UFSM

CIÊNCIA e NATURA

ISSN 2179-460X
ACESSO ABERTO

Ci. e Nat., Santa Maria, v. 45, n esp. 2 e80997, p. 01-11, 2023 • https://doi.org/10.5902/2179460X80997
Submissão: 27/04/2023 • Aprovação: 14/06/2023 • Publicação: 22/11/2023

Edição Especial

Comparação de métodos de preenchimento de dados de fluxo de CO_2

Comparison of gap-filling methods for CO_2 flux data

Bernardo Ivo Goltz[I], Daniele Morgenstern Aimi[I], Alecsander Mergen[I], Vanessa de Arruda Souza[I, II], Gustavo Pujol Veeck[I], Tiago Bremm[I], Michel Baptistella Stefanello[I], Débora Regina Roberti[I]

[I] Universidade Federal de Santa Maria, Santa Maria, RS, Brasil
[II] Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil

RESUMO

Dados coletados por sensores estão sempre sujeitos a possíveis falhas, seja por falta de energia, interferências externas, entre outros. Além disso, muitos dados também são excluídos no processo de filtragem por serem fisicamente inconsistentes. Essas falhas geram a necessidade da implementação de diferentes métodos de tratamento de dados com ênfase em preencher os registros ausentes. No caso de dados de fluxo de CO_2 o preenchimento dos dados faltantes é extremamente importante para obtenção de acumulados anuais e balanço de carbono. O pacote REddyProc é amplamente utilizado e documentado a respeito do preenchimento deste tipo de dados. No entanto, diferentes e modernos métodos têm sido cada vez mais explorados, buscando otimizar este processo. Neste trabalho, comparamos o preenchimento de dados entre o pacote REddyProc e o método KNN Imputer. Os resultados preliminares mostram que o pacote REddyProc possui melhores índices estatísticos no preenchimento de fluxos de CO_2 em comparação com o método KNN.

Palavras-chave: Fluxo de CO_2; Preenchimento de falhas; Dados faltantes

ABSTRACT

Data collected by sensors are always subject to possible failures, whether due to power failure, external interference, among others. Moreover, much of the data is not considered during the filtering process because it is physically inconsistent. These failures result in the need to implement various methods of data processing, with a focus on filling in missing records. In the case of CO_2 flux data, filling in the missing data is crucial to obtain annual data and the carbon balance. The REddyProc package is widely used and documented in terms of filling this type of data. However, modern methods have been increasingly explored to optimize this process. In this study, we compare data filling between

the REddyProc package and the KNN Imputer method. Preliminary results show that the REddyProc package has better statistical indices when filling CO_2 streams compared to the KNN method.

Keywords: CO_2 flux; Gap filling; Missing data

1 INTRODUÇÃO

O estudo dos fluxos superficiais que ocorrem entre a biosfera e a atmosfera tem grande aplicação em problemas ambientais. A mudança do uso da terra devido a pecuária e a agricultura tem influenciado diretamente em alterações dos fluxos biogeoquímicos e contribuído para o aumento da concentração dos gases de efeito estufa (GEE), principalmente o dióxido de carbono (CO_2) (IPCC, 2019). O CO_2 tem sofrido um aumento considerável em sua concentração na atmosfera durante as últimas décadas, devido às atividades antropogênicas, sendo que a mudança no uso da terra e a agricultura tem contribuído com 24% (IPCC, 2014; Panchasara; Samrat; Islam, 2021). Determinar as fontes e sumidouros de CO_2 para a atmosfera causados pelo uso da terra, vem se tornando cada vez mais importante visto o cenário de aumento da temperatura global, no qual afeta diretamente no aquecimento global.

Determinar as trocas líquidas de CO_2 em áreas de uso da terra é um fator limitante, devido à disponibilidade de dados observacionais. A metodologia de covariância dos vórtices (EC, do inglês *Eddy Covariance*) é o estado da arte na estimativa das trocas líquidas de CO_2 (Baldocchi *et al.*, 2012). O método EC estima os fluxos de gases entre a superfície do solo e a atmosfera, através de uma covariância estatística entre as flutuações temporais da velocidade vertical do vento com as flutuações temporais da concentração de gases. A turbulência é o processo físico responsável por estes fluxos na atmosfera, e é estudado principalmente na área de micrometeorologia. Esta técnica faz uso de sensores muito sensíveis para medidas de alta frequência (10 Hz = 10 medidas por segundo) das componentes do vento, temperatura e umidade do ar, através de anemômetros sônicos 3D e da concentração de gases e vapor d'água. Em geral, estes equipamentos são reunidos em uma torre, chamada torre de fluxo.

Embora o método EC exija equipamentos muito sensíveis e de alto custo, ele tem a significativa vantagem de permitir medidas rápidas (um ou dois anos) do balanço de carbono em um ecossistema, se comparado aos outros métodos de quantificação do SOC (cinco a dez anos). Além disso, possibilitam inúmeras correlações com outras variáveis ambientais, como as práticas de uso e manejo do solo, possibilitando, com isso, avaliar formas de mitigar as emissões de carbono para a atmosfera. Essas medidas experimentais frequentemente apresentam falhas de dados, devido a vários fatores, tais como: queda de energia, falha ou problemas nos sensores, interferência causadas por diversos agentes externos, entre outros. As perdas de dados geralmente chegam a 35-55% dos registros (Falge *et al.* 2001). Por esta razão, uma análise de dados, preenchimento de falhas e metodologias que descrevam o comportamento das séries temporais com informações físicas e ambientais são de extrema importância para estudar os fenômenos da atmosfera.

Existem diferentes maneiras de lidar com dados faltantes, ausentes. Alguns métodos, como remover toda a observação se ela tiver um valor ausente ou substituir os valores ausentes por valores médios, medianos ou moda. No entanto, esses métodos podem desperdiçar dados valiosos ou reduzir a variabilidade do seu conjunto de dados. Este trabalho tem como objetivo avaliar métodos de preenchimentos de falha de dados baseados em técnicas estatísticas, sendo eles: REddyProc e KNN, com a finalidade de buscar métodos de implementação mais simples para o preenchimento do fluxo de CO_2, a fim de preservar o comportamento físico da variável de estudo.

2 MATERIAIS E MÉTODOS

2.1 Sítio experimental ACE

O sítio experimental (ACE) está localizado no município de Aceguá em uma área particular destinada à criação de gado, Estância Cinco Salsos (31,65° S; 54,17° O; altitude: 170 m), no estado do Rio Grande do Sul, Brasil. Esta região é composta

por campos nativos com vegetação predominante de gramínea pertencente ao bioma Pampa, localizado no extremo sul do Brasil. A vegetação é caracterizada pelo domínio de *Paspalum notatum*, *Axonopus affinis*, *Mnesithea selloana*, *Paspalum dilatatum*, *Nassela sp.* e *Piptochaetium sp.*, *Baccharis coridifolia*, *B. crispa* (Baggio, 2017).

Na área experimental, uma torre de fluxo (Figura 1) foi instrumentada com sensores que medem variáveis atmosféricas, variáveis de solo e concentração de dióxido de carbono (CO_2). As componentes da velocidade do vento (*u*, *v* e *w*) foram medidas com um anemômetro sônico tridimensional (CSAT3, Campbell Scientific Inc., USA) e as concentrações de CO_2 foram medidas utilizando um analisador de gás LI-7500 (LI-COR Inc., Lincoln, NE, USA). Estas medidas foram realizadas a 2,5 m acima da superfície do solo e registradas a 10 Hz por meio de uma unidade de interface do analisador (LI-7550, LI-COR Inc., Lincoln, NE, USA) e armazenadas em um pendrive de 16 GB.

Figura 1 – Torre de fluxo do sítio experimental de Aceguá (ACE)

Fonte: Acervo particular dos autores (2022)

O fluxo de CO_2 foi obtido através da técnica de EC utilizando o software EddyPro, v7.0.6 (LI-COR Biosciences, Lincoln, Nebraska, EUA). Para o cálculo dos fluxos foi adotada

a média em bloco de 30 minutos e aplicadas as seguintes correções: rotação dupla (Wilczak; Oncley; Stage, 2001); correções para efeitos de densidade (Webb; Pearman; Leuning, 1980); atenuação de fluxo devido à configuração instrumental (Gash; Culf, 1996); correções devido a filtros de passa alta (Moncrieff *et al.*, 1997) e passa baixa (Moncrieff *et al.*, 2004); filtragem de dados de alta frequência (Vickers; Mahrt, 1997). A metodologia de Foken *et al.* (2004), foi utilizada na remoção dos fluxos de baixa qualidade, que também foram descartados em eventos de precipitação e na meia hora posterior (para secagem do instrumento). Um controle estatístico foi aplicado seguindo Béziat; Ceschia; Dedieu (2009), ou seja, dados fora de uma faixa de desvio padrão de ±2,5 da janela móvel de 200 pontos de dados (separadamente para dados diurnos e noturnos) foram identificados como espúrios e removidos.

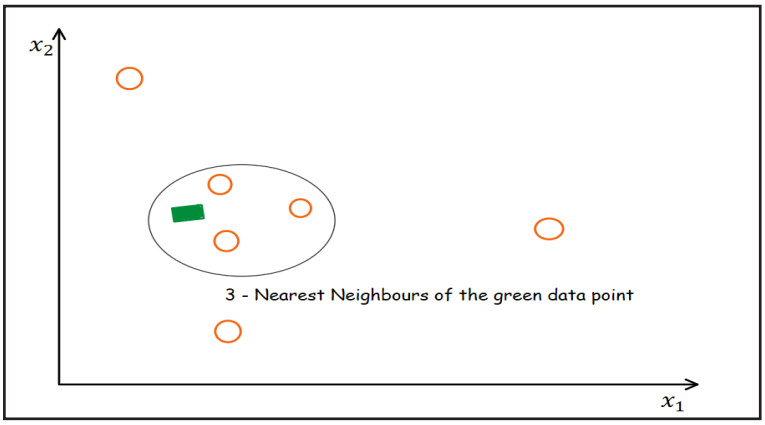
2.2 Preenchimento de falhas pelo pacote REddyProc

A metodologia estatística descrita por Reichstein *et al.* (2005) foi utilizada para o preenchimento de dados faltantes no fluxo de CO_2 (troca líquida do ecossistema – NEE), sendo realizado através do pacote REddyProc (Max Planck Institute for Biogeochemistry, Germany), disponível para o software RStudio, utilizando os seguintes parâmetros: temperatura do ar (T), radiação global (Rg) e déficit de pressão de vapor (VPD).

2.3 Preenchimento de falhas pelo método K-Nearest Neighbours

O algoritmo utilizado para preencher séries de dados com valores faltantes foi o algoritmo de aprendizagem supervisionada K-Nearest Neighbours (K-ésimos vizinhos mais próximos), implementado a partir do módulo KNN Imputer (Scikit-Learn, 2011). A ideia geral do algoritmo no ajuste de valores faltantes é computar a distância euclidiana entre uma dada quantidade K de pontos vizinhos ao valor falho e estimar o registro ausente com base no valor médio de sua vizinhança, ponderado pela distância entre os pontos (Figura 2).

Figura 2 – Diagrama representando os K-ésimos vizinhos mais próximos



Fonte: Acervo particular dos autores (2022)
Legenda: Diagrama representando os K-ésimos vizinhos mais próximos

Para estimar os valores ausentes, o algoritmo gera uma matriz a partir do conjunto de dados e, quando houver um registro faltante, considera os 'K' valores vizinhos e a distância euclidiana na matriz. Nesse caso, considerando uma série temporal é esperado que uma amostra possua certa similaridade com os registros próximos em certo período de tempo, portanto, é válido considerar os valores adjacentes para estimar um valor central.

2.4 Aplicação dos métodos

Para comparar os modelos de preenchimento, utilizamos um conjunto de dados de NEE, obtido através da técnica EC, chamados dados originais pois passaram por um processo de filtragem resultando em falhas no conjunto de dados processados. Primeiramente, realizamos um preenchimento destas falhas iniciais com os dois métodos, a fim de obter o conjunto de um ano de dados completo, para cada método de preenchimento (estatístico e machine learning). Para controle, criamos um flag de "0" para quando o dado fosse empírico, ou seja, o dado original existia, e "1" para dados resultantes de preenchimento prévio, ou seja, dado original apresentava falha. Em seguida, geramos falhas em períodos aleatórios em cada conjunto de dados apenas quando flag igual a "0", obtendo dois novos conjuntos com dados faltantes.

Foram inseridas falhas em 20% das linhas de forma aleatória nos dois conjuntos de dados, porém rigorosamente nos mesmos registros. O processo de geração de falhas aleatórias e preenchimento foi repetido por cinco rodadas por cada metodologia e então obtidas as métricas do preenchimento em comparação aos dados originais. As métricas estatísticas utilizadas para análise dos resultados foram: percentual de viés (PBIAS), raiz quadrada do erro médio (RMSE), erro quadrático médio (MSE) e coeficiente de determinação (R^2), conforme descrito em Marshall (2007).

3 RESULTADOS E DISCUSSÃO

Implementando ambos os modelos de preenchimento e comparando com os dados empíricos, obtivemos as métricas dispostas na Tabela 1. Cada método apresentou resultados similares de RMSE, MSE e R^2 para as diferentes rodadas. No entanto, o pacote REddyProc apresenta uma melhor performance estatística sobre o preenchimento, apresentando valores de RMSE variando entre 0,89 e 0,91 μmol CO_2 m^{-2} s^{-1}, enquanto para a KNN estes valores variaram de 1,22 a 1,25 μmol CO_2 m^{-2} s^{-1}. R^2, foi o mesmo para as diferentes rodadas em cada método, sendo maior para ReddyProc (0,98) que para o KNN (0,96). PBIAS apresentou grande variação para ambos os métodos, variando de -4,45 a 1,97 para o ReddyProc e entre 1,50 a 6,60 para o KNN.

De modo geral, o ReddyProc apresentou melhores índices estatísticos, representando melhor os dados originais. Este resultado pode ser devido ao fato de que o REddyProc utiliza fluxos de CO_2 obtidos em situações de similar condições ambientais de Rg, T e VPD, enquanto o KNN utiliza apenas o valor dos “k” vizinhos mais próximos no processo de preenchimento.

Os resultados aqui apresentados são preliminares. Para uma análise mais detalhada seriam necessários testes relacionados a quantidade ótima de vizinhos no método KNN, além de iterar uma quantidade maior de testes e comparar seus resultados estatísticos. Para um trabalho futuro, é possível analisar a qualidade dos métodos quando ocorrem falhas mais extensas, além de avaliações para diferentes integrações sazonais.

Tabela 1 – Métricas estatísticas dos modelos em cada rodada do preenchimento de falha de dados

Rodadas	Método	PBIAS (%)	RMSE (µmol CO_2 m^{-2} s^{-1})	MSE (µmol CO_2 m^{-2} s^{-1})	R^2
1° rodada	REddyProc	-4,45	0,89	0,80	0,98
	KNN	2,65	1,24	1,55	0,96
2° rodada	REddyProc	0,65	0,91	0,82	0,98
	KNN	5,02	1,24	1,55	0,96
3° rodada	REddyProc	1,80	0,89	0,79	0,98
	KNN	1,60	1,22	1,49	0,96
4° rodada	REddyProc	1,97	0,91	0,82	0,98
	KNN	1,50	1,23	1,51	0,96
5° rodada	REddyProc	1,83	0,90	0,82	0,98
	KNN	6,06	1,25	1,57	0,96

Fonte: Acervo particular dos autores

4 CONSIDERAÇÕES FINAIS

Os resultados obtidos nos mostraram que o pacote REddyProc destaca-se na qualidade do preenchimento de dados de fluxo de CO_2. No entanto, o método KNN Imputer mostra grande potencial no preenchimento de dados faltantes, sendo uma alternativa a ser melhor explorada.

AGRADECIMENTOS

Os autores agradecem às Agências Brasileiras de Pesquisa: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS).

REFERÊNCIAS

BAGGIO, R. **Estratégias de manejo adaptativo para os Campos Sulinos**. 2017. 129 p. Tese (Doutorado em Ecologia) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2017.

BALDOCCHI, D. *et al.* The challenges of measuring methane fluxes and concentrations over a peatland pasture. **Agricultural and Forest Meteorology**, v. 153, p. 177–187, 2012.

BÉZIAT, P.; CESCHIA, E.; DEDIEU, G. Carbon balance of a three crop succession over two cropland sites in South West France. **Agricultural and Forest Meteorology**, v. 149, n. 10, p. 1628– 1645, mar. 2009.

FALGE, E., *et al.*, Gap filling strategies for defensible annual sums of net ecosystem exchange. **Agricultural and Forest Meteorology**, v. 107, p. 43–69. 2001

FOKEN, T *et al.*, 2004. Handbook of Micrometeorology: A Guide for surface flux measurement and analysis: **Chapter 9: POST-FIELD DATA QUALITY CONTROL**, Handbook of Micrometeorology.

IPCC. **Climate Change and Land**: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. [s.l: s.n.].

IPCC. I. P. ON C. C. **Mitigation of climate change**. In Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, Switzerland: [s.n.].

MARSHALL, G. Statistical methods in the atmospheric sciences, second edition D. S. Wilks. 1995. International Geophysics Series, Vol 59, Academic Press, 464pp. ISBN-10: 0127519653. ISBN-13: 978-0127519654. **Meteorological Applications**, v. 14, n. 2, p. 205–205, jun. 2007.

MONCRIEFF, J. B. *et al.* A system to measure surface fluxes of momentum, sensible heat, water vapour and carbon dioxide. **Journal of Hydrology**, 1997.

MONCRIEFF, J. *et al.* Averaging, detrending, and filtering of eddy covariance time series, in Handbook of micrometeorology. **Handbook of Micrometeorology**: A Guide for surface flux measurement and analysis, 2004.

PANCHASARA, H.; SAMRAT, N. H.; ISLAM, N. Greenhouse Gas Emissions Trends and Mitigation Measures in Australian Agriculture Sector—A Review. **Agriculture**, v. 11, n. 2, p. 85, 20 jan. 2021.

REICHSTEIN, M. *et al.* On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. **Global Change Biology**, v. 11, n. 9, p. 1424– 1439, set. 2005.

SCIKIT-LEARN: **Machine Learning in Python**, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

VICKERS, D.; MAHRT, L. Quality control and flux sampling problems for tower and aircraft data. **Journal of Atmospheric and Oceanic Technology**, 1997.

WEBB, E. K.; PEARMAN, G. I.; LEUNING, R. Correction of flux measurements for density effects due to heat and water vapour transfer. Quarterly Journal of the Royal Meteorological Society, 1980. WILCZAK, J. M.; ONCLEY, S. P.; STAGE, S. A. Sonic anemometer tilt correction algorithms. **Boundary-Layer Meteorology**, 2001.

WUTZLER, T. *et al.* Basic and extensible post-processing of eddy covariance flux data with REddyProc. **Biogeosciences**, 2018.

Contribuição de autoria

1 – Bernardo Ivo Goltz

Universidade Federal de Santa Maria, Estudante de Engenharia Elétrica
https://orcid.org/0009-0002-3373-5874 • bergoltzx2@gmail.com
Contribuição: Curadoria de Dados; Análise Formal; Escrita – Primeira Redação

2 – Daniele Morgenstern Aimi

Universidade Federal de Santa Maria, Física Médica, Doutora em Física
https://orcid.org/0000-0002-6383-6572 • danielefm@gmail.com
Contribuição: Escrita – Primeira Redação; Escrita – Revisão e Edição; Análise Formal

3 – Alecsander Mergen

Universidade Federal de Santa Maria, Físico, Mestre em Física
https://orcid.org/0000-0001-7126-8694 • alecsandermergen@hotmail.com
Contribuição: Curadoria de Dados; Análise Formal; Escrita – Primeira Redação

4 – Vanessa de Arruda Souza

Universidade Federal do Rio Grande do Sul, Meteorologista, Doutora em Sensoriamento Remoto
https://orcid.org/0000-0002-8518-1271 • v.arruda.s@gmail.com
Contribuição: Escrita – Revisão e Edição

5 – Gustavo Pujol Veeck

Universidade Federal de Santa Maria, Físico, Mestre em Física
https://orcid.org/0000-0002-1444-0360 • veeckgp@gmail.com
Contribuição: Curadoria de Dados

6 – Tiago Bremm

Universidade Federal de Santa Maria, Físico, Mestre em Meteorologia
https://orcid.org/0000-0003-1564-1014 • bremm.tiago@gmail.com
Contribuição: Curadoria de Dados

7 – Michel Baptistella Stefanello

Universidade Federal de Santa Maria, Físico, Doutor em Física
https://orcid.org/0000-0002-6380-3252 • michelstefanello@gmail.com
Contribuição: Escrita – Revisão e Edição

8 – Débora Regina Roberti

Universidade Federal de Santa Maria, Física, Doutora em Física
https://orcid.org/0000-0002-3902-0952 • debora@ufsm.br
Contribuição: Escrita – Revisão e Edição

Como citar este artigo

GOLTZ, B. I.; AIMI, D. M.; MERGEN, A.; SOUZA, V. A.; VEECK, G. P.; BREMM, T.; STEFANELLO, M. B.; ROBERTI, D. R. Comparação de métodos de preenchimento de dados de fluxo de CO_2. **Ciência e Natura**, Santa Maria, v. 45, n. esp. 2, e80997, 2023. DOI: https://doi.org/10.5902/2179460X80997. Disponível em: https://periodicos.ufsm.br/cienciaenatura/article/view/80997. Acesso em: dia mês abreviado ano.