UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

CENTRO ESTADUAL DE PESQUISAS EM SENSORIAMENTO REMOTO E METEOROLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM SENSORIAMENTO REMOTO

MACLEIDI VARNIER

AVALIAÇÃO DE MODELOS DE PREDIÇÃO DE MUDANÇAS DE USO DO SOLO EM TRÊS BIOMAS BRASILEIROS

PORTO ALEGRE

2024

MACLEIDI VARNIER

AVALIAÇÃO DE MODELOS DE PREDIÇÃO DE MUDANÇAS DE USO DO SOLO EM TRÊS BIOMAS BRASILEIROS

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Sensoriamento Remoto como requisito parcial para a obtenção do título de mestre em Sensoriamento Remoto e Geoprocessamento.

Orientador: Prof. Dr. Eliseu José Weber

PORTO ALEGRE

2024

CIP - Catalogação na Publicação

Varnier, Macleidi
AVALIAÇÃO DE MODELOS DE PREDIÇÃO DE MUDANÇAS DE USO DO SOLO EM TRÊS BIOMAS BRASILEIROS / Macleidi Varnier. -- 2024.
109 f.
Orientador: Eliseu José Weber.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Centro Estadual de Pesquisas em Sensoriamento Remoto e Meteorologia, Programa de Pós-Graduação em Sensoriamento Remoto, Porto Alegre, BR-RS, 2024.

1. Modelos de Mudanças de Uso do Solo. 2. Fatores Condicionantes de mudanças. 3. Avaliação de Acurácia. I. Weber, Eliseu José, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Dedico este trabalho ao meu irmão, Matheus Varnier (*in Memorian*), com lembranças dos sorrisos e alegrias compartilhadas.

AGRADECIMENTOS

Aos meus pais, Zelinde Fátima Gollo Varnier e Lidomar Antonio Varnier, que mesmo de longe me auxiliaram de inúmeras formas ao longo de todo o mestrado. Sou grato por acreditarem em mim mesmo nos momentos em que eu mesmo duvidei.

À minha namorada, Dandara Rossato da Silva, que esteve comigo durante esta jornada. Sua companhia tornou meu trabalho mais leve e prazeroso.

Ao meu orientador, Eliseu José Weber, por todo auxílio e disponibilidade durante o trabalho. Agradeço por aceitar a orientação mesmo com tempo reduzido para a execução da pesquisa, bem como proporcionar experiências e contatos valiosos que contribuíram para a melhoria desta dissertação.

Ao meu primeiro orientador, Daniel Capella Zanotta, pelo auxílio inicial na formulação do tema desta pesquisa.

Aos professores Guilherme Garcia de Oliveira e Heinrich Hasenack, pela participação na defesa do plano de dissertação e pelos comentários relevantes que auxiliaram na melhoria do trabalho.

Aos professores Guilherme Garcia de Oliveira, Heinrich Hasenack e Laurindo Antonio Guasselli; por aceitarem participar da banca de defesa da dissertação, pela leitura criteriosa do trabalho e pelos comentários e sugestões que auxiliaram na melhoria do texto final.

Ao professor Robert Gilmore Pontius Jr, por reuniões e sugestões que auxiliaram na melhoria da comunicação dos resultados.

Ao professor Ujaval Gandhi, por auxílios na formulação de alguns códigos computacionais.

Ao Laboratório de Geoprocessamento do Centro de Ecologia da Universidade Federal do Rio Grande do Sul (UFRGS) e ao técnico Allan de Oliveira, por auxílios na instalação do software TerrSet.

Aos meus colegas, que tornaram esta etapa acadêmica prazerosa e divertida. Especialmente a Angelo Del Mestre da Rocha, Beatriz da Rosa Cargnin, Daniele Ghellar de Almeida e Juan Carlos Guerra Blas.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por concessão de bolsa de estudos que possibilitou a realização desta pesquisa.

À Universidade Federal do Rio Grande do Sul (UFRGS), ao Programa de Pós-Graduação em Sensoriamento Remoto (PPGSR), e ao Centro Estadual de Pesquisas em Sensoriamento Remoto e Meteorologia (CEPSRM); pela disponibilização do espaço, professores, funcionários e infraestrutura necessária para cursar o mestrado.

RESUMO

Modelos de mudanças de uso do solo são empregados para predizer alterações esperadas para o futuro. Estes modelos se baseiam no entendimento dos padrões de mudanças ocorridas entre os pontos de tempo t_0 e t_1, seguida de extrapolação para t_x. Apesar da popularidade desta abordagem, em muitos trabalhos a comunicação dos resultados é realizada de maneira inapropriada. Esta situação é liderada por dois motivos: I) a identificação dos padrões de mudanças é realizada de maneira automatizada nos programas de modelagem. E II), a avaliação da acurácia dos resultados dos modelos é comumente realizada utilizando métricas potencialmente enganosas. Como consequência destas práticas, o entendimento do funcionamento e da acurácia dos modelos é dificultada, impossibilitando conhecer a capacidade das predições em indicar as mudanças investigadas. Tendo em vista este cenário, o objetivo principal deste trabalho é avaliar a capacidade preditiva de modelos de mudanças de uso do solo a partir de métricas rigorosas. Para isso, a dissertação foi dividida em dois artigos, em ambos investigou-se áreas de estudo em três biomas brasileiros. No primeiro artigo, foi avaliado como variáveis espaciais preditivas descreveram a supressão vegetal ocorrida em diferentes períodos de tempo. No segundo artigo, foram comparadas as acurácias dos resultados de modelos baseados em aprendizado de máquina e de linha de base para diferentes períodos de extrapolação. Entre os principais resultados encontrados, pode-se destacar que: I) As variáveis preditivas apresentaram diferentes capacidades em descrever as mudanças. II) Os resultados de modelos baseados em aprendizado de máquina e de linha de base obtiveram acurácias similares. E III), a validade dos padrões de mudanças e a acurácia das predições é deteriorada ao afastar-se do período de treinamento.

Palavras-chave: Modelos de Mudanças de Uso do Solo. Fatores Condicionantes de mudanças. Avaliação de Acurácia.

ABSTRACT

Land use change models are employed to predict expected future changes. These models are based on understanding the patterns of changes that occurred between time points t_0 and t_1, followed by extrapolation to t_x. Despite the popularity of this approach, in many studies, the communication of the results is conducted inappropriately. This situation is driven by two main reasons: I) the identification of change patterns is performed automatically in modeling programs. And II), the accuracy of model results is commonly assessed using potentially misleading metrics. As a consequence of these practices, understanding the functioning and accuracy of models is hindered, making it impossible to know the ability of predictions to indicate the modeled changes. Given this scenario, the main objective of this work is to evaluate the predictive capacity of land use change models using rigorous metrics. To this end, the dissertation was divided into two articles, both of which investigated study areas in three Brazilian biomes. The first article evaluated how predictive spatial variables described vegetation suppressions that occurred in different time periods. In the second article, the accuracy of the results from machine learning and baseline models were compared for different extrapolation periods. Among the main results found, it can be highlighted that: I) The predictive variables had different capacities to describe changes. II) The results of machine learning and baseline models obtained similar accuracies. And III), the change patterns and the accuracy of predictions deteriorate as the model moves further from the training period.

Keywords: Land use Change Models. Conditioning Factors. Accuracy Assessment.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	Artificial Neural Network
AUC	Area Under the Curve
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
GLAS	Geoscience Laser Altimeter System
IBGE	Instituto Brasileiro de Geografia e Estatística
ICESat	Ice, Cloud, and Land Elevation Satellite
MLP	Multi-Layer Perceptron
ROC	Receiver Operating Characteristic
SRTM	Shuttle Radar Topographic Mission
TOC	Total Operating Characteristic

SUMÁRIO

1 INTRODUÇÃO

Mudanças de uso do solo são alterações na cobertura da superfície terrestre provenientes de ações humanas e naturais. A redução da cobertura vegetal, o crescimento das cidades e a expansão agrícola são exemplos destas mudanças. Por demonstrarem dinâmicas e formas de apropriação do espaço, o entendimento das transições auxilia no planejamento e organização do território.

Em decorrência disso, no âmbito do sensoriamento remoto, os dados satelitais são utilizados para o mapeamento do uso do solo e identificação de mudanças (JENSEN, 2000; ROGAN; CHEN, 2004). Por este meio, além de espacializar as transições ocorridas no passado, é possível investigar seus padrões de ocorrência e projetar cenários de uso do solo futuro.

Neste sentido, os modelos de predição de mudanças de uso do solo são empregados em diversos campos de estudo (ABURAS; AHAMAD; OMAR, 2019). A elaboração destes modelos se baseia na identificação de padrões e taxas de transição ocorridas entre pontos de tempo t_0 e t_1, e predição de mudanças de t_1 para t_x. Dentre os motivos que levam a sua utilização, em comum há o desejo de identificar padrões eficientes e prever mudanças de maneira acurada (PONTIUS; HUFFAKER; DENMAN, 2004).

A identificação de padrões de transições é realizada a partir de variáveis espaciais preditivas e amostras de mudanças e persistências de uso do solo do período t_0 a t_1. Com base nestes dados, para cada ponto amostral são extraídos os atributos presentes nas variáveis preditivas. Na sequência, utilizando algum método de machine learning, é realizado um treinamento para a identificação das características capazes de diferenciar as amostras. Como resultado deste processo, é obtida uma superfície probabilística indicando os espaços considerados mais e menos propensos a sofrerem mudanças (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; SANGERMANO; EASTMAN; ZHU, 2010; LIN et al., 2011).

Em uma etapa secundária, a espacialização das mudanças esperadas para o futuro é comumente realizada através de algoritmos de alocação de mudanças. Nestes algoritmos, a quantidade de mudanças projetada é derivada da taxa de mudanças obtida no período de t_0 a t_1. Enquanto que a definição das células que passarão por mudanças é normalmente estabelecida através da combinação da superfície de probabilidade, do contexto de vizinhança e de uma superfície com valores aleatórios. Ao fim deste processo, obtém-se um mapa de uso do solo com as mudanças previstas de t_1 para t_x. (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; SHAFIZADEH-MOGHADAM et al., 2021)

Apesar destas etapas serem fundamentais para a elaboração dos modelos, em grande parte dos trabalhos a comunicação dos resultados é precária (VAN VLIET et al., 2016, ABURAS; AHAMAD; OMAR, 2019). Para a identificação de padrões, são raros os artigos que examinam a forma como as variáveis explicam as mudanças (TRIGUEIRO; NABOUT; TESSAROLO, 2020). Em geral, apenas é mencionada a importância relativa das variáveis para o treinamento dos modelos (SANGERMANO; EASTMAN; ZHU, 2010). Já para os mapas de predição de mudanças de uso do solo, em muitos trabalhos não é realizada qualquer tipo de avaliação de acurácia. Enquanto, que em outra parcela são utilizadas métricas potencialmente enganosas (VAN VLIET et al., 2016). Considerando a utilização dos resultados dos modelos como auxílio ao entendimento do uso do solo futuro, é essencial avaliar suas potencialidades e limitações. Apenas deste modo pode-se mensurar a aplicabilidade dos produtos, permitindo conhecer sua utilidade para diferentes objetivos.

Partindo deste contexto, o objetivo geral deste trabalho é avaliar a capacidade preditiva de modelos de mudanças de uso do solo. Para alcançar este objetivo, foram realizados os seguintes objetivos específicos:

- Avaliar o potencial de variáveis espaciais preditivas em explicar a ocorrência de supressão e persistência de vegetação natural;
- Comparar o desempenho de modelos de mudanças de uso do solo baseados em machine learning e modelos de linha de base;
- Analisar o desempenho dos modelos em diferentes períodos de extrapolação;
- Comparar métodos de avaliação da acurácia das superfícies de probabilidade e dos mapas de predição de uso do solo.

2 DESENVOLVIMENTO

2.1 Referencial teórico

2.1.1 *Mapeamento de mudanças de uso do solo*

Mudanças de uso do solo podem ser detectadas através da comparação de mapas de diferentes tempos. Ao comparar uma imagem, índice ou classificação de tempo t_0 com outra de t_1, é possível identificar e quantificar as mudanças ocorridas no período. Dentre os diversos dados que podem ser empregados nesta abordagem, as classificações de uso do solo são corriqueiras (GREEN; KEMPKA; LACKEY, 1994).

Classificações de uso do solo são mapas que detalham as características de cobertura da superfície terrestre. Para sua elaboração, são utilizadas amostras de diferentes classes de uso do solo (floresta, infraestruturas urbanas, corpos d'água, agricultura, etc.) e imagens de sensores remotos e subprodutos (dados espectrais, dados de radar, índices, etc.). Sabendo que as diferentes classes possuem atributos particulares (tonalidade/cor, textura, tamanho, forma, sombra, padrão, localização, etc.), é possível diferenciá-las através de dados de sensores remotos (JENSEN, 2000). O processo de classificação de uso do solo se inicia pela caracterização da amostragem a partir das imagens empregadas. Em seguida, normalmente utilizando algoritmos de aprendizado de máquina, é realizado um treinamento para identificar propriedades capazes de diferenciar cada classe amostrada. Ao fim, as características identificadas no treinamento são extrapoladas para as variáveis, obtendo como resultado um mapa de uso do solo (TALUKDAR et al., 2020).

Apesar da possibilidade de elaboração das classificações, em muitas situações a diferenciação de algumas classes é difícil. Isso ocorre porque algumas classes de uso do solo possuem atributos espectrais e/ou de retroespalhamento similares, não permitindo uma distinção perfeita. Somado a isso, em algumas situações no interior de uma mesma célula podem ocorrer diferentes usos do solo (SHIMABUKURO; SMITH, 1991). Desta forma, é natural que as classificações possuam erros, sendo necessária sua análise.

Para mensurar a capacidade representativa de mapas de uso do solo são empregadas métricas de avaliação de acurácia. Entre eles, pode-se citar a matriz de confusão e suas métricas derivadas, como a acurácia global, o coeficiente kappa, o F1 score, a acurácia do usuário e do produtor, a acurácia específica por classe, os erros de omissão e comissão, etc. Apesar da existência de diversos artigos que demonstram a importância de avaliar a acurácia dos mapas, em muitos trabalhos não é realizado qualquer tipo de investigação. Enquanto que em outros são utilizados métodos potencialmente enganosos, como os problemáticos índices da família Kappa (CONGALTON, 1991; FOODY, 2002; PONTIUS; MILLONES, 2011; FOODY, 2020).

Dentre os métodos citados, Pontius e Millones (2011) recomendam a utilização de matrizes de confusão e a computação das discordâncias de quantidade e de alocação. Nesta abordagem, após a elaboração do mapa de uso do solo, utiliza-se um conjunto de amostras de referência para quantificar as concordâncias e discordâncias da classificação em relação à 'verdade' de campo. Esse conjunto amostral precisa ser definido com base em algum método estatístico de forma a garantir uma adequada representação das classes, além de ser elaborado de maneira independente ao treinamento. Isso permite que a validação seja imparcial e forneça informações específicas para cada classe. Com base nas matrizes de confusão, é possível calcular as métricas de acurácia global, acurácia específica por classe, acurácia do produtor e do usuário, e estimar áreas de maneira não tendenciosa.

No entanto, quando se objetiva avaliar a acurácia de mudanças de uso do solo, é necessário que o conjunto de amostras de referência represente especificamente as transições ocorridas entre t_0 e t_1. A partir desta amostragem, é possível quantificar as concordâncias e discordâncias da referência em relação às mudanças mapeadas, permitindo avaliar as transições de maneira particularizada e informando sobre erros de inclusão e omissão. Entretanto, mesmo que esta técnica seja conhecida e represente um avanço na validação de mudanças, ainda é pouco difundida (PONTIUS, 2022).

Levando em conta que mapas de uso do solo e de mudanças são utilizados em modelos de predição de uso do solo, avaliar sua acurácia possibilita inferir a quantidade de incerteza introduzida na modelagem por meio destes dados.

2.1.2 *Modelagem de mudanças de uso do solo*

Modelos espaciais estão presentes em diversos campos científicos. A compreensão e representação dos padrões de fenômenos espacialmente explícitos motiva sua formulação e utilização. No âmbito do Sensoriamento Remoto, são observadas aplicações em abordagens como inundações, deslizamentos de terra, temperatura de superfície, queimadas, distúrbios de carbono, erosão, mudanças de uso do solo, entre outros (PONTIUS et al., 2007; ADHIKARI; SOUTHWORTH, 2012; PIAO et al., 2022).

Especificamente para os modelos de mudanças de uso do solo, sua formulação é dividida em duas etapas: identificação dos padrões que explicam as mudanças ocorridas no período de t_0 a t_1; e extrapolação dos padrões encontrados para o período de t_1 para t_x. Como entradas necessárias para a modelagem, são utilizadas amostras de ocorrência e não ocorrência de mudanças, e variáveis espaciais com poder explicativo. A criação das amostras baseia-se no mapeamento de mudanças de t_0 a t_1. Já as variáveis explicativas são formuladas a partir de dados espaciais ou de sensoriamento remoto considerando as características da área de estudo e o objetivo do modelo, sendo comum o emprego de variáveis como altitude, declividade, mapas de uso do solo, mapas de distância euclidiana para categorias de interesse; etc. (PONTIUS; HUFFAKER; DENMAN, 2004, SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002, SANGERMANO; EASTMAN; ZHU, 2010.)

A partir destes dados, para cada ponto amostrado são extraídos os valores contidos nas variáveis preditivas com base em suas posições x e y. Desta forma, as amostras de mudanças e não mudanças são caracterizadas de maneira particular, permitindo descrever cada grupo independentemente. Com base nestes valores e utilizando métodos de machine learning, é realizado um treinamento para detectar as propriedades capazes de separar o conjunto amostral. Nesta etapa, a função do classificador é diferenciar, a partir dos valores amostrados, as características que explicam a ocorrência e não ocorrência de mudanças. Como resultado, é obtida uma superfície probabilística variando de 0 (baixa probabilidade de mudanças) a 1 (alta probabilidade de mudanças) (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; SANGERMANO; EASTMAN; ZHU, 2010).

Contudo, a superfície de probabilidade não indica quais células passarão por mudanças, apenas assinala as localidades consideradas mais e menos suscetíveis a alterações. Para a elaboração do mapa de predição de mudanças, é necessário o emprego de algum algoritmo de alocação. Entre os disponíveis, os modelos de autômatos celulares são os mais utilizados (SOARES-FILHO et al 2007). Neste procedimento, a taxa de mudanças projetada de t_1 a t_2 é derivada da taxa de mudanças observada entre t_0 a t_1, assumindo um ritmo de mudanças constante. Já a definição espacial das células que passarão por mudanças é geralmente estabelecida por uma combinação da superfície de probabilidade, do contexto de vizinhança e de uma superfície de valores aleatórios.

O contexto de vizinhança é elaborado a partir de filtros espaciais que contabilizam para cada célula a quantidade de vizinhos classificados como determinada classe de uso do solo em t_1. Normalmente, são contabilizados usos do solo considerados indutores de mudanças, como o caso de usos antrópicos para a predição de desmatamentos e as infraestruturas urbanas para a predição de crescimento das cidades (SHAFIZADEH-MOGHADAM et al., 2021).

Já a superfície de valores aleatórios é utilizada para introduzir variabilidade na modelagem, assumindo que as variáveis preditivas não abrangem todas as fontes indutoras de mudanças. Neste cenário, considera-se que a utilização de autômatos celulares gera melhores resultados do que a mera alocação das mudanças nos maiores graus da superfície de probabilidade (SOARES-FILHO et al 2007). Isso decorreria das mudanças de uso do solo acontecerem em um contexto de paisagem, onde as células transformadas são influenciadas pela sua vizinhança neste processo. Desta forma, a maioria das mudanças poderiam ser explicadas devido à expansão/redução por contato em relação a alguma classe específica.

Ainda, os modelos de mudanças de uso do solo podem ser divididos em dois grupos: modelos preditores de mudanças múltiplas e modelos preditores de mudanças binárias. Modelos de mudanças múltiplas predizem a transição entre diversas classes de uso do solo, com a possibilidade de transições de várias naturezas (ADHIKARI; SOUTHWORTH, 2012; BARLOW et al., 2016; CUSHMAN et al., 2017; CHAVAN et al., 2018; KUCSICSA et al., 2019; VOIGHT et al., 2019). Enquanto que, modelos binários permitem apenas a predição de transições de uma classe para outra, como de não urbano para urbano ou de floresta para não floresta

(SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; HARATI et al., 2019; SHAFIZADEH-MOGHADAM et al., 2021).

2.1.3 *Identificação e avaliação de padrões de mudanças*

A identificação de padrões de mudanças busca encontrar propriedades que explicam as alterações de maneira condicional. Nos modelos de predição de uso do solo, as variáveis espaciais são empregadas para identificar atributos correlacionados a ocorrência e não ocorrência de mudanças. Com base em informações como a proximidade ou distanciamento para objetos espaciais, busca-se caracterizar e diferenciar o grupo amostral (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; SANGERMANO; EASTMAN; ZHU, 2010).

Nos softwares comerciais de modelagem, e mesmo em modelos de código aberto, a identificação de padrões de mudança ocorre de maneira automatizada. Os resultados deste processo são comumente aceitos sem avaliação subsequente, assumindo que os padrões encontrados são válidos (LIN et al., 2011; VAN VLIET et al., 2016).

Porém, conhecendo como ocorre o treinamento dos modelos, algumas dificuldades podem ser elencadas: 1) As variáveis preditivas utilizadas no treinamento podem não possuir relação com as mudanças de uso do solo; 2) As variáveis preditivas podem induzir a identificação de padrões de mudança irreais; 3) As variáveis empregadas no treinamento podem possuir diferentes capacidades de predizer mudanças; 4) Os padrões de mudança identificados no treinamento podem não persistir no período de predição; 5) Diferentes transições de uso do solo possuem padrões de ocorrência distintos; 6) Diferentes áreas de estudo possuem padrões de mudança distintos; E 7) no interior de uma mesma área de estudo os padrões de mudança podem variar.

A partir destas dificuldades, avaliar a coerência dos padrões identificados no treinamento dos modelos possibilita melhor conhecer o seu funcionamento. Um método de investigação possível é a avaliação de fatores condicionantes. Esta técnica se baseia na representação estatística de como as amostras utilizadas no treinamento são descritas pelas variáveis preditivas. Apesar de ser difundida em abordagens como deslizamentos de terra (BRITO et al., 2016), em modelos de predição de mudanças de uso do solo são raras as iniciativas de avaliação das mudanças (TRIGUEIRO; NABOUT; TESSAROLO, 2020). Em decorrência disso, é notável a existência de uma lacuna no entendimento do funcionamento dos modelos por parte da comunidade científica. Como consequência, tem-se que em grande parte dos trabalhos a modelagem é realizada por meio de métodos prontos e os seus resultados não passam por uma avaliação crítica.

Neste contexto, observa-se a necessidade de pesquisas que investiguem as dificuldades expostas, visando a melhor descrição do funcionamento dos modelos. Como método de análise, recomenda-se a investigação dos fatores condicionantes de mudanças.

2.1.4 Avaliação da acurácia de modelos de mudanças de uso do solo

A construção e utilização de modelos de predição de mudanças de uso do solo não garante qualquer tipo de ganho de informação. Para conhecer a capacidade dos modelos em prever mudanças é necessário avaliar seus resultados. As métricas de acurácia empregadas nesta abordagem se dividem em duas partes: avaliação da superfície probabilística e avaliação do mapa de predição de uso do solo.

2.1.4.1 Avaliação da superfície de probabilidade

A avaliação da superfície de probabilidade serve como uma medida do desempenho do treinamento do modelo. Espera-se que no treinamento as amostras representativas de mudanças estejam alocadas nos maiores graus de probabilidade, enquanto as amostras que representam persistência estejam alocadas nos menores graus. Devido à complexidade envolvida na identificação de padrões, é normal que o conjunto amostral de referência não seja perfeitamente dividido pela superfície probabilística. Neste sentido, avaliar a superfície de probabilidade possibilita conhecer a capacidade da modelagem em diferenciar mudanças e persistências (PONTIUS; PARMENTIER, 2013; CUSHMAN *et al.*, 2017).

O método mais utilizado para essa avaliação é a curva ROC (Receiver Operating Characteristic) (Fielding and Bell 1997). A Curva ROC quantifica a taxa de amostras de mudanças classificadas de forma correta (verdadeiros positivos) e incorreta (falsos positivos) para cada grau da superfície de probabilidade. A taxa de verdadeiros positivos é calculada pela equação 1, enquanto que a taxa de falsos positivos é medida pela equação 2.

$$TVP = VP/(VP + FN) \tag{1}$$

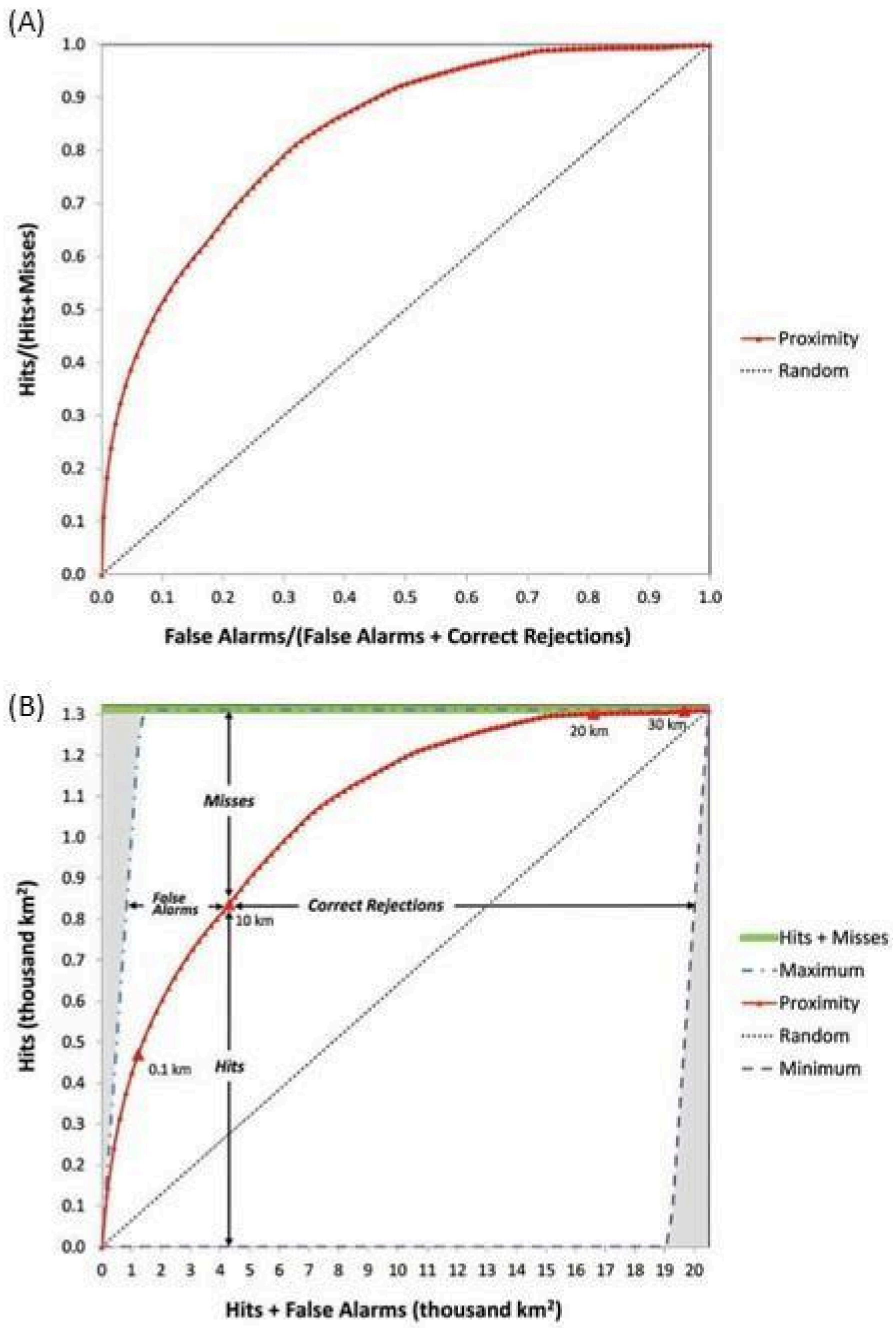
$$TFP = FP/(FP + VP) \tag{2}$$

Em que: a taxa de verdadeiros positivos para cada grau de probabilidade (TVP) é obtida pela divisão da quantidade de verdadeiros positivos (VP) pela soma de verdadeiros positivos e falsos negativos (VP + FN). Enquanto que a taxa de falsos positivos para cada grau de probabilidade (TFP) é obtida pela divisão de falsos positivos (FP) pela soma de falsos positivos (FP) e verdadeiros positivos (VP) (PONTIUS; PARMENTIER, 2014).

Uma maneira de simplificar a análise da curva ROC é a Area Under the Curve (AUC). Nesta abordagem, é agregado o valor das taxas de verdadeiros e falsos positivos de todos os limiares. Em um modelo capaz de diferenciar perfeitamente as amostras de mudanças e não mudanças, o valor de AUC será 1. Enquanto que em um modelo com desempenho igual à aleatoriedade, o valor de AUC será 0.5. Apesar da análise e comparação dos valores de AUC ser muito recorrente, a utilização exclusiva deste método é potencialmente enganosa. Isso se deve a curvas com diferentes formatos poderem receber o mesmo valor de AUC. Desta forma, recomenda-se que além do valor de AUC, também seja analisado o formato da curva e o histograma das amostras de validação em relação à superfície de probabilidade (PONTIUS; PARMENTIER, 2014).

Como aprimoramento da curva ROC, foi proposta a curva TOC (Total Operating Characteristic) (PONTIUS; SI, 2014). Na curva TOC, para cada limite de probabilidade é possível quantificar acertos, erros, falsos alarmes e rejeições corretas. No eixo ‘X’ é apresentada a soma de acertos e falsos alarmes, enquanto no eixo ‘Y’ é apresentado os acertos. A figura 2.1 apresenta as curvas ROC e TOC .

Figura 2.1: (A) ROC curve e (B) TOC curve.



Fonte: PONTIUS; SI, 2014. Adaptado.

Comparando os dois métodos, é notável que a curva TOC possibilita extrair mais informações do que a curva ROC, permitindo ao pesquisador estabelecer um limite de probabilidade que melhor convém ao seu objetivo. Desta forma, é recomendado a utilização da curva TOC em detrimento da curva ROC.

2.1.4.2 Avaliação dos mapas de predição de uso do solo

A avaliação do mapa de predição de uso do solo tem a finalidade de quantificar as concordâncias e discordâncias do modelo em relação ao observado na superfície. Considerando a metodologia da modelagem, o processo de validação possui duas dificuldades principais: 1) Devido aos modelos objetivarem predizer o uso do solo futuro, não há como comparar os resultados com a mudança ‘verdadeira’ observada. Desta forma, a alternativa normalmente utilizada é realizar uma estimativa indireta da acurácia da predição. Por exemplo, para estimar a acurácia de um modelo treinado com dados do período t_2 a t_3 para predizer o uso do solo em t_4, é elaborado um modelo treinado com dados do período t_1 a t_2 para predizer o uso do solo em t_3. A predição de uso do solo no tempo t_3 é comparada com um mapa de uso do solo disponível para o mesmo tempo e assume-se que a acurácia encontrada é válida para a predição do tempo t_4 (VOIGHT et al., 2019) Entretanto, não há garantia de que os padrões e o ritmo de transição encontrados no treinamento de t_1 a t_2 permaneçam em t_2 a t_3, gerando incertezas sobre a validade desta abordagem. E 2) a modelagem de mudanças tem como objetivo predizer transições de uso do solo que na maioria das situações ocupam porções restritas do espaço. Essa condição é relevante pois o mapa preditivo deve ser capaz de representar corretamente tanto transições quanto persistências de classe. Porém, o motivo que justifica a formulação dos modelos é a indicação de lugares que sofrerão mudanças. Logo, para uma validação criteriosa é preciso avaliar as áreas indicadas como mudanças e persistências de maneira particular, possibilitando compreender detalhadamente as discordâncias presentes na modelagem (PONTIUS et al., 2007).

Pode-se ainda, elencar duas práticas da comunidade científica que dificultam a compreensão dos modelos: 1) Muitos modeladores ignoram a avaliação da acurácia ou utilizam métodos reconhecidamente enganosos. Van Vliet et al. (2016) avaliaram práticas de calibração e validação em artigos publicados de 2010 a 2014. Para um total de 114 trabalhos, concluiu-se que em 31% não foi realizado qualquer tipo de validação. Por outro lado, um exemplo notável de utilização de métodos considerados enganosos é o índice Kappa. Apesar de já ter sido extensivamente descrito como insignificante para dados de sensoriamento remoto, continua sendo empregado por muitos pesquisadores (PONTIUS; MILLONES, 2011). Neste contexto, a omissão da avaliação de acurácia ou a utilização de métodos enganosos impossibilita conhecer a utilidade do modelo, reduzindo sua aplicabilidade.

E 2) em muitas predições o tempo de extrapolação dos padrões (t_1 a t_x) é maior que o do treinamento (t_0 a t_1) (KUCSICSA *et al*., 2019, COLMAN et al., 2024). Considerando que as mudanças de uso do solo são dinâmicas e promovidas por inúmeros fatores, quanto mais distante do treinamento for a predição, menor a chance dos padrões encontrados permanecerem válidos. Deste modo, é esperado que a acurácia dos mapas de predição de mudanças diminua ao se distanciar do período de treinamento. Todavia, apesar desta suposição existem poucos trabalhos que investigam essa temática, carecendo de mais estudos para a sua melhor descrição.

Os mapas de predição de mudanças resultantes da modelagem possuem discordâncias que podem ser resumidas em dois grupos: de quantidade e de alocação. Discordâncias de quantidade estão relacionadas à área total predita de mudança. Se no período de treinamento a área de mudanças identificada for a mesma que a do período de extrapolação, as discordâncias de quantidade serão nulas. Porém, na prática isso é muito difícil de ocorrer, visto que os fenômenos modelados geralmente possuem ritmos de mudanças inconstantes no tempo. Se a quantidade de mudanças identificadas no treinamento for maior que a ocorrida no período de extrapolação, haverá uma superestimativa das mudanças. Em sentido oposto, se as transições na etapa de treinamento ocuparem uma área menor que a ocorrida na extrapolação, as mudanças preditas serão subestimadas.

Já as discordâncias de alocação estão associadas à identificação de padrões que explicam a localização das mudanças. Se houver padrões relevantes e identificáveis no período de treinamento e estes se mantiverem no período de extrapolação, haverá uma maior chance das mudanças serem alocadas corretamente. Do contrário, se não houver padrões detectáveis ou se os padrões encontrados no treinamento não se mantiverem na extrapolação, haverá uma maior chance de ocorrerem discordâncias de alocação (PONTIUS et al., 2007; OLMEDO; PONTIUS; PAEGELOW; MAS, 2015).

Na sequência, são apresentadas algumas das principais abordagens de avaliação da acurácia de mapas de predição. Descreve-se a metodologia, as potencialidades e as limitações de cada técnica.

2.1.4.2.1 Acurácia global

A acurácia global é uma medida em porcentagem da concordância do mapa de predição de uso do solo em t_x com um mapa de referência em t_x. Por ser uma abordagem simples e de fácil entendimento, é muito utilizada na validação e comparação de modelos. Todavia, possui algumas limitações (PONTIUS et al., 2007).

Neste método, a contabilização das concordâncias e discordâncias não é estratificada pela área de mudança e persistência. Desta forma, a acurácia global geralmente transmite uma ideia inflada da capacidade dos modelos em prever alterações (PONTIUS; HUFFAKER; DENMAN, 2004). Por exemplo, supondo que as mudanças ocorridas no período de predição de um modelo correspondam a 5% da área de estudo analisada, a projeção de mudanças nulas seria capaz de concordar com 95% do uso do solo. Deste modo, mesmo com o modelo alocando todas as mudanças de forma incorreta, a acurácia global apresenta um valor muito alto. Assim, recomenda-se que a utilização da acurácia global seja acompanhada de outras abordagens, sob o risco de induzir uma interpretação equivocada dos resultados.

2.1.4.2.2 Comparação do mapa de predição de uso do solo com modelo de mudanças nulo, aleatório e de linha de base

A comparação da acurácia do mapa de predição de uso do solo com a de um modelo de mudanças nulo, aleatório ou de linha de base funciona como uma validação relativa. Espera-se que o esforço depositado na formulação dos modelos de predição forneça resultados consideravelmente mais acurados do que predizer a não ocorrência de mudanças, alocar as mudanças de forma aleatória ou alocar as mudanças considerando a proximidade para objetos geográficos. Com essa técnica, torna-se possível analisar se/quanto o mapa de predição de uso do solo foi mais/menos acurado que outros modelos de formulação mais simples (PONTIUS; HUFFAKER; DENMAN, 2004, SHAFIZADEH-MOGHADAM et al., 2021).

2.1.4.2.3 Três possíveis comparações de dois mapas

A avaliação por três mapas é realizada a partir de dois mapas de referência nos estados de tempo t_1 e t_x, e do mapa de predição de uso do solo no estado t_x. Comparando os mapas de referência em t_1 e t_x, indica-se as mudanças e as persistências 'verdadeiras'. Comparando o mapa de referência em t_1, e o resultado da predição em t_x, tem-se a área predita a mudanças pelo modelo. E comparando o mapa de referência em t_x, com o resultado da predição em t_x, obtêm-se a acurácia global do modelo. A partir da análise destas três comparações em conjunto, pode-se quantificar as persistências corretamente preditas pelo modelo (áreas que nos mapas de referência t_1 e t_x não houve mudanças de classe, e foram indicadas pelo modelo como não mudança), os falsos alarmes (áreas onde o modelo previu mudanças, porém na referência houve persistência); as omissões (áreas onde o modelo previu persistência de classe, mas na referência ocorreu mudança) e os acertos (áreas onde o modelo previu mudanças, e na referência ocorreu mudança) (PONTIUS et al., 2007). Com esta abordagem, é possível quantificar os componentes de concordância e discordância de uso e do solo e das mudanças, fornecendo informações que permitem interpretar o resultado dos modelos de maneira mais detalhada que a acurácia global.

2.1.4.2.4 Figura do mérito

Com base na área dos componentes mapeados pela técnica de comparação por três mapas, é possível calcular a figura do mérito. A equação 3 apresenta a formulação desta avaliação.

$$Figura\ do\ mérito = Acertos\ /\ (Omissões + Acertos + Falsos\ Alarmes) \tag{3}$$

Em que: a figura do mérito representa a proporção da área de mudanças previstas corretamente, pela soma da área das discordâncias e mudanças preditas corretamente (PONTIUS, 2018). Comparada à acurácia global, a figura do mérito permite analisar e comparar o desempenho dos modelos especificamente para as áreas de mudanças, fornecendo resultados mais relevantes ao considerar o objetivo dos modelos.

2.1.4.2.5 Três possíveis comparações de dois mapas em múltiplas resoluções[1]

Na avaliação por três mapas em múltiplas resoluções é comparado a predição de uso do solo em t_x com mapas de referência de t_1 e t_x considerando agrupamentos de células. A lógica desta abordagem é que as mudanças ocorrem em um contexto de paisagem, sendo espacialmente relacionadas à vizinhança. Desta forma, mesmo que os modelos possuam erros de alocação de mudanças, em muitas situações a mudança verdadeira ocorre na vizinhança da célula prevista.

Assim, ao avaliar os modelos considerando conjuntos de células, a tendência é que ao aumentar o tamanho do agrupamento a acurácia da predição também aumente. Isso ocorre, pois ao comparar grupos de células os falsos alarmes e as omissões são compensados, restando majoritariamente os erros de quantidade. Todavia, ao degradar a resolução espacial da avaliação também é reduzindo o grau de detalhamento dos dados, impossibilitando sua aplicação em alguns objetivos (PONTIUS; PEETHAMBARAM; CASTELLA, 2011).

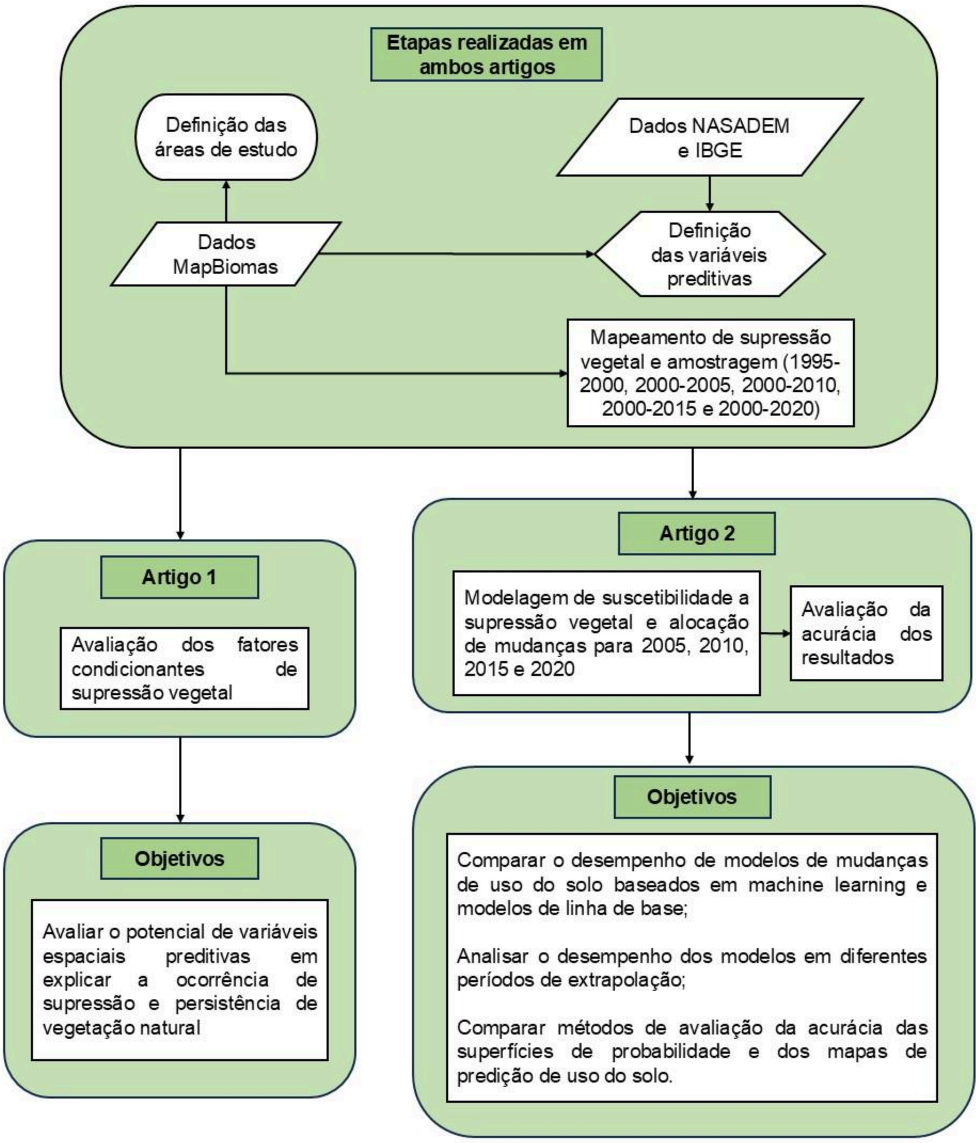
2.2 Metodologia

Para alcançar os objetivos mencionados, esta dissertação foi estruturada em dois artigos. Em cada um dos artigos a metodologia empregada é descrita detalhadamente. Nesta seção, apresenta-se os procedimentos realizados de

[1] Dentre os métodos de avaliação de acurácia apresentados, o único não utilizado neste trabalho foi o de três possíveis comparações de dois mapas em múltiplas resoluções. Isso ocorreu devido a limitação de tempo na formulação desta dissertação. No futuro, pretende-se avaliar os resultados dos modelos a partir desta técnica.

maneira resumida. A figura 2.2 apresenta o fluxograma geral da dissertação.

Figura 2.2: Fluxograma geral da dissertação.



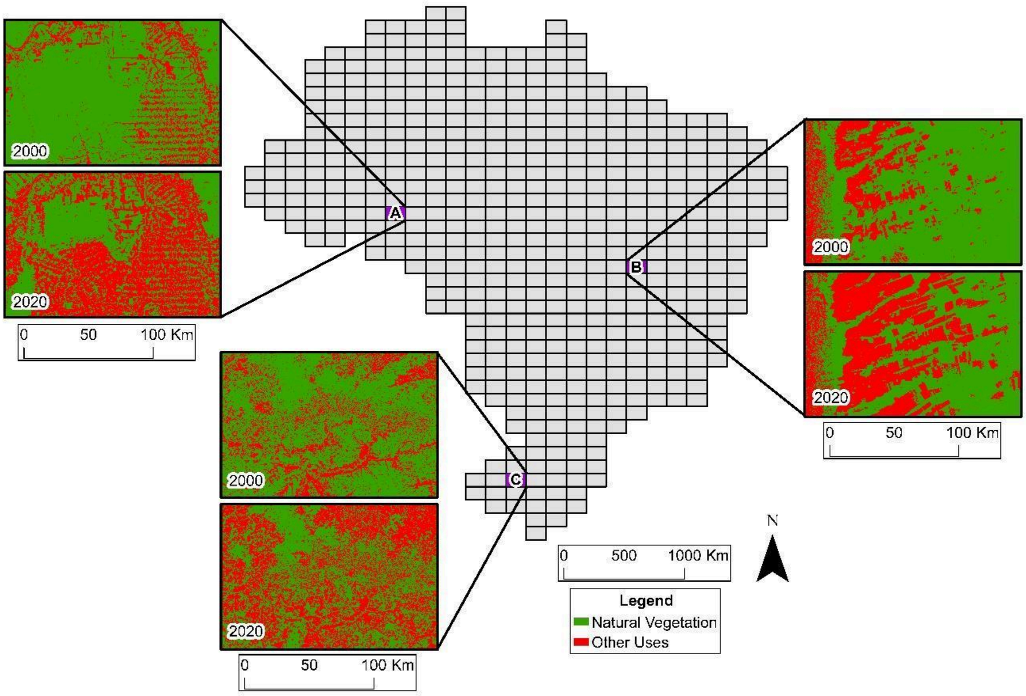
Fonte: Elaborado pelo autor.

Como etapas comuns aos dois artigos, a partir dos dados da coleção 8 do projeto MapBiomas (SOUZA et al., 2020) e na grade 1:250.000 utilizada no processamento destes dados, foi escolhida a célula com a maior supressão vegetal no período de 2000 a 2020 nos biomas Amazônia, Cerrado e Pampa (Figura 2.3). Ainda com base nos dados do MapBiomas, foram mapeadas e amostradas a supressão vegetal ocorrida nos períodos de 1995 a 2000, 2000 a 2005, 2000 a

2010, 2000 a 2015 e 2000 a 2020.

Com base nos dados do MapBiomas, no modelo digital de elevação NASADEM, e em rodovias mapeadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE), foram definidas variáveis espaciais preditivas para representar as mudanças ocorridas no período de 1995 a 2000. As variáveis preditivas foram escolhidas com o objetivo de reproduzir abordagens frequentemente empregadas em trabalhos de modelagem de mudanças.

Figura 2.3: Áreas de estudo e Uso e cobertura do solo em 2000 e 2020. (A) Amazônia, (B) Cerrado, and (C) Pampa.



Fonte: MapBiomas, 2024.

No artigo 1 foi avaliada a forma como as variáveis preditivas descrevem as supressões de vegetação natural ocorridas nos diferentes períodos. O objetivo desta avaliação foi investigar as potencialidades e limitações da identificação e extrapolação de padrões de mudanças em diferentes horizontes temporais.

No artigo 2 foram avaliadas e comparadas as acurácias de modelos de predição de mudanças de uso do solo baseados em machine learning e de linha de base. Os modelos foram treinados utilizando as variáveis preditivas do período de 1995 a 2000, para predizer o uso do solo em 2005, 2010, 2015 e 2020. Os resultados da modelagem foram avaliados através da comparação com dados de referência do MapBiomas. O objetivo desta abordagem foi analisar o desempenho dos modelos utilizando métodos rigorosos, além de compreender a acurácia das predições em diferentes horizontes de extrapolação.

3 ARTIGO 1: Spatial Variables and Land Use Change Models: A Study on Conditioning Patterns of Natural Vegetation Suppression and Persistence

Abstract

Land use change models are formulated based on the identification of patterns associated with changes and persistences. In modeling software, this stage is carried out by characterizing changes and persistences based on spatial variables. Despite being an essential stage in land use change prediction, the evaluation of the change and persistence patterns is often neglected by the scientific community. As a result, it is difficult to understand how the models work and their limitations. Thus, this study evaluated the conditioning factors of vegetation suppression and persistence in three study areas in different Brazilian biomes. The characteristics of vegetation suppressions and persistence were investigated for five different time periods, 1995 to 2000 (representing training) and 2000 to 2005, 2000 to 2010, 2000 to 2015 and 2000 to 2020 (representing the extrapolation of the patterns identified in the training). The spatial variables used to identify the patterns were formulated to represent the environmental context of the training period (1995 - 2000). The method used to analyze the data was Violin Plot graphs. Among the main results, it was shown that 1) In a group of variables, some have a greater ability to differentiate between vegetation suppression and persistence; 2) The farther the extrapolation is from the training period, the ability of the variables to differentiate between vegetation suppression and persistence decreases; and 3) Vegetation suppression and persistence in different study areas are described by the variables in distinct ways. As possible recommendations, it is highlighted that modelers analyze patterns of change and persistence using statistical techniques. This makes it possible to understand how the models work and to define the variables used in training based on measurable data.

Keywords: Predictive variables, Land use change mapping, Spatial modeling, Violin Plot graphs.

3.1. Introduction

Land use change models are used to project expected changes in the future. These models are developed by identifying patterns that explain changes between two time points, t_0 and t_1, and then extrapolating from t_1 to t_x (PONTIUS; HUFFAKER; DENMAN, 2004). To build these models, samples of change and persistence, spatial variables with explanatory capabilities, and machine learning methods are utilized. For each sampled point, the information from the spatial variables is extracted based on the corresponding location. Next, a machine learning method is commonly used to identify the characteristics that can differentiate the training sample set. The result of this process is a probabilistic surface indicating the areas considered most and least susceptible to future changes (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; SANGERMANO; EASTMAN; ZHU, 2010; LIN et al., 2011).

Normally, the definition of the spatial variables used to train the models is based on researchers' knowledge of the processes of change in the study area being analyzed. A wide range of variables can be found, with a particular emphasis on geospatial data that varies continuously across space. The most commonly used variables include Euclidean distance to spatial objects such as roads, conservation units, and specific land uses, as well as terrain attributes like altitude and slope (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; VAN VLIET et al., 2016).

The identification of change and persistence patterns aims to find properties that conditionally explain these occurrences. In land use change models, spatial variables are employed to identify attributes correlated with occurrence and non-occurrence of changes. Using information such as proximity or distance to spatial objects, the goal is to characterize and differentiate the sample group (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; SANGERMANO; EASTMAN; ZHU, 2010). In commercial and open-source modeling software, the identification of change patterns is automated. The results of this process are commonly accepted without further evaluation, assuming that the patterns found are valid (LIN et al., 2011; VAN VLIET et al., 2016).

However, knowing how the training process works, some difficulties can be listed: 1) The variables used in the training may have different capacities for predicting change (ADHIKARI; SOUTHWORTH, 2012; CHAVAN et al., 2018); 2) The predictive variables used in the training may not be related to land use changes (ADHIKARI; SOUTHWORTH, 2012; CHAVAN et al., 2018); 3) The patterns of change identified in the training may not persist over the prediction period; 4) Different study areas have different patterns of change (TRIGUEIRO et al., 2019); 5) within the same study area, patterns of change may vary (TRIGUEIRO et al., 2019); and 6) Different land use transitions have different patterns of occurrence (KUCSICSA et al., 2019).

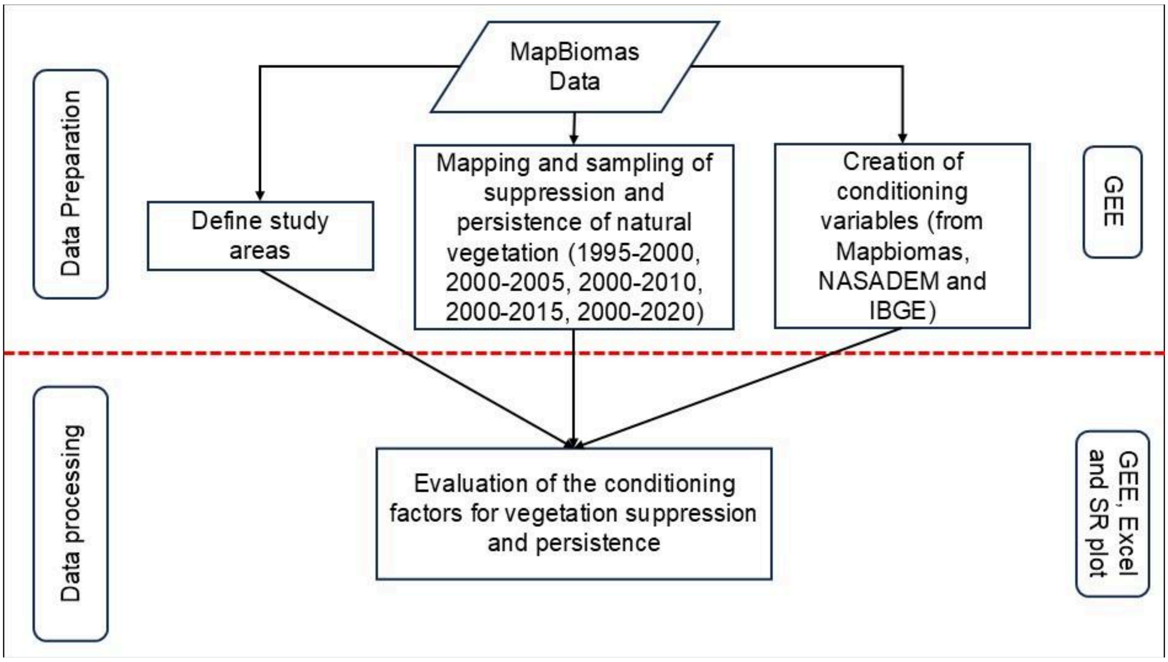
Given these complications, assessing the coherence of the identified patterns enables a better understanding of how the models work. One possible method of investigation is the evaluation of conditioning factors. This technique is based on the statistical representation of how the samples used in training are described by the predictive variables. While this technique is widely used in studies like landslide assessment (BRITO et al., 2016), initiatives to evaluate patterns of land use change are rare (TRIGUEIRO; NABOUT; TESSAROLO, 2020). Consequently, there is an important gap in the scientific community's understanding of how models work. Therefore, most modeling is carried out using ready-made methods and the results are not critically assessed.

So, it is necessary to carry out research that investigates the difficulties exposed in order to better describe the functioning of the models. Considering the above, this work aims to provide a basis for evaluating the behavior of predictive variables in land use change models under different environmental conditions and time periods.

3.2. Materials and methods

The study was conducted in selected areas of three Brazilian biomes, using land cover and land use data from MapBiomas and predictor variables derived from MapBiomas, NASADEM (digital elevation model) and the Brazilian Institute of Geography and Statistics (IBGE). The methodology comprised two main stages, executed with the assistance of Google Earth Engine (GEE) (GORELICK et al., 2017), Excel (Microsoft Corporation, 2016), and SR plot (Tang et al., 2023). Figure 3.1 shows the flowchart of the work methodology. In the central part of the figure are the processes carried out, on the left are the stages of the work, and on the right are the software used.

Figure 3.1: Flowchart of the methodology.



Source: Prepared by the authors.

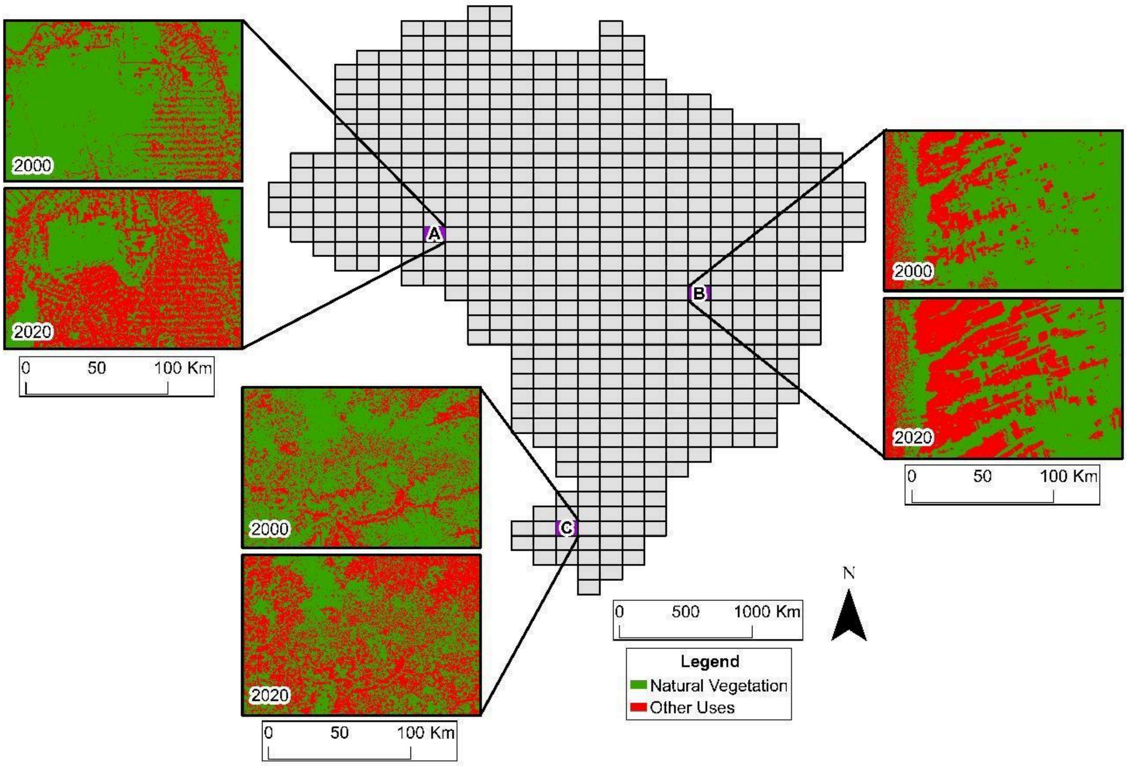
3.2.1 Study area

The study areas were selected to represent different environmental contexts that experienced land use changes and reductions in natural vegetation between 2000 and 2020. In three Brazilian biomes, areas were selected that exhibited large land use changes during this period. The areas were chosen based on land use maps from collection 8 of the MAPBIOMAS project (SOUZA et al., 2020) and the 1:250,000 grid used to generate this data.

To select the study areas, we first calculated the area of the natural vegetation classes for the years 2000 and 2020 in each grid cell of the map. Subsequently, we calculated the difference in the area of these vegetation formations between 2000 and 2020. Finally, we chose the grid cell with the largest absolute reduction in natural vegetation area during the analyzed period in the Amazon, Cerrado, and Pampa biomes. These three biomes were selected because they have distinct environmental characteristics and have undergone major land use changes during the period analyzed.

The selected study areas cover a variety of spaces and dynamics of land use change. In this way, it is possible to examine the factors that condition the process of vegetation suppression in areas with diverse characteristics. Figure 3.2 shows the location of the three study areas.

Figure 3.2: Study areas and land use and land cover in 2000 and 2020. (A) Amazon, (B) Cerrado and (C) Pampa.



Source: Prepared by the authors.

The Amazon biome study area is located in the state of Rondônia, in northern Brazil, and is characterized by the predominance of dense, humid tropical forest, contrasting with extensive areas of pasture derived from recent deforestation. In the period from 2000 to 2020, there was a reduction of approximately 40 per cent in the area of natural vegetation, with deforestation advancing from south to north. The suppression of the Amazon rainforest is primarily driven by timber harvesting and the expansion of livestock and agricultural production. Today, this process is regarded as a global environmental issue, attracting attention from various sectors of society (SOUZA et al., 2020).

The Cerrado biome study area is located on the border of the states of Bahia and Goiás, in the central portion of Brazil, and is characterized by mosaics of savannah, grassland, agriculture and pasture. In the period from 2000 to 2020, there was a reduction of approximately 28 per cent in the area of natural vegetation, with suppression progressing from west to east. The reduction of natural vegetation in this part of the Cerrado is motivated by the implementation of large-scale monocultures, a process that has received attention from the scientific community in order to understand its environmental impacts (SOUZA et al., 2020; PONTIUS et al., 2023).

The Pampa biome study area is located in the state of Rio Grande do Sul, in southern Brazil, and is characterized by grassland vegetation, pastures, agricultural uses, and forested areas. Between 2000 and 2020, natural vegetation decreased by approximately 29%, with suppression occurring without a defined spatial pattern. The conversion of natural vegetation in this part of the Pampa is primarily driven by the establishment of agricultural crops, particularly the shift from grassland formations to soybean plantations. Between 2000 and 2020, soybean production expanded extensively in the biome, considerably altering the landscape. Another important process is the periodic alternation of land use classes, with transitions from grassland vegetation to pasture and rice production, and vice versa (SOUZA et al., 2020).

3.2.2 Data used

All the geospatial data used in this study comes from open sources and is accessible for consultation. The land use maps, used to map change and persistence and to formulate Euclidean distance variables, were derived from the MAPBIOMAS project. MAPBIOMAS is a collaborative effort involving researchers from various Brazilian institutions, including universities, NGOs, research institutes, and technology start-ups. One of its main products is the annual land use and land cover maps, available since 1985 for the entire Brazilian territory. These maps are generated using data from the Landsat program, offering a spatial resolution of 30 meters. A key strength of this mapping is the robust accuracy assessment, allowing users to understand the disagreements in the classification for each land use category (SOUZA et al., 2020; MAPBIOMAS, 2023).

The altitude and slope data were derived from the NASADEM digital elevation model. NASADEM is a product resulting from the reprocessing of the Shuttle Radar Topographic Mission (SRTM) with a spatial resolution of 30 meters. Compared to SRTM, NASADEM provides greater accuracy due to processing improvements and the incorporation of auxiliary data from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) and the Ice, Cloud, and Land Elevation Satellite (ICESat)/Geoscience Laser Altimeter System (GLAS) (NASA, 2023).

The road network was derived from the Brazilian Institute of Geography and Statistics (IBGE), representing federal and state roads. The scale of this data is 1:250,000 and the features were delimited using images from the Sentinel-2, Planet and Maxar sensors. During the validation of the product, an average positional error of 125 meters was considered acceptable (IBGE, 2024).

3.2.3 Data preparation

Data preparation was conducted on the GEE platform and involved three main stages: (1) reclassification of land cover and land use maps from the MapBiomas project; (2) mapping and sampling of natural vegetation suppression and persistence; and (3) preparation of predictive variables.

3.2.3.1 Reclassification of land use and land cover maps

The land use and land cover maps from the MapBiomas project were reclassified in order to group the original classes into just two: natural vegetation and other uses. The land use maps for the years 1995, 2000, 2005, 2010, 2015, and 2020 were selected for reclassification. These specific years and intervals were selected to ensure that the training period would match the duration of each extrapolation period, enabling the evaluation of prediction results across time intervals of equivalent durations.

3.2.3.2 Mapping and sampling of natural vegetation suppression and persistence

The mapping of vegetation suppression and persistence was based on the reclassified land use maps. Five different periods were mapped: 1995 to 2000, representing training; and 2000 to 2005, 2000 to 2010, 2000 to 2015 and 2000 to 2020, representing extrapolation.

The method used was based on comparing two land use maps from different times. Equation 1 illustrates this process.

$$NVSP = Xt_0 - Xt_1 \tag{1}$$

where natural vegetation suppression and persistence (NVSP) is obtained by comparing (−) the state of the cells at time point t_0 (Xt_0) with the state of the cells at time point t_1 (Xt_1). By applying this function to the entire study area, a mask is obtained indicating the areas of vegetation suppression and persistence for the period analyzed.

Based on this mapping, a random sampling of 10,000 points was carried out for each class (suppression and persistence of natural vegetation).

3.2.3.3 Preparation of predictive variables

From the MapBiomas project, Euclidean distance surfaces were created for the water bodies in 1995, urban space in 1995, anthropogenic uses in 1995 and vegetation suppression between 1990 and 1995. NASADEM provided altitude and slope data. And from the IBGE, a Euclidean distance surface was calculated for the roads existing in 1995.

3.2.4 Evaluation of the conditioning factors for natural vegetation suppression and persistence

Based on the set of samples and the predictive variables, the attributes that explain the occurrence and non-occurrence of vegetation suppression were evaluated. For each suppression sample, the values contained in the predictive variables were extracted based on their x and y positions. This data was then organized in a table, and violin plots were generated to represent the distribution of values received by the samples.

Figure 3.3 shows a comparison between Box Plot and Violin Plot graphs.

Figure 3.3: Comparison between Box Plot (graphs on the left) and Violin Plot (graphs on the right).

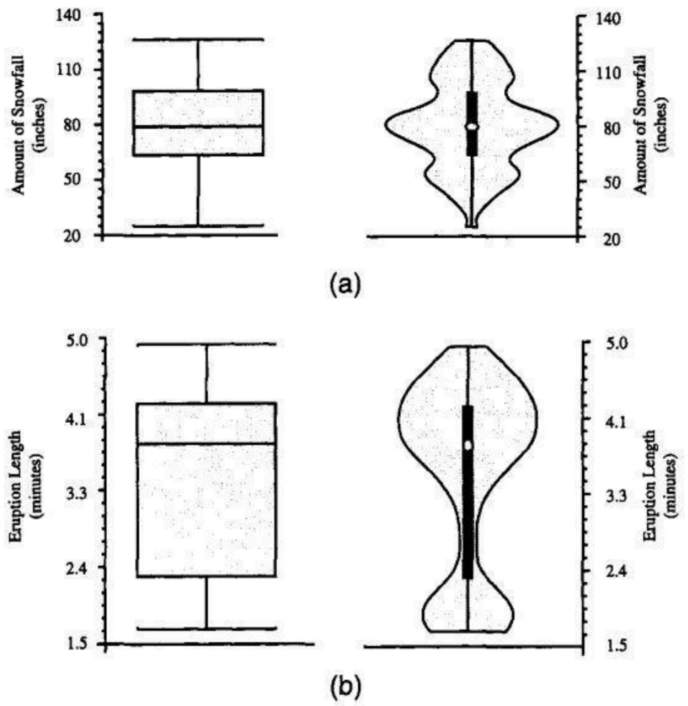


Figure 3. Additional Information in Violin Plots. Two examples from the density estimation literature: (a) annual snowfall for Buffalo, NY, 1910—1972; (b) Old Faithful eruption length.

Source: HINTZE; NELSON, (1998).

Violin Plot graphs show the distribution of a set of samples in relation to their values. The ‘y’ axis shows the values of the variables, while the ‘x’ axis shows the kernel density for the set of samples. Kernel density is a way of distributing data in a continuous space. Each sample is distributed along a vertical axis based on its value. The different samples that have the same value are superimposed and added. The result is a curve that estimates the density of the samples for all values in relation to the sample population. Inside the violin graph is a box plot, displayed in black. The white circle represents the median. The space between the circle and the lower horizontal line represents the lower quartile (25% of the samples). And the space between the circle and the upper horizontal line represents the upper quartile (25% of the samples). The samples below the lower horizontal line and above the upper horizontal line represent 25% of the samples with the lowest and highest values, respectively (TURKEY, 1977; HINTZE; NELSON, 1998).

Violin plots are considered an evolution of the Box Plot, as they allow for a more detailed analysis of sample distribution. Additionally, the Violin Plot technique encompasses all the traditional statistics of the Box Plot, providing a detailed description of the phenomenon under investigation (HINTZE & NELSON, 1998).

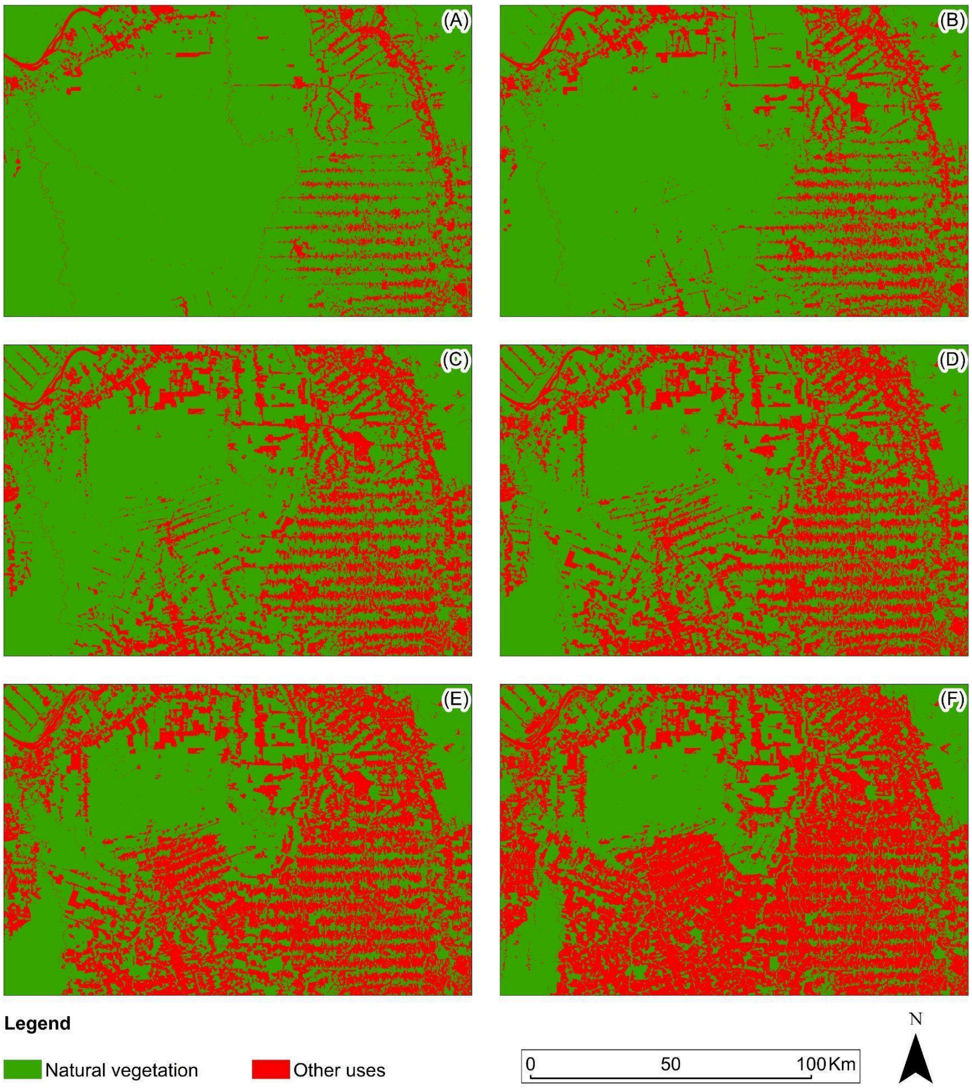
For the application used in this work, analyzing Violin plot graphs makes it possible to identify and differentiate the potential of variables to describe the occurrence and non-occurrence of natural vegetation suppression. It is expected that, for a variable with high descriptive potential, samples of vegetation suppression and persistence will be described differently.

3.3. Results

3.3.1 Amazon biome

Figure 3.4 shows the temporal evolution of land use and land cover for the study area in the Amazon biome based on MapBiomas data. In 1995, the natural vegetation class covered 89.6% of the study area. In 2000, this percentage had decreased to 82.3%. 73.02% in 2005. 66.8% in 2010. 59.7% in 2015. And 49.3% in 2020.

Figure 3.4: Temporal evolution of land use and land cover in the Amazon Biome study area. (A) 1995. (B) 2000. (C) 2005. (D) 2010. (E) 2015. (F) 2020.



Source: Prepared by the authors.

Figure 3.5 shows the results found for the study area in the Amazon biome. For all variables, the ability to differentiate between suppression and persistence of natural vegetation was greater during the training period compared to the extrapolation period. It is also observed that as the extrapolation periods advance, the ability of the variables to discriminate between suppression and persistence of natural vegetation decreases.

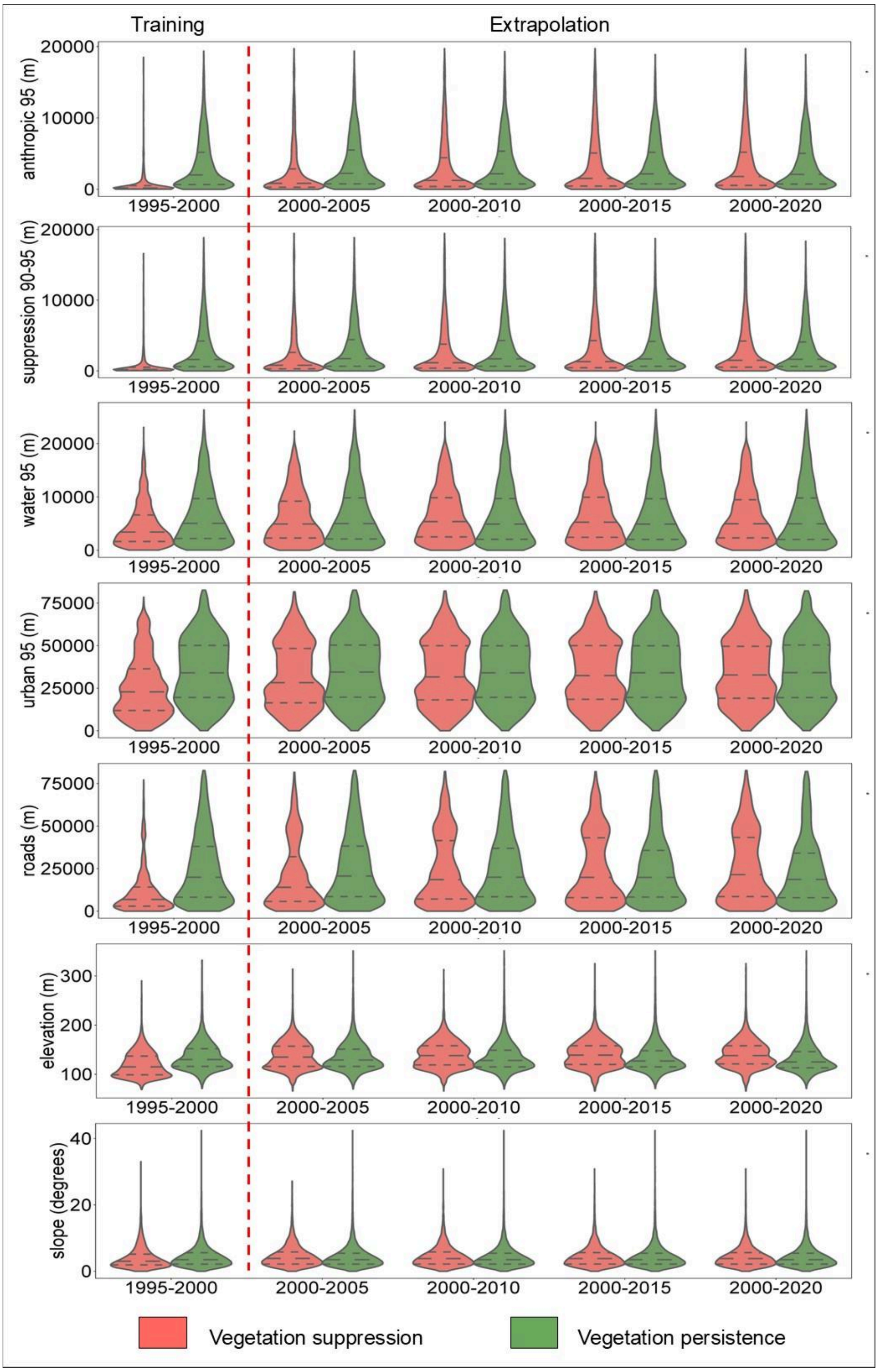
When analyzing the data for the variables of Euclidean distance for anthropic uses in 1995 (anthropic 95) and Euclidean distance for suppressions of natural vegetation between 1990 and 1995 (suppression 90-95), it is evident that vegetation suppression is concentrated in the lowest values across all periods. This indicates that the suppression of natural vegetation predominantly occurred near areas of anthropic use in 1995 and areas of suppression between 1990 and 1995.

For the variables of Euclidean distance to roads (roads), Euclidean distance to urban areas in 1995 (urban 95), and Euclidean distance to water bodies in 1995 (water 95), the suppression of natural vegetation was distributed across a wide range of values in all periods analyzed. This wide distribution makes it difficult to identify a clear pattern that explains the occurrence of vegetation suppressions in a simple and effective way. There is also a similarity between the way vegetation suppression and persistence were described, indicating a low discriminative ability of these variables.

And for the topographical variables of elevation and slope, vegetation suppression during the analyzed periods was primarily concentrated in areas with low altitudes and gentle slopes. However, knowing that the study area in the Amazon biome is predominantly low and flat terrain, the results obtained using these variables do not reveal efficient vegetation suppression patterns.

For this study area, the variables with the greatest ability to differentiate between suppression and persistence of natural vegetation were Euclidean distance for anthropic uses in 1995 (anthropic 95) and Euclidean distance for suppressions of natural vegetation between 1990 and 1995 (suppression 90-95).

Figure 3.5: Violin Plot graphs for the Amazon Biome study area.

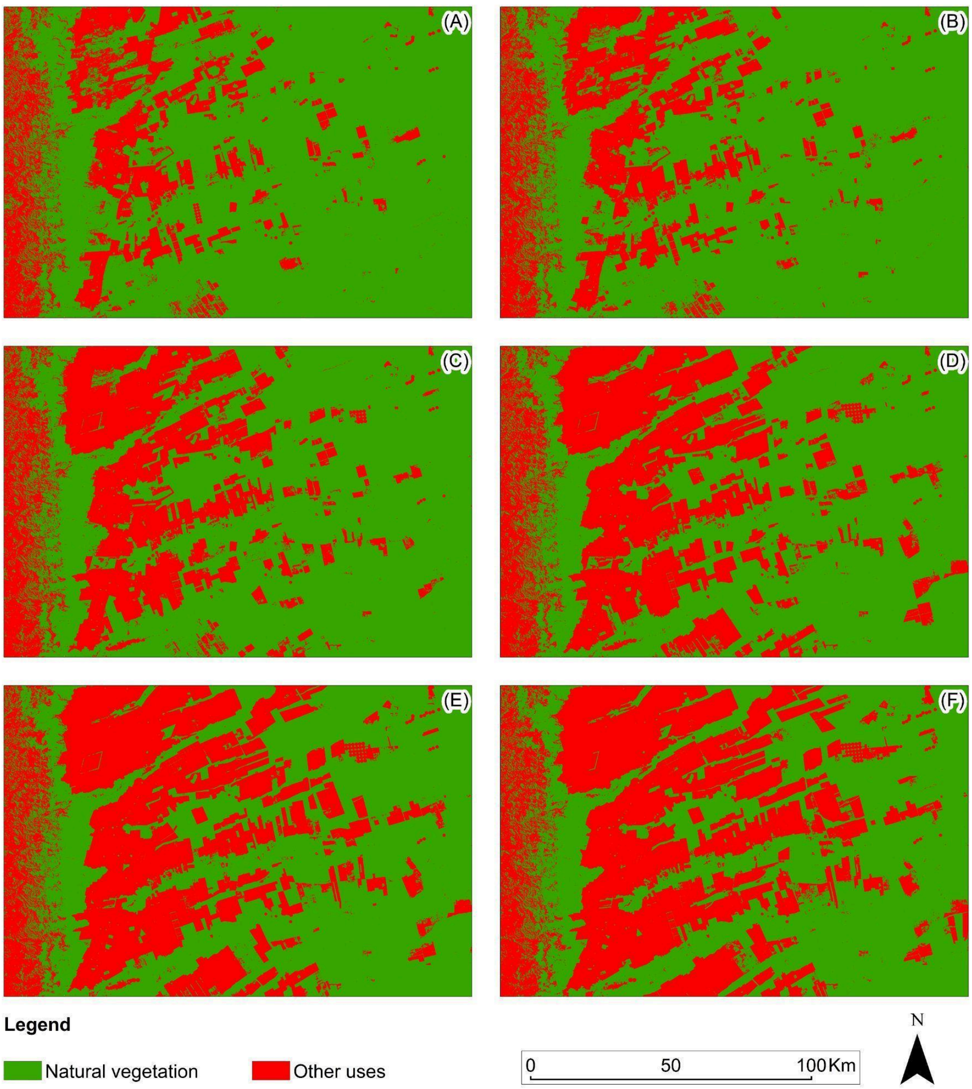


Source: Prepared by the authors.

3.3.2 Cerrado biome

Figure 3.6 shows the temporal evolution of land use and land cover for the study area in the Cerrado biome based on MapBiomas data. In 1995, the natural vegetation class covered 79.1% of the study area. In 2000, this percentage had decreased to 76.2%. 69.2% in 2005. 62.8% in 2010. 58.5% in 2015. And 55.1% in 2020.

Figure 3.6: Temporal evolution of land use and land cover in the Cerrado Biome study area. (A) 1995. (B) 2000. (C) 2005. (D) 2010. (E) 2015. (F) 2020.



Source: Prepared by the authors.

Figure 3.7 shows the results found for the study area in the Cerrado biome. For all variables, the ability to differentiate between suppression and persistence of natural vegetation was greater during the training period compared to the extrapolation period. It is also observed that as the extrapolation periods increase, the ability of the variables to discriminate between suppression and persistence of natural vegetation decreases.

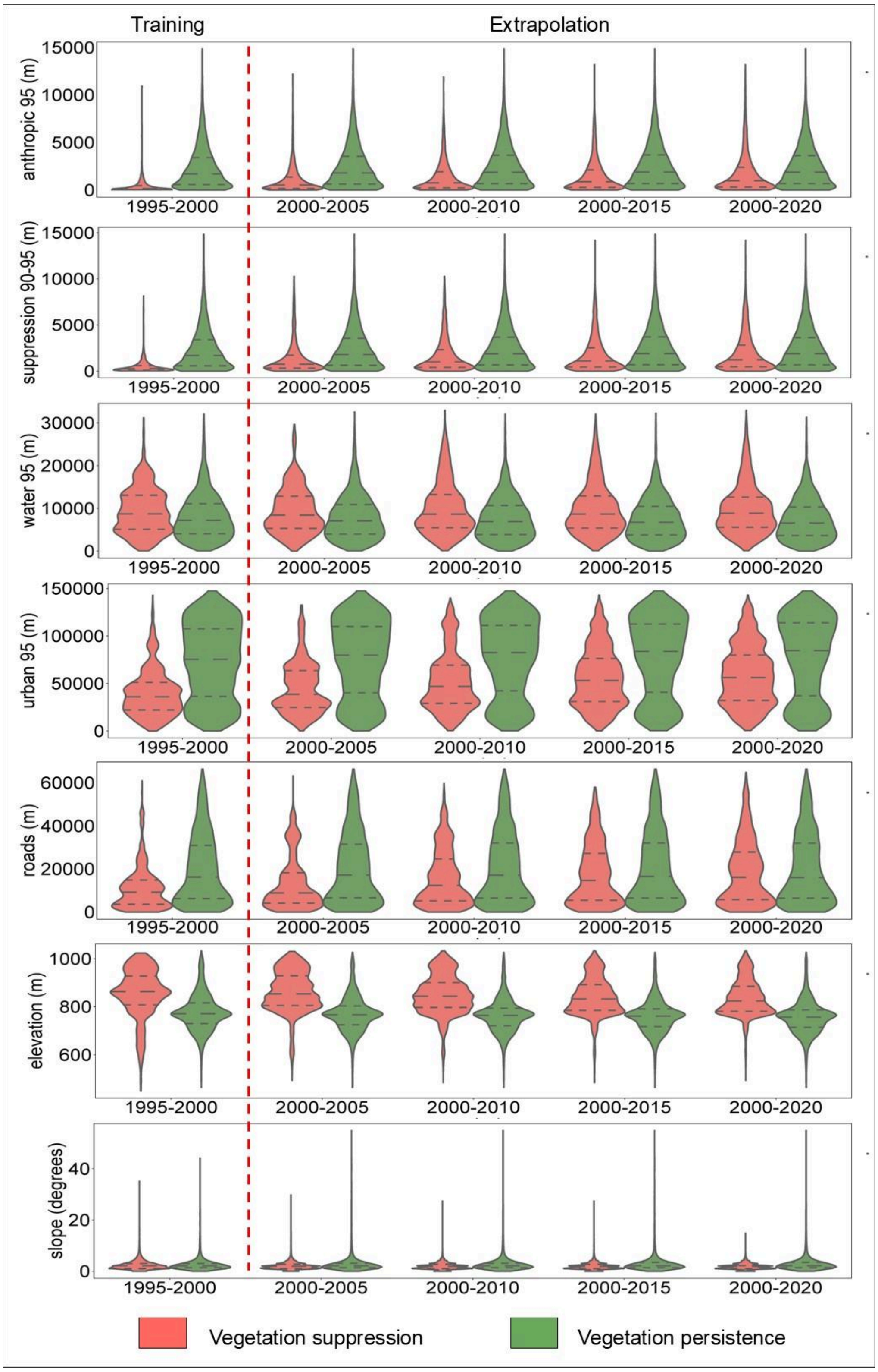
When analyzing the data for the variables of Euclidean distance to anthropic uses in 1995 (anthropic 95) and Euclidean distance to vegetation suppressions between 1990 and 1995 (suppression 90-95), it is evident that vegetation suppression is concentrated in the lowest values across all periods. This indicates that the suppression of natural vegetation primarily occurred near areas of anthropogenic uses in 1995 and areas of vegetation suppression between 1990 and 1995.

For the variables of Euclidean distance to roads (roads), Euclidean distance to urban areas in 1995 (urban 95), and Euclidean distance to water bodies in 1995 (water 95), the suppression of natural vegetation was distributed across a wide range of values in all the periods analyzed. Despite this, it is possible to see significant differences between the patterns of suppression and persistence of natural vegetation. In this way, these variables help to differentiate the two groups.

For topographical variables, in general, vegetation suppression occurred in higher areas than vegetation persistence during all periods. For slope, vegetation suppression and persistence were concentrated in areas with low slopes, not allowing easy differentiation.

For this study area, the variables with the greatest ability to differentiate between suppression and persistence of natural vegetation were Euclidean distance for anthropic uses in 1995 (anthropic 95), Euclidean distance for suppressions of natural vegetation between 1990 and 1995 (suppression 90-95), Euclidean distance for urban in 1995 (urban 95) and elevation.

Figure 3.7: Violin Plot graphs for the Cerrado Biome study area.

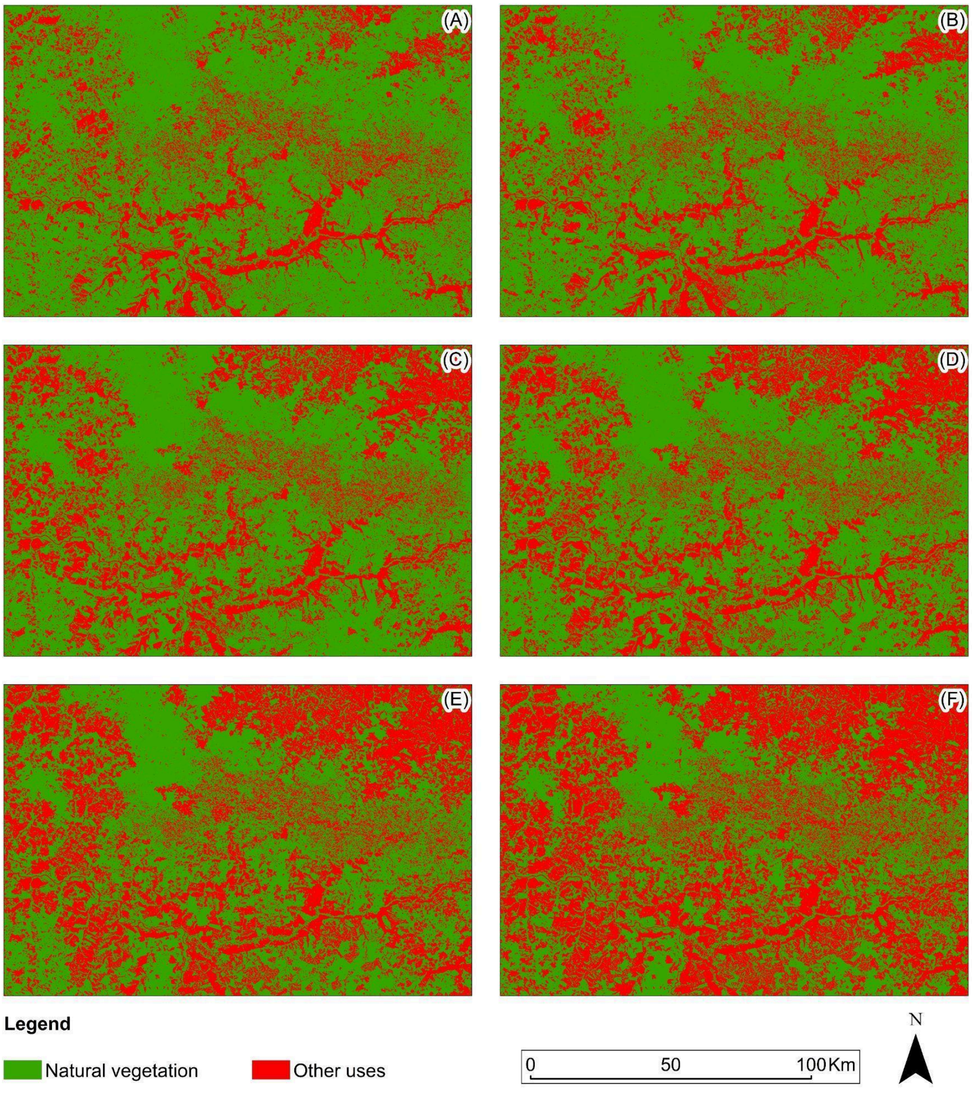


Source: Prepared by the authors.

3.3.3 Pampa biome

Figure 3.8 shows the temporal evolution of land use and land cover for the study area in the Pampa biome based on MapBiomas data. In 1995, the natural vegetation class covered 75.3% of the study area. In 2000, this percentage had decreased to 73.5%. 68.1% in 2005. 65.6% in 2010. 59.2% in 2015. And 52.2% in 2020.

Figure 3.8: Temporal evolution of land use and land cover in the Pampa Biome study area. (A) 1995. (B) 2000. (C) 2005. (D) 2010. (E) 2015. (F) 2020.



Source: Prepared by the authors.

Figure 3.9 shows the results found for the study area in the Pampa biome. For all variables, the ability to differentiate between suppression and persistence of natural vegetation was greater during the training period compared to the extrapolation period. It is also observed that as the extrapolation periods increase, the ability of the variables to discriminate between suppression and persistence of natural vegetation decreases.

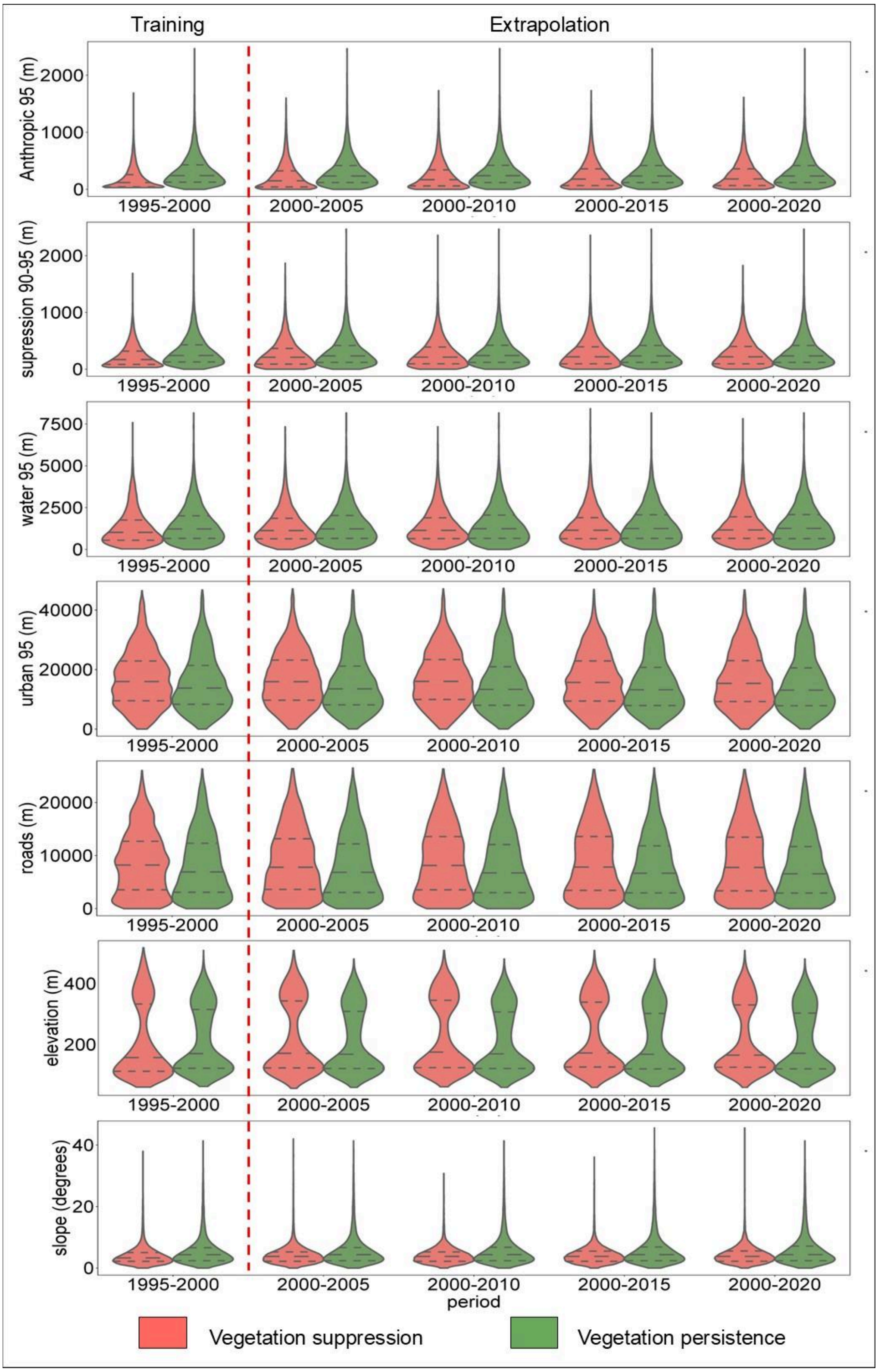
Analyzing the data for the variables of Euclidean distance for anthropic uses in 1995 (anthropic 95), Euclidean distance for vegetation suppressions between 1990 and 1995 (suppression 90-95) and Euclidean distance for water bodies in 1995 (water 95), it is evident that vegetation suppression is concentrated in the lowest values across all periods. Nevertheless, it is noted that, except for the variable anthropic 95, the others did not show significant differences between suppression and persistence of natural vegetation.

For the variables of Euclidean distance to roads (roads) and Euclidean distance to urban areas in 1995 (urban 95), the suppression of natural vegetation was distributed across a wide range of values in all the periods analyzed. In addition, there were no considerable differences between suppression and persistence of natural vegetation. In this way, it is impossible to identify a pattern that explains the occurrence of vegetation suppressions in a simple and effective way, indicating that there is no relationship between these variables and the suppressions.

As for the topographical variables, It can be observed that during the analyzed periods, vegetation suppression was concentrated in two altitude groups, around 130 meters and 350 meters. These results reflect the occurrence of suppression in two distinct relief units: the Paraná basin plateau and the central depression. In this sense, it is not possible to define an effective suppression pattern using these data, but only to demonstrate the spatial variability of suppressions in the study area. Regarding slope, there are no expressive differences between the suppression and persistence of natural vegetation.

For this study area, the variable with the greatest ability to differentiate between suppression and persistence of natural vegetation were Euclidean distance for anthropic uses in 1995 (anthropic 95).

Figure 3.9: Violin Plot graphs for the Pampa Biome study area.



Source: Prepared by the authors.

3.4. Discussions

The results presented in this article make it possible to provide some answers to the difficulties outlined in the introduction:

1) *"The variables used in the training may have different capacities for predicting change".* Analyzing the results, it can be seen that the predictive variables used had different abilities to represent vegetation suppressions. When examining the data obtained for all the study areas (Figures 3.5, 3.7, and 3.9), it is evident that in all situations some variables were superior to the others. These results corroborate with data found in other studies, leading to the conclusion that among various variables used to train the models, some usually have a greater ability to describe the changes (VOIGHT et al., 2019; KUCSICSA et al., 2019).

2) *"The predictive variables used in the training may not be related to land use changes"*. When analyzing the histograms of some variables (Figures 3.5, 3.7, and 3.9), it is remarkable the absence of relation with vegetation suppressions. One example of this issue can be demonstrated by the altitude variable in the Pampa biome (Figure 9). When analyzing the histograms for the different periods, it is observed that the suppression samples are allocated into two well-defined groups: from 80 to 190 m and from 280 to 450 m. This result evidenced the occurrence of vegetation suppressions in different relief compartments, the Sul Rio Grandense plateau and the central depression.

Thus, the data range observed in the histograms is primarily justified by the geomorphological conditions of the study area. This illustrates that the patterns identified from predictive variables may not necessarily express the conditions that determine changes, but simply represent the diversity of characteristics in the study area. In this sense, one possibility to circumvent this problem is to train the models for regions with similar environmental characteristics, facilitating the identification of valid patterns. If there is a desire to model land use changes for large areas, segmentation and training in homogeneous sub-regions may be an alternative (KUCSICSA et al., 2019).

3) "The patterns of change identified in the training may not persist over the prediction period". Observing the graphs and median values in figures 3.5, 3.7 and 3.9, it is noticed that changes are represented in a variable way in different periods, being detachable the differences between training and extrapolation. This variation can be elucidated by factors inherent to the modeling process, three main reasons can be listed: 1) it is known that the mapping and prediction of changes in the training and extrapolation periods are based on land use maps from different years. For training, changes are mapped based on the initial year t_0, and for extrapolation, the base year considered is t_1. In this sense, it is natural to expect that using maps from different time periods will result in the mapping and prediction of different areas of change. This is justified because maps from different times express distinct scenarios and have different classification disagreements (PONTIUS et al., 2017). 2) The predictive variables are developed for a time point immediately before the training period. Thus, the patterns indicated by the variables tend to be more relevant for the training period than for extrapolation. It is expected that over time the chances of changes in patterns that explain land use changes will increase. And 3), as time periods advance, the amount of mapped/projected changes tends to increase (as identified in all study areas). Thus, it is natural to expect that further away from the training period, more changes will be allocated in areas considered less susceptible during the training.

4) *"Different study areas have different patterns of change" and* 5) *"within the same study area, patterns of change may vary".* When comparing the patterns of suppression of natural vegetation identified by the same variable in the different study areas (Figures 3.5, 3.7 and 3.9), different behaviors can be observed. This demonstrates that the predictive characteristics of changes are particular to each space because each study area has specific environmental conditions. This question can be illustrated with the predictive variable of distance to anthropogenic uses in 1995. For the study area in the Amazon biome, the median vegetation suppression for the analyzed periods ranged from 22,861 to 32,791 meters. In the Cerrado biome, it was between 35,933 and 56,177 meters, while for the Pampa biome, it ranged from 15,264 to 16,037 meters.

Considering this data, the patterns of suppression found in each study area differ from the others, being representative of suppressions only in their specific context. In this sense, it is difficult to generate land use change models for large areas. The variability of the characteristics that condition the changes, combined with the environmental diversity of the different locations, restricts the identification of valid patterns for large areas. In a similar assessment, Trigueiro et al. (2019) highlighted the existence of spatial variability in the predictors of land use change in the Brazilian Cerrado biome. For the different sub-regions of the biome, it was demonstrated the relationship of vegetation suppression with different variables. In view of this, it is important to highlight that when developing a model, the statistical method used will adjust the patterns of change considering the values found in all the samples employed (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; SANGERMANO; EASTMAN; ZHU, 2010; LIN et al., 2011). In this sense, the pattern of change for the entire study area will be an adjustment derived from the specific values of each sample. As a result, modeling is unable to represent the spatial variability of all the factors of change.

6) "*Different land use transitions have different patterns of occurrence"*. The type of change to be modeled is crucial for pattern identification. In this study, only a one-way transition, from natural vegetation to other uses, is evaluated. However, it is common to project multiple changes simultaneously (VOIGHT et al, 2019; KUCSICSA et al, 2019). Considering that these models use the same set of variables to identify the patterns of all transitions, modeling multiple changes at the same time increases the chance of errors. This happens because there is no way to control which variables will be used in the training of specific changes, and data that have no relation to the changes may be used. Similarly, the evaluation of predictive factors of specific changes is hindered, requiring additional methodological steps and generating a large amount of data. As a result, the analysis of identified change patterns is commonly ignored.

Finally, the variables used to identify patterns represent only a portion of the factors driving environmental change, failing to capture the full complexity of human and natural systems. Consequently, model results provide a simplified approximation of reality and are subject to inaccuracies. Furthermore, predictive variables used in model training often contain inherent errors, such as disagreements in land use and land cover classifications and errors in the location, delineation, and referencing of geographic features. When generating Euclidean distance surfaces based on such data, the resulting change patterns will inevitably contain imperfections.

3.5. Conclusions

The results obtained in this study highlight important points about the use of predictive variables in the representation of land use changes: (1) In a group of variables used to represent vegetation suppression, each variable showed a different degree of representativeness. (2) The same set of variables used to represent changes in different locations showed specific behavior in each area. (3) When the same set of variables was applied to characterize vegetation suppression over different time intervals, each period displayed distinct behavior. (4) Some predictive variables showed no relationship with vegetation suppression. And (5), assessing the ability of predictive variables to represent land use changes provides valuable insights into how the models work.

Considering the main conclusions listed above, it is recommended that when defining predictive variables for training land use change models, researchers conduct a statistical assessment of the variables' ability to represent the investigated changes. This approach makes it possible to analyze predictive variables from a robust database, reducing the complexity of the models and improving understanding of how they work.

Supplementary materials

The computer code, spatial variables, and the reclassified land use and land cover maps used in this article can be accessed via the following link: https://github.com/macleidivarnier/Assessment-of-the-conditioning-factors-of-natural-vegetation-suppression-in-land-change-models

References

ADHIKARI, Sanchayeeta; SOUTHWORTH, Jane. Simulating Forest Cover Changes of Bannerghatta National Park Based on a CA-Markov Model: a remote sensing approach. **Remote Sensing**, [S.L.], v. 4, n. 10, p. 3215-3243, 22 out. 2012. MDPI AG. http://dx.doi.org/10.3390/rs4103215.

BRITO, Mariana Madruga de; WEBER, Eliseu José; KRIGGER, Vania Silva; LEITZKE, Felipe Padilha. ANÁLISE DOS FATORES CONDICIONANTES DE

MOVIMENTOS DE MASSA NO MUNICÍPIO DE PORTO ALEGRE A PARTIR DE REGISTROS HISTÓRICOS. **Revista Brasileira de Cartografia**, [S.L.], v. 68, n. 9, p. 1853-1872, 16 mar. 2017. EDUFU - Editora da Universidade Federal de Uberlandia. http://dx.doi.org/10.14393/rbcv68n9-44450.

CHAVAN, Sapana B.; REDDY, C. Sudhakar; RAO, S. Srinivasa; RAO, K. Kameswara. Assessing and Predicting Decadal Forest Cover Changes and Forest Fragmentation in Kinnerasani Wildlife Sanctuary, Telangana, India. **Journal Of The Indian Society Of Remote Sensing**, [S.L.], v. 46, n. 5, p. 729-735, 10 jan. 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s12524-017-0739-x.

GORELICK, Noel; HANCHER, Matt; DIXON, Mike; ILYUSHCHENKO, Simon; THAU, David; MOORE, Rebecca. Google Earth Engine:planetary-scale geospatial analysis for everyone. **Remote Sensing Of Environment**, v. 202, p. 18-27, 2017. DOI:http://dx.doi.org/10.1016/j.rse.2017.06.031

HINTZE, Jerry; NELSON, Ray. Violin Plots: a box plot-density trace synergism. **The American Statistician**, [S.L.], v. 52, n. 2, p. 181-184, maio 1998. Informa UK Limited. http://dx.doi.org/10.1080/00031305.1998.10480559.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Bases Cartográficas Contínuas (Versão 2021)**. https://geoftp.ibge.gov.br/cartas_e_mapas/bases_cartograficas_continuas/bc250/ (2024)

KUCSICSA, Gheorghe; POPOVICI, Elena-Ana; BăLTEANU, Dan; DUMITRAşCU, Monica; GRIGORESCU, Ines; MITRICă, Bianca. Assessing the Potential Future Forest-Cover Change in Romania, Predicted Using a Scenario-Based Modelling. **Environmental Modeling & Assessment**, [S.L.], v. 25, n. 4, p. 471-491, 12 nov. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10666-019-09686-6.

LIN, Yu-Pin; CHU, Hone-Jay; WU, Chen-Fa; VERBURG, Peter H.. Predictive ability of logistic regression, auto-logistic regression and neural network models in empirical land-use change modeling – a case study. **International Journal Of Geographical Information Science**, [S.L.], v. 25, n. 1, p. 65-87, 23 fev. 2011. Informa UK Limited. http://dx.doi.org/10.1080/13658811003752332.

MAPBIOMAS. **O Projeto**. 2023. Avaliable in: https://mapbiomas.org/o-projeto. Acess em: 25 jun. 2023.

MAPBIOMAS. **Análise de acurácia**. 2024. Available in: https://brasil.mapbiomas.org/analise-de-acuracia/. Access in: 28 abr. 2024.

MICROSOFT CORPORATION. **Microsoft Excel [Software]**. Versão 16.0. Redmond, WA: Microsoft, 2023.

NASA. NASADEM: **Creating a New NASA Digital Elevation Model and Associated Products**. Available in:<https://www.earthdata.nasa.gov/esds/competitive-programs/measures/nasadem>. Access: 17/09/2023

PONTIUS, Robert Gilmore; HUFFAKER, Diana; DENMAN, Kevin. Useful techniques of validation for spatially explicit land-change models. **Ecological Modelling**, [S.L.], v. 179, n. 4, p. 445-461, dez. 2004. Elsevier BV. http://dx.doi.org/10.1016/j.ecolmodel.2004.05.010.

PONTIUS, Robert Gilmore; KRITHIVASAN, Roopa; SAULS, Laura; YAN, Yan; ZHANG, Yujia. Methods to summarize change among land categories across time intervals. **Journal Of Land Use Science**, [S.L.], v. 12, n. 4, p. 218-230, 30 jun. 2017. Informa UK Limited. http://dx.doi.org/10.1080/1747423x.2017.1338768

SANGERMANO, Florencia; EASTMAN, J Ronald; ZHU, Honglei. Similarity Weighted Instance-based Learning for the Generation of Transition Potentials in Land Use Change Modeling. **Transactions In Gis**, [S.L.], v. 14, n. 5, p. 569-580, out. 2010. Wiley. http://dx.doi.org/10.1111/j.1467-9671.2010.01226.x.

SOARES-FILHO, Britaldo Silveira; CERQUEIRA, Gustavo Coutinho; PENNACHIN, Cássio Lopes. Dinamica—a stochastic cellular automata model designed to simulate the landscape dynamics in an Amazonian colonization frontier. **Ecological Modelling**, [S.L.], v. 154, n. 3, p. 217-235, set. 2002. Elsevier BV. http://dx.doi.org/10.1016/s0304-3800(02)00059-5.

SOUZA, Carlos. et al. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. **Remote Sensing**, [S.L.], v. 12, n. 17, p. 2735, 25 ago. 2020. MDPI AG. http://dx.doi.org/10.3390/rs12172735.

TANG, Doudou; CHEN, Mingjie; HUANG, Xinhua; ZHANG, Guicheng; ZENG, Lin; ZHANG, Guangsen; WU, Shangjie; WANG, Yewei. SRplot: a free online platform for data visualization and graphing. **Plos One**, [S.L.], v. 18, n. 11, p. 1-8, 9 nov. 2023. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0294236.

TRIGUEIRO, Werikson Rodrigues; NABOUT, João Carlos; TESSAROLO, Geiziane. Uncovering the spatial variability of recent deforestation drivers in the Brazilian Cerrado. **Journal Of Environmental Management**, [S.L.], v. 275, p. 111243, dez. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.jenvman.2020.111243.

TURKEY, Jhon. **Exploratory Data Analysis**, Reading, MA: AddisonWesley. 1977.

VAN VLIET, Jasper; BREGT, Arnold K.; BROWN, Daniel G.; VAN DELDEN, Hedwig; HECKBERT, Scott; VERBURG, Peter H.. A review of current calibration and validation practices in land-change modeling. **Environmental Modelling & Software**, [S.L.], v. 82, p. 174-182, ago. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.envsoft.2016.04.017.

VOIGHT, Carly; HERNANDEZ-AGUILAR, Karla; GARCIA, Christina; GUTIERREZ, Said. Predictive Modeling of Future Forest Cover Change Patterns in Southern Belize. **Remote Sensing**, [S.L.], v. 11, n. 7, p. 823, 5 abr. 2019. MDPI AG. http://dx.doi.org/10.3390/rs11070823.

4 ARTIGO 2: How accurate are land use change models? Assessment of natural vegetation suppression predictions in three brazilian biomes

Abstract

Land use change models are used to project future land use scenarios. Various methods for predicting changes can be found, which can be divided into two groups: Baseline models and machine learning-based models. Baseline models use clear change logics, such as proximity or distance to spatial objects. While machine learning-based models use computational methods and spatial variables to identify patterns that explain the occurrence of changes. Considering these two groups of models, machine learning-based models are much more widely used, even though their formulation is considerably more complex. However, the lack of studies comparing the performance of models from these two groups makes it impossible to determine the superiority of one over the other. Therefore, this article aims to evaluate and compare the accuracy of baseline and machine learning-based models for study areas in three Brazilian biomes. Four baseline models and four machine learning-based models were trained considering the environmental context of the period from 1995 to 2000. The objective of these models was to predict natural vegetation suppression for the years 2005, 2010, 2015, and 2020. The results of these predictions were evaluated by comparing them with reference land use maps using rigorous accuracy methods. The results of the accuracy assessment show that regardless of the modeling method used, the performance of the models is similar in all situations. Finally, the results and discussions presented in this work provide a database that allows understanding the performance of different types of models in different environmental contexts.

Keywords: Machine learning models, Baseline models, Accuracy assessment.

4.1. Introduction

Land use change models are employed in various scientific fields to project expected future alterations (VAN VLIET et al., 2016. ABURAS; AHAMAD; OMAR, 2019). The development of such models is based on the identification of patterns and transition rates between time points t_0 and t_1, and then the prediction of changes from t_1 to t_x (PONTIUS et al., 2007). Two groups of modeling methods can be cited: statistical models, commonly based on machine learning methods; and baseline models, generally based on Euclidean distance surfaces (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; SANGERMANO; EASTMAN; ZHU, 2010; PONTIUS; PARMENTIER, 2014; SHAFIZADEH-MOGHADAM et al., 2021).

To formulate machine learning-based models, a training process is required using samples of changes and persistence of land use that occurred between time points t_0 e t_1, and spatial predictive variables with power to explain the phenomenon. For each sample, information contained in the explanatory variables is extracted. Then, a machine learning method is used to identify the characteristics that differentiate samples of occurrence and non-occurrence of changes. As a result, a probabilistic surface is generated indicating the susceptibility to land use change (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002; SANGERMANO; EASTMAN; ZHU, 2010; LIN et al., 2011).

Baseline models do not require a training process. The probability surface in these models is generated by calculating the Euclidean distance to a spatial object that can explain the land use change studied. The proximity or distance to this object determines the probability of change. This approach is frequently used as a baseline model to compare the performance of machine learning-based models (SHAFIZADEH-MOGHADAM et al., 2021; HARATI et al., 2021).

In both methods, after creating the probability surface, it is necessary to apply a change allocation algorithm to predict the land use map at time t_x. Cellular automata algorithms are commonly used in this step. Equation 1 describes the generic operation of this type of algorithm.

$$S^{tx}_{\ ij} = f(S^{t1}_{\ ij}, \Omega^{t1}_{\ ij}, C, N) \qquad (1)$$

where S^{tx}_{ij} represents the state of cell ij at time t_x. f is the transition function; S^{t1}_{ij} refers to the state of cell ij at time t1. $\mathbf{\Omega}^{t1}_{ij}$ is the neighborhood function of cell ij at time t1; C represents the transition constraints; N represents the number of neighboring cells considered (SHAFIZADEH-MOGHADAM et al., 2021).

The use of cellular automata, which integrate neighborhood context, is considered more effective than simply allocating changes in cells classified as most susceptible by the model. This is due to the fact that changes occur in a landscape context, where transformed cells are influenced by their neighborhood during the process. In this sense, some of the land use changes being predicted could be explained by the expansion or reduction of a specific class through contact (SOARES-FILHO, CERQUEIRA, PENNACHIN, 2002).

In addition to model formulation, accuracy assessment is an essential step. Only through this can the predictive capacity to explain the modeled changes be estimated. Many evaluation methods are available. However, while some approaches provide detailed information about the predictions, others offer potentially misleading results (PONTIUS et al., 2007; PONTIUS; MILLONES, 2011; PONTIUS; SI, 2014; PONTIUS; PARMENTIER, 2014; PONTIUS, 2018; FOODY, 2020; HARATI et al., 2021). When investigating recent work, it is observed that some accuracy assessment methods continue to be used despite having been described as potentially misleading many times (VAN VLIET et al., 2016; ABURAS; AHAMAD; OMAR, 2019).

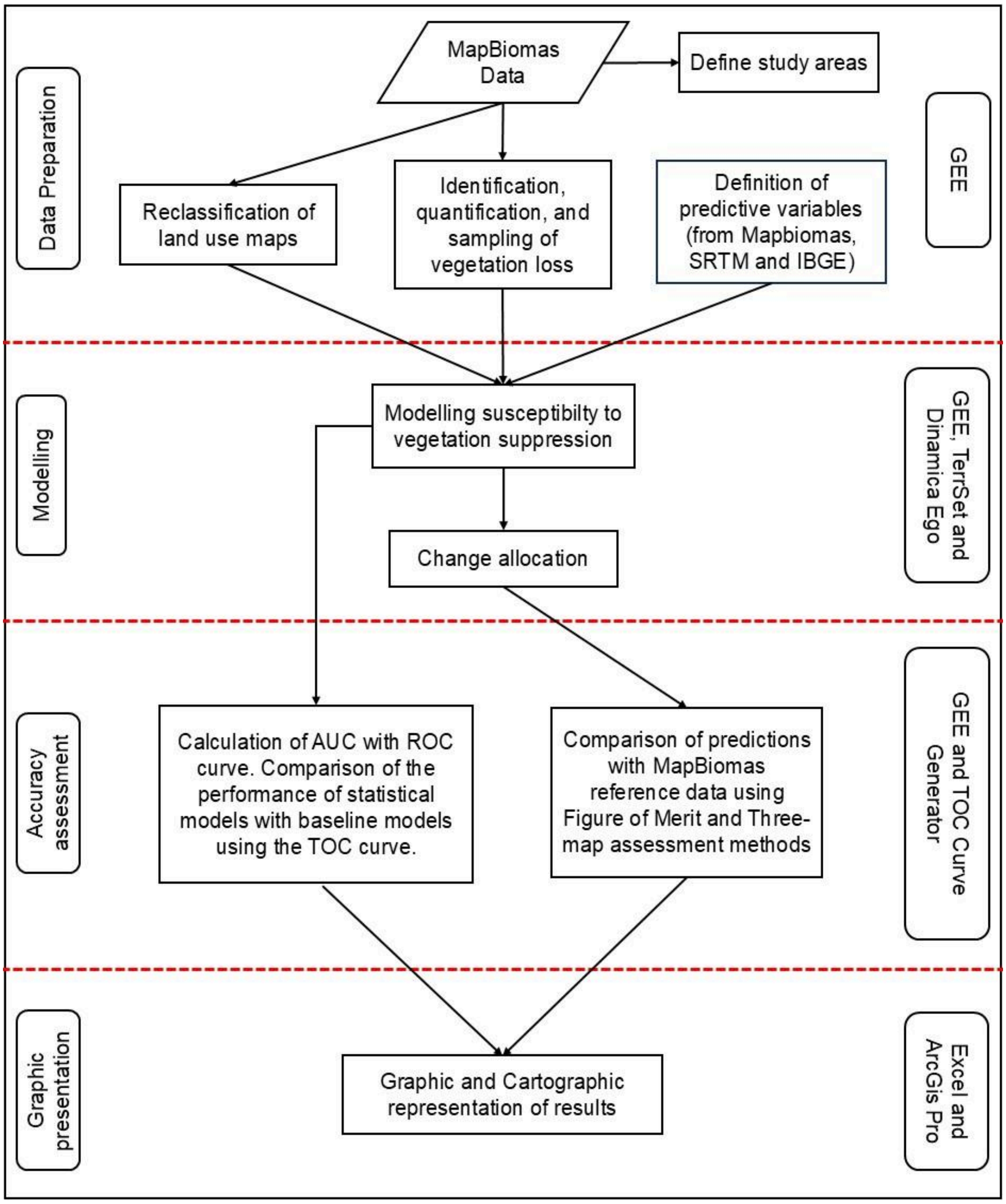
Given the formulation characteristics of the models, machine learning-based models are expected to perform considerably better than baseline models. This expectation is justified by the greater complexity of their formulation and the use of modern training methods. Despite this hypothesis, there are no detailed studies exploring this topic. In this sense, the main objective of this study is to assess whether machine learning-based models are really more accurate than baseline models. Additionally, comparisons are made between the accuracy evaluation methods used in this study and the methods generally used in this field of research.

The results of this article aim to contribute to the scientific community by providing a foundation for comparing the performance of baseline and machine learning-based land use change models.

4.2. Materials and Methods

This study evaluated the performance of eight land use change models (four baseline and four machine learning-based) in selected areas of three Brazilian biomes. The purpose of the models is to predict the occurrence of vegetation suppression using a one-way approach. The training was conducted for the period from 1995 to 2000, while land use change predictions were made for the years 2005, 2010, 2015 and 2020. Figure 4.1 presents the workflow of the methodology. In the central part of the figure are the processes carried out, on the left are the stages of the work, and on the right are the software used. In the central part of the figure are the processes carried out, on the left are the stages of the work, and on the right are the softwares used. All the computer codes used are presented in the supplementary materials.

Figure 4.1: Methodology flowchart.



Source: Prepared by the authors.

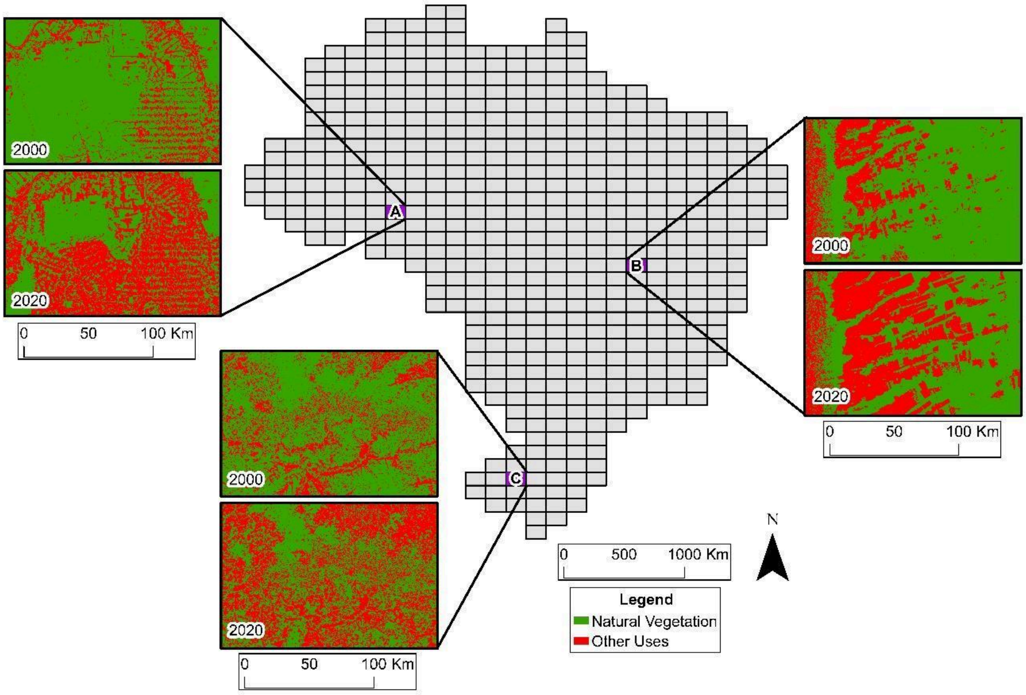
4.2.1 Study Area

The study areas were selected to represent different environmental contexts that experienced land use changes and suppression of natural vegetation between 2000 and 2020. In three Brazilian biomes, areas were selected that exhibited major land use changes during this period. The areas were chosen based on land use maps from collection 8 of the MAPBIOMAS project (SOUZA et al., 2020) and the 1:250,000 grid used to generate this data.

To select the study areas, we first calculated the area of the natural vegetation classes for the years 2000 and 2020 in each grid cell of the map. Subsequently, we calculated the difference in the area of these vegetation formations between 2000 and 2020. Finally, we chose the grid cell with the largest absolute reduction in natural vegetation area during the analyzed period in the Amazon, Cerrado, and Pampa biomes. These three biomes were selected because they have distinct environmental characteristics and have undergone major land use changes during the period analyzed in this study.

The selected study areas cover a wide variety of spaces and dynamics of land use change. In this way, it is possible to examine the factors that condition the process of vegetation suppression in areas with diverse characteristics. Figure 4.2 shows the location of the three study areas.

Figure 4.2: Study areas and land use in 2000 and 2020. (A) Amazon Biome, (B) Cerrado Biome, and (C) Pampa Biome.



Source: Prepared by the authors.

The Amazon biome study area is located in the state of Rondônia, in northern Brazil, and is characterized by the predominance of dense, humid tropical forest, contrasting with extensive areas of pasture derived from recent deforestation. In the period from 2000 to 2020, there was a reduction of approximately 40 per cent in the area of natural vegetation, with deforestation advancing from south to north. The suppression of the Amazon rainforest is primarily driven by timber harvesting and the expansion of livestock and agricultural production. Today, this process is regarded as a global environmental issue, attracting attention from various sectors of society (SOUZA et al., 2020).

The Cerrado biome study area is located on the border of the states of Bahia and Goiás, in the central portion of Brazil, and is characterized by mosaics of savannah, grassland, agriculture and pasture. In the period from 2000 to 2020, there was a reduction of approximately 28 per cent in the area of natural vegetation, with suppression progressing from west to east. The reduction of natural vegetation in this part of the Cerrado is motivated by the implementation of large-scale monocultures, a process that has received attention from the scientific community in order to understand its environmental impacts (SOUZA et al., 2020; PONTIUS et al., 2023).

The Pampa biome study area is located in the state of Rio Grande do Sul, in southern Brazil, and is characterized by grassland vegetation, pastures, agricultural uses, and forested areas. Between 2000 and 2020, natural vegetation decreased by approximately 29%, with suppression occurring without a defined spatial pattern. The conversion of natural vegetation in this part of the Pampa is primarily driven by the establishment of agricultural crops, particularly the shift from grassland formations to soybean plantations. Between 2000 and 2020, soybean production expanded extensively in the biome, considerably altering the landscape. Another important process is the periodic alternation of land use classes, with transitions from grassland vegetation to pasture and rice production, and vice versa (SOUZA et al., 2020).

4.2.2 Data Used

All the geospatial data used in this study comes from open sources and is accessible for consultation. The land use maps were derived from the MAPBIOMAS project, a collaborative effort involving researchers from various institutions in Brazil, including universities, NGOs, research institutes, and technology start-ups. One of its main products is the annual land use and land cover maps, available since 1985 for the entire Brazilian territory. These maps are generated using data from the Landsat program, offering a spatial resolution of 30 meters. A key strength of this mapping is the robust accuracy assessment, allowing users to understand the disagreements in the classification for each land use class mapped (SOUZA et al., 2020. MAPBIOMAS, 2023).

The altitude and slope data were derived from the NASADEM digital elevation model. NASADEM is a product resulting from the reprocessing of the Radar Shuttle Topographic Mission (SRTM) with a spatial resolution of 30 meters. Compared to SRTM, NASADEM provides greater accuracy due to processing improvements and the incorporation of auxiliary data from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) and the Ice, Cloud, and Land Elevation Satellite (ICESat)/Geoscience Laser Altimeter System (GLAS) (NASA, 2023).

The road network was derived from the Brazilian Institute of Geography and Statistics (IBGE), representing federal and state roads. The scale of this data is 1:250,000 and the features were delimited using images from the Sentinel-2, Planet and Maxar sensors. During the validation of the product, an average positional error of 125 meters was considered acceptable (IBGE, 2024).

4.2.3 Data preparation

The data preparation process was divided into three main stages: (1) reclassification of land use maps from the MapBiomas project; (2) identification, quantification, and sampling of vegetation loss; and (3) definition of predictive variables for modeling.

4.2.3.1 Reclassification of land use maps from the MapBiomas project

The reclassification of land use maps from the MapBiomas project was performed with the aim of grouping the original classes into only two: natural vegetation and other uses. For the reclassification, land use maps from the years 1995, 2000, 2005, 2010, 2015, and 2020 were selected. These years and time intervals were defined with the objective of the training period having the same duration as each extrapolation period, allowing the model's results to be evaluated for extrapolation phases with equivalent time periods.

4.2.3.2 Identification of the amount of vegetation suppression projected for the future

The amount of change expected in the future was defined using the two land use maps from the training period. From these maps, it is possible to identify and quantify the areas of vegetation suppression during the training, as well as allocate samples of occurrence and non-occurrence of the phenomenon. Equation 2 exemplifies this process:

$$VS = VCt_1 - VCt_0 \tag{2}$$

where VS represents the vegetation suppression during the training period, VCt_0 represents the vegetation area in 1995, and VCt_1 represents the vegetation area in 2000. Knowing the size of the suppressed area in the training period and assuming that in the future the rate of suppression will remain the same, it is possible to predict the expected area of suppression for any period of time. Equation 3 presents this possibility.

$$VS_x = \left((VCt_1 - VCt_0\right)/\left(Vt_0\right)) * \left(100/ Vt_x\right) \tag{3}$$

Where future vegetation suppression (VS_x) is calculated based on the total suppressed area during the training period (VCt_1 - VCt_0), divided by the total vegetation area in t_0 (1995), and multiplied by 1% of the vegetation area from the period preceding the extrapolation. For example, if the model's training occurred between 1995 and 2000 and it was identified that vegetation suppression during this period corresponded to 1% of the existing vegetation in 1995, to estimate the changes for the year 2005, 1% of the vegetation area in 2000 was used. For 2010, 1% of the estimated vegetation area for 2005 would be applied, and so on.

This method was used to estimate the expected changes in the baseline models and in the model using Random Forest. In the other models used, this step occurs automatically.

4.2.3.3 Predictive variables of land use changes

To prepare the predictive variables for natural vegetation suppression, data from the MapBiomas, NASADEM and IBGE were used. Using data from the MapBiomas project, Euclidean distance surfaces were created for the water bodies in 1995, urban space in 1995, anthropogenic uses in 1995 and vegetation suppression between 1990 and 1995. NASADEM provided altitude and slope data. And from the IBGE, a Euclidean distance surface was calculated for the roads existing in 1995.

Due to the model training being carried out with samples of vegetation suppression that occurred between 1995 and 2000, the predictive variables cannot be correlated with the data from this period. Therefore, the training input variables derived from MapBiomas were created based on the 1995 land use map or the vegetation suppressions that occurred in the preceding period (1990-1995). After the creation of the predictive variables, they were grouped into a data cube and, together with the land use maps, served as input data for the modeling stage.

*4.*2.4 Modeling

In the model development stage, four different statistical methods were used: from the TerrSet software, the Artificial Neural Network and SimWeight methods were used; from the Dinamica EGO software, the weights of evidence method; and from the Google Earth Engine platform (GORELICK et al., 2017), the Random Forest method.

For the TerrSet and Dinamica EGO models, the required inputs were land use maps from 1995 and 2000, and predictive variables. In these software, the creation of occurrence and nonoccurrence samples; the calculation of transition rates during the training and extrapolation periods; and the allocation of changes by cellular automata are performed automatically. In the case of the model using Google Earth Engine (GEE) and Random Forest, these steps were carried out through computer programming. Each of the four models is described in detail below.

4.2.4.1 TerrSet - Artificial Neural Network (ANN)

The ANN method used in the TerrSet software is inspired by the functioning of the human brain, with artificial neurons organized into layers (input, hidden, and output) for data processing. The architecture used is Multi-Layer Perceptron (MLP), adjusting the training data in relation to a validation set. Neurons are connected, and synaptic weights are adjusted by backpropagation to maximize convergence (LIN et al., 2011). In this method, the user can define the number of samples used in each class and the number of iterations. In this study, 10,000 samples per class and 10,000 iterations were used.

4.2.4.2 TerrSet – SimWeight

The SimWeight method of the TerrSet software is an algorithm based on similarity-weighted instances. It simplifies land use change modeling by requiring only two parameters: the number of training samples and the number of neighboring samples considered in the training process (parameter K). Using the K-Nearest Neighbor logic, for each cell to be evaluated, the k nearest samples (change or persistence) are identified. Then, the distance in the variable space of each unknown location is calculated in relation to the change samples that are within the k interval. An exponential weighting function is then used to associate labels with the cells, indicating the potential for transition based on environmental similarity with altered locations. SimWeight generates a continuous probability surface using only change samples. This is based on the assumption that persistence samples are not effectively examples of locations that have not changed, only that they are known to be locations that have not yet changed (SANGERMANO; EASTMAN; ZHU, 2010). In the modeling, the parameter K was adjusted to 100 neighboring samples, and 10,000 samples were used in training.

4.2.4.3 Dinamica Ego – Weights of evidence

The evidence weights method used in the Dinamica EGO software is a Bayesian approach where the probability of an event occurring is defined given a spatial pattern. The weights of evidence for land use changes are determined by the ability of ranges of values of the predictive variables to distinguish between areas where changes occurred and where they did not. The land use change susceptibility surface is generated by summing the evidence weights that each cell received from all the spatial variables used (SOARES-FILHO; CERQUEIRA; PENNACHIN, 2002). In the Dinamica Ego software, the adjustment of evidence weights occurs automatically, not allowing any user definition.

4.2.4.4 Google Earth Engine – Random Forest

The Random Forest classifier employed in the GEE platform is a non-parametric algorithm capable of estimating values or assigning objects to classes without making assumptions about the spatial distribution or structure of the data. Using training samples and regression analysis, the classifier divides the samples into increasingly homogeneous subsets based on values extracted from explanatory variables. The parameter for subdivision is the selection of predictive attributes that minimize uncertainty. In a decision tree, a sample is classified by the mean of the subset where it is allocated in the regression analysis. Its final label in a Random Forest classification is the result of the classification in the various regression trees, being the value that has been repeated the most times (BREIMAN, 2001; GOOGLE, 2024; GOOGLE EARTH ENGINE, 2024).

In this method, 10,000 samples were used for each class (occurrence of natural vegetation suppression and non-occurrence) and 28 decision trees. The number of decision trees was determined following the principle of parsimony, aiming for the best possible performance at the lowest computational cost.

After generating the susceptibility surface, a cellular automata algorithm was adapted for allocating the predicted changes. A simple multiplication was performed between the probability surface for vegetation suppression, a 5x5 neighborhood filter counting the number of pixels classified by MapBiomas as anthropogenic uses in 2000, and a surface with random values ranging from 0 to 1. Using the result of this multiplication, the pixels with the highest susceptibilities were selected based on the rates of change derived from Equation 2.

4.2.5 Baseline models

The Baseline models used were: Euclidean distance for anthropogenic uses in 2000, Euclidean distance for vegetation suppressions between 1995 and 2000, surface with random values from 0 to 1, and the land use map reclassified for the year 2000. As there is no training process in these models, there is no correlation problem when using land use data from this period. .

For the Euclidean distance to anthropogenic uses in 2000 and Euclidean distance to vegetation suppression occurring between 1995 and 2000, the lowest values were considered the most susceptible to vegetation suppression. For the surface with random values from 0 to 1, the highest values were considered the most susceptible. The 2000 land use map was used as a null change model.

4.2.6 Accuracy assessment

The accuracy assessment was divided into two parts: the accuracy assessment of probability surfaces and the accuracy assessment of the land use change prediction maps. The accuracy assessment of probability surfaces serves as a way to assess the accuracy of the probabilistic surface in differentiating areas where suppression did and did not occur during the prediction periods. While the accuracy assessment of the land use change prediction maps determines their ability to spatialize the alterations that actually occurred.

4.2.6.1 Accuracy assessment of probability surfaces

4.2.6.1.1 Area Under the Curve - ROC Curve

The ROC curve was used to evaluate the rate of vegetation suppression samples correctly classified (true positives) and incorrectly classified (false positives) by the models for each probability level. The true positive rate can be calculated using equation 4, while the false positive rate is measured by equation 5.

$$TVP = VP/(VP + FN) \quad (4)$$

$$TFP = FP/(FP + VP) \quad (5)$$

where the true positive rate for each probability level (TPR) is obtained by dividing the quantity of true positives (TP) by the sum of true positives and false negatives (TP + FN). While the false positive rate for each probability level (FPR) is obtained by dividing false positives (FP) by the sum of false positives (FP) and true positives (TP) (PONTIUS; PARMENTIER, 2014).

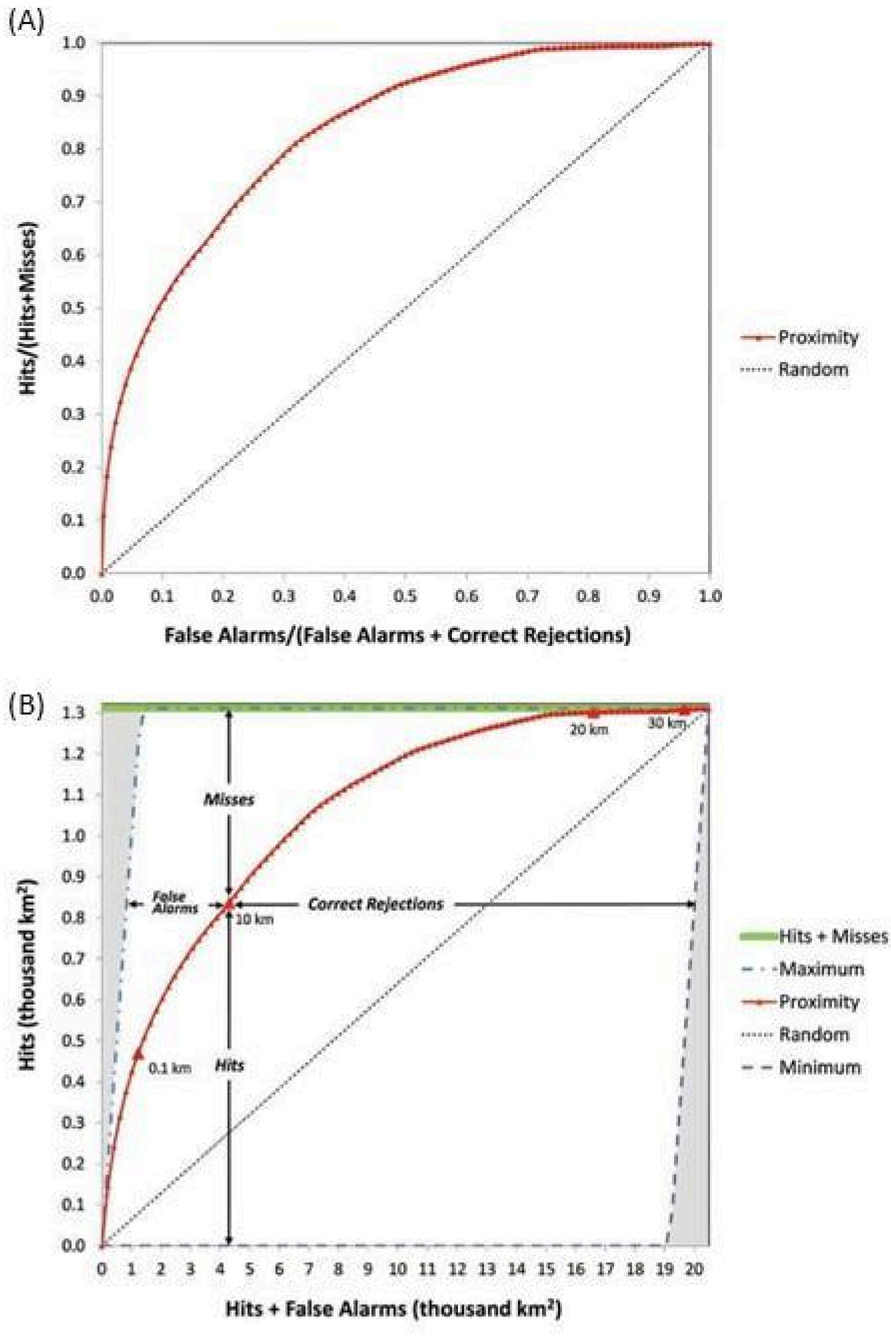
The Area Under the Curve (AUC) is a way to simplify the ROC curve analysis by aggregating the value of true and false positive rates across all thresholds. In a model capable of differentiating all occurrence and nonoccurrence samples of the phenomenon, the AUC value will be 1. In contrast, for a model with performance equal to randomness, the AUC value will be 0.5.

The AUC calculation was performed on the GEE platform. For each analyzed period, 10,000 samples of vegetation suppression and 10,000 samples of non-suppression were used. The probability surfaces were divided into 100 equal parts based on their histograms. For each of these parts, the true and false positive rates were calculated. The AUC value was defined by the average performance of the models at each of the 100 thresholds.

4.2.6.1.2 TOC Curve

The TOC curve method was proposed as an improvement over the ROC curve. For each probability threshold, this approach allows the quantification of hits, misses, false alarms and rejections. In this sense, the TOC curve enables the extraction of more information than the ROC curve. Figure 4.3 presents the ROC and TOC curves (PONTIUS; SI, 2014).

Figure 4.3: (A) ROC curve e (B) TOC curve.



Source: PONTIUS; SI, (2014).

To generate the curves, we used the TOC Curve Generator v1.2.7 software (LIU; PONTIUS, 2021). In this software, it is necessary to provide the probability surface to be evaluated, a mask of the study area's boundary, and the reference map of the suppressions that occurred during the predicted periods.

As the limit mask of the study areas, the natural vegetation areas of the year 2000 were defined. This is because the pixels inside this delimitation are the only ones that can be considered by the models as future suppressions. Furthermore, the probability surfaces were divided into 100 equal parts for the calculation of the true positive and false positive rates.

4.2.6.2 Accuracy Assessment of the Land Use Change Prediction Maps

4.2.6.2.1 Evaluation by the three possible comparisons of two maps

The three possible comparisons between two maps were used to quantify the accuracy of the models in representing land use for the years 2005, 2010, 2015, and 2020. In this approach, using two reference maps at time states t_1 and t_x, plus the land use change prediction map for state t_x, it is possible to quantify the components of agreement and disagreement of the prediction.

By comparing the reference maps at t_1 and t_x, the observed changes and persistences are indicated. When comparing the reference map at t_1 with the prediction result at t_x, the areas predicted as changes by the model are obtained. In addition, by comparing the reference map at t_x with the prediction result at t_x, the overall accuracy of the model is indicated. By analyzing these three comparisons together, it is possible to quantify the persistences correctly predicted by the model (areas that did not change class in the reference maps and were indicated by the model as persistence), false alarms (areas where the model predicted changes, but in the reference there was class persistence), misses (areas where the model predicted class persistence, but the reference showed a change), and hits (areas where the model predicted changes, and the reference showed a change).

This approach was implemented on the GEE platform using the reclassified land use and land cover maps from the MAPBIOMAS project as reference data. In addition to the components of agreement and disagreement, areas classified in the reference maps as other uses at t_1 and as natural vegetation at t_x were computed as regeneration; areas of anthropogenic vegetation at t_1 that remained in the same class at t_x were classified as excluded.

After calculating the area of the components derived from the three maps approach, this data was grouped in Excel software and presented as graphs.

4.2.6.2.2 Figure of Merit

Based on the area of the components mapped by the three-map comparison, the Figure of Merit was calculated. Equation 6 presents the formulation of this evaluation.

$$Figure\ of\ Merit\ =\ Hits\ /\ (Misses\ +\ Hits\ +\ False\ alarms) \qquad (6)$$

where the Figure of Merit represents the proportion of hits relative to the sum of misses, hits, and false alarms (PONTIUS, 2018). This approach simplifies the comparison of model performance and focuses on areas of change, distinguishing itself from overall accuracy.

In addition, for each study area and extrapolation period, the model with the highest figure of merit was selected to be presented as a map. Due to the large number of results (96 maps resulting from the evaluation by three maps), it was not possible to present all the maps in the article.

4.3. Results

4.3.1 Accuracy assessment of the vegetation suppression probability surfaces

Table 4.1 presents the AUC results found for the susceptibility surfaces of the study area in the Amazon biome. It is observed that the value obtained for the baseline models is higher than that obtained for the machine learning-based models in all extrapolation periods. Among the machine learning-based models, the Random Forest (RF) method achieved the best performance, while the Sim Weight - Terrset obtained the lowest AUC value.

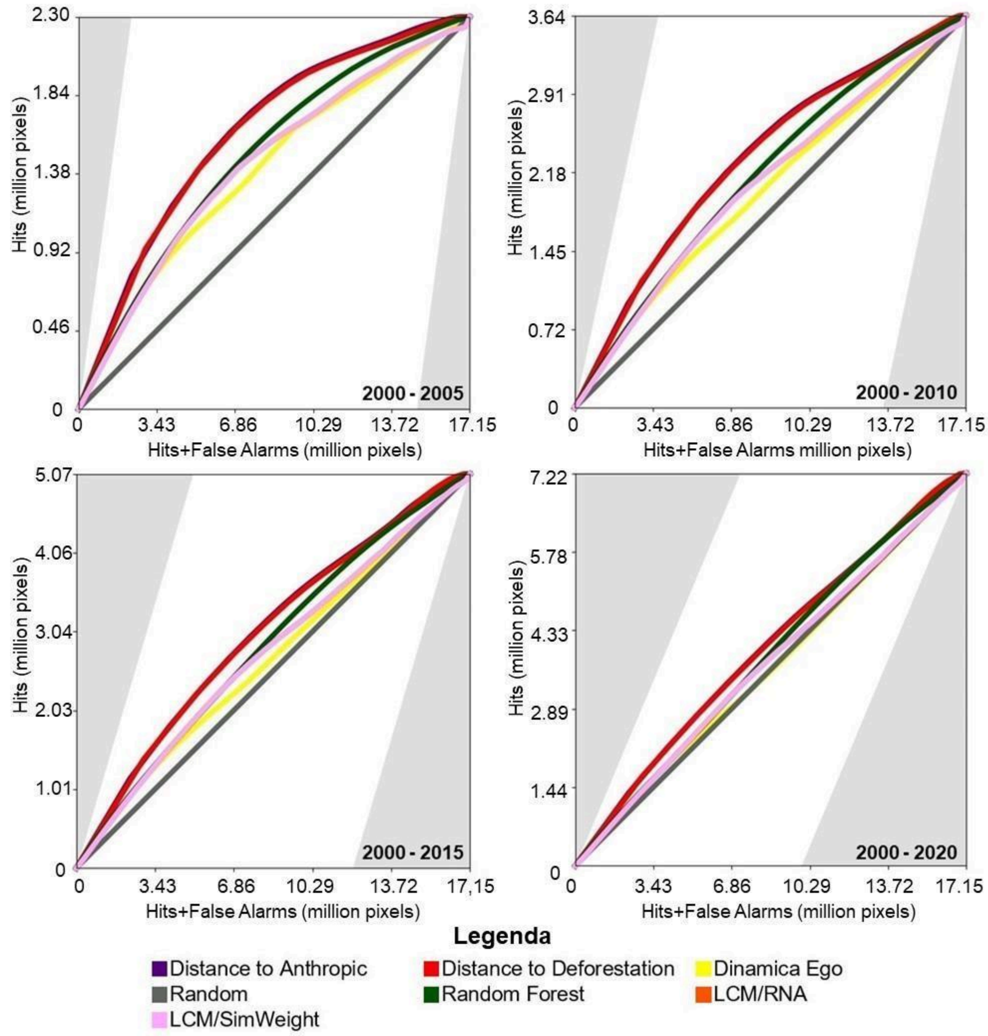
Table 4.1: AUC values for the models of the study area in the Amazon biome.

	Baseline models			**Machine Learning-based models**			
	Anthropic	**Suppression**	**Random**	**ANN**	**Sim Weight**	**Random Forest**	**Dinamica Ego**
2000 - 2005	0.737	0.732	0.5	0.638	0.575	0.715	0.636
2000 - 2010	0.673	0.669	0.5	0.585	0.535	0.650	0.570
2000 - 2015	0.640	0.638	0.5	0.561	0.529	0.615	0.548
2000 - 2020	0.595	0.595	0.5	0.532	0.494	0.567	0.515

Source: Prepared by the authors.

Figure 4.4 shows the results of the TOC curve evaluation. It is observed that in all extrapolation periods, the performance of the baseline models was slightly superior to the machine learning-based models.

Figure 4.4: TOC curve of the models for the study area in the Amazon biome.



Source: Prepared by the authors.

For the study area in the Cerrado biome, Table 4.2 presents the AUC results found for the susceptibility surfaces. The two models with the best performance were Random Forest and ANN - TerrSet. The baseline models and the method using Dinamica Ego achieved slightly lower performance. While the model using Sim Weight - TerrSet obtained the worst performance.

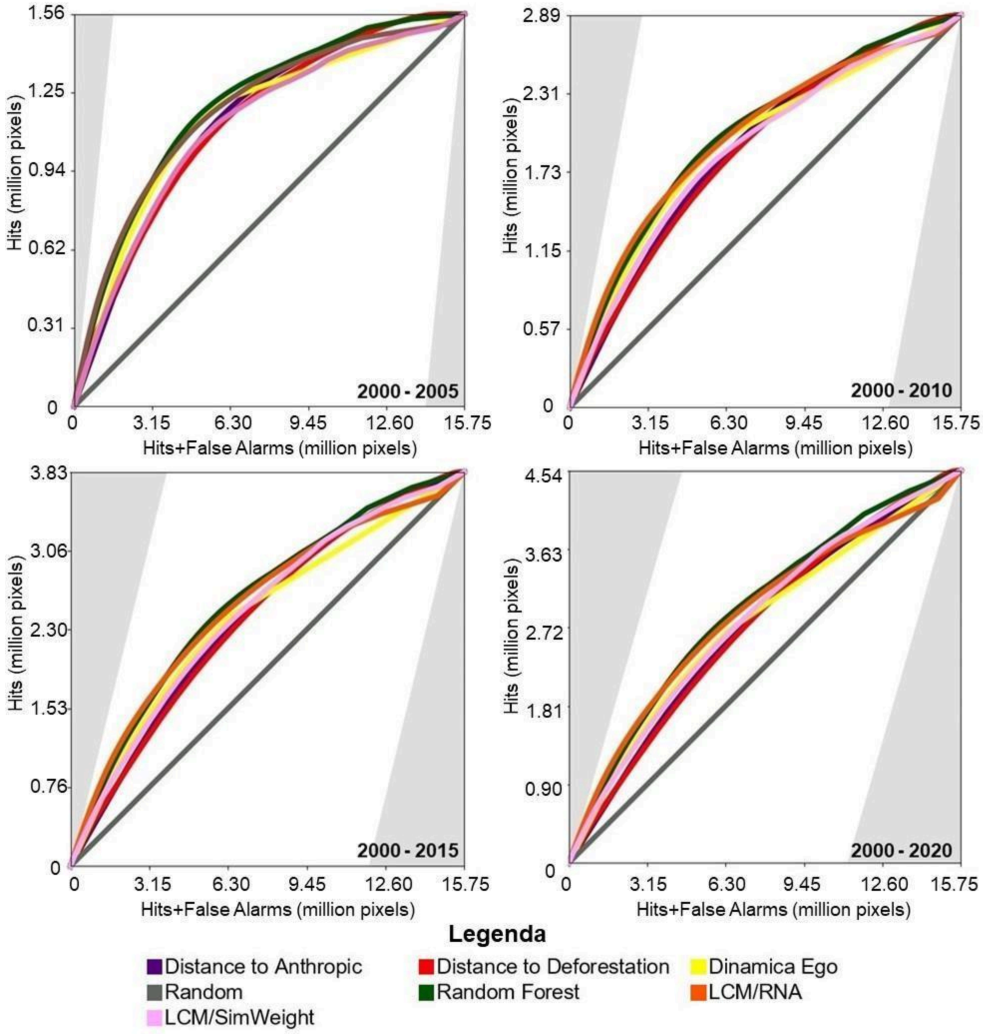
Table 4.2: AUC values for the models of the study area in the Cerrado biome.

	Baseline models			Machine Learning-based models			
	Anthropic	**Suppression**	**Random**	**ANN**	**Sim Weight**	**Random Forest**	**Dinamica Ego**
2000 - 2005	0.757	0.752	0.5	0.786	0.669	0.826	0.741
2000 - 2010	0.698	0.692	0.5	0.736	0.627	0.761	0.689
2000 - 2015	0.679	0.675	0.5	0.719	0.618	0.747	0.655
2000 - 2020	0.644	0.650	0.5	0.693	0.595	0.723	0.631

Source: Prepared by the authors.

Figure 4.5 presents the results of the TOC curve evaluation. For this study area, the baseline models achieved intermediate performance, outperforming two of the four machine learning-based models.

Figure 4.5: TOC curve of the models for the study area in the Cerrado biome.



Source: Prepared by the authors.

For the study area in the Pampa biome, the AUC results for the susceptibility surfaces are presented in Table 3. The best performances were from the Random Forest, Weights of Evidence - Dinamica Ego, and ANN - TerrSet methods. The baseline model of distance to anthropogenic uses in 2000 showed a slightly lower performance. And the two models with the lowest values were the distance to vegetation suppression between 1995 and 2000 and Sim Weight - TerrSet.

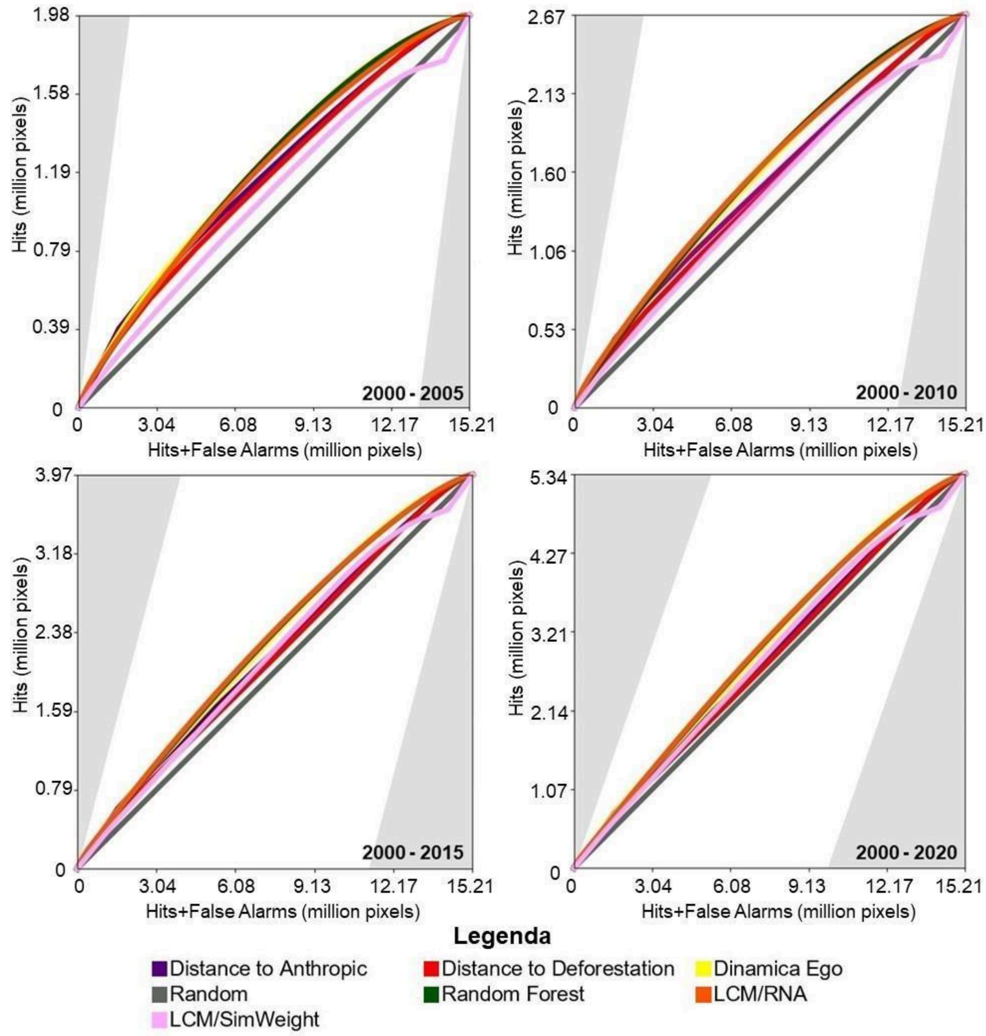
Table 4.3: AUC values for the models of the study area in the Pampa biome.

	Baseline models			Machine Learning-based models			
	Anthropic	**Suppression**	**Random**	**ANN**	**Sim Weight**	**Random Forest**	**Dinamica Ego**
2000 - 2005	0.606	0.483	0.5	0.624	0.485	0.733	0.633
2000 - 2010	0.578	0.475	0.5	0.618	0.486	0.709	0.612
2000 - 2015	0.556	0.470	0.5	0.597	0.493	0.658	0.589
2000 - 2020	0.548	0.474	0.5	0.590	0.496	0.627	0.586

Source: Prepared by the authors.

Figure 4.6 presents the results of the TOC curve evaluation. For this study area, with the exception of the Sim Weight - TerrSet model, machine learning models demonstrated superior performance compared to baseline models.

Figure 4.6: TOC curve of the models for the study area in the Pampa biome.

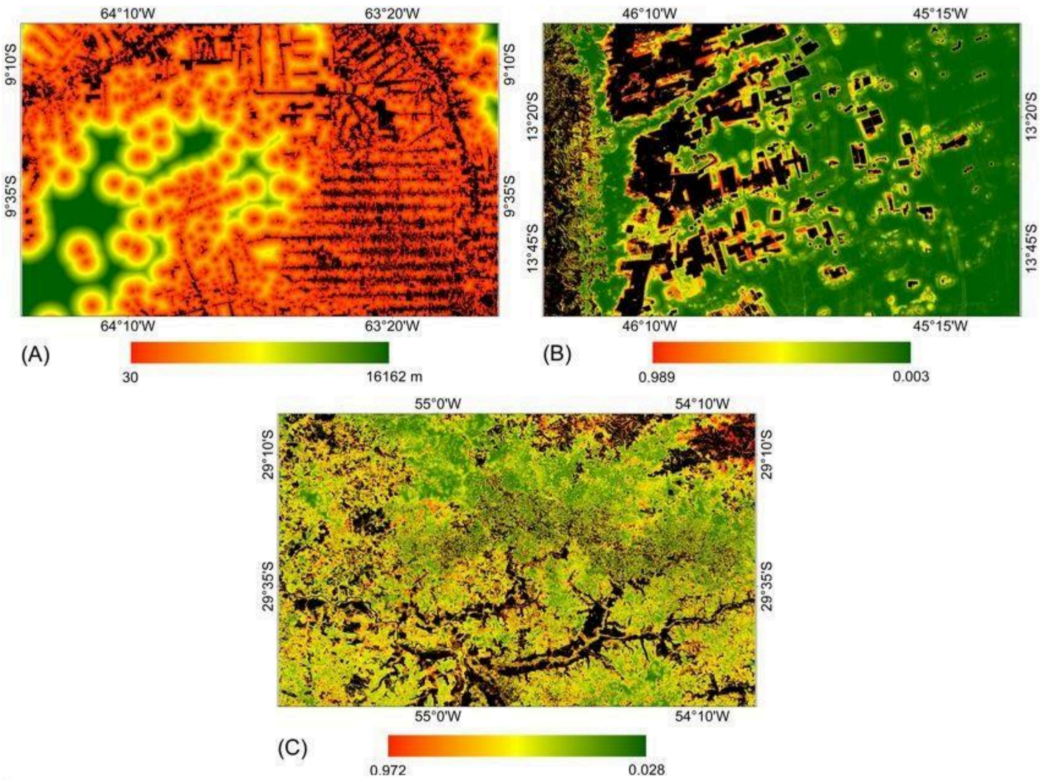


Source: Prepared by the authors.

In general, it is observed that for all study areas, regardless of the modeling method, the performance of the models was similar across all tested extrapolation periods. Additionally, it can be seen that the further away from the training period, the worse the model's performance tends to be. This is observed in all tested conditions, regardless of the method or study area, demonstrating a general behavioral pattern.

Figure 4.7 presents the probability surfaces with the highest average AUC values across the four extrapolation periods for each study area. For each evaluated method, the same susceptibility model was used to allocate changes in the four prediction periods. It is noted that for the study area in the Amazon biome, the areas considered most susceptible to vegetation suppression are located near the areas defined as anthropic uses in the year 2000. In contrast, for the study areas in the Pampa and Cerrado biomes, the areas considered most susceptible to vegetation suppression were defined by training using machine learning methods.

Figure 4.7: Probability surfaces with the best AUC for each study area. (A) Amazon: distance to anthropic uses from 2000. (B) Cerrado: Random Forest. (C) Pampa: Random Forest.

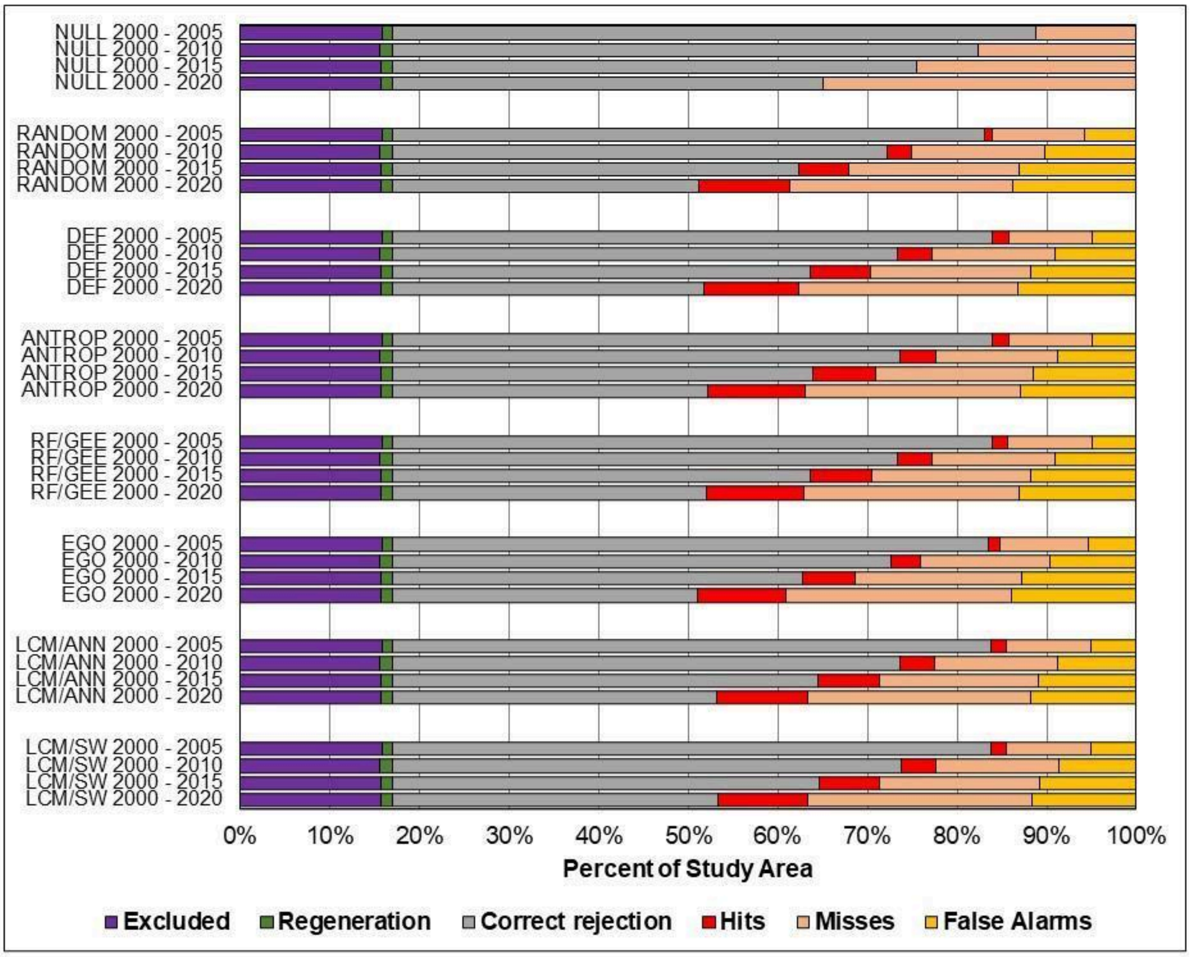


Source: Prepared by the authors.

4.3.2 Accuracy Assessment of the Land Change Prediction Maps

Figures 4.8, 4.9, and 4.10 present, respectively, for the study areas in the Amazon, Cerrado, and Pampa biomes, the results of the evaluation of land use prediction maps using the three-maps technique. The components of excluded areas, shown in purple, and regeneration, shown in green, have the same size for each study area and prediction period, regardless of the modeling method used. For null change models, misses, false alarms, and hits are not presented because these models do not project changes.

Figure 4.8: Three-map evaluation of the models for the study area in the Amazon biome.

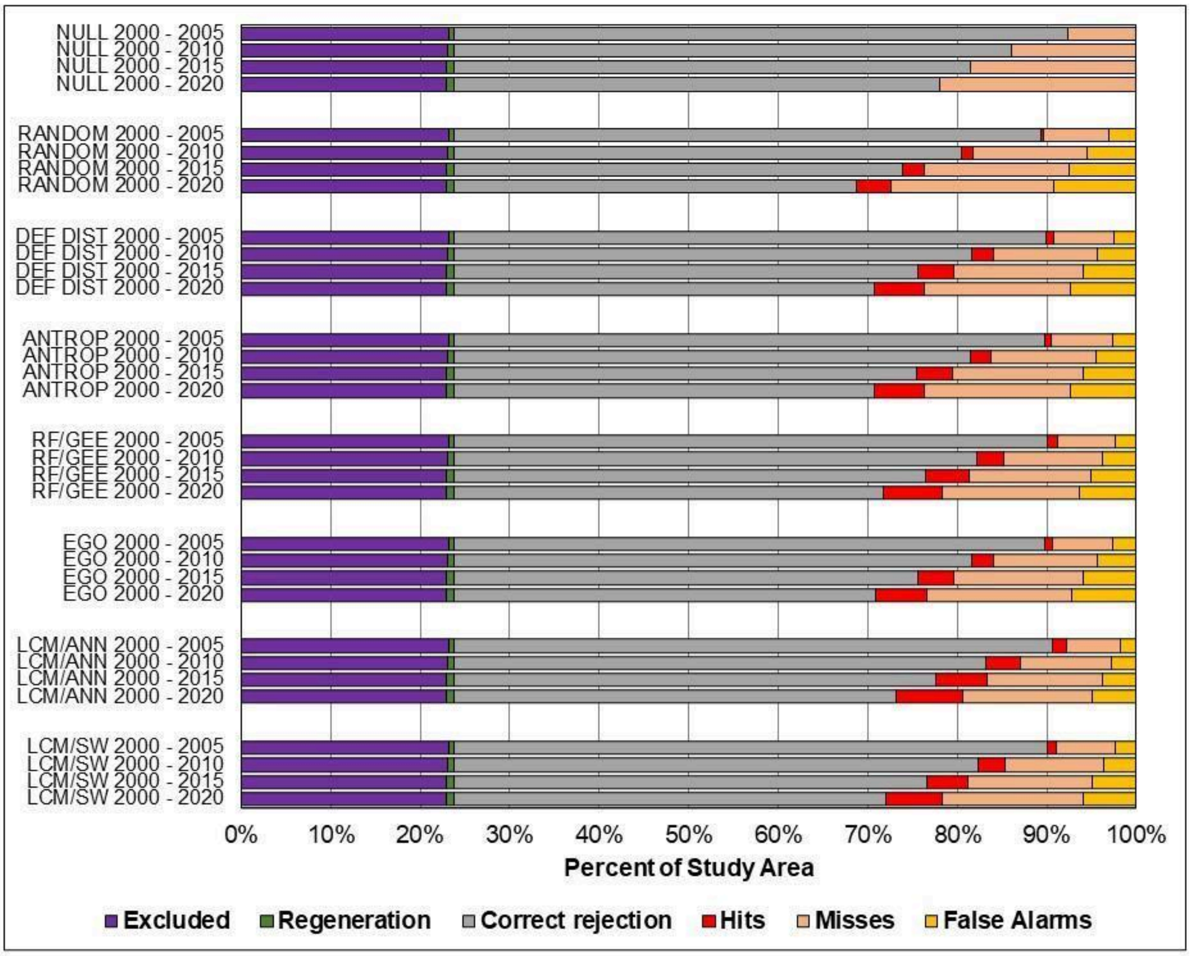


Source: Prepared by the authors.

In general, it is observed across all the study areas that regardless of the method used, the models behave very similarly in all prediction periods. Also, as noted in the TOC curve evaluation, it is observed that the farther away from the training period, the worse the performance of the models.

It can be seen that the null change models outperform the others in most cases. The exceptions occurred in the study area of the Cerrado biome for the predictions in the years 2010, 2015, and 2020. For 2010 and 2015, the ANN - TerrSet model outperformed the null change model. For 2020, the ANN - TerrSet, Sim Weight - TerrSet and Random Forest models outperformed the null change model.

Figure 4.9: Three-map evaluation of the models for the study area in the Cerrado biome.



Source: Prepared by the authors.

For the comparison between baseline and machine learning-based models, it is observed that the performance of both was similar in all situations, with one or the other performing slightly better or worse.

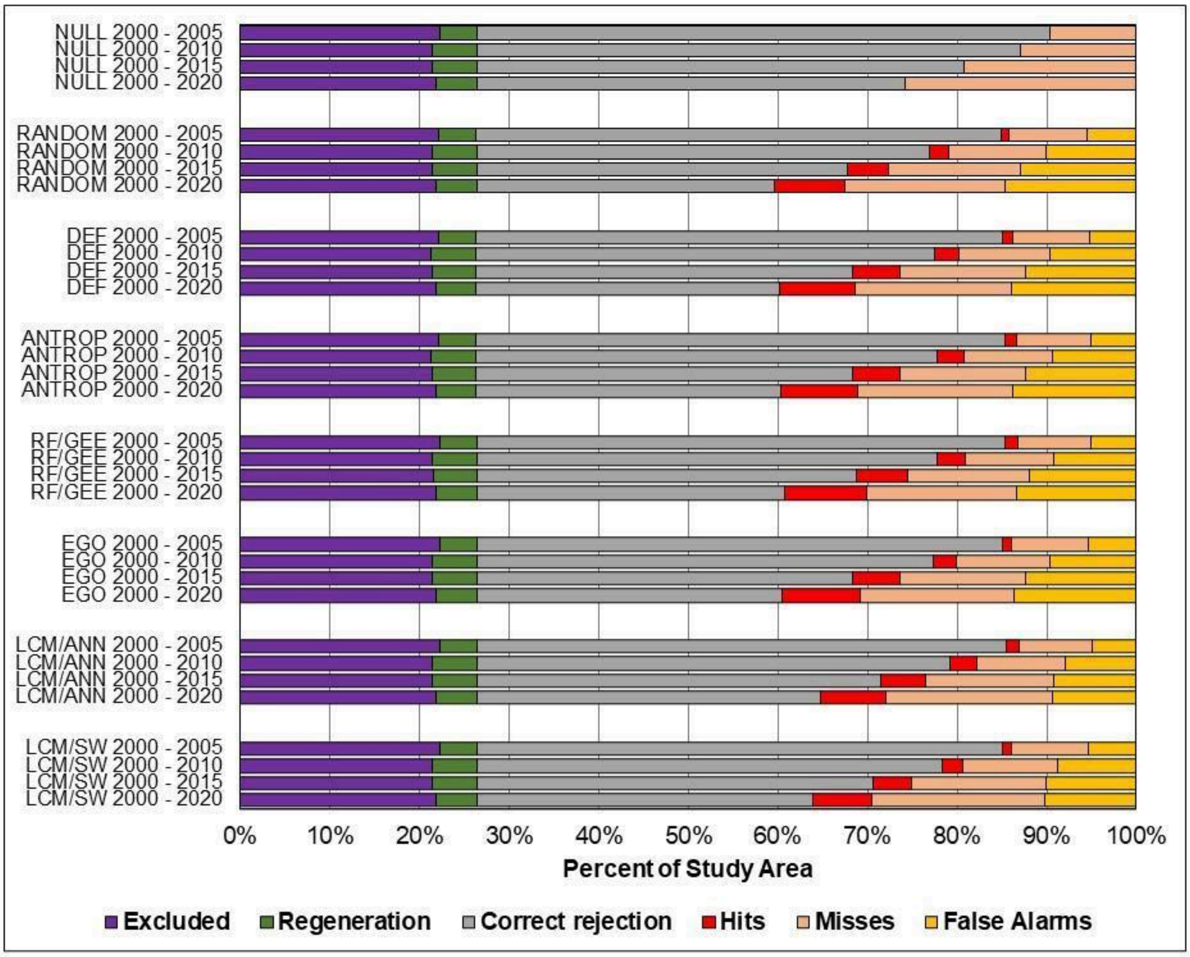
It can be observed that the quantity of Misses is greater than False alarms for all study areas, methods, and prediction periods. This indicates that disagreements due to the omission of changes contained in the reference are greater than disagreements due to the allocation of incorrectly predicted changes.

It is noted that the component of correct rejection is the most abundant in all scenarios, indicating that the main source of agreement is the prediction of persistence.

It is verified that the performance of the random model follows the same pattern as the other methods across all study areas and prediction periods.

Furthermore, when comparing the results of the Sim Weight and ANN methods with the other models, a larger area of correct rejections and fewer hits, omissions, and false alarms is observed. This occurs mainly because these models project a smaller amount of changes when compared to the others.

Figure 4.10: Three-map evaluation of the models for the study area in the Pampa biome.



Source: Prepared by the authors.

Table 4.4 presents the figure of merit values for all models and prediction periods in the study area of the Amazon biome. The models that achieved the highest figure of merit values for the prediction of 2005 were distance to anthropogenic uses and distance to vegetation suppression, while the models with the lowest values were random and Dinamica EGO. For the 2010 and 2015 predictions, the highest values were found for the distance to anthropogenic uses and ANN-TerrSet models, while the lowest values were obtained for the Random and Dinamica EGO models. And for the 2020 prediction, the highest values were found for the distance to anthropogenic uses and Random Forest models, while the lowest values were found for the Random and Dinamica EGO models.

Table 4.4: Figure of Merit for the models of the study area in the Amazon biome.

	Baseline models			**Machine Learning-based models**			
	Anthropic	**Suppression**	**Random**	**ANN**	**Sim Weight**	**Random Forest**	**Dinamica Ego**
2000 - 2005	0.114	0.114	0.053	0.104	0.103	0.107	0.079
2000 - 2010	0.155	0.143	0.098	0.148	0.147	0.146	0.117
2000 - 2015	0.195	0.184	0.146	0.191	0.189	0.187	0.157
2000 - 2020	0.228	0.219	0.205	0.216	0.214	0.225	0.201

Source: Prepared by the authors.

Table 4.5 presents the figure of merit values for all models and prediction periods in the study area of the Cerrado biome. The models that obtained the highest figure of merit values for the 2005 prediction were ANN-TerrSet and distance to vegetation suppression, while the models with the lowest values were random and Dinamica EGO. For the 2010 prediction, the highest values were found for the ANN-TerrSet and Sim Weight-TerrSet models, while the lowest values were for the Random and Dinamica EGO models. And for the 2015 and 2020 predictions, the highest values were found for the ANN-TerrSet and Random Forest models, while the lowest values were for the Random and Dinamica EGO models.

Table 4.5: Figure of Merit for the models of the study area in the Cerrado biome.

	Baseline models			Machine Learning-based models			
	Anthropic	Suppression	Random	ANN	Sim Weight	Random Forest	Dinamica Ego
2000 - 2005	0.074	0.084	0.032	0.172	0.108	0.111	0.080
2000 - 2010	0.124	0.129	0.063	0.231	0.168	0.167	0.129
2000 - 2015	0.161	0.165	0.092	0.258	0.200	0.207	0.165
2000 - 2020	0.193	0.191	0.119	0.276	0.225	0.235	0.196

Source: Prepared by the authors.

Table 4.6 presents the figure of merit values for all models and prediction periods in the study area of the Pampa biome. The models that achieved the highest figure of merit values for the predictions of 2005, 2010, and 2015 were ANN - TerrSet and Random Forest, while the lowest values were the random and Sim Weight - TerrSet models. And for the 2020 prediction, the highest values were found for the Random Forest and Dinamica EGO models, while the lowest values were for the Sim Weight-TerrSet and Random models.

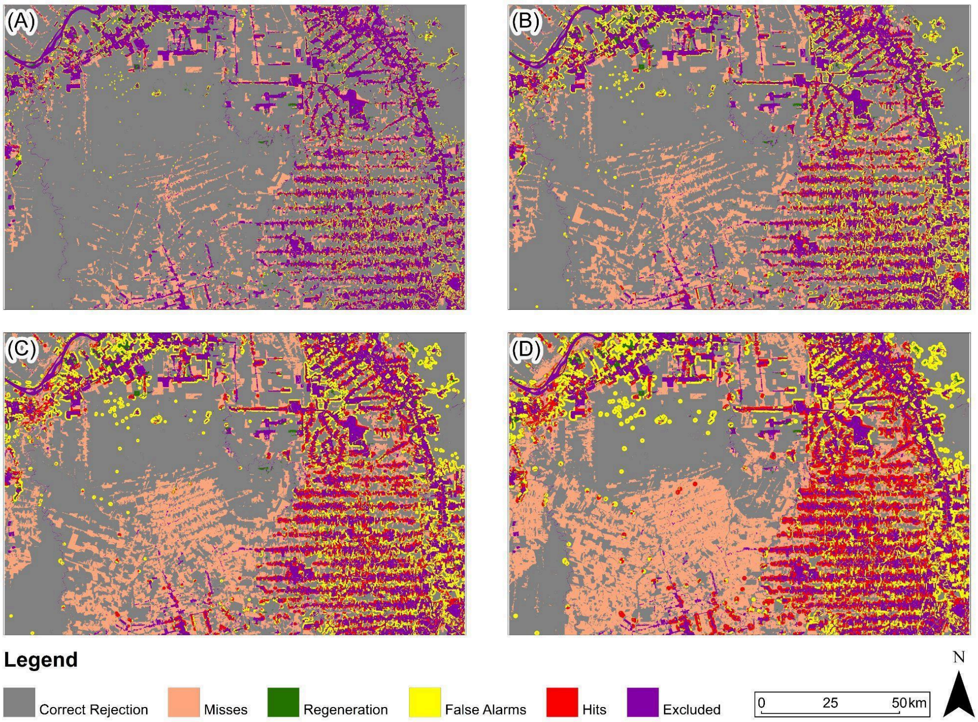
Table 4.6: Figure of Merit for the models of the study area in the Pampa biome.

	Baseline models			Machine Learning-based models			
	Anthropic	Suppression	Random	ANN	Sim Weight	Random Forest	Dinamica Ego
2000 - 2005	0.091	0.076	0.055	0.102	0.069	0.093	0.071
2000 - 2010	0.135	0.119	0.093	0.146	0.105	0.136	0.113
2000 - 2015	0.166	0.164	0.142	0.176	0.144	0.181	0.165
2000 - 2020	0.217	0.214	0.195	0.205	0.179	0.233	0.221

Source: Prepared by the authors.

Figures 4.11, 4.12, and 4.13 present, for the three study areas, the predictive map that obtained the highest Figure of Merit value for each analyzed period. For the study area in the Amazon biome, the land use change prediction map with the highest Figure of Merit value for all evaluated periods was generated based on the Euclidean distance for anthropogenic uses in 2000.

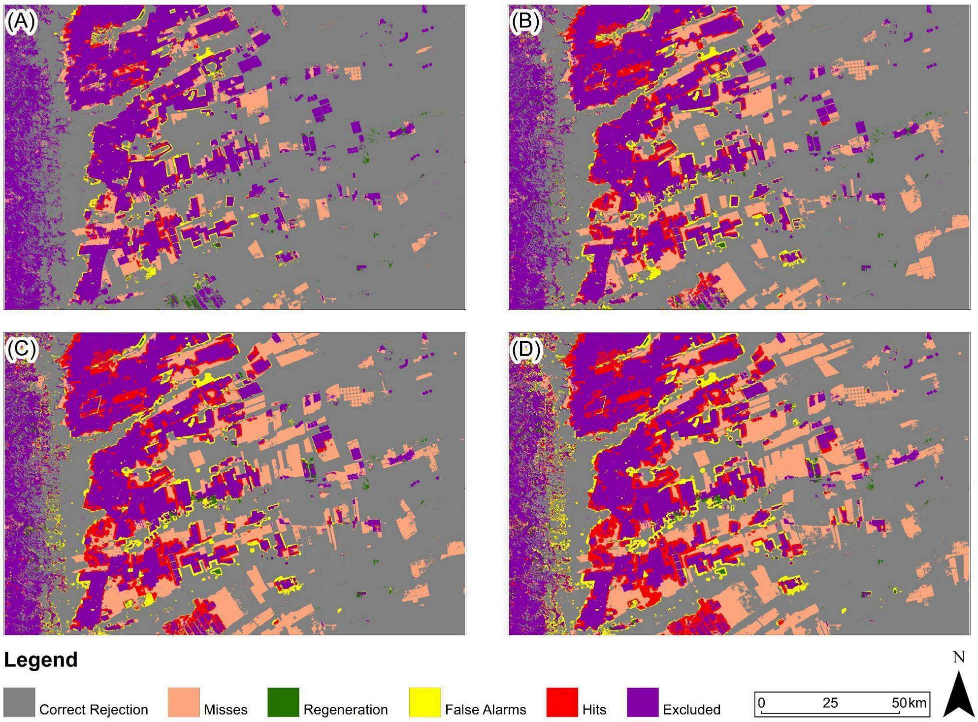
Figure 4.11: Three-map evaluation: Amazon Biome. (A) 2005 (B) 2010 (C) 2015 (D) 2020. The model presented for all periods is based on Euclidean distance for anthropogenic in 2000.



Source: Prepared by the authors.

For the study area in the Cerrado biome, the land use change prediction map with the highest Figure of Merit value for all periods was generated based on the ANN - TerrSet model.

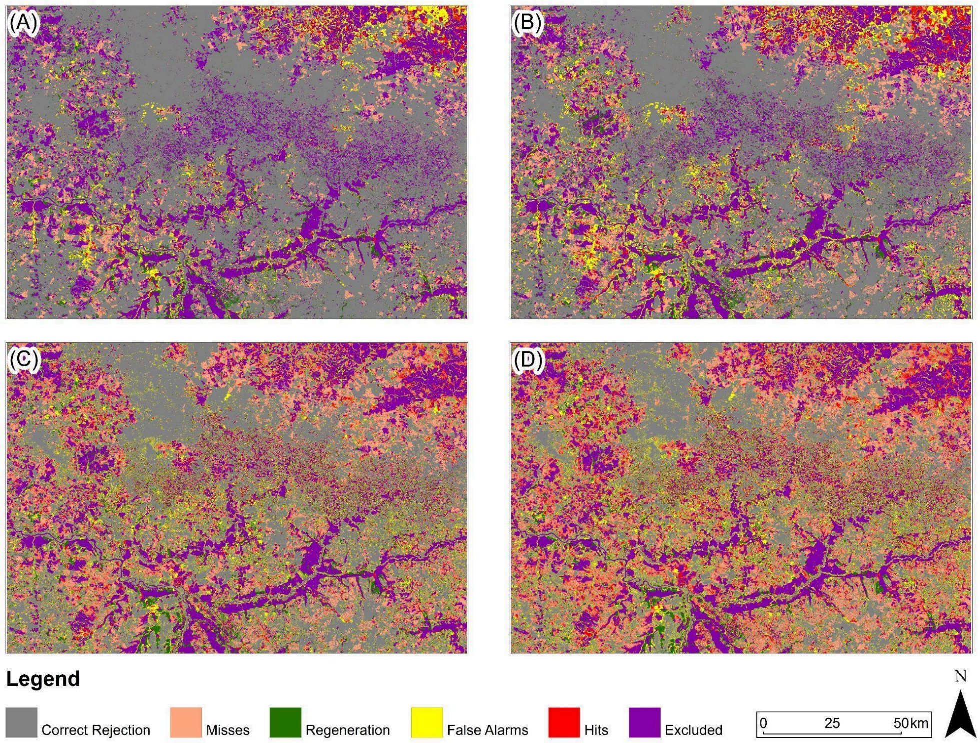
Figure 4.12: Three-map evaluation: Cerrado Biome. (A) 2005 (B) 2010 (C) 2015 (D) 2020. The model presented for all periods is ANN - TerrSet.



Source: Prepared by the authors.

For the study area in the Pampa biome, the land use change prediction map with the highest Figure of Merit value for 2005 and 2010 was generated based on the ANN-TerrSet model, and for the years 2015 and 2020 based on the Random Forest model. Letter A shows the map for the year 2005, B = 2010, C = 2015 and D = 2020.

Figure 4.13: Three-map evaluation: Pampa Biome. (A) 2005 (B) 2010 (C) 2015 (D) 2020. The model presented for 2005 and 2010 is ANN - TerrSet, and for 2015 and 2020 it is Random Forest.



Source: Prepared by the authors.

4.4. Discussion

The results presented in this article provide an opportunity to discuss relevant topics regarding land use change models. Among the items that can be analyzed, the discussion focuses on three main topics: the importance of evaluating the accuracy of model results using rigorous methods, the sources of disagreement in modeling, and the comparison between baseline models and machine learning-based models.

4.4.1 The importance of evaluating the accuracy of model results with rigorous methods

Accuracy assessment is a crucial step in land use change modeling, enabling users to understand strengths and weaknesses of each method. Unfortunately, the land use change modeling community tends to underestimate the importance of accuracy assessment. Many studies do not perform any type of accuracy assessment, while others frequently use flawed or potentially misleading methods. The ROC curve and overall accuracy stand out as the accuracy metrics most commonly employed (VAN VLIET et al., 2016; ABURAS; AHAMAD; OMAR, 2019).

The area under the ROC curve (AUC) is used in many studies as the sole approach for evaluating the accuracy of probability surfaces. Examples of this can be found in PARK et al. (2010), LIN et al. (2011), LIAO; WEI (2014), KUCSICSA et al. (2019), VOIGHT et al. (2019), and COLMAN et al. (2024). This practice is problematic because the ROC curve is generally used to evaluate the susceptibility surface relative to the change data observed during the training period. Under these conditions, the evaluation only serves to indicate the fit of the training, not providing information about the ability of the susceptibility surface to differentiate changes and no changes in the prediction periods.

Additionally, this approach does not provide any information about the spatial accuracy of change allocation. The results presented in this work show that the highest AUC value does not always coincide with the best land use change prediction map. In the study area of the Cerrado biome, for example, the Random Forest model obtained the highest AUC values for all periods. However, when evaluating the land use change prediction map by figure of merit, the ANN - TerrSet model showed the best performance. This demonstrates the importance of evaluating both the susceptibility surface and the prediction map.

Overall accuracy is a metric used to estimate the percentage of pixels correctly allocated by the prediction map. In many studies, this is the only approach used to evaluate the land use prediction map. Examples include the work of PARK et al. (2010), LIN et al. (2011), BALLESTORES JUNIOR; Zeyuan (2012), CUSHMAN et al. (2017), KUCSICSA et al. (2019). This method tends to be potentially misleading because it often considers the correct rejection component and excluded areas as model successes. However, the primary goal of developing a land use change model is to predict future changes. Therefore, a rigorous evaluation method should analyze the hits, misses, and false alarms components. This issue can be illustrated by the land use change prediction map derived from the Euclidean distance model for anthropogenic uses in the Amazon biome study area in 2005. The overall accuracy of the map is 85.6%, of which 96.6% is composed of correct rejections and excluded areas. When analyzing the hits, we observe that it is approximately seven times smaller than the sum of the misses and false alarms. Thus, analyzing only the global accuracy gives a misleading impression of the prediction's ability to represent future land use changes.

For the accuracy evaluation of the susceptibility surface, it has been demonstrated that the TOC curve provides more information compared to the ROC curve. Another good practice is the use of baseline models. The results of this work demonstrate that baseline models are more rigorous comparison methods than randomization, aiding in the interpretation of the performance of machine learning-based models (PONTIUS; SI, 2014; SHAFIZADEH-MOGHADAM et al., 2021; LIU; PONTIUS, 2021; HARATI et al., 2021).

For the evaluation of land use change prediction maps, among the pixel-based methods, the three-map comparison and figure of merit approaches provide results that allow understanding the size and proportion of agreement and disagreement components. This enables a detailed evaluation of the model's performance in capturing the components related to changes. (PONTIUS et al., 2007; PONTIUS, 2018).

Additionally, comparing prediction maps with a null change model is another recommended practice (PONTIUS; HUFFAKER; DENMAN, 2004; PONTIUS et al., 2007). It is expected that a prediction model will produce more accurate results than a null change model. However, this is not always true. The results of this work demonstrate that, in most modeled situations, the performance of the prediction map was inferior to the use of a null change model.

Another comparison approach used in this study was random-based models. Although the random models generally performed worse, the size of their agreement and disagreement components follows the same pattern as the other models. In this sense, it is evident that the majority of the agreements (hits and correct rejections) across all models are due to chance. This is an important contribution, as the agreement components are typically attributed to the good performance of the models.

Based on these results, it is clear that some common practices for evaluating models are flawed. In order to maximize the usefulness of predictions, it is necessary to understand their disagreements in detail. Although the models do not perfectly represent reality, some of them can be useful for specific objectives.

4.4.2 Sources of disagreement in land use change modeling

The disagreements observed in the model results originate from different sources. Among the possible origins, we can highlight land use maps, the amount of expected changes for the future, predictive variables, and the identification of patterns of change.

Land use maps are utilized in various stages of modeling, including the identification of changes during training, the formulation of predictive variables, and the evaluation of accuracy. Regardless of the method used to classify land use, the resulting maps do not represent the space perfectly (CONGALTON, 1991; FOODY, 2002; PONTIUS; MILLONES, 2011; FOODY, 2020). For the case of the MapBiomas project, the overall accuracy of the products is 96.8% for the Amazon biome, 84.7% for the Cerrado biome, and 85.8% for the Pampa biome (MAPBIOMAS, 2024). This shows that the use of land use maps adds confusion to the models, which is an inherent limitation of any modeling process. Knowing this difficulty, it is important that land use maps with known accuracy are used in modeling, which makes it possible to identify the uncertainties that the use of these products introduces into the models.

The predicted amount of change is another point of difficulty. When defining the expected amount of future changes based on the context of the present and past, there is no guarantee that the transition rates will remain the same (PONTIUS; SPENCER, 2005). The results found for the study area of the Cerrado Biome are good examples of this issue. The amount of vegetation suppression found for the training period (1995-2000) corresponds to 3.59% of the total study area. This rate was assumed to define the expected amount of change for the future. However, from 2000 to 2005, the amount of vegetation suppression corresponds to 7.58% of the total area. For 2005 to 2010, this percentage was 7.05%. Between 2010 and 2015, it was 5.25%. And from 2015 to 2020, it was 4.03%. In this sense, it is clear that the amount of vegetation suppression that occurred during the training period was less than in all the extrapolation periods. As a result, the predictions of vegetation suppression for the Cerrado biome study area projected fewer changes than actually occurred. This resulted in a large number of misses, surpassing the number of false alarms in all situations. For this problem, regardless of the strategy employed, there are no guarantees of better results.

The spatial variables used to identify the patterns that explain land use changes may have various errors. In addition to the variables derived from land use maps, it is common to use Euclidean distance surfaces for elements such as highways, conservation units, protected areas, etc. If these elements have errors in location, delineation, georeferencing, etc., their use adds confusion to the modeling process. In this sense, this stage deserves special attention, as it is necessary to evaluate whether the products used accurately represent what is desired.

Furthermore, identifying patterns of land use change during the training period does not guarantee that these patterns will remain the same during the prediction stage. Between the prediction and extrapolation periods, political, economic, social, and infrastructural changes may occur, driving the occupation of areas not identified in the training as susceptible to change. Additionally, when modeling changes for extensive areas, different locations may have different factors that induce the occurrence of changes (TRIGUEIRO; NABOUT; TESSAROLO, 2020).

In this sense, the results show that for all models and study areas, the further away from the training period, the worse the models perform. Knowing these limitations, the models should be interpreted considering that the patterns of change may shift. Additionally, knowing the spatial variability of the factors that explain the changes, the definition of the extrapolation area of the patterns identified in the training should be done with caution.

4.4.3 Comparison between machine learning-based and baseline models

For the predictions presented in the results, it is observed that the models of both natures obtained similar performances in all the evaluation methods. The works of SHAFIZADEH-MOGHADAM et al. (2021) and HARATI et al. (2021) can be cited, which corroborate what was found. These studies show the similarity of performance between baseline and machine learning-based models for predicting urban sprawl and forest insect disturbance.

Thus, the results found show that the effort invested in developing machine learning-based models did not provide accuracy better than baseline models. Furthermore, it is observed that in baseline models there is a direct relationship between the prediction of vegetation suppression and the proximity to the class of anthropogenic uses and past vegetation suppressions. While for machine learning-based models it is not explicit which criteria were used to define susceptibility to vegetation suppression. In view of this, the use of baseline models is recommended due to their ease of formulation, understanding and similar performance to machine learning-based models. If machine learning-based models are used, the baseline models can be used to compare the results.

4.5. Conclusions

The results and discussions presented in this work highlight important points about land use change modeling: (1) The use of baseline land use change models should be encouraged due to their quick and easy formulation process, requiring few data and not needing powerful hardware; their results are easily interpretable; and their performance is comparable to methods based on machine learning. (2) Land use change models have a limited ability to predict future scenarios. All the land use change prediction models applied in this work, regardless of the formulation method, obtained more disagreements than correctly predicted changes. Additionally, analyzing the performance of randomness-based models shows that a considerable portion of the hits of all models is due to chance. (3) Predicting land use change for periods far from the training tends to generate less accurate results than for closer periods. And (4), the evaluation of model results using inappropriate methods leads to an interpretation that generally overestimates the accuracy of predictions. It was evident that the exclusive use of evaluation methods such as the ROC curve and overall accuracy generated misleading results.

Given the increasing use of land use change models in a wide variety of applications, it is important for the scientific community to recognize the limitations of this technique to better utilize it. So, it is recommended that researchers consider the indications presented in this document to better understand and communicate the results of this type of model.

Supplementary Materials

The computational codes, the reclassification of land use maps, the spatial variables, the susceptibility models for vegetation suppression, and the land use prediction maps used in this article can be accessed through the link: https://github.com/macleidivarnier/How-accurate-are-land-change-models

References

ABURAS, Maher Milad; AHAMAD, Mohd Sanusi S.; OMAR, Najat Qader. Spatio-temporal simulation and prediction of land-use change using conventional and machine learning models: a review. **Environmental Monitoring And Assessment**, [S.L.], v. 191, n. 4, p. 191-205, 5 mar. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10661-019-7330-6.

BALLESTORES JUNIOR, Florencio; QIU, Zeyuan. An integrated parcel-based land use ch ange model using cellular automataand decision tree. **Proceedings Of The International Academy Of Ecology And Environmental Sciences**, v. 2, n. 2, p. 53-69, 2012.

BREIMAN, Leo. Random Forests, **Machine Learning**, 2001, vol. 45, pp. 5–32.

CHAVAN, Sapana; REDDY, C. Sudhakar; RAO, S. Srinivasa; RAO, K. Kameswara. Assessing and Predicting Decadal Forest Cover Changes and Forest Fragmentation in Kinnerasani Wildlife Sanctuary, Telangana, India. **Journal Of The Indian Society Of Remote Sensing**, [S.L.], v. 46, n. 5, p. 729-735, 10 jan. 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s12524-017-0739-x.

COLMAN, Carina Barbosa; GUERRA, Angélica; ALMAGRO, André; ROQUE, Fabio de Oliveira; ROSA, Isabel M. D.; FERNANDES, Geraldo Wilson; OLIVEIRA, Paulo Tarso S.. Modeling the Brazilian Cerrado land use change highlights the need to account for private property sizes for biodiversity conservation. **Scientific Reports**, [S.L.], v. 14, n. 1, p. 1-11, 24 fev. 2024. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-024-55207-1.

CONGALTON, Russell. A review of assessing the accuracy of classifications of remotely sensed data. **Remote Sensing Of Environment**, [S.L.], v. 37, n. 1, p. 35-46, jul. 1991. Elsevier BV. http://dx.doi.org/10.1016/0034-4257(91)90048-b.

FOODY, Giles. Status of land cover classification accuracy assessment. **Remote Sensing Of Environment**, [S.L.], v. 80, n. 1, p. 185-201, abr. 2002. Elsevier BV. http://dx.doi.org/10.1016/s0034-4257(01)00295-4.

FOODY, Giles. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification.

Remote Sensing Of Environment, [S.L.], v. 239, p. 111630, mar. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.rse.2019.111630.

GOOGLE. **Machine learning glossary**. Available in: <https://developers.google.com/machine-learning/glossary.>. Access: 17/04/2024.

GOOGLE EARTH ENGINE. **Smile Random Forest**. Available in: <https://developers.google.com/earth-engine/apidocs/ee-classifier-smilerandomforest>. Access: 17/04/2024.

GORELICK, Noel; HANCHER, Matt; DIXON, Mike; ILYUSHCHENKO, Simon; THAU, David; MOORE, Rebecca. Google Earth Engine:planetary-scale geospatial analysis for everyone. **Remote Sensing Of Environment**, v. 202, p. 18-27, 2017. DOI:http://dx.doi.org/10.1016/j.rse.2017.06.031

HARATI, Saeed; PEREZ, Liliana; MOLOWNY-HORAS, Roberto; PONTIUS, Robert Gilmore. Validating models of one-way land change: an example case of forest insect disturbance. **Landscape Ecology**, [S.L.], v. 36, n. 10, p. 2919-2935, 25 jun. 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10980-021-01272-0.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Bases Cartográficas Contínuas (Versão 2021)**. https://geoftp.ibge.gov.br/cartas_e_mapas/bases_cartograficas_continuas/bc250/ (2024)

KUCSICSA, Gheorghe; POPOVICI, Elena-Ana; BăLTEANU, Dan; DUMITRAşCU, Monica; GRIGORESCU, Ines; MITRICă, Bianca. Assessing the Potential Future Forest-Cover Change in Romania, Predicted Using a Scenario-Based Modelling. **Environmental Modeling & Assessment**, [S.L.], v. 25, n. 4, p. 471-491, 12 nov. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10666-019-09686-6.

LIAO, Felix; WEI, Dennis. Modeling determinants of urban growth in Dongguan, China: a spatial logistic approach. **Stochastic Environmental Research And Risk Assessment**, [S.L.], v. 28, n. 4, p. 801-816, 28 set. 2012. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00477-012-0620-y.

LIN, Yu-Pin; CHU, Hone-Jay; WU, Chen-Fa; VERBURG, Peter H. Predictive ability of logistic regression, auto-logistic regression and neural network models in empirical land-use change modeling – a case study. **International Journal Of Geographical Information Science**, [S.L.], v. 25, n. 1, p. 65-87, 23 fev. 2011. Informa UK Limited. http://dx.doi.org/10.1080/13658811003752332.

LIU, Zhen; PONTIUS, Robert Gilmore. The Total Operating Characteristic from Stratified Random Sampling with an Application to Flood Mapping. **Remote Sensing**, [S.L.], v. 13, n. 19, p. 3922, 30 set. 2021. MDPI AG. http://dx.doi.org/10.3390/rs13193922.

MAPBIOMAS. **O Projeto**. 2023. Avaliable in: https://mapbiomas.org/o-projeto. Acess in: 25 jun. 2023.

MAPBIOMAS. Análise de acurácia. 2024. Avaliable in: https://brasil.mapbiomas.org/analise-de-acuracia/. Acess in: 28 abr. 2024.

NASA. 2023. **NASADEM: Creating a New NASA Digital Elevation Model and Associated Products**. Available in: <https://www.earthdata.nasa.gov/esds/competitive-programs/measures/nasadem>. Access: 17/09/2023

PARK, Soyoung; JEON, Seongwoo; KIM, Shinyup; CHOI, Chuluong. Prediction and comparison of urban growth by land suitability index mapping using GIS and RS in South Korea. **Landscape And Urban Planning**, [S.L.], v. 99, n. 2, p. 104-114, fev. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.landurbplan.2010.09.001.

PONTIUS, Robert Gilmore; HUFFAKER, Diana; DENMAN, Kevin. Useful techniques of validation for spatially explicit land-change models. **Ecological Modelling**, [S.L.], v. 179, n. 4, p. 445-461, dez. 2004. Elsevier BV. http://dx.doi.org/10.1016/j.ecolmodel.2004.05.010.

PONTIUS, Robert Gilmore; SPENCER, Joseph. Uncertainty in Extrapolations of Predictive Land-Change Models. **Environment And Planning B: Planning and Design**, [S.L.], v. 32, n. 2, p. 211-230, abr. 2005. SAGE Publications. http://dx.doi.org/10.1068/b31152.

PONTIUS, Robert Gilmore; BOERSMA, Wideke; CASTELLA, Jean-Christophe; CLARKE, Keith; NIJS, Ton de; DIETZEL, Charles; DUAN, Zengqiang; FOTSING, Eric; GOLDSTEIN, Noah; KOK, Kasper. Comparing the input, output, and validation maps for several models of land change. T**he Annals Of Regional Science**, [S.L.], v. 42, n. 1, p. 11-37, 16 ago. 2007. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00168-007-0138-2.

PONTIUS, Robert Gilmore; MILLONES, Marco. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. **International Journal Of Remote Sensing**, [S.L.], v. 32, n. 15, p. 4407-4429, 10 ago. 2011. Informa UK Limited. http://dx.doi.org/10.1080/01431161.2011.552923.

PONTIUS, Robert Gilmore; PARMENTIER, Benoit. Recommendations for using the relative operating characteristic (ROC). **Landscape Ecology**, [S.L.], v. 29, n. 3, p. 367-382, 28 jan. 2014. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10980-013-9984-8.

PONTIUS, Robert. Criteria to Confirm Models that Simulate Deforestation and Carbon Disturbance. **Land**, [S.L.], v. 7, n. 3, p. 105, 10 set. 2018. MDPI AG. http://dx.doi.org/10.3390/land7030105.

PONTIUS, Robert Gilmore; BILINTOH, Thomas; OLIVEIRA, Gustavo de L. T.; SHIMBO, Julia Z.. TRAJECTORIES OF LOSSES AND GAINS OF SOYBEAN CULTIVATION DURING MULTIPLE TIME INTERVALS IN WESTERN BAHIA, BRAZIL. In: **PROCEEDINGS OF SPACE WEEK NORDESTE 2023**, 1., 2023, Fortaleza. Anais [...] . Fortaleza: Galoá, 2023. p. 1-4.

SANGERMANO, Florencia; EASTMAN, J Ronald; ZHU, Honglei. Similarity Weighted Instance-based Learning for the Generation of Transition Potentials in

Land Use Change Modeling. **Transactions In Gis**, [S.L.], v. 14, n. 5, p. 569-580, out. 2010. Wiley. http://dx.doi.org/10.1111/j.1467-9671.2010.01226.x.

SHAFIZADEH-MOGHADAM, Hossein; MINAEI, Masoud; PONTIUS JUNIOR, Robert Gilmore; ASGHARI, Ali; DADASHPOOR, Hashem. Integrating a Forward Feature Selection algorithm, Random Forest, and Cellular Automata to extrapolate urban growth in the Tehran-Karaj Region of Iran. Computers, **Environment And Urban Systems**, [S.L.], v. 87, p. 101595, maio 2021. Elsevier BV. http://dx.doi.org/10.1016/j.compenvurbsys.2021.101595.

SOARES-FILHO, Britaldo Silveira; CERQUEIRA, Gustavo Coutinho; PENNACHIN, Cássio Lopes. Dinamica—a stochastic cellular automata model designed to simulate the landscape dynamics in an Amazonian colonization frontier. **Ecological Modelling**, [S.L.], v. 154, n. 3, p. 217-235, set. 2002. Elsevier BV. http://dx.doi.org/10.1016/s0304-3800(02)00059-5.

SOUZA, Carlos. et al. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. **Remote Sensing**, [S.L.], v. 12, n. 17, p. 2735, 25 ago. 2020. MDPI AG. http://dx.doi.org/10.3390/rs12172735.

TRIGUEIRO, Werikson Rodrigues; NABOUT, João Carlos; TESSAROLO, Geiziane. Uncovering the spatial variability of recent deforestation drivers in the Brazilian Cerrado. **Journal Of Environmental Management**, [S.L.], v. 275, p. 111243, dez. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.jenvman.2020.111243.

VAN VLIET, Jasper; BREGT, Arnold K.; BROWN, Daniel G.; VAN DELDEN, Hedwig; HECKBERT, Scott; VERBURG, Peter H.. A review of current calibration and validation practices in land-change modeling. **Environmental Modelling & Software**, [S.L.], v. 82, p. 174-182, ago. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.envsoft.2016.04.017.

5 CONCLUSÕES E RECOMENDAÇÕES

As avaliações realizadas neste trabalho auxiliam no melhor entendimento do funcionamento e da acurácia de modelos de predição de mudanças de uso do solo. A partir dos resultados encontrados, pode se concluir que:

I) As variáveis preditivas utilizadas no treinamento dos modelos demonstraram diferentes capacidades de descrever as mudanças. Em sentido semelhante, nas diferentes áreas de estudo investigadas, as variáveis espaciais representaram as supressões de maneira particular em cada espaço.

II) Na avaliação da acurácia dos modelos, ficou demonstrada que a utilização exclusiva de métodos como a curva ROC e a acurácia global podem fornecer resultados potencialmente enganosos. Por outro lado, os métodos de curva TOC, avaliação por três mapas e figura do mérito forneceram resultados rigorosos e de alto detalhamento.

III) Tanto na avaliação das superfícies de suscetibilidade, quanto na avaliação dos mapas de predição de uso do solo, o desempenho dos modelos baseados em métodos de linha de base e em machine learning foi muito similar. Tendo em vista esta similaridade, percebe-se que o impacto das variáveis preditivas na acurácia dos modelos é maior do que o método de modelagem.

IV) Na avaliação dos mapas de predição de uso do solo, os modelos de mudanças nulas e baseados na aleatoriedade obtiveram acurácias similares aos demais métodos. Isso demonstra que boa parte dos acertos dos modelos de mudanças se deve ao acaso, não necessariamente indicando um bom desempenho do método empregado.

V) Em todos os modelos avaliados a quantidade de mudanças preditas de maneira correta foi menor do que a quantidade de falsos alarmes e omissões, evidenciando que a alocação espacial incorreta é a principal fraqueza dos modelos.

A partir destas conclusões, recomenda-se que a definição das variáveis utilizadas no treinamento dos modelos seja baseada em testes estatísticos. Essa medida auxilia na definição de variáveis com capacidade de representar as mudanças, além de permitir a redução da dimensionalidade dos dados e o tempo de processamento computacional. Para a definição da área de estudo a ser modelada, indica-se a escolha de regiões com características humanas e naturais homogêneas, facilitando a identificação de padrões de mudanças eficientes.

Quanto aos métodos de modelagem, considerando a similaridade de desempenho entre as abordagens de linha de base e baseadas em machine learning, recomenda-se a utilização de modelos de linha de base. Isso se deve pela maior facilidade de formulação e compreensão da lógica de funcionamento destes modelos. Caso se opte pela utilização de modelos baseados em machine learning, os modelos de linha de base devem ser empregados como linha de base para comparação de desempenho.

Sobre a avaliação da acurácia dos modelos, recomenda-se a utilização dos métodos de curva TOC, avaliação por três mapas e figura do mérito. Outra boa prática recomendada é a utilização dos modelos de linha de base, como o modelos de mudanças nulas e baseadas na aleatoriedade. Esta medida possibilita avaliar a modelagem de maneira comparativa, informando sobre o ganho de informação obtido pela elaboração de um modelo complexo.

FINANCIAMENTO

Este trabalho foi financiado pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior do Brasil) através de uma bolsa de mestrado (processo número: 88887.721596/2022-00).

REFERÊNCIAS

ABURAS, Maher Milad; AHAMAD, Mohd Sanusi S.; OMAR, Najat Qader. Spatio-temporal simulation and prediction of land-use change using conventional and machine learning models: a review. **Environmental Monitoring And Assessment**, [S.L.], v. 191, n. 4, p. 191-205, 5 mar. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10661-019-7330-6.

ADHIKARI, Sanchayeeta; SOUTHWORTH, Jane. Simulating Forest Cover Changes of Bannerghatta National Park Based on a CA-Markov Model: a remote sensing approach. **Remote Sensing**, [S.L.], v. 4, n. 10, p. 3215-3243, 22 out. 2012. MDPI AG. http://dx.doi.org/10.3390/rs4103215.

BALLESTORES JUNIOR, Florencio; QIU, Zeyuan. An integrated parcel-based land use change model using cellular automata and decision tree. **Proceedings Of The International Academy Of Ecology And Environmental Sciences**, v. 2, n. 2, p. 53-69, 2012.

BARLOW, Jos; LENNOX, Gareth D.; FERREIRA, Joice; BERENGUER, Erika; LEES, Alexander C.; NALLY, Ralph Mac; THOMSON, James R.; FERRAZ, Silvio Frosini de Barros; LOUZADA, Julio; OLIVEIRA, Victor Hugo Fonseca. Anthropogenic disturbance in tropical forests can double biodiversity loss from deforestation. **Nature**, [S.L.], v. 535, n. 7610, p. 144-147, 29 jun. 2016. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nature18326.

BREIMAN, Leo. Random Forests, **Machine Learning**, 2001, vol. 45, pp. 5–32.

BRITO, Mariana Madruga de; WEBER, Eliseu José; KRIGGER, Vania Silva; LEITZKE, Felipe Padilha. ANÁLISE DOS FATORES CONDICIONANTES DE MOVIMENTOS DE MASSA NO MUNICÍPIO DE PORTO ALEGRE A PARTIR DE REGISTROS HISTÓRICOS. **Revista Brasileira de Cartografia**, [S.L.], v. 68, n. 9, p. 1853-1872, 16 mar. 2017. EDUFU - Editora da Universidade Federal de Uberlândia. http://dx.doi.org/10.14393/rbcv68n9-44450.

CHAVAN, Sapana; REDDY, C. Sudhakar; RAO, S. Srinivasa; RAO, K. Kameswara. Assessing and Predicting Decadal Forest Cover Changes and Forest Fragmentation in Kinnerasani Wildlife Sanctuary, Telangana, India. **Journal Of The Indian Society Of Remote Sensing**, [S.L.], v. 46, n. 5, p. 729-735, 10 jan. 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s12524-017-0739-x.

COLMAN, Carina Barbosa; GUERRA, Angélica; ALMAGRO, André; ROQUE, Fabio de Oliveira; ROSA, Isabel M. D.; FERNANDES, Geraldo Wilson; OLIVEIRA, Paulo Tarso S.. Modeling the Brazilian Cerrado land use change highlights the need to account for private property sizes for biodiversity conservation. **Scientific Reports**, [S.L.], v. 14, n. 1, p. 1-11, 24 fev. 2024. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-024-55207-1.

CONGALTON, Russell. A review of assessing the accuracy of classifications of remotely sensed data. **Remote Sensing Of Environment**, [S.L.], v. 37, n. 1, p. 35-46, jul. 1991. Elsevier BV. http://dx.doi.org/10.1016/0034-4257(91)90048-b.

CUSHMAN, Samuel; MACDONALD, Ewan; LANDGUTH, Erin; MALHI, Yadvinder; MACDONALD, David. Multiple-scale prediction of forest loss risk across Borneo. **Landscape Ecology**, [S.L.], v. 32, n. 8, p. 1581-1598, 18 maio de 2017. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10980-017-0520-0.

FIELDING, Alan H.; BELL, John F.. A review of methods for the assessment of prediction errors in conservation presence/absence models. **Environmental Conservation**, [S.L.], v. 24, n. 1, p. 38-49, mar. 1997. Cambridge University Press (CUP). http://dx.doi.org/10.1017/s0376892997000088.

FOODY, Giles. Status of land cover classification accuracy assessment. **Remote Sensing Of Environment**, [S.L.], v. 80, n. 1, p. 185-201, abr. 2002. Elsevier BV. http://dx.doi.org/10.1016/s0034-4257(01)00295-4.

FOODY, Giles. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. **Remote Sensing Of Environment**, [S.L.], v. 239, p. 111630, mar. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.rse.2019.111630.

GOOGLE. **Machine learning glossary**. Disponível em: <https://developers.google.com/machine-learning/glossary.>. Accesso: 17/04/2024.

GOOGLE EARTH ENGINE. **Smile Random Forest**. Disponível em: <https://developers.google.com/earth-engine/apidocs/ee-classifier-smilerandomforest>. Accesso: 17/04/2024.

GORELICK, Noel; HANCHER, Matt; DIXON, Mike; ILYUSHCHENKO, Simon; THAU, David; MOORE, Rebecca. Google Earth Engine:planetary-scale geospatial analysis for everyone. **Remote Sensing Of Environment**, v. 202, p. 18-27, 2017. DOI:http://dx.doi.org/10.1016/j.rse.2017.06.031

GREEN, Kass; KEMPKA, Dick; LACKEY, Lisa. Using remote sensing to detect and monitor land-cover and land-use change. **Photogrammetric Engineering And Remote Sensing**, [S.L.], 1994.

HARATI, Saeed; PEREZ, Liliana; MOLOWNY-HORAS, Roberto; PONTIUS, Robert Gilmore. Validating models of one-way land change: an example case of forest insect disturbance. **Landscape Ecology**, [S.L.], v. 36, n. 10, p. 2919-2935, 25 jun. 2021. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10980-021-01272-0.

HINTZE, Jerry; NELSON, Ray. Violin Plots: a box plot-density trace synergism. **The American Statistician**, [S.L.], v. 52, n. 2, p. 181-184, maio 1998. Informa UK Limited. http://dx.doi.org/10.1080/00031305.1998.10480559.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Bases Cartográficas Contínuas (Versão 2021)**. https://geoftp.ibge.gov.br/cartas_e_mapas/bases_cartograficas_continuas/bc250/ (2024)

JENSEN, John. **Remote Sensing of the Environment: an earth resource perspective**. Nova York: Prentice Hall, 2000.

KUCSICSA, Gheorghe; POPOVICI, Elena-Ana; BăLTEANU, Dan; DUMITRAşCU, Monica; GRIGORESCU, Ines; MITRICă, Bianca. Assessing the Potential Future Forest-Cover Change in Romania, Predicted Using a Scenario-Based Modelling. **Environmental Modeling & Assessment**, [S.L.], v. 25, n. 4, p. 471-491, 12 nov. 2019. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10666-019-09686-6.

LIAO, Felix; WEI, Dennis. Modeling determinants of urban growth in Dongguan, China: a spatial logistic approach. **Stochastic Environmental Research And Risk Assessment**, [S.L.], v. 28, n. 4, p. 801-816, 28 set. 2012. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00477-012-0620-y.

LIN, Yu-Pin; CHU, Hone-Jay; WU, Chen-Fa; VERBURG, Peter H. Predictive ability of logistic regression, auto-logistic regression and neural network models in empirical land-use change modeling – a case study. **International Journal Of Geographical Information Science**, [S.L.], v. 25, n. 1, p. 65-87, 23 fev. 2011. Informa UK Limited. http://dx.doi.org/10.1080/13658811003752332.

LIU, Zhen; PONTIUS, Robert Gilmore. The Total Operating Characteristic from Stratified Random Sampling with an Application to Flood Mapping. **Remote Sensing**, [S.L.], v. 13, n. 19, p. 3922, 30 set. 2021. MDPI AG. http://dx.doi.org/10.3390/rs13193922.

MAPBIOMAS. **O Projeto**. 2023. Disponível em: https://mapbiomas.org/o-projeto. Acesso em: 25 jun. 2023.

MAPBIOMAS. **Análise de acurácia**. 2024. Disponível em: https://brasil.mapbiomas.org/analise-de-acuracia/. Acesso em: 28 abr. 2024.

MICROSOFT CORPORATION. **Microsoft Excel [Software]**. Versão 16.0. Redmond, WA: Microsoft, 2023.

NASA. 2023. NASADEM: **Creating a New NASA Digital Elevation Model and Associated Products**. Available in:<https://www.earthdata.nasa.gov/esds/competitive-programs/measures/nasadem>. Access: 17/09/2023

OLMEDO, María Teresa Camacho; PONTIUS, Robert Gilmore; PAEGELOW, Martin; MAS, Jean-François. Comparison of simulation models in terms of quantity and allocation of land change. **Environmental Modelling & Software**, [S.L.], v. 69, p. 214-221, jul. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.envsoft.2015.03.003.

PARK, Soyoung; JEON, Seongwoo; KIM, Shinyup; CHOI, Chuluong. Prediction and comparison of urban growth by land suitability index mapping using GIS and RS in South Korea. **Landscape And Urban Planning**, [S.L.], v. 99, n. 2, p. 104-114, fev. 2011. Elsevier BV. http://dx.doi.org/10.1016/j.landurbplan.2010.09.001.

PIAO, Yong; LEE, Dongkun; PARK, Sangjin; KIM, Ho Gul; JIN, Yihua. Forest fire susceptibility assessment using google earth engine in Gangwon-do, Republic of Korea. **Geomatics, Natural Hazards And Risk**, [S.L.], v. 13, n. 1, p. 432-450, 2 fev. 2022. Informa UK Limited. http://dx.doi.org/10.1080/19475705.2022.2030808.

PONTIUS, Robert Gilmore; HUFFAKER, Diana; DENMAN, Kevin. Useful techniques of validation for spatially explicit land-change models. **Ecological Modelling**, [S.L.], v. 179, n. 4, p. 445-461, dez. 2004. Elsevier BV. http://dx.doi.org/10.1016/j.ecolmodel.2004.05.010.

PONTIUS, Robert Gilmore; SPENCER, Joseph. Uncertainty in Extrapolations of Predictive Land-Change Models. **Environment And Planning B: Planning and Design**, [S.L.], v. 32, n. 2, p. 211-230, abr. 2005. SAGE Publications. http://dx.doi.org/10.1068/b31152.

PONTIUS, Robert Gilmore; BOERSMA, Wideke; CASTELLA, Jean-Christophe; CLARKE, Keith; NIJS, Ton de; DIETZEL, Charles; DUAN, Zengqiang; FOTSING, Eric; GOLDSTEIN, Noah; KOK, Kasper. Comparing the input, output, and validation maps for several models of land change. T**he Annals Of Regional Science**, [S.L.], v. 42, n. 1, p. 11-37, 16 ago. 2007. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s00168-007-0138-2.

PONTIUS, Robert Gilmore; PEETHAMBARAM, Smitha; CASTELLA, Jean-Christophe. Comparison of Three Maps at Multiple Resolutions: a case study of land change simulation in cho don district, vietnam. **Annals Of The Association Of American Geographers**, [S.L.], v. 101, n. 1, p. 45-62, 21 jan. 2011. Informa UK Limited. http://dx.doi.org/10.1080/00045608.2010.517742.

PONTIUS, Robert Gilmore; MILLONES, Marco. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. **International Journal Of Remote Sensing**, [S.L.], v. 32, n. 15, p. 4407-4429, 10 ago. 2011. Informa UK Limited. http://dx.doi.org/10.1080/01431161.2011.552923

PONTIUS, Robert Gilmore; PARMENTIER, Benoit. Recommendations for using the relative operating characteristic (ROC). **Landscape Ecology**, [S.L.], v. 29, n. 3, p. 367-382, 28 jan. 2014. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10980-013-9984-8.

PONTIUS, Robert Gilmore; SI, Kangping. The total operating characteristic to measure diagnostic ability for multiple thresholds. **International Journal Of Geographical Information Science**, [S.L.], v. 28, n. 3, p. 570-583, 6 jan. 2014. Informa UK Limited. http://dx.doi.org/10.1080/13658816.2013.862623.

PONTIUS, Robert Gilmore; KRITHIVASAN, Roopa; SAULS, Laura; YAN, Yan; ZHANG, Yujia. Methods to summarize change among land categories across time intervals. **Journal Of Land Use Science**, [S.L.], v. 12, n. 4, p. 218-230, 30 jun. 2017. Informa UK Limited. http://dx.doi.org/10.1080/1747423x.2017.1338768

PONTIUS, Robert. Criteria to Confirm Models that Simulate Deforestation and Carbon Disturbance. **Land**, [S.L.], v. 7, n. 3, p. 105, 10 set. 2018. MDPI AG. http://dx.doi.org/10.3390/land7030105.

PONTIUS, Robert Gilmore. **Metrics That Make a Difference**: how to analyze change and error. Gewerbestrasse: Springer, 2022.

PONTIUS, Robert Gilmore; BILINTOH, Thomas; OLIVEIRA, Gustavo de; SHIMBO, Julia. TRAJECTORIES OF LOSSES AND GAINS OF SOYBEAN CULTIVATION DURING MULTIPLE TIME INTERVALS IN WESTERN BAHIA, BRAZIL. In: **PROCEEDINGS OF SPACE WEEK NORDESTE 2023**, 1., 2023, Fortaleza. Anais [...] . Fortaleza: Galoá, 2023. p. 1-4.

ROGAN, John; CHEN, Dongmei. Remote sensing technology for mapping and monitoring land-cover and land-use change. **Progress In Planning**, [S.L.], v. 61, n. 4, p. 301-325, maio 2004. Elsevier BV. http://dx.doi.org/10.1016/s0305-9006(03)00066-7.

SANGERMANO, Florencia; EASTMAN, J Ronald; ZHU, Honglei. Similarity Weighted Instance‐based Learning for the Generation of Transition Potentials in Land Use Change Modeling. **Transactions In Gis**, [S.L.], v. 14, n. 5, p. 569-580, out. 2010. Wiley. http://dx.doi.org/10.1111/j.1467-9671.2010.01226.x.

SHAFIZADEH-MOGHADAM, Hossein; MINAEI, Masoud; PONTIUS JUNIOR, Robert Gilmore; ASGHARI, Ali; DADASHPOOR, Hashem. Integrating a Forward Feature Selection algorithm, Random Forest, and Cellular Automata to extrapolate urban growth in the Tehran-Karaj Region of Iran. Computers, **Environment And Urban Systems**, [S.L.], v. 87, p. 101595, maio 2021. Elsevier BV. http://dx.doi.org/10.1016/j.compenvurbsys.2021.101595.

SHIMABUKURO, Yosio; SMITH, James. The least-squares mixing models to generate fraction images derived from remote sensing multispectral data. **Ieee Transactions On Geoscience And Remote Sensing**, [S.L.], v. 29, n. 1, p. 16-20, 1991. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/36.103288.

SOARES-FILHO, Britaldo Silveira; CERQUEIRA, Gustavo Coutinho; PENNACHIN, Cássio Lopes. Dinamica—a stochastic cellular automata model designed to simulate the landscape dynamics in an Amazonian colonization frontier. **Ecological Modelling**, [S.L.], v. 154, n. 3, p. 217-235, set. 2002. Elsevier BV. http://dx.doi.org/10.1016/s0304-3800(02)00059-5.

SOARES FILHO, Britaldo Silveira; CERQUEIRA, Gustavo Coutinho; ARAðJO, William Leite; VOLL, Eliane. Modelagem de dinâmica de paisagem: concepção e potencial de aplicação de modelos de simulação baseados em autômato celular. **Megadiversidade**, Belo Horizonte, v. 3, n. -, p. 74-83, dez. 2007.

SOUZA, Carlos. et al. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. **Remote Sensing**, [S.L.], v. 12, n. 17, p. 2735, 25 ago. 2020. MDPI AG. http://dx.doi.org/10.3390/rs12172735.

TALUKDAR, Swapan; SINGHA, Pankaj; MAHATO, Susanta; SHAHFAHAD; PAL, Swades; LIOU, Yuei-An; RAHMAN, Atiqur. Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. **Remote Sensing**, [S.L.], v. 12, n. 7, p. 1135, 2 abr. 2020. MDPI AG. http://dx.doi.org/10.3390/rs12071135.

TANG, Doudou; CHEN, Mingjie; HUANG, Xinhua; ZHANG, Guicheng; ZENG, Lin; ZHANG, Guangsen; WU, Shangjie; WANG, Yewei. SRplot: a free online platform for data visualization and graphing. **Plos One**, [S.L.], v. 18, n. 11, p. 1-8, 9 nov. 2023. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0294236.

TRIGUEIRO, Werikson Rodrigues; NABOUT, João Carlos; TESSAROLO, Geiziane. Uncovering the spatial variability of recent deforestation drivers in the Brazilian Cerrado. **Journal Of Environmental Management**, [S.L.], v. 275, p. 111243, dez. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.jenvman.2020.111243.

TURKEY, John. **Exploratory Data Analysis**, Reading, MA: AddisonWesley. 1977.

VAN VLIET, Jasper; BREGT, Arnold K.; BROWN, Daniel G.; VAN DELDEN, Hedwig; HECKBERT, Scott; VERBURG, Peter H.. A review of current calibration and validation practices in land-change modeling. **Environmental Modelling & Software**, [S.L.], v. 82, p. 174-182, ago. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.envsoft.2016.04.017.

VOIGHT, Carly; HERNANDEZ-AGUILAR, Karla; GARCIA, Christina; GUTIERREZ, Said. Predictive Modeling of Future Forest Cover Change Patterns in Southern Belize. **Remote Sensing**, [S.L.], v. 11, n. 7, p. 823, 5 abr. 2019. MDPI AG. http://dx.doi.org/10.3390/rs11070823.