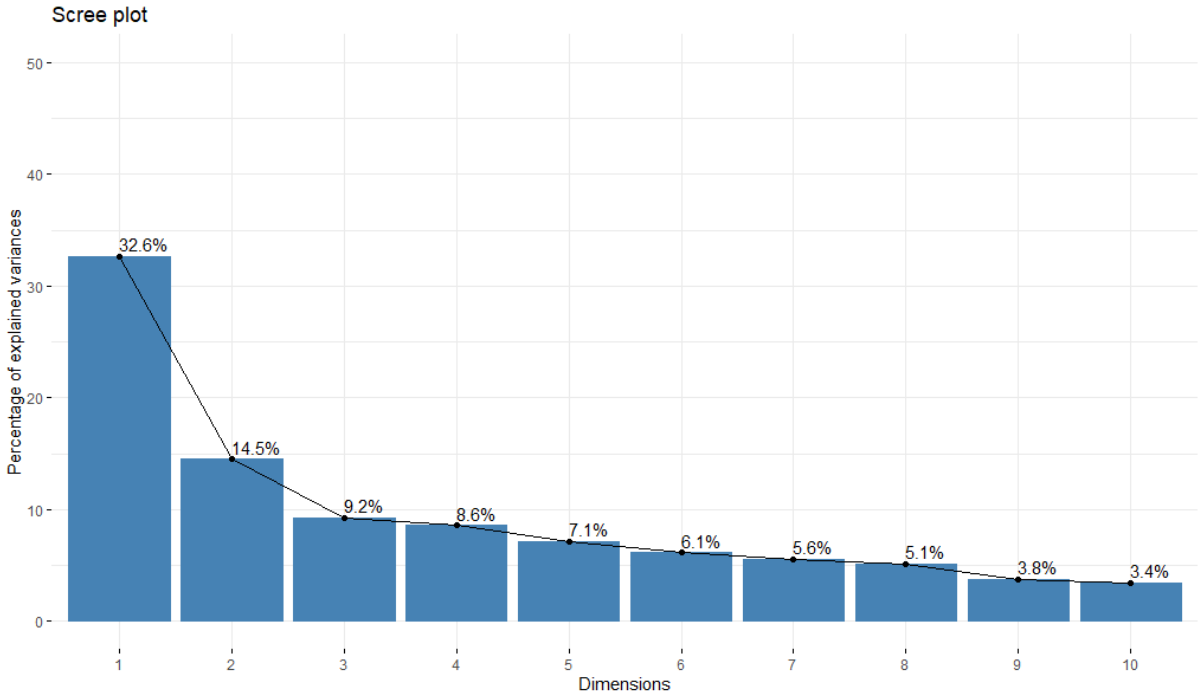
O *screen plot* foi utilizado para indicar o número de componentes principais (dimensões) a serem utilizados na análise. Tanto em 2014 quanto em 2018, foi observado que a partir da quinta dimensão em diante, não houve um aumento significativo de ganho de percentual de explicação da variabilidade total (Figuras 13 e 14). Desta maneira, foram utilizadas quatro dimensões para a análise dos componentes principais, explicando 63,8% em 2014 e 64,9% da variabilidade total em 2018.

Figura 13 - Agrupamento de Cluster nos dados de 2014.



Fonte: Produzido pelos próprios autores.

Figura 14 - Agrupamento de Cluster nos dados de 2018.



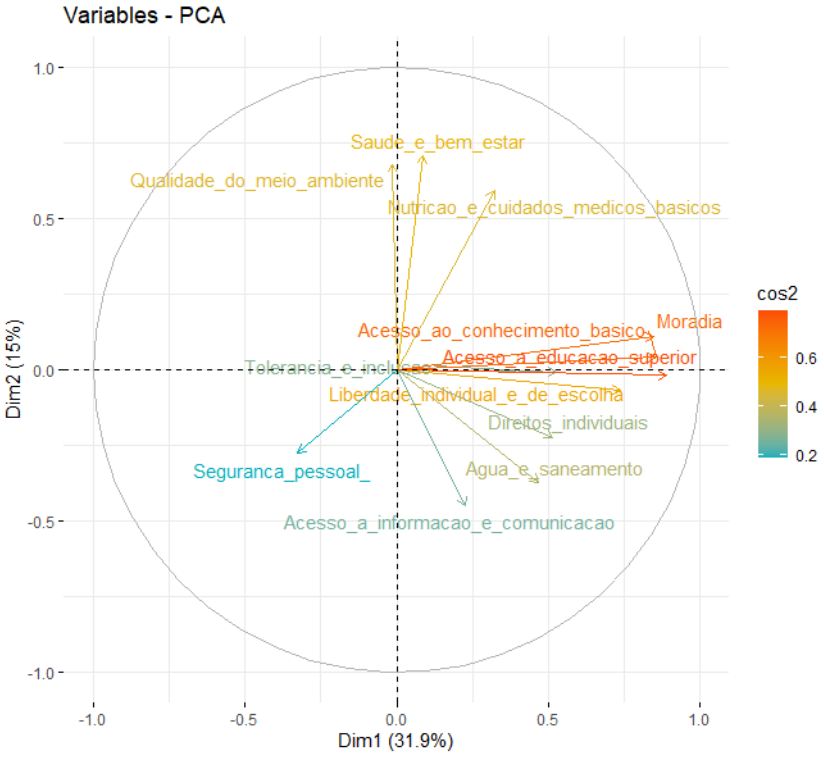
Fonte: Produzido pelos próprios autores.

O resultado referente à análise dos componentes principais esclarece a transformação do conjunto dos componentes IPS em novos conjuntos de dimensão reduzida. Por meio do valor do cosseno ao quadrado (cos^2), é possível visualizar quais variáveis compõem as dimensões criadas (Figuras 15 e 16). Tanto em 2014 quanto em 2018 temos a formação de quatro dimensões nomeadas como: Segurança, Acesso ao conhecimento e tolerância e Necessidades humanas básicas, Bem-estar e Oportunidade. Cada dimensão é formada pelas seguintes variáveis (Tabela 3):

Tabela 3 - Resumo das dimensões criadas de acordo com as variáveis.

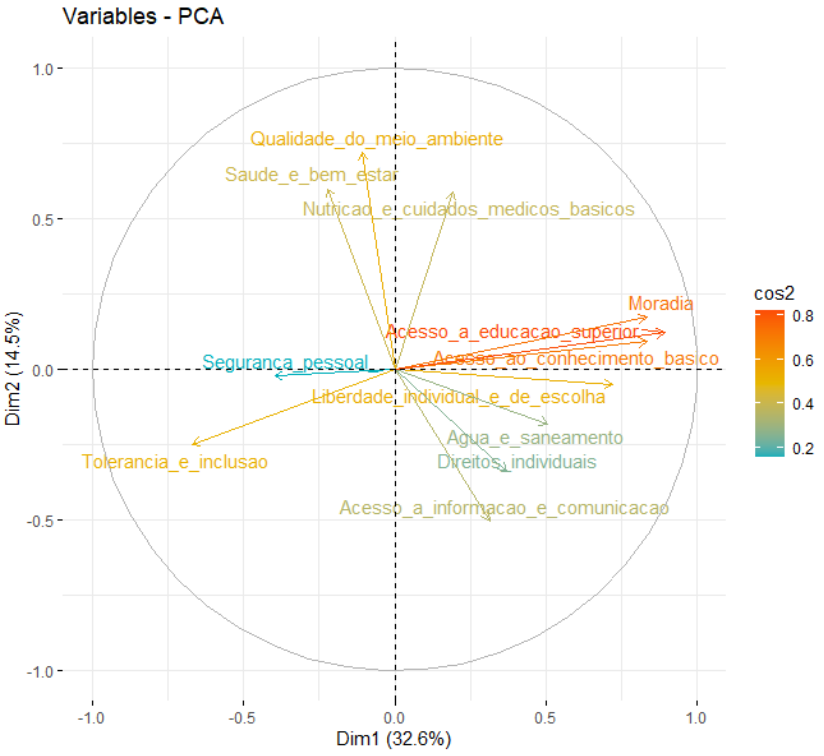
Oportunidade (Dimensão 1)	**Bem-estar (Dimensão 2)**	**Acesso ao conhecimento e tolerância e Necessidades humanas básicas (Dimensão 3)**	**Segurança (Dimensão 4)**
Moradia	Nutrição e cuidados médicos básicos	Água e saneamento	Segurança pessoal
Acesso ao conhecimento básico	Acesso à informação e comunicação (apenas em 2018)	Acesso à informação e comunicação (apenas em 2014)	
Liberdade individual e de escolha	Saúde e bem-estar	Direitos individuais	
Acesso à educação superior	Qualidade do meio ambiente	Tolerância e inclusão	

Figura 15 - Análise dos componentes principais IPS 2014.



Fonte: Produzido pelos próprios autores.

Figura 16 - Análise dos componentes principais IPS 2018.



Fonte: Produzido pelos próprios autores.

A contribuição das variáveis para cada componente principal pode ser notada por meio das cargas fatoriais, em que quanto maior o valor, maior a contribuição da variável para a dimensão (Tabelas 4 e 5). Dessa maneira, a maior carga fatorial da variável traduz em que dimensão ela tem mais impacto.

Tabela 4 - Contribuição de cada variável para os componentes principais no ano de 2014.

2014	**Dim.1**	**Dim.2**	**Dim.3**	**Dim.4**
Nutricao_e_cuidado_medicos_basicos	2,71	19,52	1,8	0,65
Agua_e_saneamento	5,74	7,71	19,74	0,97
Moradia	18,79	0,68	0,07	0,09
Seguranca_pessoal	2,8	4,26	0,89	80,2
Acesso_ao_conhecimento_basico	19,2	0,09	0,09	0,34
Acesso_a_informacao_e_comunicacao	1,34	11,21	30,01	0,67
Saude_e_bem_estar	0,18	27,79	7,61	0,33
Qualidade_do_meio_ambiente	0,005	25,65	3,52	9,69
Direitos_individuais	6,92	2,74	14,94	3,53
Liberdade_individual_e_de_escolha	14,26	0,26	2,16	1,25
Tolerancia_e_inclusao	7,26	0,002	17,75	2,21
Acesso_a_educacao_superior	20,74	0,019	1,38	0,01

Tabela 5 - Contribuição de cada variável para os componentes principais no ano de 2018.

2018	**Dim.1**	**Dim.2**	**Dim.3**	**Dim.4**
Nutricao_e_cuidado_medicos_basicos	0,95	19,92	17,47	2,34
Agua_e_saneamento	6,56	1,89	29,44	0,1
Moradia	17,75	1,73	0,36	0,04
Seguranca_pessoal	4,02	0,03	0,18	33,73
Acesso_ao_conhecimento_basico	17,84	0,49	1,9	0
Acesso_a_informacao_e_comunicacao	2,49	14,51	2,07	13,78
Saude_e_bem_estar	1,24	20,58	1,63	17,78
Qualidade_do_meio_ambiente	0,29	29,74	1,39	13,57
Direitos_individuais	3,5	6,53	25,27	16,78
Liberdade_individual_e_de_escolha	13,31	0,15	4,85	0,22
Tolerancia_e_inclusao	11,51	3,54	14,74	1,1
Acesso_a_educacao_superior	20,47	0,85	0,64	0,49

III. Matriz de correlação entre as variáveis

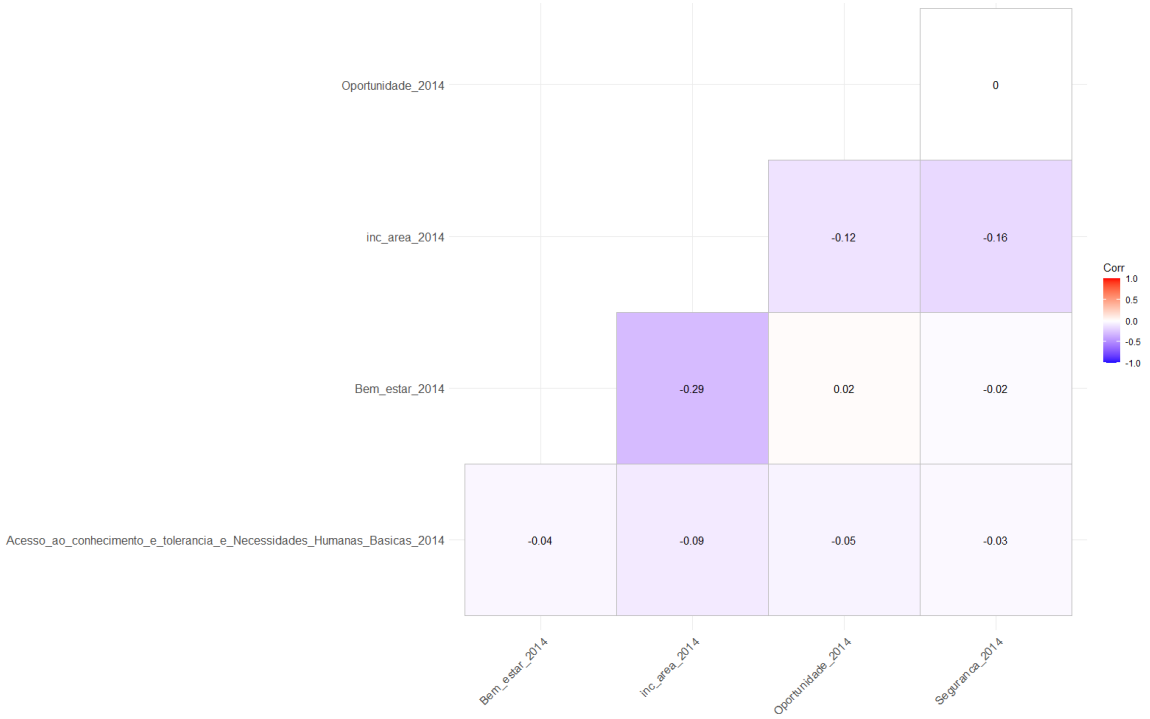
Para analisarmos a relação entre as variáveis do conjunto de dados, foram calculados os coeficientes de correlação de *Spearman*, devido à não normalidade dos dados estudados. A correlação mede a dependência entre duas variáveis. O valor calculado indica quantitativamente a relação entre as variáveis, não implicando, necessariamente, em causalidade, variando entre –1 e 1.

No ano de 2014, os testes de correlação entre a variável resposta (IDAM) e as explicativas (oportunidade, bem-estar, acesso e tolerância, segurança) foram estatisticamente significativos (p-valor: 0,0012; < 0,0001; 0,01; < 0,0001, respectivamente). Os coeficientes de correlação (-0,12; -0,29; -0,16; -0,04, respectivamente) indicam que o nível de correlação entre tais variáveis é fraco (entre 0 e 0,3), porém, significativo. A correlação da variável resposta com todas as explicativas é negativa, indicando que as variáveis são inversamente proporcionais, isto é, conforme uma aumenta, outra tende a reduzir numericamente (Figura 17).

Em relação ao ano de 2018, os testes também apresentaram significância do ponto de vista estatístico (p-valor: 0,0074; <0,0001; <0,0001; <0,0001, respectivamente) e níveis de correlação baixos (-0,06; 0,03; -0,17; -0,15, respectivamente), porém, desta vez, ocorrendo correlação positiva entre a variável alvo e a coluna indicativa de bem-estar (Figura 18).

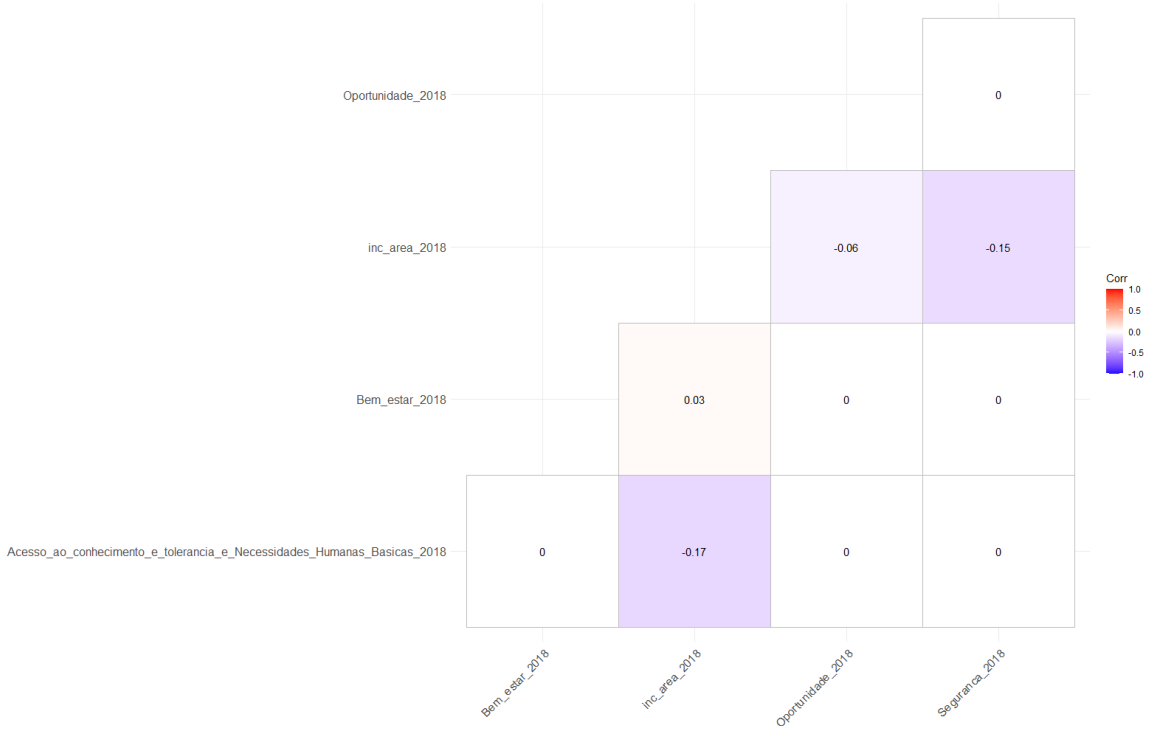
Tais dados sugerem que, pelo fato de a variável resposta, em quase todas as ocasiões, ser inversamente proporcional em relação às variáveis explicativas, quantos maiores forem os níveis dos indicadores de oportunidades, bem-estar, acesso e tolerância e segurança, menores tendem a ser os valores de IDAM.

Figura 17 – Matriz de correlação entre os indicadores no ano de 2014.



Fonte: Produzido pelos próprios autores.

Figura 18 – Matriz de correlação entre os indicadores no ano de 2018.



Fonte: Produzido pelos próprios autores.

IV. Mapas

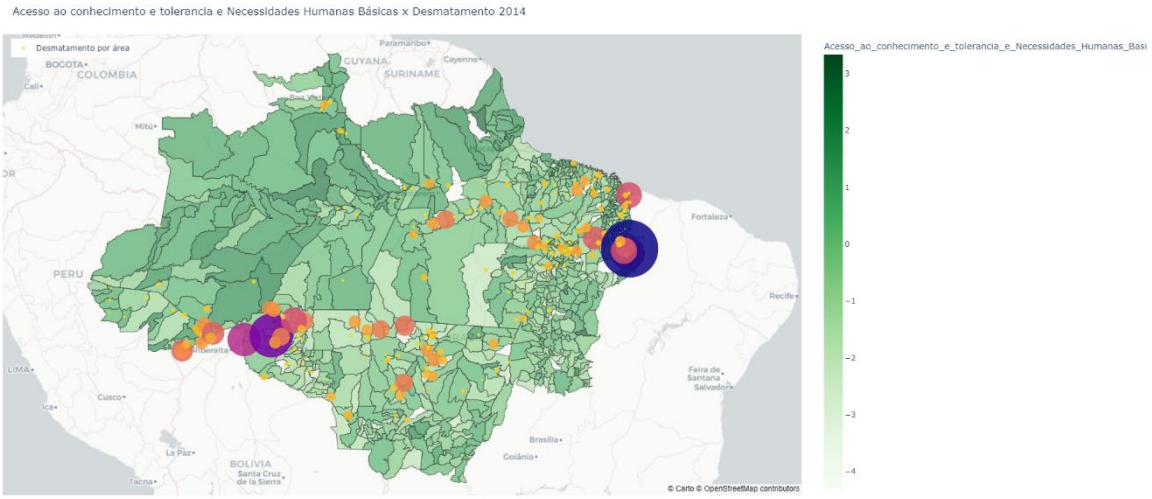
Com o auxílio da biblioteca *plotly*, foi possível construir no Google Colaboratory mapas dinâmicos contendo, tanto para 2014 quanto para 2018, informações acerca das variáveis explicativas a partir do preenchimento dos mapas de cada município, possuindo gradação de cores indicando o nível do indicador. Cada par de mapas (levando em conta ano e indicador) possui cores distintas, sendo verde para o indicador de acesso, azul para o bem-estar, cinza para oportunidade e vermelho para o indicador de segurança.

As bolhas presentes nos mapas representam, de acordo com o tamanho do raio da circunferência, o grau de IDAM em cada município.

Tais artefatos permitem que se observe, de forma visual, os dados de correlação calculados e destacados anteriormente.

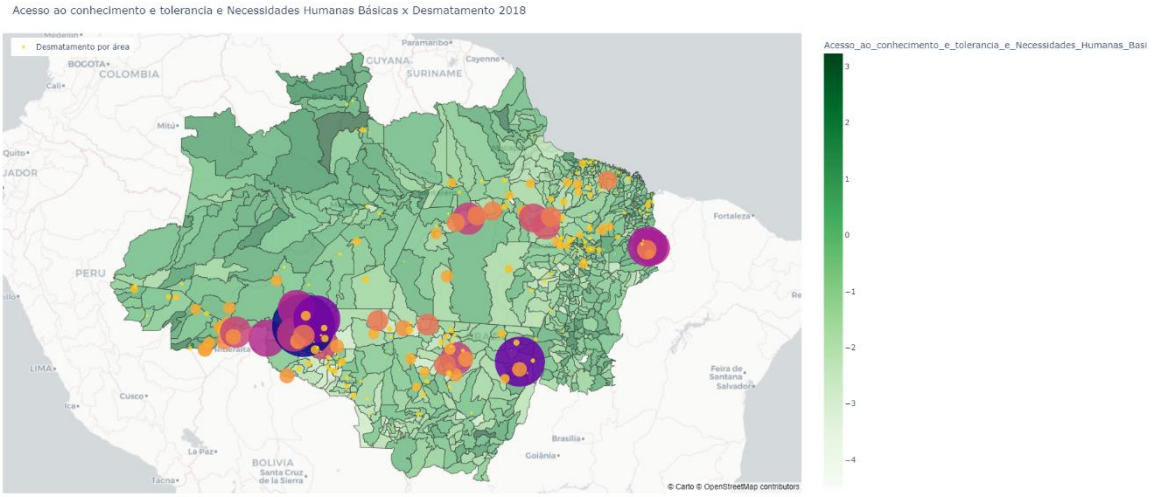
Na Figuras 19 e 20 é possível observar que nas áreas com escurecimento do tom da cor verde encontrada, que indica um maior nível nos indicadores de acesso e tolerância, possui menores valores de IDAM, e nas áreas com maior IDAM, as cores tendem a ser mais claras referentes ao indicador em questão. É possível explorar os mapas com mais detalhamentos a partir do link https://share.streamlit.io/gbuzak/mapas_tcc/main/mapas.py .

Figura 19 – Mapa contendo dados do indicador de acesso (indicado pela gradação da cor ao fundo do mapa) x IDAM (indicado pelo tamanho e cor das bolhas) em 2014.



Fonte: Produzido pelos próprios autores.

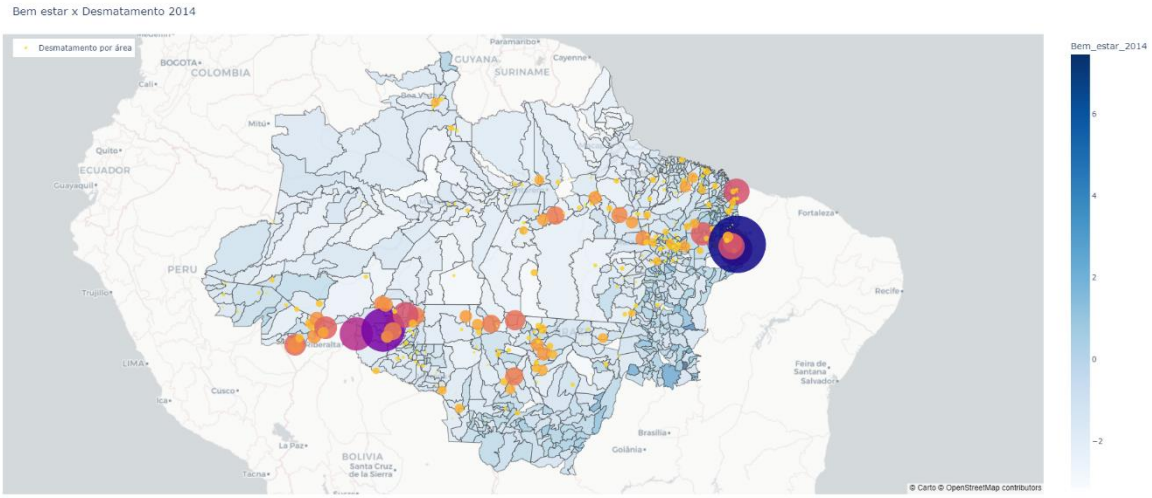
Figura 20 – Mapa contendo dados do indicador de acesso (indicado pela gradação da cor ao fundo do mapa) x IDAM (indicado pelo tamanho e cor das bolhas) em 2018.



Fonte: Produzido pelos próprios autores.

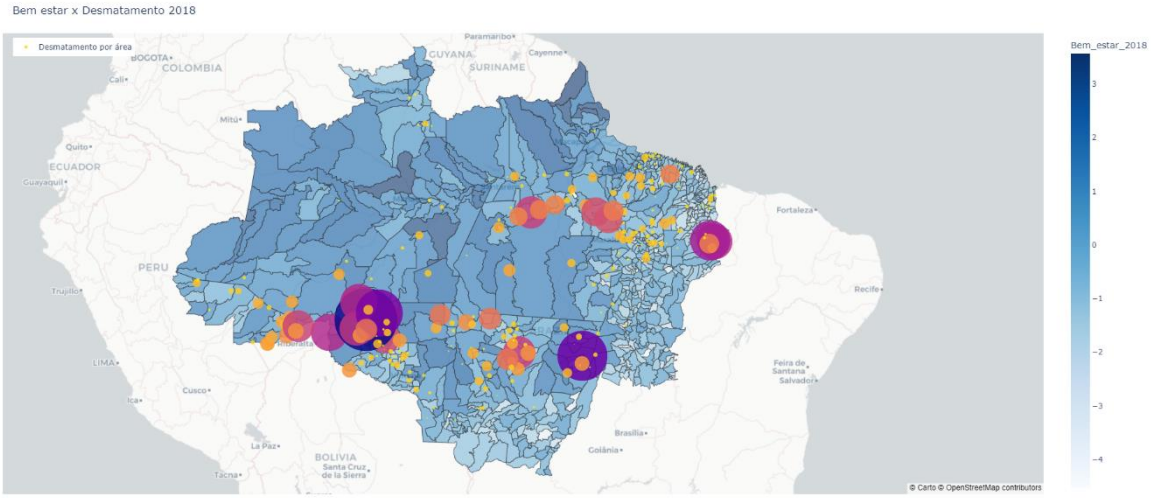
Na Figura 21 conseguimos observar que nos locais com maiores índices de IDAM no ano de 2014, o tom da cor azul indicando o bem-estar tende a estar mais claro. Na Figura 22, é possível atestar que, no ano de 2018, os índices de IDAM mais altos também se encontram em regiões mais claras do mapa do indicador de bem-estar, assim como a presença de regiões com tom de cor mais escuro em regiões com menores dados de desmatamento.

Figura 21 – Mapa contendo dados do indicador de bem-estar (indicado pela gradação da cor ao fundo do mapa) x IDAM (indicado pelo tamanho e cor das bolhas) em 2014.



Fonte: Produzido pelos próprios autores.

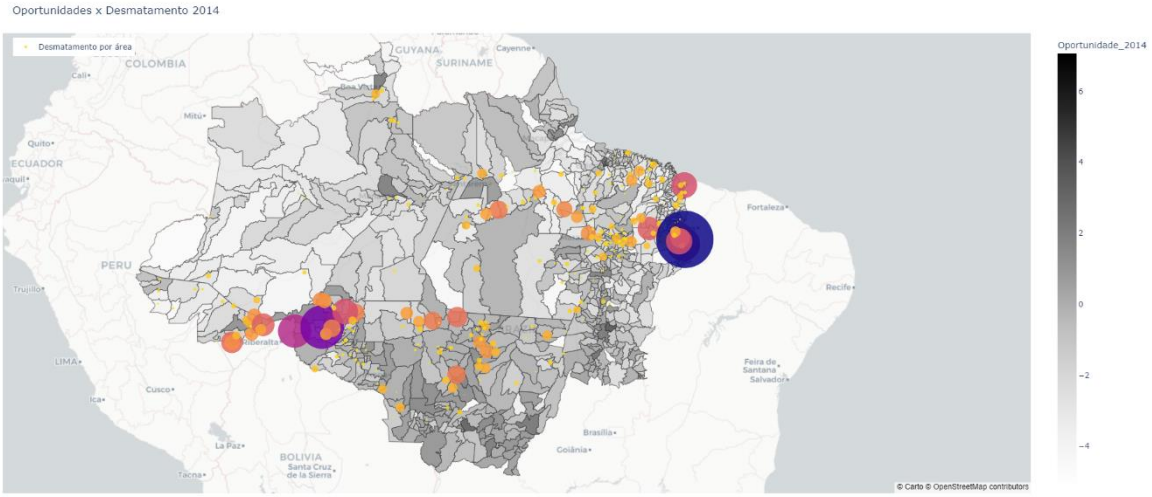
Figura 22 – Mapa contendo dados do indicador de bem-estar (indicado pela gradação da cor ao fundo do mapa) x IDAM (indicado pelo tamanho e cor das bolhas) em 2018.



Fonte: Produzido pelos próprios autores.

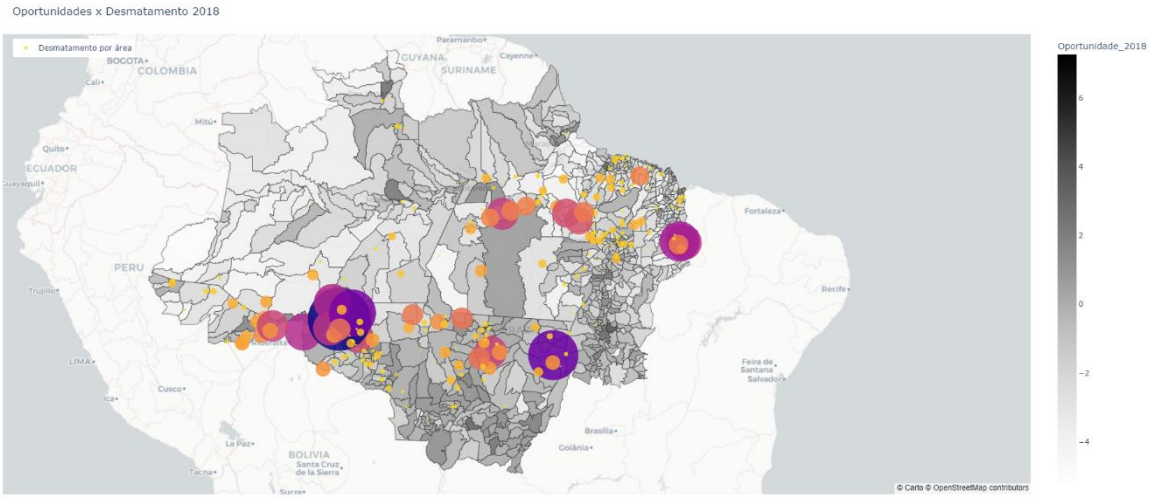
Nas Figuras 23 e 24 conseguimos observar o mesmo padrão entre os níveis de IDAM e o tom da cor acinzentada representando os indicadores das oportunidades, indicando que os mesmos são inversamente proporcionais.

Figura 23 – Mapa contendo dados do indicador de oportunidades (indicado pela gradação da cor ao fundo do mapa) x IDAM (indicado pelo tamanho e cor das bolhas) em 2014.



Fonte: Produzido pelos próprios autores.

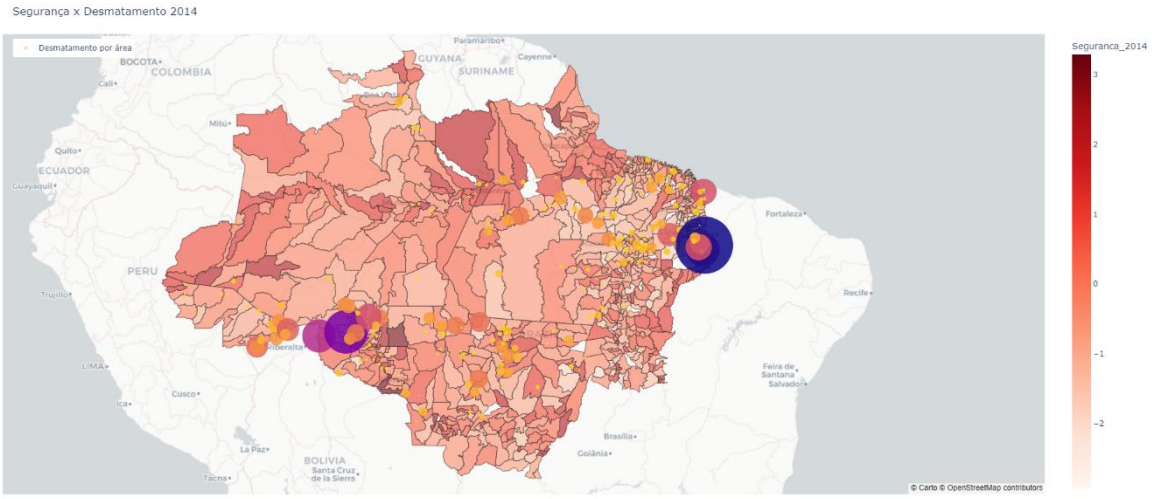
Figura 24 – Mapa contendo dados do indicador de oportunidades (indicado pela gradação da cor ao fundo do mapa) x IDAM (indicado pelo tamanho e cor das bolhas) em 2018.



Fonte: Produzido pelos próprios autores.

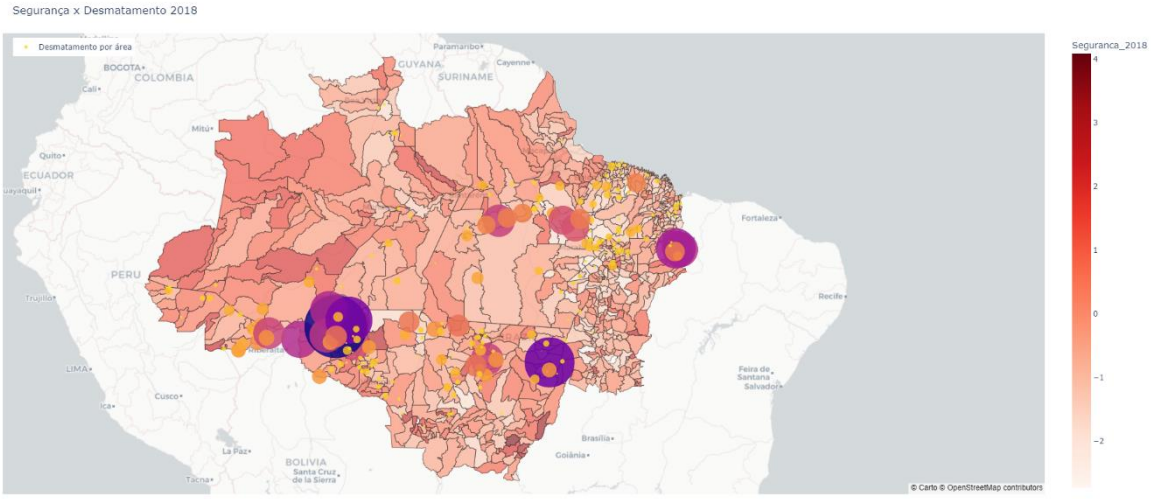
As Figuras 25 e 26 mostram, para os anos de 2014 e 2018, respectivamente, que os índices de desmatamento se apresentam inversamente proporcionais aos níveis do indicador de segurança. Tais dados corroboram visualmente o que os níveis de correlação apresentam anteriormente.

Figura 25 – Mapa contendo dados do indicador de segurança (indicado pela gradação da cor ao fundo do mapa) x IDAM (indicado pelo tamanho e cor das bolhas) em 2014.



Fonte: Produzido pelos próprios autores.

Figura 26 – Mapa contendo dados do indicador de segurança (indicado pela gradação da cor ao fundo do mapa) x IDAM (indicado pelo tamanho e cor das bolhas) em 2018.



Fonte: Produzido pelos próprios autores.

V. Modelo Linear de Regressão Quantílica

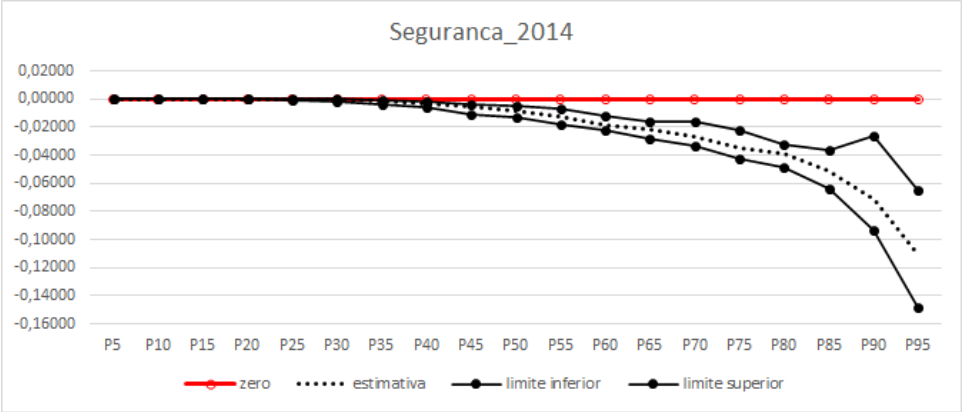
Com o propósito de estimar o comportamento do IDAM de acordo com as variáveis independentes referentes aos níveis de bem-estar, segurança, acesso e tolerância e oportunidade, foram desenvolvidos modelos de regressão quantílica. Frise-se que a causalidade descrita no título deste trabalho não é necessariamente o fator causal principal, visto que o desmatamento é um problema de natureza multifatorial.

Em 2014, podemos observar que as estimativas da variável de segurança seguem abaixo da reta de ponto zero (Figura 27) e que o *p-valor* indica significância estatística entre os percentis P50 a P95 (Tabela 6). Estão marcados em negrito os p-valores abaixo de 0,05. Da mesma forma, a variável de acesso e tolerância (Figura 28) e de oportunidade (Figura 30) seguem o mesmo padrão da anterior em relação à reta do valor zero. O indicador de acesso e tolerância apresenta *p-valor* significativo entre os percentis 50% a 95% e 55% a 90% no indicador de oportunidade. O indicador de bem-estar (Figura 29) indica estabilidade do modelo, com *p-valor* significativo entre 35% e 75% dos dados.

Tabela 6 – *p-valor* de cada percentil no modelo de regressão quantílica no ano de 2014.

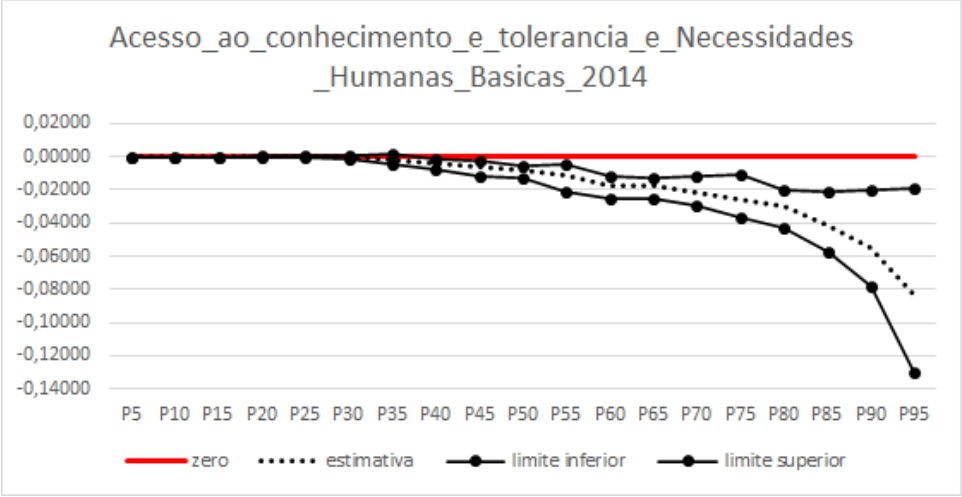
	Bem_estar_2014	**Seguranca_2014**	**Oportunidade_2014**	**Acesso_2014**
P5	1.00000	1.00000	1.00000	1.00000
P10	1.00000	1.00000	1.00000	1.00000
P15	1.00000	1.00000	1.00000	1.00000
P20	0.71959	0.98946	0.89548	0.98095
P25	0.46408	0.97403	0.81323	0.95320
P30	0.14686	0.85662	0.58000	0.97646
P35	**0.02201**	0.60027	0.48009	0.66384
P40	**0.00480**	0.29349	0.35608	0.29248
P45	**0.00015**	0.09395	0.23023	0.08530
P50	**0.00010**	**0.02050**	0.07799	**0.02861**
P55	**0.00004**	**0.00099**	**0.01968**	**0.00352**
P60	**0.00005**	**0.00003**	**0.00239**	**0.00007**
P65	**0.00014**	**0.00001**	**0.00073**	**0.00011**
P70	**0.00375**	**0.00000**	**0.00357**	**0.00004**
P75	**0.00564**	**0.00000**	**0.00241**	**0.00008**
P80	0.13118	**0.00000**	**0.00002**	**0.00007**
P85	0.39706	**0.00000**	**0.00025**	**0.00000**
P90	0.97898	**0.00000**	**0.00046**	**0.00027**
P95	0.67638	**0.00000**	0.28900	**0.01176**

Figura 27 – Modelo de regressão quantílica do indicador de segurança no ano de 2014.



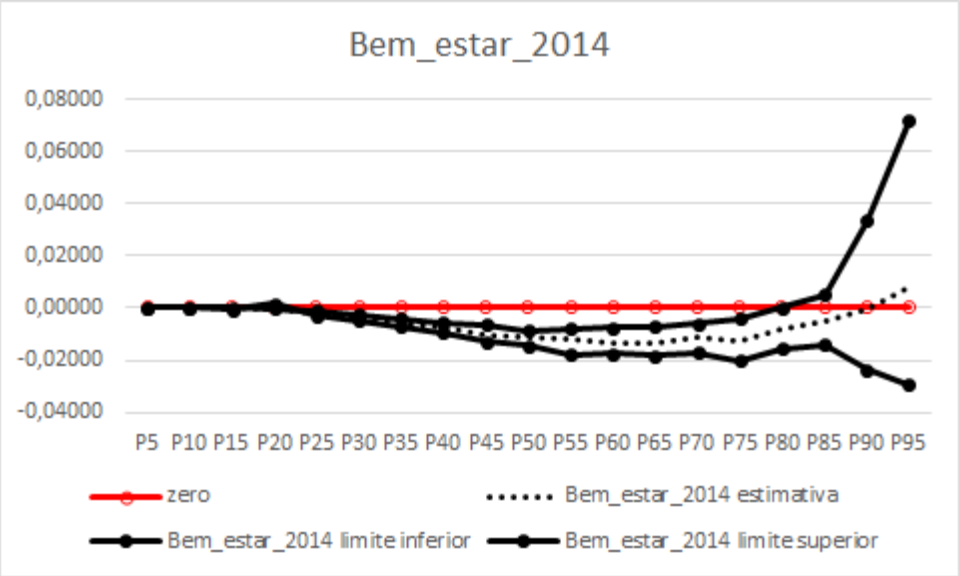
Fonte: Produção própria dos autores

Figura 28 – Modelo de regressão quantílica do indicador de acesso e tolerância no ano de 2014.



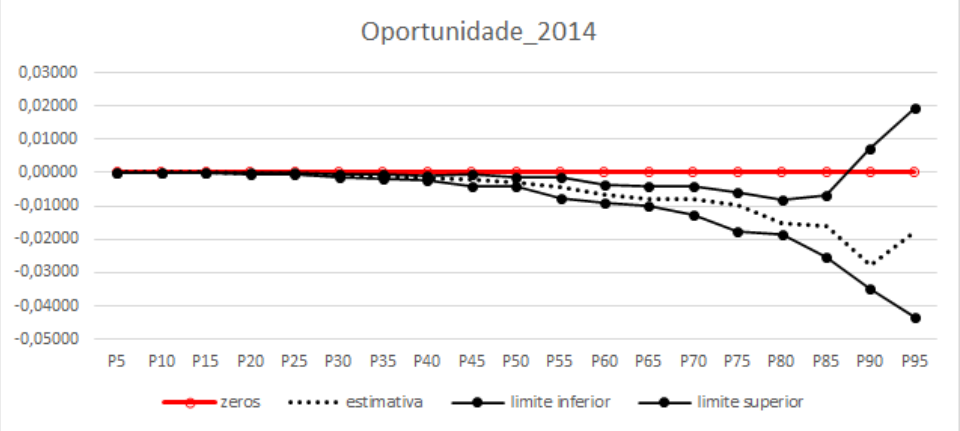
Fonte: Produção própria dos autores

Figura 29 – Modelo de regressão quantílica do indicador de bem-estar no ano de 2014.



Fonte: Produção própria dos autores.

Figura 30 – Modelo de regressão quantílica do indicador de oportunidade no ano de 2014.



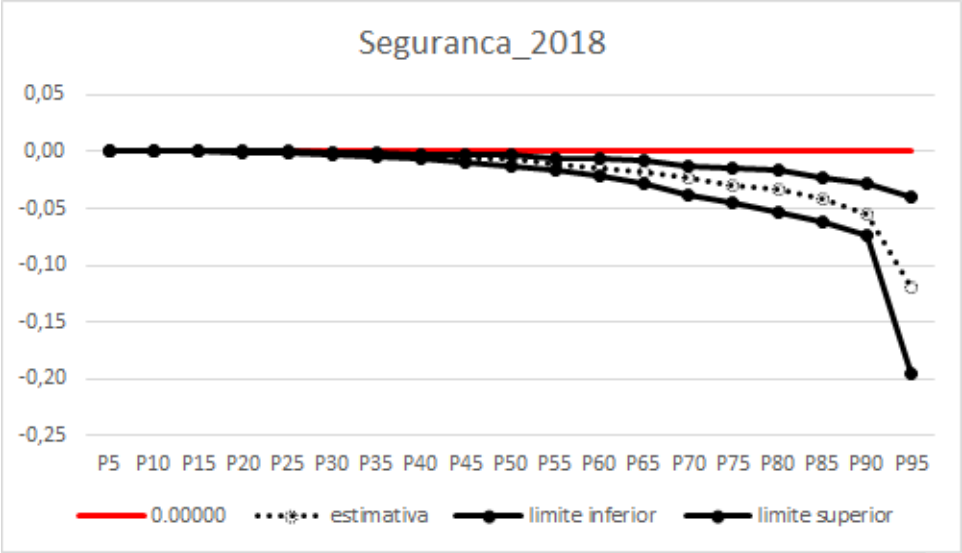
Fonte: Produção própria dos autores.

Observou-se padrão semelhante no ano de 2018. Variáveis de segurança (Figura 31), acesso e tolerância (Figura 32) e oportunidade (Figura 34) com valores abaixo do zero, apresentando, respectivamente, *p-valores* significativos nos intervalos 45% a 95%; 40% a 95% e 65% a 70% e no ponto 85% (Tabela 7). Estão marcados em negrito os p-valores abaixo de 0,05. O indicador de bem-estar (Figura 33) mais uma vez denota estabilidade do modelo, com *p-valor* significativo entre 40% e 70% dos dados.

Tabela 7 – *p-valor* de cada percentil no modelo de regressão quantílica no ano de 2018.

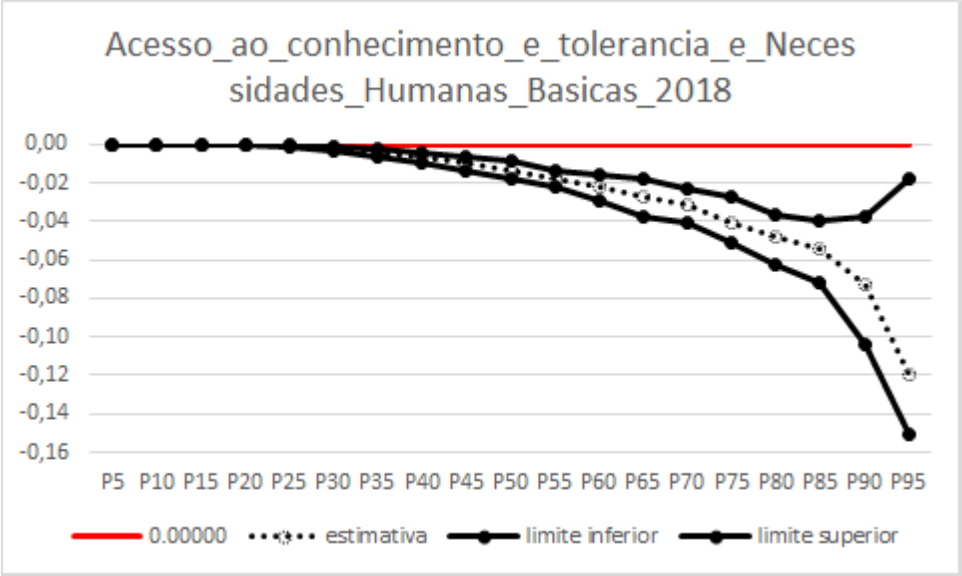
	Bem_estar_2018	**Seguranca_2018**	**Oportunidade_2018**	**Acesso_2018**
P5	1.00000	1.00000	1.00000	1.00000
P10	1.00000	1.00000	1.00000	1.00000
P15	1.00000	1.00000	1.00000	1.00000
P20	0.89589	0.99838	0.95794	0.98418
P25	0.70910	0.97685	0.90875	0.87570
P30	0.39948	0.77166	0.76139	0.60161
P35	0.10172	0.43444	0.44577	0.15675
P40	**0.04696**	0.16131	0.28086	**0.04234**
P45	**0.03797**	**0.04878**	0.21208	**0.00688**
P50	**0.02696**	**0.07174**	0.09742	**0.00026**
P55	**0.01815**	**0.00672**	0.10214	**0.00000**
P60	**0.00426**	**0.00081**	0.06264	**0.00000**
P65	**0.01023**	**0.00037**	**0.04352**	**0.00000**
P70	**0.03129**	**0.00017**	**0.02368**	**0.00000**
P75	0.06777	**0.00019**	0.53130	**0.00000**
P80	0.18780	**0.00093**	0.26023	**0.00000**
P85	0.31646	**0.00126**	**0.04486**	**0.00000**
P90	0.28476	**0.00301**	0.08594	**0.00014**
P95	0.43107	**0.00555**	0.53290	**0.00017**

Figura 31 - Modelo de regressão quantílica do indicador de segurança no ano de 2018.



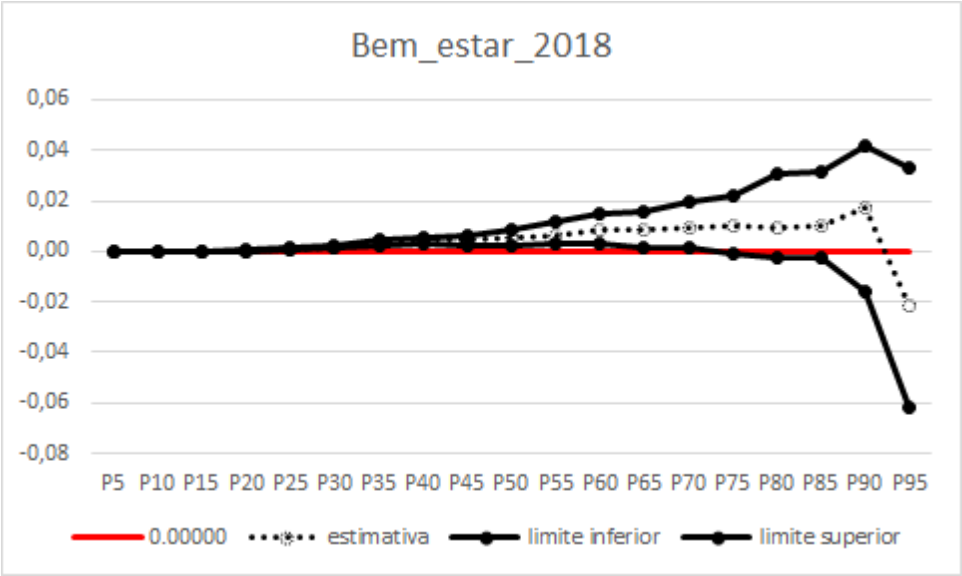
Fonte: Produção própria dos autores.

Figura 32 – Modelo de regressão quantílica do indicador de acesso e tolerância no ano de 2018.



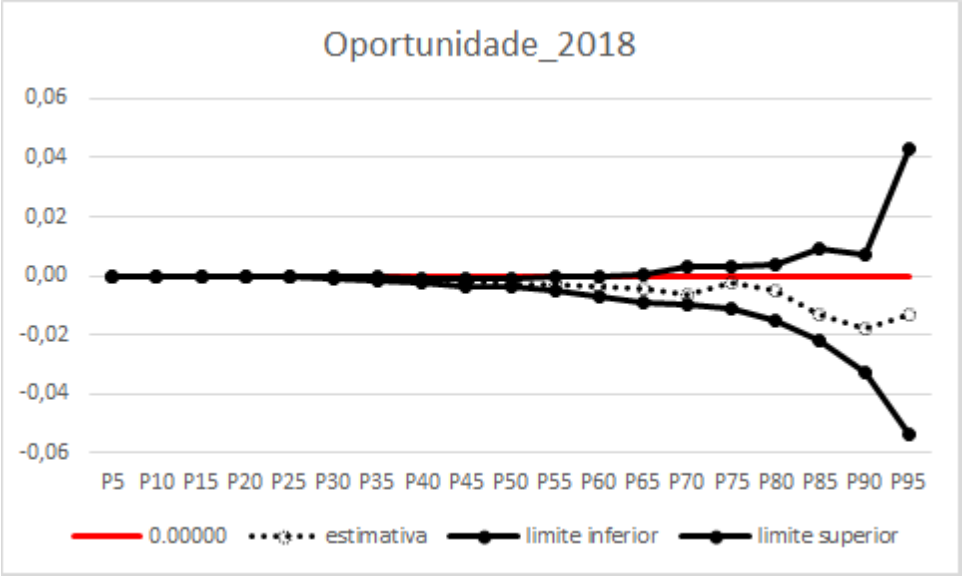
Fonte: Produção própria dos autores.

Figura 33 – Modelo de regressão quantílica do indicador de bem-estar no ano de 2018.



Fonte: Produção própria dos autores.

Figura 34 – Modelo de regressão quantílica do indicador de oportunidade no ano de 2018.



Fonte: Produção própria dos autores.

Visando a avaliação do erro dos modelos desenvolvidos, foram calculadas estimativas para o RMSE em cada percentil avaliado para cada ano do estudo. Esta métrica dá um bom indicativo do desempenho do modelo. A interpretação para tais resultados é a de que, quanto menor o valor encontrado, melhor é o ajuste ao modelo para o percentil específico. Pode ser observado (Tabela 8) que os percentis com menores valores do ajuste, e, portanto, mais adequados ao modelo, convergem com os estratos em que a significância estatística foi averiguada.

Tabela 8 - RMSA da regressão quantílica para os anos de 2014 e 2018.

	2014	**2018**
P0	0,2269	0,28798
P5	0,2269	0,28798
P10	0,2269	0,28798
P15	0,2269	0,28798
P20	0,2265	0,28784
P25	0,2258	0,28737
P30	0,2241	0,28578
P35	0,2213	0,28250
P40	0,2184	0,27932
P45	0,2143	0,27645
P50	0,2106	0,27233
P55	0,2060	0,26711
P60	0,2007	0,26266
P65	0,1987	0,25816
P70	0,1973	0,25431
P75	0,1984	0,25367
P80	0,2039	0,25750
P85	0,2174	0,27212
P90	0,2692	0,32209
P95	0,4113	0,52014
P100	1,4680	1,78873

As estimativas apresentadas nos modelos de regressão quantílica remetem às chances de que quanto maior é o valor encontrado para os indicadores, menor pode ser o IDAM, causando um incremento negativo do desfecho, no caso, uma redução nos índices de desmatamento nos municípios que estão entre os percentis com valores significativos.

Resultados semelhantes foram encontrados por ARRAES et al. (2012), ao verificar que o aumento do Índice de Desenvolvimento Humano (IDH) nos municípios da Amazônia Legal possui um efeito positivo e estatisticamente significante para a redução da degradação ambiental. Dessa maneira, pôde ser visto que melhores condições de desenvolvimento humano na Amazônia Legal tendem a aumentar a probabilidade de ocorrência de menores taxas de desmatamento.

A partir de outro ponto de vista, PRATES e BACHA (2010) analisaram a relação do desmatamento da floresta amazônica e o bem-estar da população local no sentido da geração de renda pelo uso da terra. O estudo concluiu que gerar mais desmatamento não

necessariamente é a solução para o problema de renda, pois existem terras desmatadas subutilizadas.

Em vista disso, mostra-se relevante entender a relação entre o incremento do desmatamento na Amazônia Legal e indicadores sociais dos municípios analisados.

5. CONCLUSÃO

I. Considerações Finais

O desmatamento na Amazônia Legal é um problema recorrente e que pode causar transtornos ambientais e sociais nos municípios que a permeiam. Este atual trabalho possibilitou uma análise estatística da relação entre os dados socioambientais do IPS e o incremento do desmatamento na Amazônia Legal.

Foi possível observar que existe correlação entre os fatores mais relevantes do IPS e o incremento do desmatamento nos anos de 2014 e 2018. Fatores como acesso ao conhecimento e necessidades humanas básicas, segurança e oportunidade podem contribuir para a diminuição do incremento do desmatamento. Porém não foi possível ter interpretações claras referente ao fator bem-estar, já que apresentou resultados distintos nos anos observados.

Tais resultados são similares ao de ARRAES et al. (2012), que concluiu que melhores condições de desenvolvimento humano nos municípios da Amazônia Legal tendem a aumentar a chance de ocorrer menores taxas de desmatamento.

Desta maneira, é provável que, por meio das técnicas de Aprendizado Estatístico, possa ser identificado que fatores socioambientais afetam o incremento do desmatamento na Amazônia Legal. Tais descobertas podem auxiliar na tomada de decisões por gestores públicos visando a diminuição da problemática exposta.

II. Limitações do projeto

O fato de, até o presente momento só haver dados do IPS de dois anos (2014 e 2018), diminuiu o período de amostragem dos dados e a capacidade de realizar estimativas mais fidedignas.

III. Trabalhos futuros

Estudos posteriores com a obtenção de mais dados podem contribuir em um maior entendimento da problemática. Dados do IPS para o ano de 2021 serão divulgados no segundo semestre do atual ano, o que possibilita maior visão da questão.

REFERÊNCIAS BIBLIOGRÁFICAS

AMAZON. **Serviços de nuvem – Amazon Web Services (AWS)**. Disponível em: <https://aws.amazon.com/pt/>. Acesso em: 2 fev. 2022.

BARROSO, L. M. A. *et al*. Metodologia para análise de adaptabilidade e estabilidade por meio de regressão quantílica. **Pesquisa Agropecuaria Brasileira**, v. 50, n. 4, p. 290–297, 2015.

BOGAERTS, L. *et al*. Is there such a thing as a 'good statistical learner'? **Trends in Cognitive Sciences**, v. 26, n. 1, p. 25–37, 1 jan. 2022.

BRAGA, L. F. Regressão quantílica aplicada ao potencial de mercado. 2019.

CRAN. **The Comprehensive R Archive Network**. Disponível em: <https://cran.r-project.org/>. Acesso em: 3 fev. 2022.

DE ALBUQUERQUE ARRAES, R.; ZILANIA MARIANO, F.; GOMES SIMONASSI, A. Causas do Desmatamento no Brasil e seu Ordenamento no Contexto Mundial. **Revista de Economia e Sociologia Rural [online]**, v. v. 50, p. 119–140, 2012.

FIOCRUZ. Queimadas na Amazônia e seus impactos na saúde: A incidência de doenças respiratórias no sul da Amazônia aumentou significativamente nos últimos meses. **Observatório de Clima e Saúde**, 2019.

FOX, J.; BOUCHET-VALAT, M. **RCMDR. American Statistical Association**, 2022.

GONÇALVES, K. DOS S.; DE CASTRO, H. A.; HACON, S. DE S. As queimadas na região amazônica e o adoecimento respiratório. **Ciência e Saúde Coletiva**, v. 17, n. 6, p. 1523–1532, 2012.

GOOGLE. **Olá, este é o Colaboratory**. Disponível em: <https://colab.research.google.com/>. Acesso em: 2 fev. 2022.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Overview of Supervised Learning. In: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. New York, NY: Springer New York, 2009. p. 9–41.

HOUSEH, M.; KUSHNIRUK, A. W.; BORYCKI, E. M. **Big Data, Big Challenges: A Healthcare Perspective**. Cham: Springer International Publishing, 2019.

HUSSON, F. *et al*. **FACTOMINER**. [s.l: s.n.]. Disponível em: <http://factominer.free.fr>.

IBGE. **AMAZÔNIA LEGAL**. Disponível em:< https://www.ibge.gov.br/geociencias/cartas-e-mapas/mapas-regionais/15819-amazonia-legal.html?=&t=downloads>. Acesso em: 12 fev. de 2022.

INPE. **PRODES — Coordenação-Geral de Observação da Terra**. Disponível em: <http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes>. Acesso em: 6 fev. 2022.

IPS AMAZÔNIA. **Índice de Progresso Social.** Disponível em:<http://www.ipsamazonia.org.br/>. Acesso em: 12 fev. 2022.

JAMES, G. *et al*. Statistical Learning. In: **An Introduction to Statistical Learning: with Applications in R**. New York, NY: Springer US, 2021. p. 15–57.

JOHNSON *et al*. **Applied multivariate statistical analysis**. London, UK:: Pearson, 2014.

JOSSE, J.; HUSSON, F. **MISSMDA American Statistical Association**, 2020.

JUPYTER. **Project Jupyter**, 2022. Disponível em: <https://jupyter.org/>. Acesso em: 2 fev. 2022

KASSAMBARA, A. **GGCORRPLOT**. [s.l: s.n.].

KASSAMBARA, A.; MUNDT, F. **FACTOEXTRA**. GITHUB. Disponível em: <https://github.com/kassambara/factoextra/issues>.

KENT, J. T.; BIBBY, John; MARDIA, K. V. Multivariate analysis. Amsterdam: Academic Press, 1979.

KOENKER, R. **QUANTREG**. [s.l: s.n.].

KHOURY, M. J.; IOANNIDIS, J. P. A. **Big data meets public health**. Science, v. 346, n. 6213, p. 1054–55, 2014

MAURANO, L. E. P.; ESCADA, M. I. S.; RENNO, C. D. Padrões espaciais de desmatamento e a estimativa da exatidão dos mapas do PRODES para Amazônia Legal Brasileira. **Ciência Florestal**, v. 29, n. 4, p. 1763–1775, 10 dez. 2019.

NUMPY. **NumPy**. Disponível em: <https://numpy.org/>. Acesso em: 2 fev. 2022.

PANDAS. **PANDAS - Python Data Analysis Library**. Disponível em: <https://pandas.pydata.org/>. Acesso em: 2 fev. 2022.

PLOTLY. **Plotly: The front end for ML and data science models**. Disponível em: <https://plotly.com/>. Acesso em: 2 fev. 2022.

POSTGRESQL. **PostgreSQL: The world's most advanced open source database**. Disponível em: <https://www.postgresql.org/>. Acesso em: 2 fev. 2022.

PRATES, R. C.; CAETANO BACHA, C. J. **Análise da relação entre desmatamento e bem-estar da população da Amazônia Legal**. Revista de Economia e Sociologia Rural [online], v. v. 48, p. 165–193, 2010.

PRODES. **Desmatamento nos Municípios.** PORTAL PRODES. Disponível em:<http://www.dpi.inpe.br/prodesdigital/prodesmunicipal.php>. Acesso em: 12 fev. 2022.

PYPI. **The Python Package Index**. Disponível em: <https://pypi.org/>. Acesso em: 2 fev. 2022.

PYTHON. **Welcome to Python.org**. Disponível em: <https://www.python.org/>. Acesso em: 2 fev. 2022.

R SPECIAL INTEREST GROUP ON DATABASES; WICKHAM, H.; MÜLLER, K. **DBI**. [s.l: s.n.].

RASTEIRO, L. R. **Regressão quantílica para dados censurados**. [s.l: s.n.].

SAMBASIVAN, R.; DAS, S.; SAHU, S. K. A Bayesian perspective of statistical machine learning for big data. **Computational Statistics**, v. 35, n. 3, p. 893–930, 2020.

SARKER, I. H. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. **SN Computer Science**, v. 2, n. 5, p. 377, 2021.

SANTOS, D. *et al*. Índice de Progresso Social na Amazônia Brasileira IPS Amazônia 2018. **Imazon**, 2019.

SANTOS, B. R. DOS. **Modelos de Regressão Quantílica**. [s.l: s.n.].

SCHAPIRO, A.; TURK-BROWNE, N. Statistical Learning. **Brain Mapping: An Encyclopedic Reference**, v. 3, p. 501–506, 2015.

SQLALCHEMY. **SQLAlchemy - The Database Toolkit for Python**. Disponível em: <https://www.sqlalchemy.org/>. Acesso em: 2 fev. 2022.

STEINER, M.; GRIEDER, S. **EFATOOLS**. [s.l: s.n.].

THE R FOUNDATION. **R: The R Project for Statistical Computing**. Disponível em: <https://www.r-project.org/>. Acesso em: 3 fev. 2022.

VARELLA, C. A. A. **Análise de Componentes Principais**. [s.l: s.n.]. Disponível em: <http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada%20aplicada%20as%20ciencias%20agrarias/Aulas/analise%20de%20componentes%20principais.pdf>. Acesso em: 4 fev. 2022.

WEI, T.; SIMKO, V. **CORRPLOT**. [s.l: s.n.].

WICKHAM, H. *et al*. **DPLYR**. [s.l: s.n.].

ZHANG, G. *et al*. **EFAUTILITIES**. [s.l: s.n.].

APÊNDICE 1 – SCRIPT DE CRIAÇÃO DE TABELAS E VIEWS NO SQL

```
CREATE TABLE public."IPS"
(
    index bigint,
    "IBGEDados" bigint,
    "Município" text COLLATE pg_catalog."default",
    "Estado" text COLLATE pg_catalog."default",
    "Ano" bigint,
    "Índice de Progresso Social" double precision,
    "Necessidades Humanas Básicas" double precision,
    "Fundamentos para o Bem-Estar " double precision,
    "Oportunidades" double precision,
    "Nutrição e cuidados médicos básicos" double precision,
    "Água e saneamento" double precision,
    "Moradia" double precision,
    "Segurança pessoal " double precision,
    "Acesso ao conhecimento básico" double precision,
    "Acesso à informação e comunicação" double precision,
    "Saúde e bem-estar" double precision,
    "Qualidade do meio ambiente" double precision,
    "Direitos individuais" double precision,
    "Liberdade individual e de escolha" double precision,
    "Tolerância e inclusão" double precision,
    "Acesso à educação superior" double precision,
    "Mortalidade infantil até 5 anos (Óbitos por mil nascidos vivo" double precision,
    "Mortalidade materna (Óbitos maternos por 100 mil nascidos vivo" double precision,
    "Mortalidade por desnutrição  (Óbitos por 100 mil habitantes)" double precision,
    "Mortalidade por doenças infecciosas (Óbitos por 100 mil habit" double precision,
    "Subnutrição (% da população)" double precision,
    "Abastecimento de água  (% da população) " double precision,
    "Esgotamento sanitário  (% da população) " double precision,
    "Saneamento rural (diferença entre a % da pop. Rural com acesso" double precision,
```

"Acesso à energia elétrica (% da população)" double precision,
"Coleta de lixo (% da população)" double precision,
"Moradia adequada (% da população)" double precision,
"Assassinatos de jovens (Óbitos por 100 mil habitantes de 15 a " double precision,
"Homicídios (Óbitos por 100 mil habitantes. Pontuados em uma e" double precision,
"Mortes por acidente no trânsito (Óbitos por 100 mil habitante" double precision,
"Acesso ao ensino fundamental (% de frequência líquida ao ensi" double precision,
"Acesso ao ensino médio (% de frequência líquida ao ensino m" double precision,
"Analfabetismo (% da população de 15 anos ou mais)" double precision,
"Qualidade da educação Ideb (escala de 0-10)" double precision,
"% de conexão efetuadas com sucesso. Pontuados em uma escala de" double precision,
"Conexão de voz (% de ligações realizadas com sucesso. Pontua" double precision,
"Expectativa de vida ao nascer (número de anos) " double precision,
"Mortalidade por doenças crônicas (Óbitos por 100 mil habitan" double precision,
"Mortalidade por doenças respiratórias (Óbitos por 100 mil ha" double precision,
"Obesidade (% da população) " double precision,
"Suicídio (Óbitos por 100 mil habitantes) " double precision,
"Área degradada (%)" double precision,
"Áreas Protegidas (%)" double precision,
"Desmatamento acumulado (%)" double precision,
"Desmatamento recente (% do desmatamento de 2015, 2016, 2017 em " double precision,
"Desperdício de água (%)" double precision,
"Diversidade partidária (%)" double precision,
"Mobilidade urbana (número de ônibus por mil habitantes)" double precision,
"Pessoas ameaçadas (número de ameaçados de morte por 100 mil " double precision,
"Acesso a cultura, lazer e esporte (Categórica. Pontuado em: 0 " double precision,
"Gravidez na infância e adolescência (% de mulheres de 15 a 17" double precision,
"Trabalho infantil (% da população entre 10 a 14 anos de idade" double precision,
"Vulnerabilidade familia (% de mães)" double precision,
"Desigualdade racial na educação (% da população com 15 anos" double precision,
"Violência contra a mulher (casos por 100 mil mulheres) " double precision,
"Violência contra indígena (casos por mil indígenas. Pontuado" double precision,
"Educação feminina (% da população feminina com 15 anos ou m" double precision,

```
    "Frequência ao ensino superior (% da população entre 18-24 an" double precision,
    "Pessoas com ensino superior (% da população com mais de 25 an" double precision
)
-----------------------------------------------------------------------------------------------------------
CREATE TABLE public.desmatamento_terra_brasilis
(
    ano bigint,
    id_municipio text COLLATE pg_catalog."default",
    area double precision,
    desmatado double precision,
    incremento double precision,
    floresta double precision,
    nuvem double precision,
    nao_observado double precision,
    nao_floresta double precision,
    hidrografia double precision
)
-----------------------------------------------------------------------------------------------------------
CREATE TABLE public.ibge
(
    nm_regiao character varying COLLATE pg_catalog."default",
    cd_uf character varying COLLATE pg_catalog."default",
    nm_uf character varying COLLATE pg_catalog."default",
    sigla character varying COLLATE pg_catalog."default",
    cd_mun character varying COLLATE pg_catalog."default",
    nm_mun character varying COLLATE pg_catalog."default",
    area_tot numeric,
    area_int numeric,
    perc_int numeric,
    lat_sede numeric,
    lng_sede numeric,
    sede_al boolean
)
```

```
-----------------------------------------------------------------------------------------------------------
CREATE TABLE public.norm_componente_df_desmatamento_ips_2014_1
(
    "IBGEDados" text COLLATE pg_catalog."default",
    "Municipio" text COLLATE pg_catalog."default",
    "Estado" text COLLATE pg_catalog."default",
    "Ano" bigint,
    inc_area_2014 double precision,
    "Oportunidade_2014" double precision,
    "Bem_estar_2014" double precision,
    "Acesso_ao_conhecimento_e_tolerancia_e_Necessidades_Humanas_Basi" double
precision,
    "Seguranca_2014" double precision
)
-----------------------------------------------------------------------------------------------------------
CREATE TABLE public.norm_componente_df_desmatamento_ips_2018_1
(
    "IBGEDados" text COLLATE pg_catalog."default",
    "Municipio" text COLLATE pg_catalog."default",
    "Estado" text COLLATE pg_catalog."default",
    "Ano" bigint,
    inc_area_2018 double precision,
    "Oportunidade_2018" double precision,
    "Bem_estar_2018" double precision,
    "Acesso_ao_conhecimento_e_tolerancia_e_Necessidades_Humanas_Basi" double
precision,
    "Seguranca_2018" double precision
)
-----------------------------------------------------------------------------------------------------------
CREATE OR REPLACE VIEW public.map_comp_desmatamento_ips_2014
 AS
 SELECT b.lat_sede AS lat,
    b.lng_sede AS long,
```

```
    a."IBGEDados",
    a."Municipio",
    a."Estado",
    a."Ano",
    a.inc_area_2014,
    a."Oportunidade_2014",
    a."Bem_estar_2014",
    a."Acesso_ao_conhecimento_e_tolerancia_e_Necessidades_Humanas_Basi",
    a."Seguranca_2014"
   FROM norm_componente_df_desmatamento_ips_2014_1 a
     LEFT JOIN ibge b ON a."IBGEDados" = b.cd_mun::text;
-----------------------------------------------------------------------------------------------------------
CREATE OR REPLACE VIEW public.map_comp_desmatamento_ips_2018
 AS
 SELECT b.lat_sede AS lat,
    b.lng_sede AS long,
    a."IBGEDados",
    a."Municipio",
    a."Estado",
    a."Ano",
    a.inc_area_2018,
    a."Oportunidade_2018",
    a."Bem_estar_2018",
    a."Acesso_ao_conhecimento_e_tolerancia_e_Necessidades_Humanas_Basi",
    a."Seguranca_2018"
   FROM norm_componente_df_desmatamento_ips_2018_1 a
     LEFT JOIN ibge b ON a."IBGEDados" = b.cd_mun::text;
```

APÊNDICE 2 – SCRIPT PYTHON GERAL

```
## Importação de bibliotecas necessárias
import pandas as pd
import numpy as np
from functools import reduce
from unidecode import unidecode
from sqlalchemy import create_engine
import sweetviz
from pandas_profiling import ProfileReport
from statsmodels.tsa.seasonal import seasonal_decompose
from pandas.plotting import register_matplotlib_converters
register_matplotlib_converters()
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import spearmanr
sns.set_style()
#--------------------------------------------------------------------#
#--------------------------------------------------------------------#
## Configuração do Banco de dados
postgres_str =
('postgresql://{username}:{password}@{ipaddress}:{port}/{dbname}'.format(username=PO
STGRES_USERNAME, password=POSTGRES_PASSWORD,
ipaddress=POSTGRES_ADDRESS, port=POSTGRES_PORT,
dbname=POSTGRES_DBNAME))

# Criação da conexão com o banco
db_connection = create_engine(postgres_str)
#--------------------------------------------------------------------#
#--------------------------------------------------------------------#
## IBGE
# Coleta de dados pelo banco
df_municipios = pd.read_sql_query('"SELECT * FROM ibge;"', db_connection)
```

```
# Criação da coluna auxiliar referente aos códigos da coluna IBGE do SUS (nelas há a falta do último dígito do CD_MUN), para facilitar a junção das tabelas.
df_municipios['cd_mun_sus'] = df_municipios.cd_mun.map(lambda x: str(x)[:-1])

# Colocando os nomes em maiúsculo e sem acento
df_municipios[['nm_uf', 'nm_mun']] = df_municipios[['nm_uf', 'nm_mun']].apply(lambda x: x.str.upper())
df_municipios.nm_mun = df_municipios.nm_mun.apply(lambda x: unidecode(x))
df_municipios.nm_uf = df_municipios.nm_uf.apply(lambda x: unidecode(x))
#-----------------------------------------------------------------#
#-----------------------------------------------------------------#
## PRODES
# Coleta de dados pelo banco
df_desmatamento_prodes = pd.read_sql_query('''SELECT * FROM desmatamento_terra_brasilis;''', db_connection)
db_connection.dispose()

# Renomeando as colunas
df_desmatamento_prodes = df_desmatamento_prodes.rename(columns={'id_municipio': 'cd_mun'})

# Trocando os tipos das colunas
df_desmatamento_prodes.cd_mun = df_desmatamento_prodes.cd_mun.apply(str)
df_desmatamento_prodes.ano = df_desmatamento_prodes.ano.apply(str)
df_desmatamento_prodes.area = df_desmatamento_prodes.area.str.replace(',', '.').astype(float)

### adicionando coluna calculada
df_desmatamento_prodes['incremento_por_area'] = df_desmatamento_prodes['incremento'] / df_desmatamento_prodes['area'] * 100
#-----------------------------------------------------------------#
#-----------------------------------------------------------------#
## IPS
```

```
df_ips = pd.read_sql_query('"SELECT * FROM "IPS";"', db_connection)
db_connection.dispose()


df_ips = df_ips.iloc[:,1:]
df_ips.replace(0, np.nan, inplace=True)
df_ips_2014 = df_ips[df_ips.Ano == 2014] ## IPS 2014
df_ips_2018 = df_ips[df_ips.Ano == 2018] ## IPS 2018
#-----------------------------------------------------------------#
#-----------------------------------------------------------------#
cols_desmat = ['IBGEDados', 'Município', 'Estado', 'Ano', 'Índice de Progresso Social',
               'Necessidades Humanas Básicas', 'Fundamentos para o Bem-Estar ',
               'Oportunidades', 'Nutrição e cuidados médicos básicos',
               'Água e saneamento', 'Moradia', 'Segurança pessoal ',
               'Acesso ao conhecimento básico', 'Acesso à informação e comunicação',
               'Saúde e bem-estar', 'Qualidade do meio ambiente',
               'Direitos individuais', 'Liberdade individual e de escolha',
               'Tolerância e inclusão', 'Acesso à educação superior',
               'Mortalidade infantil até 5 anos (Óbitos por mil nascidos vivo',
               'Mortalidade materna (Óbitos maternos por 100 mil nascidos vivo',
               'Mortalidade por desnutrição  (Óbitos por 100 mil habitantes)',
               'Mortalidade por doenças infecciosas (Óbitos por 100 mil habit',
               'Subnutrição (% da população)',
               'Abastecimento de água  (% da população) ',
               'Esgotamento sanitário  (% da população) ',
               'Saneamento rural (diferença entre a % da pop. Rural com acesso',
               'Acesso à energia elétrica (% da população)',
               'Coleta de lixo (% da população)', 'Moradia adequada (% da população)',
               'Assassinatos de jovens (Óbitos por 100 mil habitantes de 15 a ',
               'Homicídios (Óbitos por 100 mil habitantes. Pontuados em uma e',
               'Mortes por acidente no trânsito (Óbitos por 100 mil habitante',
               'Acesso ao ensino fundamental (% de frequência líquida ao ensi',
               'Acesso ao ensino médio (% de frequência líquida ao ensino m',
```

```
        'Analfabetismo (% da população de 15 anos ou mais)',
        'Qualidade da educação Ideb (escala de 0-10)',
        '% de conexão efetuadas com sucesso. Pontuados em uma escala de',
        'Conexão de voz (% de ligações realizadas com sucesso. Pontua',
        'Expectativa de vida ao nascer (número de anos) ',
        'Mortalidade por doenças crônicas (Óbitos por 100 mil habitan',
        'Mortalidade por doenças respiratórias (Óbitos por 100 mil ha',
        'Obesidade (% da população) ',
        'Suicídio (Óbitos por 100 mil habitantes) ', 'Área degradada (%)',
        'Áreas Protegidas (%)', 'Desmatamento acumulado (%)',
        'Desmatamento recente (% do desmatamento de 2015, 2016, 2017 em ',
        'Desperdício de água (%)', 'Diversidade partidária (%)',
        'Mobilidade urbana (número de ônibus por mil habitantes)',
        'Pessoas ameaçadas (número de ameaçados de morte por 100 mil ',
        'Acesso a cultura, lazer e esporte (Categórica. Pontuado em: 0 ',
        'Gravidez na infância e adolescência (% de mulheres de 15 a 17',
        'Trabalho infantil (% da população entre 10 a 14 anos de idade',
        'Vulnerabilidade familia (% de mães)',
        'Desigualdade racial na educação (% da população com 15 anos',
        'Violência contra a mulher (casos por 100 mil mulheres) ',
        'Violência contra indígena (casos por mil indígenas. Pontuado',
        'Educação feminina (% da população feminina com 15 anos ou m',
        'Frequência ao ensino superior (% da população entre 18-24 an',
        'Pessoas com ensino superior (% da população com mais de 25 an', 'ano',
        'cd_mun', 'area', 'desmatado', 'incremento', 'floresta', 'nuvem',
        'nao_observado', 'nao_floresta', 'hidrografia', 'incremento_por_area']

cols_desmat = [item.replace(" ", "_") for item in cols_desmat]
cols_desmat = [item.replace("-", "_") for item in cols_desmat]
cols_desmat = [item.replace("(", "_") for item in cols_desmat]
cols_desmat = [item.replace(")", "_") for item in cols_desmat]
cols_desmat = [item.replace("%", "percent") for item in cols_desmat]
cols_desmat = [unidecode(item) for item in cols_desmat]
```

```
df_desmatamento_2014 = df_desmatamento_prodes[df_desmatamento_prodes.ano == 2014]
df_desmatamento_2018 = df_desmatamento_prodes[df_desmatamento_prodes.ano == 2018]

df_ips_2014['IBGEDados'] = df_ips_2014['IBGEDados'].apply(str)
df_ips_2018['IBGEDados'] = df_ips_2018['IBGEDados'].apply(str)

df_desmatamento_ips_2014 = df_ips_2014.merge(df_desmatamento_2014, how='left',
left_on=['IBGEDados','Ano'], right_on=['cd_mun','ano']);
df_desmatamento_ips_2014 = df_desmatamento_ips_2014[df_desmatamento_ips_2014.ano
== 2014]
df_desmatamento_ips_2014.columns = cols_desmat

df_desmatamento_ips_2018 = df_ips_2018.merge(df_desmatamento_2018, how='left',
left_on=['IBGEDados','Ano'], right_on=['cd_mun','ano']);
df_desmatamento_ips_2018 = df_desmatamento_ips_2018[df_desmatamento_ips_2018.ano
== 2018]
df_desmatamento_ips_2018.columns = cols_desmat

#-----------------------------------------------------------------#
#-----------------------------------------------------------------#
## ADIÇÃO AO BANCO DE DADOS
df_desmatamento_ips_2014.to_sql('df_desmatamento_ips_2014', db_connection,
index=False, if_exists='replace')
df_desmatamento_ips_2018.to_sql('df_desmatamento_ips_2018', db_connection,
index=False, if_exists='replace')
#-----------------------------------------------------------------#
#-----------------------------------------------------------------#
## A partir deste ponto, dados foram tratados no RStudio
#-----------------------------------------------------------------#
#-----------------------------------------------------------------#
```

APÊNDICE 3 – SCRIPT RSTUDIO – LIMPEZA DAS BASES

```
## Instalação e importação dos pacotes necessários
# install.packages('RPostgreSQL')
# install.packages('devtools')
# install.packages('remotes')
# remotes::install_github('r-dbi/RPostgres')
# install.packages('RPostgres')
# install.packages('corrr')
# install.packages(c("FactoMineR", "factoextra"))
# install.packages("MVN")
# install.packages('missMDA')
# install.packages("EFAutilities")
# install.packages("EFAtools")

library(EFAutilities)
library(DBI)
library(FactoMineR)
library(factoextra)
library(corrplot)
library(factoextra)
library(EFAtools)
library(ggcorrplot)
library(quantreg)
library(missMDA)

#############################################################################

## Conexão com o banco de dados Postgres na AWS.
db <- DB_NAME
host_db <- DB_HOST
db_port <- DB_PORT
db_user <- DB_USER
```

```
db_password <- DB_PASSWORD
con <- dbConnect(RPostgres::Postgres(), dbname = db, host=host_db, port=db_port,
user=db_user, password=db_password)

## SELECT na base de dados conectada
df_desmatamento_ips_2014 <- dbGetQuery(con, 'SELECT * FROM
df_desmatamento_ips_2014')
df_desmatamento_ips_2018 <- dbGetQuery(con, 'SELECT * FROM
df_desmatamento_ips_2018')

##Separando as colunas de ano, nome dos municipios e cód IBGE para adição ao final
mun_ano_codIBGE_2014 <- df_desmatamento_ips_2014[,1:4]
mun_ano_codIBGE_2018 <- df_desmatamento_ips_2018[,1:4]

# Normalizando a variável resposta incremento por área
inc_area_2014 <- df_desmatamento_ips_2014$incremento_por_area
colnames(inc_area_2014) <- 'inc_area_2014'

inc_area_2018 <- df_desmatamento_ips_2018$incremento_por_area
colnames(inc_area_2018) <- 'inc_area_2018'

################################################################################

## Parte relativa ao desmatamento:
desm_df_desmatamento_ips_2014 <- df_desmatamento_ips_2014[,67:74]
desm_df_desmatamento_ips_2018 <- df_desmatamento_ips_2018[,67:74]

## A base de dados é composta por 3 classificações (No IPS):
## Dimensão > Componente > Indicador
## Dividiremos a base, então, nestas 3 possíveis classificações.

### Dimensão ###
dimensao_df_desmatamento_ips_2014 <- df_desmatamento_ips_2014[,6:8]
```

```
dimensao_df_desmatamento_ips_2018 <- df_desmatamento_ips_2018[,6:8]

### Componente ###
componente_df_desmatamento_ips_2014 <- df_desmatamento_ips_2014[,9:20]
componente_df_desmatamento_ips_2018 <- df_desmatamento_ips_2018[,9:20]

### Indicadores ###
indicadores_df_desmatamento_ips_2014 <- df_desmatamento_ips_2014[,21:63]
indicadores_df_desmatamento_ips_2018 <- df_desmatamento_ips_2018[,21:63]

## Instanciando tabelas de slices
sl_indicadores_df_desmatamento_ips_2014 <- indicadores_df_desmatamento_ips_2014
sl_indicadores_df_desmatamento_ips_2018 <- indicadores_df_desmatamento_ips_2018

## Retirando colunas com muitos NA's
names(sl_indicadores_df_desmatamento_ips_2014) <-
  make.names(names(sl_indicadores_df_desmatamento_ips_2014))
sl_indicadores_df_desmatamento_ips_2014 <-
within(sl_indicadores_df_desmatamento_ips_2014, {
  Mortalidade_materna__Obitos_maternos_por_100_mil_nascidos_vivo <- NULL
  Violencia_contra_a_mulher__casos_por_100_mil_mulheres__ <- NULL
  Pessoas_ameacadas__numero_de_ameacados_de_morte_por_100_mil_ <- NULL
  Desmatamento_recente__percent_do_desmatamento_de_2015._2016._20 <- NULL
  Desmatamento_acumulado__percent_ <- NULL
  Areas_Protegidas__percent_ <- NULL
  Suicidio__Obitos_por_100_mil_habitantes__ <- NULL
  Mortalidade_por_doencas_infecciosas__Obitos_por_100_mil_habit <- NULL
  Mortalidade_por_desnutricao___Obitos_por_100_mil_habitantes_ <- NULL
  Pessoas_ameacadas__numero_de_ameacados_de_morte_por_100_mil_ <- NULL
  Desmatamento_recente__percent_do_desmatamento_de_2015._2016._20 <- NULL
  Areas_Protegidas__percent_ <- NULL
})
```

```
## Retirando colunas com muitos NA's
names(sl_indicadores_df_desmatamento_ips_2018) <-
  make.names(names(sl_indicadores_df_desmatamento_ips_2018))
sl_indicadores_df_desmatamento_ips_2018 <-
within(sl_indicadores_df_desmatamento_ips_2018, {
  Mortalidade_materna__Obitos_maternos_por_100_mil_nascidos_vivo <- NULL
  Violencia_contra_a_mulher__casos_por_100_mil_mulheres__ <- NULL
  Pessoas_ameacadas__numero_de_ameacados_de_morte_por_100_mil_ <- NULL
  Desmatamento_recente__percent_do_desmatamento_de_2015._2016._20 <- NULL
  Desmatamento_acumulado__percent_ <- NULL
  Areas_Protegidas__percent_ <- NULL
  Suicidio__Obitos_por_100_mil_habitantes__ <- NULL
  Mortalidade_por_doencas_infecciosas__Obitos_por_100_mil_habit <- NULL
  Mortalidade_por_desnutricao___Obitos_por_100_mil_habitantes_ <- NULL
  Pessoas_ameacadas__numero_de_ameacados_de_morte_por_100_mil_ <- NULL
  Desmatamento_recente__percent_do_desmatamento_de_2015._2016._20 <- NULL
  Areas_Protegidas__percent_ <- NULL
})
################################################################################
```

APÊNDICE 4 – SCRIPT RSTUDIO – PCA

```
### Normalização das bases ###
norm_componente_df_desmatamento_ips_2014 <-
as.data.frame(scale(componente_df_desmatamento_ips_2014))
norm_componente_df_desmatamento_ips_2018 <-
as.data.frame(scale(componente_df_desmatamento_ips_2018))

#########################################################################
#######

### PCA ###
  ### Componentes ###
res.comp_2014 = imputePCA(norm_componente_df_desmatamento_ips_2014,ncp=5)
res.comp_2018 = imputePCA(norm_componente_df_desmatamento_ips_2018,ncp=5)
res.pca_comp_2014 = PCA(res.comp_2014$completeObs, ncp=4)
res.pca_comp_2018 = PCA(res.comp_2018$completeObs, ncp=4)

eig.val_comp_2014 <- get_eigenvalue(res.pca_comp_2014)
#n_val_comp_2014 <- 3

eig.val_comp_2018 <- get_eigenvalue(res.pca_comp_2018)
#n_val_comp_2018 <- 4

res.pca_comp_2014$var$contrib
res.pca_comp_2014$ind

#########################################################################
#######

# Cotovelo
factoextra::fviz_eig(res.pca_comp_2014, addlabels = TRUE, ylim = c(0, 50))
```

```
# Quadrantes coloridos
fviz_pca_var(res.pca_comp_2014, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping
)

## Estimação
  local({
    .PC_comp_2014 <-

princomp(~Acesso_a_educacao_superior+Acesso_a_informacao_e_comunicacao+Acesso_ao
_conhecimento_basico+Agua_e_saneamento+Direitos_individuais+Liberdade_individual_e_
de_escolha+Moradia+Nutricao_e_cuidados_medicos_basicos+Qualidade_do_meio_ambiente
+Saude_e_bem_estar+Seguranca_pessoal_+Tolerancia_e_inclusao,
               cor=FALSE, data=norm_componente_df_desmatamento_ips_2014)
    norm_componente_df_desmatamento_ips_2014 <<-
      within(norm_componente_df_desmatamento_ips_2014, {
        Seguranca_2014 <- .PC_comp_2014$scores[,4]
        Acesso_ao_conhecimento_e_tolerancia_e_Necessidades_Humanas_Basicas_2014
<- .PC_comp_2014$scores[,3]
        Bem_estar_2014 <- .PC_comp_2014$scores[,2]
        Oportunidade_2014 <- .PC_comp_2014$scores[,1]
      })
  })



#############################################################################
#######

# Cotovelo
factoextra::fviz_eig(res.pca_comp_2018, addlabels = TRUE, ylim = c(0, 50))
```

```
# Quadrantes coloridos
fviz_pca_var(res.pca_comp_2018, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping
)

res.pca_comp_2018$var$contrib
res.pca_comp_2018$ind

## Estimação
  .PC_comp_2018 <-

princomp(~Acesso_a_educacao_superior+Acesso_a_informacao_e_comunicacao+Acesso_ao
_conhecimento_basico+Agua_e_saneamento+Direitos_individuais+Liberdade_individual_e_
de_escolha+Moradia+Nutricao_e_cuidados_medicos_basicos+Qualidade_do_meio_ambiente
+Saude_e_bem_estar+Seguranca_pessoal_+Tolerancia_e_inclusao,
             cor=FALSE, data=norm_componente_df_desmatamento_ips_2018)

    norm_componente_df_desmatamento_ips_2018 <<-
within(norm_componente_df_desmatamento_ips_2018, {
      Seguranca_2018 <- .PC_comp_2018$scores[,4]
      Acesso_ao_conhecimento_e_tolerancia_e_Necessidades_Humanas_Basicas_2018
<- .PC_comp_2018$scores[,3]
      Bem_estar_2018 <- .PC_comp_2018$scores[,2]
      Oportunidade_2018 <- .PC_comp_2018$scores[,1]
  })
```

APÊNDICE 5 – SCRIPT RSTUDIO – INSERÇÃO NO BANCO DE DADOS

```
## Juntando a coluna-resposta de incremento às vars explicativas (todas em escala):

# Comp 2014
norm_componente_df_desmatamento_ips_2014_1 <- bind_cols(mun_ano_codIBGE_2014,
inc_area_2014_scale, norm_componente_df_desmatamento_ips_2014)
colnames(norm_componente_df_desmatamento_ips_2014_1)[5] <- 'inc_area_2014_scale'

names(norm_componente_df_desmatamento_ips_2014_1) <-
  make.names(names(norm_componente_df_desmatamento_ips_2014_1))
norm_componente_df_desmatamento_ips_2014_1 <-
within(norm_componente_df_desmatamento_ips_2014_1, {
  Nutricao_e_cuidados_medicos_basicos <- NULL
  Agua_e_saneamento <- NULL
  Moradia <- NULL
  Seguranca_pessoal_ <- NULL
  Acesso_ao_conhecimento_basico <- NULL
  Acesso_a_informacao_e_comunicacao <- NULL
  Saude_e_bem_estar <- NULL
  Qualidade_do_meio_ambiente <- NULL
  Direitos_individuais <- NULL
  Liberdade_individual_e_de_escolha <- NULL
  Tolerancia_e_inclusao <- NULL
  Acesso_a_educacao_superior <- NULL
})

# Comp 2018
norm_componente_df_desmatamento_ips_2018_1 <-
bind_cols(mun_ano_codIBGE_2014,inc_area_2018_scale,
norm_componente_df_desmatamento_ips_2018)
colnames(norm_componente_df_desmatamento_ips_2018_1)[5] <- 'inc_area_2018_scale'
```

```
names(norm_componente_df_desmatamento_ips_2018_1) <-
  make.names(names(norm_componente_df_desmatamento_ips_2018_1))
norm_componente_df_desmatamento_ips_2018_1 <-
within(norm_componente_df_desmatamento_ips_2018_1, {
  Nutricao_e_cuidados_medicos_basicos <- NULL
  Agua_e_saneamento <- NULL
  Moradia <- NULL
  Seguranca_pessoal_ <- NULL
  Acesso_ao_conhecimento_basico <- NULL
  Acesso_a_informacao_e_comunicacao <- NULL
  Saude_e_bem_estar <- NULL
  Qualidade_do_meio_ambiente <- NULL
  Direitos_individuais <- NULL
  Liberdade_individual_e_de_escolha <- NULL
  Tolerancia_e_inclusao <- NULL
  Acesso_a_educacao_superior <- NULL
})
#####
## Insert no BD
dbRemoveTable(con, "norm_componente_df_desmatamento_ips_2014_1",
norm_componente_df_desmatamento_ips_2014_1)
dbCreateTable(con, "norm_componente_df_desmatamento_ips_2014_1",
norm_componente_df_desmatamento_ips_2014_1)
dbWriteTable(con, "norm_componente_df_desmatamento_ips_2014_1",
norm_componente_df_desmatamento_ips_2014_1, overwrite=T)
dbRemoveTable(con, "norm_componente_df_desmatamento_ips_2018_1",
norm_componente_df_desmatamento_ips_2018_1)
dbCreateTable(con, "norm_componente_df_desmatamento_ips_2018_1",
norm_componente_df_desmatamento_ips_2018_1)
dbWriteTable(con, "norm_componente_df_desmatamento_ips_2018_1",
norm_componente_df_desmatamento_ips_2018_1, overwrite=T)
dbDisconnect(con)
```

APÊNDICE 6 – SCRIPT RSTUDIO – MODELAGEM

```
multi_rqfit14 <- rq(inc_area_2014~
                    Bem_estar_2014+
                    Seguranca_2014+
                    Oportunidade_2014+

Acesso_ao_conhecimento_e_tolerancia_e_Necessidades_Humanas_Basicas_2014,
                    data = norm_componente_df_desmatamento_ips_2014, tau = seq(0, 1, by =
0.05))
###estimativas, limites inferiores e superiores
summary(multi_rqfit14)

colMeans(multi_rqfit14$residuals^2)

##### p-valor
summary(multi_rqfit14,se="ker")

multi_rqfit18 <- rq(inc_area_2018~
                    Bem_estar_2018+
                    Seguranca_2018+
                    Oportunidade_2018+

Acesso_ao_conhecimento_e_tolerancia_e_Necessidades_Humanas_Basicas_2018,
                  data = norm_componente_df_desmatamento_ips_2018, tau = seq(0, 1, by = 0.05))
###estimativas, limites inferiores e superiores
summary(multi_rqfit18)

colMeans(multi_rqfit18$residuals^2)

##### p-valor
summary(multi_rqfit18,se="ker")
```

APÊNDICE 7 – SCRIPT PYTHON MAPA

```
fig = px.choropleth_mapbox(map_2014,
 geojson = br_json_map,
 locations='IBGEDados',
 featureidkey = 'properties.id',
 color = 'Bem_estar_2014',
 labels = 'Bem_estar_2014',
 hover_name = 'Municipio',
 hover_data = ['Bem_estar_2014', 'inc_area_2014'],
 color_continuous_scale='blues',
 mapbox_style = 'carto-positron',
 center = {'lat':-8.2665485718, 'lon': -50.9861401688},
 zoom = 3,
 opacity = 0.6 )
fig.update_geos(fitbounds = 'locations', visible = False)

fig.add_trace(go.Scattermapbox(
        lat=map_2014.lat,
        lon=map_2014.long,
        mode='markers',
        marker=go.scattermapbox.Marker(
            color=map_2014['inc_area_2014'] ,
            size=map_2014['inc_area_2014'] * 40,
            opacity=0.85,
            reversescale=True
        ),
        text=map_2014['Municipio'],
        hoverinfo='text',
        name = 'Desmatamento por área'
    ))

fig.update_layout(
        title_text = 'Bem estar x Desmatamento 2014',
        showlegend=True,
        legend=dict(
             yanchor="top",
             y=0.99,
             xanchor="left",
             x=0.01
        ))fig.show()
```