



ESTIMATIVA DE BIOMASSA DE PASTAGEM UTILIZANDO SÉRIES TEMPORAIS DE IMAGENS DE SENSORIAMENTO REMOTO E APRENDIZAGEM DE MÁQUINA

Vitor H. Nisieimon¹, Aliny A. Dos Reis², Gleyce K. D. A. Figueiredo³

¹Faculdade de Engenharia Agrícola– FEAGRI, Universidade Estadual de Campinas – UNICAMP, 13083-875 Campinas, Brasil
v178454@dac.unicamp.br¹, aareis@unicamp.br², gleyce@unicamp.br³

RESUMO:

Medições precisas e objetivas de biomassa em pastagens constituem uma etapa fundamental para a melhoria da produtividade do sistema de pastejo, permitindo assim que pecuaristas consigam suprir melhor as necessidades de alimentação de seus animais. Devido às características sinópticas, multiespectrais e de revisita dos satélites, as imagens de sensoriamento remoto permitem a obtenção de informações fundamentais para a caracterização espacial e temporal das pastagens, assim como a estimativa de biomassa ao longo do desenvolvimento da forragem utilizando séries temporais. No entanto, dados de sensoriamento remoto permitem gerar grande quantidade de variáveis preditoras (como por exemplo, diversos índices de vegetação), o que torna a modelagem da estimativa de biomassa um desafio. Diante deste cenário, esse projeto teve como objetivo testar três métodos de seleção de atributos (Algoritmo Genético - GA, Random Forest - RF e Relief) e estimar a biomassa de pastagem em sistema de integração lavoura-pecuária utilizando índices de vegetação derivados de imagens Sentinel-2A e aprendizagem de máquina (Random Forest). Foram selecionados 104 índices de vegetação para serem utilizados como variáveis preditoras nesse estudo. Utilizando-se como referência o Erro Absoluto Médio (*Mean Absolute Error - MAE*) comparou-se as 3 técnicas de seleção de atributos, as quais apresentaram bons resultados. A GA foi a que obteve o menor MAE, 24.55 g/m², seguida da RF com 26.46 g/m², a Relief com 28.84 g/m² e, por fim, um MAE de 29.87 g/m² quando utilizado todos os 104 índices de vegetação extraídos das imagens disponibilizadas pelo Sentinel-2. Além disso, a técnica de modelagem Random Forest apresentou grande potencial no entendimento de padrões da realidade mesmo com diferentes tipos de atributos, evidenciando assim a sua capacidade em gerar modelos mais generalistas e que se adaptam a diferentes problemas.

PALAVRAS-CHAVE: *Sistemas integrados, Sentinel-2, Seleção de atributos, Aprendizagem de máquina, Índices vegetativo.*

INTRODUÇÃO

Medições precisas e objetivas de biomassa em pastagens constituem uma etapa fundamental para a melhoria da produtividade do sistema de pastejo por partes dos produtores, permitindo assim que os pecuaristas consigam suprir melhor às necessidades de alimentação de seus animais (ANDERSSON et al., 2017). Assim, estimativas de biomassa em tempo real contribuem para a otimização das estratégias de pastejo, permitindo que os produtores adequem as taxas de lotação por piquete em função da real produtividade daquela área, resultando no aumento da utilização das pastagens, e subsequentemente as taxas de crescimento das pastagens junto com a produtividade geral (SCHAEFER; LAMB, 2016).

As técnicas de sensoriamento remoto têm sido fundamentais por fornecer informações valiosas para a avaliação das condições da vegetação, auxiliando em tomadas de decisões que visam a melhoria dos sistemas produtivos locais (MANABE; MELO; ROCHA, 2018).

Devido às características sinópticas, multiespectrais e de revisita dos satélites, estes permitem a obtenção de informações fundamentais para a caracterização espacial e temporal das pastagens, assim como a estimativa de biomassa ao longo do desenvolvimento da forragem utilizando séries temporais de imagens de sensoriamento remoto (BARRACHINA; CRISTÓBAL; TULLA, 2015; MICHEZ et al., 2019; OTGONBAYAR et al., 2018).

O monitoramento do crescimento das pastagens via sensoriamento remoto requer imagens coletadas com uma alta resolução temporal. No entanto, o processo de estimativa de biomassa de pastagens utilizando séries temporais não é um processo simples, uma vez que envolve diversos fatores, como experiência do intérprete, conhecimento das áreas de estudo, da assinatura temporal dos alvos, e também da metodologia utilizada (MANABE; MELO; ROCHA, 2018). Atualmente, os métodos mais avançados de estimativa de biomassa utilizando imagens de sensoriamento remoto fazem uso de algoritmos de aprendizagem de máquina



(OTGONBAYAR et al., 2018). No entanto, ainda existem limitações envolvidas ao utilizá-los para a estimativa de biomassa levando em consideração a variação espaço-temporal das pastagens, como por exemplo, a enorme quantidade de variáveis preditoras geradas para o treinamento dos modelos de inteligência artificial.

A crescente na dimensionalidade de dados disponíveis, no entanto, traz junto um custo computacional elevado, já que torna a procura de padrões e correlações dos modelos estatísticos muito mais complexa. Apresenta-se assim como um grande desafio encontrar quais são os dados mais relevantes para a estimativa da biomassa dada a elevada capacidade espectral do Sentinel-2. Os métodos de seleção de atributos aparecem como técnicas de mineração de dados para a escolha de subconjuntos de dados mais correlacionados e relevantes à modelagem e que buscam portanto a melhora no desempenho de modelos como demonstrado em Rasel et al. (2019).

Nesse sentido, este estudo teve como objetivo avaliar como diferentes métodos de seleção de atributos afetam a estimativa de biomassa em pastagem utilizando índices de vegetação derivados de imagens Sentinel-2, bem como avaliar o algoritmo Random Forest de modelagem.

METODOLOGIA

A área de estudo corresponde a quatro talhões de integração lavoura-pecuária (sucessão Soja – Milheto + *Brachiaria*), submetidos a diferentes níveis de pastejo, com aproximadamente 50 hectares cada, localizados na Fazenda Campina, no município de Caiuá – SP (Figura 1). Essa fazenda possui extensas áreas com sistemas de integração lavoura-pecuária.

Dados de biomassa de pastagem foram coletados no campo entre Maio e Novembro de 2019 utilizando como base uma grade de 100 pontos amostrais, distribuídos sistematicamente na área de estudo (Figura 1). As coletas de biomassa em campo ocorreram nas seguintes datas: 17 e 25 de Maio, 18 de Junho, 14 de Julho, 12 de Agosto e 02 de Novembro. O número de pontos amostrais coletados em cada medição variou em função da presença de gados nos piquetes, totalizando ao final 346 pontos amostrados. Em cada ponto amostral, a biomassa fresca da pastagem foi coletada utilizando um esquadro de 1m² e pesada em campo utilizando uma balança de mão. Para a determinação do peso seco, a massa fresca da pastagem foi seca em laboratório por 72 horas, e pesadas novamente para a obtenção da biomassa seca (g/m²) por ponto amostral e medição.

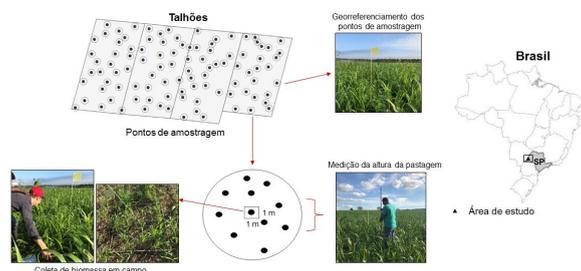


Figura 1. Localização da área de estudo e distribuição dos pontos amostrais coletados em campo.

Em paralelo, foi organizado um banco de dados com imagens multiespectrais do satélite Sentinel-2, as quais foram adquiridas no banco de dados da ESA - *European Space Agency* (<https://scihub.copernicus.eu/dhus/#/home>) para datas próximas das datas de coleta dos dados em campo. O satélite Sentinel-2 apresenta um sensor multiespectral de alta e média resolução espacial (10, 20 e 60 m). Seu sensor *MultiSpectral Instrument* (MSI) adquire imagens com resolução temporal de cinco dias, considerando seus dois satélites 2A e 2B, e possui treze bandas espectrais que abrangem comprimentos de onda entre 0,4 a 2,2 μm como demonstrado na Tabela 1.

Tabela 1. Características espectrais das imagens obtidas pelo Sentinel-2.

Resolução	Nº da Banda	Código	Nome da Banda	Comprimento de onda central (nm)
10m	B02	b	Blue (Azul)	490
	B03	g	Green (Verde)	560
	B04	r	Red (Vermelho)	665
	B08	nir	NIR (Infravermelho Próximo)	842
20m	B05	re1	Red Edge 1	705
	B06	re4	Red Edge 2	740
	B07	re3	Red Edge 3	783
	B08A	re4	Red Edge 4	865
	B11	swir1	SWIR 1	1610
	B12	swir2	SWIR 2	2190
60m	B01	a	Aerosol	443
	B09	-	Water Vapor	940
	B10	-	Citrus	1375

Fonte. ESA.



Em sequência, foi realizada a correção atmosférica das imagens e, posteriormente, a obtenção dos valores de reflectância na superfície nas diferentes bandas espectrais. Para isso utilizou-se o pacote 'sen2r' durante todo o processo, desde o download das imagens até o processamento. O ideal seria que as imagens coletadas fossem obtidas na mesma data da coleta de biomassa em campo, porém devido à resolução temporal de 5 dias do sensor Sentinel-2, isso não foi possível e por isso selecionou-se as imagens com a data mais próxima para uma melhor aproximação do modelo com relação à realidade, conforme mostrado na Tabela 2.

Tabela 2. Relação temporal entre as imagens do sensor e o período de coleta em campo.

Banco de Imagens (2020)	
Data de coleta	S2A Images
17 mai	✓ (21 mai)
25 mai	✓ (28 mai)
18 jun	✓ (17 jun)
14 jul	✓ (17 jul)
18 ago	✓ (19 ago)
02 nov	✓ (02 nov)

Após o download e processamento de todas as imagens, selecionou-se um total de 104 índices, todos disponibilizados no Index DataBase (IDB), <https://www.indexdatabase.de/db/s-single.php?id=96>, e assim elaborou-se as 104 respectivas funções na linguagem R para cálculo de cada um, de acordo com a fórmula também disponibilizada no IDB.

Com todas as funções prontas, foi possível aplicá-las nas imagens pixel a pixel para obtenção dos índices. Porém, além da correspondência temporal feita nos downloads, foi preciso fazer uma correspondência espacial entre os pixels das imagens e o ponto exato de coleta em campo, para que a modelagem fosse também precisa quanto à localização. Na coleta de campo, por meio de um aparelho GPS, obteve-se as coordenadas de cada ponto amostrado, possibilitando a extração dos valores dos índices de vegetação dos pontos corretamente.

Em seguida, foi realizada a seleção de atributos utilizando 2 técnicas por filtragem: Relief e *Random Forest Importance* (RF), além do Algoritmo Genético (GA), o qual utiliza-se da metodologia *Wrapper* de seleção. Com base nas variáveis preditoras extraídas das imagens do Sentinel-2 e previamente selecionadas foi realizada a modelagem e estimativa de biomassa da pastagem utilizando o algoritmo de aprendizagem de máquina: *Random Forest*. A RF foi proposta por Breiman (2001), e corresponde a um algoritmo de aprendizagem que utiliza o método ensemble em suas camadas finais da regressão em árvore. É um método que pode ser utilizado tanto para regressão quanto para

classificação (BREIMAN, 2001). O algoritmo utiliza de uma floresta de árvores formadas através de entradas aleatórias, com ou sem reposição nas mesmas e, por fim, a combinação dessas árvores por meio da média.

Para validação da eficácia das técnicas de seleção de atributos, realizou-se ainda uma modelagem com os 104 atributos selecionados previamente do IDB totalizando, assim, 4 banco de dados diferentes: os 3 provenientes da seleção de atributos, RF, Relief e GA, bem como esse último, no qual fora incluído todos atributos e nomeado de *All*. Utilizou-se uma divisão de 80% dos dados para treinamento e 20% para teste dos modelos. Todo este processo pode ser visualizado no fluxograma da Figura 2.

Por fim, como métricas de avaliação da estimativa de biomassa das pastagens foram utilizados o erro absoluto médio (MAE – *Mean Absolute Error*) e a raiz do erro quadrático médio (RMSE – *Root Mean Square Error*). O processamento das imagens, correção atmosférica e extração dos valores de reflectância, assim como a modelagem, predição da estimativa de biomassa e cálculo de erros foram realizados utilizando a linguagem de programação R (R CORE TEAM, 2019).

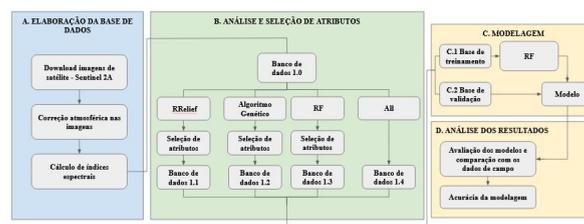


Figura 2. Fluxograma das etapas metodológicas seguidas neste estudo.

RESULTADOS E DISCUSSÃO

A partir da coleta de biomassa em campo e o banco estruturado, foi possível extrair alguns dados estatísticos dos valores de biomassa observados em campo, como demonstrado na Tabela 3. Ressalta-se o valor médio de biomassa, quando analisado todos os 346 pontos coletados, de 164,82 g/m², o desvio padrão de 84,80 g/m² com os valores mínimo e máximo de coleta de 24,04 e 656,08 g/m², respectivamente.



Tabela 3. Dados de biomassa coletados em campo

Mês	Nº pontos amostrados	Média biomassa (g/m ²)	Desvio padrão biomassa (g/m ²)
Maio	100	207,68	101,50
Junho	50	142,17	65,84
Julho	100	110,71	54,44
Agosto	38	163,58	52,58
Novembro	58	204,57	65,08
Total	346	164,82	84,80

Na etapa de seleção de atributos, obteve-se como resultado para as técnicas de filtragem, RF e Relief, as distribuições de importância dos índices demonstradas na Figura 3. Os resultados do algoritmo genético, por sua vez, não são ranqueados de acordo com a importância dos atributos, uma vez que trata-se da metodologia *Wrapper* e, portanto, apenas determinou-se qual o subconjunto de atributos mais adequado para se realizar a modelagem, sem demonstração gráfica.

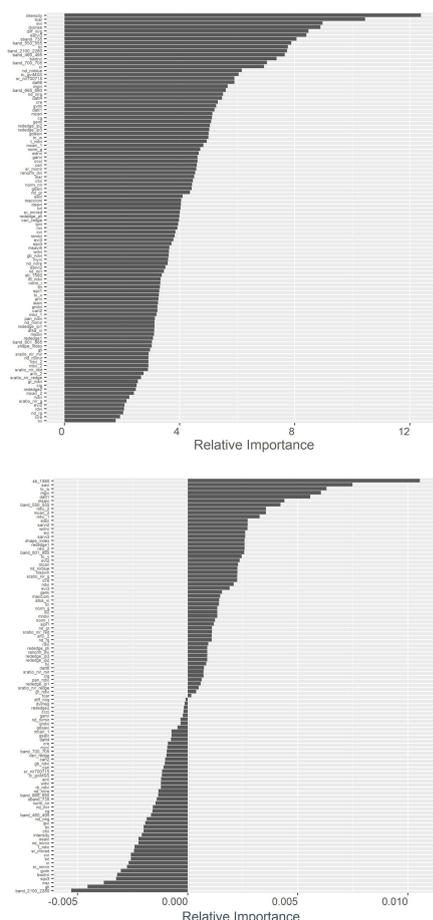


Figura 3. Importância dos atributos em RF (cima) e Relief (baixo), respectivamente.

Os resultados dos erros MAE e RMSE no conjunto de teste obtidos a partir da modelagem com o algoritmo RF podem ser observados na Tabela 4.

Tabela 4. Resultado dos 4 modelos treinado com RF e avaliados com MAE (g/m²) e RMSE (g/m²).

	GA	RF	Relief	All
MAE	24,55	26,46	28,84	29,87
RMSE	35,90	38,56	40,03	43,38

A seleção de atributos se mostrou uma técnica eficaz na melhora de desempenho do modelo, uma vez que selecionou um subconjunto de dados com atributos de maior correlação com a variável de interesse (biomassa). As modelagens que tiveram um algoritmo de seleção de atributos aplicado previamente (GA, RF e Relief) obtiveram erros menores em comparação com o modelo no qual utilizou-se todos os atributos disponibilizados (All).

Entre as 3 técnicas utilizadas na seleção de atributos, o Algoritmo Genético foi o que obteve o melhor resultado, seguido da seleção pelo próprio algoritmo da Random Forest e, por fim, o Relief. A metodologia *Wrapper* do GA, como esperado, foi a que demandou maior tempo de processamento justificando o seu resultado com o menor erro. Porém analisando-se a margem de erro entre as técnicas, verifica-se que a metodologia por filtro (Relief e RF), mesmo com um custo de processamento menor, obteve resultados bem próximos ao do GA.

Além disso, outro ponto a ser destacado é a elevada capacidade espectral do Sentinel-2, o qual proporcionou uma grande quantidade de dados disponíveis como base para cálculo dos índices utilizados, já que cobre diversas regiões do espectro, desde a banda azul (b) até infravermelho médio (swir1 e swir2). No entanto, algumas regiões ganham destaque, já que as bandas que as cobrem foram bastante utilizadas no cálculo dos índices, como é o caso da região do visível e do infravermelho próximo, onde encontram-se as bandas re1, re2, re3, re4 e nir, respectivamente.

Por fim, com relação à modelagem a partir da RF, verifica-se os resultados obtidos a partir das métricas MAE e RMSE na Tabela 4. Mesmo com o maior erro sendo com a utilização de todos os atributos (All), devido à alta complexidade de padrão e correlação entre si, a RF conseguiu se adaptar aos diferentes cenários, gerando erros próximos tanto para o MAE quanto para RMSE nas 4 diferentes modelagens, o que mostra a sua capacidade de generalização.



CONCLUSÃO

O Sentinel-2 com sua alta capacidade espectral possibilitou a geração de 104 atributos para a modelagem da biomassa de pastagem, além de disponibilizar imagens próximas às datas de coleta em campo e de boa qualidade espacial. A elevada quantidade de dados gerada pelo sensor, proporcionou uma boa base para que a modelagem da biomassa fosse realizada e com resultados coerentes com a realidade, o que demonstra todo o potencial do Sentinel-2 no fornecimento de dados para a modelagem da biomassa de pastagem.

As técnicas de seleção de atributos mostraram-se eficazes selecionando subconjuntos de dados mais correlacionados com a biomassa e aumentando portanto a acurácia do modelo. A determinação de qual a melhor técnica, no entanto, fica a critério do responsável pela modelagem uma vez que os resultados entre as 3 técnicas estudadas foram bem próximos, se diferenciando apenas pela metodologia de seleção e também pelo custo de processamento.

A RF mostrou-se, como esperado, uma técnica de modelagem que consegue se adaptar a diferentes tipos de dados gerando modelos relativamente menos complexos e mais generalistas. Mesmo com quantidades e tipos de atributos diferentes utilizados nas 4 modelagens feitas neste trabalho, a técnica gerou erros bem próximos entre si, mostrando portanto sua capacidade de aplicação em diversos cenários da realidade.

APOIO

Este trabalho foi realizado com o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) e da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (2017/50205-9), além do apoio em infra-estrutura do laboratório de Geoprocessamento da Faculdade de Engenharia Agrícola (FEAGRI) da Universidade Estadual de Campinas (UNICAMP), e da colaboração do dono e gerente da Fazenda Campina (Nelore CV) na coleta dos dados em campo.

BIBLIOGRAFIA

- ANDERSSON, K. et al. Estimating pasture biomass with active optical sensors. **Advances in Animal Biosciences**, v. 8, n. 02, p. 754–757, 2017.
- BARRACHINA, M.; CRISTÓBAL, J.; TULLA, A. F. Estimating above-ground biomass on mountain meadows and pastures through remote sensing. **International Journal of Applied Earth Observation and Geoinformation**, v. 38, n. 5, p. 184–192, jun. 2015.
- ESA - Sentinel Online. Radiometric Resolutions. Página Disponível em: <<https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/radiometric>> Acesso em: 27 de set. de 2020
- ESA - Sentinel Online. Spatial Resolutions. Página Disponível em: <<https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/spatial>> Acesso em: 27 de set. de 2020
- MANABE, V.; MELO, M.; ROCHA, J. Framework for Mapping Integrated Crop-Livestock Systems in Mato Grosso, Brazil. **Remote Sensing**, v. 10, n. 9, p. 1322, 21 ago. 2018.
- OTGONBAYAR, M. et al. Mapping pasture biomass in Mongolia using Partial Least Squares, Random Forest regression and Landsat 8 imagery. **International Journal of Remote Sensing**, v. 1161, 2018.
- RASEL, M. M. S. et al. Application of feature selection methods and machine learning algorithms for saltmarsh biomass estimation using Worldview-2 imagery. **Geocarto International**, 2019.
- R CORE TEAM. **R: a language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing. 2018.
- SCHAEFER, M. T.; LAMB, D. W. A combination of plant NDVI and LiDAR measurements improve the estimation of pasture biomass in tall fescue (*festuca arundinacea* var. *fletcher*). **Remote Sensing**, v. 8, n. 2, 2016.