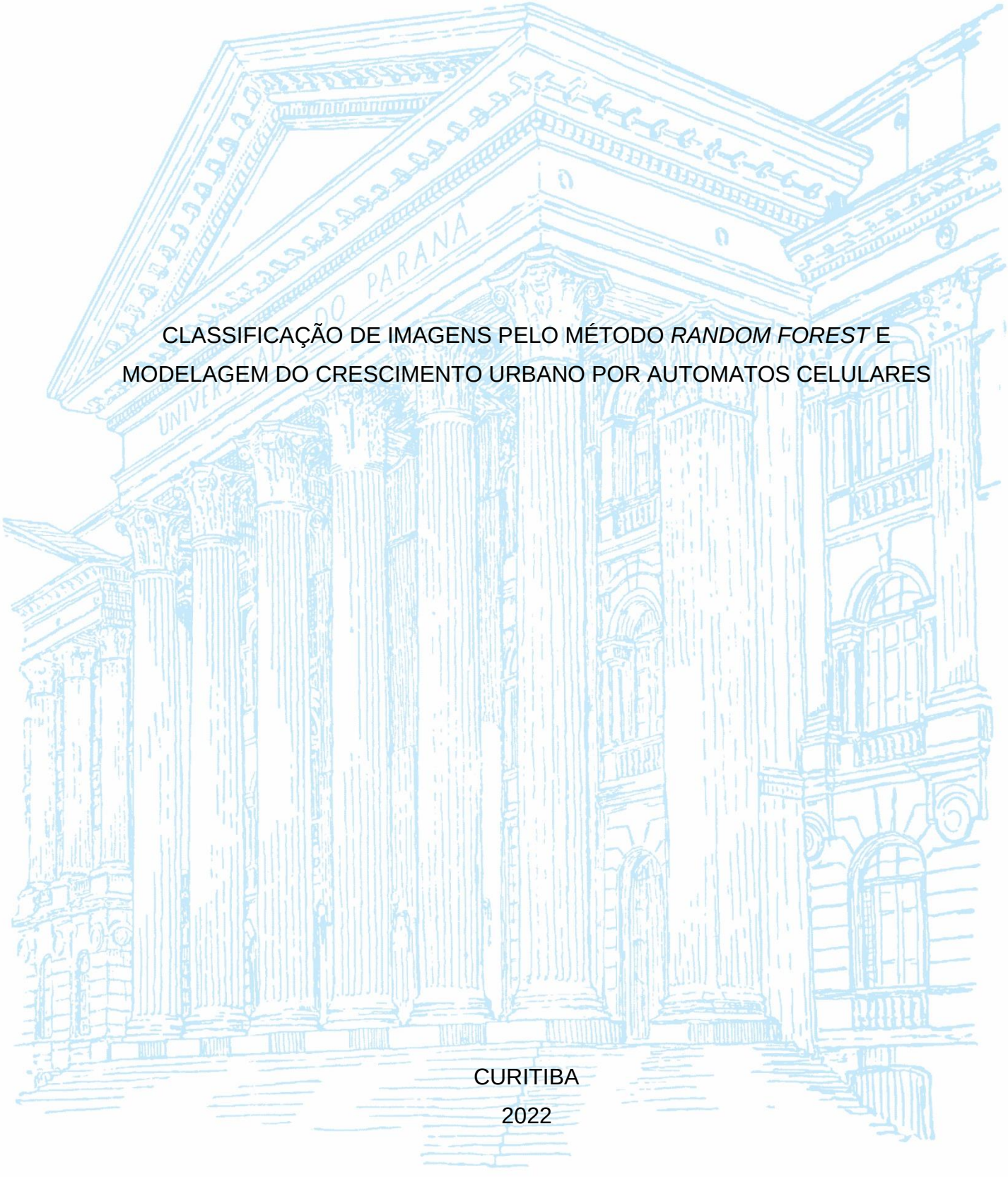
UNIVERSIDADE FEDERAL DO PARANÁ

GABRIEL CAMPIGOTO



CLASSIFICAÇÃO DE IMAGENS PELO MÉTODO *RANDOM FOREST* E MODELAGEM DO CRESCIMENTO URBANO POR AUTOMATOS CELULARES

CURITIBA

2022

GABRIEL CAMPIGOTO

CLASSIFICAÇÃO DE IMAGENS PELO MÉTODO *RANDOM FOREST* E MODELAGEM DO CRESCIMENTO URBANO POR AUTOMATOS CELULARES

Projeto final apresentada ao curso de Graduação em Engenharia Cartográfica e de Agrimensura, Setor de Ciências da Terra, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Engenharia Cartográfica e de Agrimensura.

Orientador Prof. Dr. Hideo Araki

CURITIBA

2022

AGRADECIMENTOS

Agradeço aos meus pais, João e Elenita, pelo confiança e suporte que me deram através de todos os anos.

Agradeço ao Henrique Peschl, por me fornecer dados que me possibilitaram fazer esse projeto final.

Agradeço também ao meu orientador Hideo Araki e a professora Silvana Camboim, pelo conhecimento e dicas fornecidos a mim para que eu conseguisse prosseguir nesse trabalho.

Escolhe um trabalho de que gostes e não terás que trabalhar nem um dia na tua vida (Confúcio)

RESUMO

A análise multitemporal de imagens de satélite permite verificar as mudanças de um fenômeno através dos anos. Neste trabalho se propôs a utilizar um método de *machine learning* chamado de *random forest*, para a criação de um modelo de classificação, a fim de gerar dados que permitissem de maneira rápida realizar análises sobre a urbanização em determinadas regiões. Outro propósito desse projeto é a modelagem do crescimento urbano de Curitiba e de parte de sua região metropolitana – Pinhais, Piraquara e São José dos Pinhais – por autômatos celulares (AC). Essa técnica consiste em modelos simples que representam a evolução temporal de comportamentos complexos, a partir de regras de transição (*layers* no formato *raster* que apresentam alguma restrição ou condição) e um estado inicial, imagem classificada contendo as classes: área urbana, vegetação, água e *background*. Com a modelagem e a classificação da imagem, para a época posteriori, é possível comparar a projeção do autômato. Descrito isso houve êxito em realizar ambas as atividades, contudo os resultados alcançados não foram os esperados no início do projeto, portanto foram analisados e discutidos ao término deste projeto.

Palavras-chave: Crescimento urbano. Classificação. *Random forest.* Curitiba. Região Metropolitana. Autômatos celulares

ABSTRACT

The multitemporal analysis of satellite images allows verifying the changes of a phenomenon over the years. In this project, it was proposed to use a machine learning method called random forest, toc reate a classification model, in order to generate data that would permit fast analysis of urbanization in certain regions. Another purpose of this Project is to model the urban growth of Curitiba and part of its metropolitan region – Pinhais, Piraquara and São José dos Pinhais – by cellular automata (CA). This technique consists of simple models that represent the temporal evolution of complex behaviors, based on transition rules (layers in raster format that presente some restriction or condition) and an initial state, classified iamge containing the classes: urban área, vegetation, water and background. With the modeling and classification of the image, for the posterior epoch, it is possible to compare the projection of the automaton. As described, there was success in carrying out both activities, however the results achieved were not as expected at the beginning of the project, therefore they are analyzed and discussed at the end of this project.

Keywords: Urban Growthing. Classification. Random Forest. Curitiba. Metropolitan Region. Celluar Automata

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS OU SIGLAS

COMEC	- Coordenação da Região Metropolitana de Curitiba
IBGE	- Instituto Brasileiro de Geografia e Estatística
AC	- Autômato Celular
TIN	- Malha Irregular Triangular
IPPUC	- Instituto de Pesquisa e Planejamento Urbano de Curitiba
IAT	- Instituto Água e Terra
ML	- *Machine Learning*
EMBRAPA	- Empresa Brasileira de Pesquisa Agropecuária
RMC	- Região Metropolitana de Curitiba

SUMÁRIO

1 INTRODUÇÃO

Segundo Santos e Júnior (2014) a urbanização consiste no desenvolvimento das cidades. Entre os séculos XVI até XIX o Brasil era majoritariamente um país rural, sendo as cidades locais secundários para habitação. Contudo com a chegada do séc. XX começa a ter um alto desenvolvimento industrial no país, que ocasionou na automatização do campo e na geração de novas oportunidades nas cidades, levando ao êxodo rural.

Essa mudança na distribuição espacial da população, fez com que as cidades crescessem de maneira muito acelerada e desordenada, resultando em problemas de infraestrutura, serviços e na degradação do meio ambiente. Com esses impactos se fez necessário compreender o espaço de onde as cidades se estabeleciam, com o intuito de criar diretrizes para que pudesse ocorrer a gestão urbana.

Descrito isso, uma região que se encaixa nessa situação é a cidade de Curitiba, o qual segundo o IBGE em 1960 sua população era composta por 361.309 habitantes e nos anos 80 atingiu 1.052.147 habitantes. Esse processo intenso de ocupação fez com que a população começasse a migrar para além dos limites da cidade, como se observou em São José dos Pinhais, o qual na década de 60 possuía 28.888 habitantes e em 1980 alcançou 70.634 moradores, um crescimento de mais de 240%.

Sabendo desse tipo de fenômeno, o presente trabalho se propôs a analisar esse crescimento urbano, utilizando imagens de sensores orbitais e autômatos celulares (AC), uma vez que desde a década de 80 eles vêm sido difundidos para a modelagem de fenômenos com o auxílio de dados geoespacias. No autômato é preciso a aplicação de regras de transição, que são as condições necessárias para que uma célula/pixel tenha uma mudança de seu estado em uma iteração, entretanto o estabelecimento dessas restrições não é algo simples, pois a introdução de diferentes regras de transição pode resultar em produtos distintos, como exemplo disso tem a distância das ruas principais, o qual por mais que aparenta ter influência, no dia a dia, não se sabe qual a sua real região de impacto e se necessariamente são elas que estão ocasionando mais urbanização. Para tentar contornar isso é preciso de pesquisas, como pode ser observado no trabalho de Tripathy e Kumar

(2019) que realizaram um estudo na cidade de Delhi, a fim de determinar quais são as ruas principais propicias ao desenvolvimento da cidade.

Além dessas adversidades, existem problemas na disponibilização de dados de qualidade no Brasil, segundo TEIXEIRA et al. (2017) a popularização do uso de dados espaciais, por meio da internet, banalizou esse tipo de produto, resultando no aumento de usuários sem conhecimento em cartografia que geram informações sem levar em consideração a qualidade do resultado final. Essa falta de cuidados resulta na disseminação de dados de má qualidade, que consequentemente traz insegurança em saber se um determinado dado atende ou não sua demanda.

Apresentadas essas dificuldades, para a realização da modelagem por autômato celular, é indispensável ter um estado inicial que descreva como é a disposição atual do espaço urbano de determinada região. Sendo assim, nesse trabalho se fez necessário realizar a classificação de imagens de satélite. Para isso decidiu-se utilizar aprendizado de máquina, para a criação de um modelo de classificação, que faria essa tarefa.

Com isso em mente, nesse projeto, foi proposto um estudo de autômatos celulares na região de Curitiba, Pinhais, Piraquara e São José do Pinhais, área também considerada no projeto desenvolvido por Peschl (2021) que obteve em suas pesquisas arquivos em formato *raster* que serviriam como regras de transição para o AC. Com essas informações e imagens classificadas, como estado inicial, seria possível realizar a modelagem do crescimento urbano.

1.1 JUSTIFICATIVA

Compreender como uma cidade irá se desenvolver, tem sido um dos objetivos atuais para a realização da modelagem por autômatos celulares em regiões urbanas, exemplos desses estudos podem ser vistos por Zhou e Chen (2020) ou Feng, Liu e Batty (2015).

Um modelo de autômato celular simples foi desenvolvido por Tripathy e Kumar (2019) para uma análise do crescimento urbano de Delhi, que disponibilizaram seu código implementado em linguagem *Python*. Sendo assim se propôs utilizar esse algoritmo para realizar um estudo sobre seus efeitos no crescimento urbano de Curitiba e região metropolitana.

1.2 OBJETIVOS

1.2.1 Objetivo geral

Classificar imagens de satélites que abrangem a região de Curitiba, Piraquara, Pinhais e São José dos Pinhais, a fim de verificar se a previsão urbana realizada pelo autômato celular simples de Tripathy e Kumar (2019) proporcionam resultados condizentes com a realidade, utilizando como base as regras de transição que tiveram o tratamento de Peschl (2021)

1.2.2 Objetivos específicos

1. Desenvolver modelo de classificação, pelo método *Random Forest*, para imagens multitemporais;
2. Realizar operações com as imagens classificadas, visando refina-las;
3. Utilizar imagens classificadas no modelo de autômato celular desenvolvido por Triphaty e Kumar (2019);
4. Testar diferentes limiares de parâmetros no modelo do autômato celular desenvolvido por Triphaty e Kumar (2019);
5. Realizar análise dos resultados obtidos.

2 REVISÃO DE LITERATURA

2.1 APRENDIZADO DE MÁQUINA

O termo aprendizado de máquina ou *machine learning* (ML), surgiu em 1952, quando um dos pioneiros da inteligência artificial, Arthur Samuel, descreveu o conceito de ML como a habilidade dos computadores aprenderem conceitos ou padrões, sem existir a necessidade de que tais aptidões fossem programadas. Entretanto, por mais que essa metodologia tenha sido desenvolvida na década de 50, foi somente com o advento da internet e a evolução do poder computacional, que o aprendizado de máquina passou ser mais estudado e expandido.

Os quatro tipos principais de modelos de classificação são:

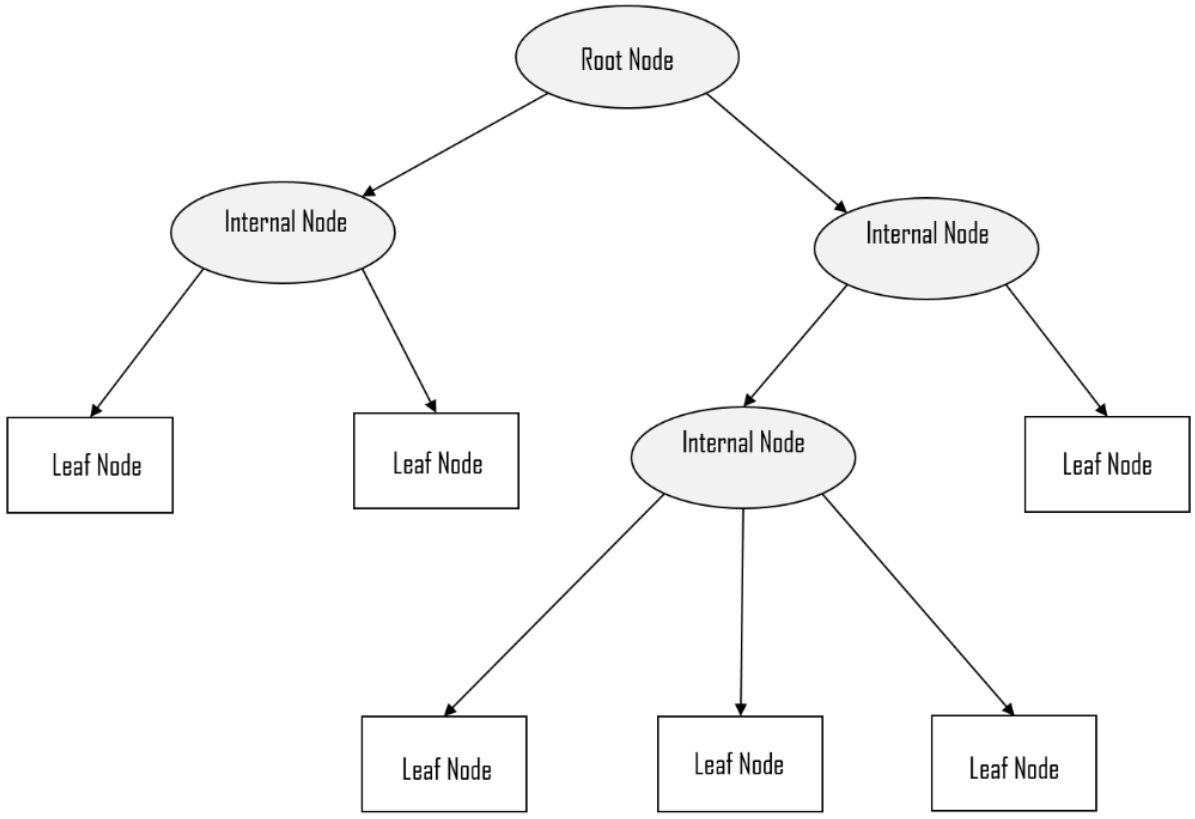
- Supervisionado: A máquina recebe um conjunto de dados, previamente classificado, que gera um modelo que tenta se aproximar ao máximo do resultado das amostras recebidas;
- Não supervisionado: O computador não recebe informações rotuladas, dessa maneira ele tenta encontrar padrões, a partir das amostragens que ele tem;
- Semi supervisionado: São recebidos dados previamente classificados, contudo isso não ocorre para todo sua informação, sendo assim ele necessita encontrar parâmetros que possibilitem descrever corretamente o tipo da amostra desconhecida;
- Reforço: O modelo toma decisões a respeito de um conjunto de dados, caso a informação esteja correta ele é recompensado, caso não, ele recebe alguma penalidade. Sendo assim o objetivo final dessa técnica é que na tentativa e erro ele obtenha o melhor resultado possível no final.

2.1.1 Árvore de decisões

Dos tipos de modelo que foram apresentados anteriormente, um dos principais tipos para a classificação supervisionado é a árvores de decisões, o qual em um contexto visual, é similar a um fluxograma.

De maneira simplificada as árvores particionam um conjunto de dados, executando a aplicação em cada nó de regras, o qual caso ela seja cumprida ou não, será direcionada a um outro nó ou a um nó folha, sendo que na primeira opção, o ciclo de decisões continuará, entretanto se for a segunda, aquela linha de raciocínio terá um fim. Na figura a seguir é possível ver uma representação do que foi mencionado.

FIGURA 1 - Fluxo de uma árvore de decisão



FONTE: Hauwei Community

Em um contexto geral as árvores de decisões funcionam da mesma maneira, contudo sua construção pode se dar de forma diferente, dependendo do algoritmo, a seguir serão apresentados dois exemplos.

- Árvore do tipo CART (*Classification and Regression Trees*): É uma árvore de decisão binária, que em cada nó as amostras são divididas em dois grupos disjuntos considerando o valor do índice de *Gini* da partição;
- Árvore do tipo ID3 (*Iterative Dichtomiser* 3): Cria uma árvore multidirecional, que gera em cada nó a característica que fornecerá o maior ganho de informações para os valores rotulados.

2.1.2 Random Forest Classifier

O *Random Forest* é um algoritmo do tipo *ensemble*, isso é, um método que se origina da composição de programas mais básicos, com a particularidade que combina diversos modelos para obter um único resultado final. Nesse caso o algoritmo apresentado utiliza um grande número de árvores de decisões, para obter o melhor modelo de classificação.

2.1.3 Hiperparâmetros

São parâmetros que devem ser definidos previamente ao treinamento de um modelo. No caso do *Random Forest*, alguns exemplos são o tamanho das árvores de decisão ou quantas amostras cada nó folha pode ter.

2.2 MORFOLOGIA MATEMÁTICA

A morfologia matemática surgiu em 1964 pelas pesquisas de Jean Serra e Georges Matheron e tinha sido desenvolvida – inicialmente – para a análise de imagens microscópicas.

Apesar de ter como objetivo extrair informações de imagens, utilizando um elemento estruturante, os operadores morfológicos permitem restaurar imagens, aumentar sua qualidade, compreender sua estrutura, entre outras funcionalidades. A seguir serão apresentadas algumas funções da morfologia matemática, sendo elas erosão, dilatação e fechamento.

2.2.1 Erosão

A erosão é uma operação que consiste em eliminar ou reduzir conjuntos em uma imagem, dependendo do tamanho do seu elemento estruturante. Matematicamente a erosão é definida por:

$$A\theta B = \{c \in \mathrm{Z}^2 | \mathrm{c} + \mathrm{b} \in \mathrm{A}, \mathrm{para\ b} \in \mathrm{B}\}$$

Sendo que:

- A e B: São conjuntos pertencentes a Z^2;
- A: É a imagem a ser erodida;
- B: É o elemento estruturante, que irá percorrer a imagem.

2.2.2 Dilatação

A dilatação tem por objetivo aumentar, conectar e reparar conjuntos. Sua expressão matemática é dada por:

$$A \oplus B = \{c \in \mathrm{Z}^2 | \mathrm{c} = \mathrm{a} + \mathrm{b}, \mathrm{a} \in \mathrm{A\ e\ b}\ \in \mathrm{B}\}$$

2.2.3 Fechamento

É a operação da dilatação seguida da erosão, tem a função de suavizar contornos, conectar conjuntos e manter objetos maiores que o elemento estruturante. Sua definição matemática é demonstrada por:

$$A \bullet B = (A \oplus B)\theta B$$

2.3 AUTÔMATO CELULAR

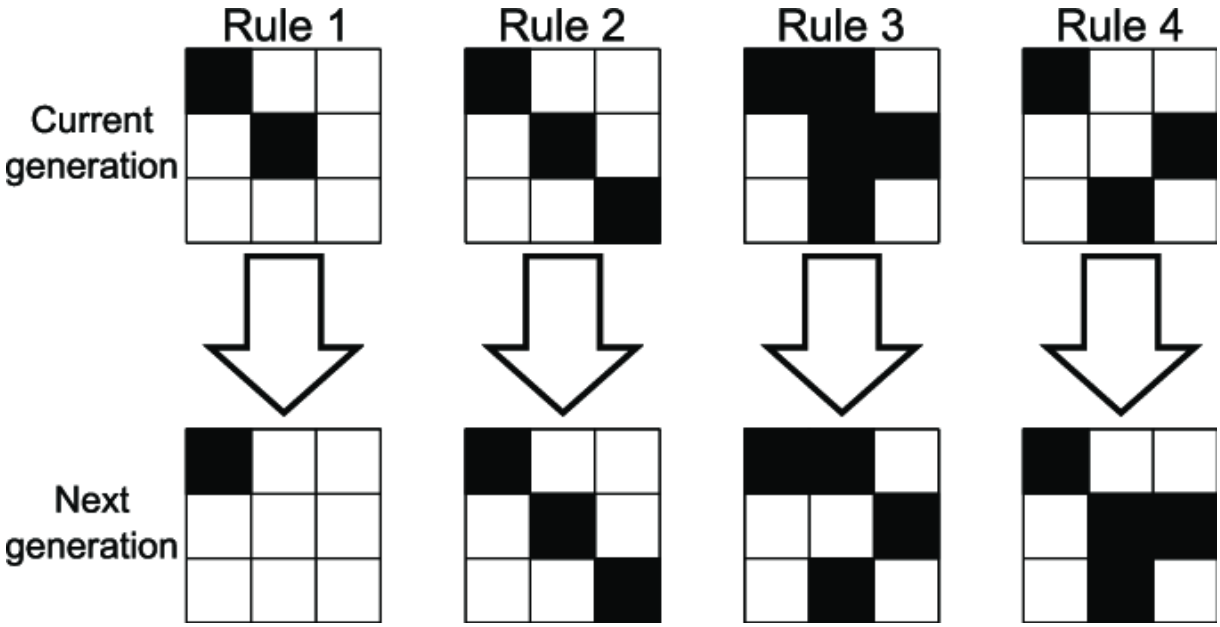
O primeiro autômato celular (AC) proposto foi por John von Neumann (1903 - 1957) com a proposta de modelar a autorreprodução biológica, a partir de iterações e regras que permitiram modelar comportamentos complexos. Contudo esse tipo de

modelagem só foi se popularizar em 1970 quando John Holton Conway propôs o autômato celular como um jogo, que ficou conhecido como “jogo da vida” – *game of life.*

Esse AC tinha por objetivo representar as alterações e mudanças de um determinado grupo, utilizando regras simples. Nesse modelo uma célula poderia morrer ou viver dependo da condição em que ela estava exposta. A seguir estão as regras definidas para o seu comportamento.

1. Qualquer célula com menos de dois vizinhos morre de solidão;
2. Qualquer célula com dois vizinhos vivos continua no mesmo estado para a próxima iteração.
3. Qualquer célula viva com mais de três vizinhos vivos morre de superpopulação;
4. Qualquer célula com exatamente três vizinhos vivos se torna uma célula viva;

FIGURA 2 - Regras do jogo da vida



FONTE: Takayuki Hirose e Tetsuo Sawaragi (2020)

2.4 CURITIBA E SUA REGIÃO METROPOLITANA

Segundo Pitz (2021) antes da ocupação dos colonizadores portugueses, o primeiro planalto paranaense, por muito tempo chamado de planalto de Curitiba, foi uma região que apresentava vasta vegetação rasteira, grandes bacias hidrográficas e grande presença da mata atlântica. Nesse local viviam a princípio as etnias indígenas Tupi Guarani (localizados na Serra do Mar e litoral) e os Jê Meridional (presentes em regiões mais altas). Ambos os grupos apresentados eram conhecidos pela coleta, caça e por não apresentarem aldeias fixas, dessa maneira a etimologia da palavra Curitiba, na linguagem tupi (Coritiba), evidencia o motivo da região ter sido ocupada, pois “coré” significa cateto, coleta e caça, enquanto que “tiba” muita, abundância e fartura.

Por volta de 1640 teve início a chegada de portugueses e imigrantes a região de Curitiba, devido a busca do ouro e as excursões dos bandeirantes, o qual fez com que o local deixasse de ser uma região de ocupação majoritariamente indígena, assumindo em 1693 a posição de vila. Apesar disto, constata-se que foi apenas em 1850 que houve um aumento considerável na população dessa região, devido as políticas de imigrações de colonos europeus, o qual grande parte dos imigrantes ao invés de serem direcionados a Curitiba, eram realocados para núcleos coloniais, os quais futuramente dariam origem a novos munícipios ou ampliariam o território da cidade. Nas próximas décadas que passariam, devido ao declínio do cultivo de café, a mecanização do campo e a industrialização, verificou-se um êxodo rural, os quais as pessoas se dirigiam as cidades em busca de novos empregos e oportunidades.

Mais tarde, por volta dos anos de 1970, a Região Metropolitana de Curitiba começou a ter um alto crescimento urbano, quando o Paraná passa a ser uma das federações em que a modernização se fez expressiva. Nos anos seguintes foram implantadas novas rodovias e novas oportunidades começaram a surgir longe da capital, o qual gerou o surgimento de cidades-polos.

Atualmente a Região Metropolitana de Curitiba é formada por 29 municípios e segundo Nojima et al (2004) é o polo da economia paranaense. Este fato está vinculado a instalação de empresas de grande porte, como é o caso da Renault em São José dos Pinhais e a New Holland em Curitiba. Essa concentração de

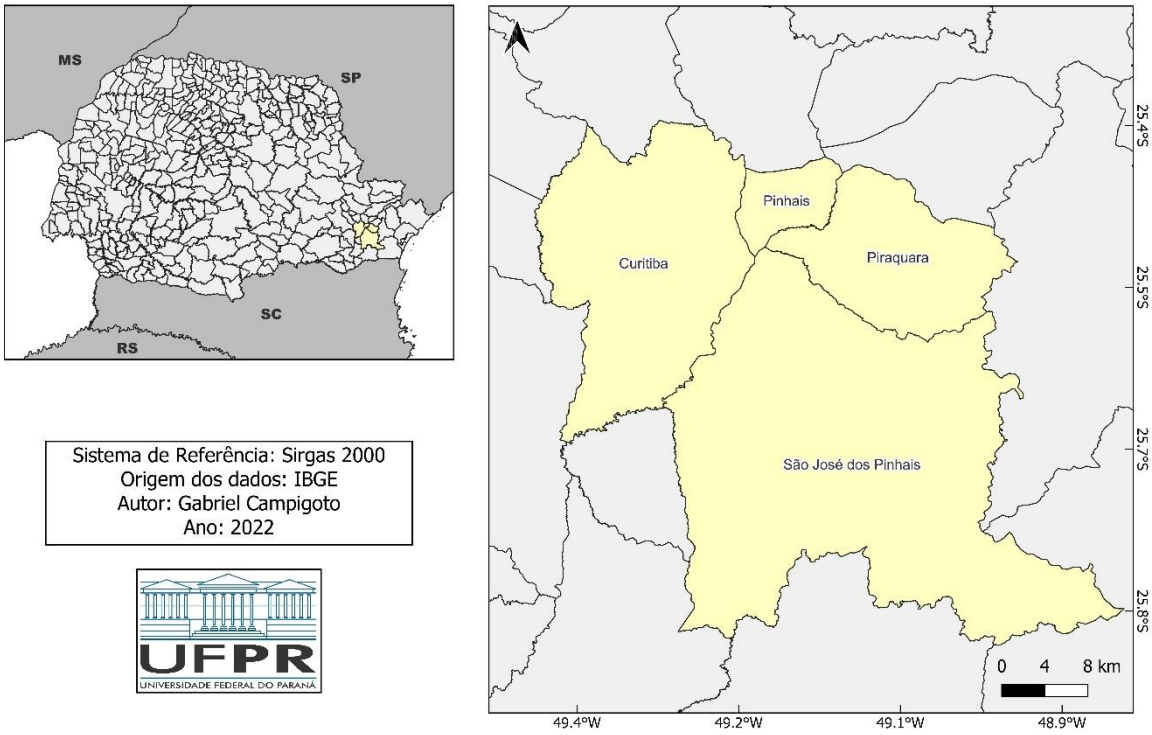
investimentos econômicos na região mudou o perfil socioespacial e ambiental da região (NOJIMA et. al, 2004, p.5).

Com todo o crescimento dessa região houve o aumento na ocupação e uso sobre área visualmente vulneráveis, particularmente os mananciais de abastecimento hídrico. Esse intenso e contínuo processo de apropriação vem agudizando as contradições socioespaciais da Região Metropolitana de Curitiba (NOJIMA et al, 2004, p. 34)

3 MATERIAL E MÉTODOS

Os materiais e métodos que serão apresentados no decorrer desse capítulo, foram limitados a área de estudo, que corresponde a Curitiba, Pinhais, Piraquara e São José dos Pinhais. A sua disposição espacial pode ser visto na FIGURA 3.

FIGURA 3 - Área de Estudo



FONTE: O autor (2022)

Os municípios expostos fazem parte da Região Metropolitana de Curitiba (RMC) desde a lei estadual n° 11.027/94. Segundo LIMA (2004) Curitiba sedia e mantém as principais funções do governo estadual, possui os principais espaços educacionais, comerciais e culturais, que somados conferem um posicionamento de relevância no contexto estadual e do país.

A Região Metropolitana de Curitiba é constituída por áreas com algum grau de urbanização, assim cabe destacar que um dos aspectos que mais caracterizam a Região Metropolitana de Curitiba são os diferentes contextos espaciais resultantes

da transformação e das tendências da base produtiva regional e estadual (LIMA, 2004).

Francisco (2005, p.49) menciona que as alterações em um ambiente dizem respeito à evolução conjunta das condições sociais e ecológicas estimuladas pelos impulsos das relações entre forças externas e internas de uma unidade espacial e ecológica, histórica ou socialmente determinada. É a relação entre sociedade e natureza que se transforma diferencialmente e dinamicamente, reestruturando o espaço (COELHO, 2000, p.25)

3.1 MATERIAIS

Os materiais usados para esse trabalho foram desde *softwares*, programas em linguagem *Python*, dados desenvolvidos por órgãos governamentais, por Peschl (2021), Tripathy e Kumar (2019) e o próprio autor. A seguir serão apresentados cada uma dessas informações e a procedência delas.

- Imagens de satélite: Tiveram origem do *Earth Explorer*, disponibilizada pela USGS;
- QGIS 3.22.0 (Białowieża): Sistema de informações geográficas, de código aberto, com o objetivo de permitir a visualização, edição e análise de dados geoespaciais;
- IDE *Spyder* versão 5: Ambiente gratuito para o desenvolvimento de algoritmos em linguagem *Python;*
- Algoritmo de modelagem do autômato celular: Código originalmente desenvolvido por Tripathy e Kumar (2019);
- Dados para aplicar as regras do autômato celular: Foram utilizados arquivos *rasters* que foram criados e tratados noo trabalho de conclusão de curso de Peschl (2021). A fonte original dessas informações, entretanto vem de órgãos públicos (IBGE, COMEC e IPPUC) e também de inciativas privadas – *Open Street Map*;
- Bibliotecas para o desenvolvimento dos códigos em Python:

- Numpy: Possui uma vasta gama de funções, que permitem operações matemáticas entre arrays;
- GDAL: Permite a visualização e operação com dados vetoriais e matriciais, mantendo suas informações geoespaciais;
- OpenCV: Desenvolvida principalmente para o processamento de imagem e visão computacional;
- Tkinter: Utilizada para trabalhar com interfaces gráficas no *Python*;
- Scikit – Learn: *Software* gratuito para aprendizado de máquina.
- Pandas: Usado para a manipulação de dados e análises.

3.2 METODOLOGIA

A seguir serão apresentadas as etapas, desde a geração do modelo de classificação até o resultado final do autômato celular.

3.2.1 Classificação das imagens

3.2.1.1 Seleção das imagens

Antes de começar a realizar qualquer atividade desse trabalho, foi necessário obter imagens de satélite Landsat 4,5,7,8 e 9, de um período que abrange desde 1991 até 2022. Como critério a área de estudo deveria estar isenta de nuvens ou ter uma quantidade relativamente pequena, que de preferência não interferisse nas áreas urbanas.

A seguir será apresentada a TABELA 1 que indica todas as imagens propícias ao trabalho. Essa análise teve que ser manual e tiveram sua origem do USGS (*United States Geological Survey*) com o auxílio da ferramenta *Earth Explorer*, o qual permite a visualização e *download* gratuito de imagens, ao fornecer a posição, data e o satélite desejado.

TABELA 1 - Conjunto de imagens para treinamento e verificação

Data	Satélite
12/09/1991	Landsat 5

04/11/1993	Landsat 5
18/07/1994	Landsat 5
03/08/1994	Landsat 5
16/04/1995	Landsat 5
18/04/1996	Landsat 5
24/06/1997	Landsat 5
30/06/1999	Landsat 5
26/09/1999	Landsat 7
19/08/2000	Landsat 5
31/05/2000	Landsat 5
24/06/2000	Landsat 7
07/05/2000	Landsat 7
02/01/2001	Landsat 7
02/09/2002	Landsat 7
26/03/2002	Landsat 7
11/07/2003	Landsat 5
20/12/2004	Landsat 5
04/12/2004	Landsat 5
15/09/2004	Landsat 5
30/08/2004	Landsat 5
02/09/2005	Landsat 5
21/11/2005	Landsat 5
24/11/2006	Landsat 5
05/09/2006	Landsat 5
20/08/2006	Landsat 5
19/07/2006	Landsat 5
06/07/2007	Landsat 5
31/10/2009	Landsat 5
05/03/2009	Landsat 5
01/02/2009	Landsat 5
19/11/2010	Landsat 5
08/09/2013	Landsat 8
29/10/2014	Landsat 8
26/08/2014	Landsat 8
30/01/2014	Landsat 8
13/08/2015	Landsat 8
12/06/2016	Landsat 8
20/01/2016	Landsat 8
14/05/2017	Landsat 8
26/02/2018	Landsat 8
23/07/2019	Landsat 8
14/07/2019	Landsat 8
18/04/2019	Landsat 8
02/04/2019	Landsat 8
13/10/2020	Landsat 8
26/08/2020	Landsat 8

09/07/2020	Landsat 8
07/06/2020	Landsat 8
04/04/2020	Landsat 8
25/05/2021	Landsat 8
13/02/2022	Landsat 8

FONTE: O autor (2022)

Sobre as imagens presentes na TABELA 1: a princípio pensava-se em aproveitar todas durante a execução das atividades. Contudo, somente as em amarelo foram utilizadas. Isso ocorreu devido ao tempo de processamento e as dificuldades para se treinar o modelo de classificação. Assim decidiu-se reduzir o volume de dados. No caso das imagens selecionadas, elas seriam utilizadas para desenvolver e verificar o modelo. Das que foram escolhidas, nenhuma é proveniente do Landsat 7, uma vez que a partir do dia 31 de maio de 2003 o *Scan Line Corrector* (SLC) falhou, resultando em lacunas nos produtos gerados, dessa maneira existem muito poucas imagens que não apresentam esse defeito.

Determinado o *dataset* que seria utilizado para a coleta de amostras, foi escolhido quais bandas seriam usadas para a obtenção dos valores de pixels. No caso das que estavam antes de 2010 (imagens Landsat 5), foram utilizadas as 3 bandas do visível, 3 do infravermelho e 1 do termal, enquanto as imagens posteriores a 2010 (imagens Landsat 8) houve a adição de mais uma banda do termal, a Coastal Aerosol e a Cirrus. Essa escolha de bandas se deu para que o modelo tivesse o máximo de informações espectrais, a fim de que ele pudesse diferenciar as classes com mais facilidade.

Além disso necessitou-se separar os tipos de bandas para cada um dos satélites, pois seria preciso treinar dois modelos de classificação, uma vez que o Landsat 5 e Landsat 8 tem diferenças nos níveis radiométricos – 8 e 12 bits respectivamente – e também possuem diferenças na resolução espectral (TABELA 2), essas informações podem ser encontradas no site da USGS e da EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária), que podem ser acessadas via referência.

TABELA 2 – Resoluções espectrais do Landsat 5 e 8

Landsat 5		**Landsat 8**	
Banda	**Resolução espectral (μm)**	**Banda**	**Resolução espectral (μm)**
Azul	0,45 – 0,52	**Azul**	0,45 – 0,51
Verde	0,52 – 0,60	**Verde**	0,53 – 0,59
Vermelho	0,63 – 0,69	**Vermelho**	0,64 – 0,67
Infravermelho próximo	0,76 – 0,90	**Infravermelho próximo**	0,85 – 0,88
Infravermelho médio	1,55 – 1,75	**Infravermelho médio 1**	1,57 – 1,65
Termal	10,40 – 12,50	**Termal 1**	10,6 – 11,19
Infravermelho médio	2,08 – 2,35	**Infravermelho médio 2**	2,11 – 2,29
-	-	**Termal 2**	11,50 – 12,51
-	-	**Cirrus**	1,36 – 1,38
-	-	**Costa Aerosol**	0,43 – 0,45

FONTE: O autor (2022)

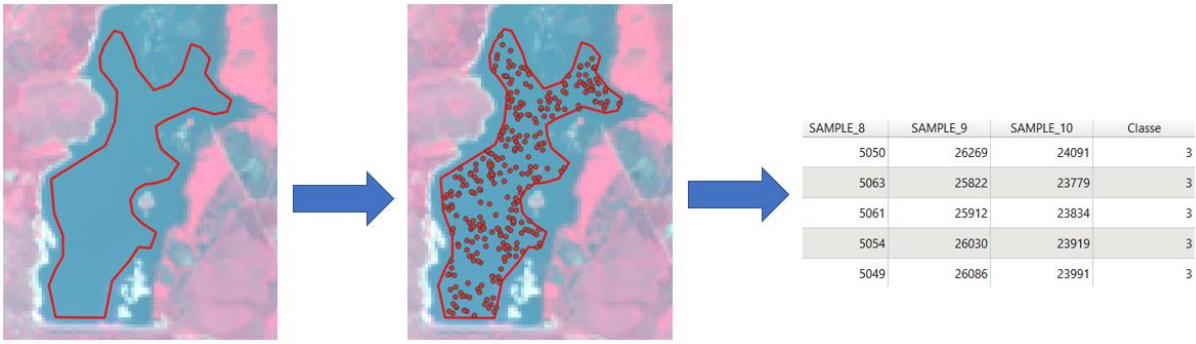
3.2.1.2 Criação das amostras de treinamento

Segundo o estudo de Tripathy e Kumar (2019) o modelo de autômato celular, criado por eles, teriam as seguintes classes: área urbana, vegetação, água e outros. Nesse trabalho, com o intuito de utilizar o algoritmo desenvolvido por Tripathy e Kumar (2019), decidiu-se utilizar os mesmos elementos, contudo com algumas diferenças. A região urbana seria um local mais antropizado, abrangendo regiões de solo exposto e áreas rurais, o qual no texto base seria vegetação. Por fim a classe outros seria o *background* das imagens, o qual possui valores de pixel igual a 0 em todas as bandas.

A seguir será apresentado o passo a passo de como foi feita a coleta das amostras.

1. Criação de polígonos: Utilizando o *software* QGIS definiram-se áreas que representassem as classes dentro das imagens utilizadas como treino;
2. Criação de pontos: Com os polígonos criados, dentro deles foram inseridos dados pontuais – distribuídos de maneira aleatória – com o auxílio do QGIS. A quantidade máxima de pontos destruídos ficou limitado a 500, por área, e tinha por objetivo extrair os valores de pixels nas bandas das imagens;
3. Rotulação das classes: Com os pontos criados, foi necessário rotular cada um deles, sendo eles 1 para área urbana, 2 para vegetação, 3 para água e 4 para o *background*;
4. Exportar para arquivo CSV: Com as amostras prontas, exportou-se o dado vetorial como uma tabela, para que futuramente isso fosse utilizado no algoritmo de treinamento.

FIGURA 4 - Exemplo de coleta de amostras



FONTE: O autor (2022)

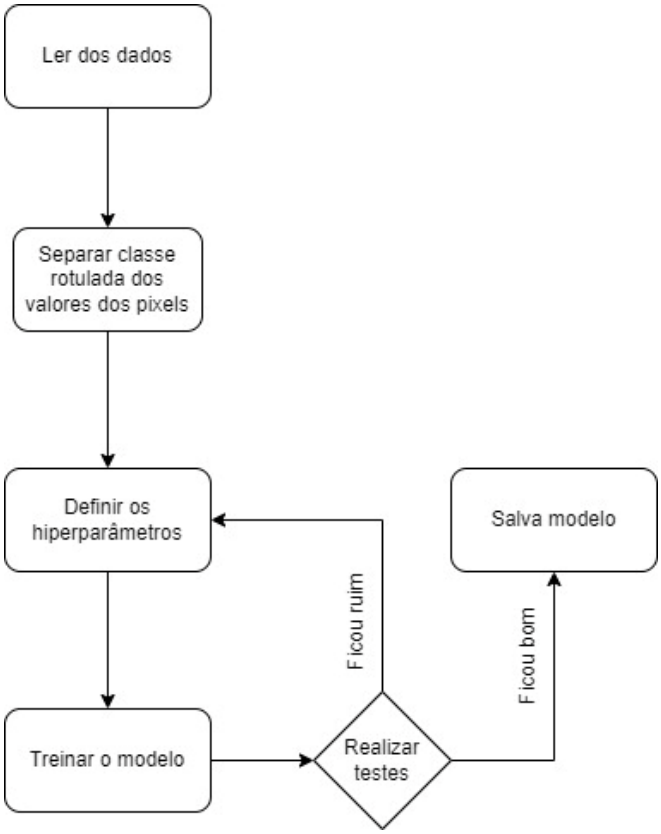
Após a coleta de amostras, nas etapas de criação de polígonos e pontos, eram analisadas regiões nas quais as classes eram iguais para ambas as imagens de treinamento, pois mesmo que a posição do pixel seja igual, existirá diferença em seus valores, uma vez que a origem do dado nunca será igual a de outro.

Por fim em relação a todas as etapas apresentadas, esse processo foi repetido mais de uma vez durante esse trabalho, visto que nos modelos de classificação treinados notou-se uma deficiência em certas regiões, necessitando coletar mais amostras.

3.2.1.3 Treinamento do modelo

Para realizar a fase de treinamento foi criado um algoritmo em linguagem *Python* que conseguisse gerar um modelo de classificação pelo método *Random Forest Classifier*. A seguir será apresentado um fluxograma que exemplifica as etapas tomadas.

FIGURA 5 - Fluxograma para o treinamento do modelo de classificação



FONTE: O autor (2022)

Do que foi apresentado anteriormente a etapa mais importante foi a definição dos hiperparâmetros. O Scikit – Learn possui uma ferramenta chamada de *GridSearch*, a qual analisa parâmetros de entradas e cria o máximo de combinações com esses valores para entrar no treinamento, isso impacta diretamente quanto tempo irá demorar para gerar o modelo de classificação, assim como na sua acurácia final, visto que quanto mais possibilidades mais análises necessitam ser feitas.

Abaixo serão descritos os hiperparâmetros foram fornecidos ao algoritmo.

- Número de árvores de decisões para o *Random Forest*;
- A métrica para as árvores de decisão (*gini* e *entropy*);
- Amostras mínimas para dividir um nó interno;
- Profundidade máxima que uma árvore pode ter;
- Método para dividir as amostras em cada árvore;
- Mínimo de amostras para dividir um nó externo.

Dos hiperparâmetros acima, o das métricas são funções que medem a qualidade da melhor divisão de um nó. No Scikit-Learn usar esses critérios resulta na construção de uma árvore CART (*gini*) e ID3, caso seja do tipo *entropy*.

Além do que foi apresentado, o *GridSearch* necessita também de uma metodologia para retornar o melhor modelo possível, nesse caso foi escolhido a acurácia.

3.2.1.4 Classificação pelo modelo

Após a etapa de treinamento e da geração de um modelo de predição, era necessário classificar as imagens de satélite, para verificar se o resultado obtido estava condizente com a realidade. Para isso utilizou-se a biblioteca GDAL, uma vez que as imagens trabalhadas eram muito grandes e outros pacotes não tinham capacidade de realizar a leitura delas, além de que ela consegue salvar informações de sistema de referência e projeção, que podem ser inseridas novamente no produto final.

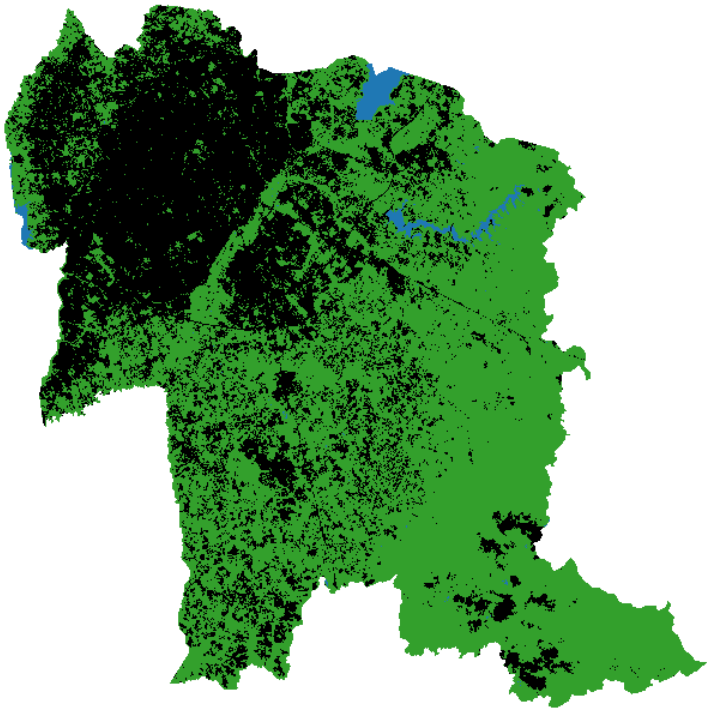
O processo seguido para classificar as imagens se deu da seguinte maneira:

1. Abrir a imagem;
2. Definir suas dimensões;
3. Guardar suas informações geoespaciais;
4. Transformar o dado inicial em um *array* de linhas únicas para cada banda disponível;
5. Utilizar o modelo para predizer o dado transformado;
6. Salvar resultado final, restaurando suas informações geoespaciais.

Todo esse processo descrito resultou em um algoritmo que tem um tempo de execução rápido ou lento, dependendo do tamanho do seu dado bruto, contudo como a região estudada é apenas Curitiba e parte da região metropolitana, o dado classificado era retornada de maneira ágil.

Dessas imagens as primeiras a serem analisadas eram as que foram utilizadas como fontes de treinamento, pois caso houvesse *underfitting* (modelo não se ajustou nem aos de treino e nem os testes) o modelo de predição era descartado imediatamente, retornando a fase de treino novamente.

FIGURA 6 - Exemplo de imagem classificada



FONTE: O autor (2022)

3.2.1.5 Pós-classificação

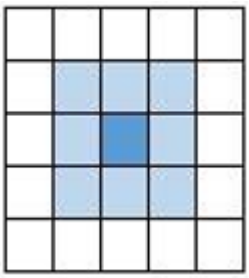
Durante as atividades, notou-se que os modelos de classificação, treinados, geravam muitos ruídos na imagem, além de apresentarem constantemente problemas na generalização de suas classes, logo alguns dados obtinham resultados melhores que outros em determinados modelos

Para resolver esse impasse foi proposto realizar um pós-processamento nas predições obtidas com o operador morfológico fechamento, utilizando um elemento estruturante 3 x 3 – FIGURA 7. Outro processamento visou minimizar o efeito

sazonal nas regiões de agricultura e represas, utilizando como máscara imagens mais recentes e com classificação aparente mais correta, seguindo as seguintes regras.

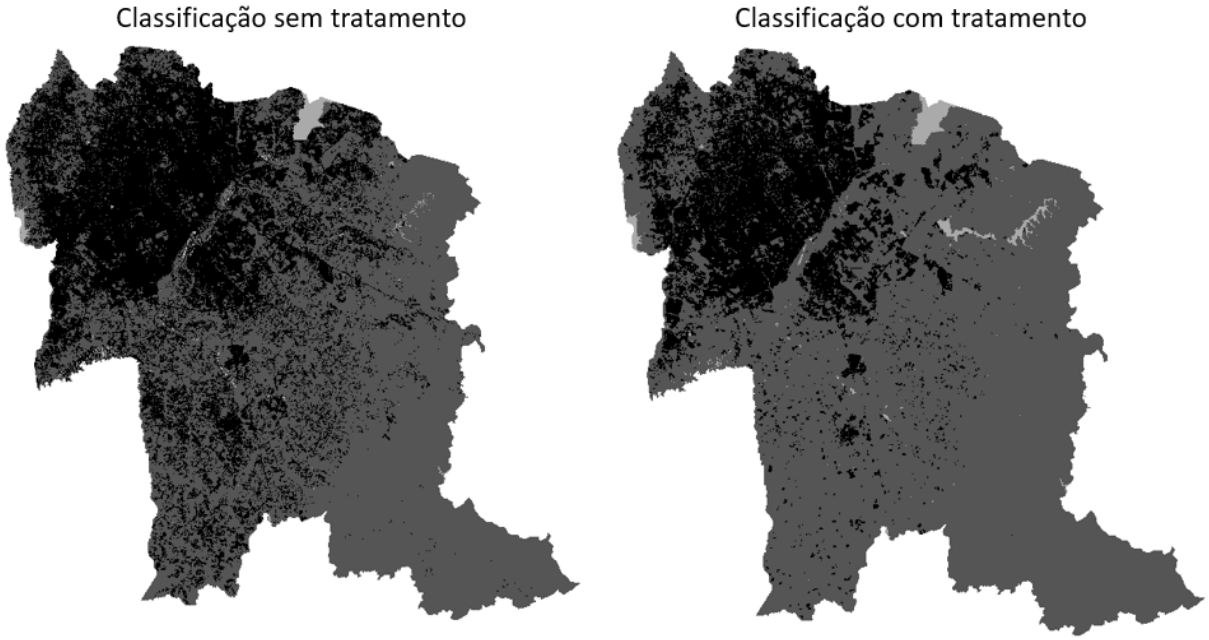
- Caso na imagem mais recente ainda exista vegetação e na anterior não, o valor será alterado para se tornar igual ao da mais atual, uma vez que a chance de haver um crescimento de vegetação é menor do que ter urbano. Isso pode ser observado no estudo de Stanganini e Lollo (2018) que dizem que a degradação ambiental está associada ao crescimento das cidades.
- Caso na imagem mais atual exista água e na anterior não, o valor será alterado para se tornar igual ao da mais atual, isso se deve pois em certas imagens não existiam reservatórios, logo para evitar que o modelo do autômato previsse crescimento urbano nessas regiões, elas já seriam alteradas.

FIGURA 7 - Elemento estruturante 3 x 3



FONTE: Fioravanti (2019)

FIGURA 8 - Imagem pré e pós-processamento



FONTE: O autor (2022)

3.2.2 Modelo do autômato celular

O algoritmo do autômato celular, utiliza uma matriz 3 x 3 que analisa os pixels vizinhos, observando o atual estado do pixel central. A expressão que considera a vizinhança de um pixel i,j é:

$$\begin{bmatrix} a_{i-1,j-1}^{(t)} & a_{i-1,j}^{(t)} & a_{i-1,j+1}^{(t)} \\ a_{i,j-1}^{(t)} & a_{i,j}^{(t)} & a_{i,j+1}^{(t)} \\ a_{i+1,j-1}^{(t)} & a_{i+1,j}^{(t)} & a_{i+1,j+1}^{(t)} \end{bmatrix}$$

Segundo Tripathy e Kumar (2019) o modelo depende do estado inicial do pixel, da vizinhança e suas regras de transição, por isso é essencial que a posição espacial de cada *raster* esteja se sobrepondo no local correto, para evitar erros na previsão.

O estado futuro do valor do pixel é determinado pelas suas regras de transições e seu atual estado, cuja a expressão matemática é:

$$a_{i,j}^{t+1} = \phi(A_{i,j}^{t})$$

A regra de transição (ϕ) é uma função que representa o limiar dos valores, essa expressão é dada por:

$$\phi = \mathrm{f}(\mathrm{T}, \mathrm{B})$$

A equação acima nos diz que as regras de transição estão em função de T, os quais os limites dos conjuntos afetam os parâmetros e B (número de pixels de classe urbano), para cada elemento de T. T e B são expressos por:

$$T = \{T_R, T_C, T_P, T_S\}$$
$$B = \{B_R, B_C, B_P, B_S\}$$

Sendo que:

- T_R, T_C, T_P, T_S: São os valores limiares para proximidade das ruas, distância do centro, densidade populacional e declividade;
- B_R, B_C, B_P, B_S: Corresponde aos pixels de região urbana para cada elemento que pertence a T;

Do que foi apresentado as regras do AC são:

- Caso o pixel seja área urbana e água, não haverá modificação de classe na modelagem;
- Caso o pixel contenha vegetação ele pode ser alterado devido aos limiares propostos e a sua vizinhança, desde que não exista uma regra que o restrinja a modificação para a classe urbano.

Ao que foi apresentado nessa seção, sua base teórica foi adaptada de Tripathy e Kumar (2019), uma vez que o algoritmo utilizado nesse trabalho para a previsão do crescimento urbano, foi desenvolvido por eles.

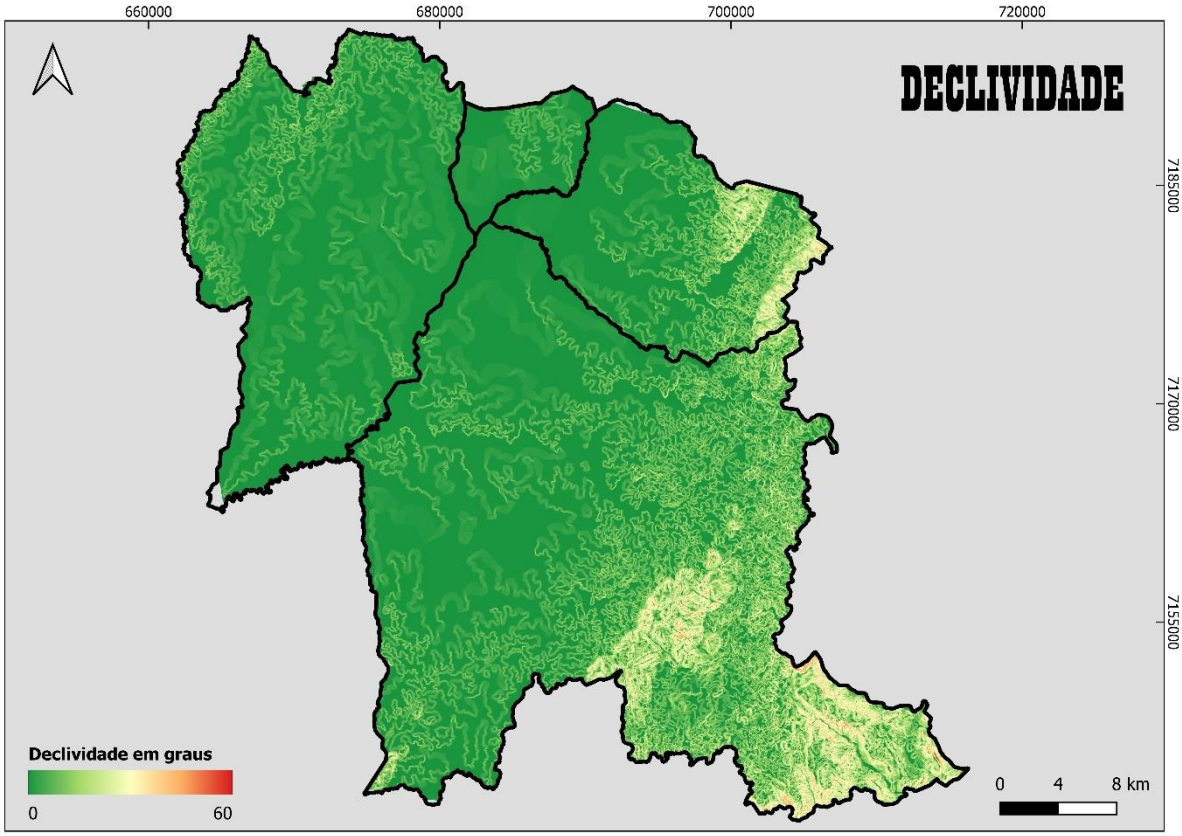
3.2.2.1 Regras do autômato celular

Tendo já discutido a parte teórica e a metodologia aplicada no código do autômato celular, foram obtidos arquivos no formato *raster* que serviriam como dados paras as regras de transição. Essas informações foram disponibilizadas por órgãos governamentais e por empresas, que fornecem dados livres, e tiveram o tratamento de Peschl (2021). A seguir será explicado como ele desenvolveu cada um desse dados.

3.2.2.1.1 Declividade

As curvas de nível tiveram sua origem do IAT (Instituto Água e Terra), espaçadas de 20 em 20 metros. A partir desse dado foi realizado a interpolação pelo método TIN – Malha Irregular Triangular – para obter a declividade da região de interesse.

FIGURA 9 - Mapa de declividade

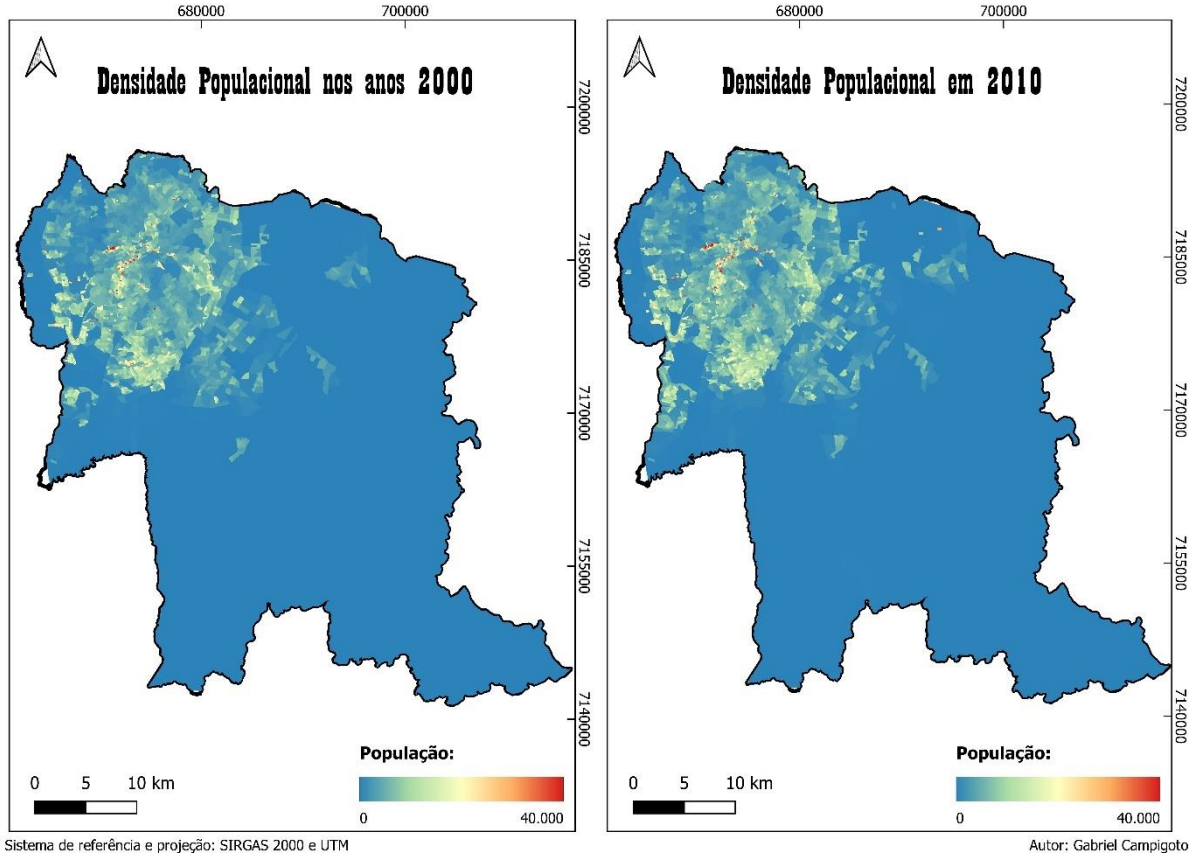


FONTE: O autor (2022)

3.2.2.1.2 Densidade Populacional

Foi obtido a partir do censo de 2000 e 2010, realizado pelo IBGE (Instituto Brasileiro de Geografia e Estatística). A princípio esses eram dados tabulares e de área (malha censitária), que a partir do comando *join* do QGIS foram combinadas essas informações e em seguida foi convertido o arquivo *shapefile* para o formato *raster*.

FIGURA 10 - Mapas de densidade populacional



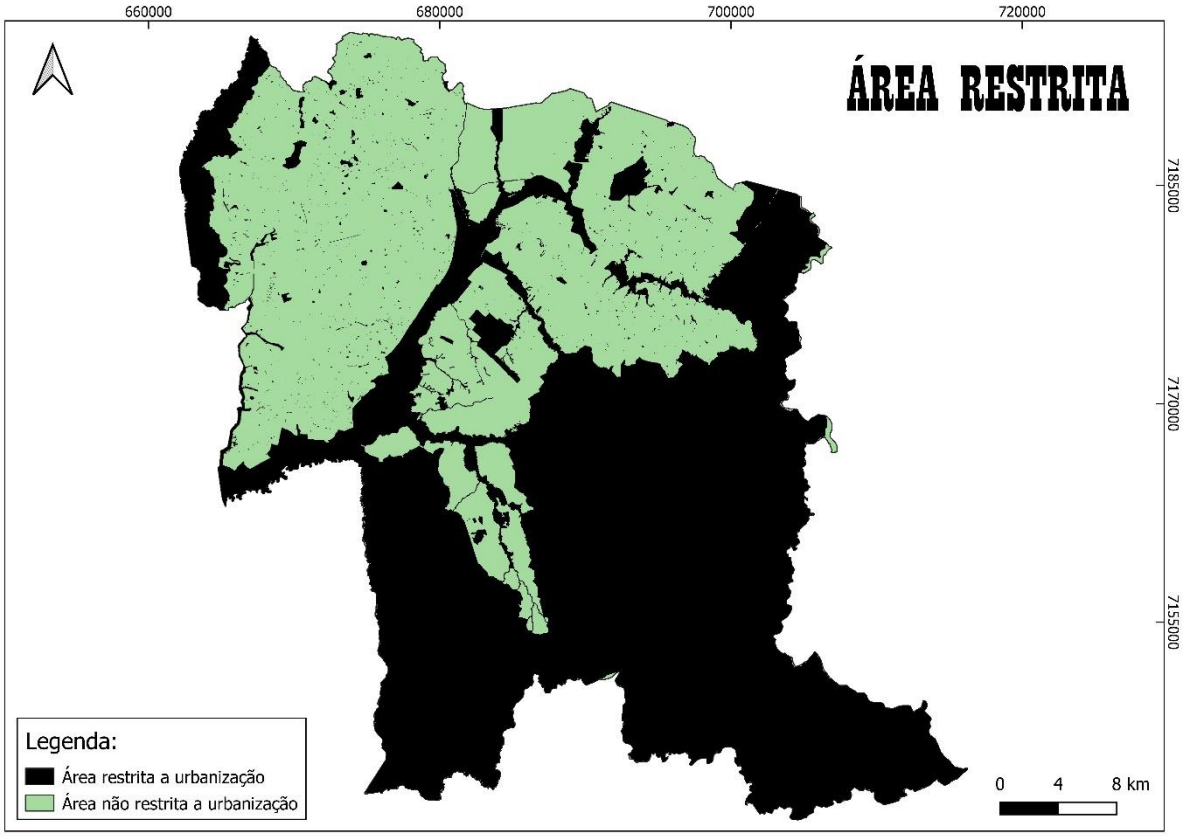
FONTE: O autor (2021)

3.2.2.1.3 Área de restrição para a urbanização

Foram gerados a partir dos dados de zoneamento do IPPUC (Instituto de Pesquisa e Planejamento Urbano de Curitiba) e da COMEC (Coordenação da Região Metropolitana de Curitiba), assim como parques e áreas de conservação. Além disso ainda foi gerado um dado para a área de proteção de nascentes e rios, a partir de um *buffer* de 50 e 30 metros, respectivamente, originados de arquivos no

formato *shapefiles* de hidrografia do IAT, AGUASPARANÁ e COPEL. Os respectivos raios dos *buffers* foram definidos seguindo as leis federais n.º 12.651 e n.º 12.727/2012.

FIGURA 11 - Mapa de áreas restritas



FONTE: O autor (2022)

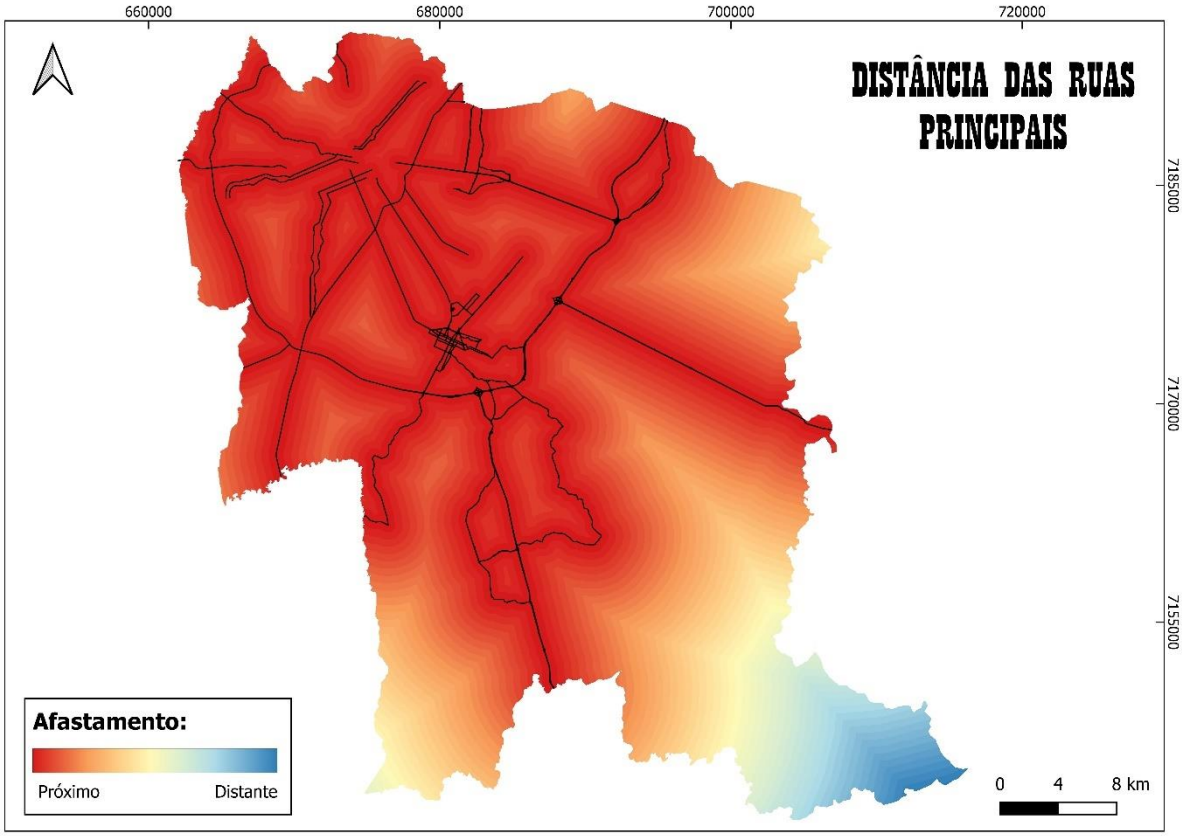
3.2.2.1.4 Distância das ruas principais

Foram obtidas as vias primárias e secundárias pelo *Open Street Map*, devido a seguinte definição para esses dados.

- Primárias: Vias pavimentadas que formam a malha principal de circulação, ligando bairros em uma cidade grande ou média;
- Secundária: Vias em geral pavimentadas que formam a malha secundária de circulação, ligando bairros de uma cidade de grande ou médio porte, ou forma uma malha principal de circulação para cidades pequenas.

Com essas informações criou-se um *buffer* de 500 metros sobre elas, para obter aquilo que está perto ou longe delas.

FIGURA 12 - Mapa das distâncias entre as ruas principais

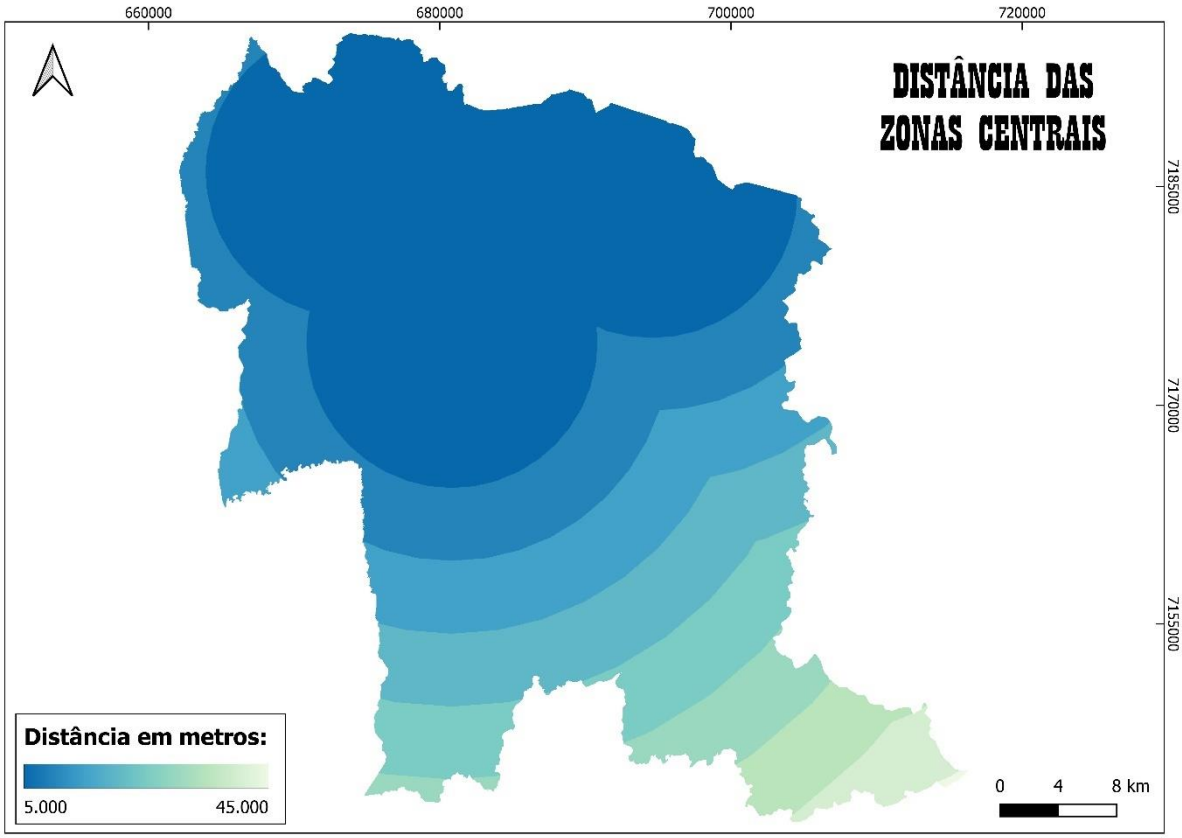


FONTE: O autor (2022)

3.2.2.1.5 Distância das zonas centrais

A partir do zoneamento (obtido do COMEC) da região de interesse, foi encontrada a zona central para cada município, utilizando os seus centroides, em seguida calculou-se uma matriz euclidiana que representasse a distância das zonas centrais através dos seus pixels.

FIGURA 13 - Distância das Zonas Centrais



FONTE: O autor (2022)

4 RESULTADOS E DISCUSSÕES

Nessa seção serão apresentados os resultados obtidos e as análises acerca dos produtos gerados. Em relação aos algoritmos desenvolvidos todos eles estão disponibilizados no Github (site que permite o armazenamento de códigos, por desenvolvedores do mundo inteiro, a fim de que eles possam ser compartilhados e até melhorados pela comunidade). O *link* do repositório para encontrar os *scripts* desse trabalho é https://github.com/GabrielCampigoto/Classifcacao-e-modelagem-de-dados-geoespaciais

4.1 MODELOS GERADOS

Durante todo o período do trabalhado foram gerados diversos modelos de classificação, que levavam desde 30 minutos a várias horas para serem treinados, dependendo dos parâmetros inseridos previamente. Entretanto, no final dessa atividade foram mantidos apenas dois modelos, sendo um para imagens Landsat 5 e outro para o 8.

A princípio esperava-se chegar a apenas um, porém devido as diferenças radiométricas e espectrais, ao realizar os treinamentos, em nenhum momento houve um modelo que conseguisse se adaptar para ambos os casos e tivesse um bom retorno em seus resultados, por isso foi necessário manter essa separação.

Além do que foi mencionado, houve uma grande dificuldade de gerar modelos de classificação que conseguissem predizer corretamente todas as classes definidas anteriormente, no conjunto de imagens selecionadas. Esse fato pode se dar devido a algumas circunstâncias:

- O Landsat 5, ficou em operação mais tempo do que era previsto, pois necessitou substituir a ausência do Landsat 7, logo durante os anos suas imagens tiveram um degradamento de qualidade;
- O Landsat 8, por ter imagens 12 bits, possui muito mais possibilidades de cores, o qual a coleta de amostras as vezes ocasionava no *underfitting*;
- A coleta de amostras se deu de maneira automática, dentro de regiões pré-estabelecidas pelo autor, dessa forma o pixel de

determinados locais podiam não corresponder à classe correta, o qual pode ter influenciado o resultado.

Do que foi comentado, tentou-se ao máximo reduzir os efeitos desses problemas, tomando as seguintes ações:

- Utilizar imagens Landsat 5 que abrangesse tanto a época em que o sensor estava em suas melhores condições, quanto em suas piores;
- Utilizar uma grande quantidade de amostras, a fim de evitar que pontos posicionados em pixels errados, devido a função do QGIS para criar pontos dentro de polígonos de forma aleatória, gerasse muitos erros. Essa solução de obter mais pontos também resolve o problema da resolução radiométrica do Landsat 8, pois haverá mais dados para o modelo de classificação diferenciar as classes.

Descritos estes problemas e suas soluções, durante o treinamento também foram notados alguns padrões, sendo que a quantidade de amostras, por mais que fosse alta, nem sempre resultava em um bom modelo, pois existia a possibilidade de ter *overfitting* (quando o modelo se adapta muito bem as amostras, mas não se adequa a outros dados), além disso notou-se que a inclusão de mais possibilidades nos hiperparâmetros, em geral apenas resultava em mais tempo de processamento, uma vez que os resultados finais obtidos das imagens chegavam a ter resultados muito próximos aos que tinham menos, logo conclui-se que em determinado momento o modelo chegava a um limite, se estagnando em um resultado.

Mencionada isso, a seguir será apresentado quais foram as imagens utilizadas para gerar os modelos.

TABELA 3 - Imagens usadas no treinamento

Satélite	**Imagens**
Landsat 5	• 16/04/1995 • 31/05/2000 • 02/09/2005 • 19/11/2010

Landsat 8	• 13/08/2015 • 14/07/2019 • 13/10/2022

FONTE: O autor (2022)

Desse *dataset* para o Landsat 5, foram coletados 4501 valores amostrais de pixels, enquanto que para o Landsat 8, foram coletadas 33.350. Abaixo é possível observar a aparência dos dados para ambos os casos.

FIGURA 14 - Tabela de amostragem

Amostra tabela Landsat 5

SAMPLE_1	SAMPLE_2	SAMPLE_3	SAMPLE_4	SAMPLE_5	SAMPLE_6	SAMPLE_7	Classe
40	16	13	9	7	118	5	3
40	15	11	9	6	116	3	3
41	16	12	8	7	117	5	3
40	16	12	8	7	118	4	3
40	16	12	8	6	115	5	3

Amostra tabela Landsat 8

SAMPLE_1	SAMPLE_2	SAMPLE_3	SAMPLE_4	SAMPLE_5	SAMPLE_6	SAMPLE_7	SAMPLE_8	SAMPLE_9	SAMPLE_1	Classe
8384	7672	6895	6186	5475	5120	5078	5028	25213	23556	3
8415	7693	7000	6254	5549	5156	5089	5030	25252	23616	3
8396	7675	6902	6214	5535	5162	5101	5030	25166	23497	3
8392	7680	6936	6228	5536	5156	5097	5036	25252	23574	3
8393	7668	6926	6206	5487	5130	5071	5030	25224	23565	3

FONTE: O autor (2022)

Ao final do treinamento, além de verificar os resultados da classificação visualmente, a algoritmo retornava a média da acurácia dos modelos na *cross validation* (técnica para avaliar o desempenho de modelos treinados), o qual permitia ter uma ideia prévia de como os resultados seriam. Os valores encontrados foram:

TABELA 4 - Acurácia dos modelos de classificação treinados

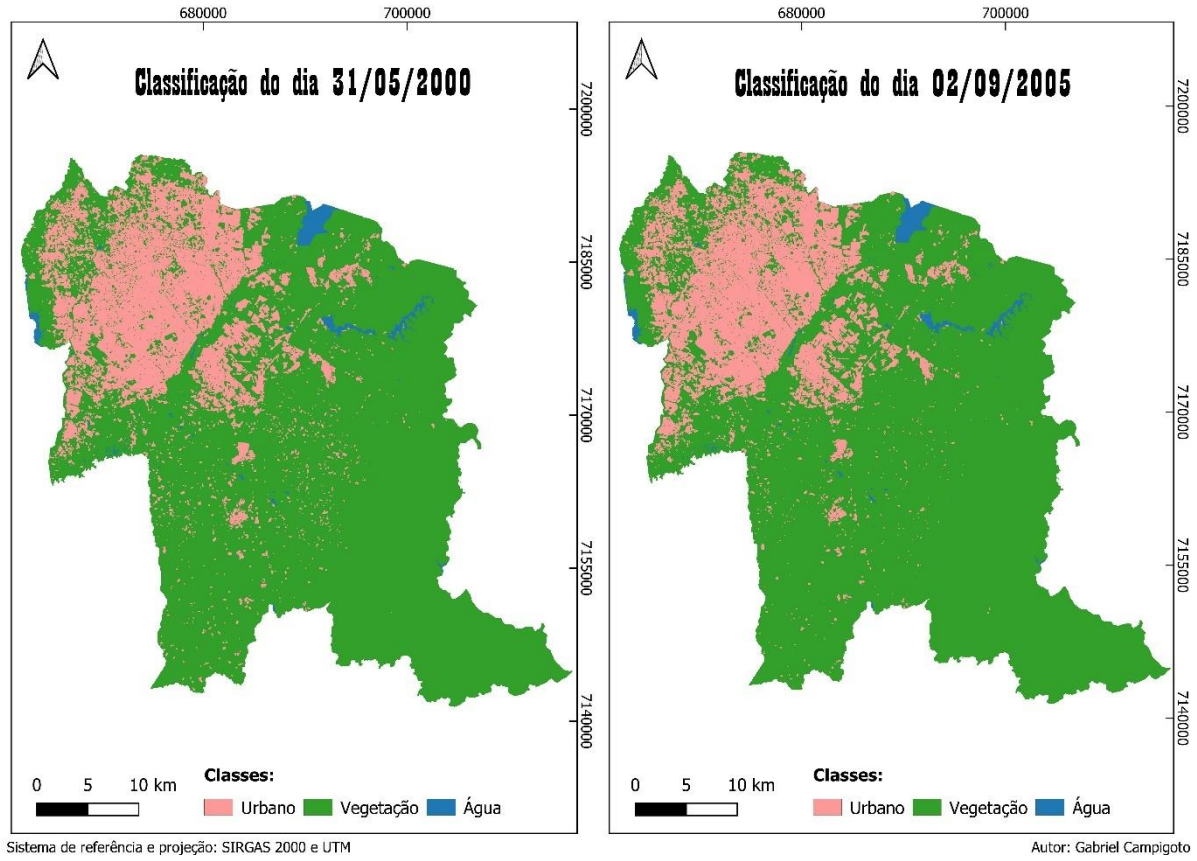
Modelo	**Acurácia**
Landsat 5	95,86%
Landsat 8	72,59%

FONTE: O autor (2022)

4.2 IMAGENS CLASSIFICADAS

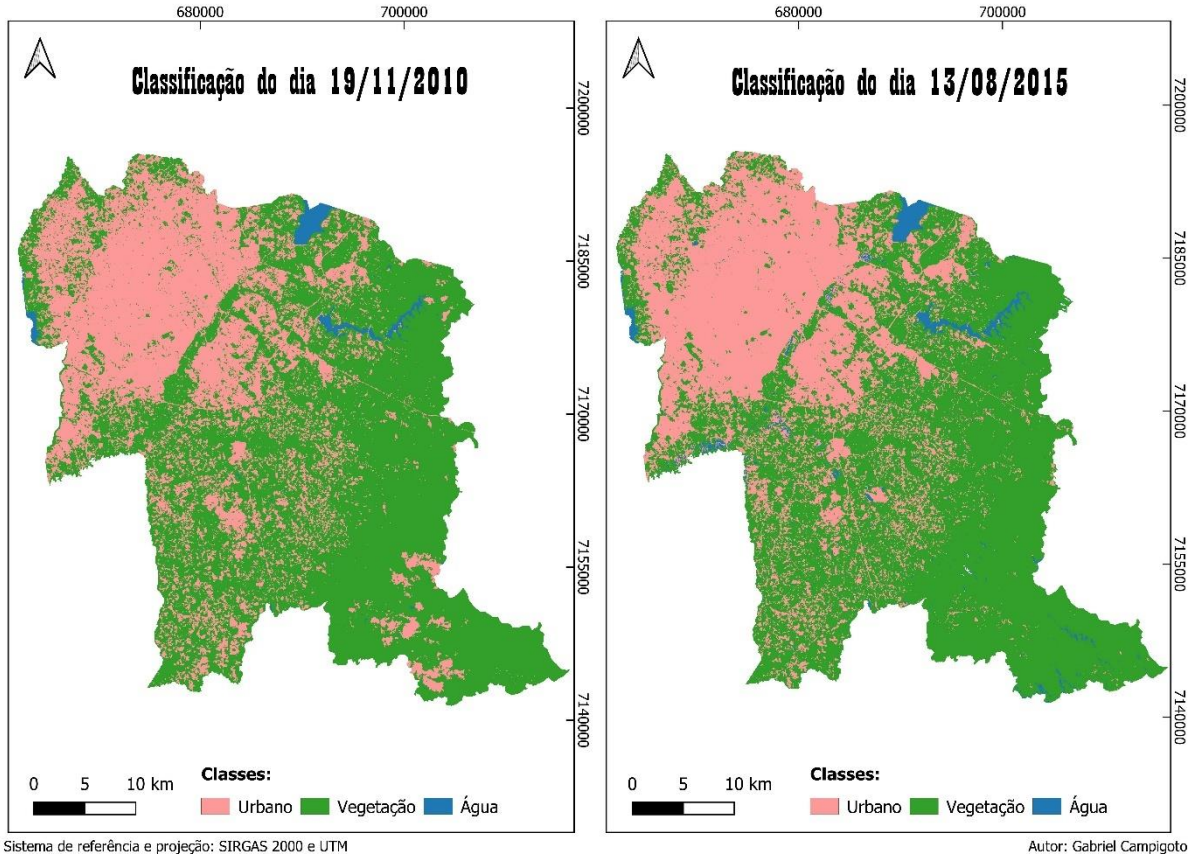
Treinados os modelos para a predição, se deu o processo de classificação de imagens dos anos de 2000, 2005, 2010, 2015, 2019 e 2020 (treino) e dados que estavam fora do *dataset* amostral, como 2009 e 2018. Após realizar a essa etapa ainda foi necessário realizar reclassificação das imagens Landsat 5, a fim de melhorar os resultados obtidos. A seguir será apresentado algumas imagens das imagens classificadas para futuras análises a serem apresentadas.

FIGURA 15 - Classificação de 2000 e 2005



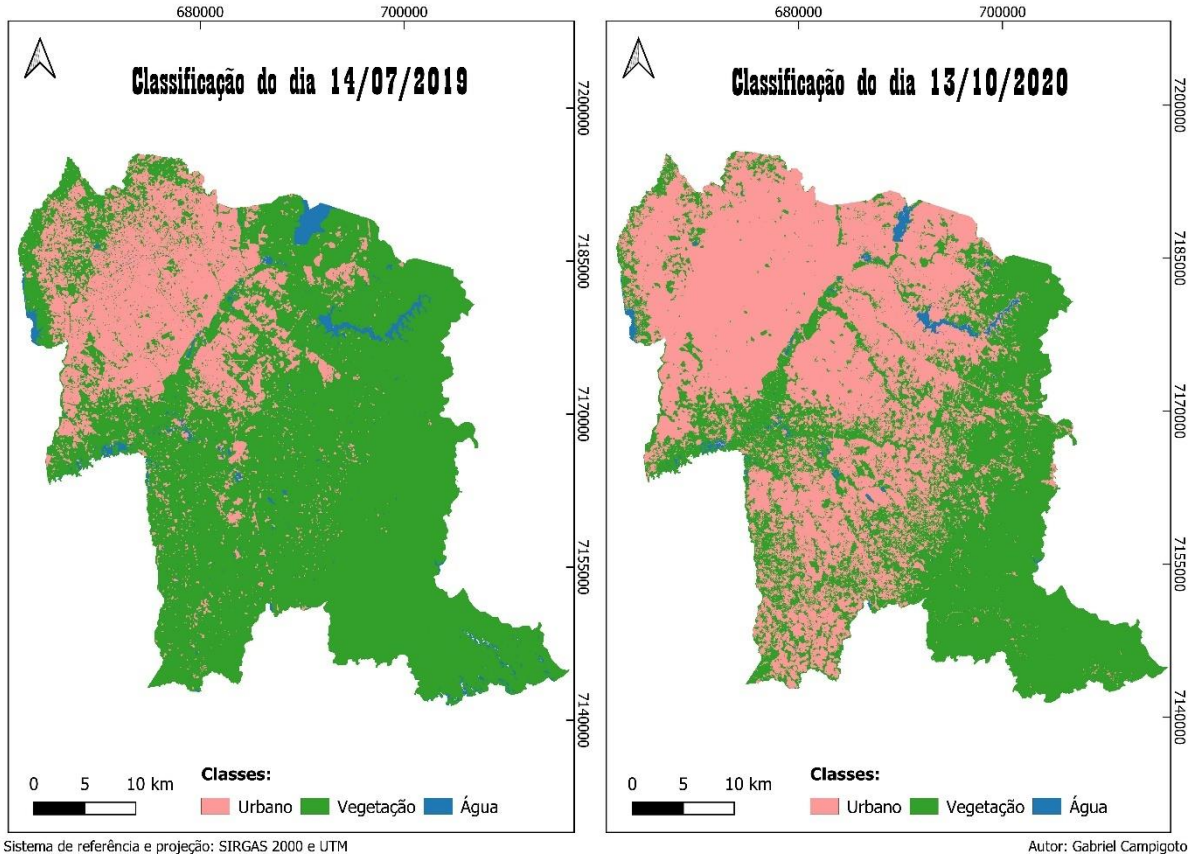
FONTE: O autor (2022)

FIGURA 16 - Classificação de 2010 e 2015



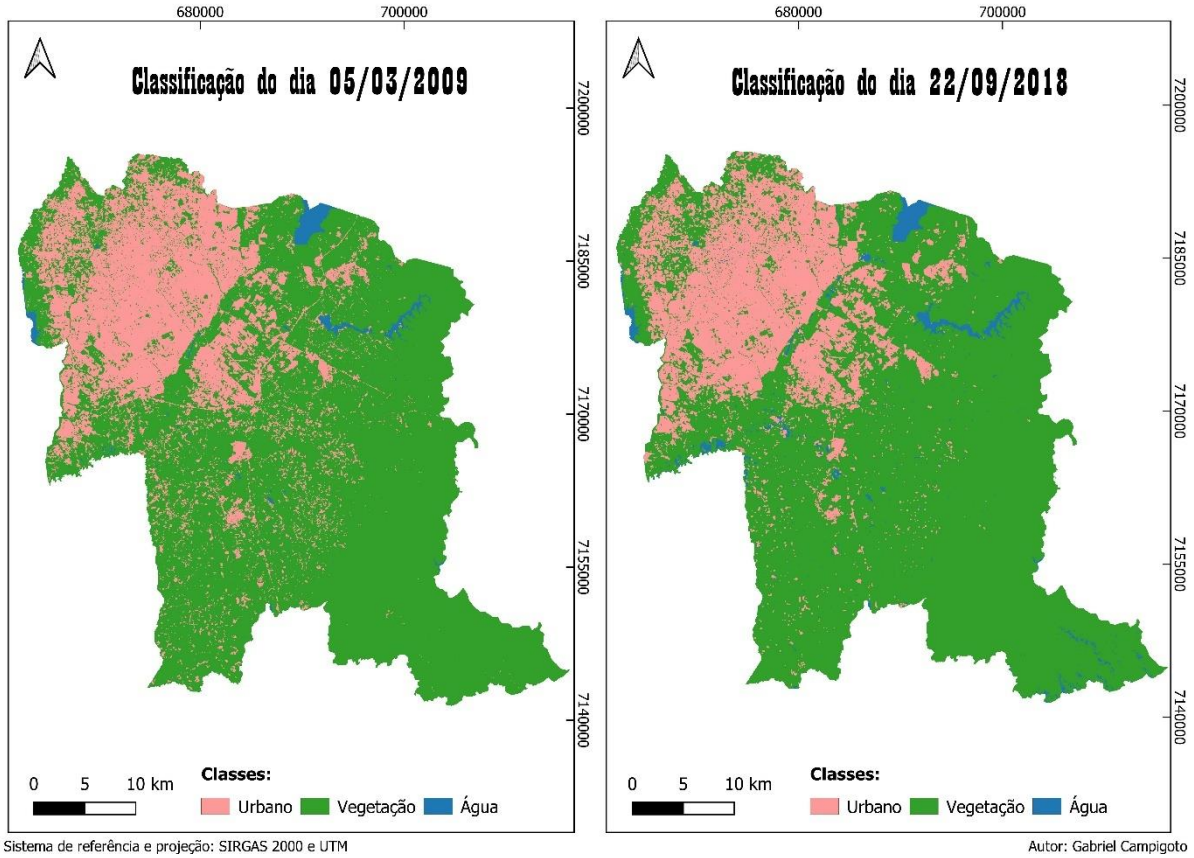
FONTE: O autor (2022)

FIGURA 17 - Classificação de 2019 e 2020



FONTE: O autor (2022)

FIGURA 18 - Classificação de 2009 e 2018

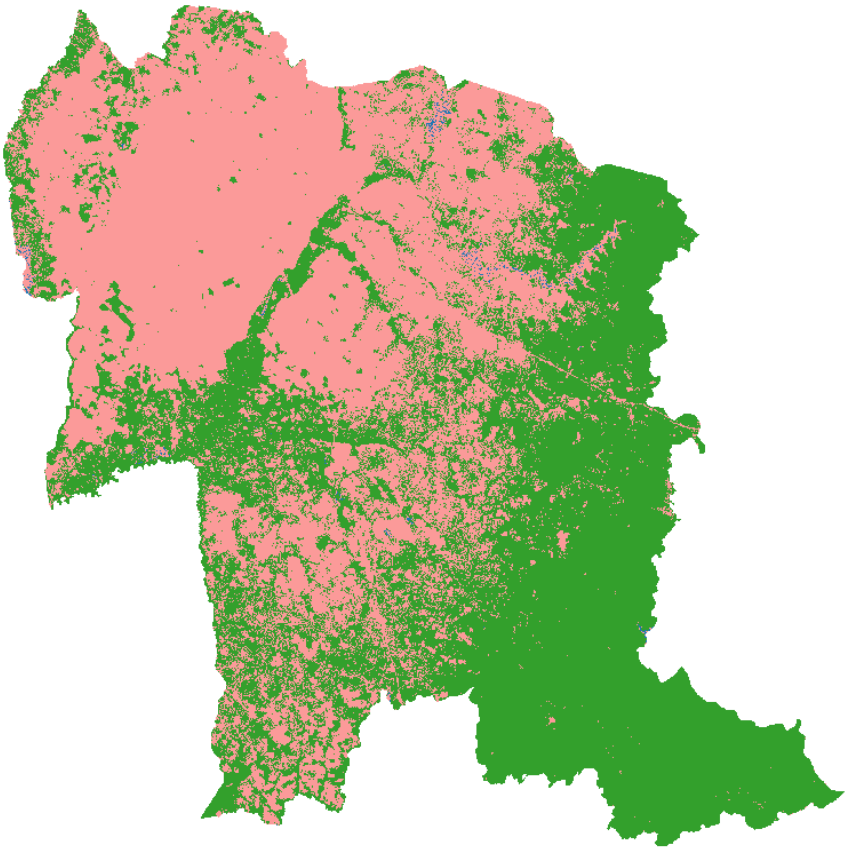


FONTE: O autor (2022)

Como é possível verificar nas imagens classificadas, até 2010, elas apresentaram um comportamento aparentemente gradual através dos anos, com exceção de alguns erros sistemáticos, como a classificação errada das nuvens – como área urbana – porém tirando esses detalhes, pode-se dizer que visualmente elas têm bons resultados. Entretanto as imagens que foram classificadas usando o modelo do Landsat 8, não apresentaram consistência nos seus resultados, acredita-se que como ela possui mais combinações de cores, a amostragem tomada por ela não foi suficiente para separar de maneira adequada as classes, mesmo que seu *dataset* de amostras já seja maior que a do Landsat 5, além disso tem o fator sazonal que impacta nos resultados também.

Durante a geração do modelo para as imagens 12 bits, foram tomadas algumas tentativas, porém cada uma com um resultado pior do que o apresentado, FIGURA 19, sendo assim as análises para o autômato celular ficaram limitadas as de 8 bits, uma vez que o outro modelo de classificação resultaria em informações muito discrepantes da realidade.

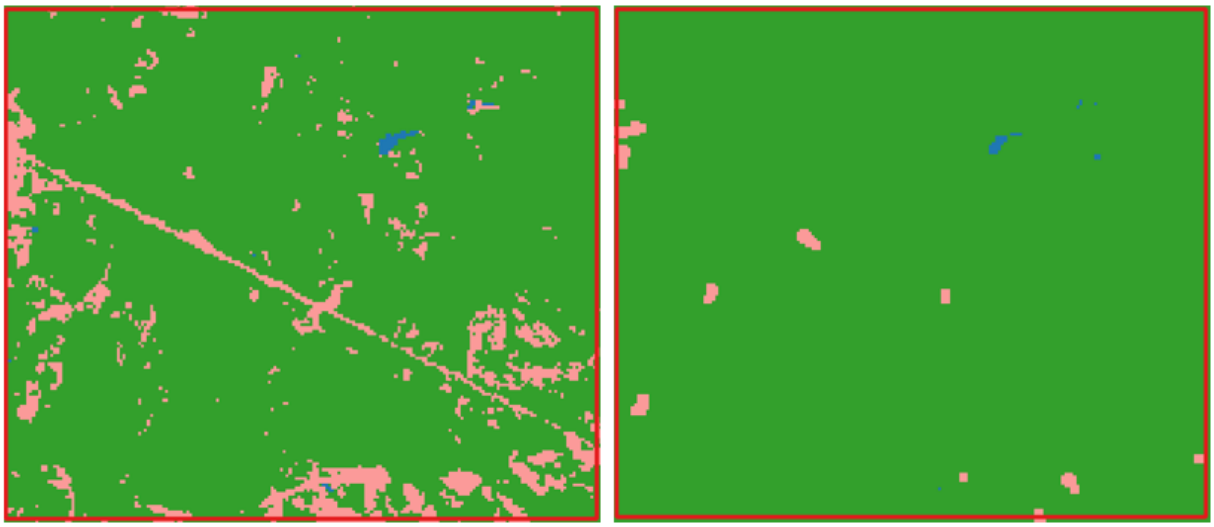
FIGURA 19 - Classificação de 2020, sem água



FONTE: O autor (2022)

Além desses problemas citados, mais voltados a classificação do modelo, no pós-processamento, ao aplicar o fechamento, por mais que muitos ruídos fossem eliminados, informações também eram perdidas, como é o caso das estradas e águas, em que muitos casos trechos delas eram generalizados como vegetação (FIGURA 20). Isso ocorre, pois por mais que o elemento estruturante seja pequeno – 3 x 3 – a resolução do pixel é de 30 metros, o qual acaba abrangendo uma área relativamente grande.

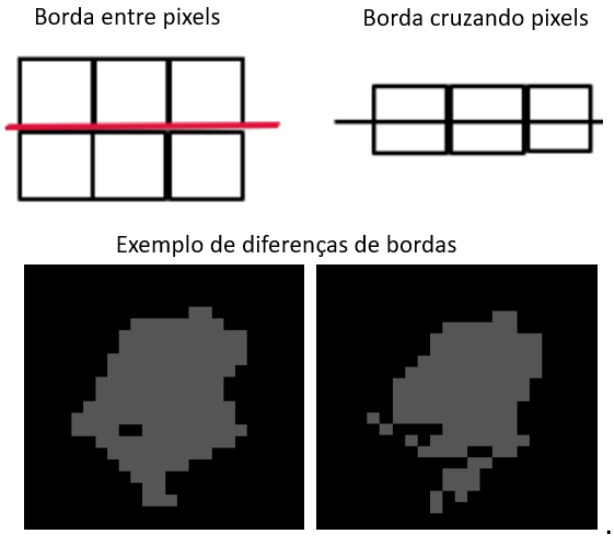
FIGURA 20 - Generalização das estradas e água



FONTE: O autor (2022)

Outra dificuldade enfrentada também, é a diferença do posicionamento dos pixels de uma imagem para outra, pois certas estruturas – em suas bordas – acabavam sendo classificadas de maneira diferente, dependendo de como seu valor foi registrado. Na FIGURA 21 é representado esse fenômeno.

FIGURA 21 - Classificação de bordas



FONTE: O autor (2022)

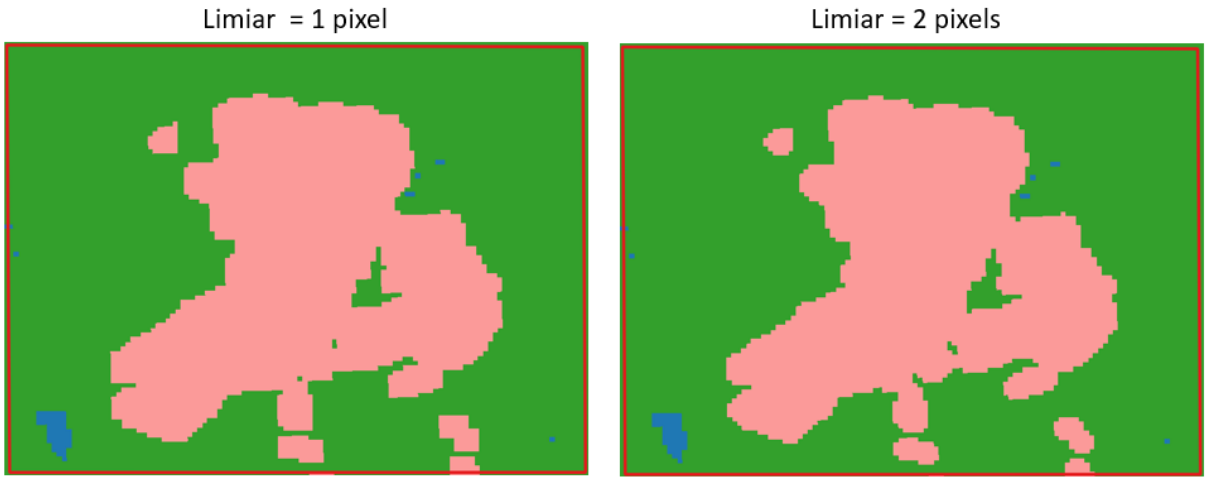
4.3 MODELAGEM DO AUTÔMATO CELULAR

Para realizar a modelagem do autômato celular, primeiramente foi necessário estudar os limiares das regras de transição. Os valores que foram usados são:

- Proximidade das ruas principais: 1500 m;
- Proximidade das zonas centrais: 1500 m;
- Declividade: Analisando a área de estudo, verificou-se que valores menores que 7° é onde está estabelecido a maior parte da área urbana;
- Densidade populacional: 5000 habitantes por polígono da malha censitária, convertida para o formato *raster*;
- Áreas restritas: O limiar nessa regra não foi alterado, pois o valor 1 representa se o pixel está ou não está em uma região restritiva;
- Pixels de área urbana necessários para o crescimento urbano: 3

Nesse estudo foram experimentados diversos valores, contudo não foi notado diferenças em sua modelagem ao mudar os limiares, com exceção ao que é referente a quantidade de pixels necessários para crescimento, que traz alterações. Esse parâmetro ocasiona em maiores áreas urbanas quanto menor seu valor, na FIGURA 22 é possível ver um exemplo disso.

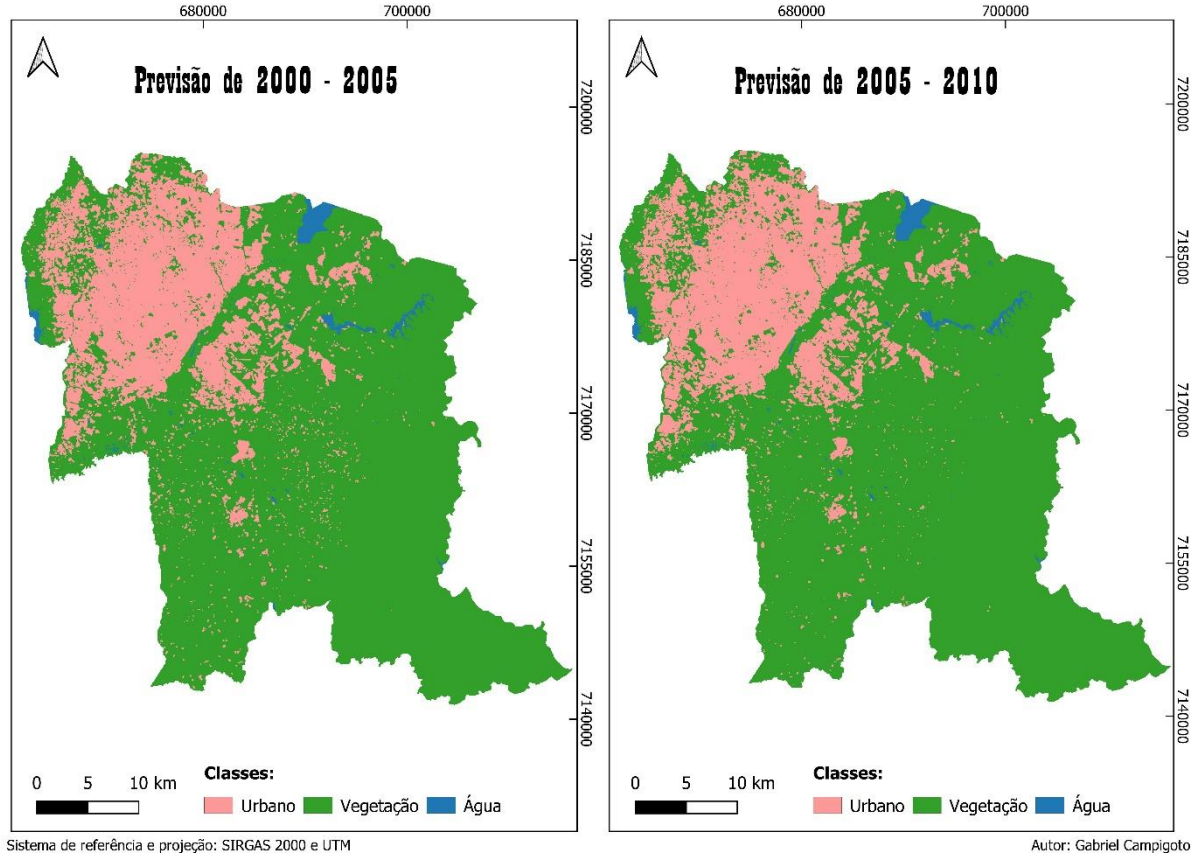
FIGURA 22 - Diferença de limiares



FONTE: O autor (2022)

Sabendo disso, a seguir será apresentado a modelagem com os limiares apresentados anteriormente, uma vez que foi aquele que gerou melhores resultados.

FIGURA 23 - Modelagem para 2005 e 2010



FONTE: O autor (2022)

Além das imagens modeladas pelo autômato celular, o algoritmo de Tripathy e Kumar (2019) também retorna informações de quanto da área urbana foi predita e qual foi realmente o crescimento. Esse processo se dá pela diferença entre as imagens, subtraindo os valores de regiões construídas. A seguir é possível ver esses resultados.

TABELA 5 - Crescimento pelo autômato celular

Período	**Valor predito**	**Crescimento verdadeiro**
2000 – 2005	35 km²	39 km²
2005 – 2010	33 km²	260 km²

FONTE: O autor (2022)

Após o cálculo do valor predito e o crescimento verdadeiro, o código verifica a acurácia espacial do resultado, que é a diferença entre as imagens no local onde é área urbana que resulta em 1 ou -1, essa operação matemática é calculada com base na seguinte equação:

$$acurácia = 100 - [(\sum_{i=1}^{n}\sum_{i=1}^{m}|p_{i,j}^{t_2} - c_{i,j}^{t_2}|/\sum_{i=1}^{n}\sum_{i=1}^{m}(c_{i,j}^{t_2}))x100]$$

Sabendo disso o resultado obtido dessa operação foi:

TABELA 6 - Acurácia da previsão

Período	**Acurácia Espacial**
2000 – 2005	80,782 %
2005 – 2010	59,399 %

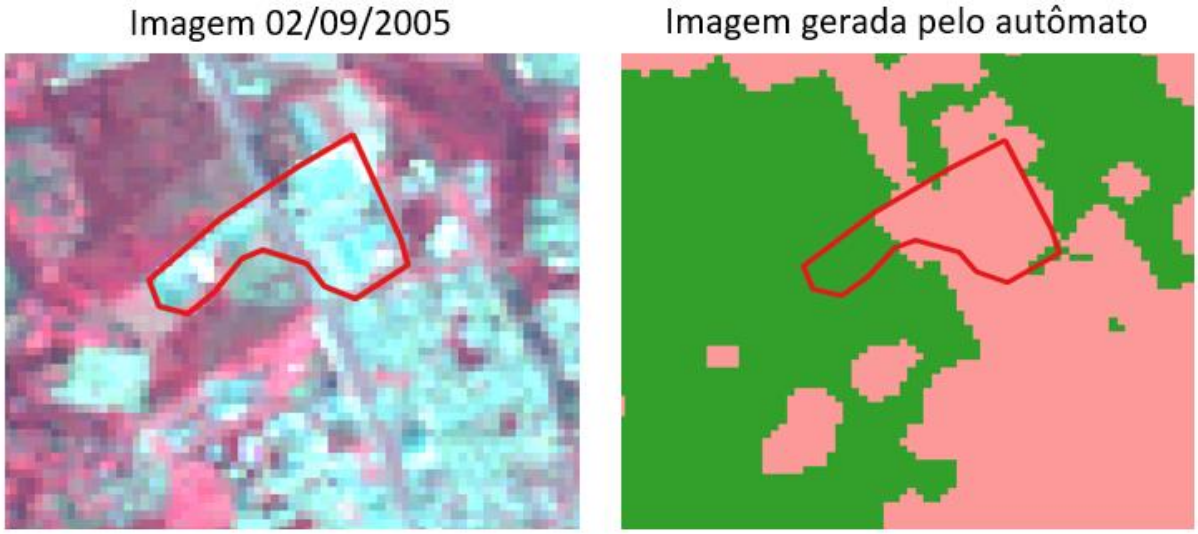
FONTE: O autor (2022)

Analisando os resultados obtidos, percebe-se que 2005 – 2010 teve uma grande discrepância no resultado final, isso é consequência de como o modelo foi treinado, uma vez que nele está generalizado para que regiões de solo exposto e regiões agrícolas com baixa densidade de vegetação, sejam classificadas como região urbana, logo a imagem tomada em 2010, acabou tendo mais áreas urbanas do que a de 2005.

Com essa verificação constata-se que o modelo treinado não é indicado para a classificação de regiões agrícolas, podendo gerar interpretações erradas de como a área urbana está distribuída.

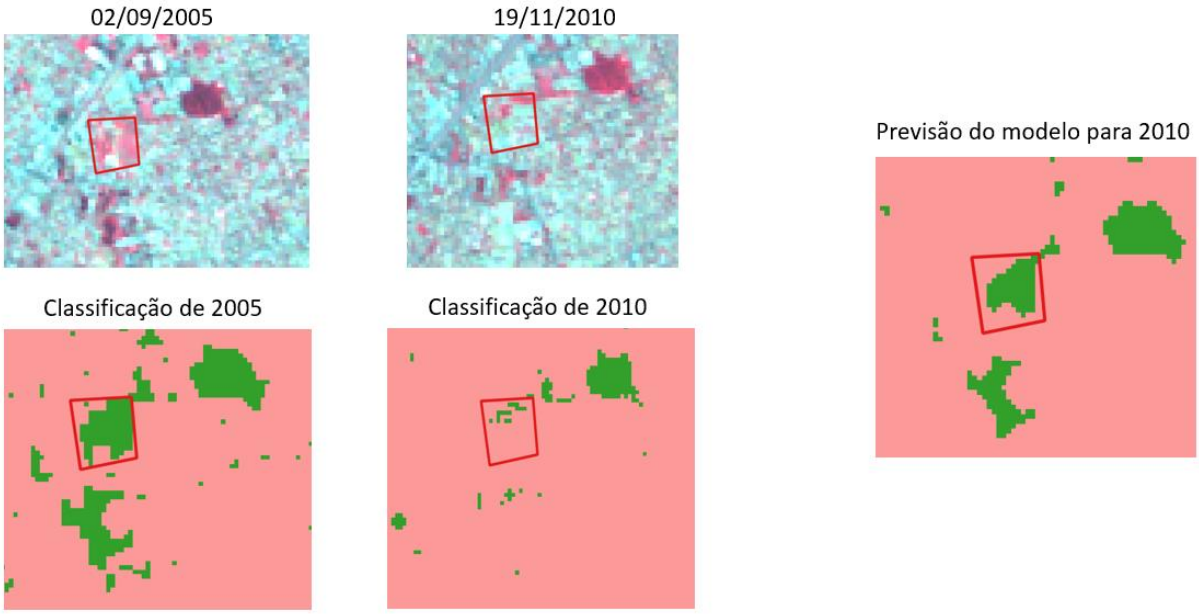
Comparando as imagens modeladas pelo autômato notou-se uma limitação de sua previsão, em respeito ao crescimento da área urbana, pois em certos locais a urbanização foi mais intensa do que o modelo previa. Um exemplo disso está situado ao norte da Cidade Industrial de Curitiba (CIC), como pode ser visto abaixo.

FIGURA 24 - Crescimento urbano inesperado pelo modelo



Em contrapartida foi notado que o autômato conseguiu, de maneira correta, classificar regiões onde futuramente se perderia vegetação, por mais que a região expressa por ele não tenha sido totalmente similar.

FIGURA 25 - Perda de vegetação ao norte do bairro Xaxim em Curitiba



FONTE: O autor (2022)

Sendo assim, os resultados obtidos não são perfeitos e apresentam bastante discrepância, contudo o comportamento da modelagem condiz com a realidade, por mais que a maneira que ele prevê o crescimento urbano aconteça de forma mais

suave. Isso pode estar ocorrendo, pois as regras de transição não são suficientes para predizer essas alterações, assim como elas podem não ter sido geradas de uma maneira que realmente represente o verdadeiro crescimento urbano da região, exemplo disso são as ruas principais e o distanciamento das zonas centrais, o qual as distâncias fornecidas não vieram de um estudo especifico para Curitiba e região metropolitana.

Outra coisa que mostra que o autômato desenvolvido por Tripathy e Kumar (2019) necessita de mais refinamento, foi em relação aos testes feitos para os limiares das regras de transição, o qual em geral não apresentam impactos significativos no resultado final.

5 CONSIDERAÇÕES FINAIS

Dos objetivos propostos a esse trabalho cumpriram-se ambos, entretanto os resultados não foram tão precisos quanto era inicialmente esperado. Uma das consequências disso é que no treinamento dos modelos de classificação, por mais que eles forneçam resultados satisfatórios (não em todos os casos), eles necessitam passar por um pós-processamento, o qual mesmo que retire ruídos (pixels classificados como outra classe em relação ao seu entorno) e corrija dados preditos de maneira incorreta, existe perda de informação pela generalização desse fenômeno.

Além disso existe alguns problemas sistemáticos na classificação das imagens, como a diferença de bordas entre feições, a classificação de nuvens como área urbana e a variação sazonal da assinatura espectral das áreas de cultivo agrícola, que quando as culturas estão desenvolvidas são rotuladas como vegetação, contudo quando se está no início de uma plantação, acaba-se por se confundir com o solo exposto e criar casos de muita área urbana, gerando uma classificação errada na imagem, dificultando na sua interpretação e em seus resultados, que não corresponderão com o desejado. Em relação a sazonalidade a separação da agricultura do que é vegetação e área urbana poderia reduzir os erros da classificação.

Em relação ao autômato ele fez o que era esperado, porém como sua modelagem é muito simples e não foram consideradas todas as condições nas regras de transição, os seus resultados seguiam a tendência de mudanças, observadas nas análises multitemporais, todavia ele não conseguia prever tão precisamente caso houvesse mudanças muito bruscas. Para contornar isso se faz necessário realizar um estudo específico para a região trabalhada, para definir novos limiares que se adaptem melhor ao modelo e a criação de camadas de regras. Outra possibilidade também seria realizar pesquisas para a utilização de outros métodos para a modelagem do crescimento da cidade, por exemplo utilizar redes neurais artificiais que aprendam a detectar e compreender como ocorre a urbanização.

E vale ressaltar que todas as modelagens foram feitas em *Python*, com o auxílio de bibliotecas de funções. A disponibilização dos códigos para a comunidade

pode incentivar a implementação de modificações que resultem em produtos melhores.

5.1 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Como recomendação para futuros trabalhos, acredita-se que trabalhar com reflectância ao invés dos valores do pixel pode gerar melhores modelos de classificação do *Random Forest*. Além disso a realização de estudos para implementação de mais regras para o autômato o tornara cada vez mais complexo, o qual por consequência acredita-se que trará resultados melhores.

Recomenda-se também a realização de estudos para a melhor definição de parâmetros para o autômato na região de interesse, uma vez que encontrando quais são realmente os limiares e locais que tem mais impacto no crescimento, o modelo tende-se a representar mais fielmente a realidade.

REFERÊNCIAS

PESCHL, Henrique. **MODELAGEM DO CRESCIMENTO URBANO COM AUTÔMATOS CELULARES: ESTUDO DE CASO PARA CURITIBA E REGIÃO.** 2021

JUNIOR, Rui; SANTOS, Moacir. **A Urbanização das Cidade.** 2014. Disponível em: <http://www.unitau.br/files/arquivos/category_154/MPH1081_1427392152.pdf>. Acesso em: 14 abr. 2022

Prefeitura de São José dos Pinhais. **Evolução da População.** Disponível em: <http://www.sjp.pr.gov.br/evolucao-da-populacao/>. Acesso em: 14 abr. 2022

KUMAR, A.; TRIPATHY P., **Monitoring and modelling spatio-temporal urban growth of Delhi using Cellular Automata and geoinformatics**, Cities, Volume 90, 2019.

LIMA, Cristina. **A OCUPAÇÃO DE ÁREA DE MANANCIAIS NA REGIÃO METROPOLITANA DE CURITIBA: DO PLANEJAMENTO À GESTÃO AMBIENTAL URBANA - METROPOLITANA**. 2000. 406 f. Tese (Doutorado) - Programa de Doutorado em Meio Ambiente e Desenvolvimento, UFPR, Paraná, Curitiba, 2000.

ZHOU, Yuyu; LI, Xuecao; CHEN, Wei. **An improved urban cellular automata model by using the trend-adjusted neighborhood.** Disponível em: <https://ecologicalprocesses.springeropen.com/articles/10.1186/s13717-020-00234-9> Acesso em 14 de abr. 2022

OTGONBAYAR, Mendbayar; BADARIFU; RANATUNGA, Tharangika; ONISH, Takeo; HIRAMATSU, Ken. **Cellular Atomata Modelling Approach for Urban Growth**. Disponível em: <https://www.jstage.jst.go.jp/article/ras/6/0/6_93/_pdf/-char/en>. Acesso em 14 de abr. 2022

FENG, Yongjiu; LIU,Yan; BATTY, Michael. **Modeling urban growth with GIS based cellular automata and least squares SVM rules: a case study in Qingpu-**

Songjiang área of Shanghai, China. Disponível em: <http://www.spatialcomplexity.info/files/2015/07/Feng-Liu-and-Batty.pdf >. Acesso em 14 de abr. 2022

O que é Machine Learning. Disponível em: https://www.hpe.com/br/pt/what-is/machine-learning.html?jumpid=ps_wm5abvqyzs_aid-520061736&ef_id=Cj0KCQjwjN-SBhCkARIsACsrBz6jDGeU-nxg_Iuuc3DxI8bb1CGYD-rp- >. Acesso em 14 de abr. 2022

A Brief History of Machine Learning. Disponível em: <https://pandio.com/blog/when-was-machine-learning-invented/#:~:text=Many%20genius%20individuals%20contributed%20to,%E2%80%9CMachine%20Learning%E2%80%9D%20in%201952.>. Acesso em 14 de abr. 2022

CAMPOS, Raphael. **Árvores de Decisão**. Disponível em: <https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69 >. Acesso em 14 de abr. 2022

O que é e como funciona o algoritmo RandomForest. Disponível em: <https://didatica.tech/o-que-e-e-como-funciona-o-algoritmo-randomforest/ >. Acesso em 15 de abr. 2022

NETO, Raul. **Introdução à Morfologia Matemática Binária e em Tons de Cinza**. Disponível em: http://www.ime.unicamp.br/~valle/PDFfiles/valente10.pdf. Acesso em 15 de abr. 2022

FERREIRA, Giordano. **Modelos Baseados em Autômatos Celulares para o Planejamento de Caminhos em Robôs Autônomos**. Disponível em: <https://repositorio.ufu.br/bitstream/123456789/12573/1/ModelosBaseadosAutomatos.pdf>. Acesso em 15 de abr. 2022

PITZ, Gustavo. **Curitiba antes da fundação: uma história indígena**. Disponível em: <https://www.turistoria.com.br/curitiba-antes-da-fundacao-uma-historia-indigena>. Acesso em: 16 de abr. 2022

USGS – Earth Explorer. Disponível em:<https://earthexplorer.usgs.gov/>

EMBRAPA. **LANDSAT – Land Remote Sensing Satellite**. Disponível em: <https://www.embrapa.br/satelites-de-monitoramento/missoes/landsat >. Acesso em 19 de abr. 2022

STANGANINI, Fábio; LOLLO, José. **O crescimento urbano da área da cidade de São Carlos/SP entre os anos de 2010 e 2015: o avanço da degradação ambiental**. Disponível em < https://www.scielo.br/j/urbe/a/JvMqH7837GprwMhNd6pVsYw/?lang=pt > Acesso em: 07 de maio de 2022

TEIXEIRA, Isabella. **Avaliação de qualidade de dados geoespaciais: uma abordagem moderna**. 2017

LIMA, Cristina. **REGIÃO METROPOLITANA DE CURITIBA – DESAFIOS SÓCIO-AMBIENTAIS E DE GESTÃO DO DESENVOLVIMENTO SUSTENTÁVEL**. Disponível em: < http://www.tecnologia.ufpr.br/portal/lahurb/wp-content/uploads/sites/31/2019/06/Regiao-Metropolitana-de-Curitiba-Desafios-Socioambientas-e-e-de-Gestao-Rumo-ao-Desenvolvimento-Sustentavel.pdf > Acesso em: 09 de maio de 2022

FRANCISCO, Denise. **DANOS SOCIOAMBIENTAIS URBANOS EM CURITIBA: UMA ABORDAGEM GEOGRÁFICA**. Disponível em: <https://core.ac.uk/download/pdf/328067448.pdf > Acesso em: 11 de maio de 2022

NOJIMA, Daniel. **DINÂMICA RECENTE DA ECONOMIA E TRANSFORMAÇÕES NA CONFIGURAÇÃO ESPACIAL DA REGIÃO METROPOLITANA DE CURITIBA**. Disponível em: <https://www.ipardes.pr.gov.br/sites/ipardes/arquivos_restritos/files/documento/2019-09/dinamica_RMC_primeira_versao_2004.pdf> Acesso em: 11 de maior de 2022