Estimativa da troca líquida de carbono a partir dos produtos MODIS e dados meteorológicos aplicados a modelos de aprendizado de máquina

Aline A. Nascimento[1], Lucas O. Bauer[1], Alan J. P. Calheiros[1], Luciana V. Rizzo[2]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos, SP

[2]Instituto de Física – Universidade de São Paulo (USP)
São Paulo, SP

{aline.andrade,lucas.bauer,alan.calheiros}@inpe.br, lrizzo@usp.br

Abstract. *The Fluxcom project employed machine learning and surface data to estimate the global carbon balance. Nevertheless, these estimates are less accurate in tropical regions, including the Amazon. This study focuses on estimating the net ecosystem exchange (NEE) in a 0.25° cell located at the K67 Tower in the Tapajós National Forest, Santarém. We use data from the ERA-5 reanalysis model, MODIS products, and the BrSa1 Fluxnet tower. The data from the former sets were used to train machine learning models, while the latter served as the target (NEE) for estimating its time series from 2002 to 2011. The estimation results closely matched those of the Fluxcom project.*

Resumo. *O projeto Fluxcom usou machine learning e dados de superfície para estimar o balanço de carbono global, porém as estimativas são menos precisas em regiões tropicais, incluindo a Amazônia. Este trabalho foca na estimativa da troca líquida de carbono (NEE) em uma célula de 0.25° junto à Torre K67, na Floresta Nacional dos Tapajós, Santarém. Utilizaram-se dados do modelo de reanálise ERA-5, produtos do MODIS e da torre de fluxo BrSa1 da Fluxnet. Os dados dos primeiros conjuntos foram usados para treinar modelos de machine learning, enquanto os últimos foram utilizados como alvo (NEE) para estimar sua série temporal de 2002 a 2011. Os resultados da estimativa se aproximaram dos da Fluxcom.*

1. Introdução

As emissões crescentes de gases de efeito estufa, resultantes de atividades humanas como queimadas e desmatamento, causam desequilíbrios climáticos e eventos extremos, ameaçando a vida e os ecossistemas. O CO_2 desempenha um papel fundamental na problemática das mudanças climáticas e sua remoção da atmosfera por ecossistemas terrestres pode contribuir para mitigar as emissões de gases de efeito estufa no Brasil e no mundo [Baldocchi 2003].

A partir da análise da troca líquida de carbono, expressa pela variável NEE (Net Ecosystem Exchange), é possível identificar áreas que atuam como fontes e sumidouros de CO_2 e aplicar ações de mitigação onde necessário. No entanto, as medições da NEE, obtidas por torres de fluxo, possuem representatividade espacial local, tornando difícil

identificar padrões espaciais no comportamento dos fluxos de CO_2. Nesse contexto, surge a iniciativa Fluxcom com o uso de técnicas de *machine learning* (ML) com a integração de diversas fontes de dados terrestres, para promover a estimativa de três variáveis fundamentais para o estudo de fluxos de carbono na interface entre a biosfera e a atmosfera, que são as variáveis da equação $NEE = -GPP + R_e$ [Tramontana and Jung 2016]. NEE é a variável Net Ecosystem Exchange, onde GPP representa a produção primária bruta, relacionada à absorção de carbono por fotossíntese, e R_e é a respiração do ecossistema, envolvendo a emissão de carbono por processos autotróficos e heterotróficos [Baldocchi 2003]. A iniciativa Fluxcom empregou dados de modelos de reanálise, como ERA-5, e produtos de sensores MODIS como preditores. Eles usaram a variável NEE do conjunto de dados global de torres de fluxo, Fluxnet, como alvo. Dessa forma, realizaram a estimativa do balanço de carbono global, empregando apenas dados espacializados de superfície [Tramontana and Jung 2016, Jung 2020].

A Fluxcom alcançou resultados notáveis em regiões como os Estados Unidos e a Europa, atingindo uma métrica estatística R^2 de até 0.99. No entanto, enfrentou desafios ao obter resultados satisfatórios em regiões tropicais, incluindo a designada ”América do Sul tropical”, que abrange a região Amazônica. O maior coeficiente de determinação encontrado para a área pela Fluxcom foi, a partir dados de sensoriamento remoto, 0.1 e a partir dos dados meteorológicos foi de 0.33 [Jung 2020]. Os modelos utilizados pela Fluxcom obtiveram os piores resultados para as regiões trópicais e com tipo funcional de planta EBF (Floresta perene de folhas largas) [Tramontana and Jung 2016].

Portanto, dados os resultados na região Amazônica e o reconhecimento da importância do entendimento do balanço de carbono nessa região, este estudo se propôs a utilizar dados semelhantes aos empregados pela Fluxcom e implementar outras técnicas de aprendizagem de máquina para promover melhores estimativas de NEE (Net Ecosystem Exchange) e identificar se as técnicas de ML que foram escolhidas e o ”ajuste fino”de hiperparâmetros especificamente para a região Amazônica poderia levar a um resultado melhor do que o da Fluxcom. A principal meta foi estimar o fluxo líquido de dióxido de carbono (NEE) para uma célula de 0.25°, localizada na região da Torre K67, conhecida como BrSa1 nos dados do Fluxnet. Essa torre está posicionada na Floresta Nacional dos Tapajós, em Santarém.

2. Material e Métodos

Inicialmente, optamos pela grade de 0,25° do ERA-5 para a integração dos dados. Em seguida, foram extraídos os dados horários do período de 2002 a 2011 do ERA-5 através da API disponibilizada pelo Copernicus [Hersbach and Bell 2023], seguido pelo cálculo das médias diárias de cada uma das variáveis.

Foi efetuada a extração da geometria da célula da grade do ERA5 e essa foi usada para adquirir dados dos produtos MODIS através da API Python do Google Earth Engine. Devido à maior resolução espacial dos produtos MODIS (conforme indicado na Tabela 1), calcularam-se as médias diárias dos valores contidos nas células da grade do ERA5 para o período de 2002 a 2011.

Os produtos MODIS apresentam uma resolução temporal que varia entre 4, 8 ou 16 dias, variando de acordo com o produto. Por isso, para a obtenção dos dados diários foi adotada a mesma abordagem metodológica utilizada pela Fluxcom, a qual considerou

o comportamento sazonal das variáveis, por exemplo, no caso do NDVI, cada medição no dia 16 representa as medições dos 15 dias anteriores. Essa técnica foi aplicada de acordo com a sazonalidade de cada variável e assim foi preenchida a série temporal diária de 2002 a 2011 com os dados do MODIS.

Os dados do Merge foram baixados a partir da API e foram extraídos a partir da média dos valores contidos na célula de 0.25°, semelhante ao que foi feito com os dados do MODIS [Rozante José 2010] . Na tabela 1 há a descrição e abreviaturas de todos os dados usados nesse artigo. Os atributos foram escolhidos com base nas variáveis do ERA-5 e do MODIS utilizadas pela iniciativa Fluxcom e com base nas variáveis explicitadas como importantes no comportamento biogeoquímico de florestas de acordo com [Waring and Running 2007]. Realizou-se a integração dos dados e o cálculo da correlação

Tabela 1. Abreviatura, nomes, resoluções temporais/espaciais e fontes dos dados utilizados nesse trabalho.

Abreviatura	Atributo	Resolução Temporal	Resolução Espacial	Fonte
t2m	2m temperature	horária	0.25°	ERA-5
d2m	2m dewpoint temperature	horária	0.25°	ERA-5
aluvd	UV visible albedo for diffuse radiation	horária	0.25°	ERA-5
alnid	Near IR albedo for diffuse radiation	horária	0.25°	ERA-5
e	Evaporation	horária	0.25°	ERA-5
stl1 e stl4	Soil Temperature Level 1 e 4	horária	0.25°	ERA-5
swvl4	Volumetric soil water layer 4	horária	0.25°	ERA-5
ro	Runoff	horária	0.25°	ERA-5
ndvi	NDVI	16 dias	1km	MODIS MOD13A2
evi	EVI	16 dias	1km	MODIS MOD13A2
lai	Lai	4 dias	500m	MODIS MCD15A3H
fpar	Fpar	4 dias	500m	MODIS MCD15A3H
lst_day_1km	Temperatura de Superfície Dia	8 dias	1km	MODIS MOD11A3
lst_night_1km	Temperatura de Superfície Noite	8 dias	1km	MODIS MOD11A3
prec	Precipitation	diário	0.1°	MERGE

de Pearson entre as séries temporais. Essa análise teve como objetivo compreender a relação dos dados ao longo do tempo e determinar a possível necessidade de eliminar alguns atributos devido a correlações fortes. O atributo-alvo utilizado é o NEE_VUT_REF fornecido pelo Fluxnet [Pastorello 2020]. Esse atributo é a medição de NEE efetuada na torre KM67, como parte do projeto LBA [Avissar 2002], que foi disponibilizado para a Fluxnet e passou por um processo de padronização com outros dados de fluxo globais. Isso foi realizado para preenchimento de falhas e padronização dos algoritmos [Pastorello 2020].

Para realizar a estimativa da série temporal de NEE foram utilizados os algoritmos de aprendizado de máquina Gradient Boost (GBoost), Support Vector Macchine (SVM),

Random Forest (RF) e a rede neural Multilayer Perceptron (MLP). Os três últimos foram utilizados pela fluxcom [Tramontana and Jung 2016, Jung 2020] e outras iniciativas para estimativa global de NEE [Zhuravlev 2022]. Foram gerados dois conjuntos de dados de entrada para esses modelos: o Meteo, que continha apenas dados meteorológicos (ERA-5 e Merge), e o Meteo-RS, que incluiu dados meteorológicos e de sensoriamento remoto (MODIS).

Após avaliar a correlação entre variáveis, as com alta correlação positiva e negativa foram removidas para evitar interferências nos algoritmos de Machine Learning e facilitar a generalização. Cada variável foi normalizada de 0 a 1 e dividida em conjuntos de treino (67%) e validação (33%). Cerca de 1000 testes foram realizados com a biblioteca Python Optuna para ajustar hiperparâmetros dos modelos. A escolha dos melhores hiperparâmetros baseou-se em três métricas: R^2, MSE e RMSE, selecionando os que obtiveram melhor desempenho.

3. Resultados

A correlação final entre as variáveis pode ser visualizada na Figura 1. Em relação à variável alvo NEE_VUT_REF, foi possível identificar correlações positivas com todas as variáveis, algumas em menor intensidade, como a precipitação e temperatura do ponto de orvalho e outras em maior intensidade, como a evaporação e quantidade de água no solo. O albedo da radiação difusa no infravermelho próximo se mostrou com correlação negativa e com menor intensidade. Dentre os algoritmos de aprendizado de máquina,

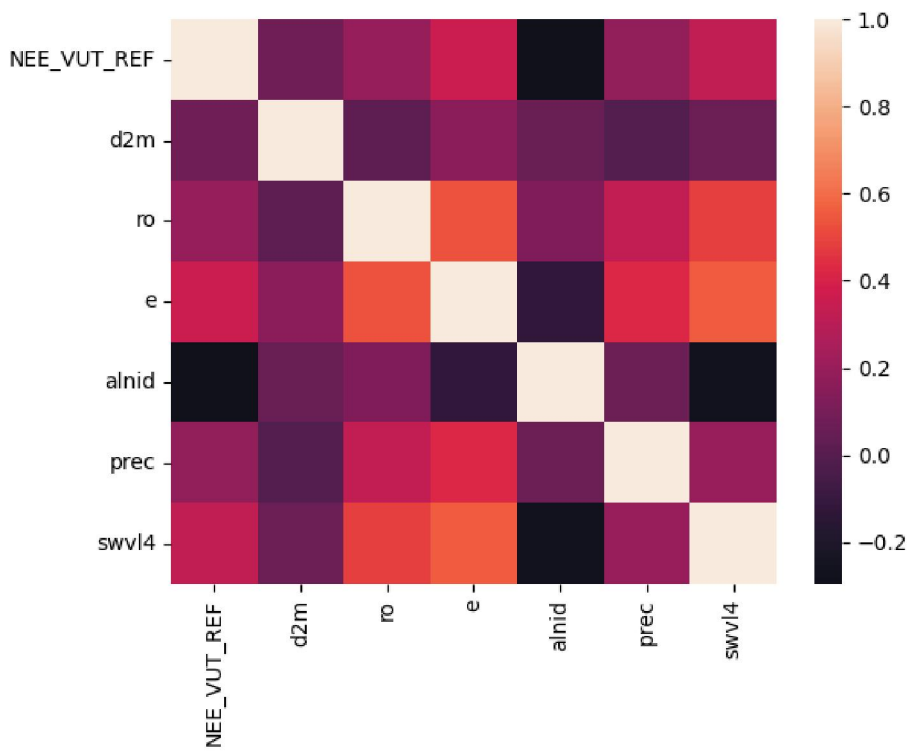

Figura 1. Heatmap de Correlação das variáveis.

o Multilayer Perceptron alcançou o melhor desempenho em termos do coeficiente de determinação R^2 no conjunto de dados Meteo, conforme pode ser visto nas tabelas 2 e 3. Entretanto, o mesmo apresentou a maior diferença nos erros, tanto no MSE quanto no RMSE. Os modelos de Random Forest (RF), Gradient Boosting (GBoost) e Support Vector Machine (SVM) apresentaram valores mais baixos de R^2, além de registrarem os menores valores de Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio

Tabela 2. Resultados dos Modelos de Machine Learning para o conjunto de dados METEO, métricas R^2, MSE e RMSE.

METEO			
Modelo	R^2	MSE	RMSE
RF	0.2679	0.8726	0.9341
GBoost	0.2332	0.9141	0.9561
SVM	0.2401	0.9058	0.9517
MLP	0.2879	1.8878	1.3740

(RMSE), os quais se mostraram bastante próximos entre si. Em todos os algoritmos de

Tabela 3. Resultados dos Modelos de Machine Learning para o conjunto de dados METEO-RS, métricas R^2, MSE e RMSE.

METEO-RS			
Modelo	R^2	MSE	RMSE
RF	0.2533	0.8921	0.9445
GBoost	0.2304	0.9195	0.9589
SVM	0.2341	0.9151	0.9566
MLP	0.2793	1.8120	1.3461

aprendizado de máquina, os resultados mais eficazes foram obtidos por meio de estruturas mais simples, ou seja, modelos menos elaborados. De fato, quando houve um aumento excessivo na complexidade, como por exemplo, ao adicionar um grande número de árvores nos modelos Random Forest e GBoost, as métricas obtidas começaram a deteriorar, e algumas delas se aproximaram consideravelmente do melhor desempenho alcançado com a estrutura mais simples, com apenas 30 árvores. No caso do MLP, foi observado que com o aumento das camadas ocultas e épocas ocorreu piora no treinamento e piores métricas. No entanto, com apenas 14 épocas com o conjunto de dados Meteo e 30 épocas com o Meteo-RS foi possível encontrar os melhores resultados de estimativa e generalização das redes. O diferencial necessário no MLP incluiu o uso de uma taxa de aprendizagem variável a cada 5 épocas para ambos os conjuntos de dados, o uso de inicialização aleatória de pesos e uma camada oculta com 100 e 50 neurônios, nos conjuntos Meteo e Meteo-RS, respectivamente.

O modelo Support Vector Machine (SVM) teve desempenho semelhante aos outros modelos, mas precisou de uma redução na função de custo para 0.03 para otimizar seu desempenho. O uso do algoritmo de RBF implicou a aplicação de uma função de kernel para criar fronteiras de decisão mais complexas em um espaço dimensional superior, lidando com relações não lineares nos dados. No entanto, ao aumentar a função de custo, observaram-se métricas de desempenho inferiores, sugerindo possível sobreajuste do modelo.

4. Conclusões

As variáveis abordadas neste estudo refletem o microclima de um ecossistema. A análise de correlação de Pearson, conduzida ao longo de um período de 10 anos, revelou as

relações lineares entre as variáveis ambientais e a NEE. Esse processo possibilitou a identificação dos atributos de maior impacto no balanço de carbono, permitindo-nos selecionar os mais relevantes para o treinamento dos modelos e eliminar atributos redundantes.Os resultados mais promissores dos diversos modelos foram detalhados nas Tabelas 2 e 3, após a avaliação de várias configurações de hiperparâmetros. Esses resultados estão em concordância com as estimativas da iniciativa Fluxcom na América do Sul [Jung 2020, Tramontana and Jung 2016].

Contudo, há margem para aprimorar essas estimativas. O menor valor de R^2 nos modelos deste estudo em comparação com os resultados da Fluxcom pode ser atribuído à menor quantidade de dados utilizados, restritos apenas aos dados da torre BrSa1 da Amazônia e sua célula co-localizada. Além disso, a discrepância na resolução espacial tanto dos dados de entrada quanto de saída pode ter influenciado, visto que a Fluxcom utilizou uma grade espacial de 0.083° e 0.5°, divergente da adotada neste estudo. É crucial aprimorar e expandir as estimativas do NEE em toda a região Amazônica para compreender o equilíbrio de carbono. Planejamos utilizar métodos inovadores, como a expansão da coleta de dados por meio de outras torres e o emprego de redes neurais, visando alcançar estimativas mais precisas das séries temporais do NEE. Essa melhoria é fundamental para apoiar políticas e iniciativas de combate às mudanças climáticas.

Referências

Avissar, e. a. (2002). The large-scale biosphere-atmosphere experiment in amazonia (lba). 107, article 8086.

Baldocchi, D. D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, 9(4):479–492.

Hersbach, H. and Bell, e. a. (2023). Era5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

Jung, M. e. a. (2020). Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the fluxcom approach. *Biogeosciences*, 17(5):1343–1365.

Pastorello, e. a. (2020). The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific data*, 7(1):1–27.

Rozante José, e. a. (2010). Combining trmm and surface observations of precipitation: Technique and validation over south america. *Weather and Forecasting*, 25(3):885 – 894.

Tramontana, G. and Jung, e. a. (2016). Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms. *Biogeosciences*, 13(14):4291–4313.

Waring, R. H. and Running, S. W. (2007). *Forest Ecosystems*. Academic Press, San Diego, third edition edition.

Zhuravlev, e. a. (2022). Globally scalable approach to estimate net ecosystem exchange based on remote sensing, meteorological data, and direct measurements of eddy covariance sites. *Remote Sensing*, 14(21).