ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

CURSO DE ENGENHARIA DE COMPUTAÇÃO

MEGUMI TSURU
SAMUEL VIEIRA DUCCA
VITOR DIAS SOUZA

PLATAFORMA ONLINE PARA INFERÊNCIA DE DADOS DE INDICADORES AMBIENTAIS E SOCIOECONÔMICOS UTILIZANDO REDES NEURAIS EM IMAGENS DE SATÉLITE

SÃO PAULO - SP

2021

MEGUMI TSURU
SAMUEL VIEIRA DUCCA
VITOR DIAS SOUZA

PLATAFORMA ONLINE PARA INFERÊNCIA DE DADOS DE INDICADORES AMBIENTAIS E SOCIOECONÔMICOS UTILIZANDO REDES NEURAIS EM IMAGENS DE SATÉLITE

Trabalho apresentado à Escola Politécnica da Universidade de São Paulo como requisito para obtenção do título de Bacharel em Engenharia de Computação.

Área de concentração: Engenharia de Computação

Orientador: Prof. Dr. Pedro Luiz Pizzigatti Corrêa

Co-Orientadora: Dra. Marina Jeaneth Machicao Justo

Prof. Dr. Pedro Luiz Pizzigatti Corrêa (orientador)

Escola Politécnica da USP

PhD. Marina Jeaneth Machicao Justo (co-orientador)

Escola Politécnica da USP

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catalogação-na-publicação

PLATAFORMA ONLINE PARA INFERÊNCIA DE DADOS DE INDICADORES AMBIENTAIS E SOCIOECONÔMICOS UTILIZANDO REDES NEURAIS EM IMAGENS DE SATÉLITE / M. Tsuru, S. Vieira Ducca, V. Dias Souza -- São Paulo, 2021.
121 p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.redes neurais 2.análise de imagens de satélite 3.indicadores ambientais 4.indicadores socioeconômicos 5.previsão de dados I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t. III.Vieira Ducca, Samuel IV.Dias Souza, Vitor

Agradecimentos

Agradecemos à nossas famílias pelo apoio incondicional e por todo o suporte dado ao longo da nossa graduação.

Agradecemos ao Prof. Dr. Pedro Luiz Pizzigatti Corrêa e Dra. Marina Jeaneth Machicao Justo por nos aceitarem como orientados e nos guiar nessa árdua e recompensadora jornada que foi desenvolver este projeto. Agradecemos por todo apoio, pela atenção e sobretudo pela imensa paciência conosco.

Agradecemos à parceria NEXUS-PARSEC e a todos os seus pesquisadores pelo apoio na condução deste projeto e também à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) por permitir que tal parceria fosse possível. Em especial, gostaríamos de agradecer ao Dr. Pedro Ribeiro de Andrade Neto e à Danton Ferreira Vellenich por todo o auxílio prestado na execução deste projeto.

Resumo

O Brasil é o país com a quinta maior extensão territorial do mundo, e que contém uma grande diversidade ecológica e de biomas em seu território. Para que esses recursos naturais sejam manejados de forma sustentável, é fundamental que haja um monitoramento adequado das condições socioambientais do país — o que é um grande desafio, dada a sua dimensão continental e dificuldade de acesso a zonas de preservação. O emprego de tecnologias que permitam inferir a situação socioambiental de uma região sem a necessidade de trabalho de campo pode ajudar a fornecer informações valiosas para facilitar iniciativas voltadas ao desenvolvimento sustentável e à preservação da biodiversidade.

Este trabalho usa metodologias de aprendizado de máquina para análise de imagens de satélite para estimar indicadores ambientais e socioeconômicos. O modelo de rede neural desenvolvido foi treinado e aplicado em quatro estados do Nordeste brasileiro, com foco no bioma da Caatinga. Os resultados do trabalho foram compilados e apresentados em uma plataforma *web* de mapas interativos, de modo a disponibilizar os dados usados e facilitar a comparação e compreensão dos indicadores estimados pelo modelo.

Palavras Chave: aprendizado de máquina, imagens de satélite, previsão de dados, indicadores ambientais, indicadores socioeconômicos

Abstract

Brazil is the fifth largest country in the world, a country that has great ecological and biological diversity within its territory. In order for these natural resources to be sustainably managed, it is essential to adequately monitor the country's social and environmental conditions — which is a major challenge given the country's territorial extent and the difficulty of accessing some of its protected areas. Using technologies that make it possible to assess the social and environmental situation of a region without having to do field work can help to provide valuable information to promote initiatives for sustainable development and biodiversity conservation on the national territory.

This work uses machine learning methods for the analysis of satellite images with the goal of predicting environmental, social, and economic indicators. The neural network model developed was trained and applied in four estates of the brazilian northwest, focusing on the Caatinga biome. The results were compiled and presented on a web platform in the form of interactive maps to make data accessible as well as facilitate comparison and understanding the indicators predicted by the model.

Keywords: machine learning, satellite images, data prediction, environmental indicators, socioeconomic indicators

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AL	Alagoas
API	*Application Programming Interface* - Interface de programação de aplicação.
CNN	*Convolutional Neural Networks* - Redes Neurais Convolucionais
DHS	*Demographic and Health Surveys* - Programa de Pesquisas Demográficas e de Saúde
DMSP	*Defense Meteorological Satellite Program* - Programa de Satélite de Defesa Metereológica
EM	*Expectation-Maximization* - Especulação-Maximização
GAUL	*The Global Administrative Unit Layers* - Unidades Administrativas Globais
GPU	*Graphics Processing Unit* - Unidade de Processamento Gráfico
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
INPE	Instituto Nacional de Pesquisas Espaciais
LSMS	*Living Standards Measurement Study* - Estudo de Medição de Padrões de Vida
MAE	*Mean Absolute Error* - Erro Médio Absoluto
NCEI	*National Center for Environmental Information* - Centro Nacional de Informação Ambiental
NOAA	*National Oceanic and Atmospheric Administration* - Administração Atmosférica e Oceânica Nacional
NTSG	*Numerical Terradynamic Simulation Group* - Grupo de Simulação Numérica e Termodinâmica
PB	Paraíba
PIB	Produto Interno Bruto
PRONAF	Programa Nacional de Fortalecimento da Agricultura Familiar
RMSE	*Root Mean Square Error* - A Raiz do Erro Quadrático Médio
RN	Rio Grande do Norte
RSS	*Residual Sum of Squares* - Soma Residual de Quadrados
SE	Sergipe
TMF	Taxa de mortalidade fetal
TSS	*Total Sum of Squares* - Soma Total de Quadrados
UNCS	*UN Cartographic Unit* - Unidade Cartográfica das Nações Unidas

Sumário

1. Introdução

1.1. Motivação

O assunto deste trabalho consiste na utilização de metodologias de aprendizado de máquina para realizar a previsão de indicadores ambientais e socioeconômicos através da análise de imagens de satélite.

O Brasil é um país de vasta extensão territorial, com mais de 8 milhões de km^2 (IBGE, 2021) [1], que o tornam a nação com a quinta maior área do mundo. Neste amplo território, que compreende mais da metade da América do Sul, se encontra uma grande variedade de biomas e a maior biodiversidade do planeta, com mais de 116.000 espécies de animais e 46.000 espécies vegetais conhecidas, 20% da biodiversidade mundial [2]. A abundante riqueza da fauna e flora brasileiras é fonte de recursos importantes para o país, não apenas em serviços ecossistêmicos, mas também pelas oportunidades apresentadas em sua conservação, uso sustentável e no seu patrimônio genético.

Entretanto, para que esses recursos naturais sejam manejados de forma sustentável, é fundamental que haja um monitoramento adequado das condições socioambientais do país — o que é um grande desafio no caso do Brasil, dada sua extensão territorial e dificuldade de acesso a zonas de preservação. Neste contexto, o emprego de metodologias de aprendizado de máquina para análise de imagens de satélite pode ajudar a fornecer informações valiosas para o monitoramento ambiental e socioeconômico, facilitando assim iniciativas voltadas para o desenvolvimento sustentável e preservação da biodiversidade em território nacional.

A motivação para o emprego da associação dessas duas tecnologias vem da recente aplicação de técnicas de redes neurais profundas para estimar critérios socioeconômicos em países africanos com base em imagens de satélite de seus territórios, demonstrada por Jean et al. (2016) [3] e Yeh et al. (2020) [4].

As vantagens fornecidas pelo emprego desta metodologia, sobretudo a não necessidade de deslocamento para a área que se deseja estudar, combinada com a possibilidade de se cobrir vastas zonas territoriais e a capacidade de automatização do processo análise, tornam o uso de redes neurais em imagens de satélite uma técnica ideal para estudar a situação socioambiental de um território, sobretudo no caso brasileiro.

1.2. Objetivo

O objetivo do projeto é desenvolver um sistema que empregue métodos baseados em redes neurais para identificar padrões em imagens de satélite de forma a prever indicadores socioeconômicos e ambientais. Esse sistema será concretizado em uma plataforma web que permita o fácil acesso e visualização dos indicadores obtidos pelo modelo.

No contexto do trabalho, “previsão” não significa tentar traçar tendências futuras, mas sim estimar certos indicadores socioambientais em uma determinada área ou período, mesmo sem a coleta destes dados em pesquisas de campo. Assim, a proposta é construir um modelo que seja ‘treinado’ para uma região, e que possa ser utilizado futuramente para prever indicadores na mesma área, quando novas imagens de satélite estiverem disponíveis.

Os indicadores e a área para treinamento do modelo são fornecidos pelo projeto NEXUS-PARSEC [5-6], uma parceria entre a Escola Politécnica da Universidade de São Paulo e o Instituto Nacional de Pesquisas Espaciais (INPE), que visa aproveitar dados coletados pelo projeto NEXUS [5] para a realização de projetos de aprendizado profundo. A Área Nexus, foco do trabalho, consiste nos biomas brasileiros da Caatinga e do Cerrado, englobando grande parte do Nordeste e Centro-Oeste do país.

O projeto também visa compartilhar os resultados alcançados com o consórcio internacional de compartilhamento de dados socioeconômicos de áreas protegidas PARSEC [6], do qual este trabalho faz parte, bem como disponibilizar através da plataforma online os dados obtidos para que possam servir de base para novos estudos e pesquisas relacionadas à predição de indicadores ambientais e socioeconômicos.

1.3. Justificativa

Nos últimos séculos, um dos principais mecanismos de impulso ao desenvolvimento econômico de uma nação tem sido a expansão de sua produção industrial e de seu consumo de insumos. Com isso, a exploração de recursos naturais e a ação antrópica sobre os diversos biomas do planeta têm se intensificado exponencialmente, produzindo desequilíbrios ecológicos cada vez mais intensos e muitas vezes até irreversíveis. Além da perda de biodiversidade, estes desequilíbrios têm o potencial de prejudicar seriamente a qualidade de vida de gerações futuras, restringindo seu acesso a recursos e afetando atividades econômicas, como por exemplo a agricultura. Por isso, a preservação ambiental

e o desenvolvimento sustentável são assuntos que têm se tornado cada vez mais importantes nas pautas das políticas nacionais.

O Brasil é o país com a quinta maior extensão territorial do mundo [7] e que contém uma grande diversidade ecológica e de biomas em seu território. Para efetivamente promover a preservação de sua fauna e flora e para se tomar medidas que favoreçam um desenvolvimento sustentável da nação, é necessário que haja um monitoramento adequado da situação ambiental e socioeconômica nos vários biomas do Brasil. Isso se mostra um grande desafio no cenário brasileiro, dada a sua vasta extensão territorial.

Nesse contexto, o emprego de metodologias de aprendizado de máquina para analisar imagens de satélite dos diversos biomas do território nacional pode se mostrar uma ferramenta de grande potencial não só para monitorar uma área efetivamente maior, mas também para tornar esse monitoramento automatizado e mais eficiente. Assim, seria possível prever indicadores socioambientais de forma mais fácil, complementando os dados obtidos através de pesquisas de campo e construindo um conjunto mais amplo e detalhado de informações, que pode auxiliar a fomentar políticas públicas de desenvolvimento sustentável e preservação ecológica no país.

Nos últimos anos, o projeto NEXUS [5] têm trabalhado para estabelecer indicadores socioeconômicos na região da Bacia do Rio São Francisco e em torno de regiões brasileiras que cobrem os biomas da Caatinga e Cerrado, compondo uma quantidade significativa de dados para as municipalidades contidas nessa região. A Área Nexus, selecionada como região de estudo para o trabalho, engloba os biomas da Caatinga e do Cerrado, que contém os principais estoques de terras disponíveis para expansão agrícola no Brasil, além de áreas de elevado potencial solar e eólico [5] — sendo assim de grande importância tanto no contexto socioeconômico, como no ambiental.

Em paralelo, o projeto PARSEC [6] tem trabalhado no último ano para averiguar o impacto de ações conduzidas na região nas populações das municipalidades locais em termos dos indicadores mencionados através de métodos de aprendizado profundo e sensoriamento remoto.

Dessa forma foi possível identificar um ponto de intersecção entre os dois projetos que constituiu a parceria NEXUS-PARSEC, da qual este projeto faz parte, que visa utilizar as metodologias de análise e monitoramento propostas pelo projeto PARSEC em conjunto com os dados obtidos pelo projeto NEXUS.

1.4. Organização do Trabalho

Esta monografia encontra-se dividida em oito capítulos, descritos a seguir:

No primeiro capítulo é feita a introdução para o trabalho, citando sua motivação, objetivo, justificativa e forma de organização.

No segundo capítulo são descritos os aspectos conceituais do projeto, incluindo definições de metodologias e revisão da literatura de referência.

No terceiro capítulo são apresentadas as tecnologias utilizadas para o desenvolvimento do trabalho.

No quarto capítulo é apresentada a metodologia do trabalho, especificando as etapas tomadas para sua execução.

No quinto capítulo é descrita a especificação de requisitos para o projeto, tanto no âmbito do modelo de rede neural, como para a plataforma online desenvolvida.

No sexto capítulo são apresentados os detalhes do projeto e da implementação realizadas, descrevendo o treinamento do modelo de aprendizado de máquina e o desenvolvimento da plataforma online.

No sétimo capítulo são apresentados os resultados obtidos pelo modelo treinado, bem como a análise dos dados resultantes.

No oitavo e último capítulo são feitas as considerações finais sobre o projeto realizado, com as conclusões alcançadas e perspectivas de continuidade para trabalhos futuros.

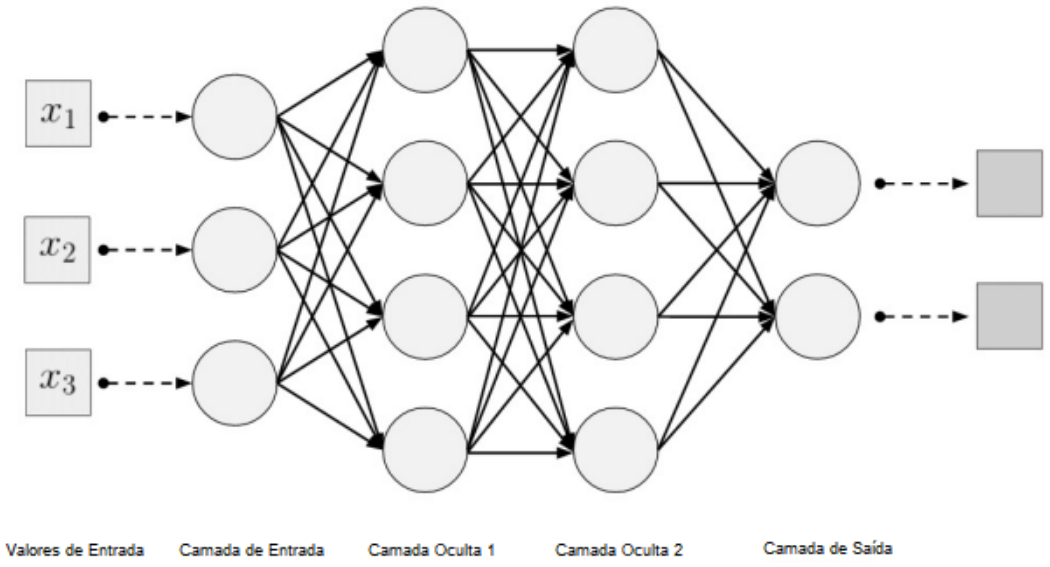
2. Aspectos Conceituais

2.1. Redes Neurais Artificiais

Uma Rede Neural Artificial é um modelo computacional inspirado na estrutura neural de animais, composta por uma série de nós (ou neurônios) - unidades de processamento de funcionamento simples trabalhando em paralelo, em vez de existir uma única unidade de controle centralizada. Esses neurônios se interconectam entre si através de canais de comunicação, cada um com um determinado peso (*weight*) associado. Os pesos nas conexões em uma rede neural são coeficientes que amplificam ou minimizam o sinal de entrada de um determinado neurônio na rede (PATTERSON; GIBSON, 2017) [8].

Segundo Patterson e Gibson (2017) [8], o comportamento de uma rede neural é moldado pela sua arquitetura de rede, que pode ser essencialmente definida através do: número de neurônios, número de camadas e tipos de conexões entre as camadas. A Figura 1 exemplifica uma possível topologia de uma rede neural.

Figura 1: Exemplo de topologia de uma rede neural, neste caso, de uma rede neural com múltiplas camadas do tipo *feed-forward.* Ela tem uma camada de entrada, uma ou mais camadas ocultas e uma única camada de saída. Cada camada pode ter um número diferente de neurônios e está totalmente conectada à camada adjacente.



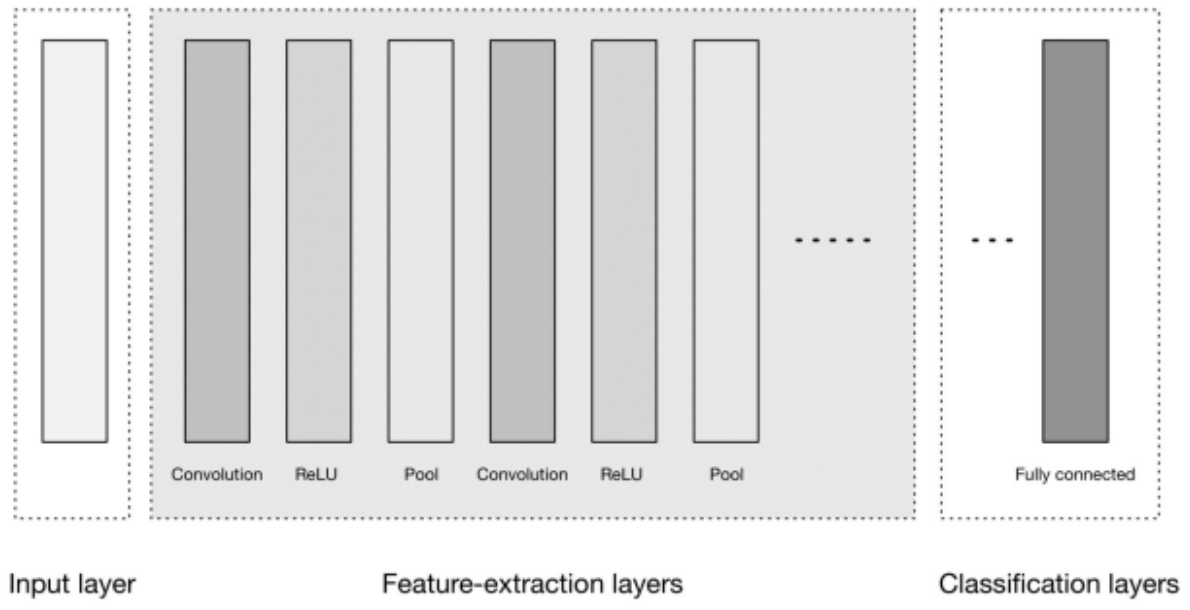
Fonte: modificado de Patterson e Gibson (2017) [8].

Técnicas de Aprendizado Profundo (*Deep Learning*) utilizam redes neurais artificiais com um grande número de camadas, sendo capazes de processar um grande número de parâmetros (PATTERSON; GIBSON, 2017) [8]. As Redes Neurais Convolucionais (*Convolutional Neural Networks,* ou CNN) são um exemplo de um modelo de Aprendizado Profundo voltado especificamente para o processamento de dados que possuem uma topologia conhecida em forma de *grid* - por exemplo, dados de imagem, que podem ser pensados como uma grade de pixels de duas dimensões. O nome “Rede Neural Convolucional” indica que a rede emprega uma operação matemática chamada convolução, um tipo especializado de operação linear. As redes convolucionais são simplesmente redes neurais artificiais que utilizam a operação de convolução em pelo menos uma de suas camadas (GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A., 2016) [9].

Segundo Patterson e Gibson (2017) [8], apesar de existirem diversas variações na arquitetura de uma rede neural convolucional, no geral ela é baseada no padrão de camadas apresentado na Figura 2, composto essencialmente de três grandes grupos:

1. Camada de entrada (*input layer*);
2. Camada(s) de extração de características (*feature-extraction layers*);
3. Camada de classificação (*classification layers*).

Figura 2: Arquitetura geral em alto nível de uma rede neural convolucional.



Fonte: Patterson e Gibson (2017) [8].

A camada de entrada é a camada onde são carregados e armazenados os dados brutos de entrada para processamento na rede.

As camadas de extração de características processam os dados de entrada, extraindo uma série de características (*features*) e gerando progressivamente características de ordem superior - as CNNs costumam aprender características gerais visualmente genéricas nas primeiras camadas e gradualmente acumulam características específicas do conjunto de dados nas camadas posteriores. Em geral, as camadas de extração de características são compostas pela repetição da seguinte sequência de subcamadas:

1. **Camada de convolução (*convolution layer*)**, que processa os dados com operações de convolução e passam o resultado para a próxima camada;
2. **Camada de agrupamento (*pooling layer*)**, que são comumente inseridos entre as camadas de convolução para reduzir as dimensões dos dados progressivamente ao longo da rede neural e com isso ajudar a controlar o sobreajuste (*overfitting*).

Finalmente, a camada de classificação possui uma ou mais camadas completamente conectadas (*fully connected layers*) para extrair as características de maior ordem e produzir resultados da classificação.

2.2. Transferência de Aprendizado

Devido à grande quantidade de tempo e de poder de processamento necessários para treinar um modelo CNN do começo, uma abordagem comum é utilizar um modelo CNN previamente treinado que tenha tido bons resultados para um certo problema e treiná-lo em um conjunto adicional de dados - diferente daquele usado para pré-treinar o modelo - para resolver um problema distinto, mas relacionado. Esta técnica é chamada de Transferência de Aprendizado (PATTERSON; GIBSON, 2017) [8].

Segundo Patterson e Gibson (2017) [8], a Transferência de Aprendizado pode ser eficaz nas seguintes situações:

- O conjunto de dados de treinamento é pequeno;
- O conjunto de dados de treinamento compartilha características visuais com o conjunto de dados base.

Existem dois principais casos de uso para a técnica de Transferência de Aprendizado:

- **Utilização de um modelo convolucional existente como um extrator de características:** nesta abordagem, retiramos a última camada totalmente conectada de um modelo pré treinado - a camada de saída da camada de classificação. Dessa forma, esse CNN pré treinado e sem a última camada

pode atuar como um extrator de características - gerando um vetor de características (*features*) correspondentes a um novo conjunto de dados.

- **Realização de ajuste fino de um modelo existente:** nesta segunda abordagem, além de se substituir a última camada classificadora deste modelo, algumas das camadas anteriores são seletivamente retreinadas com um novo conjunto de dados, para ajustar o modelo de acordo com as necessidades.

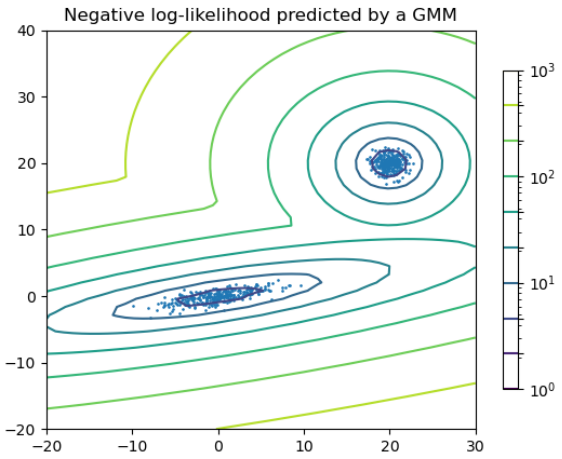
2.3. Modelo de Misturas de Gaussianas

Um Modelo de Misturas de Gaussianas (REYNOLDS, 2015) [10] é um modelo probabilístico que assume que todos os pontos de dados são gerados a partir de uma mistura de um número finito de distribuições gaussianas com parâmetros desconhecidos. Pode-se pensar em modelos de mistura como uma generalização do método de agrupamento *k-means*[1] para incorporar informações sobre a estrutura de covariância dos dados.

A principal dificuldade em gerar os Modelos de Mistura de Gaussianas a partir de dados não rotulados é que geralmente não se sabe quais pontos provêm de qual componente gaussiano. O algoritmo iterativo de Maximização de Expectativa (*Expectation-Maximization* ou EM) é um algoritmo estatístico bem fundamentado para contornar este problema através de um processo iterativo [14]. Este algoritmo assume componentes aleatórios (centrados aleatoriamente em pontos de dados, aprendidos com o método *k-means*, ou mesmo distribuídos normalmente em torno da origem) e calcula para cada ponto uma probabilidade gerada por cada componente do modelo. Em seguida, ajusta-se os parâmetros para maximizar a probabilidade dos dados por essas atribuições. A repetição deste processo é garantida de forma a sempre convergir para um ótimo local. A Figura 3 exemplifica um Modelo de Mistura de Gaussianas com dois componentes.

[1] Método que consiste na separação de um conjunto de dados em k grupos, onde cada amostra pertence ao grupo mais próximo da média.

Figura 3: Gráfico exemplificando os pontos de dados e superfícies de equi-probabilidade de um Modelo de Mistura de Gaussianas de dois componentes. Os dados foram gerados a partir de duas Gaussianas com centros e matrizes de covariância diferentes [14].



Fonte: *Scikit-learn Developers* (2021) [11].

2.4. Regressão Ridge

Modelos de Regressão Linear (JAMES et al., 2013) [12] são utilizados para prever valores de uma variável de interesse (chamada de variável resposta ou dependente) a partir dos valores de outras variáveis (chamadas de variáveis preditoras, independentes ou explicativas), além de serem úteis para estudar a relação entre as variáveis preditoras com a variável resposta. Essa relação pode ser estudada através da estimativa dos chamados coeficientes de regressão.

O procedimento de Regressão Ridge (HOERL; KENNARD, 2000) [13] é um método de estimação útil na presença de multicolinearidade entre as variáveis preditoras, ou então em casos em que o número de variáveis preditoras (p) é muito maior que o tamanho da amostra (n), para evitar problemas de sobreajuste do modelo.

2.4.1. Modelo de Regressão Linear

O modelo de regressão linear múltipla é dado pela seguinte equação.

$$y_i = \widehat{y_i} + \epsilon_i,\ i = 1,\ 2,\ \ldots,\ n \qquad (2.4.1.1)$$

Com:

$$\widehat{y_i} = \beta_0 + \sum_{j=1}^{p} \beta_0 X_{ij},\ i = 1, 2, \ldots, n \qquad (2.4.1.2)$$

Onde:

- y_i representa a i-ésima variável resposta;
- $\widehat{y_i}$ representa a i-ésima variável prevista pelo modelo;
- $\beta_0, \beta_1, \ldots, \beta_p$ são os coeficientes de regressão que devem ser estimados;
- X_{ij} é a variável preditora;
- ϵ_i é o resíduo (ou erro) associado à i-ésima observação;
- n é o número de elementos da amostra;
- p é o número de variáveis preditoras.

Segundo James et al. (2013) [12], uma abordagem comum para estimar os coeficientes de regressão é chamada de método de mínimos quadrados (*least squares*), que consiste no cálculo dos valores de $\beta_0, \beta_1, \ldots, \beta_p$ que minimizam a soma residual dos quadrados (*residual sum of squares* ou RSS), dada pela equação:

$$RSS = \sum_{i=1}^{n} \left(\epsilon_i\right)^2 \qquad (2.4.1.3)$$

Com:

$$\epsilon_i = y_i - \widehat{y_i} + , \ i = 1, 2, \ldots, n \qquad (2.4.1.4)$$

Onde:

- ϵ_i é o resíduo associado à i-ésima observação;
- y_i representa a i-ésima variável resposta;
- $\widehat{y_i}$ representa a i-ésima variável prevista pelo modelo;

No entanto, na presença de multicolinearidade entre as variáveis preditoras, o estimador obtido pelo método de mínimos quadrados pode apresentar grande variância, sendo inadequado para estimar os coeficientes de regressão. Além disso, pode-se ter problemas de sobreajuste quando o número de variáveis preditoras (p) é muito maior que o tamanho da amostra (n), ou seja, em casos em que p >> n. Para esses casos, a Regressão *Ridge* é uma alternativa para tentar solucionar tais problemas (CASAGRANDE, 2016) [14].

2.4.2. Regressão *Ridge*

A Regressão *Ridge* é um método de regressão linear que usa a regularização para reduzir o sobreajuste do modelo. Segundo James et al. (2013) [12], a Regressão Ridge é muito semelhante ao método de mínimos quadrados, exceto que os coeficientes de regressão são estimados pela minimização de um valor ligeiramente diferente - a Regressão *Ridge* reduz os coeficientes de regressão impondo uma penalidade em seu tamanho. Dessa forma, a Regressão *Ridge* estima os valores de β_0, β_1, ..., β_p que minimizam uma soma residual de quadrados penalizada, dada pela expressão:

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (2.4.2.1)$$

Onde:

- RSS é a soma residual de quadrados definida pela Equação.(2.4.1.3);
- $\lambda \geq 0$ é um parâmetro de ajuste (chamado de *tuning parameter* ou *complexity parameter*) que controla o quanto os coeficientes de regressão serão reduzidos;
- p é o número de variáveis preditoras;
- β_j é o j-ésimo coeficiente de regressão a ser estimado.

Vale comentar que quanto maior o valor de λ, maior é a redução dos coeficientes de regressão, os quais se tornam cada vez mais robustos à multicolinearidade. O valor de λ mais adequado para o modelo de predição pode ser estimado através de métodos como a validação cruzada, descrito na Seção 2.6.

2.4.3. Avaliação do desempenho de um modelo de regressão

A fim de avaliar o desempenho de um método de aprendizado estatístico sobre um determinado conjunto de dados, é necessário alguma forma para medir o quão bem suas previsões realmente correspondem aos dados observados. Ou seja, precisa-se quantificar até que ponto o valor de resposta previsto para uma determinada observação está próximo do valor de resposta verdadeiro para essa observação. Nesta seção apresentam-se algumas métricas utilizadas neste projeto.

2.4.3.1. Erro Absoluto Médio (MAE)

Entende-se como Erro Absoluto Médio (*mean absolute error* ou MAE) [18] a média aritmética dos resíduos de um modelo de regressão linear. O valor de MAE pode ser calculado pela seguinte expressão:

$$MAE(y, \widehat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right| \qquad (2.4.3.1.1)$$

Onde:

- y_i representa a i-ésima variável resposta;
- $\widehat{y_i}$ representa a variável prevista para a i-ésima amostra, definida pela Equação (2.4.1.2);
- n é o número de total elementos da amostra.

2.4.3.2. Raiz do Erro Quadrático Médio (RMSE)

No ajuste de um modelo de regressão, a medida mais comumente utilizada é o Erro Médio Quadrático (*mean squared error* ou MSE) [15]. Ele mede a variância dos resíduos do modelo de regressão. O valor de MSE é calculado pela expressão:

$$MSE(y, \widehat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2 \qquad (2.4.3.2.1)$$

Onde:

- y_i representa a i-ésima variável resposta;
- $\widehat{y_i}$ representa a variável prevista para a i-ésima amostra, definida pela Equação.(2.4.1.2);
- n é o número de total elementos da amostra.

A Raiz do Erro Quadrático Médio (*root mean square error* ou RMSE) [15] é obtido calculando-se a raiz quadrada de MSE, fornecendo a medida do desvio padrão dos resíduos do modelo de regressão. O valor de RMSE é calculado pela expressão:

$$RMSE(y, \widehat{y}) = \sqrt{MSE(y, \widehat{y})} \qquad (2.5.3.2.2)$$

2.4.3.3. Coeficiente de Determinação (R^2)

O coeficiente de determinação (*coefficient of determination*, ou R^2) representa a proporção da variância (da variável dependente) que é explicada por um modelo de regressão linear. Ela fornece uma indicação da qualidade do ajuste e, portanto, uma medida

de quão bem as amostras não vistas são previsíveis pelo modelo, através de uma proporção — a proporção da variância explicada.

A melhor pontuação possível para o R² é de 1, podendo também atingir valores negativos. Um modelo constante que sempre prevê o valor esperado de uma variável *y*, independentemente das características dos dados de entrada, obteria uma pontuação R² de 0.

Segundo James et al. (2013) [12], o R² pode ser calculado pela seguinte expressão:

$$R^2(y, \hat{y}) = 1 - \frac{RSS}{TSS} \qquad (2.4.3.3.1)$$

Com:

$$TSS = \sum_{i=1}^{n} \left(y_i - \overline{y_i}\right)^2 \qquad (2.4.3.3.2)$$

E:

$$\overline{y_i} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad (2.4.3.3.3)$$

Onde:

- y_i representa a i-ésima variável resposta;
- $\hat{y}$ representa a variável prevista para a i-ésima amostra, definida pela Equação (2.4.1.2);
- RSS representa soma residual dos quadrados definida pela Equação (2.4.1.3);
- TSS representa a soma total dos quadrados (*total sum of squares*);
- $\overline{y_i}$ é a média aritmética de todos os valores da variável resposta da amostra;
- n é o número de total de elementos da amostra.

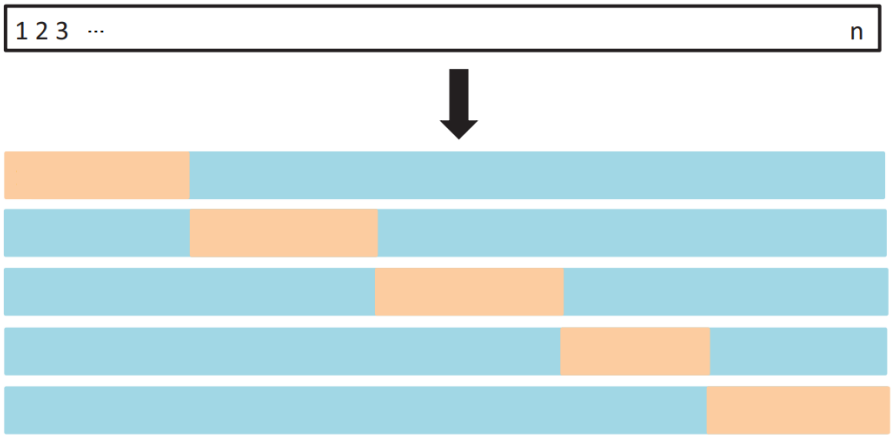
É importante notar que quando RSS > TSS, isso pode resultar em R² com valores negativos. Isso justifica o fato de ter sido obtido valores negativos menores que -1 para alguns casos nos resultados apresentados para este projeto.

2.5. Validação Cruzada

A validação cruzada é uma técnica utilizada para avaliar a capacidade de generalização de um modelo de predição (JAMES et al., 2013) [12]. Neste projeto, adota-se a validação cruzada *k-fold*, que consiste na divisão dos dados de entrada de um modelo em

k subconjuntos de dados (também chamados de *folds*). Um dos *folds* é utilizado como um subconjunto de teste e os demais k-1 são utilizados como subconjuntos de treino do modelo de predição. Esse processo é repetido k vezes, cada vez com um *fold* diferente sendo utilizado para teste, e calculando-se o erro para cada um dos testes. Ao final, pode-se estimar o erro do modelo através da média dos erros de todos os testes realizados. A validação cruzada *k-fold* também pode ser utilizada para a escolha de hiperparâmetros do modelo de predição - no caso da Regressão de Ridge, para a escolha do melhor parâmetro de ajuste (λ) e para o cálculo de métricas como o coeficiente de determinação (R^2).

Figura 4: Desenho esquemático de uma validação cruzada com 5 *folds*. Um conjunto de n observações são divididos aleatoriamente em cinco grupos não sobrepostos. Cada um desses cinco grupos atua como um conjunto de validação (mostrado em bege), e o restante como um conjunto de treinamento (mostrado em azul).



Fonte: modificado de James et al. (2013) [12].

2.5.1. Validação Cruzada com Segregação dos Dados em Subconjuntos de Treino, Validação e Teste

Em algumas abordagens como a proposta por Hastie, Tibshirani e Friedman (2009) [16], além da segregação do conjunto de dados em subconjuntos de treino e de validação (conforme exemplificado na Figura 5), propõe-se a separação de um subconjunto de teste para ser utilizado em uma validação final. Nesta abordagem, após a aplicação da validação cruzada nos subconjuntos de treino e de validação para realizar o ajuste e a seleção do modelo que minimize o erro de predição, realiza-se uma validação final em um terceiro subconjunto de teste para estimar o erro de generalização do modelo preditivo final com uma amostra diferente daqueles utilizados no processo de validação cruzada.

Figura 5: Separação da amostra em subconjuntos de treino, validação e teste.

Fonte: HASTIE, TIBSHIRANI e FRIEDMAN (2009) [16].

2.6. Normalização

A normalização de um conjunto de dados é um requisito comum para muitos estimadores de aprendizagem de máquinas. Eles podem se comportar mal se as características individuais não se parecerem com os dados distribuídos normalmente - dados com média zero e variância unitária.

2.6.1. Normalização Min-Max

A normalização min-max consiste no redimensionamento dos dados em uma certa escala, geralmente em uma escala de 0 a 1, e pode ser calculada pela fórmula:

$$x_i' = \frac{x_i - min(x)}{max(x) - min(x)},\ i = 0, 1,..., n \qquad (2.6.1.1)$$

Onde:

- x_i é o i-ésimo valor original de uma amostra de n elementos;
- x_i' é o i-ésimo valor normalizado a partir de uma amostra de n elementos;
- max(x) o valor máximo da amostra x;
- min(x) o valor mínimo da amostra x.

2.6.2. Método de Z-Score

O método de Z-score se baseia na média e desvio padrão dos dados para normalizar os dados, através da fórmula:

$$z_i = \frac{x_i - \mu}{\sigma},\ i = 0, 1,..., n \qquad (2.6.2.1)$$

Onde:

- x_i é o i-ésimo valor original de uma amostra de n elementos;
- z_i é o i-ésimo valor normalizado a partir de uma amostra de n elementos;
- μ é a média da amostra;
- σ é o desvio padrão da amostra.

2.7. Revisão da Literatura

Os artigos mencionados a seguir são relacionados ao treinamento de modelos de aprendizado de máquina utilizando imagens de satélite para estimar indicadores socioeconômicos, e foram utilizados como principal base de consulta para a realização do trabalho.

2.7.1. Yeh, C. et al. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa (2020) [4]

Neste artigo, imagens de satélite disponíveis publicamente foram utilizadas como base para treinar um modelo de aprendizado de máquina com o objetivo de estimar as condições socioeconômicas em diversas regiões da África subsaariana. Os autores utilizaram como base os dados de riqueza de bens levantados em pesquisas de campo do *Demographic and Health Surveys* (DHS), para gerarem um modelo CNN combinando dois modelos treinados separadamente, um com imagens de luzes noturnas e outro com imagens diurnas, para depois juntá-los em uma camada final totalmente conectada. O objetivo do modelo é de aprender características nas imagens diurnas e noturnas que são preditivas da riqueza de bens, sem previamente ter estabelecido quais as características que o modelo deve procurar.

O modelo foi capaz de explicar em média 70% (R^2 = 0,7) da variação de riqueza de bens aferida nas pesquisas em solo, com resultados nunca abaixo de 50% (R^2 = 0,5).

2.7.2. Jean, N. et al. Combining satellite imagery and machine learning to predict poverty (2016) [3]

O artigo tinha como objetivo construir um modelo preciso, barato e escalável capaz de estimar índices socioeconômicos através da análise de imagens de satélite de alta resolução, empregando um método com um bom grau de acurácia a partir apenas de dados disponíveis em domínio público.

Para atingir esse objetivo foram utilizados dados censitários e imagens de satélite referentes ao território de cinco nações africanas: Nigéria, Tanzânia, Uganda, Malawi e Rwanda. O algoritmo empregou uma técnica de redes neurais convolucionais treinadas para identificar aspectos das imagens analisadas e explicar variações locais de níveis socioeconômicos. O algoritmo proposto tinha como entrada imagens de satélite e como saída dados censitários, mais especificamente índices de pobreza correspondentes ao local analisado.

Os dados necessários para o funcionamento do algoritmo são divididos em três tipos: indicadores socioeconômicos (referentes ao grau de riqueza e pobreza das populações locais), imagens de satélite e imagens de satélite de luzes noturnas. Os dados socioeconômicos foram obtidos através do projeto do Banco Mundial, o *Living Standards Measurement Study* (LSMS) (World Bank, 2020) [17], convertidos em formato **.csv** *(comma-separated value*), enquanto as imagens de satélite foram obtidas por coordenadas geográficas através da API da *Planet* [18] e pelo *Google Static Maps*. Além disso, as imagens de luzes noturnas foram adquiridas pelas interfaces do *National Oceanic and Atmospheric Administration* (NOAA) em conjunto com o *National Center for Environmental Information* (NCEI).

O algoritmo empregado pelo artigo foi construído em linguagem R e funciona com base em uma metodologia de trabalho dividida em 4 etapas distintas: coleta de dados, extração de *features*, treinamento de modelo e validação.

Na etapa de coleta de dados primeiramente uma região de estudo é selecionada e subdivida em *clusters*, cada qual correspondente a um valor numérico do indicador que se procura estudar. Cada cluster corresponde aproximadamente a uma área de 10 km por 10 km e com base em suas coordenadas geográficas. O algoritmo utiliza as API's de imagens para adquirir fotografias de satélite correspondentes a cada cluster (um mínimo de 10 imagens de satélite diurnas por *cluster* e uma imagem de luzes noturnas).

Em posse dos dados censitários e das imagens de satélite, o algoritmo inicia o treinamento de modelo. Inicialmente é realizada a extração de *features* das imagens de satélite diurnas, para que o algoritmo seja capaz de reconhecer aspectos específicos das imagens. Depois, utilizando as imagens de luzes noturnas correspondentes a mesma região das diurnas, o algoritmo filtra e remove locais com baixa intensidade de luz. Isso ocorre pois como no artigo de referência o indicador estudado era de índices socioeconômicos, não fazia sentido estudar regiões sem presença humana significativa, o que podia ser identificado pela ausência de luz em fotos noturnas. Então, com *features* extraídas e imagens devidamente filtradas, o algoritmo efetivamente utilizava os dados censitários e as imagens de satélite diurnas para treinar o modelo de aprendizado profundo.

A ultima etapa consiste em validar resultados obtidos e refinar o modelo. Como resultado, o algoritmo foi capaz de explicar de 55% (R^2 = 0,55) a 75% (R^2 = 0,75) da variação de riqueza medida nos 5 países estudados, e de 37% (R^2 = 0,37) a 55% (R^2 = 0,55) da variação do consumo doméstico, um resultado satisfatório que este trabalho procura replicar para indicadores ambientais.

2.7.3. Triñanes et al. Aplicação de algoritmo de aprendizado profundo para predição de dados socioeconômicos através de imagens de satélite no Vale do Ribeira (2020) [19].

O artigo de Triñanes et al. (2020) [12] buscou reproduzir o artigo de referência (JEAN et al., 2016) [3] e adaptá-lo para realizar a predição de índices socioeconômicos em território brasileiro, mais especificamente para inferir valores de Índice de Desenvolvimento Humano (IDH) nos diversos municípios da região do Vale do Ribeira. Essa adaptação segue a mesma metodologia de trabalho proposta por Jean et al. (2016) [3], mas com o algoritmo, originalmente desenvolvido em linguagem R, adaptado para uma versão utilizando *Python 3*, uma linguagem mais atual e de melhor compreensão para divulgação científica.

A aquisição dos dados se divide em três etapas, assim como definido pela metodologia de Jean et al. (2016) [3]: (i) aquisição de indicadores, (ii) aquisição de imagens de satélite diurnas e (iii) aquisição de imagens de satélite noturnas. O indicador de IDH refere-se às três dimensões de longevidade, alfabetização e renda da população. A partir dos dados censitários do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2010 foram calculados tais indicadores seguindo a metodologia de Abreu et al. (2011) [20]. Nesse artigo foi utilizado o indicador de renda para estudar a região do Vale do Ribeira. Os dados censitários foram obtidos e extraídos em formato **.csv** da base de dados do órgão[2]. Os *clusters* definidos para adaptação correspondiam às subdivisões de setores censitários definidas pelo IBGE para a região do Vale do Ribeira, que também possuíam, de forma semelhante ao artigo (JEAN et al., 2016) [3], uma área de 10 Km por 10 Km.

Para cada cluster foram adquiridas um mínimo de 10 imagens, todas extraídas através do fornecimento das coordenadas geográficas do *cluster* à API da *Planet* [11]. No total 88 clusters foram adotados, com o download de 8800 imagens. Estes dados foram submetidos ao mesmo fluxo de trabalho definido pela metodologia de Jean et al. (2016) [3], com as mesmas etapas posteriores de extração de *features*, treinamento de modelo e validação. Para a região, os autores alcançaram índices de correlação de R^2 = 0,38, que eles consideram suficientes para o estudo.

---

2 https://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Resultados_do_Universo/Agregados_por_Setores_Censitarios/

3. Tecnologias Utilizadas

3.1. Aquisição e Armazenamento de Imagens de Satélite

3.1.1. Aquisição de Imagens de Satélite Diurnas

Para a aquisição de imagens de satélite diurnas, foi usado o *Static Maps API*, um serviço da plataforma *Google Maps* [21] que retorna uma imagem (em formatos .**GIF**, .**PNG** ou .**JPEG**) em resposta a uma solicitação HTTP. Para cada solicitação, é possível especificar a localização, o tamanho, a resolução espacial e o tipo de imagem desejado. As requisições feitas à API para a aquisição das imagens são precificadas no modelo *pay-as-you-go* (pague conforme o uso). Este serviço foi escolhido por fornecer imagens de satélite diurnas de alta resolução quando comparadas a serviços gratuitos como o *Google Earth Engine*.

As imagens de satélite adquiridas foram armazenadas na *Amazon S3* (*Amazon Simple Storage Service*). O serviço permite que o armazenamento e a recuperação de objetos seja realizado por meio de requisições do tipo HTTP feitas à sua API.

3.1.2. Aquisição de Imagens com dados georreferenciados

O *Google Earth Engine* disponibiliza gratuitamente imagens de satélite combinadas com uma variedade de *datasets* georreferenciados através do *Earth Engine API* [22]. O serviço foi utilizado para a aquisição de imagens do tipo GeoTIFF com dois tipos de dados georreferenciados.

- Intensidade de luzes noturnas, cujo *dataset* foi disponibilizado na plataforma pelo *Earth Observation Group* [23] da Escola de Minas do Colorado;
- Evapotranspiração, cujo *dataset* foi disponibilizado na plataforma pelo *Numerical Terradynamic Simulation Group* (NTSG) da Universidade de Montana.

Estes dois tipos de imagens foram usados para treinar modelos CNN através da técnica de Transferência de Aprendizado.

3.2. Dataset ImageNet

O *ImageNet* [24] é um dataset de imagens organizado de acordo com a hierarquia *Wordnet* [25]. Cada conceito significativo no *WordNet*, possivelmente descrito por múltiplas palavras ou frases, é chamado de “set de sinônimo”, existindo mais de cem mil desses “sets” no *WordNet*. O papel do *ImageNet* [24] é essencialmente o de proporcionar uma média de mil imagens para ilustrar cada *set* de sinônimos, realizando um controle de qualidade e aplicando anotações de pessoas nas imagens de cada conceito, de forma a proporcionar dezenas de milhares de imagens rotuladas e organizadas com os “sets de sinônimos” existentes.

Dessa forma, essa base de dados se mostra de grande ajuda para aplicações utilizando inteligência artificial, de modo que permite que tais aplicações sejam capazes de identificar mais facilmente *features* nas imagens analisadas, como por exemplo cantos ou arestas.

3.3. Processamento e Visualização de Dados

Apesar de parte do código original do artigo de referência (JEAN et al., 2016) [3] ter sido escrita com a linguagem de programação Python, existe uma grande parcela que foi desenvolvida apenas com a linguagem R. Para reproduzir a metodologia adotada pelo artigo de referência e posteriormente replicá-lo na Área Nexus, utilizou-se como base uma versão adaptada do código original implementada por Triñanes et al. (2020) [12], que foi totalmente desenvolvida com a linguagem Python. Isso facilitou o trabalho de adaptação, por ser uma linguagem mais atual e flexível do que a linguagem R, e ter melhor integração com outras ferramentas de aprendizado de máquina.

O desenvolvimento foi feito através da ferramenta *Jupyter Notebook*, uma ferramenta que fornece um ambiente de desenvolvimento interativo e que permite criar documentos contendo códigos executáveis, imagens e texto.

Dentre as principais bibliotecas de Python utilizadas no projeto para o processamento e visualização de dados, podem ser destacadas as seguintes:

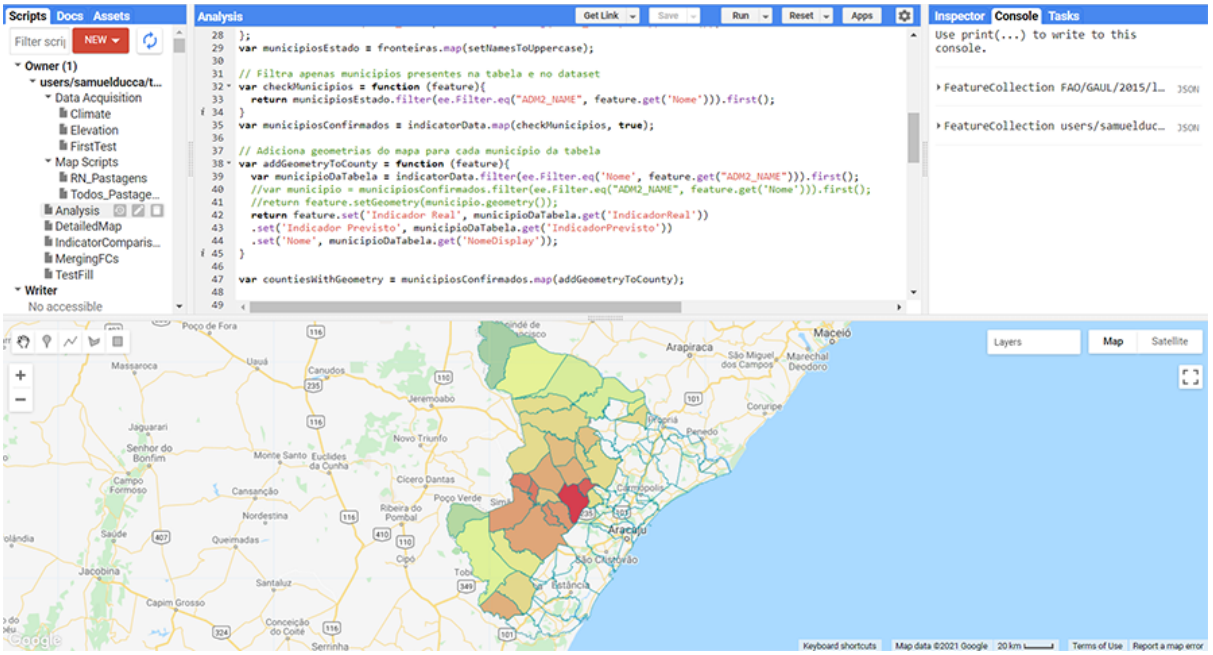
- *Pandas*: biblioteca que facilita a análise e o processamento de dados tabulares, por exemplo permitindo a leitura de arquivos no formato **.csv** que podem ser transformados em *DataFrames* facilmente manipuláveis;
- *Geopandas:* extensão da biblioteca *Pandas* que possibilita e facilita a manipulação dos dados espaciais e georreferenciados e a visualização dos dados em forma de mapas;

- *Scikit-learn:* biblioteca de aprendizado de máquina de código aberto que inclui vários algoritmos de classificação, regressão e agrupamento incluindo o Modelo de Mistura de Gaussianas e modelos de regressão linear;
- *PyTorch*: biblioteca de aprendizagem de máquinas de código aberto baseada na biblioteca Torch, usada para aplicações como visão computacional e processamento de linguagem natural, sendo capaz de realizar cálculos usando tensores (vetores n-dimensionais) com forte aceleração através de unidades de processamento gráfico (GPU);
- *Numpy*: biblioteca que permite a manipulação de arranjos, vetores e matrizes, com ferramentas de cálculo entre matrizes e também ferramentas para geração de números aleatórios, entre outras funcionalidades.
- *Matplotlib*: biblioteca usada para a criação de visualizações estatísticas, em forma de gráficos personalizáveis, interativos e exportáveis para diversos formatos.

3.4. Aplicação *Web*

Os mapas interativos da aplicação *web* foram desenvolvidos na linguagem *Javascript*, utilizando o *framework* do *Google Earth Engine* (Figura 6), uma plataforma para análise científica de *datasets* geoespaciais [22]. Além da aquisição de imagens de satélite, essa plataforma permite executar *scripts* para análise e consolidação de dados, armazenar de *datasets* e tabelas a serem referenciadas, e construir mapas interativos que podem ser incorporados em outros *websites* através do *Earth Engine Apps* [26].

Figura 6: Ambiente de desenvolvimento do *Google Earth Engine.*



Fonte: Compilação do autor.

A aplicação *web* do projeto foi desenvolvida e hospedada de forma gratuita na plataforma do *Google Sites* [27], que permite rápida prototipação e montagem de páginas, bem como a incorporação de aplicativos desenvolvidos no *Earth Engine*.

4. Metodologia do Trabalho

A metodologia de trabalho empregada é composta por duas etapas: (i) o estudo e compreensão dos métodos utilizados no artigo de referência (JEAN et al., 2016) [3] através da reprodução dos resultados obtidos e (ii) a adaptação e uso da metodologia de trabalho descrita no artigo de referência para ser usada no cenário brasileiro da área de estudo do projeto NEXUS-PARSEC [5-6].

4.1. Reprodução do Artigo de Referência

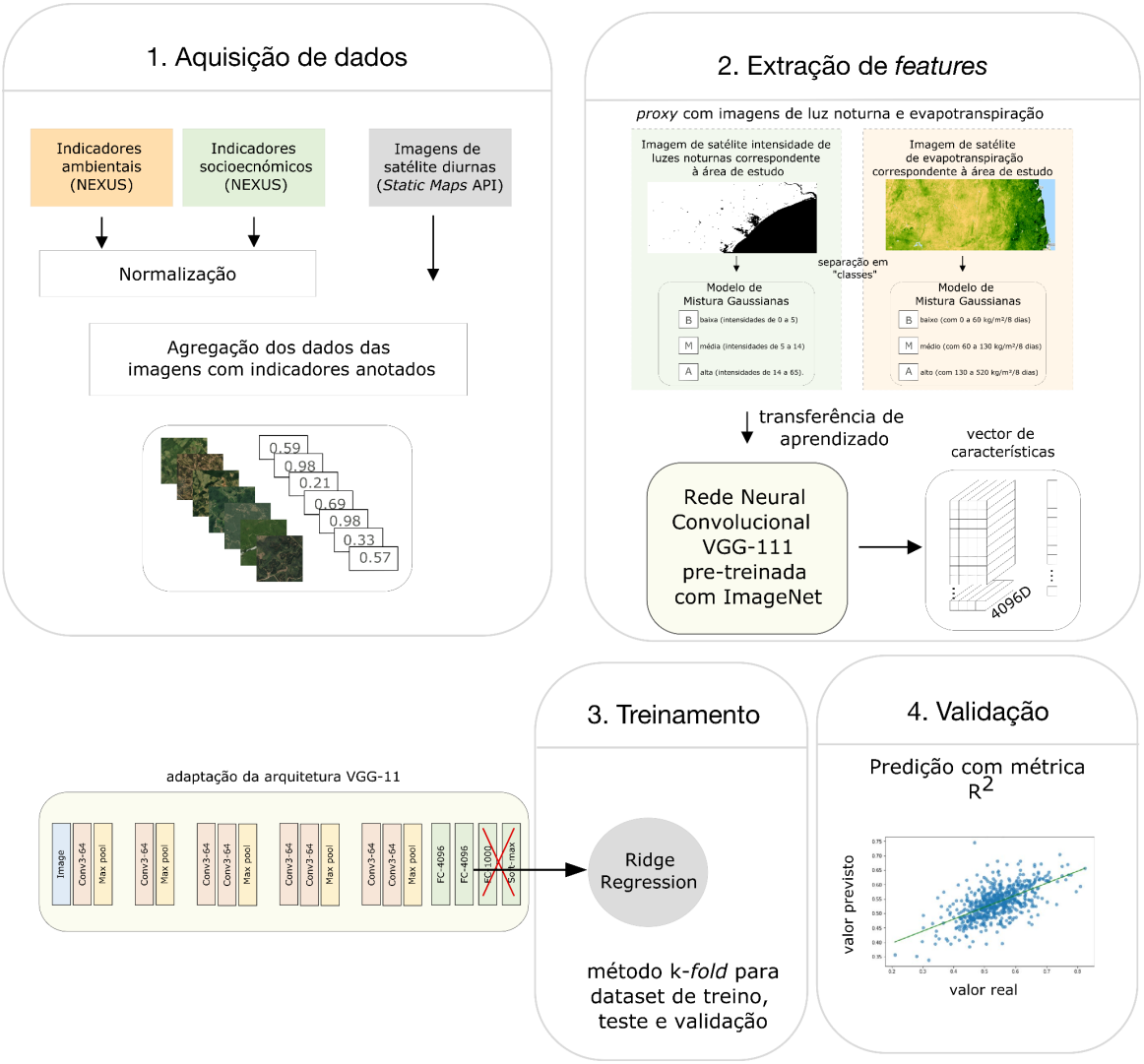
Esta primeira etapa tem como objetivo proporcionar uma melhor compreensão da metodologia empregada pelo artigo de referência (JEAN et al., 2016) [3] quanto aos métodos utilizados para a obtenção de dados, o treinamento do modelo, a validação e a apresentação dos resultados. Para que possam ser obtidos resultados similares aos atingidos pelo artigo, é necessário compreender e executar em sua totalidade o fluxo de trabalho proposto (esquematizado na Figura 7). Em posse de conhecimento prático deste fluxo de trabalho, é possível adaptá-lo para funcionar com entradas distintas para atender as necessidades do projeto.

4.2 Adaptação do Fluxo de Trabalho

O fluxo de trabalho (Figura 7) empregado consiste em quatro passos: (i) aquisição de dados, (ii) extração de *features*, (iii) treinamento de modelo e (iv) validação dos resultados.

Na etapa de aquisição de dados, é realizada a coleta dos indicadores e imagens de satélite a serem utilizados, bem como foi feita a agregação dos dados através de suas coordenadas geográficas. Com o emprego de uma arquitetura de redes neurais convolucionais, é realizada a extração de *features* das imagens através de uma estratégia de transferência de aprendizado utilizando um dataset previamente treinado. Então, é realizada a etapa de treinamento de modelo, na qual primeiramente o algoritmo aplica o modelo VGG-11 [32] como identificador de padrões nas imagens e então direciona sua saída para um segundo modelo de regressão linear, que identifica os padrões reconhecido e que emite como saída final os indicadores previstos. Por fim, é efetuada a validação do modelo, na qual a performance é analisada com métricas de coeficientes de determinação (R^2).

Figura 7: Diagrama ilustrando o fluxo de trabalho com os quatro principais passos para o processo de adaptação do algoritmo estudado e a sua aplicação na área de estudo.



Fonte: Compilação do autor. Arquitetura de rede VGG-11 adaptado de (SIMONYAN; ZISSERMAN, 2015) [32].

4.2.1 Aquisição de Dados

Assim como na metodologia adotada pelo artigo de referência Jean et al. (2016) [3] e por Triñanes et al. (2020) [12], a etapa de aquisição de dados se divide na coleta de indicadores e de imagens de satélite correspondentes à zona estudada. Essa zona é subdividida em *clusters*, cada qual sendo representado por uma coordenada geográfica central de latitude e longitude. Para cada *cluster* é atribuído um valor do indicador estudado, e a coordenada será utilizada para obter imagens de satélites correspondentes à sua área.

Obtêm-se dois tipos de imagem: uma que será propriamente utilizada no treinamento e teste do modelo (exemplos na Figura 8) e outra que servirá de representante (*proxy)* para filtrar as demais imagens. No modelo do artigo de referência (JEAN et al., 2016) [3], a imagem utilizada como *proxy* era de luzes noturnas — o que era feito porque o estudo em questão media critérios socioeconômicos e, sendo assim, não fazia sentido incluir no modelo zonas sem ocupação humana, o que podia ser determinado pela baixa intensidade ou ausência de luzes em imagens noturnas de satélite.

Na adaptação do projeto, o *proxy* de luzes noturnas foi mantido para o treinamento com os indicadores socioeconômicos, mas para os indicadores ambientais foi necessário encontrar um novo *proxy* que tivesse uma relação mais direta com esses dados. Para tal, foram escolhidas imagens de satélite com medições de evapotranspiração, por conta de sua correlação com fatores como a cobertura vegetal e tipo de superfície do solo.

Os pontos georreferenciados no estudo são delimitados por arquivos **.shp** (*shapefile*) e **.dbf** *(dBase database file),* e os indicadores disponibilizados foram dispostos em tabelas de formato **.csv**, com cada valor relacionado à coordenada geográfica de seu *cluster* correspondente.

Figura 8: Exemplo de imagens de satélite diurnas utilizadas no modelo. Estas imagens foram extraídas da área Nexus.

Fonte: Compilação do Autor.

4.2.2 Extração de *features*

Nesta etapa, o algoritmo recebe o arquivo **.csv** dos indicadores e utiliza uma técnica de transferência de aprendizado para identificar características nas imagens baixadas, de forma a facilitar o treinamento do modelo na etapa seguinte.

Assim como no artigo de referência, primeiro se utiliza uma rede neural convolucional que foi previamente treinada no *ImageNet* [24], uma base de dados de classificação de imagens que é composta essencialmente de imagens rotuladas. Isso permite ao modelo identificar certas *features* das imagens, como por exemplo cantos e arestas.

Em seguida, fazendo-se uso da técnica de transferência de aprendizado (utilizando-se dados extraídos de imagens de luzes noturnas ou de imagens contendo informações de evapotranspiração), esse modelo pré-treinado é aproveitado como um extrator de *features* que recebe como entrada imagens de satélite diurnas. Essas *features* extraídas das imagens de satélite são vetores de representação que contêm informações relevantes sobre as imagens para se inferir os indicadores.

4.2.3 Treinamento do Modelo

O terceiro passo consiste no treinamento do modelo de regressão *Ridge*, também para cada um dos indicadores, alimentado com as *features* extraídas no passo anterior através de imagens de satélite diurnas. O treinamento do modelo é feito por *cluster*, com cada um sendo treinado individualmente com as imagens disponíveis no *dataset*, um processo que é repetido até que a melhor correlação seja encontrada. O modelo tem como saída os indicadores previstos, ou seja, os dados finais.

4.2.4 Validação

Finalmente, no último passo, realiza-se a validação dos modelos treinados para cada um dos indicadores. Para isso, é adotado o método de validação cruzada *k-fold*. A medição da acurácia (ou performance) do modelo é baseada na análise do coeficiente de determinação (R^2), que indica o grau de correlação entre os indicadores reais e o previstos.

5. Especificação de Requisitos do Projeto

Os requisitos para o projeto podem ser organizados em duas frentes diferentes: uma relacionada ao algoritmo de aprendizado de máquina treinado, e outra referente à plataforma online onde os resultados obtidos são disponibilizados.

5.1 Requisitos do Modelo de Aprendizado de Máquina

A principal métrica a ser avaliada nas predições do modelo de redes neurais treinado é o coeficiente de determinação (Equação 2.4.3.3.1) entre os valores dos indicadores previstos pelo algoritmo e o dos indicadores fornecidos pelo projeto NEXUS [5]. Para cada indicador analisado, o modelo será considerado bem-sucedido — ou seja, capaz de estimar os indicadores a partir de imagens de satélite — caso o coeficiente de determinação (R^2) seja maior que 0.37, ou seja, semelhante ou superior ao valor mínimo apresentado nos trabalhos de Triñanes et al. (2020) [12] e Jean et al. (2016) [3], que serviram como base metodológica para o desenvolvimento do algoritmo de predição. Caso contrário, será considerado que o indicador em questão não tem correlação com imagens de satélite, ou que não houveram dados suficientes para que o modelo fosse devidamente treinado.

5.2 Especificação da plataforma *web*

O objetivo da plataforma online é apresentar os resultados obtidos através dos modelos treinados para facilitar sua compreensão e uso por parte do usuário, permitindo com que os indicadores previstos pelo algoritmo possam ser comparados com os de referência fornecidos pelo projeto NEXUS [5]. Para tal, foram especificados os seguintes casos de uso para o sistema:

1. **Visualizar mapa com indicadores:** O usuário utiliza um mapa interativo com possibilidade de ampliar e deslocar a visão para visualizar os indicadores em diferentes municípios da área NEXUS.
2. **Filtrar indicadores no mapa:** O usuário seleciona qual indicador socioambiental deseja visualizar no mapa, dentre uma lista de indicadores disponíveis.

3. **Comparar indicadores reais com os previstos:** O usuário pode selecionar um ou mais municípios, e comparar os valores dos indicadores previstos pelo modelo com os valores reais através de gráficos e tabelas geradas dinamicamente.
4. **Fazer download dos indicadores gerados**: O usuário pode fazer o download dos indicadores previstos e reais para os municípios que selecionar, tanto no formato de gráficos de comparação, como em tabelas no formato **.csv .**

6. Projeto e Implementação

6.1. Reprodução do Artigo de Referência

Para facilitar o processo de adaptação do algoritmo, foi também utilizado como base um outro trabalho, Triñanes et al. (2020) [12], que buscou reproduzir e adaptar o artigo de referência (JEAN et al., 2016) [3] para analisar porções do território brasileiro, modificando o algoritmo para realizar a predição de indicador de renda per capita na região do Vale do Ribeira. Essa adaptação usa o mesmo código e metodologia do artigo de referência, mas adaptada a dados brasileiros.

6.1.1 Aquisição de Dados

A adaptação de Triñanes et al. (2020) [12] segue a mesma metodologia de Jean et al. (2016) [3] para realizar a aquisição dos dados necessários para o funcionamento do algoritmo, dividindo o processo em três etapas: (i) aquisição de dados censitários, (ii) aquisição de imagens de satélite diurnas e (iii) aquisição de imagens de satélite noturnas. Na reprodução dessa adaptação os mesmos passos foram seguidos.

Os dados censitários utilizados correspondem ao censo populacional do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2010 e foram extraídos em formato **.csv** da base de dados do órgão. Os clusters definidos para adaptação correspondiam às subdivisões de setores censitários definidas pelo IBGE para a região do Vale do Ribeira, e também possuíam, de forma semelhante ao artigo (JEAN et al., 2016), uma área de 10 km por 10 km.

Seguindo novamente a metodologia do artigo original, para cada cluster foram adquiridas um mínimo de 10 imagens, todas extraídas através do fornecimento das coordenadas geográficas do cluster à API da *Planet* [18]. No total 88 clusters foram adotados, com 8800 imagens baixadas.

6.1.2. Resultados da Reprodução

A reprodução realizada da adaptação de (TRIÑANES et al., 2020) alcançou os mesmos resultados obtidos pelo autor (Figura 9). Primeiramente, das 8800 imagens que deveriam ser baixadas apenas um total de 4.296 teve algum retorno da API, cerca de 48,82% do esperado, o mesmo resultado obtido por Triñanes et al. (2020). Além disso, 510 imagens das que foram baixadas se mostraram inadequadas para o algoritmo e tiveram de

ser descartadas, possuindo bordas cortadas e/ou uma porcentagem da imagem em branco (Figura 10). No fim apenas 3.786 (43,02%) imagens de satélite puderam ser efetivamente utilizadas. Apesar disso, a adaptação de Triñanes et al. (2020) ainda possui uma porcentagem de imagens obtidas 36,56% maior que a do artigo de referência (JEAN et al., 2016), com imagens com também melhor resolução.

Figura 9: Imagens de satélite (API de dados da Planet [18]) válidas extraídas para treinamento do modelo (Vale do Ribeira).

Fonte: TRIÑANES et al. (2020).

Figura 10: Imagens de satélite descartadas por erros.

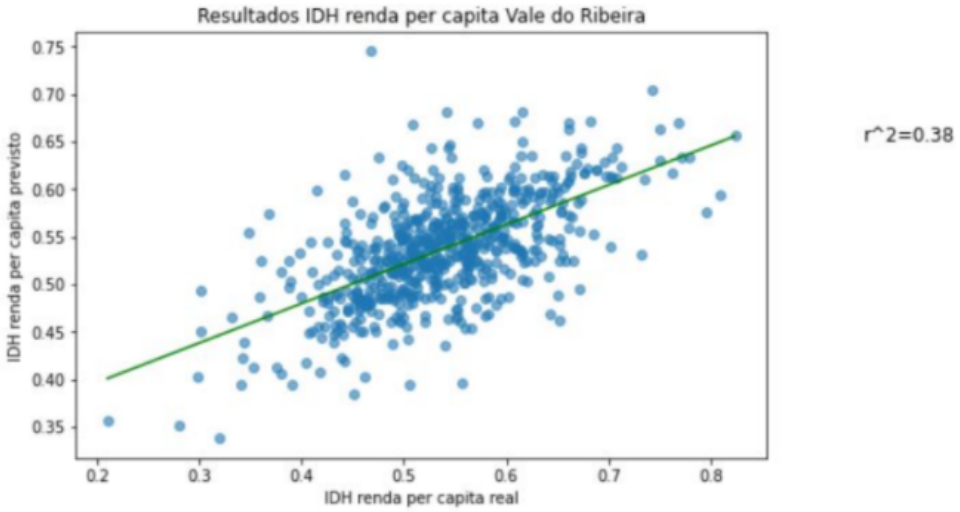
Fonte: TRIÑANES et al. (2020).

Figura 11: Imagem de luzes noturnas da região do Vale do Ribeira.

Fonte: TRIÑANES et al. (2020).

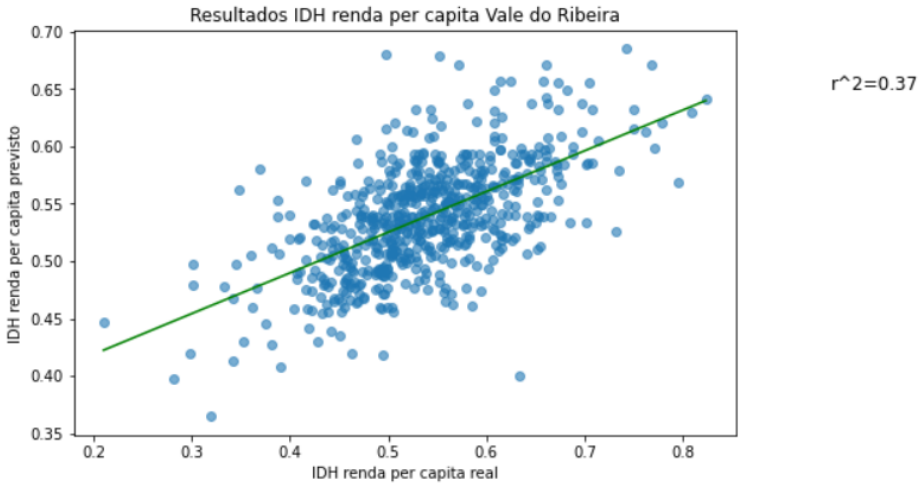
Após essa etapa, foi realizada a etapa de treinamento, com a filtragem de imagens por luzes noturnas (Figura 11) e finalmente o treinamento do modelo de aprendizado profundo. Na adaptação de Triñanes et al. (2020) foi utilizada a biblioteca de inteligência artificial para aprendizagem profunda *Torchvision* [28]. Por fim, foi realizada a validação do modelo utilizando a técnica de validação cruzada. A saída do algoritmo retorna um gráfico de comparação entre o indicador de renda previsto e o real e mapas de calor referentes ao indicador de renda previsto e ao real por setor censitário. Os resultados obtidos pela reprodução do algoritmo e os resultados originais obtidos pela adaptação se encontram nas Figuras seguintes (Figuras 12 e 13):

Figura 12: Resultados do algoritmo original (TRIÑANES et al., 2020).



Fonte: TRIÑANES et al., 2020.

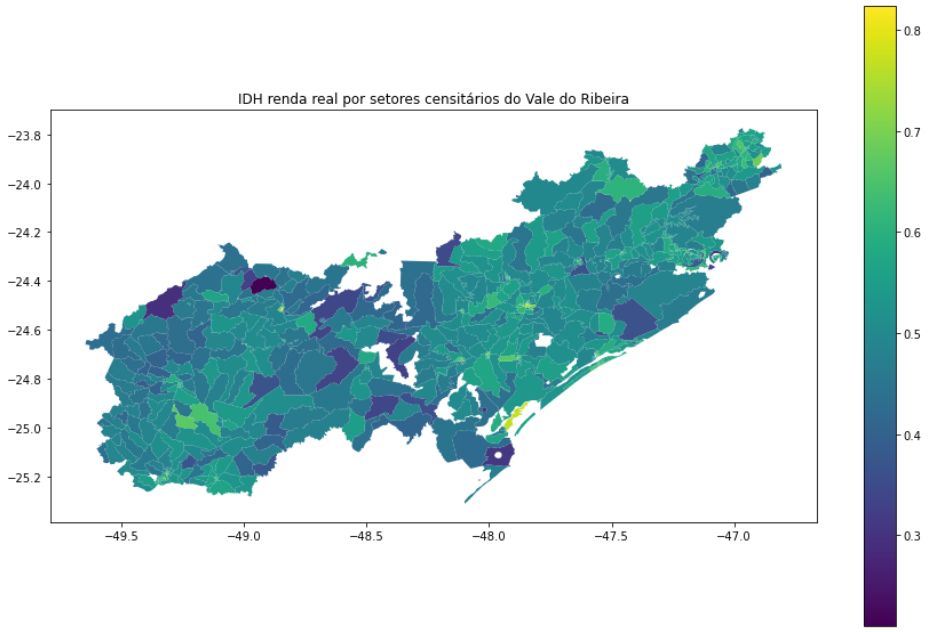
Figura 13: Resultados da reprodução do algoritmo original (TRIÑANES et al., 2020).



Fonte: Compilação do autor.

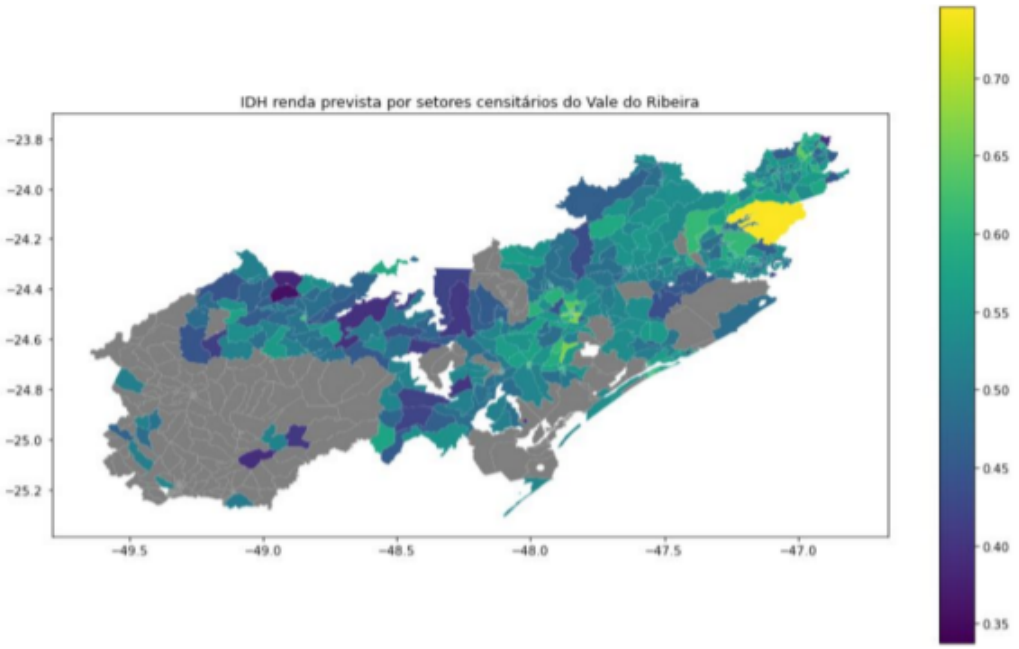
A reprodução do algoritmo alcançou resultados muito próximos aos obtidos por Triñanes et al. (2020), com uma taxa de correlação (R^2= 0,37), que é bem próxima do valor de (R^2=0,38) alcançado pelo artigo reproduzido (TRIÑANES et al., 2020). A mesma semelhança de resultados pode ser observada nos mapas de calor gerados (Figuras 15 e 16), com uma proximidade satisfatória entre ambos, demonstrando que a metodologia utilizada é consistente. Além disso, pode-se observar que há uma grande semelhança entre ambas as imagens de calor com IDH renda previsto e a imagem de calor com o IDH renda real (Figura 14), com as apenas acentuadas diferenças nas zonas em cinza por conta das imagens que não puderam ser baixadas.

Figura 14: Mapa de Calor com valor de IDH renda real por setor censitário do Vale do Ribeira.



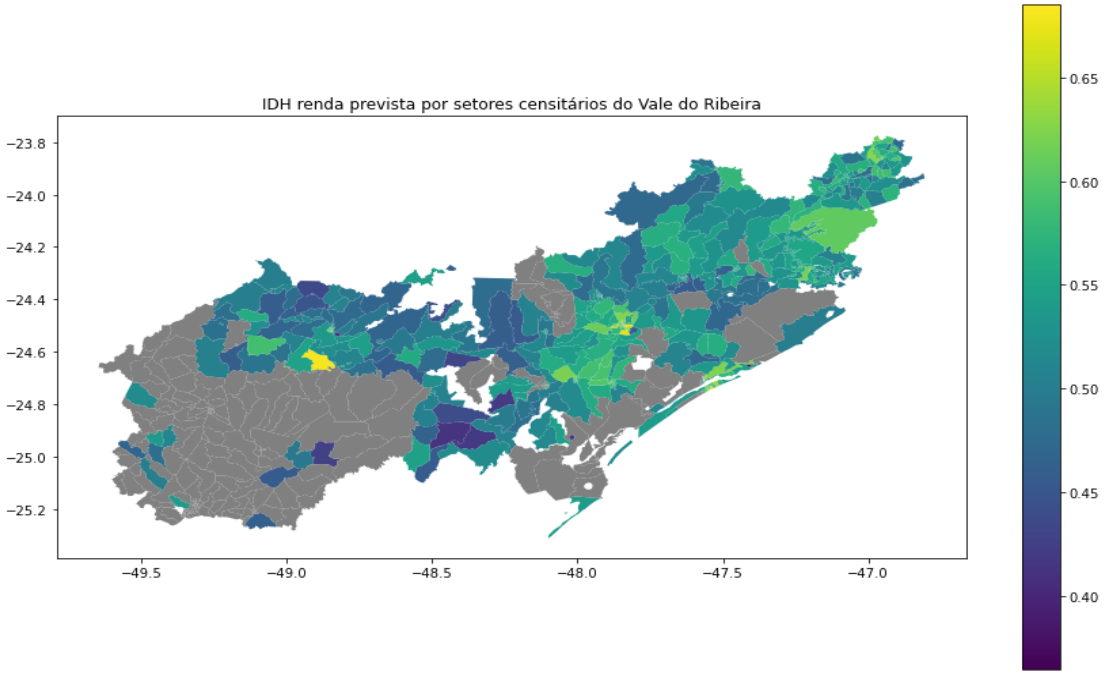
Fonte: Compilação do autor.

Figura 15: Mapa de calor com valor de IDH renda previsto pelo algoritmo de Triñanes et al. (2020).



Fonte: TRIÑANES et al., 2020.

Figura 16: Mapa de calor com valor de IDH renda previsto pela reprodução do algoritmo.



Fonte: Compilação do autor.

6.2. Aplicação do Algoritmo na Área de Estudo

Após a etapa de reprodução do experimento de Triñanes et al. (2020), aplicou-se a mesma metodologia para a Área Nexus, utilizando os indicadores socioeconômicos e ambientais fornecidos pelo projeto NEXUS-PARSEC [5-6].

6.2.1. Aquisição de Dados

Assim como na metodologia adotada pelo artigo de referência (JEAN et al., 2016) e por Triñanes et al. (2020), a etapa de aquisição de dados se divide na coleta de imagens de satélite e na coleta de indicadores correspondentes à zona estudada.

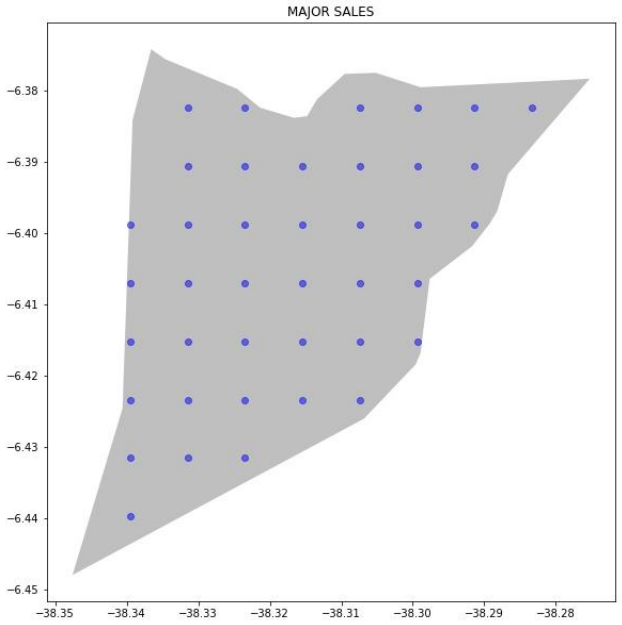
Inicialmente, a intenção era cobrir toda a Área Nexus, mas por limitação de recursos (custos para realizar a aquisição das imagens pelo *Google Static Maps API*), não foi possível adquirir todas as imagens. Dessa forma, alterou-se o escopo de forma a se trabalhar com os quatro estados nordestinos de Alagoas (AL), Paraíba (PB), Rio Grande do Norte (RN) e Sergipe (SE), que fazem parte da Caatinga, um bioma menor do que o Cerrado e que desta forma pode ser analisado utilizando um número menor de imagens.

Esses quatro estados nordestinos pertencentes à Área Nexus foram divididos em *clusters*. Como a granularidade dos indicadores fornecidos pelo projeto NEXUS-PARSEC é de nível municipal, cada *cluster* também é correspondente a um município, totalizando em 567 *clusters* na região de estudo.

6.2.1.1. Aquisição de Imagens de Satélite Diurnas

As imagens de satélite diurnas utilizadas no modelo foram obtidas através do *Static Maps API* do *Google* [28]. Para cada município, foi obtido a latitude e longitude mínimas e máximas, montando-se uma grade de pontos ao redor do contorno do município, com cada ponto representando o centro de uma imagem com dimensões fixas de 400 x 400 pixels e nível de *zoom* 16, o que resulta em imagens de 1 km² aproximadamente. As coordenadas dos pontos centrais das imagens foram calculadas de tal forma que ficassem dispostas como num quebra cabeça, uma imagem encaixada na outra. Após esse processo, os pontos que estavam fora da área do município foram retirados. A ideia é que com isso se obtivessem as coordenadas centrais das imagens cobrindo praticamente toda a área interna de cada município da região de estudo. A Figura 17 exemplifica tal procedimento, representando as coordenadas centrais das imagens de satélite que foram adquiridas para o município de Major Sales, pertencente ao estado de RN.

Figura 17: Exemplo do procedimento de definição das coordenadas centrais de imagens de satélite diurnas que foram adquiridas para o município de Major Sales. Os pontos em azul simbolizam esses pontos centrais. O eixo y do mapa indica a latitude e o eixo x, a longitude.



Fonte: Compilação do autor.

Vale comentar que as imagens de satélite adquiridas a partir da *Google Static Maps API* são atemporais — isso porque não são fornecidas informações sobre a data da imagem disponibilizada pela plataforma. As imagens fornecidas por esta API são imagens em mosaico, ou seja, são compostas por várias imagens brutas de um mesmo local agregadas.

Isso permite que se obtenha uma imagem de boa qualidade visual. No entanto, as informações temporais podem ser perdidas, pois não é possível associar a imagem a uma data específica.

6.2.1.2. Aquisição de Imagens de Satélite com dados georreferenciados

Para a aquisição de imagens do tipo GeoTIFF com dados georreferenciados, foi utilizado o *Google Earth Engine*. Dois tipos de imagens foram adquiridas:

- **Intensidade de luzes noturnas**, cujo dataset (denominado *DMSP OLS: Nighttime Lights Time Series Version 4, Defense Meteorological Program Operational Linescan System* [30]) foi disponibilizado na plataforma pelo *Earth Observation Group* da Escola de Minas do Colorado. Desde a década de 1970, o Programa Meteorológico de Defesa por Satélite (*Defense Meteorological Satellite Program* ou DMSP) da Força Aérea dos Estados Unidos tem satélites implantados equipados com sensores do Sistema Operacional *Linescan*. O processamento de dados do DMSP implica na remoção de observações distorcidas por obstrução de nuvens, luz da lua, pores-do-sol sazonalmente tardios e auroras. Para cada célula de 30 segundos de arco, todas as observações restantes sobre o ano são calculadas como média e depois convertidas para um "valor digital" inteiro entre 0 (sem iluminação) e 63 (representando a luminosidade de código superior) para produzir um conjunto de dados do ano do satélite em grade.
- **Evapotranspiração**, cujo dataset (denominado *MOD16A2: MODIS Global Terrestrial Evapotranspiration 8-Day Global 1km* [31]) foi disponibilizado na plataforma pelo *Numerical Terradynamic Simulation Group* (NTSG) da Universidade de Montana. O *dataset* fornece informações sobre a evapotranspiração terrestre global de 8 dias com resolução de 1 km pixel. Evapotranspiração (com unidade de medida *kg/m^2*) é a soma da evaporação e da transpiração da planta da superfície terrestre para a atmosfera. Com dados de evapotranspiração de longo prazo, os efeitos das mudanças no clima, uso da terra e distúrbios nos ecossistemas podem ser quantificados.

Essas imagens foram utilizadas para treinar dois modelos de CNN distintos para cada um dos tipos de imagens, através da estratégia de Transferência de Aprendizado.

6.2.1.3. Aquisição e Dados Socioeconômicos e Ambientais

Os indicadores utilizados neste projeto foram coletados a partir de estudos realizados pelo Instituto Nacional de Pesquisas Espaciais (INPE). Através do projeto NEXUS-PARSEC [5-6], mais de 130 indicadores foram identificados na Área Nexus, divididos em 7 categorias como conservação florestal, biodiversidade, degradação da terra, energia, produção agrícola, recursos hídricos e riscos climáticos e socioeconômicos.

Dentre estes indicadores, foram selecionados 9 indicadores socioeconômicos e 11 indicadores ambientais de acordo com a disponibilidade dos dados para os estados nos quais as imagens de satélite foram coletadas, para assim serem utilizados como variáveis de resposta dos modelos de regressão. Foram considerados como indicadores ambientais aqueles pertencentes às categorias: conservação florestal, biodiversidade, degradação da terra, produção agrícola e recursos hídricos. Todos os indicadores selecionados foram coletados no ano de 2015.

Os 9 indicadores socioeconômicos selecionados são:

1. *Ocorrência de doenças veiculadas com fonte hídrica*;
2. *Domicílios com renda maior que um salário mínimo*;
3. *Isolamento da população considerando a distância a corpos hídricos e estradas*;
4. *Taxa de mortalidade em menores de 5 anos de idade*;
5. *Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)*;
6. *Proporção de Cadastramento de pessoas em serviços básicos de saúde*;
7. *PIB per capita*;
8. *Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população*;
9. *Domicílios Inadequados*.

Os 11 indicadores ambientais selecionados são:

1. *Evolução dos Sistemas Agroflorestais em estabelecimentos rurais*;
2. *Pastagens degradadas*;
3. *Produtividade agrícola de alimentos básicos*;
4. *Proporção do uso de agrotóxico*;
5. *Abrangência e Diversidade do PRONAF (Programa Nacional de Fortalecimento da Agricultura Familiar)*;
6. *Uniformidade de receitas ou atividades do produtor rural*;
7. *Existência e proteção de recursos hídricos nos estabelecimentos agropecuários*;

8. *Ineficácia de maquinários à produtividade agrícola*;
9. *Abrangência do Programa Cisternas*;
10. *Produtividade Pecuária e Leiteira*;
11. *Alternativas ao abastecimento de água*.

A descrição completa de cada indicador selecionado encontra-se nos anexos A e B.

Todos os indicadores foram normalizados através do método de Normalização Max-Min (Equação 2.7.1.1) antes de serem submetidos aos modelos de regressão, assumindo valores entre 0 e 1. Os valores máximo e mínimo utilizados para realizar a normalização dos indicadores foram calculados considerando-se os dados de todos os municípios da Área Nexus.

6.2.2. Extração de *Features*

Para a extração do vetor de características das imagens de satélite diurnas provenientes do *Static Maps API* da *Google*, foi utilizada a rede neural convolucional VGG-11 (SIMONYAN; ZISSERMAN, 2015) [32], um modelo com 11 camadas previamente treinado no grande conjunto de dados do *ImageNet* [24].

A técnica de transferência de aprendizado foi aplicado através da realização de ajustes finos de 2 modelos CNN distintos, separadamente: (i) usando as imagens de intensidade de luzes noturnas como *proxy* dos indicadores socioeconômicos e (ii) usando as imagens de evapotranspiração como *proxy* dos indicadores ambientais. Para ambos os modelos, 80% das imagens foram separadas para treino e 20% para teste.

Como descrito nas Seções 6.2.1.1 e 6.2.2.2, para cada município foi adquirida uma grade de imagens (cujo número total varia de acordo com o tamanho do município - quanto maior o seu tamanho, o *cluster* tem mais imagens também). Essas imagens foram submetidas aos modelos de CNN extratores de características, e depois os vetores de características das imagens de um mesmo município foram reduzidos em um único vetor (contendo a média dos valores) com 4096 características.

6.2.2.1. Imagens de intensidade de luzes noturnas como *proxy* de indicadores socioeconômicos

Um dos CNNs foi ajustado para conseguir estimar a intensidade de luzes noturnas a partir de uma imagem de satélite diurna, assim como foi feito no artigo de referência. Esta etapa da abordagem de transferência de aprendizado trata de um problema de classificação, em que o CNN aprende a classificar as imagens de satélite diurnas em uma das três classes de intensidade de luzes noturnas separadas em: intensidades baixa (com

intensidades de 0 a 5), média (com intensidades de 5 a 14) e alta (com intensidades de 14 a 65). Essas classes foram obtidas através do ajuste das frequências relativas dos valores de intensidade noturna em um Modelo de Mistura com três distribuições Gaussianas. Depois de treinar o modelo CNN para prever a intensidade luminosa noturna, utilizamos este modelo aprendido como extrator de características das imagens de satélite diurnas, descartando a última camada do modelo, que é a camada de classificação de intensidade de luzes noturnas. Este modelo foi gerado com o intuito de extrair características que fossem relevantes para a estimativa de indicadores socioeconômicos.

6.2.2.2. Imagens de evapotranspiração como *proxy* de indicadores ambientais

Um segundo CNN foi ajustado para conseguir estimar a evapotranspiração a partir de uma imagem de satélite diurna. Da mesma forma que ocorreu com o primeiro CNN, o segundo CNN aprende a classificar as imagens de satélite diurnas em uma das três classes de evapotranspiração, também separadas em nível baixo (com 0 a 60 kg/m²/8 dias), médio (com 60 a 130 kg/m²/8 dias) e alto (com 130 a 520 kg/m²/8 dias) e essas classes foram obtidas através do ajuste das frequências relativas dos valores de evapotranspiração em um Modelo de Mistura com três distribuições Gaussianas. Este segundo modelo foi utilizado com o intuito de se extrair características que fossem relevantes para a estimativa de indicadores ambientais.

6.2.3. Treinamento e Validação do Modelo

O vetor de características extraídas das imagens de satélite no passo anterior são as variáveis preditoras do modelo e os indicadores fornecidos pelo projeto NEXUS-PARSEC são as variáveis resposta. Tendo isso em vista, foram realizados três tipos de experimentos (Seções 6.2.3.1, 6.2.3.2 e 6.2.3.3) para a avaliação dos modelos de regressão de Ridge que foram ajustados para cada um dos indicadores socioeconômicos e ambientais. Antes de ajustar o modelo, as variáveis preditoras foram normalizadas pelo método *Z-Score*, dado pela Equação (2.6.2.1).

6.2.3.1. Validação cruzada totalmente randômica

Este experimento foi aplicado simultaneamente nos quatros estados (Alagoas, Paraíba, Rio Grande do Norte e Sergipe) que compõem a área de estudo deste trabalho

A validação cruzada de *5-folds* foi aplicada em um conjunto com os dados de todos os clusters dos quatro estados agregados. O coeficiente de determinação (R^2) final obtido é o valor médio dos coeficientes de determinação que foram calculados ao longo dos *folds* da

validação cruzada. Em cada *fold*, a escolha do melhor parâmetro de ajuste (λ) foi feito através de um loop interno de validação cruzada *5-fold*.

6.2.3.2. Validação cruzada randomizada por estado

Este experimento foi aplicado separadamente em cada um dos quatros estados (Alagoas, Paraíba, Rio Grande do Norte e Sergipe).

A validação cruzada de *5-folds* foi aplicada para o cálculo do R^2 em cada *fold*, só que neste caso a validação foi aplicada separadamente para cada um dos estados. Da mesma forma que no experimento anterior, o coeficiente de determinação (R^2) final obtido é o valor médio dos coeficientes de determinação que foram calculados ao longo dos folds da validação cruzada e a escolha do melhor parâmetro de ajuste (λ) foi feito em cada fold um *loop* interno de validação cruzada *5-fold*. Este experimento foi repetido 4 vezes em 4 estados diferentes para cada indicador.

6.2.3.3. Validação transregional

Neste terceiro tipo de experimento, foi separado um estado para teste, e os demais três estados para o treinamento, validação e a escolha do melhor parâmetro de ajuste (λ) para o modelo. Tendo este modelo ajustado com λ, o cálculo do R^2 foi feito submetendo o estado de teste neste modelo. Foram feitos 4 experimentos deste tipo para cada indicador, e em cada experimento foi separado um estado diferente para teste (e os demais para a validação e treinamento do modelo).

6.3. Desenvolvimento da plataforma *web*

A implementação da plataforma *web* foi realizada em 4 etapas: (i) importação dos dados, (ii) consolidação dos dados, (iii) desenvolvimento dos mapas interativos, e (iv) a implementação do *website* para a divulgação dos resultados.

6.3.1 Importação dos dados

Os dados obtidos com o modelo treinado na etapa anterior foram importados no ambiente de desenvolvimento integrado (IDE) online da plataforma do *Earth Engine* em tabelas no formato **.csv**, contendo as seguintes colunas:

- **Código Município:** Código do IBGE para o respectivo município;
- **Nome:** Nome do município, totalmente capitalizado e sem acentos, para comparação com o *dataset* de língua inglesa;

- **NomeDisplay:** Nome do município com acentos e capitalização usual, para ser mostrado na interface do usuário;
- **Indicador Real:** Valor real do indicador para aquele município;
- **Indicador Previsto:** Valor do indicador previsto pelo modelo de rede neural para o município.

Cada tabela contém os dados referentes a um indicador, para os municípios de um estado. Foram selecionados dois indicadores com os melhores resultados de coeficiente de determinação: um socioeconômico (Isolamento da população considerando a distância a corpos hídricos e estradas) e um ambiental (Pastagens degradadas) — ambos obtidos no experimento de validação cruzada randomizada por estado.

Ao todo, foram importadas 8 tabelas, correspondendo aos 4 estados estudados e aos dois indicadores selecionados. A importação dos dados foi realizada de maneira compartimentada para que o desenvolvimento dos mapas pudesse ser realizado tão logo os novos resultados fossem obtidos do modelo, sem a necessidade de esperar com que todos os experimentos ficassem prontos de antemão.

6.3.2 Consolidação dos dados

Após a importação, as tabelas são automaticamente transformadas pelo *Earth Engine* em *FeatureCollections* [33], um conjunto de objetos (*Features*) que representam cada um dos dados ingeridos, permitindo com que sejam filtrados, ordenados e renderizados nos mapas.

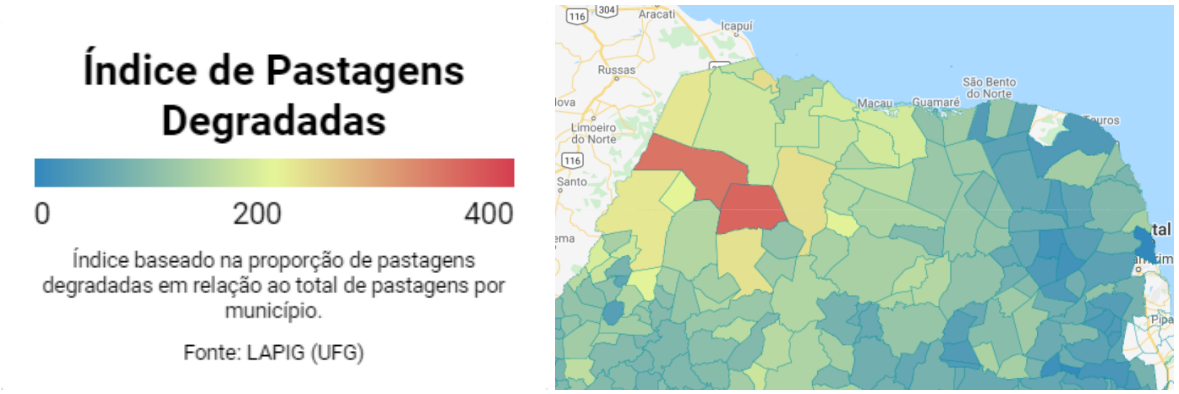
No primeiro passo da consolidação, um *script* utiliza as *FeatureCollections* dos dados importados, juntamente com o *dataset* de Unidades Administrativas Globais (*The Global Administrative Unit Layers — GAUL*) [34] fornecido pela Unidade Cartográfica das Nações Unidas (*UN Cartographic Unit — UNCS*), para adicionar a cada município as informações de geometria espacial que configuram suas fronteiras. Como não há identificação comum entre os códigos da UNCS e os do IBGE, essa associação é feita através do nome do município (sem acentos, e em letras maiúsculas). Nessa etapa também são removidos os municípios que não contém indicadores.

Em seguida, as *FeatureCollections* geradas no passo anterior são consolidadas em uma só *FeatureCollection* contendo os dados de todos os municípios estudados para cada um dos indicadores, que é exportada para ser usada posteriormente nos mapas interativos.

6.3.3 Desenvolvimento dos mapas interativos

Com os dados consolidados, a *FeatureCollection* de cada indicador é então importada em um terceiro *script*, para que seja usada como base para criar o mapa de calor dos indicadores. São criadas duas camadas para cada indicador, uma com os valores do indicador real, e outra com os valores previstos pelo modelo de rede neural. Cada município é então pintado de acordo com o valor do indicador, baseado em uma escala de cores que parte do menor valor do indicador para todo o *dataset* (em azul), e vai até o maior valor (em vermelho), como mostrado na Figura 18.

Figura 18: Escala de cor e mapa de calor para o indicador Índice de Pastagens Degradadas.



Fonte: Compilação do autor.

O *dataset GAUL* é novamente usado para desenhar as fronteiras em uma camada acima do mapa de calor, de maneira a deixar clara a distinção entre os municípios.

As demais funcionalidades do mapa foram baseadas nos mapas de código aberto disponibilizados na página do *Earth Engine Apps* [26], além das listadas na especificação de requisitos do projeto. Foram implementados:

- Lista no formato *combobox* para filtrar os indicadores mostrados no mapa;
- *Slider* de opacidade, permitindo deixar as cores do mapa de calor mais transparentes para melhor visualização do mapa base abaixo;
- Seleção de municípios ao clicar no mapa, gerando um gráfico e uma tabela de comparação com os dados do indicador real e do previsto para cada município selecionado;
- Exportar os gráficos e tabelas geradas para os municípios selecionados (formatos **.csv**, **.svg** ou **.png**)

6.3.4 Desenvolvimento do *website*

O *website* do trabalho foi desenvolvido com a ferramenta *Google Sites*, com a incorporação dos mapas interativos feitos no *Earth Engine Apps*. As páginas da plataforma foram organizadas da seguinte forma:

- *Plataforma*: A página inicial e principal do site, onde encontra-se o mapa interativo com os resultados obtidos pelo trabalho, bem como uma breve explicação sobre cada um dos indicadores contemplados;
- *O Projeto*: Página com informações sobre o objetivo e motivação do projeto;
- *Metodologia*: Página com uma breve explicação sobre a metodologia usada no trabalho;
- *Download*: Página com referências para o download dos resultados obtidos, bem como a monografia do trabalho.

7. Testes e Avaliação

7.1. Aplicação do Algoritmo na Área de Estudo

7.1.1. Aquisição de Dados

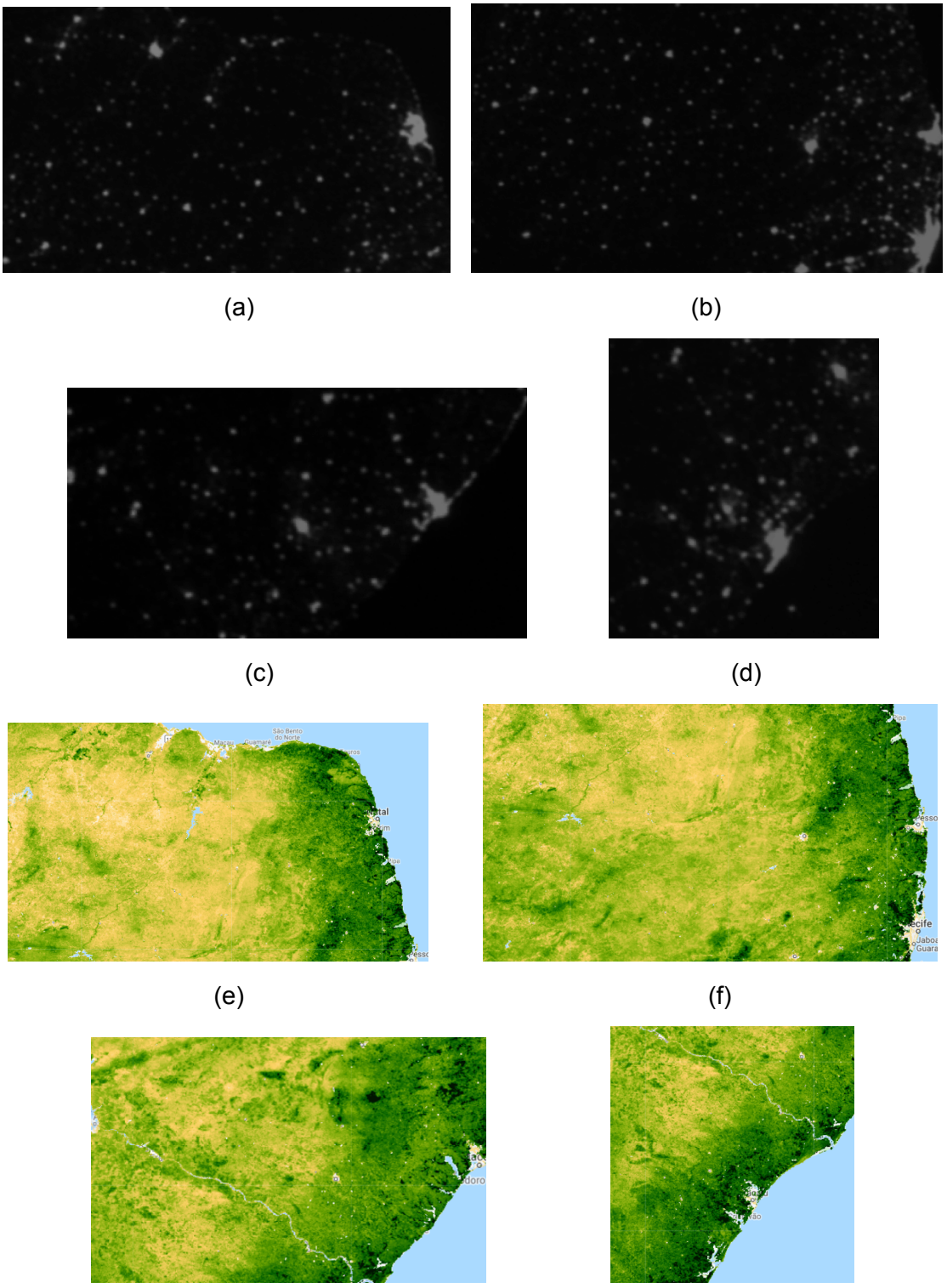
Quanto às imagens de satélite diurnas, foram adquiridas 65.030 imagens para o estado de PB, 60.180 para o estado de RN, 32.431 para o estado de AL e 15.150 para o estado de SE, com um total de 172.791 imagens obtidas a partir do *Google Static Maps API*. A Figura 19 exemplifica algumas das imagens adquiridas.

Figura 19: Imagens de satélite diurnas adquiridas do *Google Static Maps API* para o estado do Rio Grande do Norte.

Fonte: Compilação do autor.

Quanto às imagens com dados georreferenciados, para cada um dos quatro estados da área de estudo foram adquiridas quatro imagens com dados de intensidade de luzes noturnas e outras quatro com dados de evapotranspiração (Figura 20).

Figura 20: Imagens de satélite de intensidade de luzes noturnas adquiridas para os estados do Rio Grande do Norte (a), Paraíba (b), Alagoas (c) e Sergipe (d) e imagens com dados de evapotranspiração para o estado de Rio Grande do Norte (e), Paraíba (f), Alagoas (g) e Sergipe (h).



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fonte: Compilação do autor.

Os indicadores socioeconômicos e ambientais fornecidos pelo projeto NEXUS-PARSEC [5-6] em formato **.dbf** foram processados descartando-se indicadores que não fossem de 2015 (para trabalhar somente com indicadores do mesmo ano) e também aqueles que possuíssem algum indicador com valores não numéricos ou inválidos. Feito isso, os indicadores disponibilizados foram dispostos em tabelas **.csv**, relacionando-se cada valor do indicador às coordenadas geográficas centrais de latitude e longitude de seu *cluster* correspondente. Vale comentar que os indicadores de 2015 foram selecionados em detrimento de outros anos porque, entre os aproximadamente 130 indicadores disponibilizados no dataset, mais de 100 deles eram indicadores de 2015. Como já foi dito no capítulo anterior, todos os indicadores foram normalizados através do método de Normalização Max-Min (Equação 2.6.1.1) antes de serem submetidos aos modelos de predição, assumindo valores entre 0 e 1 - os valores máximo e mínimo utilizados para realizar a normalização foram calculados com base em todos os dados referentes a todos os municípios da Área Nexus e não só os dados dos municípios que foram usados para treinar o modelo.

7.1.2. Extração de Features

Nesta etapa, conforme foi descrito na seção 6.2.2, foram ajustados dois modelos distintos de CNN através de técnicas de transferência de aprendizado: um com dados georreferenciados de intensidade de luzes noturnas e outra com dados georreferenciados de evapotranspiração.

Para a execução do código de treinamento dos modelos CNN, foram usados GPUs do tipo *Nvidia P1000* disponibilizados no serviço Google Colab [35].

O primeiro modelo foi treinado e validado com um conjunto de mais de 150 mil imagens de satélite diurnas anotadas com dados de intensidade de luzes noturnas. Isso porque dentre as aproximadamente 170 mil imagens que foram baixadas, cerca de 20 mil imagens foram descartadas por não terem dados de intensidade de luzes noturnas correspondentes disponíveis. Dentre as 150 mil imagens restantes, cerca de 120 mil imagens foram separadas aleatoriamente para treino e outras 30 mil para a validação. O ajuste do modelo levou cerca de 400 minutos e a acurácia final medida para a tarefa de classificação foi de (R^2=0,7288). Ao submeter as imagens de validação no modelo CNN sem a última camada de classificação, obteve-se uma matriz de características de dimensões 30767 x 4096, sendo 30.767 o número de imagens de validação e 4.096 o número de características geradas para cada uma delas. Para cada município foi calculado

um vetor de características médio (Seção 6.2.2.). Dessa forma, a matriz de características teve suas dimensões reduzidas para 567 x 4096, sendo 567 o número de municípios.

Da mesma forma que, o segundo modelo foi treinado e validado com um conjunto de aproximadamente 150 mil imagens de satélite diurnas anotadas agora com dados de evapotranspiração. Também neste caso, dentre as cerca de 170 mil imagens que foram baixadas, aproximadamente 20 mil imagens foram descartadas por não terem dados de evapotranspiração correspondentes disponíveis. Logo, das 150 mil imagens restantes, cerca de 120 mil imagens foram separadas aleatoriamente para treino e outras 30 mil para a validação. O ajuste do modelo levou cerca de 328 minutos e a acurácia final medida para a tarefa de classificação foi de (R^2=0,7732). Ao submeter as imagens de validação no modelo CNN sem a última camada de classificação, obteve-se uma matriz de características de dimensões 29931 x 4096, sendo 29.931 o número de imagens de validação e 4.096 o número de características geradas para cada uma delas. Para cada município foi calculado um vetor de características médio (Seção 6.2.2). Dessa forma, a matriz de características teve suas dimensões reduzidas para 567 x 4096, sendo 567 o número de municípios.

7.1.3. Treinamento e Validação do Modelo

Nesta seção apresentam-se os resultados obtidos pelos modelos de regressão linear. Foram estudados os 9 indicadores socioeconômicos e 11 indicadores ambientais para os quais foram gerados os modelos. Vale ressaltar que, para treinar modelos de regressão dos indicadores socioeconômicos, foi usado o vetor de características das imagens extraídas pelo modelo CNN treinado com imagens de luzes noturnas (Seção 6.2.2.1). Para os modelos de regressão dos indicadores ambientais, foi utilizado o vetor de características das imagens extraídas pelo modelo de CNN treinado com dados de evapotranspiração (Seção 6.2.2.2). Os valores dos vetores de características (variáveis preditoras) foram normalizados pelo método *Z-Score* (Equação 2.6.2.1). Para cada modelo de regressão, foram calculadas métricas como o coeficiente de determinação (R^2, com Equação 2.4.3.3.1), erro absoluto médio (MAE, com Equação 2.4.3.1.1) e a raiz do erro quadrático médio (RMSE, com Equação 2.4.3.2.1). Os resíduos (ou valores residuais) dos modelos de regressão foram calculados com base na Equação (2.4.1.4).

Nas Seções 7.1.3.1 a 7.1.3.3 serão mostrados os resultados usando gráficos de dispersão que comparam os valores reais e os valores previstos para cada indicador socioeconômico e ambiental, assim como mapas de calor mostrando os mapas das regiões estudadas seguindo uma ordem particular: primeiro os valores previstos pelo modelo, logo depois os valores reais do indicador, e por último o mapa dos resíduos do modelo de

regressão, ou seja o valor absoluto do valor real subtraído ao valor previsto, de tal forma a mostrar o desempenho do modelo (Equação 2.4.1.4).

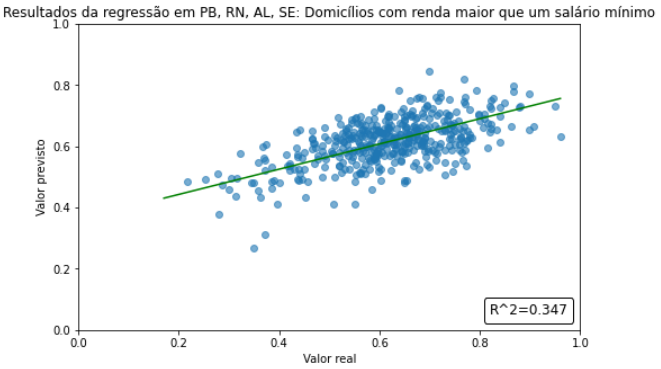
7.1.3.1. Validação cruzada totalmente randômica

Neste experimento, o método da validação cruzada foi usado para estimar o R^2 dos modelos de regressão Ridge treinados com os dados de todos os *clusters* dos estados de RN, PB, AL e SE. Vale ressaltar que o coeficiente de determinação (R^2) final obtido é o valor médio dos coeficientes de determinação que foram calculados ao longo dos *folds* da validação cruzada.

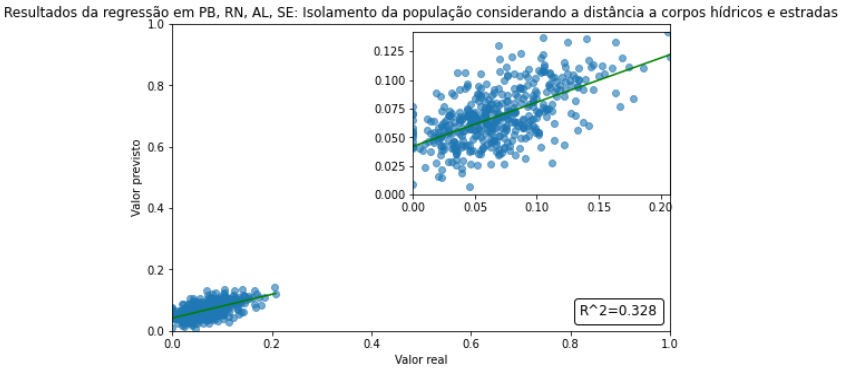
Para os indicadores socioeconômicos, os melhores resultados do R^2 foram o **R^2=0.347** (Figura 21.a) para o indicador *Domicílios com renda maior que um salário mínimo* e o **R^2=0.328** (Figura 21.b) para o indicador *Isolamento da população considerando a distância a corpos hídricos e estradas*. Para os demais indicadores, o R^2 foi menor que 0.3, atingindo em piores casos valores negativos - como no caso do indicador *Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população* que apresentou **R^2=-63.076** (Figura 21.c). A Figura 22 apresenta os mapas de calor correspondentes ao modelo de regressão com o melhor R^2 para este experimento (obtido para o indicador *Domicílios com renda maior que um salário mínimo)*, nos quais são representados os valores reais do indicador, os valores previstos pelo modelo e os resíduos calculados para cada um dos municípios. Já a Figura 23 apresenta os mapas de calor correspondentes ao modelo de regressão com o pior R^2 (obtido para o indicador *Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população*). Os resultados de R^2 correspondentes aos demais indicadores socioeconômicos encontram-se no Apêndice A.

Já no caso dos experimentos com indicadores ambientais, os melhores R^2 obtidos superaram 0.5, com **R^2=0.553** para o indicador *Pastagens degradadas* (Figura 24.a) e **R^2=0.512** para o indicador *Existência e proteção de recursos hídricos nos estabelecimentos agropecuários* (Figura 24.b). Para os demais indicadores, o R^2 foi menor que 0.3. O pior R^2 foi obtido para o indicador *Produtividade agrícola de alimentos básicos*, com **R^2=-36.293** (Figura 24.c). A Figura 25 apresenta os mapas de calor correspondentes ao modelo de regressão com o melhor R^2 para este experimento (obtido para o indicador *Pastagens degradadas)* e a Figura 26 apresenta os mapas de calor correspondentes ao modelo de regressão com o pior R^2 (obtido para o indicador *Produtividade agrícola de alimentos básicos*). Os resultados de R^2 correspondentes aos demais indicadores ambientais encontram-se no Apêndice B.

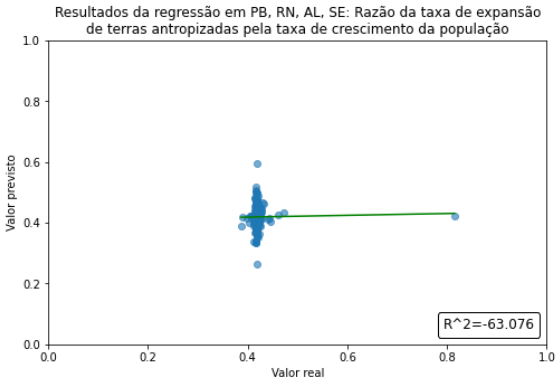
Figura 21: Gráficos de dispersão comparando os valores reais dos indicadores e os valores previstos pelo modelo de regressão gerados para o experimento de validação cruzada totalmente randômica realizado com os **indicadores socioeconômicos**. Os gráficos (a) e (b) representam os dois melhores resultados de R^2 para este experimento. O gráfico (c) representa o pior resultado de R^2.



(a)



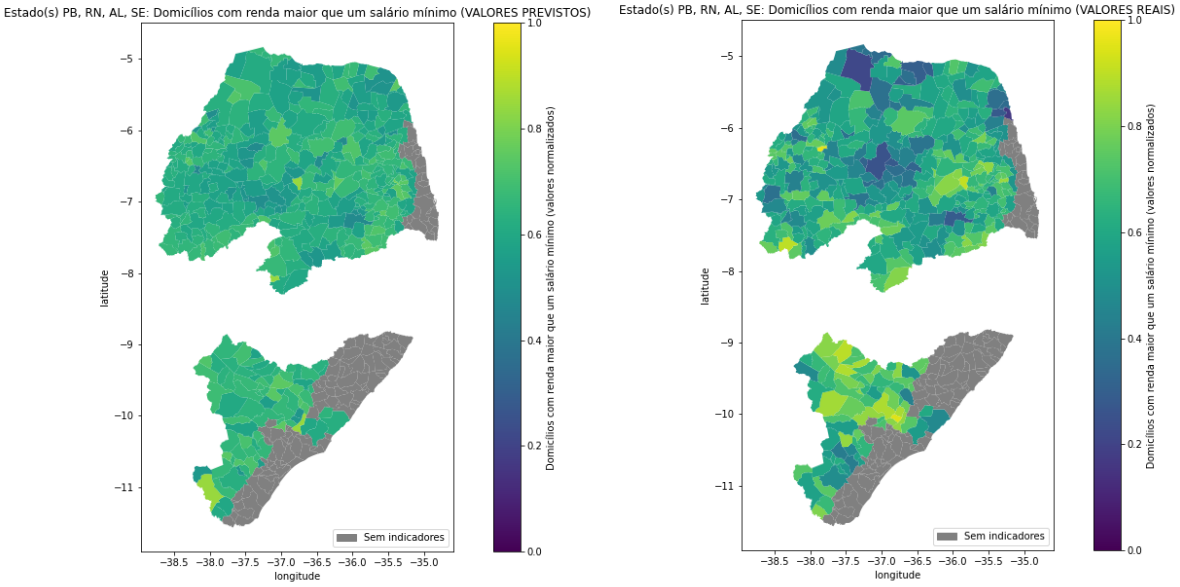
(b)



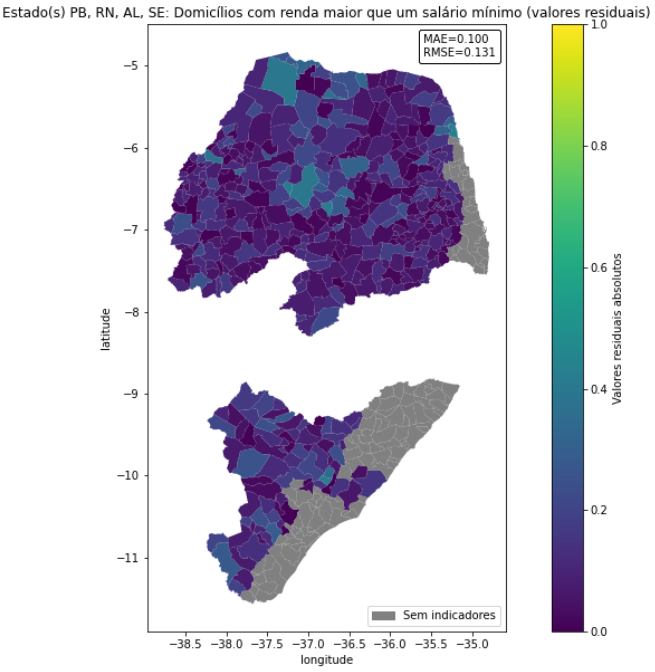
(c)

Fonte: Compilação do autor.

Figura 22: Mapas de calor gerados no experimento de validação cruzada totalmente randômica realizado com o indicador socioeconômico *Domicílio com renda maior que um salário mínimo*, para o qual foi obtido o melhor resultado de R^2 para este experimento **(R^2=0.347)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.



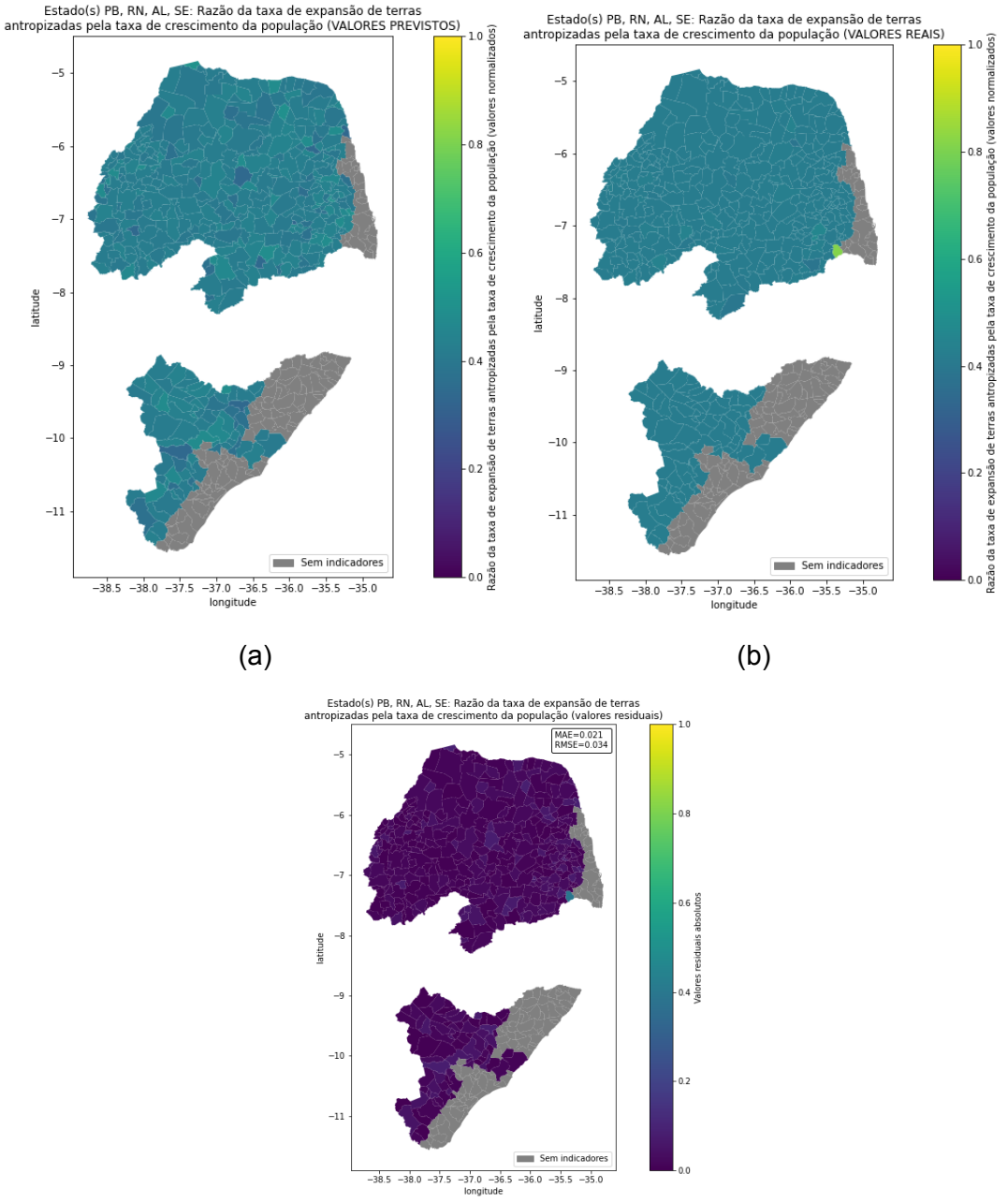
(a) (b)



(c)

Fonte: Compilação do autor.

Figura 23: Mapas de calor gerado no experimento de validação cruzada totalmente randômica realizado com o indicador socioeconômico *Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população*, para o qual foi obtido o pior resultado de R^2 para este experimento **(R^2=-63.076)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.

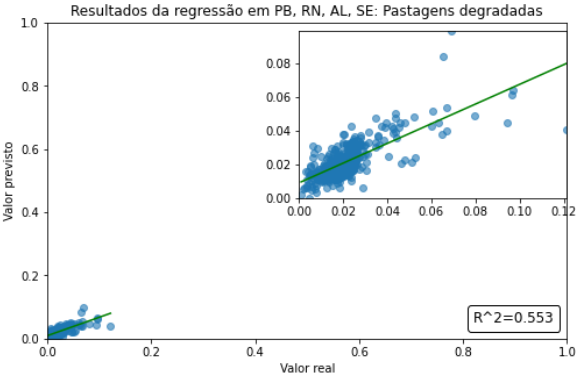


(a) (b)

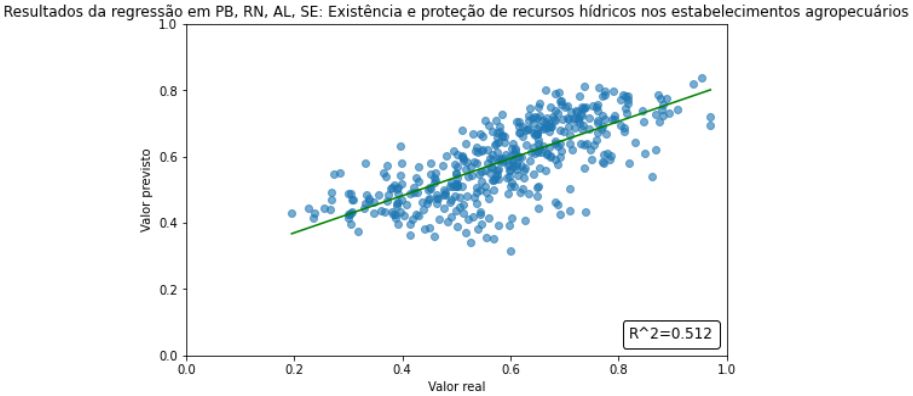
(c)

Fonte: Compilação do autor.

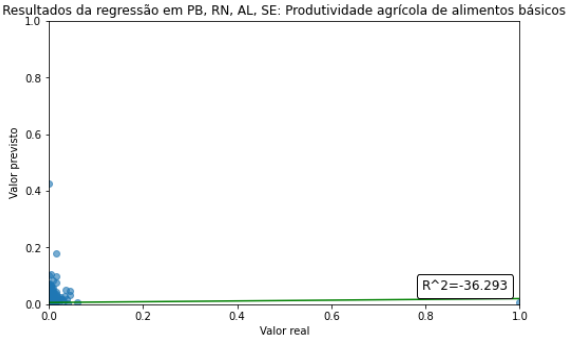
Figura 24: Gráfico de dispersão comparando os valores reais dos indicadores e os valores previstos pelo modelo de regressão, para o experimento de validação cruzada totalmente randômica realizado com os **indicadores ambientais**. Os gráficos (a) e (b) representam os dois melhores resultados de R^2 para este experimento. O gráfico (c) representa o pior resultado de R^2.



(a)



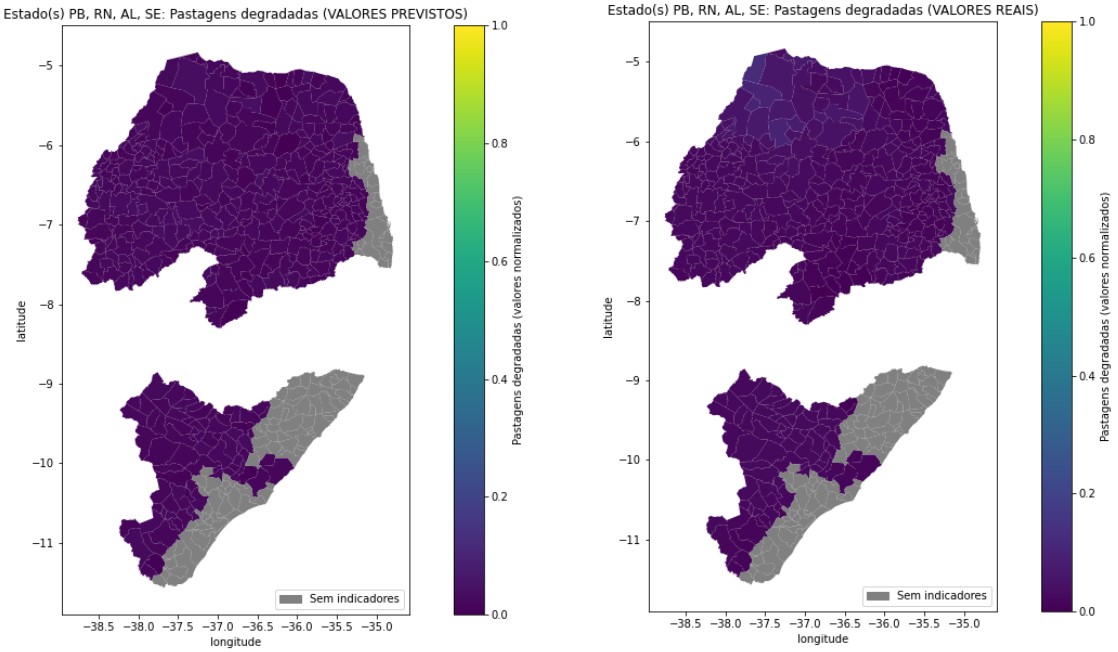
(b)



(c)

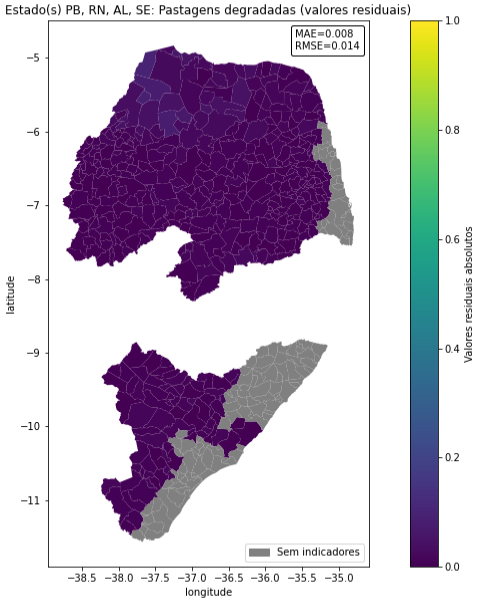
Fonte: Compilação do autor.

Figura 25: Mapas de calor gerados no experimento de validação cruzada totalmente randômica realizado com o indicador ambiental *Pastagens degradadas*, para o qual foi obtido o melhor resultado de R^2 para este experimento **(R^2=0.553)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.



(a)

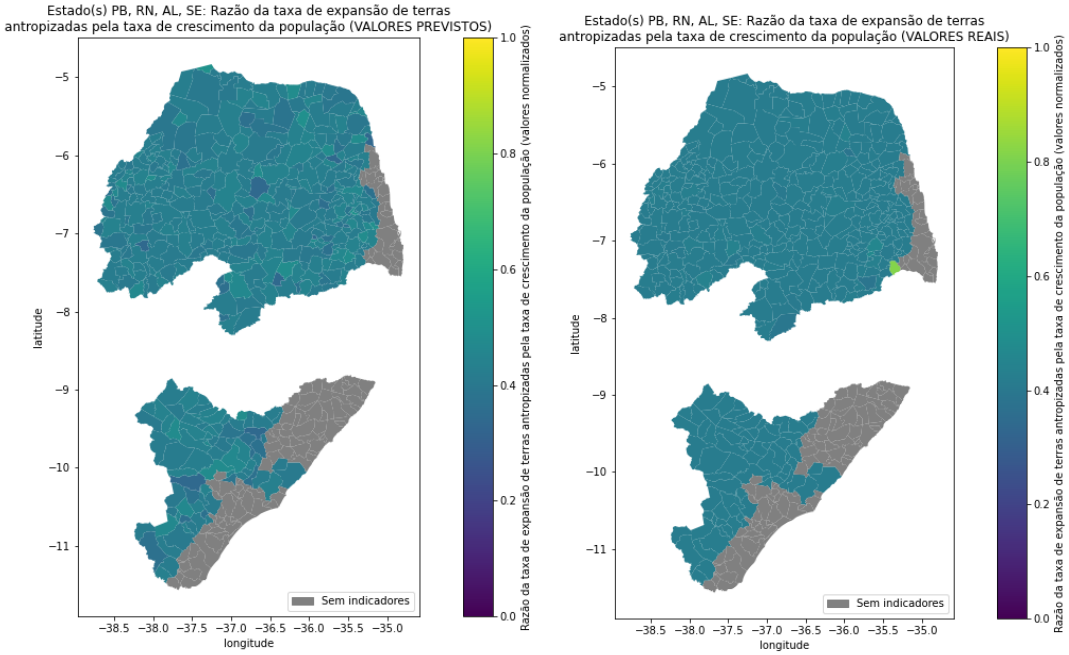
(b)



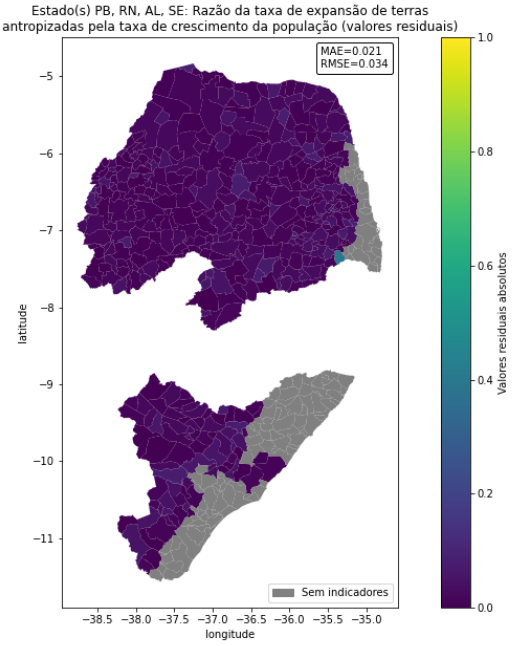
(c)

Fonte: Compilação do autor.

Figura 26: Mapas de calor gerados no experimento de validação cruzada totalmente randômica realizado com o indicador ambiental *Produtividade agrícola de alimentos básicos*, para o qual foi obtido o pior resultado de R² para este experimento **(R²=-36.293)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.



(a) (b)



(c)

Fonte: Compilação do autor.

7.1.3.2. Validação cruzada randomizada por estado

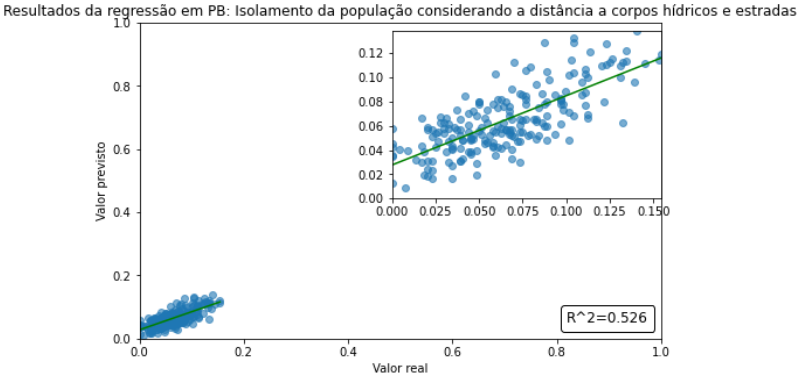
Neste experimento, o método da validação cruzada foi usado para estimar a média dos R^2 obtidos para cada um dos modelos de regressão Ridge dos indicadores ambientais e socioeconômicos, que foram treinados separadamente para cada um dos estados de RN, PB, AL e SE.

Para os indicadores socioeconômicos, os dois melhores resultados neste experimento foram obtidos para o indicador *Isolamento da população considerando a distância a corpos hídricos e estradas*, com **R^2=0.526** no estado de PB (Figura 27.a) e **R^2=0.462** no estado de RN (Figura 27.b). Vale comentar que entre os seis melhores resultados para a métrica do R^2, quatro foram obtidos para este indicador, com o R^2 variando de 0.307 a 0.526 (Apêndice C) dependendo do estado em que o modelo de regressão foi treinado. Para os modelos de regressão correspondentes aos indicadores *PIB per capita* e *Domicílios com renda maior que um salário mínimo*, foram obtidos R^2 entre 0.30 e 0.35 para o estado de RN (Apêndice C). Para os demais modelos, o R^2 encontrado foi menor do que 0.3 (Apêndice C). O pior valor de R^2 foi encontrado para o indicador *Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população* no estado de PB, com **R^2=-74.106** (Figura 27.c). A Figura 28 apresenta os mapas de calor correspondentes ao modelo de regressão com o melhor R^2 para este experimento, nos quais são representados os valores reais e previstos do indicador, e os resíduos calculados para cada um dos municípios. Já a Figura 29 apresenta os mapas de calor correspondentes ao modelo de regressão com o pior R^2. Os resultados de R^2 correspondentes aos demais indicadores socioeconômicos encontram-se no Apêndice C.

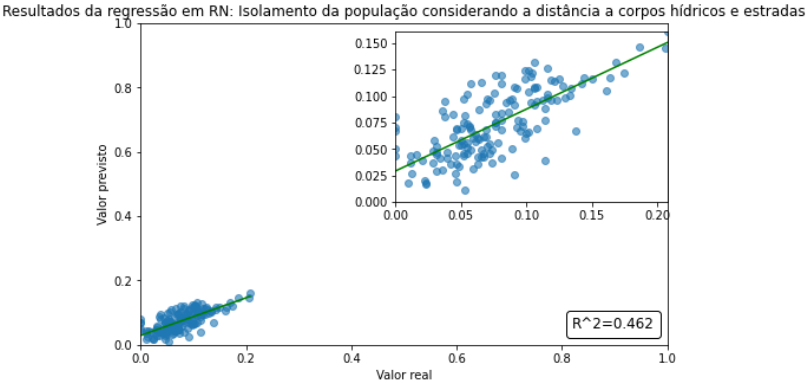
Para os indicadores ambientais, os três melhores resultados neste experimento foram obtidos para o indicador *Pastagens degradadas*, com **R^2=0.620** no estado de SE (Figura 30.a), **R^2=0.589** no estado de RN (Figura 30.b) e **R^2=0.520** no estado de RN (Figura 30.c). Para o estado de AL, obteve-se um R^2 ligeiramente menor, com R^2=0.433 (Apêndice D). Os modelos de regressão do indicador *Existência e proteção de recursos hídricos nos estabelecimentos agropecuários* apresentou R^2 maiores do que 0.4, com R^2=0.501 no estado de PB e R^2=0.439 no estado de RN (Apêndice D). Para o indicador *Produtividade agrícola de alimentos básicos*, também obteve-se um R^2 maior do que 0.4 no estado de PB, com o R^2=0.400 (Apêndice D). Para os demais modelos, o R^2 encontrado foi menor do que 0.3 (Apêndice D). O pior valor de R^2 foi encontrado para o indicador *Produtividade agrícola de alimentos básicos* no estado de RN, com **R^2=-270.160** (Figura 30.d). A Figura 31 apresenta os mapas de calor correspondentes ao modelo de regressão com o melhor R^2 para este experimento, nos quais são representados os valores reais e previstos do indicador, e os resíduos calculados para cada um dos municípios. Já a Figura 32 apresenta

os mapas de calor correspondentes ao modelo de regressão com o pior R^2. Os resultados de R^2 correspondentes aos demais indicadores socioeconômicos encontram-se no Apêndice D.

Figura 27: Gráficos de dispersão comparando os valores reais dos indicadores e os valores previstos pelo modelo de regressão, para o experimento de validação cruzada randomizada por estado realizado com os indicadores **socioeconômicos**. Os gráficos (a) e (b) representam os dois melhores resultados de R^2 para este experimento. O gráfico (c) representa o pior resultado de R^2.



(a)

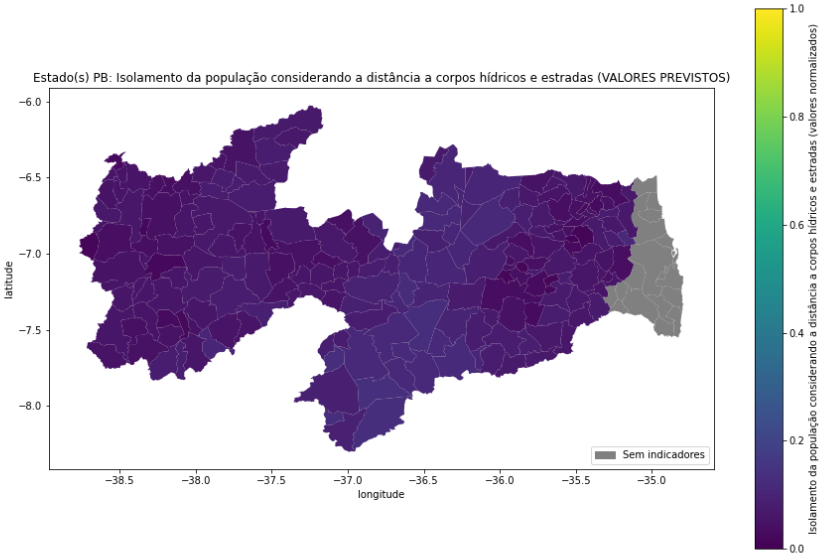


(b)

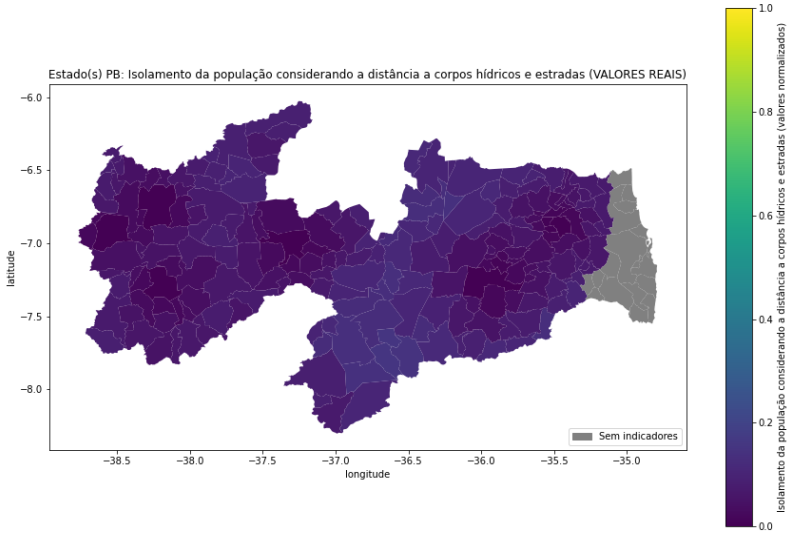
(c)

Fonte: Compilação do autor.

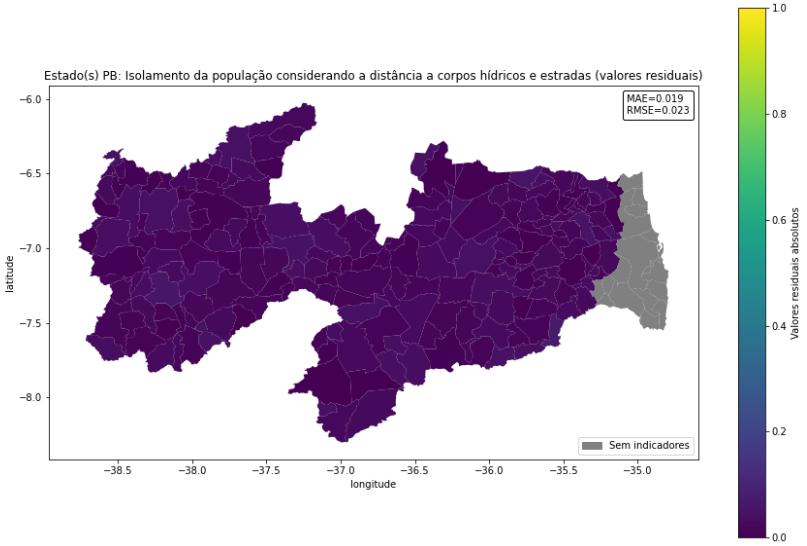
Figura 28: Mapas de calor gerado no experimento de validação cruzada totalmente randômica realizado para o indicador socioeconômico *Isolamento da população considerando a distância a corpos hídricos e estradas* no estado de PB, para o qual foi obtido o melhor resultado de R^2 para este experimento **(R^2=0.526)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.



(a)



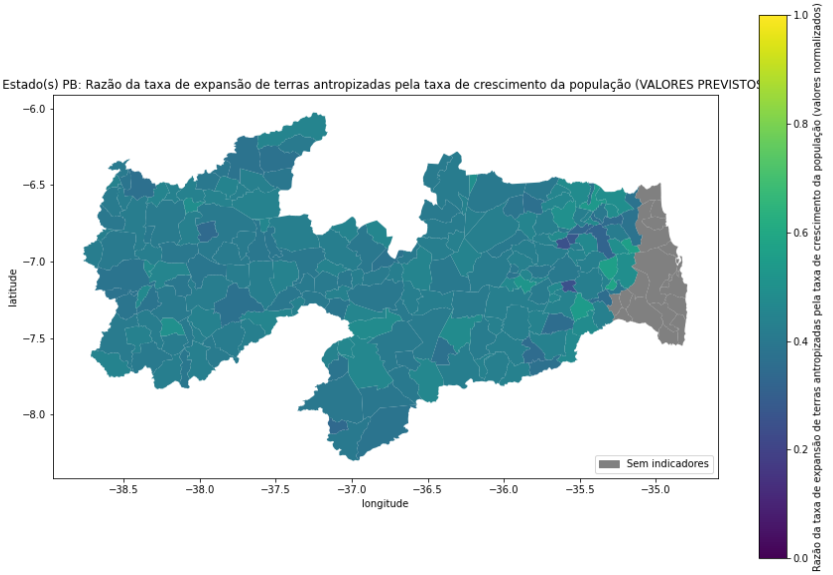
(b)



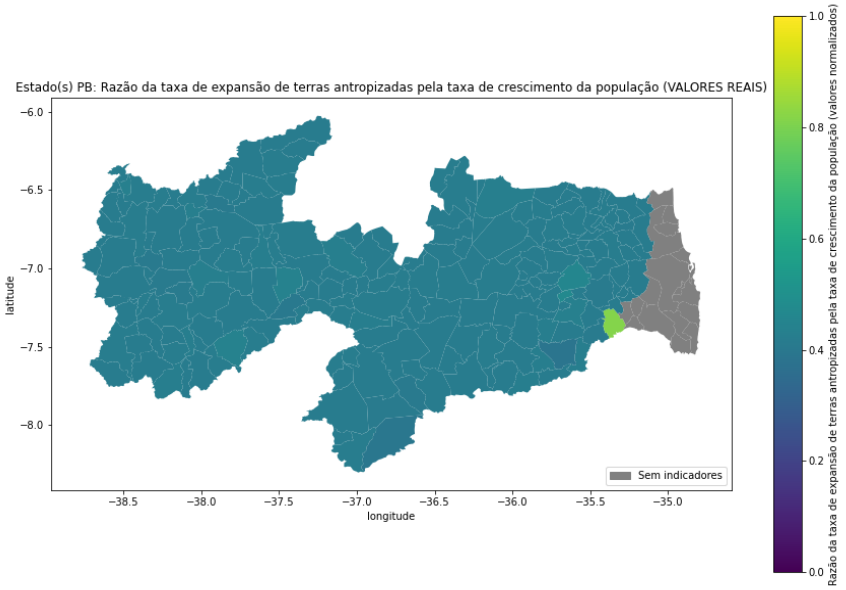
(c)

Fonte: Compilação do autor.

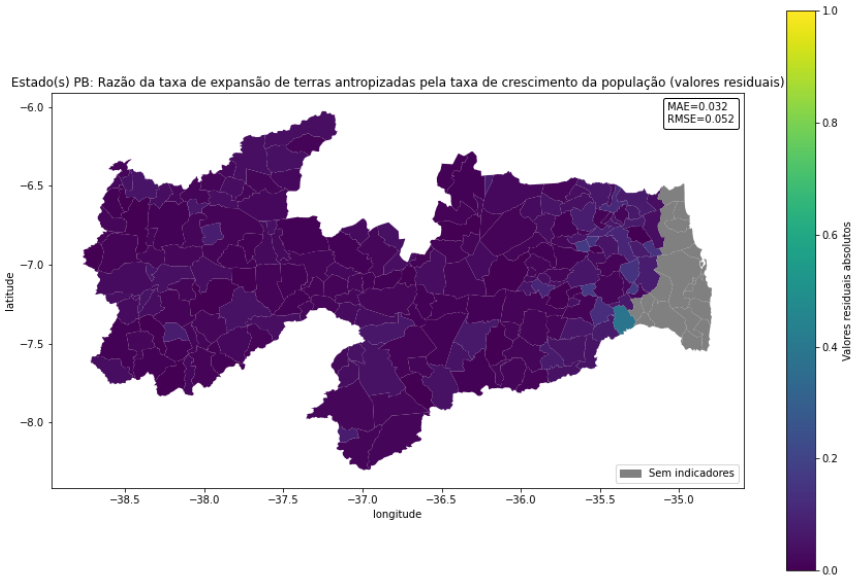
Figura 29: Mapas de calor gerado no experimento de validação cruzada totalmente randômica realizado para o indicador socioeconômico *Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população* no estado de PB, para o qual foi obtido o pior resultado de R^2 para este experimento **(R^2=-74.106)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.



(a)



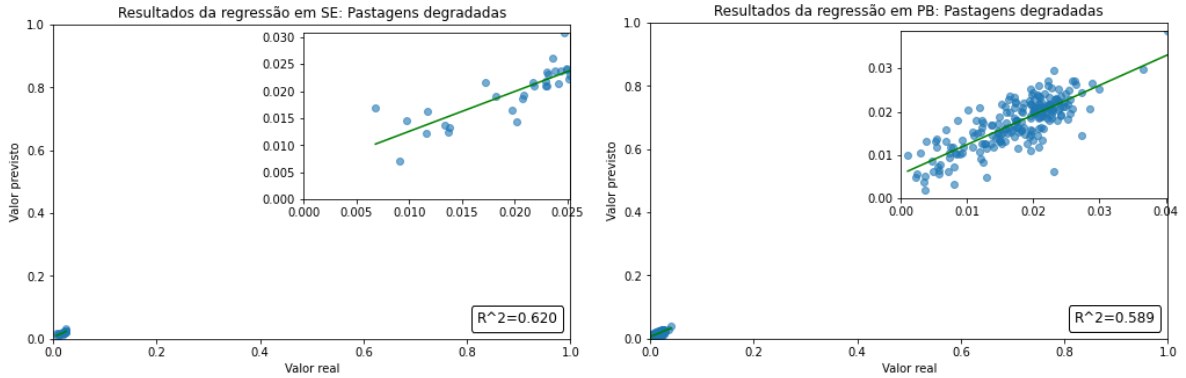
(b)



(c)

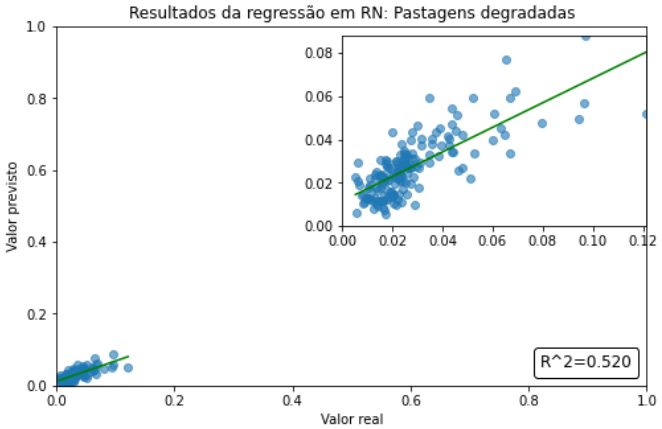
Fonte: Compilação do autor.

Figura 30: Gráficos de dispersão comparando os valores reais dos indicadores e os valores previstos pelo modelo de regressão, para o experimento de validação cruzada randomizada por estado, realizado com os indicadores **ambientais**. Os gráficos (a), (b) e (c) representam os três melhores resultados de R^2 para este experimento. O gráfico (d) representa o pior resultado de R^2.

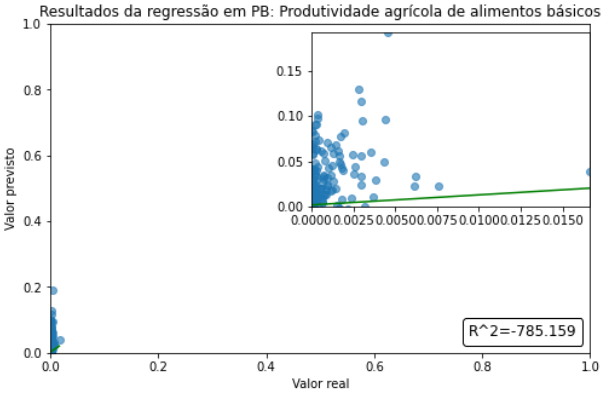


(a)

(b)



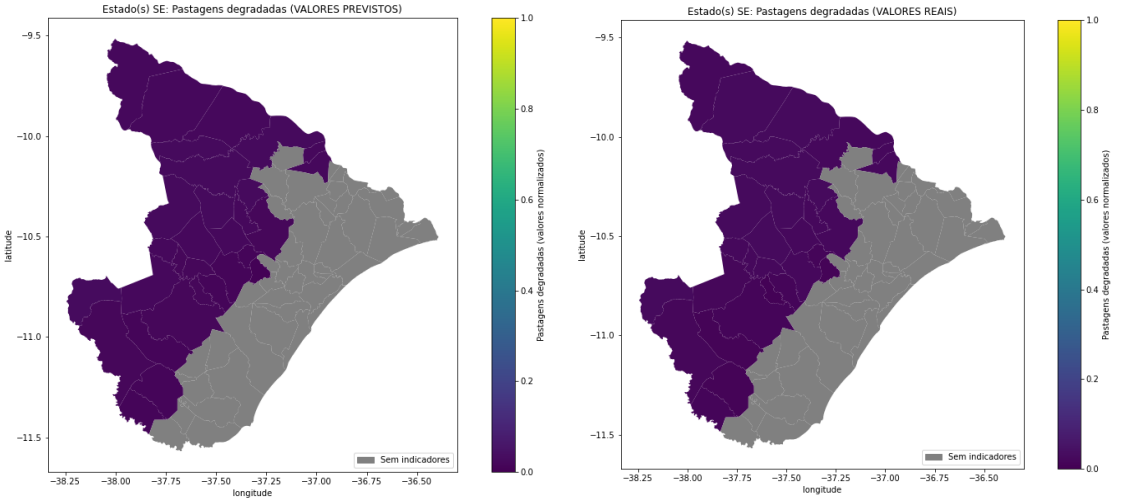
(c)



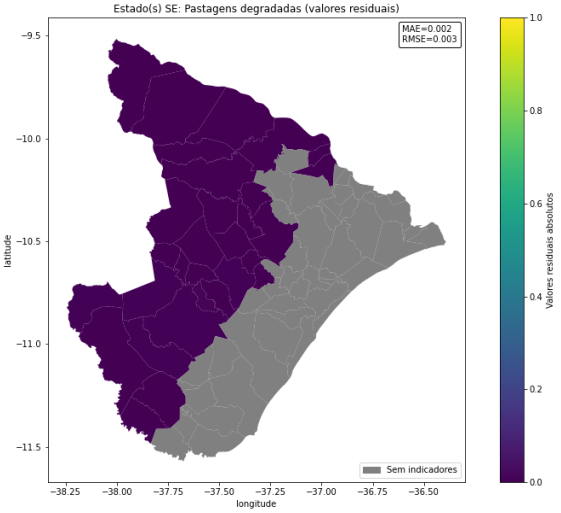
(d)

Fonte: Compilação do autor.

Figura 31: Mapas de calor gerado no experimento de validação cruzada totalmente randômica realizado para o indicador ambiental *Pastagens degradadas* no estado de SE, para o qual foi obtido o melhor resultado de R^2 para este experimento **(R^2=0.620)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.



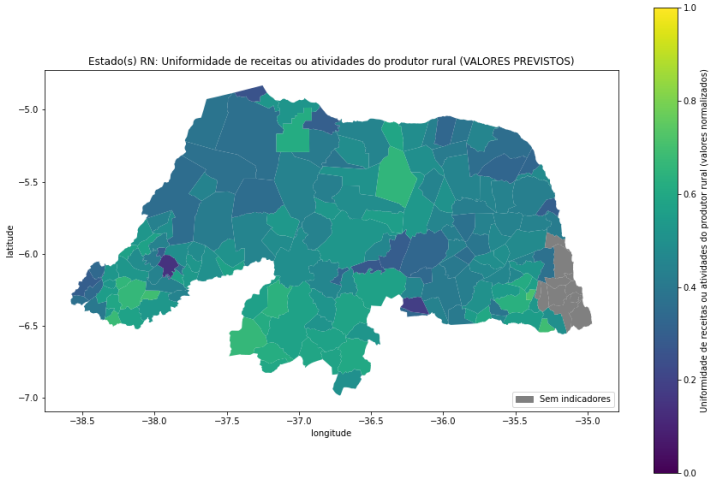
(a) (b)



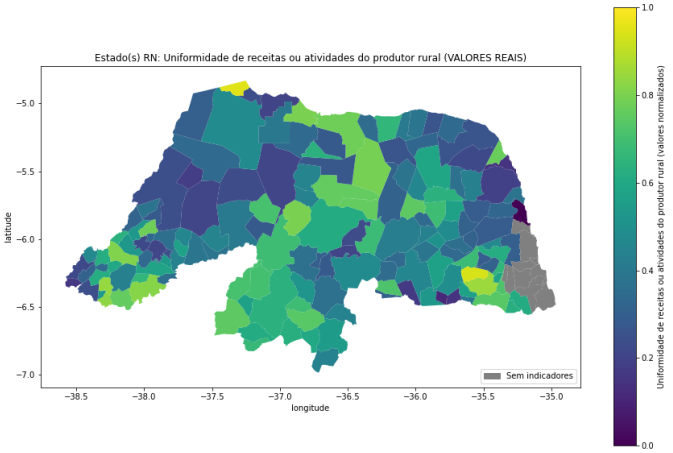
(c)

Fonte: Compilação do autor.

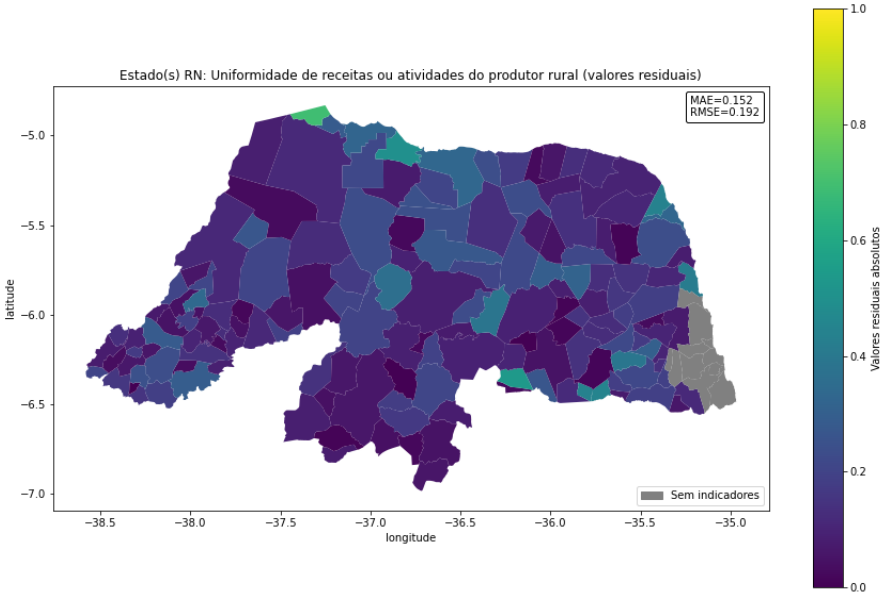
Figura 32: Mapas de calor gerado no experimento de validação cruzada totalmente randômica realizado para o indicador ambiental *Produtividade agrícola de alimentos básicos* no estado de RN, para o qual foi obtido o pior resultado de R^2 para este experimento **(R^2=-270.160)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.



(a)



(b)



(c)

Fonte: Compilação do autor.

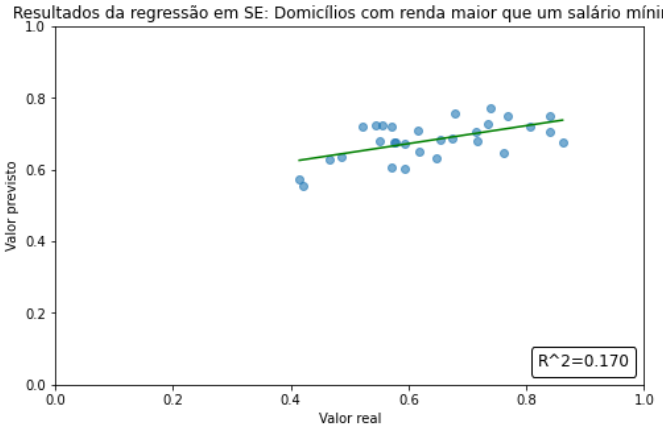
7.2.3.3. Validação transregional

No experimento de validação transregional, foi separado um estado para teste, e os demais três estados para o treinamento, validação e a escolha do melhor parâmetro de ajuste do modelo de regressão. Para cada um dos indicadores ambientais e socioeconômicos foram feitas quatro rodadas deste experimento, e em cada uma destas foi separado um estado diferente para teste (e os demais para a validação e treinamento do modelo).

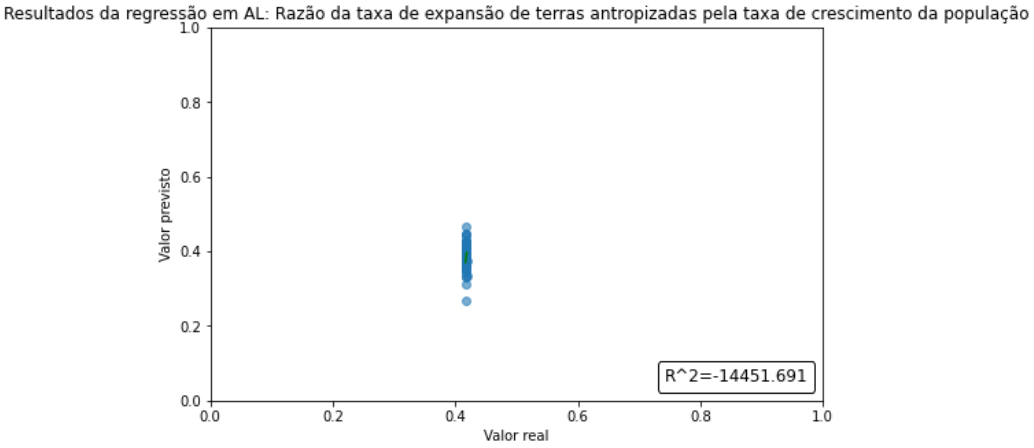
Para os indicadores socioeconômicos, o melhor resultado do R^2 foi menor do que 0.2, com **R^2=0.170** (Figura 33.a) para o modelo de regressão do indicador *Domicílios com renda maior que um salário mínimo* que foi testado no estado de SE. Para os demais indicadores, o R^2 foi menor que 0.1, atingindo muitas vezes valores negativos (Apêndice E) - como no caso do indicador *Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população* que apresentou o pior resultado, com **R^2=-14451.691** (Figura 33.b). A Figura 34 apresenta os mapas de calor correspondentes ao modelo de regressão com o melhor R^2 para este experimento, nos quais são representados os valores reais do indicador, os valores previstos pelo modelo e os resíduos calculados para cada um dos municípios. Já a Figura 35 apresenta os mapas de calor correspondentes ao modelo de regressão com o pior R^2. Os resultados de R^2 correspondentes aos demais indicadores socioeconômicos encontram-se no Apêndice E.

Já no caso dos experimentos com indicadores ambientais, os melhores R^2 obtidos superaram 0.3, com **R^2=0.405** para o modelo de regressão do indicador *Existência e proteção de recursos hídricos nos estabelecimentos agropecuários* que foi testado em PB (Figura 36.a), e **R^2=0.309** para o modelo de regressão do indicador *Pastagens degradadas* o qual foi testado em SE (Figura 36.b). Para os demais indicadores, o R^2 foi menor que 0.3. O pior R^2 foi obtido para o modelo de regressão do indicador *Produtividade agrícola de alimentos básicos* testado no estado de PB, com **R^2=-785.159** (Figura 36.c). A Figura 37 apresenta os mapas de calor correspondentes ao modelo de regressão com o melhor R^2 para este experimento (obtido para o indicador *Pastagens degradadas)* e a Figura 38 apresenta os mapas de calor correspondentes ao modelo de regressão com o pior R^2 (obtido para o indicador *Produtividade agrícola de alimentos básicos*). Os resultados de R^2 correspondentes aos demais indicadores ambientais encontram-se no Apêndice B.

Figura 33: Gráficos de dispersão comparando os valores reais dos indicadores e os valores previstos pelo modelo de regressão, para o experimento de validação cruzada transregional realizado com os indicadores **socioeconômicos**. O gráfico (a) representa o melhor resultado de R^2 para este experimento. O gráfico (b) representa o pior resultado de R^2.



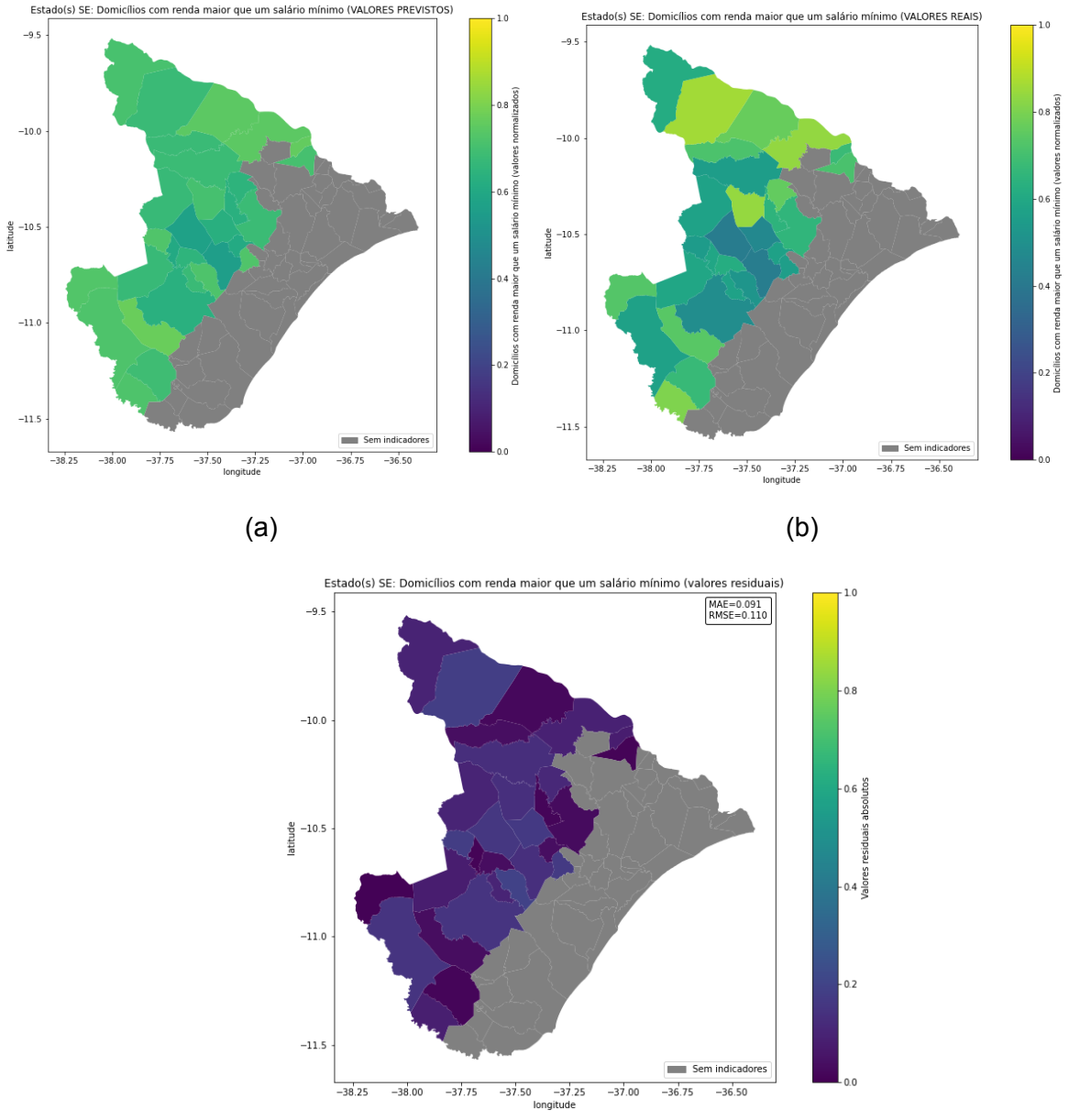
(a)



(b)

Fonte: Compilação do autor.

Figura 34: Mapas de calor gerado no experimento de validação cruzada transregional realizado para o modelo de regressão do indicador socioeconômico *Domicílios com renda maior que um salário mínimo* que foi testado no estado de SE, para o qual foi obtido o melhor resultado de R^2 para este experimento **(R^2=0.170)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.



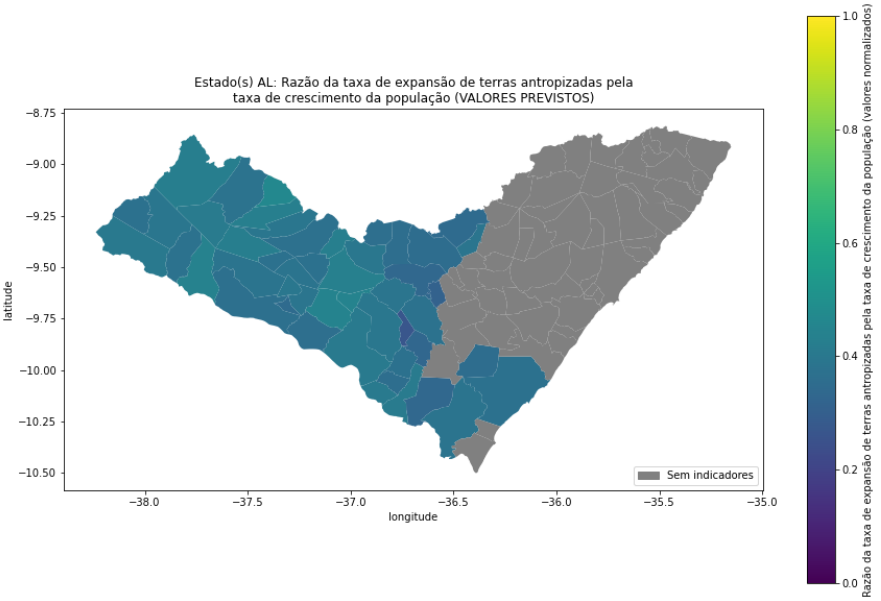
(a) (b)

(c)

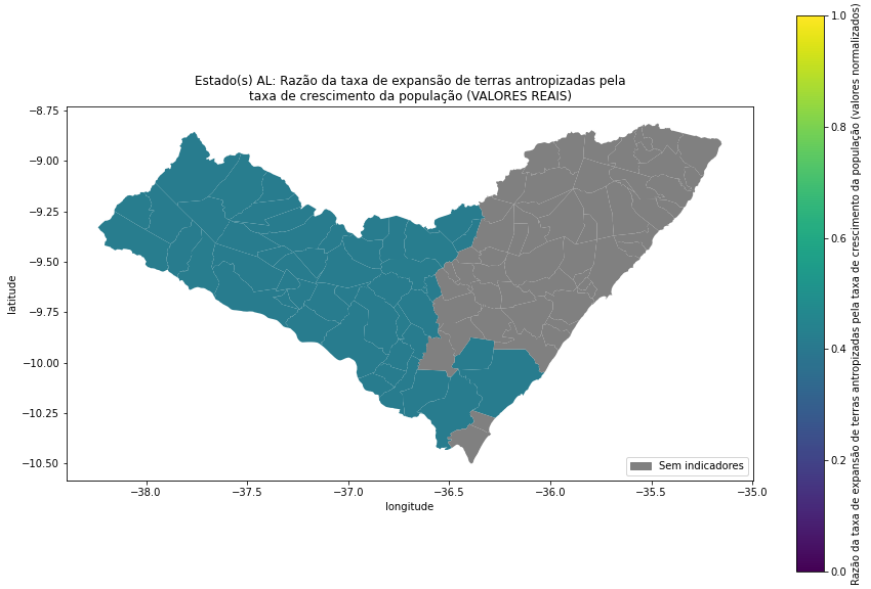
Fonte: Compilação do autor.

Figura 35: Mapas de calor gerado no experimento de validação cruzada transregional realizado para o modelo de regressão do indicador socioeconômico *Razão da taxa de*

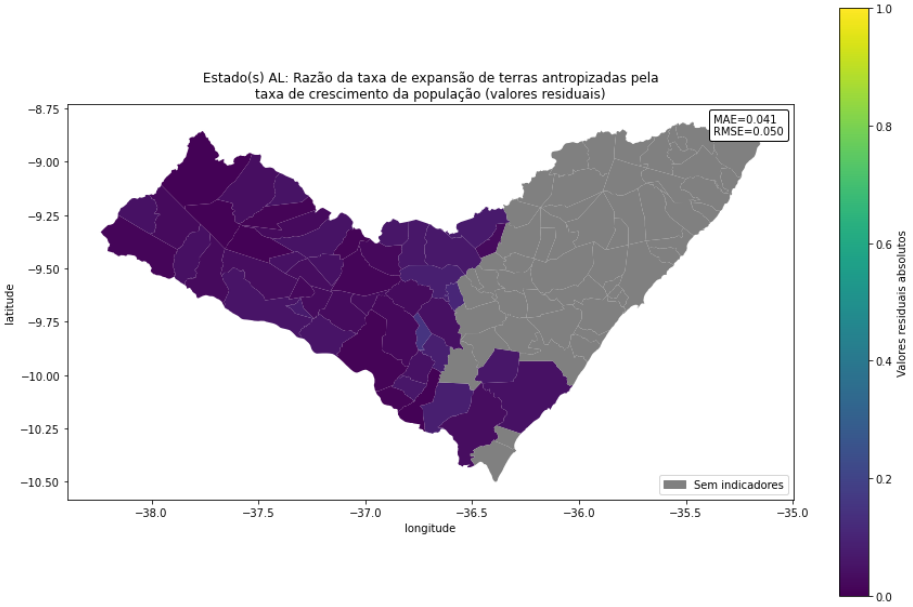
expansão de terras antropizadas pela taxa de crescimento da população que foi testado no estado de AL, para o qual foi obtido o pior resultado de R^2 para este experimento **(R^2=-14451.6912)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.



(a)



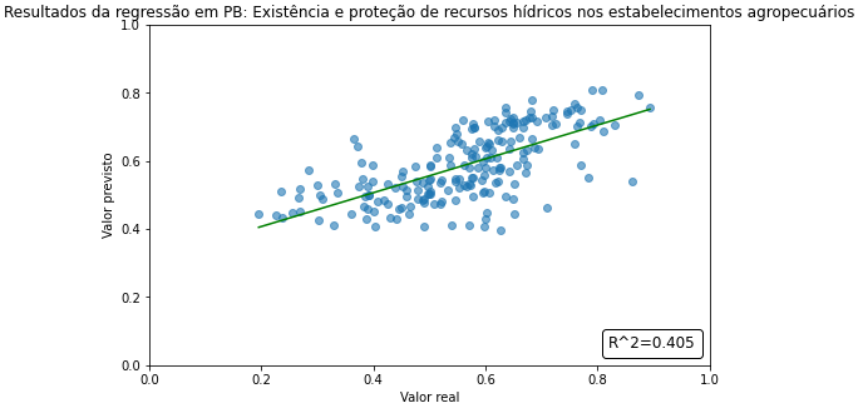
(b)



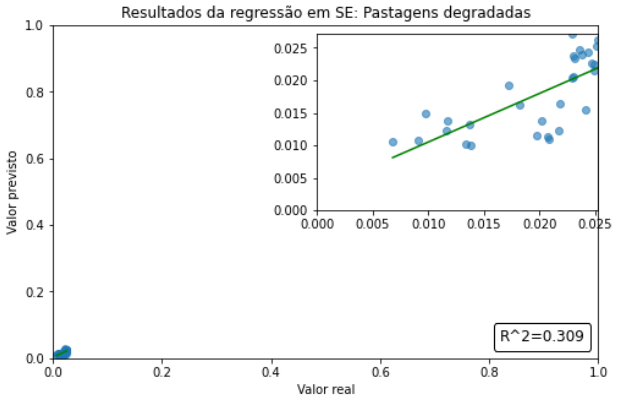
(c)

Fonte: Compilação do autor.

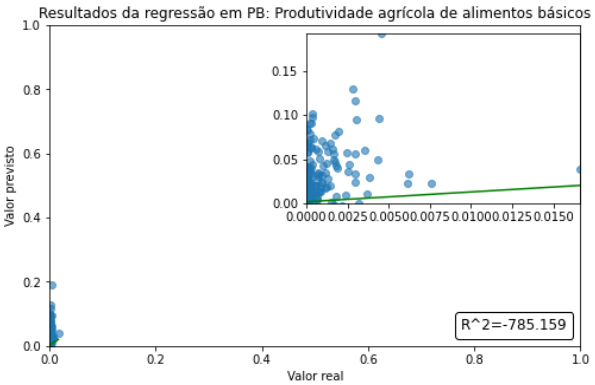
Figura 36: Gráficos de dispersão comparando os valores reais dos indicadores e os valores previstos pelo modelo de regressão, para o experimento de validação cruzada transregional realizado com os indicadores **ambientais**. Os gráficos (a) e (b) representam os dois melhores resultados de R^2 para este experimento. O gráfico (c) representa o pior resultado de R^2.



(a)



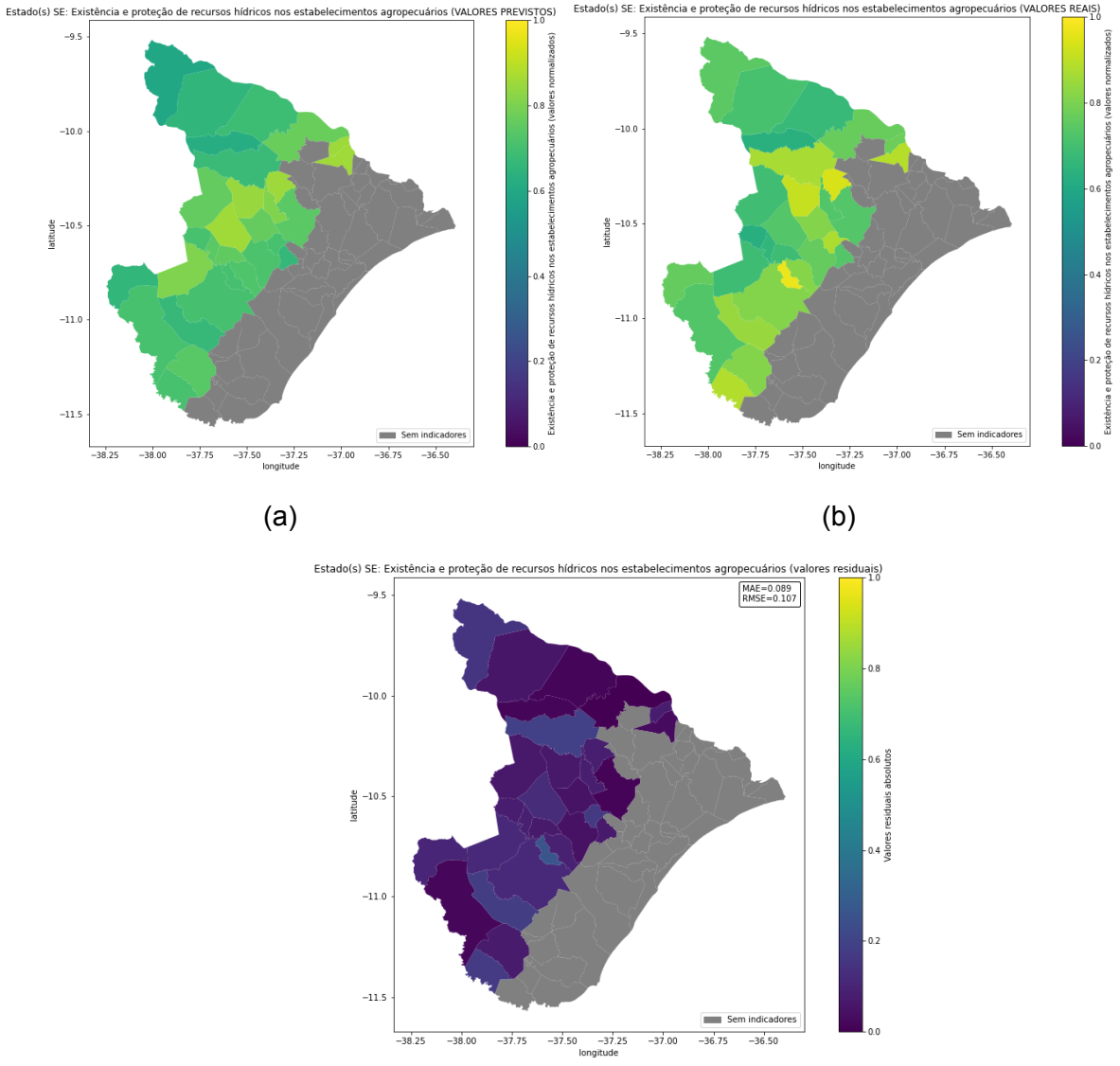
(b)



(c)

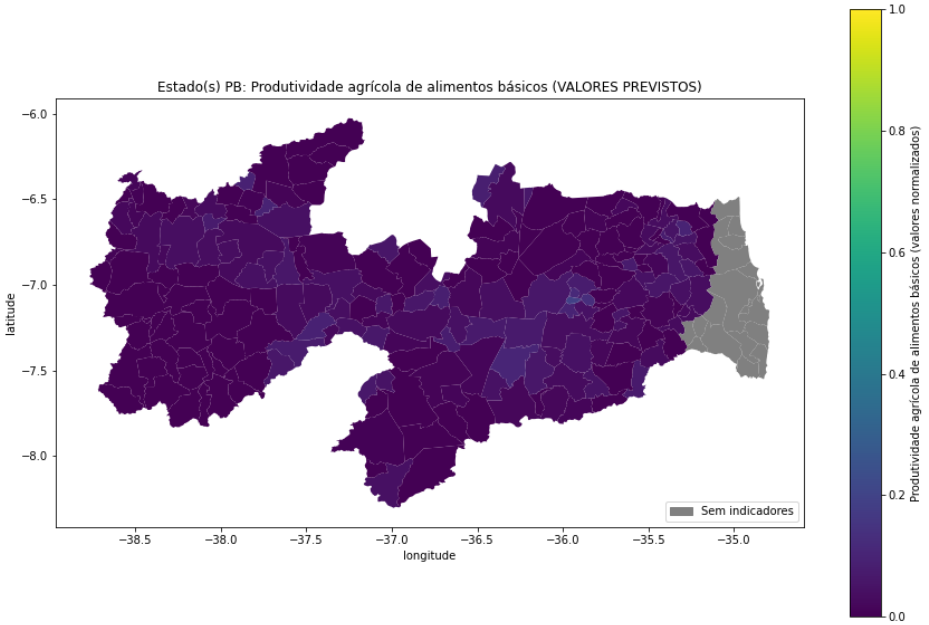
Fonte: Compilação do autor.

Figura 37: Mapas de calor gerado no experimento de validação cruzada transregional realizado para o modelo de regressão do indicador ambiental *Existência e proteção de recursos hídricos nos estabelecimentos agropecuários* que foi testado no estado de SE, para o qual se obteve o melhor resultado de R^2 **(R^2=0.405)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.

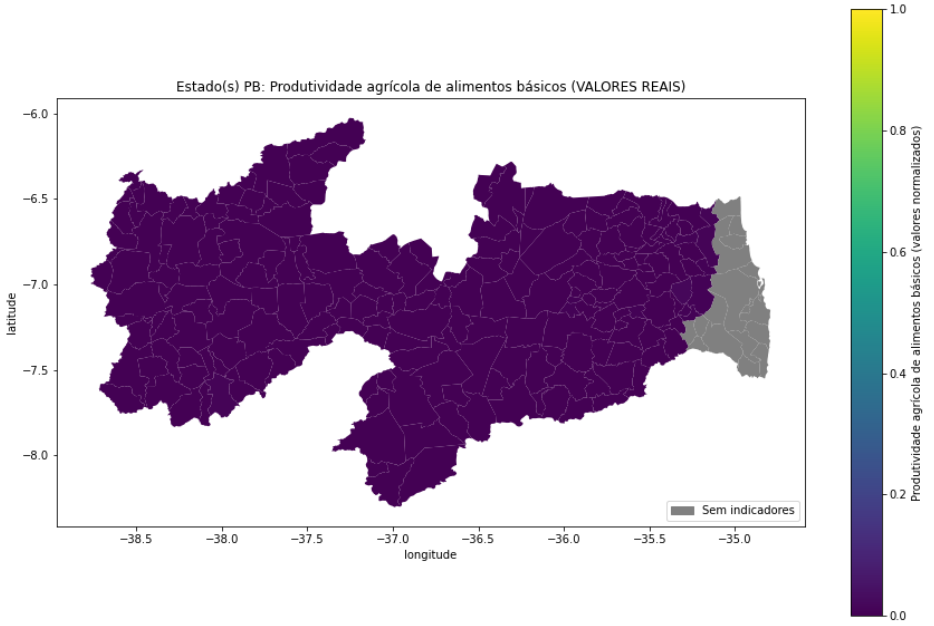


Fonte: Compilação do autor.

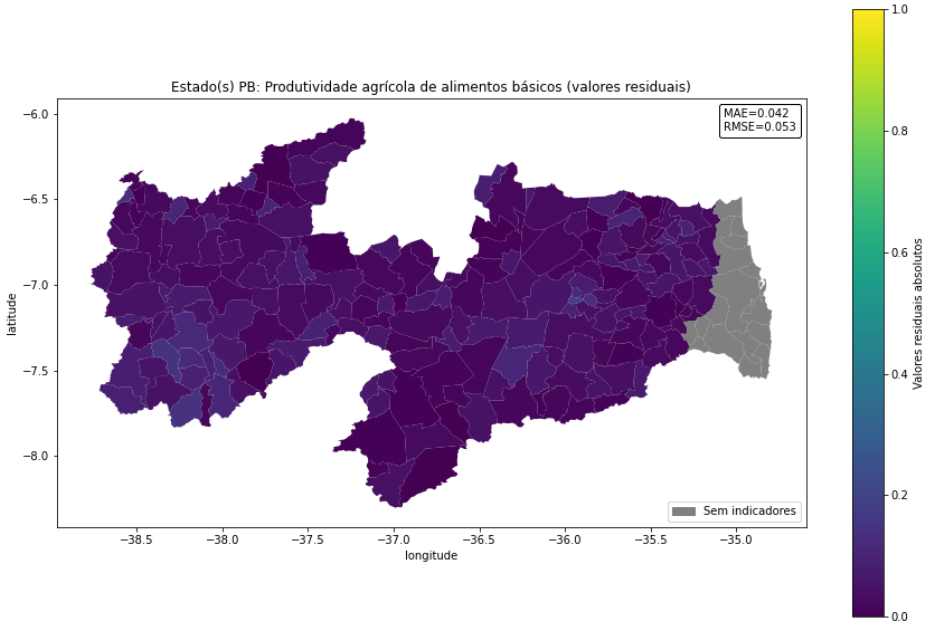
Figura 38: Mapas de calor gerado no experimento de validação cruzada transregional realizado para o modelo de regressão do indicador socioeconômico *Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população* que foi testado no estado de AL, para o qual foi obtido o pior resultado de R^2 para este experimento **(R^2=-785.158)**. A imagem (a) representa os valores previstos pelo modelo e (b) representa os valores reais do indicador. A imagem (c) representa os resíduos do modelo de regressão.



(a)



(b)



(c)

Fonte: Compilação do autor.

7.1.4. Análise geral dos resultados

Foram realizados experimentos para a previsão de 11 indicadores socioeconômicos e 9 indicadores ambientais em mais de 400 municípios pertencentes ao bioma da Caatinga, permitindo analisar quais deles tinham maior ou menor nível de correlação com as imagens de satélite registradas.

Fazendo-se uma análise geral dos resultados dos experimentos, pode-se dizer que os modelos de predição gerados pela combinação de técnicas de transferência de aprendizado e regressão linear se mostraram satisfatórios para alguns dos indicadores ambientais e socioeconômicos. Como já foi dito no Capítulo 5, considera-se que o modelo é bem sucedido se o coeficiente de determinação (R^2) obtido for maior que **0.37**, ou seja, semelhante ou superior ao mínimo apresentado como satisfatório nos trabalhos que serviram de base para esse projeto (Triñanes et al. (2020) [12] e Jean et al. (2016) [3]). Idealmente quanto mais próximo de (R^2) igual **1.0** melhor o resultado, entretanto adotamos o limiar de 0.37 como referência aos trabalhos anteriores, de forma que resultados superiores a esse limiar demonstram uma reprodução bem sucedida e/ou melhora com relação aos trabalhos base.

No caso dos experimentos de validação cruzada totalmente randômica, nos quais os modelos de predição foram treinados com todos os estados juntos, foram obtidos resultados equiparáveis aos modelos de predição dos artigos de referência para os indicadores ambientais *Pastagens degradadas* (**R^2=0.553**) e *Existência e proteção de recursos hídricos nos estabelecimentos agropecuários* (**R^2=0.512**). Para os demais indicadores, os valores de correlação obtidos não foram maiores do que 0.37.

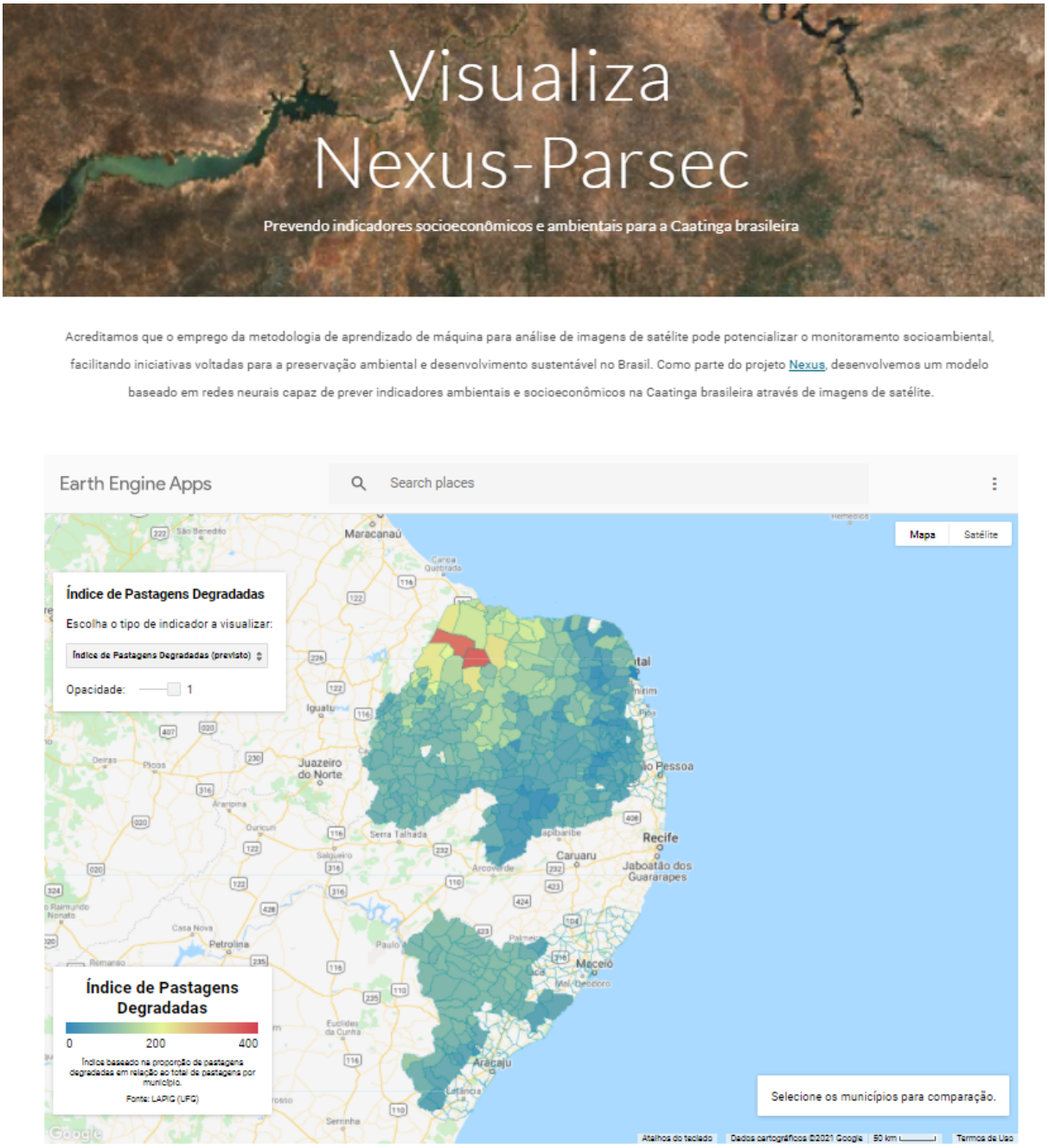
No caso dos experimentos de validação cruzada randomizada por estado, nos quais os modelos de predição foram treinados separadamente para cada estado, foram obtidos resultados similares aos dos artigos de referência para o indicador ambiental *Pastagens degradadas* (com o **R^2** variando de **0.433 a 0.620**, de acordo com o estado em que foi treinado o modelo). Para este indicador, foram obtidos resultados considerados satisfatórios para todos os estados. Já no caso do indicador ambiental *Existência e proteção de recursos hídricos nos estabelecimentos agropecuários*, foram obtidos coeficientes de determinação maiores do que 0.37 somente para os estados de PB (**R^2=0.501**) e de RN (**R^2=0.439**). Para um outro indicador ambiental denominado *Produtividade agrícola de alimentos básicos*, foi encontrado um resultado satisfatório somente para o estado de PB (**R^2=0.400**). Para o indicador socioeconômico *Isolamento da população considerando a distância a corpos hídricos e estradas*, foram obtidos resultados bem-sucedidos somente para o estado de PB (**R^2=0.526**) e para o estado de RN (**R^2=0.462**). Para os demais casos que não foram citados, os valores dos coeficientes de determinação (R^2) foram menores do que 0.37.

Finalmente, para os experimentos de validação cruzada transregional, nos quais foram conduzidos quatro rodadas com um estado separado para teste, e os demais três estados para o treinar o modelo, foram obtidos resultados satisfatórios somente para o indicador ambiental *Existência e proteção de recursos hídricos nos estabelecimentos agropecuários*, cujo modelo de predição foi testado em PB e treinado nos demais estados (com **R^2=0.405**). Para os demais modelos de predição, os valores dos coeficientes de determinação foram menores do que **0.37**. Vale destacar que não foram gerados modelos com resultados satisfatórios para os indicadores socioeconômicos neste terceiro experimento.

7.2. Plataforma *Web*

Dos resultados obtidos, foram selecionados os dois indicadores (um socioeconômico, um ambiental) com os melhores coeficientes de determinação para serem mostrados através do mapa interativo na página inicial da plataforma *web,* mostrada na Figura 39. É possível alternar entre os indicadores previstos pelo modelo e os indicadores reais, bem como comparar os dados de municípios específicos ao selecioná-los na tela, com gráficos e tabelas geradas dinamicamente. Por fim, os dados dos municípios selecionados podem ser exportados tanto na forma de imagem do gráfico, como em tabela no formato **.csv**.

Figura 39: Tela inicial da plataforma *web* Visualiza NEXUS-PARSEC.



Fonte: Compilação do autor

A plataforma também conta com explicações sobre os indicadores mostrados, além de informações sobre a motivação e a metodologia do trabalho realizado, e pode ser acessada através do endereço < https://sites.google.com/usp.br/visualizanexus/ >. Demais capturas de tela das páginas da plataforma encontram-se no Apêndice G.D

8. Considerações Finais

8.1. Conclusões

Neste trabalho parte-se da premissa de que as imagens de satélite contêm informações visuais ricas que podem ser correlacionadas com indicadores socioeconômicos e ambientais. Por esta razão, seguindo o intuito de que imagens de luzes noturnas tem relação com indicadores socioeconômicos, da mesma forma pressupõe-se que imagens de evapotranspiração teriam relação com indicadores ambientais, pelo que poderia ser demonstrado ao obter valores de coeficiente de determinação mais próximos de 1.

O trabalho conseguiu alcançar o seu objetivo, aplicando a metodologia de redes neurais para estimar indicadores socioeconômicos e ambientais em quatro estados do Nordeste brasileiro, e disponibilizando os resultados através de uma plataforma online com mapas interativos.

Inicialmente, planejava-se realizar os experimentos para todo o bioma da Caatinga e do Cerrado, mas por conta de limitações de recursos na aquisição de dados, a área de estudo teve de ser reduzida a apenas quatro estados da Caatinga. Mesmo assim, o modelo de rede neural foi treinado com sucesso, conseguindo uma correlação de até 62% entre os indicadores previstos e os reais.

Os resultados mostraram que a estrutura de aprendizagem profunda usada aqui é capaz de estimar indicadores socioeconômicos e ambientais para indicadores que são mais condizentes com as imagens de satélite utilizadas como *proxy*.

Além disso, é possível salientar que o modelo base fornecido por Jean et al. (2016) é generalizável para o cenário brasileiro, pois a reprodutibilidade e aplicabilidade de sua metodologia foi demonstrado com sucesso.

8.2. Contribuições

Embora não sejam precisos o suficiente para substituírem os dados de pesquisa de campo, os resultados já podem ser usados de maneira qualitativa para inferir a situação socioambiental atual dos municípios analisados em comparação com aquela do período no qual os indicadores foram coletados — atestando o potencial da metodologia de aprendizado de máquina empregada e estimulando trabalhos futuros no campo de predição de dados através de imagens de satélite.

Além disso, o projeto evidenciou o potencial do uso de imagens de evapotranspiração em conjunto com técnicas de aprendizado para estimar indicadores ambientais.

As mais de 170 mil imagens de satélite coletadas e os dados obtidos durante a execução do trabalho estarão disponíveis para reúso para projetos futuros, e também serão compartilhados com o projeto NEXUS-PARSEC para que possam ser usados por pesquisadores tanto no Brasil como internacionalmente.

Por fim, os *scripts* desenvolvidos para a criação dos mapas interativos estão disponíveis abertamente, e podem ser facilmente adaptados e expandidos para o uso em outras aplicações que envolvem a visualização de dados em mapas.

8.3. Perspectivas de Continuidade

Como proposta de continuidade para o presente trabalho, os experimentos propostos podem ser realizados em uma área maior, englobando desta vez os biomas do Cerrado e da Caatinga em sua totalidade, e analisando também uma gama maior de indicadores.

Além de treinar o modelo de rede neural em áreas maiores, podem ser usadas imagens de satélite com dados temporais, de modo a desenvolver um modelo de predição com indicadores obtidos em diferentes períodos, aumentando assim a quantidade de dados disponível e possivelmente melhorando a capacidade preditiva do modelo.

A plataforma online desenvolvida também pode ser continuamente expandida conforme novos resultados forem obtidos, ampliando a área de cobertura e a quantidade de indicadores disponíveis para visualização nos mapas interativos.

Referências

[1] INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). *Áreas Territoriais*. Disponível em: <https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15761-areas-dos-municipios.html?=&t=sobre>. Acesso em: 14 dez. 2021.

[2] MINISTÉRIO DO MEIO AMBIENTE. *Biodiversidade*. Disponível em: <https://www.gov.br/mma/pt-br/assuntos/biodiversidade>. Acesso em: 14 dez. 2021.

[3] JEAN, N. et al. Combining satellite imagery and machine learning to predict poverty (2016). Science 19 Aug 2016: Vol. 353, Issue 6301, pp. 790-794.

[4] YEH, C. et al. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa (2020). Nat Commun 11, 2583.

[5] INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. *NEXUS - Transição para sustentabilidade e o nexo agricultura-energia-água: uma abordagem integradora no Cerrado e Caatinga*. Disponível em: <http://nexus.ccst.inpe.br/projeto/>. Acesso em: 14 dez. 2021.

[6] PARSEC. *Building New Tools for Data Sharing and Reuse through a Transnational Investigation of the Socioeconomic Impacts of Protected Areas*. Disponível em: <http://parsecproject.org/>. Acesso em: 14 dez. 2021.

[7] IBGE. *CONCLA - O Brasil no Mundo.* Disponível em: <https://cnae.ibge.gov.br/en/component/content/article/94-7a12/7a12-vamos-conhecer-o-brasil/nosso-territorio/1461-o-brasil-no-mundo.html>. Acesso em: 14 dez. 2021.

[8] PATTERSON, J.; GIBSON, A. *Deep Learning: A Practitioner's Approach.* 1st ed. O'Reilly Media, Inc., 2017. 538 p.

[9] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. The MIT Press, 2016. 800 p. Disponível em <https://www.deeplearningbook.org/>. Acesso em: 19 abr. 2021.

[10] REYNOLDS, D. *Gaussian Mixture Models* (2015). In: Li S.Z., Jain A.K. (eds) Encyclopedia of Biometrics. Springer, Boston, MA.

[11] SCIKIT-LEARN DEVELOPERS. *Gaussian mixture models*. Disponível em: <https://scikit-learn.org/stable/modules/mixture.html>. Acesso em: 16 dez. 2021.

[12] JAMES, G. et al. *An Introduction to Statistical Learning with Applications in R.* 1st ed. Springer, 2013. 440 p.

[13] HOERL, A. E.; KENNARD R. W. *Ridge Regression: Biased Estimation for Nonorthogonal Problems* (2000). *Technometrics* 42, no. 1, pp. 80-86.

[14] CASAGRANDE, M. H. *Comparação de métodos de estimação para problemas com colinearidade e/ou alta dimensionalidade (p > n)* (2016). Dissertação (Mestrado em Estatística) - Estatística Interinstitucional do ICMC e UFSCar, Universidade de São Paulo, São Carlos, 2016.

[15] SCIKIT-LEARN DEVELOPERS. *Metrics and scoring: quantifying the quality of predictions*. Disponível em: <https://scikit-learn.org/stable/modules/model_evaluation.html>. Acesso em: 16 dez. 2021.

[16] HASTIE, T.; FRIEDMAN, J; TIBSHIRANI, R. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, 2009. 767 p.

[17] WORLD BANK, 2020. About LSMS. Disponível em <https://www.worldbank.org/en/programs/lsms/overview>. Acesso em: 14 dez. 2021.

[18] PLANET API. *OVERVIEW* em: <https://developers.planet.com/docs/apis/data/>. Acesso em: 14 dez. 2021.

[19] TRIÑANES, E.; MACHICAO, J.; CORRÊA, P. (2020). Application of a deep learning algorithm for predicting socioeconomic data through satellite images in the Vale do Ribeira. Disponível em: <https://doi.org/10.5281/zenodo.4712815>. Acesso em: 14 dez. 2021.

[20] ABREU, MARCOS & OLIVEIRA, JULIO & ANDRADE, VIVIANE & MEIRA, ANDERSON. Methodological proposal for spatial calculation and analysis of the intra-urban HDI of Viçosa, Brazil. (2011). Revista Brasileira de Estudos de População. 28. 169-186.

[21] GOOGLE. *Welcome to Google Maps Platform*. Disponível em: <https://mapsplatform.google.com/?hl=pt-br>. Acesso em: 16 dez. 2021.

[22] GOOGLE. *Google Earth Engine*. Disponível em: <https://earthengine.google.com/>. Acesso em: 16 dez. 2021.

[23] COLORADO SCHOOL OF MINES. *EARTH OBSERVATION GROUP*. Disponível em: <https://payneinstitute.mines.edu/eog/>. Acesso em: 16 dez. 2021.

[24] IMAGENET. *What is ImageNet* Disponível em: <https://image-net.org/about.php>. Acesso em: 14 dez. 2021.

[25] *What is Worldnet.* Disponível em: <https://wordnet.princeton.edu/>. Acesso em: 14 dez. 2021.

[26] GOOGLE. *Earth Engine Apps*. Disponível em: <https://www.earthengine.app/>. Acesso em: 16 dez. 2021.

[27] GOOGLE. *Google Sites*. Disponível em: <https://workspace.google.com/intl/pt-BR/products/sites//>. Acesso em: 16 dez. 2021.

[28] PYTORCH - TORCHVISION. Disponível em: <https://pytorch.org/vision/stable/index.html>. Acesso em: 16 dez. 2021.

[29] GOOGLE. *Static Maps API*. Disponível em: <https://developers.google.com/maps/documentation/maps-static/overview>. Acesso em: 14 dez. 2021.

[30] *DMSP OLS: Nighttime Lights Time Series Version 4, Defense Meteorological Program Operational Linescan System*. Disponível em: <https://developers.google.com/earth-engine/datasets/catalog/NOAA_DMSP-OLS_NIGHTTIME_LIGHTS>. Acesso em: 14 dez. 2021.

[31] *MOD16A2: MODIS Global Terrestrial Evapotranspiration 8-Day Global 1km*. Disponível em:<https://developers.google.com/earth-engine/datasets/catalog/MODIS_NTSG_MOD16A2_105>. Acesso em: 14 dez. 2021.

[32] SIMONYAN, K.; ZISSERMAN, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition* (2015). Disponível em: <https://arxiv.org/abs/1409.1556>. Acesso em 19 dez. 2021.

[33] GOOGLE. *FeatureCollection Overview*. Disponível em: <https://developers.google.com/earth-engine/guides/feature_collections>. Acesso em: 14 dez. 2021.

[34] GOOGLE. *FAO GAUL: Global Administrative Unit Layers 2015, Second-Level Administrative Units*. Disponível em: <https://developers.google.com/earth-engine/datasets/catalog/FAO_GAUL_2015_level2>. Acesso em: 14 dez. 2021.

[35] GOOGLE. *Colaboratory - Perguntas Frequentes.* Disponível em: <https://research.google.com/colaboratory/intl/pt-BR/faq.html> Acesso em: 19 dez. 2021.

[36] INPE. *NEXUS - Banco de dados - Descrição Completa dos Indicadores do Projeto Nexus*. Disponível em: <http://nexus.ccst.inpe.br/banco-de-dados/>. Acesso em: 19 dez. 2021.

Apêndice A - Resultados da validação cruzada totalmente randômica para indicadores socioeconômicos

Tabela 1: Coeficientes de determinação (R^2) obtidos para indicadores socioeconômicos no experimento de validação cruzada totalmente randômica.

Nome do indicador socioeconômico	**R^2**
Domicílios com renda maior que um salário mínimo	0.347
Isolamento da população considerando a distância a corpos hídricos e estradas	0.328
PIB per capita	0.197
Domicílios Inadequados	0.101
Ocorrência de doenças veiculadas com fonte hídrica	0.073
Taxa de mortalidade em menores de 5 anos de idade	0.013
Proporção de Cadastramento de pessoas em serviços básicos de saúde	0.002
Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)	-0.397
Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população	-63.076

Fonte: Compilação do autor.

Apêndice B - Resultados da validação cruzada totalmente randômica para indicadores ambientais

Tabela 2: Coeficientes de determinação (R^2) obtidos para indicadores ambientais no experimento de validação cruzada totalmente randômica.

Nome do indicador ambiental	Categoria indicador	R^2
Pastagens degradadas	Degradação da Terra	0.553
Existência e proteção de recursos hídricos nos estabelecimentos agropecuários	Produção Agrícola Sustentável	0.512
Produtividade Pecuária e Leiteira	Produção Agrícola Sustentável	0.237
Uniformidade de receitas ou atividades do produtor rural	Produção Agrícola Sustentável	0.151
Evolução dos Sistemas Agroflorestais em estabelecimentos rurais	Conservação Florestal e Biodiversidade	0.141
Ineficácia de maquinários à produtividade agrícola	Produção Agrícola Sustentável	0.134
Alternativas ao abastecimento de água	Recursos Hídricos	0.129
Alternativas ao abastecimento de água	Recursos Hídricos	0.129
Proporção do uso de agrotóxico	Produção Agrícola Sustentável	0.096
Abrangência e Diversidade do PRONAF	Produção Agrícola Sustentável	0.054
Abrangência do Programa Cisternas	Produção Agrícola Sustentável	0.047
Produtividade agrícola de alimentos básicos	Produção Agrícola Sustentável	-36.293

Fonte: Compilação do autor.

Apêndice C - Resultados da validação cruzada randomizada por estado para indicadores socioeconômicos

Tabela 3: Coeficiente de determinação (R^2) obtidos para indicadores ambientais no experimento de validação cruzada randomizada por estado.

Nome do indicador socioeconômico	Estado	R^2
Isolamento da população considerando a distância a corpos hídricos e estradas	PB	0.526
Isolamento da população considerando a distância a corpos hídricos e estradas	RN	0.462
PIB per capita	RN	0.349
Isolamento da população considerando a distância a corpos hídricos e estradas	AL	0.341
Domicílios com renda maior que um salário mínimo	RN	0.318
Isolamento da população considerando a distância a corpos hídricos e estradas	SE	0.307
Ocorrência de doenças veiculadas com fonte hídrica	RN	0.222
Domicílios com renda maior que um salário mínimo	AL	0.202
Domicílios Inadequados	PB	0.163
Domicílios com renda maior que um salário mínimo	SE	0.111
Domicílios com renda maior que um salário mínimo	PB	0.057
Ocorrência de doenças veiculadas com fonte hídrica	PB	0.052
Domicílios Inadequados	RN	0.032
PIB per capita	PB	-0.055
Taxa de mortalidade em menores de 5 anos de idade	AL	-0.091
Taxa de mortalidade em menores de 5 anos de idade	PB	-0.111
Domicílios Inadequados	AL	-0.249
Domicílios Inadequados	SE	-0.309
Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população	SE	-0.389
PIB per capita	AL	-0.811
Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)	AL	-0.934
Proporção de Cadastramento de pessoas em serviços básicos de saúde	AL	-1.313
Proporção de Cadastramento de pessoas em serviços básicos de saúde	PB	-1.520
Proporção de Cadastramento de pessoas em serviços básicos de saúde	SE	-1.541
Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população	AL	-1.621
Taxa de mortalidade em menores de 5 anos de idade	RN	-1.810

Proporção de Cadastramento de pessoas em serviços básicos de saúde	RN	-2.020
Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)	RN	-2.057
Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)	PB	-2.171
Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)	SE	-2.638
Ocorrência de doenças veiculadas com fonte hídrica	AL	-3.435
Taxa de mortalidade em menores de 5 anos de idade	SE	-4.453
Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população	RN	-6.438
PIB per capita	SE	-14.696
Ocorrência de doenças veiculadas com fonte hídrica	SE	-27.387
Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população	PB	-74.106

Fonte: Compilação do autor.

Apêndice D - Resultados da validação cruzada randomizada por estado para indicadores ambientais

Tabela 4: Coeficiente de determinação (R^2) obtidos para indicadores ambientais no experimento de validação cruzada randomizada por estado.

Nome do indicador ambiental	Categoria indicador	Estado	R^2
Pastagens degradadas	Degradação da Terra	SE	0.620
Pastagens degradadas	Degradação da Terra	PB	0.589
Pastagens degradadas	Degradação da Terra	RN	0.520
Existência e proteção de recursos hídricos nos estabelecimentos agropecuários	Produção Agrícola Sustentável	PB	0.501
Existência e proteção de recursos hídricos nos estabelecimentos agropecuários	Produção Agrícola Sustentável	RN	0.439
Pastagens degradadas	Degradação da Terra	AL	0.433
Produtividade agrícola de alimentos básicos	Produção Agrícola Sustentável	PB	0.400
Uniformidade de receitas ou atividades do produtor rural	Produção Agrícola Sustentável	PB	0.287
Produtividade Pecuária e Leiteira	Produção Agrícola Sustentável	RN	0.246
Produtividade Pecuária e Leiteira	Produção Agrícola Sustentável	PB	0.217
Proporção do uso de agrotóxico	Produção Agrícola Sustentável	AL	0.188
Uniformidade de receitas ou atividades do produtor rural	Produção Agrícola Sustentável	AL	0.183
Abrangência do Programa Cisternas	Produção Agrícola Sustentável	AL	0.173
Evolução dos Sistemas Agroflorestais em estabelecimentos rurais	Conservação Florestal e Biodiversidade	RN	0.125
Abrangência do Programa Cisternas	Produção Agrícola Sustentável	PB	0.098
Ineficácia de maquinários à produtividade agrícola	Produção Agrícola Sustentável	PB	0.086
Proporção do uso de agrotóxico	Produção Agrícola Sustentável	RN	0.081
Abrangência e Diversidade do PRONAF	Produção Agrícola	PB	0.076

	Sustentável		
Alternativas ao abastecimento de água	Recursos Hídricos	AL	0.075
Alternativas ao abastecimento de água	Recursos Hídricos	AL	0.075
Proporção do uso de agrotóxico	Produção Agrícola Sustentável	PB	0.058
Uniformidade de receitas ou atividades do produtor rural	Produção Agrícola Sustentável	SE	0.035
Uniformidade de receitas ou atividades do produtor rural	Produção Agrícola Sustentável	RN	0.017
Evolução dos Sistemas Agroflorestais em estabelecimentos rurais	Conservação Florestal e Biodiversidade	PB	0.015
Abrangência e Diversidade do PRONAF	Produção Agrícola Sustentável	RN	0.001
Ineficácia de maquinários à produtividade agrícola	Produção Agrícola Sustentável	RN	-0.008
Alternativas ao abastecimento de água	Recursos Hídricos	RN	-0.031
Alternativas ao abastecimento de água	Recursos Hídricos	RN	-0.031
Abrangência e Diversidade do PRONAF	Produção Agrícola Sustentável	SE	-0.048
Alternativas ao abastecimento de água	Recursos Hídricos	PB	-0.049
Alternativas ao abastecimento de água	Recursos Hídricos	PB	-0.049
Produtividade Pecuária e Leiteira	Produção Agrícola Sustentável	SE	-0.109
Existência e proteção de recursos hídricos nos estabelecimentos agropecuários	Produção Agrícola Sustentável	AL	-0.285
Produtividade agrícola de alimentos básicos	Produção Agrícola Sustentável	AL	-0.427
Abrangência do Programa Cisternas	Produção Agrícola Sustentável	RN	-0.657
Existência e proteção de recursos hídricos nos estabelecimentos agropecuários	Produção Agrícola Sustentável	SE	-1.039
Produtividade agrícola de alimentos básicos	Produção Agrícola Sustentável	SE	-1.153
Ineficácia de maquinários à produtividade agrícola	Produção Agrícola Sustentável	AL	-1.228
Produtividade Pecuária e Leiteira	Produção Agrícola Sustentável	AL	-1.398
Abrangência do Programa Cisternas	Produção Agrícola Sustentável	SE	-1.514
Abrangência e Diversidade do PRONAF	Produção Agrícola Sustentável	AL	-1.948

Alternativas ao abastecimento de água	Recursos Hídricos	SE	-2.243
Alternativas ao abastecimento de água	Recursos Hídricos	SE	-2.243
Evolução dos Sistemas Agroflorestais em estabelecimentos rurais	Conservação Florestal e Biodiversidade	SE	-4.847
Evolução dos Sistemas Agroflorestais em estabelecimentos rurais	Conservação Florestal e Biodiversidade	AL	-4.931
Proporção do uso de agrotóxico	Produção Agrícola Sustentável	SE	-6.218
Ineficácia de maquinários à produtividade agrícola	Produção Agrícola Sustentável	SE	-9.498
Produtividade agrícola de alimentos básicos	Produção Agrícola Sustentável	RN	-270.160

Fonte: Compilação do autor.

Apêndice E - Resultados da validação transregional para indicadores socioeconômicos

Tabela 5: Coeficiente de determinação (R^2) obtidos para indicadores ambientais no experimento de validação transregional.

Nome do indicador socioeconômico	Estado	R^2
Domicílios com renda maior que um salário mínimo	SE	0.170
Domicílios com renda maior que um salário mínimo	AL	0.077
Domicílios Inadequados	PB	0.065
Domicílios com renda maior que um salário mínimo	RN	0.010
Domicílios com renda maior que um salário mínimo	PB	0.001
PIB per capita	RN	-0.003
Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população	PB	-0.008
Isolamento da população considerando a distância a corpos hídricos e estradas	SE	-0.010
Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)	PB	-0.016
Domicílios Inadequados	RN	-0.026
Proporção de Cadastramento de pessoas em serviços básicos de saúde	AL	-0.027
Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)	RN	-0.049
Taxa de mortalidade em menores de 5 anos de idade	SE	-0.054
Ocorrência de doenças veiculadas com fonte hídrica	AL	-0.054
Taxa de mortalidade em menores de 5 anos de idade	PB	-0.067
Proporção de Cadastramento de pessoas em serviços básicos de saúde	RN	-0.101
Proporção de Cadastramento de pessoas em serviços básicos de saúde	PB	-0.114
Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)	SE	-0.177
Domicílios Inadequados	SE	-0.213
PIB per capita	SE	-0.244
Proporção de Cadastramento de pessoas em serviços básicos de saúde	SE	-0.254
Ocorrência de doenças veiculadas com fonte hídrica	RN	-0.298
Taxa de mortalidade em menores de 5 anos de idade	AL	-0.319
Isolamento da população considerando a distância a corpos hídricos e estradas	PB	-0.353
Isolamento da população considerando a distância a corpos hídricos e	RN	-0.449

estradas		
Taxa de mortalidade em menores de 5 anos de idade	RN	-0.465
Ocorrência de doenças veículadas com fonte hídrica	PB	-0.481
PIB per capita	AL	-0.871
Ocorrência de doenças veículadas com fonte hídrica	SE	-1.276
Domicílios Inadequados	AL	-1.524
Isolamento da população considerando a distância a corpos hídricos e estradas	AL	-3.096
PIB per capita	PB	-4.154
Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)	AL	-5.853
Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população	RN	-145.232
Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população	SE	-1373.930
Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população	AL	-14451.691

Fonte: Compilação do autor.

Apêndice F - Resultados da validação transregional para indicadores ambientais

Tabela 6: Coeficiente de determinação (R^2) obtidos para indicadores ambientais, no experimento de validação transregional.

Nome do indicador ambiental	Categoria indicador	Estado	R^2
Existência e proteção de recursos hídricos nos estabelecimentos agropecuários	Produção Agrícola Sustentável	PB	0.405
Pastagens degradadas	Degradação da Terra	SE	0.309
Alternativas ao abastecimento de água	Recursos Hídricos	AL	0.203
Alternativas ao abastecimento de água	Recursos Hídricos	AL	0.203
Abrangência do Programa Cisternas	Produção Agrícola Sustentável	AL	0.157
Ineficácia de maquinários à produtividade agrícola	Produção Agrícola Sustentável	AL	0.121
Ineficácia de maquinários à produtividade agrícola	Produção Agrícola Sustentável	RN	0.066
Abrangência e Diversidade do PRONAF	Produção Agrícola Sustentável	AL	0.065
Produtividade Pecuária e Leiteira	Produção Agrícola Sustentável	PB	0.056
Existência e proteção de recursos hídricos nos estabelecimentos agropecuários	Produção Agrícola Sustentável	RN	0.048
Pastagens degradadas	Degradação da Terra	RN	0.029
Produtividade agrícola de alimentos básicos	Produção Agrícola Sustentável	RN	0.015
Ineficácia de maquinários à produtividade agrícola	Produção Agrícola Sustentável	SE	0.000
Produtividade Pecuária e Leiteira	Produção Agrícola Sustentável	RN	-0.001
Proporção do uso de agrotóxico	Produção Agrícola Sustentável	PB	-0.013
Uniformidade de receitas ou atividades do produtor rural	Produção Agrícola Sustentável	PB	-0.044
Uniformidade de receitas ou atividades do produtor rural	Produção Agrícola Sustentável	SE	-0.049
Abrangência do Programa Cisternas	Produção Agrícola Sustentável	RN	-0.079
Produtividade Pecuária e Leiteira	Produção Agrícola	SE	-0.118

	Sustentável		
Abrangência e Diversidade do PRONAF	Produção Agrícola Sustentável	SE	-0.120
Abrangência do Programa Cisternas	Produção Agrícola Sustentável	SE	-0.121
Proporção do uso de agrotóxico	Produção Agrícola Sustentável	RN	-0.151
Proporção do uso de agrotóxico	Produção Agrícola Sustentável	AL	-0.229
Abrangência e Diversidade do PRONAF	Produção Agrícola Sustentável	RN	-0.235
Alternativas ao abastecimento de água	Recursos Hídricos	PB	-0.290
Alternativas ao abastecimento de água	Recursos Hídricos	PB	-0.290
Uniformidade de receitas ou atividades do produtor rural	Produção Agrícola Sustentável	AL	-0.299
Evolução dos Sistemas Agroflorestais em estabelecimentos rurais	Conservação Florestal e Biodiversidade	AL	-0.319
Existência e proteção de recursos hídricos nos estabelecimentos agropecuários	Produção Agrícola Sustentável	AL	-0.348
Ineficácia de maquinários à produtividade agrícola	Produção Agrícola Sustentável	PB	-0.350
Existência e proteção de recursos hídricos nos estabelecimentos agropecuários	Produção Agrícola Sustentável	SE	-0.436
Proporção do uso de agrotóxico	Produção Agrícola Sustentável	SE	-0.526
Evolução dos Sistemas Agroflorestais em estabelecimentos rurais	Conservação Florestal e Biodiversidade	SE	-0.546
Abrangência do Programa Cisternas	Produção Agrícola Sustentável	PB	-0.599
Produtividade Pecuária e Leiteira	Produção Agrícola Sustentável	AL	-0.842
Uniformidade de receitas ou atividades do produtor rural	Produção Agrícola Sustentável	RN	-0.847
Abrangência e Diversidade do PRONAF	Produção Agrícola Sustentável	PB	-0.860
Alternativas ao abastecimento de água	Recursos Hídricos	RN	-1.007
Alternativas ao abastecimento de água	Recursos Hídricos	RN	-1.007
Alternativas ao abastecimento de água	Recursos Hídricos	SE	-1.074
Alternativas ao abastecimento de água	Recursos Hídricos	SE	-1.074
Evolução dos Sistemas Agroflorestais em estabelecimentos rurais	Conservação Florestal e Biodiversidade	RN	-1.082

Pastagens degradadas	Degradação da Terra	AL	-1.149
Pastagens degradadas	Degradação da Terra	PB	-1.256
Evolução dos Sistemas Agroflorestais em estabelecimentos rurais	Conservação Florestal e Biodiversidade	PB	-2.270
Produtividade agrícola de alimentos básicos	Produção Agrícola Sustentável	AL	-6.344
Produtividade agrícola de alimentos básicos	Produção Agrícola Sustentável	SE	-27.064
Produtividade agrícola de alimentos básicos	Produção Agrícola Sustentável	PB	-785.159

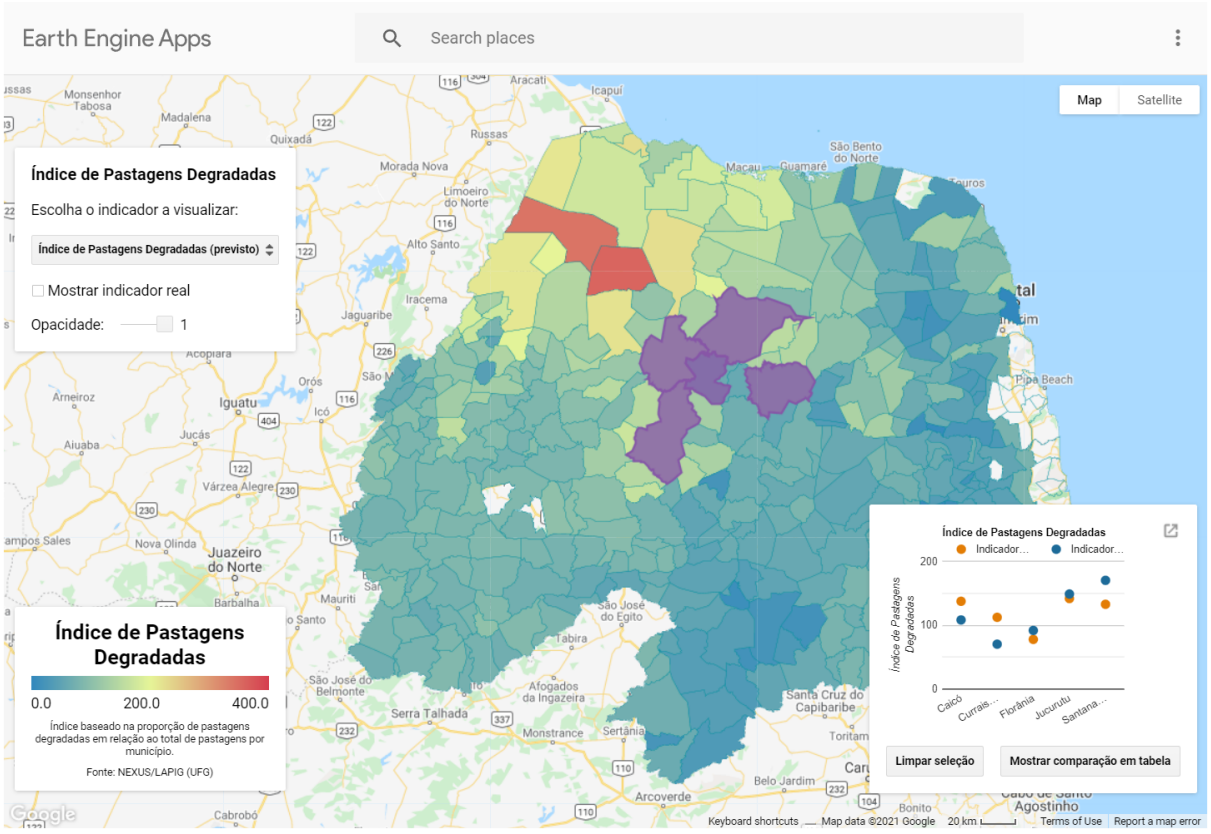
Fonte: Compilação do autor.

Apêndice G - Imagens da plataforma Visualiza Nexus-Parsec

Figura 40: Tela com informações sobre o projeto, na plataforma *web* Visualiza NEXUS-PARSEC.

Fonte: Compilação do autor.

Figura 41: Mapa de indicadores previstos, com municípios selecionados para a análise.



Fonte: Compilação do autor.

Figura 42: Tela com informações sobre a metodologia do trabalho, na plataforma *web* Visualiza NEXUS-PARSEC.

Fonte: Compilação do autor.

Anexo A - Descrição dos Indicadores Socioeconômicos do Projeto NEXUS

Tabela 7: Tabela com o nome e a descrição dos indicadores socioeconômicos fornecidos pelo Projeto Nexus.

Indicadores Socioeconômicos	
Nome	**Descrição**
Ocorrência de doenças veiculadas com fonte hídrica	O percentual de ocorrência de doenças foi obtido pela razão entre a soma dos casos confirmados notificados no sistema de informação de agravos de notificação, referente às doenças: Difteria, Cólera, Febre Tifoide, Hepatite A, Dengue, Febre Amarela, Leishmaniose Tegumentar, Leishmaniose Visceral, Malária, Doença de Chagas, Esquistossomose, Leptospirose, Peste e Hantavirose, disponibilizados em Portal da Saúde pelo departamento de informática do Sistema Único de Saúde do Brasil (DATASUS) entre 2010 e 2015, e a média da população residente ou estimada (disponibilizada em Censo Demográfico e Estimativas de População pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para o mesmo período).
Domicílios com renda maior que um salário mínimo	O indicador é resultante da parcela complementar ao percentual de domicílios particulares permanentes com rendimento nominal mensal domiciliar per capita de até 1 (um) salário mínimo. A Informação dos domicílios com rendimento per capita de até 1 salário mínimo foi obtida no Censo Demográfico disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para o ano de 2010. https://sidra.ibge.gov.br/tabela/3562
Isolamento da população considerando a distância a corpos hídricos e estradas	Isolamento geográfico da população (geralmente rural) no município, considerando a temática de demografia. Indicador resultante da distância as estradas e/ou corpos d'água ponderado pelo coeficiente médio de acessibilidade do município a outro de maior hierarquia em infraestrutura, considerando sua importância relativa à problemática de seca. O dado de distância às estradas e/ou corpos d'água foi calculado a partir da distância euclidiana média entre as unidades visitadas do Censo 2010 do IBGE em relação a estradas de terra e asfaltadas e corpos hídricos superficiais perenes da base cartográfica 1:250.000 do Instituto Brasileiro de Geografia e Estatística (IBGE). Em seguida, esta informação foi espacialmente agregada para nível municipal e a proporção zonal destas distâncias médias foram calculadas para cada município. Por fim, utilizou-se o coeficiente médio de acessibilidade derivado do dado de Acessibilidade Geográfica desenvolvido pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para o ano de 2018 para ponderar a informação de isolamento da população. Para efeitos de compreensão o dado resultante é adimensional e foi aproximado para um função logarítmica natural, portanto, quanto maior o valor mais isolada é a

	população e quanto menor o valor do indicador menos a população está distante de acesso aos recursos hídricos. Fonte: https://geoftp.ibge.gov.br/cartas_e_mapas/bases_cartograficas_continuas/bc250/versao2015/Shapefile/ https://www.ibge.gov.br/geociencias/organizacao-do-territorio/tipologias-do-territorio/26253-acessibilidade-geografica.html?=&t=downloadshttps://geoftp.ibge.gov.br/cartas_e_mapas/bases_cartograficas_continuas/bc250/versao2015/Shapefile/
Taxa de mortalidade em menores de 5 anos de idade	Número de óbitos de menores de cinco anos de idade, por mil nascidos vivos, na população residente em determinado espaço geográfico, no ano considerado. Estima o risco de morte dos nascidos vivos durante os cinco primeiros anos de vida, refletindo condições de desenvolvimento socioeconômico e a infra-estrutura ambiental precários, que condicionam a desnutrição infantil e as infecções associadas. O acesso e a qualidade dos recursos disponíveis para atenção à saúde materno-infantil são também determinantes da mortalidade nesse grupo etário. É influenciada pela composição da mortalidade no primeiro ano de vida (mortalidade infantil), amplificando o impacto das causas pós-neonatais, a que estão expostas também as crianças entre 1 e 4 anos de idade. Porém, taxas reduzidas podem estar encobrindo más condições de vida em segmentos sociais específicos. TM < 5 anos anual = Número de óbitos em menores de 1 ano e entre 1 a 4 anos ocorridos em determinado ano / Número de nascimentos totais de mães residentes no mesmo ano a) Mortalidade - 1996 a 2019, pela CID-10. Mortalidade Geral. Ministério da Saúde. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10br.def> b) Estatísticas vitais. Nascidos Vivos - 1994 a 2019. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinasc/cnv/nvbr.def> TM média < 5 anos = (tx anual 2010) + (tx anual 2011) + ... + (tx anual 2019) / 10
Taxa de mortalidade fetal (TMF) - óbitos fetais (Óbitos fetais)	Número de óbitos fetais por mil nascimentos totais, na população residente em determinado espaço geográfico, no ano considerado. A taxa de mortalidade fetal (TMF) é considerada um dos melhores indicadores de qualidade de assistência prestada à gestante e ao parto. Estima o risco de um feto nascer sem qualquer sinal de vida. De maneira geral, reflete a ocorrência de fatores vinculados à gestação e ao parto, entre eles o peso ao nascer, bem como as condições de acesso a serviços de saúde e a qualidade da assistência pré-natal e ao parto. A TMF permite analisar variações populacionais, geográficas e temporais da mortalidade fetal, identificando situações de desigualdade, bem como subsidiar a avaliação da qualidade da assistência prestada à gestação e ao parto. Também pode contribuir na avaliação dos níveis de saúde e de desenvolvimento socioeconômico da população, prestando-se para comparações nacionais e internacionais.

	a) Mortalidade - 1996 a 2019, pela CID-10. Ministério da Saúde. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/fet10br.def> b) Estatísticas vitais. Nascidos Vivos - 1994 a 2019. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinasc/cnv/nvbr.def>
Proporção de Cadastramento de pessoas em serviços básicos de saúde	Indica a quantidade de pessoas com potencial em obter o cuidado (ainda que não efetivada), sendo apenas um dos componentes da avaliação dos serviços de saúde. Considerando o objetivo 3.8 dos ODS - que busca atingir a cobertura universal de saúde - a cobertura da atenção básica pode ser um proxy do cadastramento de pessoas no serviço básico de saúde. Disponível em: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?pacto/2015/cnv/coapmunbr.def>
PIB per capita	PIB municipal per capita, calculado pela média entre os anos de 2010 a 2018
Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população	Razão da taxa de expansão de terras antropizadas pela taxa de crescimento da população entre 2010 e 2019
Domicílios Inadequados	Nível de qualidade estrutural dos domicílios que pode afetar a saúde da população. Este indicador foi obtido pela razão entre a soma do número de domicílios particulares permanentes com revestimento externo inadequado (domicílios com madeira aparelhada, com taipa revestida e com revestimento não durável) e do número de domicílios particulares permanentes com esgotamento inadequado (domicílios que tinham banheiro com outro tipo de escoadouro, que tinham sanitário com outro tipo de escoadouro e que não tinham banheiro ou sanitário) pelo total de domicílios particulares permanentes, disponibilizados em Censo Demográfico pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para 2010.

Fonte: INPE (2021) [36]

Anexo B - Descrição dos Indicadores Ambientais do NEXUS

Tabela 8: Tabela com o nome e a descrição dos indicadores ambientais fornecidos pelo Projeto Nexus.

Indicadores Ambientais	
Nome	**Descrição**
Evolução dos Sistemas Agroflorestais em estabelecimentos rurais	Evolução da proporção de estabelecimentos agroflorestais em relação ao total de estabelecimentos rurais, calculado segundo os Censos segundo os Censos Agropecuários de 2006 e 2017. (https://sidra.ibge.gov.br/tabela/3355 e https://sidra.ibge.gov.br/tabela/6883)
Pastagens degradadas	O indicador mede a proporção de pastagens degradadas em relação ao total de pastagens por município. De acordo com o LAPIG (UFG), o indicativo da qualidade de pastagens foi produzido por meio de uma análise de tendências sobre anomalias acumuladas pixel a pixel e para o período de 2011 a 2016. Esta análise utilizou dados satelitários (NDVI/MOD13Q1) e avaliou perdas ou ganhos em produtividade. As áreas com tendência significativas de perda em produtividade ($p < 0.05$) foram consideradas com indícios de degradação", em relação às áreas de pastagem do Brasil, envolvem uma "série histórica das áreas de pastagens do Brasil, produzida para toda a extensão territorial brasileira, para os últimos últimos 33 anos (1985 a 2017), no âmbito do projeto do Map Biomas". O total de pastagem compreendeu este histórico foi obtido pelo dado de “área de pastagem por município no Brasil de 2018” do LAPIG, já a pastagem degradada pelo dado de “pastagem degradada de 2018” que possui 4 classes, a classe 1 representando áreas não degradadas e 2, 3 e 4 os níveis de degradação – leve, moderada e severa). As classes 2, 3 e 4 foram somadas para obter a pastagem degradada por município. Houveram poucos casos em que o dado de área de pastagem total estava zerado na planilha do LAPIG, mas tinha valor na planilha de pastagem degradada. Nesses casos, considerou-se como pastagem total a soma das classes 1, 2, 3 e 4 da “pastagem degradada de 2018”. Por fim, foi calculado o percentual. Fonte: https://pastagem.org/atlas/map
Produtividade agrícola de alimentos básicos	Este indicador quantifica a eficiência da produção municipal dos alimentos básicos como feijão, mandioca, arroz e milho. O cálculo é dado pela razão da quantidade média colhida de feijão, mandioca, arroz e milho (em quilogramas) pelo total de área plantada (em hectare). Os dados foram obtidos no Censo Agropecuário disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para 2017.

Proporção do uso de agrotóxico	O indicador corresponde à proporção dos estabelecimentos agropecuários que utilizam (ou já utilizaram) agrotóxico. Este indicador é dado pela razão entre o número de estabelecimentos agropecuários que utilizam (ou já utilizaram) agrotóxico, e o número total de estabelecimentos, multiplicado pela área total dos estabelecimentos. Todos os dados estão disponíveis em nível municipal pelo Censo Agropecuário do Instituto Brasileiro de Geografia e Estatística (IBGE) para 2017. Fonte: https://sidra.ibge.gov.br/tabela/6852
Abrangência e Diversidade do PRONAF	O Programa de Fortalecimento da Agricultura Familiar (PRONAF) é um financiamento que visa estimular a geração de renda e melhorar o uso da mão de obra familiar, por meio de atividades e serviços rurais agropecuários e não agropecuários desenvolvidos em estabelecimento rural ou em áreas comunitárias próximas. O indicador reflete o nível de abrangência e a diversidade de categorias do Programa em relação à distribuição fundiária de cada município. O indicador foi calculado a partir da razão do montante repassado pelo PRONAF e a quantidade de estabelecimentos agropecuários multiplicado pelo Índice de Simpson gerado a partir do valor financiado nas linhas de crédito do Programa para cada município. Os subprogramas considerados foram: Agroecologia; Agroindústria; Agroindústria (investimento); Cotas Partes; Custeio; Energia Renovável e Sustentabilidade Ambiental (ECO); FGPP-RES.4801, Art. 2; Floresta; Jovem; Mais Alimentos; Microcrédito; Mulher; Produtivo Orientado; Reforma Agrária; Reforma Agrária (microcrédito); e Semiárido. Os dados sobre os valores repassados pelo PRONAF foram obtidos na Matriz de Dados do Crédito Rural (Contratações) disponibilizada pelo Banco Central do Brasil (BCB), entre os anos de 2015 a 2017. Os dados correspondentes aos números de estabelecimentos agropecuários foram incluídos no Censo Agropecuário, disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para o ano 2017. Fonte: https://sidra.ibge.gov.br/Tabela/6771 e BCB (2020)
Uniformidade de receitas ou atividades do produtor rural	Uniformidade da produção local baseada nos grupos de atividades econômicas desenvolvidas nos estabelecimentos agropecuários. Quanto maior a uniformidade menor a capacidade de adaptação do produtor às externalidades. O indicador considera a parcela complementar do valor calculado pelo índice de Diversificação de Simpson, baseado no número de estabelecimentos agropecuários e suas respectivas áreas, que possuem as seguintes atividades econômicas: Produção de lavouras temporárias, Horticultura e floricultura, Produção de lavouras permanentes, Produção de sementes e mudas certificadas, Pecuária e criação de outros animais, Produção florestal - florestas plantadas, Produção florestal - florestas nativas, Pesca e Aquicultura. Os dados em nível municipal disponíveis pelo Censo Agropecuário disponibilizado pelo pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para o ano de 2017. Fonte: https://sidra.ibge.gov.br/tabela/6879.0
Existência e proteção de recursos	Proporção de registros de estabelecimentos agropecuários com nascentes nascentes e rios/riachos protegidos por mata, poços e cisternas. Fonte: Censo Agropecuário 2017 (IBGE)

hídricos nos estabelecimentos agropecuários	
Ineficácia de maquinários à produtividade agrícola	Este indicador quantifica a ineficácia da utilização de tratores, implementos e máquinas relacionada à produtividade agrícola, ponderada pela distribuição fundiária. Este indicador é dado pela subtração de 1 menos a "razão entre o total de tratores, semeadoras/plantadeiras e colheitadeiras e a produtividade agrícola, multiplicado pela razão entre o número de estabelecimentos agropecuários e a área total destes estabelecimentos". A produtividade agrícola, em quilograma por hectare, é resultante da razão entre a quantidade produzida nas lavouras temporárias e área colhida nas lavouras temporárias, com exceção da cultura de abacaxi que apresenta a quantidade produzida na unidade de “mil frutos”. Os dados estão disponíveis em nível municipal pelo Censo Agropecuário 2017 do Instituto Brasileiro de Geografia e Estatística (IBGE). Fonte: https://sidra.ibge.gov.br/tabela/6878 e https://sidra.ibge.gov.br/tabela/6957 e https://sidra.ibge.gov.br/tabela/6641 0
Abrangência do Programa Cisternas	O Programa Nacional de Apoio à Captação de Água de Chuva e outras Tecnologias Sociais (Programa Cisternas) tem como objetivo a promoção do acesso à água para o consumo humano e para a produção de alimentos por meio da implementação de tecnologias sociais simples e de baixo custo. O indicador reflete o nível de acesso à água para produção por meio do Programa Cisternas de cada município e foi obtido pelo nível de acesso à água para produção considerando a temática de manutenção da produção agropecuária. O cálculo foi baseado na razão entre o número de cisternas para produção entregues pelo número de domicílios rurais de cada município. Dados em nível municipal, obtidos em Sistema de Informações Gerenciais (SIG Cisternas), disponibilizado pela Secretaria Especial do Desenvolvimento Social/ Ministério da Cidadania para o ano de 2016. Fonte: EMBRAPA, 20160
Produtividade Pecuária e Leiteira	Este indicador é composto pela produtividade de rebanhos e leite em pastagens em boas condições. O cálculo deste indicador foi realizado pela média entre a produtividade de rebanhos (R) e leite (L). A produtividade de rebanhos foi calculada pela razão entre a quantidade de animais de pastoreio (número total de cabeças de caprinos, ovinos e bovinos) e a área de pastagem plantada no município (em hectare), multiplicada pela porcentagem de pastagens consideradas em boas condições. A produtividade de leite foi calculada pela razão entre o total de produção de leite (mil litros) entre vacas, ovelhas e cabras e o número de cabeças desses animais, multiplicada pela porcentagem de pastagens consideradas em boas condições. O cálculo final foi realizado pela média entre os valores normalizados das duas produtividades consideradas - R e L. Dados em nível municipal obtidos do Censo Agropecuário disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) para o ano de 2017. Fonte: https://sidra.ibge.gov.br/tabela/6908 , sidra.ibge.gov.br/Tabela/6782, https://sidra.ibge.gov.br/Tabela/6720 e https://sidra.ibge.gov.br/Tabela/6719

Alternativas ao abastecimento de água	Indicador resultante do percentual da soma de domicílios com poços e nascentes fora da propriedade e carro pipa em relação ao somatório de domicílios sem rede de abastecimento geral, com poços ou nascentes na propriedade e cisternas. Os dados de número de domicílios com forma de abastecimento de água por “rede geral”, “Poço ou nascente na propriedade”, “Poço ou nascente fora da propriedade”, “carro-pipa” e “Água da chuva armazenada em cisterna” são obtidos em dados do censo demográfico do IBGE para o ano de 2010. Fonte: https: sidra.ibge.gov.br/tabela/3217

Fonte: INPE (2021) [36]