AVALIAÇÃO DO IMPACTO DAS AMOSTRAS DE TREINAMENTO NA ACURÁCIA DA CLASSIFICAÇÃO RANDOM FOREST DOS SISTEMAS INTEGRADOS DE PRODUÇÃO AGROPECUÁRIA.

Patrick Calvano Kuchler [1], Margareth Simões [2,3], Rodrigo Ferraz[3], Mateus Benchimol Ferreira de Almeida[1] , Agnès Bégué[3]

[1]Programa de Pós-graduação em Meio Ambiente (PPGMA), Doutorado Interdisciplinar, Universidade do Estado do Rio de Janeiro (UERJ), Rua São Francisco Xavier, 524, Pavilhão João Lyra Filho, 12 Andar, Bloco F, Sala 12005, Bairro Maracanã, RJ, CEP: 20550-013, e-mail: geocalvano@gmail.com; mateusbenchimol@hotmail.com; [2]Departamento de Engenharia de Sistemas e Computação, UERJ/FEN/PPGMA, Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro (UERJ), Bairro Maracanã, RJ,CEP: 20550-013, e-mail: margareth.simoes@embrapa.br; [3]Embrapa Solos, Rua Jardim Botânico, 1024, Jardim Botânico, RJ, e-mail: {margareth.simoes, rodrigo.demonte}@embrapa.br; [4]Maison de la Télédétéction, Montpellier, França, agnes.begue@teledetection.fr

RESUMO

Ao conduzir uma classificação supervisionada com algoritmos de aprendizado de máquina, como o *Random Forest*, a estratégia de balanceamento das amostras é fundamental, pois impacta diretamente nos resultados. Estes classificadores são sensíveis às proporções das amostras de treinamento das diferentes classes. Compreender como estes fatores influenciam na classificação de áreas de produção agropecuária, sobretudo de sistemas minoritários e complexos como o iLP (Integração Lavoura-Pecuária) são de extrema importância para contribuir com metodologias de monitoramento. Para avaliar o impacto do balanceamento, foram testados três grupos de dados de aprendizagem do *Random Forest*: (i) Bset01: dados balanceados entre três classes prioritárias no estado do Mato Grosso; (ii) Bset02: dados desbalanceados com as proporções refletindo a realidade de campo e (iii) Bset03: superestimando a classe rara iLP. Os melhores valores de fscore da classe iLP foram para Bset01 (0,81) e Bset02 (0,83), com um erro de comissão mais alto para Bset01, sugerindo uma melhor performance do Bset02.

Palavras-chave **—** iLP, aprendizado de máquina, dados de treinamento, agricultura de baixa emissão de carbono, séries temporais.

ABSTRACT

Sample balancing strategy is fundamental for a supervised classification with machine learning algorithms, such as Random Forest, as it directly impacts the results. These classifiers are sensitive to the proportions of training samples of different classes. Understanding these factors that influence the classifications of agricultural production areas, especially minority and complex systems such as the iLP (Crop-Livestock Integration) are extremely important to contribute to monitoring methodologies. To assess the impact of this balancing, three groups of learning data from the random forest were tested: (i) Bset01: balanced data between three priority classes in the state of Mato Grosso; (ii) Bset02: unbalanced data with the proportions reflecting the field reality and (iii) Bset03: overestimating the rare class iLP. The best fscore values of the iLP class were for Bset01 (0.81) and Bset02 (0.83), with a higher commission error for Bset01, suggesting a better performance of Bset03.

Key words *— iCL, machine learning, training data, balancing, low carbon agriculture, time series.*

1. INTRODUÇÃO

Atualmente, o aprendizado de máquina e a aprendizagem estatística vêm sendo amplamente utilizados na área de *Big Earth Observation Data (BEOD)* e podem ser divididos em dois principais grupos: Os paramétricos e os não paramétricos [1]. O primeiro grupo utiliza uma quantidade de parâmetros, ou suposições que independem do número de amostras de treinamento, conduzindo a um processamento mais rápido, porém o processo de aprendizagem pode ser mais limitado [2]. No caso dos algoritmos não paramétricos, um número flexível de parâmetros são usados e apresentam a necessidade de uma quantidade maior de amostras para o aprendizado. Como exemplo de modelos não paramétricos largamente utilizados, o *Support Vector Machine* (SVM) e o *Random Forest* (RF), são atualmente muito utilizados em BEOD por apresentarem significativa flexibilidade e a possibilidade de processar um grande volume de dados sem necessidade de conhecimento prévio. O *Random Forest* tem sido largamente utilizado e portanto tem tido significativo destaque [2].

Randon Forest é uma técnica de aprendizado de máquina que gera uma infinidade de árvores de decisão aleatórias que são agregadas, para então gerar uma classificação [3] Cada árvore de classificação é construída de um conjunto amostrado aleatoriamente composto por aproximadamente um terço do conjunto completo de dados, que será chamado no termo em inglês de *bootstraped dataset* [4]. Em estudos de classificação de uso e cobertura da terra, o classificador é considerado estável, além de envolver poucos parâmetros

definidos pelo usuário e mesmo assim, obter bons níveis gerais de acurácia [5].

Também é considerado robusto, fácil de treinar, menos sensível à qualidade dos dados de treinamento e há menos parâmetros para ajustar em comparação com outros classificadores não paramétricos [6,7]. Os autores relatam que a abordagem *Random Forest* aumenta a acurácia da classificação, especialmente, nos casos com alta dimensionalidade de dados como em séries temporais de imagens e sensores hiperespectrais. Também é apontado que os dados de treinamento apresentam grande importância na acurácia, considerando que a baixa quantidade de dados pode causar classificação incorreta.

Além da quantidade, a estratégia de balanceamento das amostras é fundamental, pois impacta diretamente no resultado dos mapas de saída. Estes tipos de classificadores são sensíveis às proporções das diferentes classes, ou seja, de um conjunto de dados balanceado e não balanceado. Dados não balanceados referem-se a uma situação em que o número de observações não é o mesmo para todas as classes em uma base de aprendizagem. Como consequência, esses algoritmos tendem a favorecer a classe com a maior proporção de observações (conhecida como classe majoritária). Isso pode ser particularmente problemático quando o interesse é na classificação de uma classe "rara" (também conhecida como classe minoritária). Dado que esses algoritmos visam minimizar a taxa de erro global, em vez de prestar atenção especial à classe minoritária, eles podem falhar em fazer uma previsão precisa para esta classe se não obtiverem a quantidade necessária de informações sobre ela. Por este motivo, é fundamental realizar uma análise da sensibilidade do classificador à distribuição das amostras de treinamento. Alguns estudos investigaram o desempenho do *Random Forest* para a classificação de imagens de satélite em diferentes estratégias de construção da base de aprendizagem, utilizando amostras balanceadas e não balanceadas. Dalponte et al., e Jin et al. [8,9] encontraram, em seus experimentos, melhores resultados com amostras de treinamento balanceadas, onde cada classe tem a mesma quantidade de amostras. Noi e Kappas [10] encontraram uma relação entre o tamanho da base de aprendizagem e a performance dos conjuntos de dados balanceados e não balanceados. Para o algoritmo *Random Forest* observou-se que ele é extremamente sensível às amostras balanceadas e não balanceadas alcançando maior acurácia global nas amostras balanceadas, porém quando o conjunto de dados é composto por uma quantidade significativa de amostras, a diferença entre os dois é insignificante, concluem os autores. Já Colditz et al. e Mellor et al. [11,12] encontraram uma tendência de melhora na acurácia quando há uma divisão de amostras desbalanceadas entre as classes, de forma que representasse melhor a proporção da área de uso e cobertura da terra.

Considerando as conclusões divergentes da bibliografia existente sobre a influência do balanço de amostras na qualidade da classificação do uso e cobertura da terra, o presente trabalho tem por objetivo a avaliação do impacto do balanceamento de amostras para a classificação de sistemas agropecuários, com ênfase no sistema integrado Lavoura-Pecuária (iLP), considerada uma classe rara dentre as demais. O desenvolvimento de metodologias que possibilitem monitorar os sistemas integrados como o iLP são de grande importância para a avaliação do cumprimento das metas firmadas pelo Brasil na redução de emissões de Gases de Efeito Estufa no setor, sendo inclusive um ponto de destaque do plano ABC (Agricultura de Baixa emissão de Carbono) [13]

2. MATERIAL E MÉTODOS

Foram testados para o estado do Mato Grosso três cenários de classificação para o mapeamento das classes de duplo cultivo: a) Soja+Cereal; b) Soja+Algodão e c) Sistema Integrado Lavoura Pecuária para o ano-safra 2016/2017. A primeira classificação foi realizada por um conjunto de dados equilibrado (Bset01), ou seja, com um número de amostras semelhantes em cada classe. Outras duas classificações foram realizadas aplicando conjuntos de dados não balanceados, um com uma distribuição de amostras por classe próxima à distribuição real encontrada no campo (estimado a partir do conjunto de dados coletado) (Bset02). O outro conjunto de dados não balanceados, foi composto por uma super-representação da classe rara iLP (Bset03) (Figura 01).

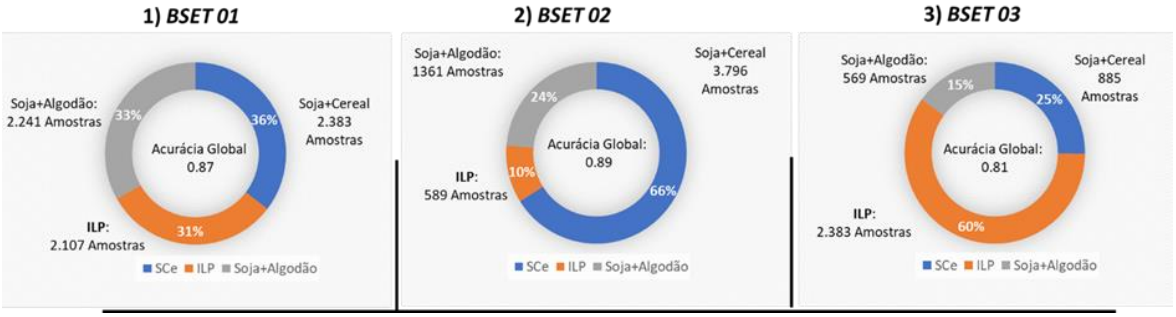

Figura 1. Composição dos três conjuntos de dados de treinamento testados (Bset01, Bset02 e Bset03), e detalhes associados às classes "soja + algodão" (soja + algodão), "soja + cereais" (SCe) e sistemas integrados (iLP).

O classificador *Random Forest* foi usado com 100 árvores aleatórias, combinadas com um conjunto de dados de aprendizagem/validação de aproximadamente 6.700 pixels disponíveis para a composição dos 3 cenários de balanceamento. Os dados de campo foram levantados de diferentes formas: a) por GPS *in situ*, b) coleta por entrevistas com produtores, c) dados de cooperativas locais. Durante as entrevistas com produtores, foram coletados dados sobre o histórico das práticas de campo e sobre o histórico das práticas de cultivos. Uma parceria importante para uma coleta significativa de amostras foi com o grupo Bom Futuro, um dos maiores grupos de produção agrícola do Brasil.

Foi adotada uma abordagem de classificação hierárquica [14], partindo da classe soja do produto anual do Mapbiomas coleção 5 [15].O processamento foi realizado na plataforma *Google Earth Engine* (GEE), que fornece fácil acesso a produtos prontos para uso e grandes volumes de dados, o que possibilitou processar 592 imagens do produto MOD13Q1 (23 datas de NDVI, EVI, NIR e MIR; 6 cenas para cobrir o Mato Grosso).

3. RESULTADOS

A tabela 1 apresenta as métricas de acurácia para os três mapas produzidos. É possível observar que dentre os sistemas sequenciais, a classe soja + algodão é a que apresenta menor variação condicionada ao conjunto de dados de treinamento.

Métricas	***BSET 01***			***BESET 02***			***BESET 03***		
	Soja Algodão	**Soja Cereal**	**iLP**	**Soja Algodão**	**Soja Cereal**	**iLP**	**Soja Algodão**	**Soja Cereal**	**iLP**
Acurácia Produtor	1	0,81	0,80	1	0,91	0,76	0,99	0,54	0,88
Acurácia Usuário	1	0,80	0,81	1	0,79	0,90	1	0,82	0,66
F-SCORE	1	0,96	0,81	1	0,85	0,83	1	0,65	0,75
Amostras	2.241	2.383	2.107	1.361	3.796	589	2.383	885	2.383
Acurácia Global	**0,87**			**0,89**			**0,81**		

Tabela 1. Métricas dos resultados de acurácia dos conjuntos balanceados e não balanceados associados às classes "soja + algodão" (soja + algodão), "soja + cereais" (SCe) e sistemas integrados (iLP).

Por outro lado, a classe cuja precisão apresenta maior variação é a soja + cereais, com valores de fscore entre 0,65 e 0,96. Os melhores valores de fscore da classe iLP são obtidas para Bset01 e Bset02 (0,81 e 0,83 respectivamente), com um erro de comissão mais alto para Bset01 (+ 9%). No Bset03, onde há uma superestimação da classe iLP, o valor de *fscore* se apresentou mais baixo, condicionado por uma acurácia do usuário (0,66) extremamente baixa. A elevada acurácia do produtor (0,88 maior entre todos os conjuntos), não foi suficiente para elevar o valor do *fscore*.

4. DISCUSSÃO

Os Sistemas iLP abrangem diferentes soluções produtivas com vários arranjos espaços-temporais de atividades agrícolas e pastoris. São sistemas complexos, difíceis de mapear por sensoriamento remoto [16–18]. Este estudo, objetivou compreender o impacto do desenho amostral no mapeamento dos sistemas integrados, uma classe rara, no classificador *Random Forest*. Pelos 3 diferentes conjuntos de dados de treinamento testados, descobrimos que o conjunto de dados balanceado, com uma composição semelhante à realidade, apresenta os melhores resultados de classificação para a classe iLP e demais. Este resultado está de acordo com Colditz et al. [11] Mellor et al. [12]. No entanto, se mostra divergente com os resultados encontrados em outros estudos [9,10]. Esses resultados opostos encontrados na bibliografia podem estar relacionados à diversidade dos dados de treinamento e até mesmo em relação à distribuição e proporcionalidade das classes minoritárias. A intepretação destes resultados sugere que, quando uma maior quantidade de amostras é conferida a classe iLP em uma abordagem não balanceada, o *Random Forest* vai supervalorizar esta classe e vai apresentar um viés com maior erro de comissão, ou seja, mais áreas que não são o iLP, serão classificadas como tal de forma errada. Neste caso, teríamos uma super estimativa de área implementada com iLP no estado do Mato Grosso, ao passo que teríamos uma área subestimada de soja + cereal com um erro de omissão de 0,46. Considerando o desenvolvimento de uma metodologia para o mapeamento da classe iLP, a omissão de áreas destes sistemas pode ser considerado como uma abordagem conservadora, mais adequada do que superestimar áreas com iLP.

5. CONCLUSÕES

Frente a estes resultados, considerando a maior acurácia global, maior fscore e menor erro de comissão (objetivando estimar áreas de forma conservadora) para a classe de interesse iLP, o conjunto de dados Bset02 (em que o número

de amostras por classe é representativo da proporção de classes no campo). Neste trabalho, testamos aspectos fundamentais na classificação de de sistemas agropecuários complexos utilizando séries temporais MODIS no modelo Random Forest. A coleta dos dados de campo dos sistemas com alto dinamismo dentro do ano-safra se reflete como um grande desafio para a coleta, necessitando abordagens otimizadas e com maior acurácia. Os resultados mostram que a série de resultados com uma base de aprendizagem balanceada foram capazes de detectar os sistemas iLP contribuindo para o desenvolvimento de uma metodologia de mapeamento destes sistemas de forma reprodutível, possibilitando a operacionalização de uma ferramenta para o monitoramento da adoção da intensificação sustentável do uso da terra, preconizado no plano ABC.

6. AGRADECIMENTOS

Esse trabalho foi realizado no ambito do projeto CAPES-COFECUB GeoABC e do projeto Europeu H2020-MSCA-RISE-2015 ODYSSEA project (Project Reference: 691053).

6. REFERÊNCIAS

[1] Holloway, J.; Mengersen, K. Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review. *Remote Sensing* **2018**, *10*, 1365, doi:10.3390/rs10091365.

[2] Tamiminia, H.; Salehi, B.; Mahdianpari, M.; Quackenbush, L.; Adeli, S.; Brisco, B. Google Earth Engine for Geo-Big Data Applications: A Meta-Analysis and Systematic Review. *ISPRS Journal of Photogrammetry and Remote Sensing* **2020**, *164*, 152–170, doi:10.1016/j.isprsjprs.2020.04.001.

[3] Breiman, L. Random Forest. *Springer Link* **2001**, *45*, 5–32, doi:10.1023/a:1010933404324.

[4] Cutler, D.; C Edwards, T.; Beard, K.; Cutler, A.; T Hess, K.; Gibson, J.; Lawler, J. Random Forests for Classification in Ecology. *Ecology* **2007**, *88*, 2783–2792, doi:10.1890/07-0539.1.

[5] Lawrence, R.L.; Wood, S.D.; Sheley, R.L. Mapping Invasive Plants Using Hyperspectral Imagery and Breiman Cutler Classifications (RandomForest). *Remote Sensing of Environment* **2006**, *100*, 356–362, doi:10.1016/j.rse.2005.10.014.

[6] Belgiu, M.; Csillik, O. Sentinel-2 Cropland Mapping Using Pixel-Based and Object-Based Time-Weighted Dynamic Time Warping Analysis. *Remote Sensing of Environment* **2018**, *204*, 509–523, doi:10.1016/j.rse.2017.10.005.

[7] Mahdianpari, M.; Salehi, B.; Mohammadimanesh, F.; Brisco, B.; Homayouni, S.; Gill, E.; DeLancey, E.R.; Bourgeau-Chavez, L. Big Data for a Big Country: The First Generation of Canadian Wetland Inventory Map at a Spatial Resolution of 10-m Using Sentinel-1 and Sentinel-2 Data on the Google Earth Engine Cloud Computing Platform. *Canadian Journal of Remote Sensing* **2020**, *46*, 15–33, doi:10.1080/07038992.2019.1711366.

[8] Dalponte, M.; Orka, H.O.; Gobakken, T.; Gianelle, D.; Naesset, E. Tree Species Classification in Boreal Forests With Hyperspectral Data. *IEEE Transactions on Geoscience and Remote Sensing* **2013**, *51*, 2632–2645, doi:10.1109/tgrs.2012.2216272.

[9] Jin, H.; Stehman, S.V.; Mountrakis, G. Assessing the Impact of Training Sample Selection on Accuracy of an Urban Classification: A Case Study in Denver, Colorado. *International Journal of Remote Sensing* **2014**, *35*, 2067–2081, doi:10.1080/01431161.2014.885152.

[10] Noi, P.T.; Kappas, M. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors* **2017**, *18*, 18, doi:10.3390/s18010018.

[11] Colditz, R. An Evaluation of Different Training Sample Allocation Schemes for Discrete and Continuous Land Cover Classification Using Decision Tree-Based Algorithms. *Remote Sensing* **2015**, *7*, 9655–9681, doi:10.3390/rs70809655.

[12] Mellor, A.; Boukir, S.; Haywood, A.; Jones, S. Exploring Issues of Training Data Imbalance and Mislabelling on Random Forest Performance for Large Area Land Cover Classification Using the Ensemble Margin. *ISPRS Journal of Photogrammetry and Remote Sensing* **2015**, *105*, 155–168, doi:10.1016/j.isprsjprs.2015.03.014.

[13] Kuchler, P.C.; Simões, M.; Begué, A.; Peçanha, R.; Arvor, D. SENSORIAMENTO REMOTO E ANÁLISE ESPACIAL: UMA CONTRIBUIÇÃO PARA O MAPEAMENTO DOS SISTEMAS INTEGRADOS DE PRODUÇÃO AGROPECUÁRIA. *Aplicações e Princípios do Sensoriamento Remoto 3* **2019**, 1–10, doi:https://doi.org/10.22533/at.ed.3791923091.

[14] Lebourgeois, V.; Dupuy, S.; Vintrou, É.; Ameline, M.; Butler, S.; Bégué, A. A Combined Random Forest and OBIA Classification Scheme for Mapping Smallholder Agriculture at Different Nomenclature Levels Using Multisource Data (Simulated Sentinel-2 Time Series, VHRS and DEM). *Remote Sensing* **2017**, *9*, 259, doi:10.3390/rs9030259.

[15] Souza, C.M.; Shimbo, J.Z.; Rosa, M.R.; Parente, L.L.; Alencar, A.A.; Rudorff, B.F.T.; Hasenack, H.; Matsumoto, M.; Ferreira, L.G.; Souza-Filho, P.W.M.; et al. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. *Remote Sensing* **2020**, *12*, 2735, doi:10.3390/rs12172735.

[16] Kuchler, P.C.; Simões, M.; Ferraz, R.; Arvor, D.; de Almeida Machado, P.L.O.; Rosa, M.; Gaetano, R.; Bégué, A. Monitoring Complex Integrated Crop–Livestock Systems at Regional Scale in Brazil: A Big Earth Observation Data Approach. *Remote Sensing* **2022**, *14*, doi:10.3390/rs14071648.

[17] Kuchler, P.C.; Bégué, A.; Simões, M.; Gaetano, R.; Arvor, D.; Ferraz, R.P.D. Assessing the Optimal Preprocessing Steps of MODIS Time Series to Map Cropping Systems in Mato Grosso, Brazil. *International Journal of Applied Earth Observation and Geoinformation* **2020**, *92*, 102150, doi:10.1016/j.jag.2020.102150.

[18] Manabe, V.D.; Melo, M.R.S.; Rocha, J.V. Framework for Mapping Integrated Crop-Livestock Systems in Mato Grosso, Brazil. *Remote Sensing* **2018**, *10*, 1322, doi:10.3390/rs10091322.