CLASSIFICAÇÃO DO USO DO SOLO DE UMA BACIA HIDROGRÁFICA POR MEIO DE ALGORITMOS DE *MACHINE LARNING*

Rafael João Sampaio[1], Carol Santos Shuenck Maya[2], Gisele Dornelles Pires[3] e Fabrício Polifke da Silva[4]

[1,3,4] Professores e Pesquisadores Engenharia Civil UNIG [2] Aluna Graduação Engenharia Civil UNIG
[1,2,3,4,] Grupo de Pesquisa Engenharia e Sociedade, Faculdade de Ciências Exatas e Tecnológicas, Universidade Iguaçu - UNIG, Av. Abílio Augusto Távora, 2134 - Jardim Nova Era, 26275-580, Nova Iguaçu - RJ

samprafael@gmail.com , carolshuenck@outlook.com, unigengenharia@gmail.com, briciopolifke@gmail.com

Resumo - *O monitoramento do uso de solo em uma bacia hidrográfica é fundamental para gestão dos recursos hídricos, sendo essa importância ampliada em bacias com altas taxas de urbanização, em que se eleva o grau do efeito antrópico no ambiente. O uso de algoritmos de aprendizado de máquinas aplicado a imagens de satélite tem se apresentado nos últimos anos como uma alternativa promissora para convecção de mapas de cobertura do solo. Este trabalho analisa o desempenho dos algoritmos árvore de decisão (DT), floresta aleatória (RF), máquina de vetor suporte (SVM) e k vizinhos mais próximos (K-NN) para classificar o uso do solo na bacia do rio Iguaçu – Sarapuí – RJ por meio de uma imagem Landsat 8 OLI/TIRS. Esta bacia é fortemente degradada, possuindo grandes áreas urbanas e industriais. Foram utilizados 5 mil pontos de amostragem, divididos em um conjunto de treinamento (77%) e teste (33%). As sete primeiras bandas espectrais mais o Índice de Vegetação Normalizado (NDVI) foram combinadas em treze ensaios diferentes. O algoritmo árvore de decisão e K-NN foram que obtiveram melhores desempenhos, alcançando 75% de acerto. Entre os ensaios, os melhores ensaios consideraram as bandas 1, 2, 3, 4, 5 e 6. As maiores taxas de acerto em relação as classes foram em áreas urbanas (83%) e floresta (88%).*

*(**Palavras chaves** – Aprendizado de máquinas, Uso do solo, inteligência artificial)*

Abstract - The monitoring of soil use in a river basin is fundamental for the management of water resources, being one of the largest dimensions in basins with high rates of urbanization, in which the degree of anthropic effect on the environment increases. The use of machine learning algorithms was used as one of the satellite images in recent years as a promising alternative for the convection of soil cover maps. This work analyzes the data of quantitative methods (DT), random forest (RF), support material (SVM) and k villages plus next (KNN) to separate the soil in the Iguaçu River basin - Sarapuí - RJ by means of a image Landsat 8 OLI / TIRS. This region is heavily degraded, with large urban and industrial areas. Five thousand sampling points were used, divided into a training set (77%) and test (33%). The first seven spectral bands plus the Normalized Vegetation Index (NDVI) were combined in thirteen different assays. The algorithm decision tree and K-NN were the ones that obtained better performances, reaching 75% of correct. Among the trials, the best trials are considered bands 1, 2, 3, 4, 5 and 6. Access rates are the classes in urban areas (83%) and forest (88%).

*(**Keywords** - Machine learning, Soil use, artificial intelligence)*

1 Introdução

Nas últimas décadas ocorreram importantes avanços para desenvolvimento métodos de sensoriamento remoto que permitam a caracterização precisa da mudança da cobertura do solo (Rogan & Chen 2004), incluindo a expansão das cidades (Chan et al. 2001, Scheneider et al. 2012, Lyu et al. 2018). No entanto, o mapeamento de áreas urbanas continua vem sendo um desafio complexo devido à heterogeneidade característica de urbanizações que podem levar a diferentes variações dentro dos pixels de cada imagem espectral (Small e Lu 2006). Particularmente problemático também é o fato de que as áreas urbanas recém-desenvolvidas aparentam tipicamente às terras cultivadas em pousio em um dado momento, uma vez que ambas exibem alta refletância nos comprimentos de onda do visível-infravermelho (Long et al. 2009). Fator que dificulta a identificação da classe por meio de imagens.

A classificação de imagens espectrais utilizando algoritmos de aprendizado de máquina tem recebido atenção considerável nas últimas décadas (Hepner et al. 1990; Lawrence et al. 2004), fazendo com que pesquisadores na área de sensoriamento remoto estão adotando cada vez mais esses classificadores para mapeamento de uso do solo. Quando comparados com produtos derivados de classificadores estatísticos tradicionais (por exemplo, distância mínima, máxima verossimilhança e classificadores de paralelepípedos), os resultados de pesquisas geralmente demostram que os classificadores de aprendizado de máquina geram

resultados mais precisos e confiáveis, particularmente se estiverem disponíveis dados abundantes de treinamento (Schneider 2012, Pal e Mather 2005). Embora há uma disponibilidade de imagens remotamente capturadas, como MODIS (Shao e Lunetta, 2012) e Worldview (Zhen et al. 2013), para serem testadas como entrada para classificadores de aprendizado de máquina, as imagens da série Landsat comumente são mais utilizadas devido sua acessibilidade livre e moderada resolução temporária. Dois de muitos exemplos são Li et al. (2013) que aplicaram o aprendizado de máquina para mapear e rastrear mudanças para espécies florestais, e Rogan et al. (2008) para o mapeamento da mudança da cobertura da terra.

A bacia do rio Iguaçu-Sarapuí abrange uma parte expressiva da Região Metropolitana do Rio de Janeiro (RMRJ). A região apresenta áreas com grande crescimento urbano e industrial, além de áreas rurais em processo de urbanização e áreas onde a ocupação do solo conflita com as condições de habitabilidade, em especial nas áreas mal drenadas (Carneiro et al. 2008). A bacia também apresenta recorrentes e graves problemas de inundações e possui mananciais importantes para o abastecimento de parte da Baixada Fluminense (Paiva 2014).

À medida que o número de algoritmos de aprendizado de máquina aumenta é benéfico para a comunidade científica a avaliação desses métodos nas mais diversas circunstâncias, afim de obter um melhor conhecimento sobre os desempenhos de cada algoritmo. No campo da classificação de imagens por sensoriamento remoto é necessária uma comparação mais abrangente dos principais algoritmos de aprendizado de máquina. Isso deve ser feito com o mesmo esquema de classificação de cobertura e uso da terra e a mesma imagem de satélite (Li et al 2014). Acredita-se que os resultados finais da classificação de imagem dependem de vários fatores: esquema de classificação, dados de imagem disponíveis, seleção de amostra de treinamento, pré-processamento dos dados incluindo seleção e extração de características, algoritmo de classificação, técnicas de pós-processamento e coleta de amostra. e métodos de validação (Gong e Howarth 1990). Neste contexto, o presente estudo busca analisar o desempenho de dos algoritmos árvore de decisão, floresta aleatória, máquina de vetor suporte para a classificação do uso de solo da bacia do rio Iguaçu-Sarapuí a partir de uma imagem Landsat 8 OLI / TIRS.

2 Metodologia

2.1 Área de estudo

A bacia do rio Iguaçu-Sarapuí (BHIS) tem uma área de drenagem que mede 726 Km^2, dos quais 168 Km^2 representam a sub-bacia do Sarapuí e 558 Km^2 a do Iguaçu (Carneiro et al. 2004). Esta bacia abriga parte dos Municípios do Rio de Janeiro, Nilópolis, Mesquita, São João de Meriti, Nova Iguaçu, Belford Roxo e Duque de Caxias, todos inseridos na Região Metropolitana do Rio de Janeiro, conforme demonstra a Figura 1. Possui como limites no norte, a bacia do rio Paraíba do Sul, no sul, a bacia dos rios Pavuna/Meriti, no leste, a bacia dos rios Inhomirim/Estrela e no oeste, a bacia do Rio Guandu e afluentes da baía de Sepetiba. O rio principal, rio Iguaçu, tem suas nascentes na serra do Tinguá, a 1000 m acima do mar, aproximadamente. Possui uma extensão de 43 km e deságua na Baía de Guanabara. Seus principais afluentes são os rios: Tinguá, Pati e Capivari pela margem esquerda e Botas e Sarapuí, pela margem direita. O rio Sarapuí, segundo em importância para a bacia, nasce na serra de Bangu, no maciço da Pedra Branca, município do Rio de Janeiro, numa altitude de aproximadamente 900 m. De sua nascente até a sua foz no rio Iguaçu, percorre cerca de 36 Km.

O relevo da bacia se caracteriza principalmente por duas unidades: a serra do Mar, onde se encontra o ponto culminante da bacia, o pico do Tinguá, (1600m), e a Baixada Fluminense. O clima é quente e úmido, com estação chuvosa no verão, temperatura média em torno de 22°C e precipitação média anual em torno de 1700 mm.

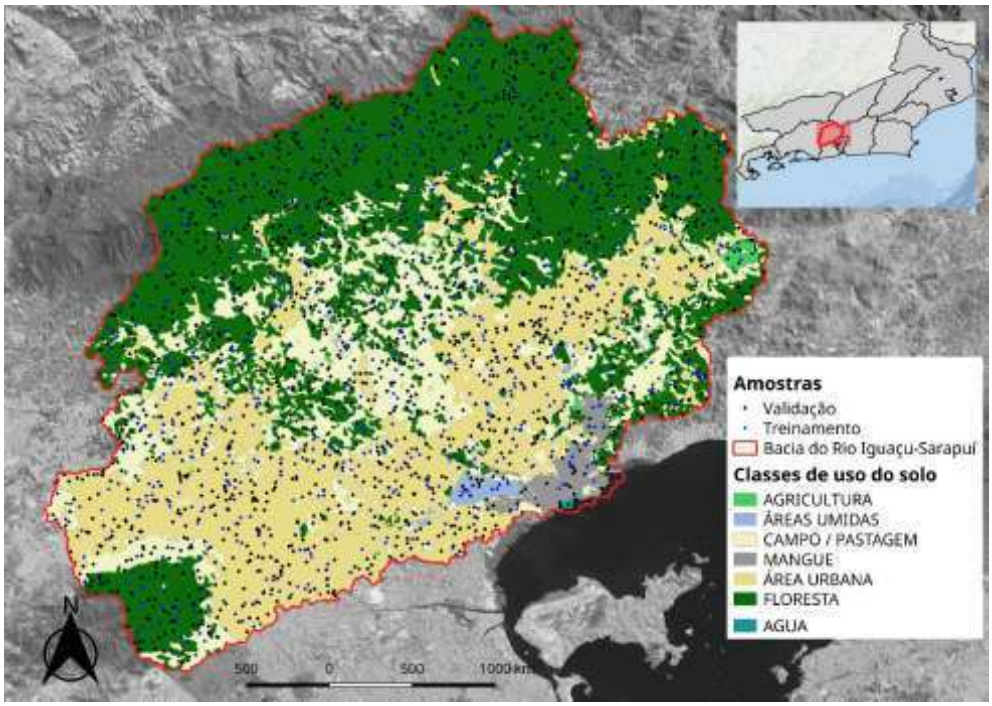

Figura 1 - Uso do solo na bacia do rio Iguaçu-Sarapuí (linha vermelha). Pontos azuis e pretos indicam, respectivamente, as amostras de treinamento e de validação do algoritmo.

2.2 Dados de treinamento e Algoritmos de classificação

O estudo teve como base uma cena Landsat 8 OLI/TIRSS (cena 217/76) coletada em 12/05/2018, obtida no endereço *online earthexplorer.usgs.gov*. As 7 primeiras bandas espectrais mais o Índice de Vegetação Normalizado (NDVI), obtidos por uma relação algébrica entre as bandas 5 e 4 (Schlerf et al. 2005), foram analisadas. Estes dados foram combinados em 13 ensaios diferentes (Tabela 1) no intuito de eleger as melhores combinações para classificação.

Tabela 1 - Ensaios analisados para

Ensaios	Bandas Landsat 8 OLI/TIRS
A1	1, 2, 3, 4, 5, 6
A2	1, 2, 3
A3	2, 3, 4
A4	3, 4, 5
A5	5, 6, 7
A6	2, 3, 6
A7	1, 2, 6
N1	2, 3, 4, NDVI
N2	1, 2, 3, NDVI
N3	2, 3, NDVI
N4	3, 4, NDVI
N5	1, 2, 6, NDVI
N6	2, 3, 6, NDVI

O treinamento dos algoritmos teve como referência o mapa de uso e cobertura do solo 2015, produzido pelo Instituto Estadual do Ambiente (INEA). A execução do mapa foi realizada em março de 2017 com detecções de mudanças bi anual, escala 1:100.000 e sistema de referência SIRGAS 2000 (EPS: 4674). O metadados pode ser acessado em *http://www.metadados.inde.gov.br/*. Cinco mil pontos amostrais cobrindo toda a área da bacia do rio Iguaçu-Sarapuí foram gerados aleatoriamente no software Qgis 2.18 (Qgis 2019). Destes, 3350 foram utilizados para treinamento e 1650 foram utilizados para validação (Figura 1).

2.2.1 Árvore de decisão

A árvores de decisão (DT), juntamente com redes neurais, são os algoritmos de aprendizado de máquinas mais utilizados na classificação de dados de sensoriamento remoto (Rogan et al. 2008, Galiano et al., 2014). Consistem em representações simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados. As árvores de decisão representam um conjunto de restrições ou condições, que são organizadas hierarquicamente e que são aplicadas sucessivamente de uma raiz a um nó terminal ou folha da árvore (Breiman et al. 1984). Para induzir a DT de um conjunto de dados, uma medida de avaliação de cada uma das variáveis é usada para maximizar a heterogeneidade interclasse. Existem muitas abordagens para selecionar atributos que podem ser usados para a indução de modelos de árvore de decisão (Galiano et al. 2014). Alguns dos mais frequentes são o índice de Gini e qui-quadrado. Uma vez que a medida de avaliação foi escolhida, a variável a partir da qual as divisões começarão é determinada (nó raiz). A partir do nó raiz, o processo de divisão de dados em cada nó interno de uma regra de classificação da árvore é repetido até que todos os exemplos mostrem o mesmo rótulo ou uma condição de parada especificada anteriormente seja atingida. Depois que o processo de indução da árvore é concluído, a poda é aplicada com o objetivo de melhorar a capacidade de generalização da árvore, reduzindo sua complexidade estrutural. O número de nós de folhas, o número de nós internos ou a profundidade da árvore podem ser considerados critérios de remoção.

2.2.2 Floresta Aleatória

O algoritmo de classificação floresta aleatória (RF) pode ser compreendida como uma combinação de várias árvores de decisão, de forma que cada árvore se relaciona com vetores aleatórios amostrados de forma independente e distribuídos igualmente para todas as arvores na floresta (Breiman, 2001). Dessa forma é definido como um conjunto de muitas classificações ou árvores de regressão projetadas para produzir previsões precisas, que não sobrecarregam os dados (Feng et al., 2017).

2.2.3 K vizinhos mais próximos - K-NN

O algoritmo *vizinho mais próximo* - K-NN é um algoritmo de aprendizado supervisionado do tipo lazy, introduzido por Aha et al. (1991). A ideia geral desse algoritmo consiste em encontrar os k exemplos rotulados mais próximos do exemplo não classificado e, com base no rótulo desses exemplos mais próximos, é tomada a decisão relativa à classe do exemplo não rotulado. Os algoritmos da família kNN requerem pouco esforço durante a etapa de treinamento. Em contrapartida, o custo computacional para rotular um novo exemplo é relativamente alto, pois, no pior dos casos, esse exemplo deverá ser comparado com todos os exemplos contidos no conjunto de exemplos de treinamento.

O algoritmo kNN classifica exemplos considerando a classe dos k vizinhos mais próximos. Se $k = 1$, então o exemplo é classificado com a mesma classe do exemplo mais próximo segundo a medida de similaridade utilizada. Se $k > 1$, então são consideradas as classes dos k exemplos mais próximos para realizar a classificação. Nesse caso, a abordagem mais simples consiste em atribui ao exemplo a classe majoritária dos k exemplos mais próximos.

2.2.4 Máquina de vetor de suporte

O algoritmo máquina de vetor de suporte (SVM) é um método não paramétrico de classificação supervisionada baseado em técnicas de aprendizagem estatística, portanto, não há suposições feitas sobre a distribuição de dados subjacentes (Mountrakis e Ogole 2011). Em sua formulação original (Vapnik, 1979), o método é apresentado com um conjunto de instâncias de dados rotulados e em que o algoritmo de treinamento SVM visa encontrar um hiperplano que separe o conjunto de dados em um número discreto de classes predefinidas de acordo com o treinamento. O termo hiperplano de separação ótima é usado para se referir ao limite de decisão que minimiza erros de classificação, obtidos na etapa de treinamento. A aprendizagem refere-se ao processo iterativo de encontrar um classificador com limite de decisão ideal para separar os padrões de treinamento (em espaço potencialmente alto-dimensional) e separar dados de simulação sob as mesmas configurações (dimensões) (Zhu e Blumberg, 2002).

2.2.5 Parametrizações dos algoritmos e implementação

Os parâmetros de entrada de cada algoritmo foram definidos através de diversos testes iniciais com dados da imagem do Landsat 8 utilizadas no estudo. Os parâmetros de cada algoritmo que obtiveram os melhores resultados foram adotados para todo o estudo. A tabela 2 apresenta os parâmetros utilizados em cada algoritmo de classificação.

Tabela 2 - Parametrização utilizada em cada algoritmo estudado

Algoritmos	**Parâmetros**
Máquina de vetores suporte (SVM)	Kernel= Radial Basis Function Gamma = 1/n
k vizinhos mais próximos (K-nn)	k = 5 metric = minkowski wight = No weightung
Árvore de decisão	Criterion = gini, n =5
Floresta aleatória	Criterion = gini, n = 50

Os algoritmos de aprendizado de máquinas foram implementados e aplicados utilizando o pacotes Scikitlearn (https://scikit-learn.org/) e rasterio (https://rasterio.readthedocs.io/), desenvolvidos em python. Os testes estatísticos foram realizados pelo pacote statsmodels (https://www.statsmodels.org/), também para python. Os mapas foram elaborados no software Qgis 2.18.

3 Avaliação do desempenho

A avaliação do desempenho de cada algoritmo foi realizada utilizando a exatidão global (G) e o índice Kappa (K). O primeiro é obtido pela razão direta entre os píxeis classificados corretamente e a quantidade total, como apresenta a equação 1

$$G = \frac{\sum x_{ii}}{n} \quad (1)$$

onde G é a exatidão global, x_{ii} são os píxels classificados corretamente e n é o número total de píxels na imagem. O coeficiente Kappa relaciona a classificação real do píxel com a probabilidade esperada, conforme a equação 3 .

$$K = \frac{n\sum x_{ii} - \sum x_{i.}x_{.i}}{n^2 - \sum x_{i.}x_{.i}} \quad (2)$$

onde k é uma estimativa para o coeficiente Kappa.

O valor do índice K varia de 0 a 1, sendo que mais próximo de um melhor a estimativa do modelo. Pode-se considerar a escala 0 – 0,2 = ruim; 0,2 – 0,4 = razoável; 0,4 – 0,6 = boa; 0,6 – 0,8 = muito boa; e 0,8 – 1,0 = excelente (Landis e Koch, 1977).

3. Resultados

Considerando o a acurácia global (G) e o índice Kappa todos os algoritmos tiveram os resultados mais significativos no ensaio A1, que considera as seis primeiras bandas do Landsat 8, e no ensaio N6, que considera as bandas 2,3 e 6 mais o NDVI, não havendo diferença significativa entre os dois (Figura 2). O pior desempenho foi no ensaio A4, que considera as bandas 3,4 e 5. Não houve um efeito significativo para o acréscimo do NDVI, exceto para o ensaio N6. Individualmente, o algoritmo k vizinhos mais próximos (K-NN) apresentou um desempenho considerável nos ensaios A2 (1, 2 e 3), A6 (2, 3 e 6), N2 (1, 2, 3 e NDVI) e N5 (1, 2, 6, NDVI), como mostra a tabela 3. Enquanto a árvore de decisão obteve uma performance semelhante nos ensaios A6, N1 (2, 3, 4 e NDVI) e N6 (2, 3, 6 e NDVI). O algoritmo máquina de vetor suporte (SVM) também obteve boas estimativas em N6. Este resultado contrasta com Rodriguez et al. (2014), que obtiveram as melhores classificações de uso do solo com os algoritmos floresta aleatória e SVM.

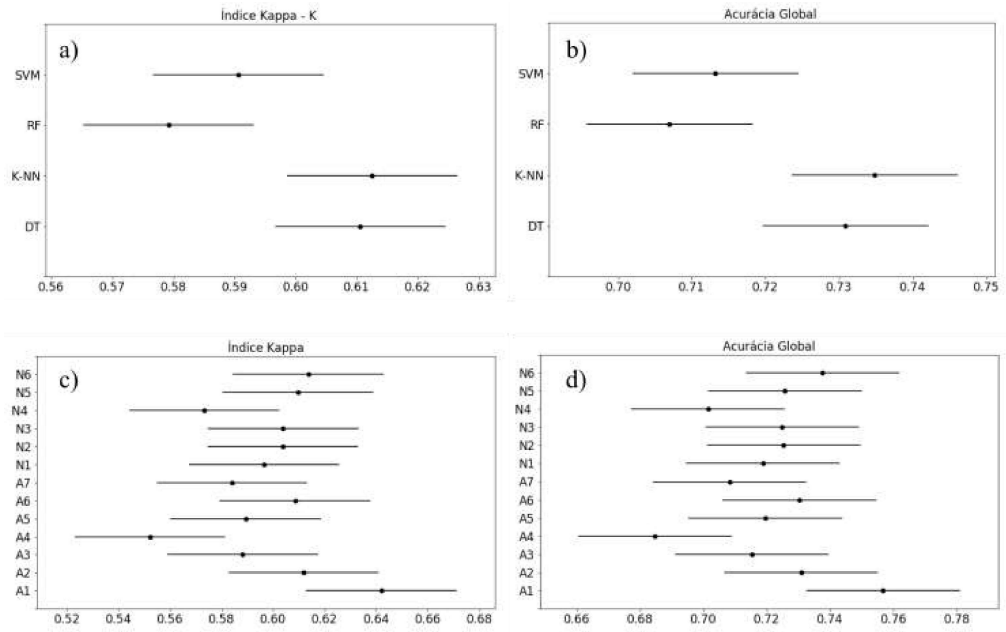

Figura 2 - Teste tukey para comparação de médias. As linhas indicam os limites de cada intervalo de confiança. Letras a e b: índice Kappa e acurácia global para os algoritmos respectivamente. Letras c e d: índice Kappa e acurácia global para os ensaios respectivamente.

As melhores classificações foram realizadas com os algoritmos K-NN e árvore de decisão DT, alcançando pelo índica Kappa um resultado muito bom (Figura 2). Entretanto não houve diferença significativa pelo teste de médias entre estes dois algoritmos e o SVM, enquanto a floresta aleatória apresentou o pior desempenho entre os quatro algoritmos. Apesar do desempenho semelhante, o algoritmo árvore de decisão tem a vantagem de ter um menor custo computacional para implementação em relação o K-NN.

Tabela 3 - Desempenho na validação de cada ensaio utilizando os algoritmos K vizinhos mais próximos (K-NN), floresta aleatória (RF), árvore de decisão (DT) e máquina de vetor de suporte (SVM).

	K-NN		**RF**		**DT**		**SVM**	
Exp	**G**	**K**	**G**	**K**	**G**	**K**	**G**	**k**
A1	**0.7573**	**0.6404**	**0.7415**	0.6226	**0.7876**	**0.6794**	**0.7403**	0.6256
A2	**0.7470**	**0.6320**	0.7136	0.5906	**0.7561**	**0.6421**	0.7069	0.5826
A3	0.7348	0.6148	0.6990	0.5633	0.7130	0.5875	0.7142	0.5874
A4	0.6996	0.5642	0.6814	0.5488	0.6833	0.5531	0.6742	0.5427
A5	0.7360	0.6123	0.6833	0.5403	0.7288	0.6011	0.7300	0.6038
A6	**0.7530**	**0.6361**	0.6960	0.5628	**0.7573**	0.6440	0.7148	0.5912
A7	0.7342	0.6118	0.6845	0.5587	0.7203	0.5982	0.6942	0.5677
N1	0.7263	0.6036	0.7124	0.5890	**0.7421**	0.6241	0.6942	0.5694
N2	**0.7415**	0.6210	0.7203	0.5970	0.7166	0.5943	0.7227	0.6031
N3	0.7282	0.6054	0.7069	0.5828	0.7397	0.6198	0.7245	0.6076
N4	0.7160	0.5871	0.7057	0.5761	0.6978	0.5718	0.6863	0.5580
N5	**0.7439**	0.6261	0.7191	0.5985	0.6996	0.5836	**0.7403**	0.6301
N6	0.7348	0.6080	0.7269	0.5994	**0.7591**	**0.6383**	0.7294	0.6087

Utilizando algoritmo árvore aleatória de decisão construiu-se a matriz confusão para uma imagem classificada da bacia do Rio Iguaçu-Sarapuí, apresentada na tabela 4. Das classes analisadas os maiores acertos foram para floresta (88%) e áreas urbanas (83,4%), enquanto o pior foi para classe áreas úmidas (23,5%). Uma provável razão é o número de amostras de treinamento para cada uma das classes. As classes floresta e área

urbana correspondem a maior parte da bacia do rio Iguaçu-Sarapuí (Figura 1) e consequentemente tiveram mais pontos de amostragem. Em contrapartida, as classes áreas úmidas e agricultura possuem pouca representatividade na bacia.

Tabela 4 - Matriz confusão para algoritmo árvore de decisão. Os resultados estão normalizados de 0 a 1.

	AGRICULTURA	AGUA	A. UMIDAS	FLORESTA	MANGUE	PASTAGEM	URBANO
AGRICULTURA	**0,471**	0,000	0,000	0,118	0,000	0,235	0,176
AGUA	0,250	**0,750**	0,000	0,000	0,000	0,000	0,000
A. UMIDAS	0,559	0,000	**0,235**	0,088	0,059	0,000	0,059
FLORESTA	0,061	0,000	0,000	**0,881**	0,000	0,045	0,013
MANGUE	0,207	0,000	0,017	0,103	**0,655**	0,000	0,017
PASTAGEM	0,229	0,000	0,002	0,173	0,000	**0,478**	0,119
URBANO	0,105	0,000	0,000	0,026	0,000	0,035	**0,834**

O baixo número de amostras pode também justificar uma baixa sensibilidade do algoritmo aos tipos vegetação. A classe pastagem, por exemplo, possui as maiores quantidades de falso positivo nas classes agricultura e floresta, resultado semelhante é visto em relação as classes agricultura e floresta. Galiano et al. (2014) analisou a sensibilidade dos métodos de aprendizagem de máquinas em relação ao número de amostras de treinamento em uma classificação de uso do solo. Classes homogêneas, como corpos hídricos, não aumentam a dificuldade de classificação com a redução de amostras de treinamento, entretanto classes heterogêneas, como é o caso de vegetações, são mais difíceis de serem classificadas corretamente com um menor número de amostras de treinamento.

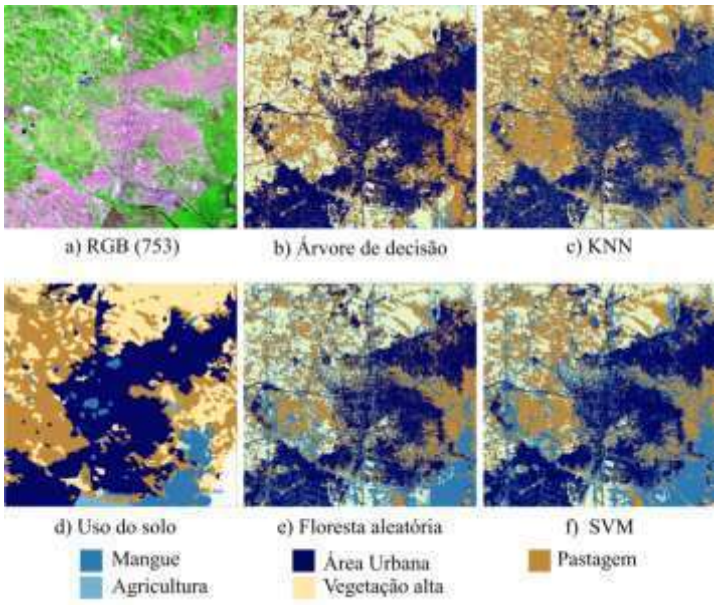

Figura 3 - Detalhe de uma região na bacia do rio Iguaçu-Sarapuí demonstrando o desempenho de cada um dos 4 algoritmos estudados na classificação automática do uso do solo de uma imagem Landsat 8 OLI/TIRS. O item 'a' apresenta uma cena Landsat 8 na com as combinações de cor/banda R = 7, G = 5 e B = 3; o item 'd' apresenta o mapa temático usado para extrair os dados de treinamento e validação; os demais item apresentam os mapas gerados pelos respectivos algoritmos.

A partir de uma cena Landsat 8, os métodos de aprendizado de máquina estudados demonstraram um potencial para produzir um mapa de uso do solo com resolução espacial superior ao dado de entrada utilizado, como demonstra a figura 3. Foram utilizadas as bandas 2,3, 6 e o NDVI. Todos algoritmos conseguiram identificar a mancha urbana no trecho da imagem, sendo a árvore de decisão apresentou o melhor resultado. Os algoritmos floresta aleatória e SVM não conseguiram classificar corretamente as classes agricultura e mangue, fazendo com que essas se comportassem como um ruído na imagem. O algoritmo K-NN demonstrou um resultado melhor em relação aos demais na classificação de píxels de pastagem, tendo como base a imagem de treinamento.

4. Conclusão

O presente trabalhou demonstrou o potencial que os algoritmos de aprendizagem de máquinas possuem para classificação de uso do solo utilizando imagens espectrais satelitais. Os algoritmos árvore de decisão e K-NN apresentaram os melhores desempenhos na classificação do uso do solo de uma imagem Landsat 8 da bacia do rio Iguaçu-Sarapuí. Os resultados demonstraram que os algoritmos conseguiram classificar automaticamente uma imagem espectral obtendo um resultado com um maior nível de detalhe que o dado de entrada. Isto é um indicativo que o aprendizado de máquinas podem reduzir custos na confecção de mapas de uso do solo. Ademais, possuem o potencial para serem uma ferramenta para monitoramento da cobertura do terreno, uma vez que, devido a automação, a periodicidade da classificação da cobertura do solo só depende da resolução temporal do sensor remoto. Há uma infinidade de alternativas para estudos futuros, como explorar outros algoritmos e uma maior quantidade de dados para treinamento, assim como a aplicação de aprendizado de máquinas em imagens com maior resolução espacial.

5. Agradecimentos

Os autores agradecem o Serviço Geológico Americano (USGS), que disponibilizou gratuitamente a imagem Landsat, a Universidade Iguaçu, que forneceu recursos para realização do trabalho, e toda comunidade *open source*, que desenvolveram de forma gratuita todos os softwares e as bibliotecas utilizadas neste estudo.

6.Referências

AHA DW (1990). A study of instance-based algorithms for supervised learning tasks:mathematical, empirical and psychological evaluations. Novembro. Tese de Doutorado, Universidade da California

Breiman L, Friedman J, Stone CJ, Olshen RA. (1984) Classification and Regression Trees. Taylor & Francis, Belmount, California.

Breiman L. (2001). Random forests. Machine Learning, Londres

Carneiro PRF, Cardoso, AL, Azevedo JPS de. (2008) O planejamento do uso do solo urbano e a gestão de bacias hidrográficas: o caso da bacia dos rios Iguaçu/Sarapuí na Baixada Fluminense. Cadernos metrópole 19: 165-190.

Chan JCW, Chan KP, Yeh AGO. (2001). Detecting the nature of change in an urban environment: A comparison ofmachine learning algorithms. Photogrammet- ric Engineering and Remote Sensing 67: 213–225.

Feng Y, Cui N, Gong D, Zhang Q, Zhao L(2017). Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling. Agricultural Water Management 193: 163–173.

Galiano VFR, Chica-Rivas M(2014). Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models. International Journal of Digital Earth 7: 492–509.

Gong P, Howarth, PJ. (1990) An assessment of some factors influencing multispectral land-cover classification. Photogramm. Engeneer Remote Sensing 56: 597–603.

Hepner GF, Logan T, Ritter N, and Bryant N. (1990) Artificial Neural Network Classification Using a Minimal Training Set- Comparison to Conventional Supervised Classification. Photogrammetric Engineering and Remote Sensing 56 4: 469–473.

Landis JR, Koch GG.(1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159-174.

Lawrence, R, Bunn A, Powell S, Zambon M. (2004) "Classification of Remotely Sensed Imagery Using Stochastic Gradient Boosting as a Refinement of Classification Tree Analysis." Remote Sensing of Environment 90: 331–336

Li M, Im J, Beier C. (2013). Machine Learning Approaches for Forest Classification and Change Analysis Using Multi-Temporal Landsat TM Images over Huntington Wildlife Forest. GIScience & Remote Sensing 50: 361–384.

Li C, Wang J, Wang L, Hu L, Gong P. (2014). Comparison of classification algorithms and training sample sizes in urban land classification with landsat thematic mapper imagery. Remote Sensing 6: 964–983.

Lyu H, Lu H, Mou L, Li W, Wright J, Li X, Gong P (2018). Long-term annual mapping of four cities on different continents by applying a deep information learning method to Landsat data. Remote Sensing, 10: 1-23.

Long H, Liua Y, Wu X, Dong G. (2009). Spatio-temporal dynamic patterns offarm- land and rural settlements in Su–Xi–Chang region: Implications for building a new countryside in coastal China. Land Use Policy 26: 322–333.

Mountrakis G, Im J, Ogole C. (2011) Support Vector Machines in Remote Sensing: A Review. ISPRS Journal of Photogrammetry and Remote Sensing 66: 247–259.

Paiva ACE (2014) Previsibilidade de cheias na bacia do rio Iguaçu/Sarapuí - RJ, usando modelagem hidrológica distribuída. Tese de doutorado, Instituto Nacional de Pesquisas Espaciais .

Rogan J, Franklin J, Stow DA, Miller J, Woodcock C, and Roberts D. (2008). Mapping Land-Cover Modifications over Large Areas: A Comparison of Machine Learning Algorithms. Remote Sensing of Environment 112: 2272–2283.

Pal M (2005). Random forest classifier for remote sensing classification. International Journal ofRemote Sensing, 26: 217–222.

QGIS Development Team (2019). QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org. Acessado em 20 de abril de 2019

Rogan J, Chen DM (2004). Remote sensing technology for mapping and monitor- ing land-cover and land-use change. Progress in Planning 61: 301–325.

Schneider A. (2012) Monitoring Land Cover Change in Urban and Peri-Urban Areas Using Dense Time Stacks of Landsat Satellite Data and a Data Mining Approach. Remote Sensing of Environment 124: 689–704.

Schlerf M., Atzberger C, Hill J. (2005) Remote Sensing of Forest Biophysical Variables Using HyMap Imaging Spec- trometer Data. Remote Sensing of Environment 95:177–194.

Shao Y, Lunetta RS. (2012) Comparison of Support Vector Machine, Neural Network, and CART Algorithms for the Land-Cover Classification Using Limited Training Data Points." ISPRS Journal of Photogrammetry and Remote Sensing 70: 78–87.

Small C, Lu J. (2006). Estimation and vicarious validation of urban vegetation abun- dance by spectral mixture analysis. Remote Sensing of Environment 100: 441–456.

Vapnik V, 1979. Estimation of Dependences Based on Empirical Data. Nauka, Moscow

Zhao H, Ma L, Wang L, Li J (2016). Examining urban expansion using multi-temporal landsat imagery: A case study of montreal census metropolitan area. 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

Zhen Z, LJ Quackenbush, SV Stehman, Zhang L. (2013) Impact of Training and Validation Sample Selection on Classification Accuracy and Accuracy Assessment When Using Reference Polygons in Object-Based Classification. International Journal of Remote Sensing 34: 6914– 6930

Zhu G, Blumberg DG. (2002) Classification using ASTER data and SVM algorithms; The case study of Beer Sheva, Israel. Remote Sensing of Environment 80: 233–240.