Revista Brasileira de Cartografia
ISSN 1808-0936 | https://doi.org/10.14393/revbrascartogr
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto

Análise Comparativa Entre o Uso de Bandas Espectrais e o Uso da Análise de Componentes Principais (ACP) na Classificação de Uso e Cobertura da Terra

Comparative Analysis Between Use of Spectral Bands and Use of Principal Component Analysis (PCA) in Land Use and Cover Classification

Leticia Figueiredo Sartorio[1], Macleidi Varnier[2], Leonardo Da Silva Felippe[3], Daniel Capella Zanotta[4], Marcos Wellausen Dias de Freitas[5] e Atílio Efrain Bica Grondona [5]

1 Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil. leticia.sartorio98@gmail.com
ORCID: https://orcid.org/0000-0001-6936-9939
2 Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil. macleidivarnier@gmail.com
ORCID: https://orcid.org/0000-0002-2255-7294
3 Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil. leonardo.felippe@ufrgs.br
ORCID: https://orcid.org/0000-0001-9917-0103
4 Universidade do Vale do Rio dos Sinos, São Leopoldo, Brasil. danielczanotta@gmail.com
ORCID: https://orcid.org/0000-0003-2959-6525
5 Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil. mfreitas@ufrgs.br; atilio.grondona@ufrgs.br
ORCID: https://orcid.org/0000-0001-9879-2584 ; https://orcid.org/0000-0002-5455-5402

Recebido: 01.2023 | Aceito: 04.2023

Resumo: Informações referentes ao uso e cobertura da terra são essenciais para a realização de um planejamento ambiental eficaz. O uso de imagens de sensoriamento remoto facilita a elaboração desses mapeamentos, pela grande disponibilidade de imagens e avanços computacionais no processamento dos dados. Portanto, o objetivo deste trabalho é comparar e avaliar a potencialidade da técnica de Análise de Componentes Principais para o aprimoramento da acurácia da classificação do uso e cobertura da terra da bacia hidrográfica Lajeado São José. Para isso, a técnica de Análise de Componentes Principais foi aplicada na imagem Sentinel-2 no Google Earth Engine. E após, na mesma plataforma, foi realizada a classificação de imagens utilizando as bandas espectrais originais e as componentes principais geradas, a partir de dois classificadores distintos: CART e Random Forest. A fim de comparar se há algum incremento na qualidade da classificação ao usar as componentes principais em relação às bandas espectrais. A avaliação de acurácia foi produzida por meio da análise de pontos de controle distribuídos de forma estratificada. Os resultados indicaram que a melhor classificação de uso e cobertura da terra foi produzida com as componentes principais combinada ao classificador Random Forest, pois apresentou 87,7% de acurácia global, e menor discordância de quantidade (5,2%) e alocação (7,1%) na avaliação da acurácia. Portanto, pode-se concluir a partir da análise de acurácia que o uso de componentes principais pode gerar melhores resultados na classificação de uso e cobertura da terra quando comparados ao uso das bandas espectrais.
Palavras-chave: Sentinel-2, Sensoriamento Remoto, Google Earth Engine.

Abstract: Information regarding land use and cover is essential for effective environmental planning. The use of remote sensing images facilitates the elaboration of these mappings, due to the large availability of images and computational advances in the data processing. Therefore, this work aims to compare and evaluate the potential of the Principal Component Analysis technique to improve the accuracy of land use and cover classification in the Lajeado São José watershed. To this end, the Principal Component Analysis technique was applied to the Sentinel-2 image in Google Earth Engine. Then, in the same platform, the image classification was performed using the original spectral bands and the principal components generated, with two different classifiers: CART and Random Forest. In order to compare whether there is any increase in classification quality when using principal components in relation to spectral bands. The accuracy evaluation was produced through the analysis of stratified distributed control points. The results indicated that the best land use and cover classification was produced with the principal components combined with the Random Forest classifier, because it presented 87.7% of overall accuracy, and lower discordance of quantity (5.2%) and allocation (7.1%) in the accuracy evaluation. Therefore, it can be concluded from the accuracy analysis that the use of principal components can generate better results in land use and land cover classification when compared to the use of spectral bands.
Keywords: Sentinel-2, Remote Sensing, Google Earth Engine.

1 INTRODUÇÃO

A expansão das áreas urbanas sem planejamento adequado implica em impactos negativos ao ambiente, de modo que o monitoramento do espaço deve ser considerado no planejamento. A classificação do uso e cobertura da terra se apresenta como ferramenta de baixo custo para o conhecimento das alterações na paisagem e nortear tomadas de decisão (SANTOS et al., 2018). Em termos de planejamento ambiental e de gerenciamento de recursos naturais a informação referente ao uso e cobertura da terra, tanto atual como histórica, é essencial para a execução de estudos fidedignos que resultem em ações eficazes. Pois, a disposição espaço-temporal dos diversos usos e coberturas revelam a dinâmica e a possível pressão humana sob a área de interesse (SANTOS, 2004).

O sensoriamento remoto é uma ferramenta muito utilizada para estudos de recursos naturais que têm atingido variado grau de sofisticação. Com relação aos estudos sobre a superfície terrestre, a ampla cobertura espacial e repetitividade das observações dos satélites permitem a capacidade de cobertura de grandes áreas em intervalos temporais constantes (GAIDA et al., 2020). O mapeamento do uso e cobertura da terra pode ser realizado de maneira automatizada através de algoritmos classificadores que possibilitam a identificação de padrões e objetos homogêneos, rotulando os *pixels* da imagem e os associando a uma classe categórica (MATHER; TSO, 2009). Desta forma, a classificação semi-automática de imagens de sensoriamento remoto possibilita a produção de mapeamentos de uso e cobertura da terra de maneira eficiente em múltiplas escalas cartográficas. Tornando essa valiosa informação disponível mais facilmente, um exemplo de aplicação é o projeto MapBiomas e o mapeamento em tempo real no *Google Earth Engine* (GEE) (SOUZA, 2020; MAPBIOMAS, 2022; BROWN et al., 2022).

Partindo do exposto ressalta-se a importância da execução de mapeamentos de uso e cobertura da terra de qualidade e acurados. visam a melhoria da classificação de imagens como em Kobayashi et al. (2020) e Ferreira et al. (2007). Podendo se destacar o uso de índices espectrais (NDVI, EVI, SAVI, NDWI, etc.) (ZANOTTA; FERREIRA; ZORTEA, 2019), de imagens fração provenientes da técnica de mistura espectral, a aplicação de classificadores robustos como o *Random Forest* e de aprendizado profundo, como as redes neurais convolucionais de arquitetura U-Net (MAPBIOMAS, 2022).

Outra técnica que pode ser aplicada em imagens de sensoriamento remoto é a Análise de Componentes Principais (ACP). Esta abordagem parte do princípio de que as bandas espectrais utilizadas estão correlacionadas entre si, havendo redundância de dados. Assim, a ACP condensa a informação disponível em um número menor de canais espectrais, totalmente descorrelacionados, realçando os alvos na imagem e melhorando o resultado da classificação. Isso ocorre, pois, a utilização de dados não redundantes é de muito interesse no processo de classificação de imagens (MATHER; TSO, 2009; ZANOTTA; FERREIRA; ZORTEA, 2019).

A ACP alcança bons resultados para reduzir dados multiespectrais e hiperespectrais. Dessa forma, melhores resultados na classificação de imagens são obtidos utilizando as componentes principais, pois, nelas é possível diferenciar objetos na cena que antes eram inseparáveis (DHARANI; SREENIVASULU, 2021). Segundo Gupta et al. (2013) a aplicação da ACP no campo do sensoriamento remoto é vantajosa devido a sua capacidade de identificar padrões nos dados, destacando as diferenças e similaridades. A técnica resulta em baixa perda de informação, com concentração das informações mais relevantes nas primeiras componentes, permitindo diminuir a demanda e tempo computacional pela redução do número de bandas e pela remoção da correlação.

Classificadores baseados em árvores de decisão são amplamente utilizados no campo do Sensoriamento Remoro. Por conta disso, os classificadores empregados neste trabalho foram o CART e o *Random Forest*, *que* se baseiam nos valores de reflectância espectral de cada classe fornecida no treinamento para a diferenciação dos objetos no terreno. O treinamento dos classificadores e o resultado da classificação é realizado na escala do *pixel*, não sendo empregado qualquer tipo de segmentação pré-classificação ou de filtros pós-classificação. A natureza destes classificadores é a de árvores de decisão, como já mencionado, sendo este método baseado na segmentação de subconjuntos de regressão linear para a determinação do rótulo de cada classe com base nos valores espectrais das amostras de treinamento. Enquanto o classificador CART utiliza de

apenas uma árvore de decisão; o classificador *Random Forest* utiliza de múltiplas árvores de decisão, sendo o seu resultado derivado da concordância da maioria das árvores de decisão (LEG - UFPR, 2022).

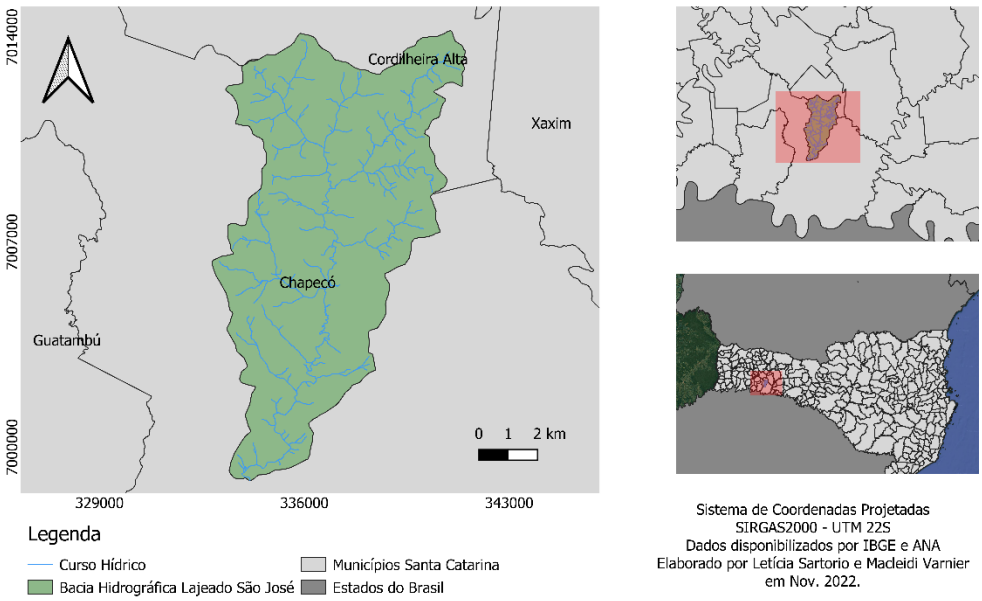
Portanto, o objetivo deste trabalho é comparar e avaliar a potencialidade da técnica de Análise de Componentes Principais (ACP) para o aprimoramento da acurácia da classificação do uso e cobertura da terra da bacia hidrográfica Lajeado São José, com uma abordagem *pixel-a-pixel*. Para isso, serão comparados, por meio métricas de acurácia, o resultado da classificação a partir das bandas espectrais e o resultado da classificação que utilizou as componentes principais. Visando, assim, contribuir para a melhoria dos mapeamentos temáticos baseados em imagens de sensoriamento remoto.

2 MATERIAIS E MÉTODOS

2.1 Área de estudo

A bacia hidrográfica do Lajeado São José (76,32 km²) (Figura 1) está localizada no Oeste de Santa Catarina. Pertence ao comitê de bacias do rio Chapecó e Irani, as quais estão inseridas na Região Hidrográfica do Uruguai (PANDOLFO, 2002). A bacia é responsável pelo abastecimento público dos municípios de Chapecó e Cordilheira Alta, além de servir para dessedentação animal e atendimento às agroindústrias inseridas na bacia. Estas, agentes da expansão urbana em áreas periféricas e próximas aos cursos d'água, contribuem para o aumento da impermeabilização do solo, redução da vegetação e deterioração da qualidade da água (OTSUSCHI, 2017). A Figura 2 apresenta uma imagem Sentinel-2 da área de estudo em uma composição de cor natural de abril de 2022. Ao observá-la podemos notar os usos e coberturas da terra que predominam na bacia, como áreas voltadas para a prática pecuária e agrícola, seguidas por áreas urbanas. Ademais, também há a presença significativa de solo exposto associados a preparação da terra para o cultivo e de locais com remanescentes florestais, compondo a mata ciliar dos cursos hídricos da bacia. Por fim, destaca-se a presença limitada de corpos hídricos.

Figura 1 – Área de Estudo.



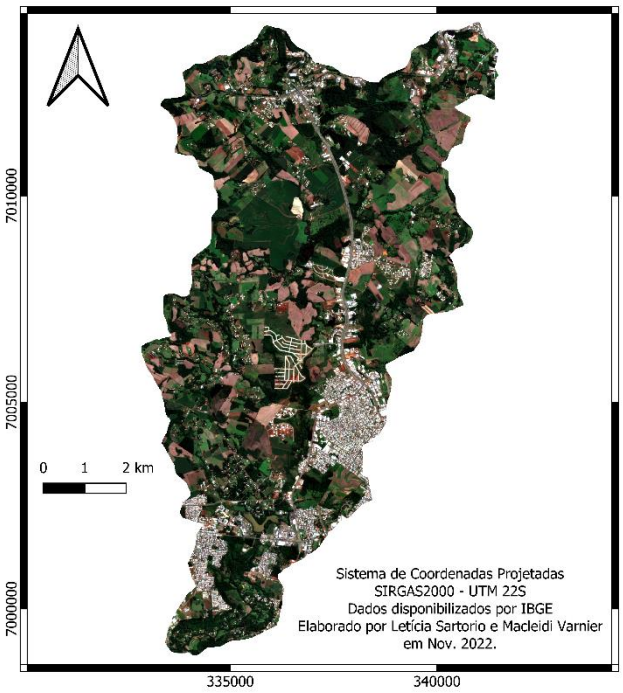
Elaboração: Autores (2022).

2.2 Materiais

As produções cartográficas foram realizadas pelo *software* QGIS 3.16.15, onde foram utilizados os *shapefiles* de malha municipal, obtido no site do IBGE e de cursos hídricos da ANA, além do modelo digital de elevação da Epagri Ciram para delimitação da bacia hidrográfica. A ACP e a classificação do uso e cobertura

da terra foi performada na plataforma *Google Earth Engine* utilizando-se como dado de entrada a imagem do satélite *Sentinel-2* sensor MSI com nível 2A de processamento, com correção atmosférica. A imagem selecionada é do dia 2 de abril de 2022, possui resolução espacial de 10 a 60 metros nas suas 12 bandas, e baixa cobertura de nuvens. A escolha de uma única data ocorreu, pois, se preferiu dar maior atenção ao uso de duas abordagens, com dois classificadores. E como a dinâmica do uso e cobertura da terra conhecida da área de estudo é baixa as classificações seriam extremamente similares, com modificações causadas pelo calendário agrícola. Neste estudo foram selecionadas as bandas espectrais com 10 metros de resolução espacial, sendo: Banda 2 do Azul (496.6nm), Banda 3 do Verde (560nm), Banda 4 do Vermelho (664.5nm) e Banda 8 do Infravermelho (835.1nm). Mais características do sensor podem ser encontradas em ESA (2015). Uma composição natural da imagem está apresentada na Figura 2. A opção pelo uso de somente das bandas espectrais com resolução espacial de 10 metros se deu com objetivo de verificar o potencial da ACP de auxiliar na classificação em situações com baixa resolução espectral, buscando utilizar as bandas disponíveis da melhor forma possível com remoção de ruídos. Ademais, a área de estudo escolhida exigia maior detalhes em termos de resolução espacial.

Figura 2 – Imagem Sentinel-2 selecionada



Elaboração: Autores (2022).

2.3 Caracterização e Análise de Componentes Principais (ACP)

Para compreender o relacionamento entre as bandas espectrais foi calculado o índice de correlação entre as bandas selecionadas no GEE, com a função ee.Reducer.pearsonsCorrelation(). Tal coeficiente indica o grau de relacionamento entre duas variáveis, medindo a dependência, positiva ou negativa entre as bandas, variando do valor -1 a 1. Uma correlação positiva aponta que o aumento dos valores em uma variável causa o aumento na outra, já uma correlação negativa apresenta o comportamento contrário, o aumento em uma variável causa a redução da outra. Portanto, em bandas espectrais a correlação ocorre quando a partir de valores digitais de uma banda pode-se deduzir os valores dos *pixels* correspondentes em outra banda, havendo, assim, redundância de dados. A existência de correlação entre bandas de imagens de satélite dificulta a identificação de pequenas diferenças em termos de reflectância dos materiais presentes na superfície terrestre, tornando mais complicado o processo de classificação de imagens, pela dificuldade de separação dos alvos (CROSTA, 1999; ZANOTTA; FERREIRA; ZORTEA, 2019). Essa caracterização inicial da imagem permite entender como estão correlacionados os valores das bandas e indicam a sua qualificação para a aplicação da técnica de ACP.

A ACP é uma técnica estatística multivariada utilizada para retirar a redundância existente entre os dados. Para isso produz as Componentes Principais (CP) que resultam da combinação linear dos dados originais. As novas CP geradas não são correlacionadas entre si e cada uma explica uma porcentagem da variância dos dados originais. Na aplicação da ACP utilizou-se a matriz de covariância (ACP não padronizada) entre os dados de entrada para determinar os autovalores e autovetores. A ACP foi aplicada na imagem selecionada na plataforma GEE utilizando uma adaptação do código disponibilizado por GEE (2022). A primeira etapa para a produção das CP foi a centralização pela média da imagem de entrada visando facilitar o cálculo da matriz de covariância, para isso a imagem foi subtraída pela sua média. Em seguida foi calculado a matriz de covariância centralizada (ee.Reducer.centeredCovariance()), a partir dessa foram encontrados os autovalores e autovetores pela função .eigen(), e as CP foram geradas por combinação linear dos dados originais com os autovetores. O autovalor indica a porcentagem de variância explicada por cada componente e o comprimento do eixo. O autovetor aponta a direção do eixo da componente, o de máxima variação dos dados. Ademais, os autovetores são utilizados como valores de ponderação que determinam a contribuição de cada banda espectral na composição de componente, formando uma combinação linear aditiva (CROSTA, 1999; GUPTA et al., 2013; GROTH et al., 2013; ZANOTTA; FERREIRA; ZORTEA, 2019; DHARANI; SREENIVASULU, 2021).

O resultado da CP foi dividido pelo seu desvio-padrão para aumentar o contraste e melhorar a visualização. Após a geração das CP, avaliou-se a porcentagem de variância explicada por cada PC para a seleção das componentes. Os autovetores foram utilizados para indicar a contribuição de cada banda na composição das CP. Por fim, as CP foram analisadas visualmente e uma composição colorida RGB com as três primeiras foi construída. Pois, as três componentes selecionadas serão utilizadas para o processo de classificação de imagem.

2.4 Classificação de Imagens e Validação

A classificação de uso e cobertura da terra foi realizada na plataforma GEE, de maneira supervisionada e com uma abordagem *pixel-a-pixel.* Para realizar a coleta de amostras definiu-se a partir da análise visual da imagem (Figura 2) as classes de uso e cobertura da terra de interesse, de acordo com a predominância e ocorrência na área de estudo. Ao total foram escolhidas cinco classes: Urbano, Água, Floresta, Pastagem, e Solo Exposto (que também pode abranger áreas agrícolas sendo preparadas para o plantio). Estas se relacionam com o nível hierárquico de subclasse (Nível II) do Manual Técnico do uso da terra (IBGE, 2013), com adaptação na nomenclatura. A escala da classificação final será de 1:95000, se enquadrando como um mapeamento de reconhecimento, indicado para distintos planejamentos, como o de bacias hidrográficas (IBGE, 2013) Em seguida foram coletadas amostras representativas destas classes na imagem base, utilizando uma composição cor natural (Figura 2). Para a coleta das amostras seguiu-se dois parâmetros básicos: os polígonos foram coletados em toda a área de estudo, sendo uma amostragem bem distribuída na bacia; e cada amostra cobria pequenas áreas onde a resposta espectral do terreno fosse bem representativa da classe de interesse, reduzindo possíveis confusões indesejadas. No total foram coletadas 726 amostras, sendo 158 no espaço urbano, 114 em Florestas, 27 em corpos d'Água, 141 em Solo Exposto e 286 em Pastagens.

Estes polígonos identificados foram utilizados no processo de classificação de imagens. Sendo que as amostras foram particionadas em treinamento do algoritmo de classificação (70%) e no teste da acurácia do classificador (30%). A próxima etapa foi a realização da classificação de imagens, para isso utilizou-se dois classificadores distintos, o CART e *Random Forest.* Estes classificadores foram escolhidos pois apresentam natureza similar com uso de árvores de decisão, no qual o CART faz uso de uma única árvore e o *Random Forest* de diversas, permitindo comparar um classificador mais simples com um mais complexo. Tais classificadores foram aplicados nas bandas espectrais e nas componentes principais, totalizando quatro mapas de uso e cobertura da terra. Utilizamos de dois algoritmos de classificação para garantir que os resultados de avaliação da acurácia dos produtos não seriam influenciados por um classificador em específico, mas representassem possíveis diferenças na qualidade da classificação em decorrência da composição de entrada utilizada.

O algoritmo CART é um tipo de classificador baseado em árvore de decisão e regressão. Breiman et al. (1984) foram os primeiros a introduzir as principais ideias da metodologia da árvore na estatística. O algoritmo CART ("*Classification and Regression Trees*") de Breiman et al. (1984) seguiu o algoritmo AID e é um dos algoritmos de árvore de decisão mais populares. A partir das amostras fornecidas inicialmente e utilizando de análise de regressão, o classificador divide as amostras em vários subconjuntos com base nos seus valores. Em nosso caso, as amostras das classes introduzidas como dados de entrada são subdivididas a partir do seu valor de resposta espectral. Além disso, as árvores de decisão têm a vantagem de raramente selecionar variáveis preditoras irrelevantes — ou seja, o algoritmo de construção de árvore recursiva estima a melhor variável para dividir em cada etapa, implicando que preditores não relacionados à(s) variável(is) de resposta tendem a não ser selecionados para divisão (BREIMAN et al. 1984; ELITH et al. 2008; YOUNG 2007). Para se predizer a classe de um *pixel* desconhecido é analisada a média do subconjunto onde o mesmo está presente dentro da análise de regressão, o caracterizando a partir do valor da sua resposta espectral em relação ao conjunto de dados de entrada. O valor das amostras fornecidas inicialmente serve como um treinamento para o algoritmo. Cada nó da árvore de regressão representa um subconjunto de dados particionado, onde os *pixels* a serem preditos são definidos com base no valor médio do subconjunto ao qual fazem parte (LEG - UFPR, 2022). Os parâmetros utilizados para o classificador CART foram o padrão especificado pelo GEE, considerando um número ilimitado de folhas na árvore de decisão e criando conjuntos de treinamento que contenham ao menos um ponto de amostra.

Por sua vez, o algoritmo de classificação *Random Forest* é baseado em múltiplas árvores de decisão. É um método de *Machine Learning* não paramétrico que usa um grupo de árvores de decisão para estimar um valor ou atribuir um objeto a uma classe (BREIMAN, 2001). Um método não paramétrico é aquele que não pressupõe a distribuição do espaço nem a estrutura do classificador. No caso de uma classificação de uso e cobertura da terra, um *pixel* desconhecido tem sua classe atribuída a partir do resultado de várias árvores de regressão, sendo o seu rótulo a classe que mais vezes se repetiu nestas várias árvores. Este método tende a produzir resultados mais acurados devido ao modelo rodar em muitas árvores, o que permite minimizar erros que poderiam ser mais comuns em um método que se utilize de apenas uma. Os parâmetros utilizados para o classificador *Random Forest* foram 500 árvores de decisão, como sugerido por Belgiu e Drăguţ (2016), a criação de conjuntos de treinamento que possuam ao menos uma amostra, um número ilimitado de folhas em cada árvore e 123 sementes de distribuição aleatória.

A próxima etapa foi a avaliação da acurácia para verificar se o uso de componentes principais melhora o resultado da classificação de imagens. Para se avaliar o quão bem uma classificação representa a realidade é necessário realizar uma análise de concordância entre a classificação e o que é encontrado na área de interesse. Para isso, é elaborada uma amostragem de pontos em que se compara *pixel* a *pixel* qual é a classe definida pela classificação e qual é a encontrada na realidade. Algumas boas práticas precisam ser tomadas para garantir a integridade da avaliação, dentre elas destacamos que: a avaliação de acurácia precisa ser realizada após a classificação, a partir do produto pronto; os pontos utilizados para esta análise precisam ser independentes dos que foram utilizados como amostras para o classificador; os pontos precisam ser distribuídos aleatoriamente e não de modo conveniente para o avaliador; todas as classes do mapeamento precisam ser representadas; e a amostragem para a avaliação de acurácia precisa ser representativa da área de interesse (FOODY, 2022).

Neste estudo optamos por uma abordagem de amostragem aleatória estratificada, em que os pontos utilizados para avaliar a acurácia da classificação são distribuídos de forma aleatória e proporcionalmente ao tamanho da área de cada classe no mapa. A classificação com *Random Forest* e composição RGB + NIR foi utilizada como a base para a amostragem dos pontos de controle. A utilização de uma estratégia de amostragem aleatória estratificada é conhecida e bastante utilizada pela comunidade de Sensoriamento Remoto (OLOFSSON et al, 2014). Esta abordagem permite a avaliação da acurácia para cada classe mapeada com a garantia de que todas as classes possuam pontos de controle distribuídos de forma proporcional ao tamanho da área de estudo. Como neste artigo realizamos a elaboração de quatro classificações, os mesmos pontos de controle foram utilizados para avaliar todas as classificações. Optou-se por essa metodologia visando reduzir possíveis influências na avaliação de acurácia dos produtos devido a configuração de amostragem, evitando que o acaso pudesse favorecer alguma das classificações. Além disso, como nosso objetivo é avaliar possíveis ganhos da classificação utilizando da técnica de ACP, a amostragem para avaliação da acurácia foi realizada

com base em uma classificação que não utiliza desta técnica como dados de entrada, visando evitar que a amostragem favorecesse as classificações que utilizam desta técnica.

Utilizamos uma imagem de alta resolução espacial disponibilizada pelo *Google Earth*, da mesma data que as utilizadas nas classificações, para avaliar a acurácia nos pontos de controle. Os parâmetros de entrada utilizados para o método de amostragem estão apresentados na Tabela 3. Sendo eles, a quantidade de *pixels* contidos em cada classe da classificação (N) e o seu peso proporcional em relação ao todo (Wi). A acurácia do usuário presumida por classe (Ui) e o erro padrão geral estipulado para a acurácia (S(Ô)). Estão disponíveis também o erro padrão presumido por classe (Si), o erro padrão presumido proporcional para cada classe em relação ao todo (WiSi) e o erro padrão presumido proporcional para cada classe em relação ao todo ao quadrado ($WiSi^2$). A partir desses parâmetros, a equação fornece informações sobre a quantidade mínima de pontos para se avaliar a acurácia dos resultados de maneira satisfatória (NA), considerando o intervalo de confiança definido no método. Cabe destacar que a amostragem para a avaliação de acurácia não é uma ciência exata, existindo variáveis que devem ser preenchidas com valores presumidos pelo avaliador. Neste sentido, o número de amostras resultado do método é um ponto de partida, tendo em conta que o conhecimento da área de estudo e a confiança no método de classificação são parâmetros importantes (STEHMAN; CZAPLEWSKI, 1998; FOODY, 2002).

Analisando a classificação, temos que a classe de Água é pouco abundante em nossa área de estudo. E o resultado da amostragem lhe dava apenas 7 pontos de controle devido a sua área reduzida. Para a melhor representação desta classe e visando garantir uma quantidade mínima aceitável de pontos em todas as classes da classificação, foram arbitrariamente adicionadas 50 amostras a mais que as calculadas pela equação na classe Água.

Para os parâmetros escolhidos foram necessários avaliar 813 pontos e validar o resultado da classificação com a imagem de alta resolução espacial disponibilizada pelo *Google Earth*. A análise das amostras ocorreu de forma visual, interpretando qual a classe de uso e cobertura da terra se aplicava para os pontos amostrados, identificando quais das cinco classes — Urbano, Água, Floresta, Pastagem, e Solo Exposto — estavam sendo representadas em cada um dos pontos. Com todos os pontos avaliados foi possível gerar uma série de análises referente a acurácia das classificações e compará-las. Foram geradas matrizes de confusão para cada classificação, visando analisar as concordâncias e discordâncias de classe para cada classificação. Tais matrizes permitem avaliar a qualidade do resultado da classificação a partir da comparação da classificação com a coleção de referência amostrada. Ao final, a avaliação de acurácia permitiu analisar a diferença na qualidade do mapa ocasionada pelo dado de entrada da classificação (PONTIUS; MILLONES, 2011).

3 RESULTADOS E DISCUSSÕES

Os resultados, assim como a metodologia, estão apresentados em dois tópicos.

3.1 Coeficiente de Correlação e Análise de Componentes Principais

O resultado do coeficiente de correlação entre as quatro bandas espectrais selecionadas da imagem *Sentinel-2* está apresentado na Tabela 1.

Tabela 1. Coeficiente de Correlação entre as Bandas Espectrais.

Bandas Espectrais	B2	B3	B4	B8
B2	1	0,973	0,923	-0,159
B3	0,973	1	0,925	-0,05
B4	0,923	0,925	1	-0,313
B5	-0,159	-0,05	-0,313	1

Elaboração: Os autores (2022).

A diagonal principal apresenta a correlação entre as mesmas bandas, e consequentemente o coeficiente é igual a 1. A parte superior da tabela é espelhada a inferior, pois resulta da mesma combinação de bandas.

Com os valores encontrados pode-se analisar a correlação entre as bandas espectrais e seu grau de relacionamento. Nota-se que as bandas do visível (B2, B3 e B4) apresentam um alto grau de correlação positiva entre si, com todos os valores superiores a 0,9. Havendo assim um grau de redundância entre os dados, devido à alta correlação entre eles. Esse comportamento pode estar associado ao fato dessas bandas estarem posicionadas em faixas próximas do espectro eletromagnético e pela presença de alvos com assinaturas espectrais similares na área de estudo. O oposto ocorre com os coeficientes de correlação entre a banda do infravermelho e as bandas do visível. O coeficiente de correlação é negativo para todos, indicando uma relação inversa, e apresenta valores mais baixos, o que aponta para uma correlação fraca. A correlação entre a B8 (infravermelho próximo) e a B3 (verde) é extremamente baixa, próxima do valor zero. Com a banda do vermelho (B4) a correlação é um pouco maior (-0,313) e com a azul menor (-0,159). Meneses e Almeida (2012) expõem que diversos alvos apresentam valores de reflectância parecidos na faixa do comprimento de onda do visível, gerando bandas espectrais semelhantes e correlacionadas. Os valores de correlação aqui encontrados demonstram isso. Portanto, a partir da Tabela 2 evidencia-se o alto grau de correlação entre as bandas espectrais, principalmente da faixa do visível, que produz redundância entre as informações que pode dificultar o processo de classificação de imagens.

A técnica ACP foi aplicada sobre as bandas selecionadas no GEE, sendo produzidas quatro componentes principais. A Tabela 2Tabela 2 apresenta a porcentagem de variância original explicada para cada componente principal e os seus autovetores correspondentes.

Tabela 2 - Autovetores e Porcentagem de Variância Explicada para as Componentes Principais.

AutoVetores	B2	B3	B4	B8	**Variância Explicada (%)**
CP 1	-0,14359	-0,111642	-0,277	0,943497	65,63
CP 2	0,480349	0,533376	0,619366	0,318058	33,36
CP 3	0,598856	0,343203	-0,71922	-0,079406	0,82
CP 4	0,624514	-0,765021	0,149592	0,04844	0,19

Elaboração: Os autores (2022).

A Figura 3 apresenta as quatro componentes principais em tons de cinza. A primeira componente principal detém a maior porcentagem de variância explicada (65,63%), seguindo pela segunda que possui 33,36%, ao total às duas somam 98,99% da variância total do banco de dados. Por outro, a terceira componente explica 0,82% e a última 0,19%, sendo que a última está provavelmente relacionada com os ruídos existentes na imagem original. Assim, quando observados os autovetores apresentados na Tabela 2 notamos que a 1º CP apresenta maior contribuição positiva da banda do infravermelho próximo (B8), e as demais bandas contribuem negativamente em menor proporção. Ao traçar um paralelo entre a Figura 3 e os autovetores desta CP, nota-se que o alto valor da B8 está associado a grande presença de vegetação na área, como pastagens e formações florestais, sendo que estes alvos são destacados com alto valor de brilho na CP 1. Tal comportamento resulta da combinação do valor alto e positivo do autovetor para a B8 com altos valores de reflexão nesta banda para alvos vegetados. Percebe-se que a contribuição da B8 reduz nas demais componentes, pois, a mesma já teve uma contribuição significativa na primeira. A 2º CP apresenta maior contribuição da banda do vermelho (B4), e todos os valores são positivos. Esta CP realça as áreas urbanas da cena, e como possui maior contribuição da banda do vermelho não destaca tanto a vegetação, que possui resposta espectral baixa nesta faixa, e a menor contribuição na construção da 2º CP é da B8. Portanto, as áreas urbanas nesta CP apresentam alto valor de brilho por conta da combinação de alta reflexão desse alvo na B4 com alto valor positivo do autovetor para a B4 nesta CP (GUPTA et al., 2013).

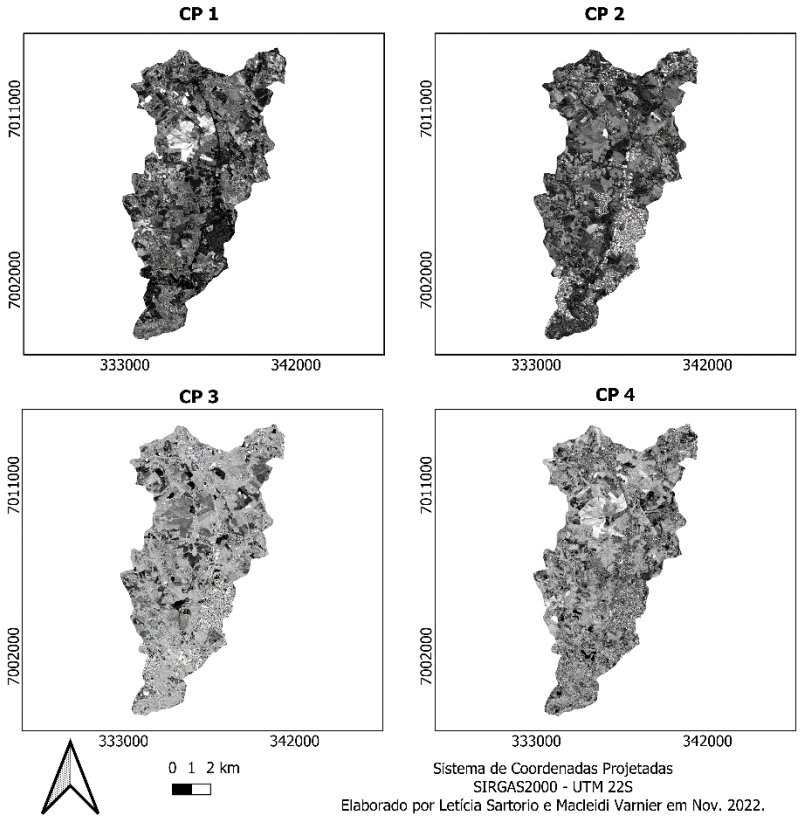
Na CP 3 a maior contribuição foi negativa e da banda do vermelho, seguida pela contribuição positiva da banda do azul. Ao visualizar essa CP na Figura 3 notamos um maior destaque nas áreas com solo exposto, justificando a contribuição praticamente nula da B8 e maior contribuição da B4. Os tons escuros que destacam o solo exposto são gerados pela combinação do valor negativo do autovetor para a B4 com valores altos de reflexão para esses alvos na B4. Por fim, a 4º CP possui contribuição negativa significativa da banda do verde, e com os demais valores positivos. Esta CP destaca áreas vegetadas e corpos hídricos, mas está acompanhada de muito ruídos. As áreas vegetadas apresentam alto valor de brilho nesta CP, por possuir um alto valor negativo de autovetor para banda do verde associada a valores baixos de reflexão nesta faixa para esse alvo.

Já, os tons escuros dos corpos hídricos ressaltados nesta CP podem estar associados ao alto valor positivo do autovetor para a B2 combinado ao baixo valor de reflexão desse alvo nessa banda. Assim, a compreensão da contribuição das bandas na construção das componentes auxilia na sua interpretação (GUPTA et al., 2013).

Normalmente a primeira componente principal apresenta a informação partilhada entre as bandas espectrais originais, podendo ser entendida como brilho médio (MATHER; TSO, 2009). A CP 1 apresenta bastante contraste, facilitando a separação visual dos alvos. Isso decorre da maior porcentagem de variância retida por essa CP, resultando em mais informação distribuída em distintos níveis de cinza, tornando a imagem mais heterogênea. A segunda componente apresenta menos contraste que a anterior, mas ainda é possível distinguir os alvos, considerando que a mesma detém 33,36% da variância original. Além disso, nota-se que essa componente realçou áreas diferentes que a CP 1. Por outro lado, as duas últimas componentes (CP 3 e CP 4) apresentam um contraste menor devido à baixa quantidade de informação, variância baixa, resultante em componentes mais homogêneas. É importante considerar que mesmo a componente apresentando um baixo valor de variância explicada, isso não torna essa informação menos importante, pois, diferenças sutis podem ser extremamente úteis no processo de classificação de imagens (CROSTA, 1999; ZANOTTA; FERREIRA; ZORTEA, 2019). Entretanto, na quarta componente (CP 4), que representa 0,19% da variância original, há presença de muitos ruídos. Os mesmos podem ser visualizados na Figura 3. A ACP concentra os ruídos nas componentes finais, facilitando a sua remoção (CROSTA, 1999). Ressalta-se que A CP 4 apresenta contribuição significativa da B4 e B2, e destaca de forma muito sutil as áreas vegetadas e os corpos hídricos presentes na cena. Ademais, os mesmos já estão representados nas outras CP. Porém, a presença de ruído é muito alta nesta última componente principal, superando possíveis benefícios provenientes do seu uso.

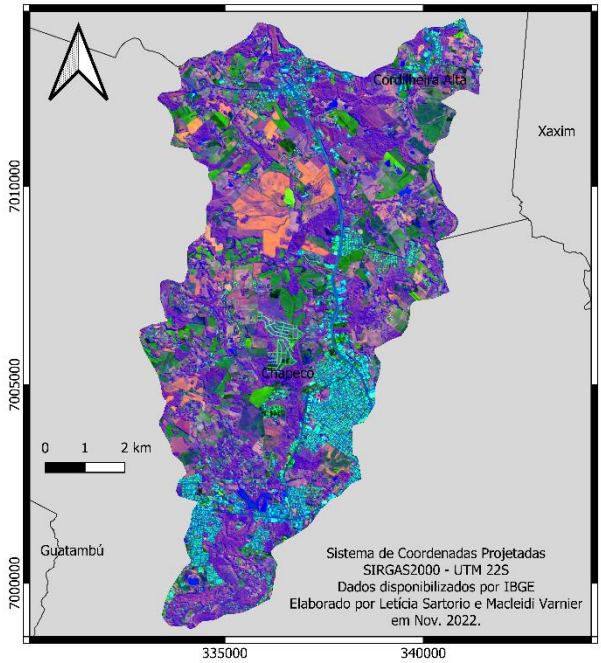
Uma composição colorida da combinação RGB da CP 1, CP 2 e CP 3, está presente na Figura 4. Tal composição apresenta grande contraste entre os alvos da imagem, os realçando e facilitando a sua distinção. Como as componentes não possuem correlação entre si, resultam em cores mais contrastantes que na imagem original (Figura 2) (MENESES; ALMEIDA, 2012). Essas três primeiras componentes foram escolhidas para o processo de classificação de imagens, sendo que a quarta componente foi descartada por representar os ruídos.

Figura 3 - Componentes Principais.



Elaboração: Os autores (2022).

Figura 4 - Composição colorida RGB entre as três primeiras componentes principais.



Elaboração: Os autores (2022).

3.2 Classificação e avaliação de acurácia

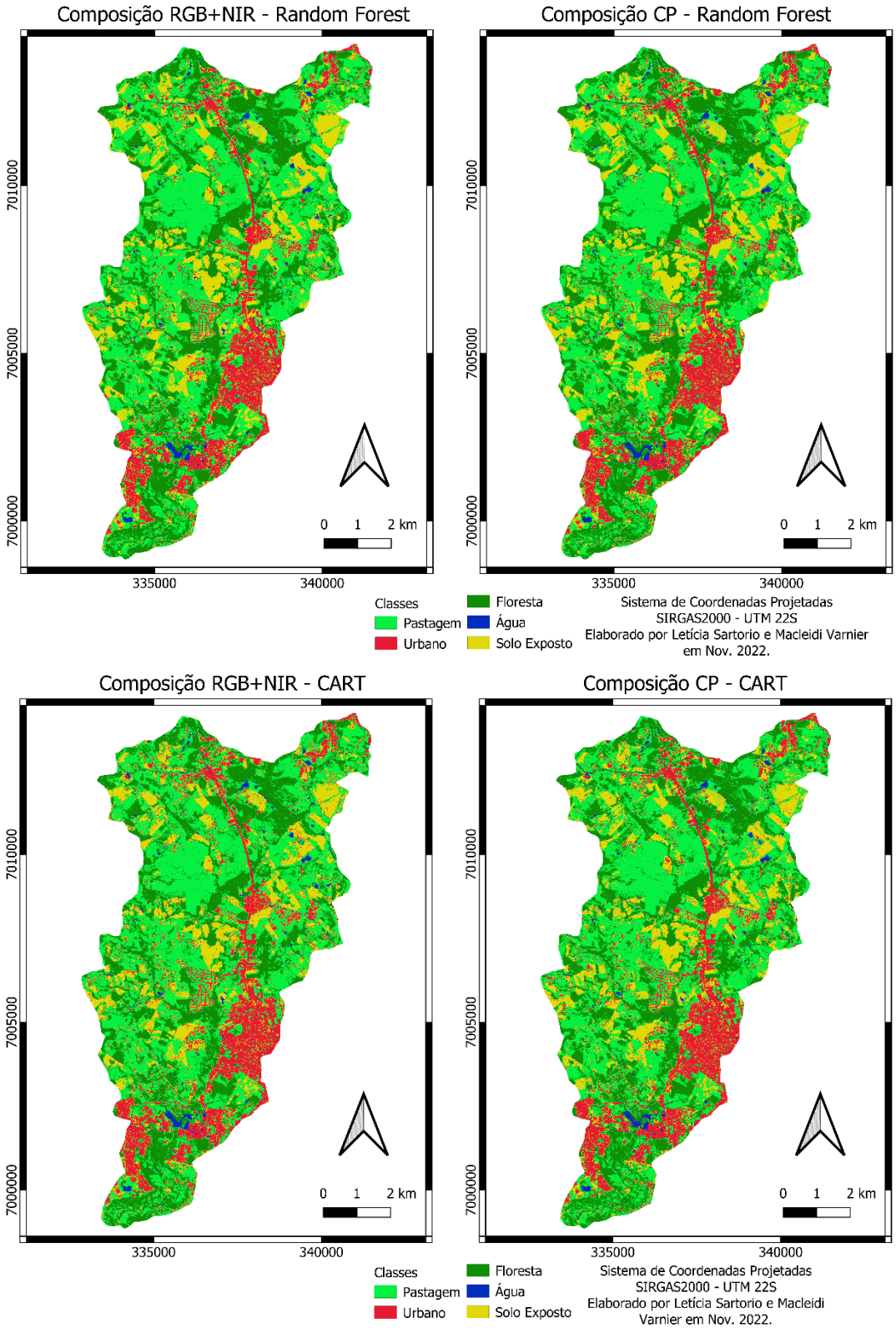
A Figura 5 apresenta os quatro mapas resultantes da classificação das imagens. Inicialmente, observando a bacia como um todo, nota-se que os mapas de uso e cobertura da terra são similares visualmente. Seguindo um padrão aproximado de distribuição espacial das classes elencadas.

Todavia, quando observamos com um grau de detalhamento maior (Figura 6) algumas diferenças ficam mais claramente visíveis. É notável que as classificações realizadas com o algoritmo CART resultaram em um produto final com um aspecto mais 'pixelado', caracterizado pelo efeito sal e pimenta. Neste caso, muitos *pixels* acabam sendo mal classificados ficando 'perdidos' entre outras classes. No caso das classificações com *Random Forest* este efeito fica sensivelmente reduzido, onde cada classe se apresenta com uma aparência mais homogênea, com menos confusões decorrentes do efeito sal e pimenta. Entretanto, algumas confusões grosseiras ainda estão presentes, como as áreas classificadas como Água dentro da área Urbana no sudeste das classificações. Possivelmente devido à presença de sombras na imagem. Estes erros podem ser explicados por problemas embutidos nas amostras de entrada, em decorrência principalmente da variabilidade da resposta espectral da água na região, e da consequente dificuldade em caracterizá-la e diferenciá-la de outros alvos. O que acaba por induzir o classificador a cometer erros ao confundir estas classes com outros objetos presentes na superfície. Pensando em melhorar esta classificação, outros dados de entrada como imagem de Radar ou um modelo digital de elevação poderiam ser fornecidos para o classificador. Porém, como neste caso o objetivo é comparar o desempenho de dois dados de entrada distintos (composição CP e RGB + NIR) não foram fornecidos ao classificador outras fontes de informação com a intenção de aprimorar as classificações. Cabe destacar que a acurácia do treinamento foi superior a 99% em todas as classificações, representando que dentro das áreas amostradas os classificadores tiveram um bom desempenho.

Contudo, uma mera análise visual dos resultados não possibilita quantificar a qualidade das classificações e nem as comparar. Por este motivo, nos propomos a realizar a avaliação da acurácia destes produtos. Para realizar esta análise, elaboramos um conjunto de dados de referência com 813 pontos expressando o tipo de uso e cobertura da terra encontrado nestes pontos. De posse destas informações, as mesmas foram comparadas com aquilo obtido nas classificações através das matrizes de confusão. As matrizes de confusão são muito utilizadas quando se objetiva comparar dois conjuntos de dados, possibilitando entender as concordâncias e discordâncias dos dados (FOODY, 2008).

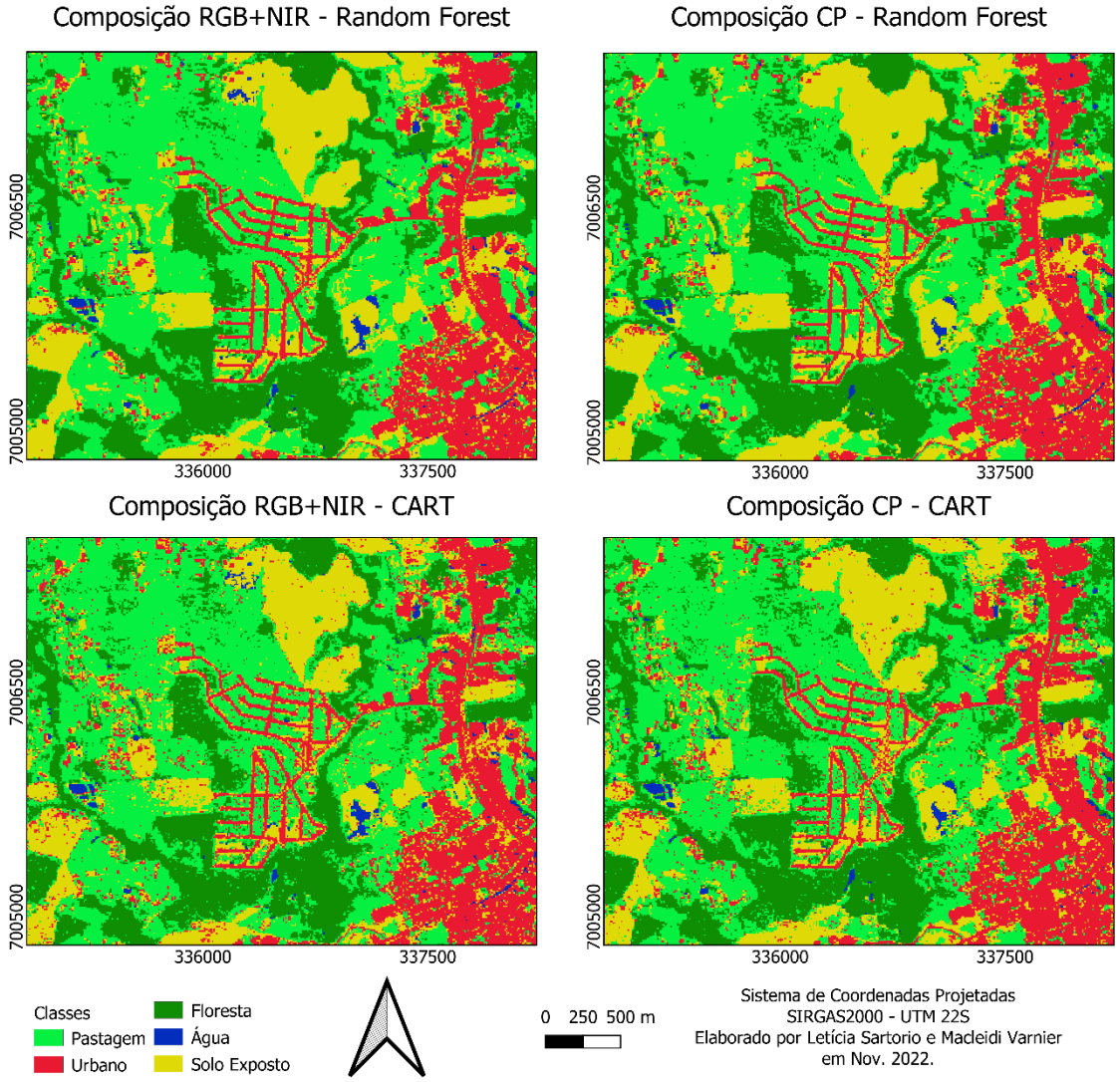
 DOI: http://dx.doi.org/10.14393/ rbcv75n0a-68010

Figura 5 - Classificações de uso e cobertura da terra da Bacia do rio Lajeado São José.



Elaboração: Os autores (2022).

Figura 6 - Classificações de uso e cobertura da terra da Bacia do rio Lajeado São José em recorte de detalhe.



Elaboração: Os autores (2022).

A Tabela 3 apresenta os parâmetros de entrada utilizados para o método de amostragem para a avaliação da acurácia.

Tabela 3 - Técnica de amostragem para avaliação de acurácia.

Classe	**N**	**Wi**	**Ui**	**Si**	**WiSi**	**WiSi2**	**NA**	**S WiSi:**	0,415
Urbano	103950	0,136	0,850	0,357	0,049	0,017	103	**S WiSi2:**	0,173
Floresta	186840	0,245	0,800	0,400	0,098	0,039	186	**S(Ô):**	0,015
Água	7300	0,010	0,650	0,477	0,005	0,002	57		
Solo Exposto	120850	0,159	0,750	0,433	0,069	0,030	121		
Pastagem	342650	0,450	0,750	0,433	0,195	0,084	344		
Total	761590	1			0,415	0,173	813		

Elaboração: Os autores (2022).

Para o caso de uma classificação de uso e cobertura da terra, são representadas na matriz de confusão as concordâncias e discordâncias entre a classificação de referência, realizada pela amostragem de pontos, e o mapa de uso e cobertura da terra, resultado da classificação. Na diagonal principal das matrizes apresentadas (destacada em verde claro) estão representadas as concordâncias, os acertos das classificações em relação ao conjunto de dados de referência. O que está acima da diagonal principal representa os erros de inclusão, sendo *pixels* classificados como sendo de uma classe, mas na realidade não são. E o que está abaixo da diagonal principal representa erros de omissão; *pixels* que na realidade fazem parte de uma classe, mas que não foram

mapeados desta forma. Também são representados os totais para as linhas e colunas, sendo uma informação importante para a definição da acurácia do usuário e produtor, respectivamente. A acurácia do produtor representa a qualidade do mapeamento de cada classe para quem elaborou o mapa, podendo ser uma informação útil pensando em melhorias. Enquanto a acurácia do usuário representa a qualidade do mapeamento de cada classe para quem utilizará estas informações, onde deve-se levar em conta cada objetivo em específico. A diferenciação dos erros de omissão e inclusão são importantes quando se pensa na elaboração de estimativas de áreas não tendenciosas, além de ser uma informação útil para quem utilizará o mapa. Também são oferecidas a acurácia geral do mapa, representando a qualidade do mapeamento como um todo; e a discordância geral, representando o total dos erros embutidos na classificação (OLOFSSON et al, 2014). A utilização do índice Kappa já se mostrou inadequada para a avaliação de acurácia de produtos de classificação de uso da terra derivados de dados de Sensoriamento Remoto. Por este motivo, optou-se por não utilizar o índice Kappa com base em Pontius e Millones (2011).

Pensando que toda a classificação de uso e cobertura da terra possui erros, a quantificação e comunicação destes erros ao público é uma etapa muito importante do trabalho. Considerando que o público pode possuir objetivos diferentes ao utilizar o mesmo mapeamento, a comunicação da acurácia por classes é de grande importância para potencializar os usos. Uma avaliação de acurácia derivada de uma classificação de referência estratificada por classes permite a estimativa de áreas não tendenciosas. Porém, como a quantificação de áreas não é nosso objetivo neste trabalho, esta etapa não será explorada (STEHMAN, 2013).

A Tabela 4 representa a matriz de confusão da classificação com o algoritmo *Random Forest* e a composição RGB + NIR. A acurácia geral desta classificação foi de 83,8%, porém percebemos que há uma variação considerável na acurácia das classes. Para o produtor, a melhor acurácia foi a da classe Pastagem, de 92%; enquanto que a menor acurácia foi a de Solo Exposto, com 75,2%. Já para o usuário, a melhor acurácia foi a da classe de Floresta com 95,9%; enquanto a menor foi a de Água, com 62,8%.

Tabela 4 - Matriz de confusão com contagens - Classificação Random Forest com composição RGB + NIR.

		Referência							
Classificação	Classes	Pastagem	Urbano	Floresta	Água	Solo exposto	Total	Acurácia do Usuário	Erros de Inclusão
	Pastagem	264	9	45	2	25	345	0,765	0,235
	Urbano	5	102	0	3	3	113	0,903	0,097
	Floresta	6	0	188	0	2	196	0,959	0,041
	Água	2	2	9	27	3	43	0,628	0,372
	Solo Exposto	10	5	1	0	100	116	0,862	0,138
	Total	287	118	243	32	133	813		
	Acurácia do Produtor	0,920	0,864	0,774	0,844	0,752		Acurácia Geral	0,838
	Erros de Omissão	0,080	0,136	0,226	0,156	0,248		Discordância Geral	0,162

Elaboração: Os autores (2022).

A Tabela 5 nos fornece uma outra visão da mesma matriz de confusão, onde os valores apresentados estão normalizados pelo total da tabela. Percebemos observando a tabela que a classe com mais erros de inclusão é a de Água, indicando que a estimativa de área desta classe está superestimada na classificação. Sendo que os maiores erros de omissão estão concentrados na classe de Floresta, indicando que esta foi a classe mais subestimada na classificação. A normalização pelo total da tabela também é importante, pois permite diferenciar justamente os erros de quantidade e alocação, sendo que para a estimativa de áreas não tendenciosas os erros de alocação não são incluídos na conta.

Tabela 5 - Matriz de confusão normalizada pelo total - Classificação Random Forest com composição RGB + NIR.

	Referência								
Classificação	Classes	Pastagem	Urbano	Floresta	Água	Solo Exposto	Total	Erros de Inclusão	Discordância de Quantidade
	Pastagem	0,325	0,011	0,055	0,002	0,031	0,424	0,235	0,071
	Urbano	0,006	0,125	0,000	0,004	0,004	0,139	0,097	0,006
	Floresta	0,007	0,000	0,231	0,000	0,002	0,241	0,041	0,058
	Água	0,002	0,002	0,011	0,033	0,004	0,053	0,372	0,014
	Solo Exposto	0,012	0,006	0,001	0,000	0,123	0,143	0,138	0,021
	Total	0,353	0,145	0,299	0,039	0,164	1,000	Discordância de quantidade	0,085
	Erros de Omissão	0,080	0,136	0,226	0,156	0,248		Discordância de alocação	0,077
	Discordância de Alocação	0,057	0,027	0,020	0,012	0,039		AG + DA	0,915

Elaboração: Os autores (2022).

Já para a classificação utilizando o algoritmo *Random Forest* e composição CP como dado de estrada, observamos na Tabela 6 que a acurácia geral foi de 87,7%. Para o produtor, a melhor acurácia foi a da classe de Pastagem, com 93,4%; enquanto que a menor acurácia foi a de Floresta, com 82,3%. Já para o usuário, a melhor acurácia foi a da classe de Floresta, com 97,1%; sendo que a menor foi a de Água, com 71,8%. Nota-se que no caso da classificação com CP a acurácia geral foi maior ao se comparar a classificação realizada com RGB + NIR, o mesmo também é observado para a acurácia da maioria das classes.

Tabela 6 - Matriz de confusão com contagens - Classificação Random Forest com composição CP.

	Referência								
Classificação	Classes	Pastagem	Urbano	Floresta	Água	Solo exposto	Total	Acurácia do Usuário	Erros de Inclusão
	Pastagem	268	5	35	0	14	322	0,832	0,168
	Urbano	5	104	0	4	5	118	0,881	0,119
	Floresta	5	0	200	0	1	206	0,971	0,029
	Água	2	3	6	28	0	39	0,718	0,282
	Solo Exposto	7	6	2	0	113	128	0,883	0,117
	Total	287	118	243	32	133	813		
	Acurácia do Produtor	0,934	0,881	0,823	0,875	0,850		Acurácia Geral	0,877
	Erros de Omissão	0,066	0,119	0,177	0,125	0,150		Discordância Geral	0,123

Elaboração: Os autores (2022).

Ao observarmos a matriz normalizada pelo total apresentada na Tabela 7, percebemos que a classe com mais erros de inclusão é a de Água, indicando que a estimativa de área desta classe está superestimada na classificação. Porém, ao compararmos esta classificação com a realizada com composição RGB + NIR, nota-se que em termos gerais houve uma redução nos erros de inclusão. No caso da classe Água, por exemplo, este erro passa de 37,2% na composição RGB + NIR, para 28,2% na composição CP. Já quanto aos erros de omissão, a maior concentração está localizada na classe de Floresta, indicando que esta foi a classe mais subestimada na classificação. Nota-se nos erros de omissão a mesma tendência observada nos erros de inclusão, havendo valores de erros menores no caso da classificação com composição CP quando comparada à realizada com composição RGB + NIR. Observamos que para o conjunto total dos dados, as discordâncias de quantidade e alocação foram menores na classificação com as CP do que na classificação com RGB + NIR.

Tabela 7 - Matriz de confusão normalizada pelo total - Classificação Random Forest com composição CP.

	Referência								
Classificação	Classes	Pastagem	Urbano	Floresta	Água	Solo Exposto	Total	Erros de Inclusão	Discordância de Quantidade
	Pastagem	0,330	0,006	0,043	0,000	0,017	0,396	0,168	0,043
	Urbano	0,006	0,128	0,000	0,005	0,006	0,145	0,119	0,000
	Floresta	0,006	0,000	0,246	0,000	0,001	0,253	0,029	0,046
	Água	0,002	0,004	0,007	0,034	0,000	0,048	0,282	0,009
	Solo Exposto	0,009	0,007	0,002	0,000	0,139	0,157	0,117	0,006
	Total	0,353	0,145	0,299	0,039	0,164	1,000	Discordância de quantidade	0,052
	Erros de Omissão	0,066	0,119	0,177	0,125	0,150		Discordância de alocação	0,071
	Discordância de Alocação	0,047	0,034	0,015	0,010	0,037		AG + DA	0,948

Elaboração: Os autores (2022).

Para a classificação com o algoritmo CART observamos resultados diferentes. Para além das diferenças visuais já descritas, percebemos que a acurácia deste classificador foi menor com ambos dados de entrada quando comparado ao *Random Forest*. Utilizando de uma composição RGB + NIR como dado de entrada, observamos na Tabela 8 que a acurácia geral desta classificação foi de 80,2%. Analisando a acurácia por classes, para o produtor a melhor classificação foi a da classe de Pastagem, com 91,3% de acurácia; enquanto que a menor acurácia foi a de Solo Exposto, com 66,2%. Já para o usuário, a melhor acurácia foi a da classe de Floresta com 96,6%; e a menor foi a de Água, com 63,6%.

Tabela 8 - Matriz de confusão com contagens - Classificação CART com composição RGB + NIR.

	Referência								
Classificação	Classes	Pastagem	Urbano	Floresta	Água	Solo exposto	Total	Acurácia do Usuário	Erros de Inclusão
	Pastagem	262	7	58	3	32	362	0,724	0,276
	Urbano	7	101	1	1	11	121	0,835	0,165
	Floresta	5	0	173	0	1	179	0,966	0,034
	Água	2	4	9	28	1	44	0,636	0,364
	Solo Exposto	11	6	2	0	88	107	0,822	0,178
	Total	287	118	243	32	133	813		
	Acurácia do Produtor	0,913	0,856	0,712	0,875	0,662		Acurácia Geral	0,802
	Erros de Omissão	0,087	0,144	0,288	0,125	0,338		Discordância Geral	0,198

Elaboração: Os autores (2022).

Observando a matriz de confusão normalizada pelo total na Tabela 9, notamos que a classe com mais erros de inclusão é a de Água, indicando que a estimativa de área desta classe está superestimada na classificação. Enquanto que os maiores erros de omissão estão concentrados na classe de Floresta, indicando que esta foi a classe mais subestimada na classificação. Nota-se que utilizando composição RGB + NIR e o classificador CART, a discordância total de quantidade e de alocação foi maior comparada a classificação elaborada com o mesmo dado de entrada utilizando o algoritmo *Random Fores*t.

Tabela 9 - Matriz de confusão normalizada pelo total - Classificação CART com composição RGB + NIR.

	Referência								
Classificação	Classes	Pastagem	Urbano	Floresta	Água	Solo Exposto	Total	Erros de Inclusão	Discordância de Quantidade
	Pastagem	0,322	0,009	0,071	0,004	0,039	0,445	0,276	0,092
	Urbano	0,009	0,124	0,001	0,001	0,014	0,149	0,165	0,004
	Floresta	0,006	0,000	0,213	0,000	0,001	0,220	0,034	0,079
	Água	0,002	0,005	0,011	0,034	0,001	0,054	0,364	0,015
	Solo Exposto	0,014	0,007	0,002	0,000	0,108	0,132	0,178	0,032
	Total	0,353	0,145	0,299	0,039	0,164	1,000	Discordância de quantidade	0,111
	Erros de Omissão	0,087	0,144	0,288	0,125	0,338		Discordância de alocação	0,087
	Discordância de Alocação	0,062	0,042	0,015	0,010	0,047		AG + DA	0,889

Elaboração: Os autores (2022).

No caso da classificação com CART e utilizando a composição CP como dado de entrada, observamos na Tabela 10 que a acurácia geral da classificação ficou em 81,8%. Para o produtor, a melhor acurácia foi a da classe Pastagem, com 90,6%; enquanto que a menor acurácia foi a de Floresta, com 75,7%. Já para o usuário, a melhor acurácia foi a da classe de Floresta, com 94,8%; e a menor foi a de Água, com 65%. Nota-se que no caso da classificação com as CP a acurácia geral foi maior ao se comparar com a classificação realizada com RGB + NIR, o mesmo também é observado para a acurácia da maioria das classes.

Tabela 10 - Matriz de confusão com contagens - Classificação CART com composição CP.

	Referência								
Classificação	Classes	Pastagem	Urbano	Floresta	Água	Solo exposto	Total	Acurácia do Usuário	Erros de Inclusão
	Pastagem	260	6	47	1	23	337	0,772	0,228
	Urbano	11	102	1	4	13	131	0,779	0,221
	Floresta	7	0	184	0	3	194	0,948	0,052
	Água	2	2	10	26	0	40	0,650	0,350
	Solo Exposto	7	8	1	1	94	111	0,847	0,153
	Total	287	118	243	32	133	813		
	Acurácia do Produtor	0,906	0,864	0,757	0,813	0,707		Acurácia Geral	0,819
	Erros de Omissão	0,094	0,136	0,243	0,188	0,293		Discordância Geral	0,181

Elaboração: Os autores (2022).

Ao observarmos a matriz normalizada pelo total apresentada na Tabela 11, percebemos que a classe com mais erros de inclusão é a de Água, indicando que a estimativa de área desta classe está superestimada na classificação. Porém, ao compararmos esta classificação com a realizada com composição RGB + NIR. nota-se que em termos gerais houve uma pequena redução nos erros de inclusão. No caso da classe Água, por exemplo, este erro passa de 36,4% na composição RGB + NIR, para 35% na composição CP. Já quanto aos erros de omissão, a maior concentração está localizada na classe de floresta, indicando que esta foi a classe mais subestimada na classificação. Nota-se nos erros de omissão a mesma tendência observada nos erros de inclusão, havendo valores de erros menores no caso da classificação com composição CP quando comparada à realizada com composição RGB + NIR. Observamos que para o conjunto total dos dados, as discordâncias de quantidade e alocação foram menores na classificação com as CP do que na classificação com RGB + NIR. Porém, também é perceptível que as classificações realizadas com CART obtiveram maiores discordâncias de quantidade e alocação quando comparadas com as realizadas com *Random Forest*, isso para ambos os dados de entrada.

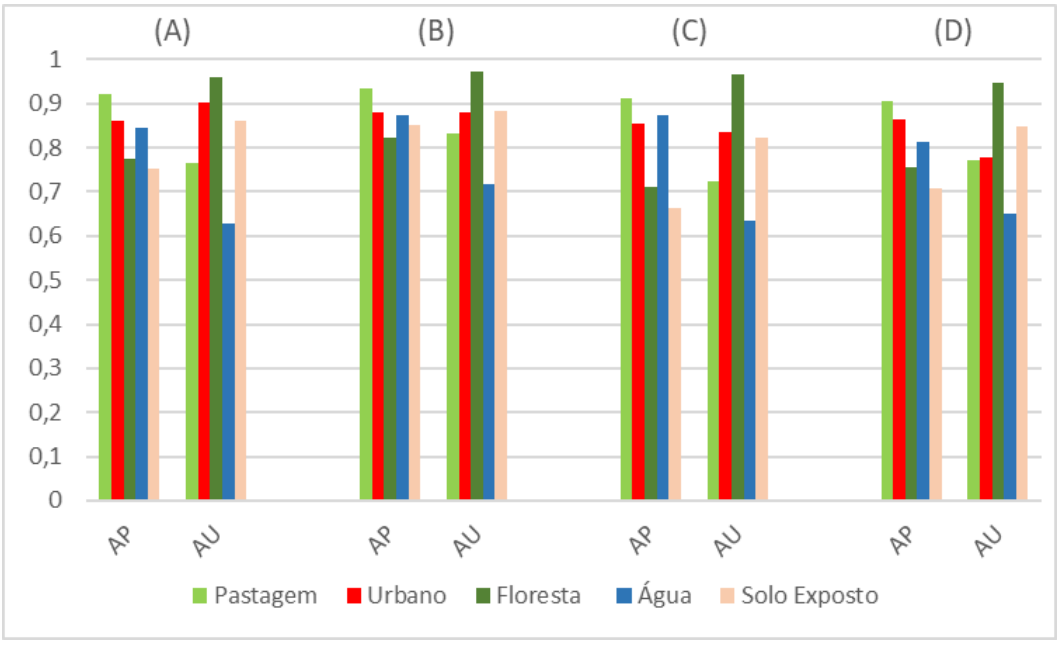
Tabela 11 - Matriz de confusão normalizada pelo total - Classificação CART com composição CP.

	Referência								
Classificação	Classes	Pastagem	Urbano	Floresta	Água	Solo Exposto	Total	Erros de Inclusão	Discordância de Quantidade
	Pastagem	0,320	0,007	0,058	0,001	0,028	0,415	0,228	0,062
	Urbano	0,014	0,125	0,001	0,005	0,016	0,161	0,221	0,016
	Floresta	0,009	0,000	0,226	0,000	0,004	0,239	0,052	0,060
	Água	0,002	0,002	0,012	0,032	0,000	0,049	0,350	0,010
	Solo Exposto	0,009	0,010	0,001	0,001	0,116	0,137	0,153	0,027
	Total	0,353	0,145	0,299	0,039	0,164	1,000	Discordância de quantidade	0,087
	Erros de Omissão	0,094	0,136	0,243	0,188	0,293		Discordância de alocação	0,093
	Discordância de Alocação	0,066	0,039	0,025	0,015	0,042		AG + DA	0,913

Elaboração: Os autores (2022).

Uma síntese das acurácias do produtor e do usuário para as quatro classificações pode ser observada na Figura 7. A análise do gráfico possibilita comparar visualmente essas métricas e compará-las.

Figura 7 - Síntese das acurácias do Produtor (AP) e Usuário (AU).
(A) Classificação *Random Forest* com composição RGB + NIR. (B) Classificação *Random Forest* com composição CP.(C) Classificação CART com composição RGB + NIR. (D) Classificação CART com composição CP.



Elaboração: Os autores (2022).

Avaliando o comportamento geral das classificações, nota-se que a classe de Água foi a que obteve os maiores erros de inclusão. Este fato está atrelado a alta variabilidade do comportamento espectral dos corpos d'água na área de estudo, o que dificulta a caracterização e discriminação dos mesmos. Os maiores erros de omissão estão associados à classe de Floresta, muito em função das confusões ocorridas entre esta classe e a classe de Pastagem. Melhorias nas amostras de entrada fornecidas para os algoritmos poderiam aumentar as respectivas acurácias. Todavia, levando em conta o objetivo deste trabalho, a elaboração de classificações com as mesmas amostras de entrada fornecidas aos classificadores possibilita comparar os resultados em função dos algoritmos e das composições de imagens utilizadas.

De forma sintética, destaca-se que os melhores resultados foram alcançados com o classificador *Random Forest*, para os dois tipos de dados de entrada (RGB+NIR e CP), com menos discordância de

quantidade e alocação quando comparadas com o CART. Quando consideramos o dado de entrada (RGB+NIR ou CP), nota-se que com o classificador *Random Forest* as componentes principais geraram melhores resultados do que as bandas espectrais utilizadas, com menores discordâncias em termo de quantidade e alocação. O classificador CART apresentou acurácia menor para os dois dados de entrada em relação ao *Random Forest*. Entretanto, também apresentou melhores resultados em termos de acurácia com a classificação a partir das CP. Pode-se notar que o uso das CP trouxe melhorias no resultado final da classificação do uso e cobertura da terra.

Alguns estudos também alcançaram bons resultados na classificação de imagens ao utilizar a técnica ACP. Dutra, Elmiro e Garcia (2022) comparou o uso de índices vegetativos, da imagem composição do infravermelho e das três primeiras componentes principais para a classificação de áreas florestadas a partir de imagens do satélite Landsat 8. E a partir da análise de acurácia e visual das classificações, concluiu-se que a ACP alcançou a melhor performance para identificar as áreas florestadas do que as outras abordagens. Arslan et al. (2023) utilizou a técnica de ACP para identificar derramamento de óleo em imagens Landsat 8 e Sentinel-2, objetivando melhorar a informação disponibilizada pelas bandas espectrais. Neste caso, as componentes principais descorrelacionadas entre si facilitam a diferenciação entre as manchas de óleo, as nuvens e a água do mar, tarefa difícil de ser alcançada somente com as bandas originais. Assim, a ACP foi aplicada como uma técnica de aprimoramento da imagem, já que remove a redundância original, gerando aumento no nível de contraste com redução dos ruídos, como aplicado também neste trabalho. Para o Sentinel-2 a CP 2 apresentou boa separação da área de óleo, e na Landsat 8 foi a CP 4 que teve um melhor contraste para este alvo. Apontando também a importância de se analisar as componentes geradas para a sua correta escolha.

Rovani, Dambros e Cassol (2013) utilizou a ACP similarmente ao aqui proposto, visando melhorar o mapeamento de uso e cobertura da terra a partir das bandas espectrais do Landsat 5. A técnica foi escolhida com objetivo de reduzir a repetição de dados, concentrando as principais informações em um conjunto menor de componentes principais. A partir dos resultados notou-se que a composição RGB construída com as componentes principais apresentava maior variabilidade de informação facilitando a discriminação dos alvos. Contribuindo assim para bons resultados na classificação, em relação ao uso tradicional das bandas espectrais. Tal resultado também foi notado neste estudo, com o alto contraste exibido pela composição RGB das três primeiras CP (Figura 4).

Ademais, aqui neste estudo utilizamos duas abordagens e dois classificadores, CART e *Random Forest*, nas imagens Sentinel-2. Ao observar os resultados alcançados, nota-se que o *Random Forest* obteve melhores resultados que o CART, isto pode estar associado ao nível de complexidade do classificador. Além disso, outros autores também encontraram resultados similares ao comparar os dois classificadores. Como em Neetu e Ray (2019) que comparou diversos classificadores baseados em aprendizado de máquina para mapeamento agrícola usando imagens do Sentinel-2, com as bandas de 10 metros de resolução espacial, no GEE. A análise de acurácia indicou que o *Random Forest* teve uma melhor desempenho que o CART.

Sujud et al. (2021) também comparou diversos classificadores do GEE com objetivo de mapear áreas agrícolas em ambiente semi-árido a partir de imagens do Landsat 8, Sentinel-2 e Sentinel-1. Considerando a acurácia geral os melhores resultados foram alcançados com *Support Vector Machine* e *Gradient Tree Boosting*. E novamente o Random Forest gerou melhores resultados que o CART. Em Praticò et al. (2021) foram analisados o desempenho do *Random Forest*, *Support Vector Machine* e CART para a realização de uma classificação supervisionada *pixel-a-pixel* com imagens Sentinel-2. Neste estudo o *Random Forest* alcançou o melhor resultado e o CART o pior. Portanto, condizendo com os resultados aqui encontrados. É esperado que o classificador *Random Forest* produza melhores resultados que o CART, já que o mesmo gera diversas árvores de decisões, como a do CART.

4 CONCLUSÕES

Este trabalho teve como proposta comparar e avaliar a potencialidade do uso de componentes principais para a classificação de imagens, em relação às bandas espectrais originais. Visando aprimorar o mapeamento do uso e cobertura da terra na bacia hidrográfica Lajeado São José no estado de Santa Catarina,

considerando o papel crucial que a informação sobre o uso e cobertura da terra detêm no processo de ordenamento territorial.

Os valores de correlação entre as bandas utilizadas no satélite *Sentinel-2*, indicaram a existência de correlação e consequentemente redundância entre os dados. Desta forma, qualificando as bandas espectrais para a aplicação da técnica de Análise de Componentes Principais. A aplicação da ACP no GEE se mostrou muito eficiente, concentrando a maioria da informação nas duas primeiras componentes principais e alocando o ruído na última componente, que foi descartada. A composição colorida gerada entre as três primeiras componentes principais apresenta um alto contraste, pois as bandas estão descorrelacionadas. Evidenciando os alvos existentes na imagem.

A etapa de classificação de imagens foi realizada de forma supervisionada no GEE com o uso de dois tipos de classificadores: CART e *Random Forest.* Os mesmos foram aplicados nas bandas espectrais e nas componentes principais selecionadas, totalizando quatro mapas de uso e cobertura da terra com cinco classes distintas. Para avaliar se o uso das componentes principais melhorou o resultado da classificação de imagens em relação ao uso das bandas espectrais originais, adotou-se uma abordagem de amostragem de pontos estratificados. Isso permitiu a avaliação do resultado das classificações em 813 pontos de observação a partir de uma imagem de alta resolução espacial. Permitindo a construção de matrizes de confusão que possibilitam comparar os resultados alcançados.

De maneira geral as classificações de uso e cobertura da terra para a área de estudo apresentaram resultados similares, mas com maior detalhe as diferenças aparecem. Os produtos gerados pelo classificador CART apresentam o efeito sal e pimenta, quando há *pixels* soltos no interior de outras classes. Por outro lado, o *Randon Forest* gerou um resultado mais homogêneo, com redução desse efeito. Para ambos os dados de entrada (RGB+NIR e CP) o algoritmo *Randon Forest produziu* melhores resultados, com menos discordância de quantidade e alocação quando comparadas com o CART.

Ao observar o do dado de entrada (RGB+NIR ou CP), nota-se que com o classificador *Random Forest* as componentes principais geraram melhores resultados que as bandas espectrais selecionadas, com menores discordâncias em termo de quantidade e alocação. Já o CART apresentou acurácia menor para os dois dados de entrada quando comparados ao *Random Forest*. Porém, também apresentou melhores resultados em termos de acurácia com a classificação a partir das CP. Assim, pode-se concluir que o uso de componentes principais apresenta potencialidade para gerar melhores resultados na classificação de uso e cobertura da terra quando comparados ao uso das bandas espectrais.

Para estudos futuros sugere-se a utilização de imagens de sensoriamento remoto com um maior número de bandas espectrais. Acredita-se que o potencial da Análise de Componentes Principais amplia-se com um maior número de bandas disponíveis. Ademais, com o uso de plataformas como o *Google Earth Engine* torna-se possível manipular e processar dados com alta dimensionalidade.

Agradecimentos

Os autores agradecem a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela concessão da bolsa de Mestrado (número dos processos: 88887.819084/2023-00, 88887.721596/2022-00).

Contribuição dos Autores

Os autores Letícia Sartorio, Macleidi Varnier e Leonardo Da Silva Felippe definiram em conjunto a conceptualização e a metodologia empregadas. Estes foram responsáveis pela curadoria dos dados, análise formal, investigação, validação, visualização e redação-revisão e edição. Os demais autores foram responsáveis pela administração e supervisão do projeto.

Conflito de interesses

Os autores declaram não haver conflitos de interesse.

Referências

ARSLAN, N.; MAJIDI NEZHAD, M.; HEYDARI, A.; ASTIASO GARCIA, D.; SYLAIOS, G. A Principal Component Analysis Methodology of Oil Spill Detection and Monitoring Using Satellite Remote Sensing Sensors. **Remote Sensing**, 2023. vol. 15, n. 5, pp. 1460.

BELGIU, M.; DRĂGUT, L. Random Forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, Abril de 2016, vol. 114, pp. 24–31.

BREIMAN, L. Random Forests, **Machine Learning**, Outubro de 2001, vol. 45, pp. 5–32.

BREIMAN, L.;, FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. **Classification and regression trees**. Nova Iorque: Wadsworth International Group. 1984.

BROWN, C.F.; BRUMBY, S.P.; GUZDER-WILLIAMS, B. *et al.* Dynamic World, Near real-time global 10 m land use land cover mapping. **Sci Data,** Julho de 2022**,** vol. 9, 251. pp. 1-17.

CROSTA, Alvaro Penteado. **Processamento digital de imagens de sensoriamento remoto**. UNICAMP/Instituto de Geociências, 1999.

DHARANI, M.; SREENIVASULU, G. Land use and land cover change detection by using principal component analysis and morphological operations in remote sensing applications. **International Journal of Computers and Applications**, Fevereiro de 2019. v. 43, n. 5, pp. 462-471.

DUTRA, D.; ELMIRO, M.; GARCIA, R. Comparative analysis of methods applied in vegetation cover delimitation using Landsat 8 images. **Sociedade & Natureza**, 2022. vol. 32, pp. 732-744.

EUROPEAN SPACE AGENCY (ESA). **Sentinel-2 User Handbook.** 2015. p.64. Disponível em: <https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook.pdf/8869acdf-fd84-43ec-ae8c-3e80a436a16c?t=1438278087000>. Acesso em: 31 de mar. 2023.

FERREIRA, M. E.; FERREIRA, L. G.; SANO, E. E.; SHIMABUKURO, Y. E. Spectral linear mixture modelling approaches for land cover mapping of tropical savanna areas in Brazil. **International Journal of Remote Sensing**, Janeiro de 2007. vol. 28, pp. 413-429.

FOODY, G. M. 2002. Status of Land Cover Classification Accuracy Assessment. **Remote Sensing of Environment.** Abril de 2002, vol. 80, pp. 185–201.

FOODY, Giles M. Harshness in image classification accuracy assessment. **International Journal Of Remote Sensing**. Maio de 2008. vol. 29, n. 11, pp. 3137-3158.

GAIDA, W.; BREUNIG, F. M.; GALVÃO, L. S.; PONZONI, F. J. Correção atmosférica em sensoriamento remoto: uma revisão. **Revista Brasileira de Geografia Física**. 2020. v.13, n. 01, pp. 229-248.

GOOGLE EARTH ENGINE (GEE). **Eigen Analysis**. 2022. Disponível em: < https://developers.google.com/earth-engine/guides/arrays_eigen_analysis>. Acesso em: 18 de jul. 2022.

GUPTA, R. P.; TIWARI, R. K.; SAINI, V.; SRIVASTAVA, N. A simplified approach for interpreting principal component images. **Advances in Remote Sensing**. 2013. vol. 2 n. 2, pp. 111-119.

GROTH, D.; HARTMANN, S.; KLIE, S.; SELBIG, J. Principal Components Analysis. In: REISFELD, B.; MAYENO, A. (eds) **Computational Toxicology. Methods in Molecular Biology**, Totowa, NJ: Humana Press, vol. 930, 2013. 527 – 547.

KOBAYASHI, N.; TANI, H.; WANG, X.; SONOBE, R. Crop classification using spectral indices derived from Sentinel-2A imagery. **Journal of Information and Telecommunication**, 2020. Vol. 4, pp. 67-90.

LEG - UFPR. **Árvores de regressão**. 2022. Disponível em: http://leg.ufpr.br/~lucambio/CE050/20211S/AR.html. Acesso em: 15 out. 2022.

MARTINEZ-IZQUIERDO, M.; MOLINA-SÁNCHEZ, I.; MORILLO-BALSERA, M. Efficient dimensionality reduction using principal component analysis for image change detection. **IEEE Latin America Transactions**. 2019, v. 17, n. 04, p. 540-547.

MAPBIOMAS. **MapBiomas General "Handbook" Algorithm Theoretical Basis Document (ATBD) - Collection 6**, Versão 1.0, pp. 1-49, 2022.

MATHER, P.; TSO, B. **Classification methods for remotely sensed data**. 2º edição. Boca Raton: CRC Press, 2009.

MENESES, Paulo Roberto; ALMEIDA, T. de. **Introdução ao processamento de imagens de sensoriamento remoto**. Brasília: Universidade de Brasília, pp. 266, 2012.

NEETU; RAY, S. S. EXPLORING MACHINE LEARNING CLASSIFICATION ALGORITHMS FOR CROP CLASSIFICATION USING SENTINEL 2 DATA. **International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences**, 2019. pp. 573–578.

OLOFSSON, P.; FOODY, G. M.; HEROLD, Martin; STEHMAN, Stephen V.; WOODCOCK, Curtis E.; WULDER, Michael A. Good practices for estimating area and assessing accuracy of land change. **Remote Sensing Of Environment**, Maio de 2014. v. 148, p. 42-57.

OTSUSCHI, C. **Alterações na vegetação florestal nativa nas bacias hidrográficas dos lajeados São José e Passo dos Índios – Oeste de Santa Catarina: Efeitos hidrológicos e na perda de solos entre 1989 e 2015**. Tese (Pós Graduação em Geografia) – Universidade Federal de Santa Maria, Santa Maria, 2017.

PANDOLFO, C. et al. **Atlas Climatológico do Estado de Santa Catarina**. Florianópolis: Epagri, 2002.

PONTIUS, R. G.; MILLONES, M. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. **International Journal Of Remote Sensing**, Agosto de 2011. vol. 32, n. 15, p. 4407-4429.

PRATICÒ, S., SOLANO, F., DI FAZIO, S., & MODICA, G. Machine learning classification of mediterranean forest habitats in google earth engine based on seasonal sentinel-2 time-series and input image composition optimisation. **Remote Sensing**, 2021. vol. 13, n. 4, pp. 586.

ROVANI, F.; DAMBROS, G.; CASSOL, R. Aplicação da análise por componentes principais para o mapeamento do uso e ocupação da terra no município de Barão de Cotegipe–RS. **Simpósio Brasileiro de Sensoriamento Remoto**,2013. vol. 16, pp. 8091-8097.

SANTOS, D. F. M.; COSTA, A. M.; OLIVEIRA, F. S.; VIANA, J. H. M. Monitoramento do uso e cobertura do solo em Sete Lagoas e Prudente de Morais - MG entre 1990-2015. **Revista Raega - O Espaço Geográfico em Análise**. Fevereiro de 2018. Curitiba, v. 43, p. 57-74.

SANTOS, Rozely Ferreira dos. **Planejamento ambiental**. São Paulo: Oficina de Textos, p. 71-135, 2004.

SOUZA, C.; SHIMBO, J. Z.; ROSA, M. R.; PARENTE, L. L.; ALENCAR, A. A.; RUDORFF, B. F. T.; ...; AZEVEDO, T. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. **Remote Sensing**, 2020. vol. 12, n. 17, pp. 27.

SUJUD, L.; JAAFAR, H.; HASSAN, M. A. H.; ZURAYK, R. Cannabis detection from optical and RADAR data fusion: A comparative analysis of the SMILE machine learning algorithms in Google Earth Engine. **Remote Sensing Applications: Society and Environment**, 2021. vol. 24.

STEHMAN, Stephen V.; CZAPLEWSKI, Raymond L. Design and Analysis for Thematic Map Accuracy Assessment. **Remote Sensing Of Environment**, Junho de 1998. vol. 64, n. 3, p. 331-344.

STEHMAN, Stephen V. Estimating area from an accuracy assessment error matrix. **Remote Sensing Of Environment**, Maio de 2013. vol. 132, p. 202-211.

ZANOTTA, D.; FERREIRA, M.; ZORTEA, M. **Processamento de Imagens de Satélite**. 1° edição. São Paulo: Oficina de Textos, 2019.

Biografia do autor principal

Letícia Figueiredo Sartorio, nascida em Rio Grande no ano de 1998. Possui técnico em Geoprocessamento pelo Instituto Federal do Rio Grande do Sul – IFRS (2016) e graduação em Geografia Bacharelado pela Universidade Federal do Rio Grande (2022) Atualmente é mestranda no Programa de Pós-Graduação em Geografia pela Universidade Federal do Rio Grande do Sul, na linha de Análise Ambiental.