*Biomatemática* ***32*** *(2022), 133–140* ***ISSN 1679-365X***
*Uma publicação do grupo de Biomatemática IMECC – Unicamp*

# Análise e agrupamento de dados da distribuição do teor de carbono em profundidade no solo

Vitor H. M. Mourão[1],
DMA, IMECC – Unicamp, 13.083-859, Campinas/SP.

Yusuf N. Karatay[2], Luis G. Barioni[3],
Embrapa Agricultura Digital, 13.083-886 - Campinas/SP.

Laércio L. Vendite[4],
DMA, IMECC – Unicamp, 13.083-859, Campinas/SP.

**Resumo**. Este trabalho tem como objetivo analisar as características da distribuição do teor de carbono no solo e através do agrupamento de dados refinar tal análise. Os dados utilizados nesse estudo originam-se de amostragens realizadas em 11 estados brasileiros entre os anos de 2020 e 2021. Foi analisada as estatisticas da distribução do teor de carbono em profundidade e para agrupar os dados foi aplicada uma técnica de mineração de dados denominada Clusterização através do uso do algoritmo *Expectation Maximization* (EM) para identificar conjuntos de alta similaridade. Foram utilizadas 5029 instâncias que representam amostras coletadas em 53 propriedades rurais pelo projeto *PRO Carbono* da Bayer em parceria com a Embrapa. Para a realização desse trabalho foram considerados dois cenários: sem preenchimento dos dados faltantes e com o preenchimento dos dados faltantes pela média global das camadas de solo. Os resultados apontam que as camadas mais profundas de solo possuem, em média, menor teor de carbono (g/kg) do que as camadas superiores e as variações dos teores de carbono em camadas mais profundas também são menores se comparado às variânciações das camadas próximas à superfície. No cenário sem preenchimento de valores faltantes, foram gerados 14 clusters que possuem diferentes combinações de camadas. No cenário onde houve preen-

[1]v137856@dac.unicamp.br
[2]yusuf.karatay@colaborador.embrapa.br
[3]luis.barioni@embrapa.br
[4]vendite@ime.unicamp.br

chimento dos dados pela média das camadas o número de clusters gerados diminuiu (de 14 para 10) e um menor coeficiente de variação médio foi obtido.

***Palavras-chave**: carbono no solo, clusterização, mineração de dados.*

# 1. Introdução

Em 2017 o Brasil possuía aproximadamente 30% do seu território ocupado por lavouras, pastagens e florestas plantadas (Miranda, 2018). Tamanha extensão de produção do setor agropecuário brasileiro permite a geração de grandes benefícios econômicos caso seja realizada a mensuração e remuneração dos estoques de carbono presentes no solo.

Estes benefícios econômicos podem ser obtidos principalmente por duas vias: (i) Quando há aumento do sequestro de $CO_2$ na forma de matéria orgânica no solo o que poderá permitir, se acuradamente mensurado, a negociação de créditos de carbono e/ou acesso à subsídios; (ii) Maior produtividade do solo, já que maior teor de carbono orgânico no solo aumenta a estabilidade estrutural do solo, melhora a capacidade de retenção de água, torna o solo mais resistente às adversidades climáticas e provê energia para processos biológicos que por sua vez acabam aumentando a presença de nutrientes (Banwart et al., 2014).

Para identificar a distribuição do teor de carbono ao longo da profundidade, foram obtidas a média, mediana, desvio padrão, valores máximos e mínimos do teor de carbono para cada camada presente no conjunto de dados. Também utilizamos o algoritmo *Expectation Maximization* (EM), implementado através do software Weka (Witten et al., 2005) para implementar o agrupamento de dados (Clusterização) (Eibe et al., 2016). Ao todo, foram analisadas 53 fazendas que juntas possuem 629 pontos de amostragem de solo, cada qual com 8 subcamadas de profundidade, totalizando mais de 5 mil instâncias.

# 2. Objetivos

- Identificar como o teor de carbono no solo é distribuido em profundidade (até 1 metro).

- Identificar se há uma regra que determina a distribuição do teor de carbono através da profundidade no solo.

## 3. Metodologia

Os dados utilizados nesse trabalho são do projeto *PRO Carbono* da Bayer que foram disponibilizados em parceria com a Empresa Brasileira de Pesquisa Agropecuária (Embrapa). São dados de 53 fazendas presentes em 11 estados brasileiros, cada uma dividida em três áreas com diferentes usos de terra. Para cada área foram amostradas 4 trincheiras de um metro de profundidade. Em cada trincheira foi realizada a análise do solo em 8 subcamadas (0-5 cm, 5-10 cm, 10-20 cm, 20-30 cm, 30-40 cm, 40-60 cm, 60-80 cm, 80-100 cm).

Foi necessário realizar o processo de limpeza dos dados. Variáveis com mais de 60% dos valores faltantes e variáveis que não são do interesse desse estudo foram removidas.

Com os dados preparados, foi realizada a análise da distribuição do teor de carbono através das camadas em profundidade, os valores médios, o desvio padrão e os valores máximos e mínimos de cada camada serviram como base para comparação dos resultados obtidos através do agrupamento de dados.

Como a técnica escolhida para análise da distribuição de carbono no solo foi o agrupamento de dados (clusterização), é essencial que apliquemos uma normalização, pois o conjunto de dados possui variáveis com diversas ordens de grandeza. Logo, para realizar a normalização foi utilizado o filtro *Normalize*, nativo do Weka.

O método escolhido para o agrupamento de dados é o *Expectation Maximization* (Do and Batzoglou, 2008), que atribui uma distribuição de probabilidade a cada instância indicando a probabilidade da mesma pertencer a cada um dos clusters e através de um processo iterativo preenche os clusters até que a ordenação das instâncias não produza melhora significativa da distribuição das mesmas. Além disso, o método decide quantos clusters formar através de validação cruzada e permite realizar a clusterização com dados faltantes, o que possibilita a comparação entre os resultados. Para aplicar o método foram utilizados os parâmetros padrão do Weka e 200 rodadas de K-Means.

Definidos os parâmetros do algoritmo, aplicamos o método de agrupamento para o conjunto de dados com valores faltantes (Cenário 1) e para o conjunto de dados onde os valores faltantes foram preenchidos pela média global das camadas (Cenário 2). Os atributos utilizados para ambos os cenários foram a quantidade de areia, silte, argila, o nível de pH, o teor de carbono e a camada em que esses dados foram coletados.

# 4. Resultados e discussão

A distribuição do teor de carbono (em g/kg) pode ser observada na figura 1. Como é possível observar, a distribuição do teor de carbono é monotonamente decrescente, sendo maior nas camadas superficiais e menor em profundidade. A variação absoluta dos teores de carbono também é maior em camadas superficiais, mas quando consideramos a variação relativa (desvio padrão/média), observamos que esta permanece estável até a quinta camada e aumenta nas camadas mais profundas. Este resultado indica uma menor certeza para camadas em profundidade e uma maior faixa de valores para camadas superfíciais.

Para a análise de agrupamento de dados, foram comparados dois cenários na análise de distribuição do teor de carbono no solo. No primeiro cenário, utilizamos o algoritmo EM sem alterar os valores faltantes presentes no conjunto de dados. Neste cenário o algoritmo EM identificou 14 clusters. No segundo cenário foram preenchidos os valores faltantes pela média de cada camada, este cenário produziu 10 clusters.

Tabela 1: Número de instâncias presentes em cada camada em cada cluster gerado.

| | **Profundidade do solo (cm)** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **0-5** | **5-10** | **10-20** | **20-30** | **30-40** | **40-60** | **60-80** | **80-100** |
| Cluster 1 | 354 | 218 | | | | | | |
| Cluster 2 | 59 | 142 | 138 | 4 | | | | |
| Cluster 3 | 105 | 146 | 32 | 3 | | | | |
| Cluster 4 | 99 | 100 | 104 | 44 | 23 | 6 | | |
| Cluster 5 | 12 | 22 | 31 | 39 | 43 | 44 | 38 | 32 |
| Cluster 6 | | 1 | 89 | 90 | 84 | 49 | 17 | 7 |
| Cluster 7 | | | 48 | 301 | 345 | | | |
| Cluster 8 | | | 188 | 149 | 134 | | | |
| Cluster 9 | | | | | | 343 | 381 | 375 |
| Cluster 10 | | | | | | 186 | 192 | 214 |

Em ambos os cenários todos os clusters gerados possuem camadas consecutivas, como observado na tabela 1 cujos clusters gerados foram reorganizados para melhor visualização. A média de teor de carbono é menor nos clusters que possuem exclusivamente instâncias das camadas mais profundas (11,36 g/kg no

Cluster 9 e 9,36 g/kg no Cluster 10) e a média de teor de carbono é maior nos clusters que possuem instâncias das camadas mais superficiais (25,53 g/kg no Cluster 1, 22,13 g/kg no Cluster 2 e 20,84 g/kg no Cluster 3). A presença de instâncias de uma mesma camada em diferentes clusters também pode indicar a variação de valores presente em cada camada. Camadas superficiais estão presentes em um número maior de clusters do que as camadas mais profundas indicando que os resultados da clusterização são consistentes com os resultados apresentados na figura 1.

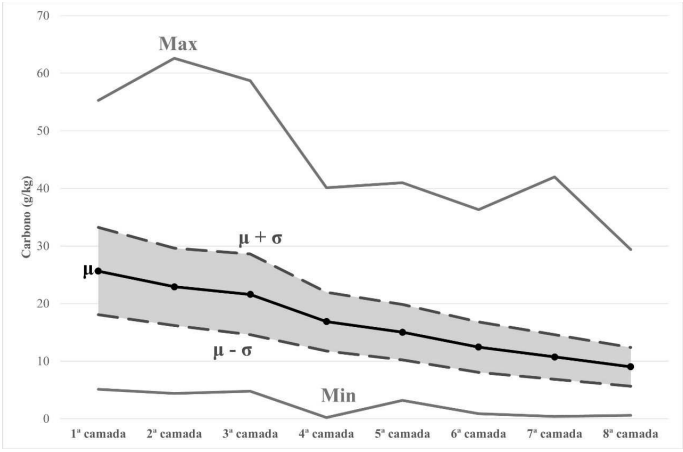


Figura 1: Distribuição do teor de carbono ao longo das camadas em profundidade.

O cenário com preenchimento dos valores faltantes pela média das camadas agrupou o conjunto de dados com um menor número de clusters (de 14 para 10) e menor coeficiente de variação médio (de 36,7% para 32,3%). Um menor coeficiente de variação indica maior coesão entre os elementos dentro de um cluster.

## 5. Conclusões

Ambos os cenários analisados apresentam entre os clusters de maior média de teor de carbono (em g/kg) aqueles que possuem instâncias das camadas superficiais, principalmente as duas primeiras. Enquanto para os clusters com menor média de teor carbono (em g/kg) estão presentes instâncias das camadas mais profundas, principalmente as três últimas.

Na comparação entre cenários, podemos destacar a diferença entre número de clusters gerados pelo mesmo método e que no caso do cenário 1 alguns clusters englobavam completamente variáveis faltantes (e.g. o cluster 4 reuniu elementos faltantes em uma única camada), algo que não poderia ocorrer no cenário 2.

Verificando as estatísticas de cada cenário obtemos que a média do coeficiente de variação do cenário 1 é maior se comparada à média do coeficiente de variação do cenário 2. Logo, o cenário 2 sem valores faltantes conseguiu agrupar o conjunto de dados em menos clusters e com uma variação menor interna em cada um.

Comparando os resultados do cenário 2 com as estatísticas dos dados por camadas descobrimos que alguns clusters encontrados conservam a distribuição média de carbono entre camadas e sua variância, preservando assim a estrutura das amostras e servindo como ferramenta importante na análise dos dados.

Este estudo verificou através do uso de agrupamento de dados que o teor de carbono no solo segue uma regra de profundidade na qual o teor de carbono, em média, decresce de maneira monótona de acordo com a profundidade do solo.

Finalmente, a técnica de clusterização se mostra uma ferramenta com potencial para a análise dos dados de teor e estoque de carbono no solo, podendo ser utilizada em estudos futuros para agrupar os dados de acordo com o tipo de manejo adotado, diferentes tipos de solo, clima e biomas.

## Agradecimentos

Gostaríamos de agradecer às empresas Bayer e EMBRAPA e seus colaboradores pela disponibilização dos dados e o compromisso com o desenvolvimento da pesquisa.

## Referências

Banwart, S. A., Noellemeyer, E., and Milne, E. (2014). *Soil carbon: Science, management and policy for multiple benefits*, volume 71. CABI.

Do, C. B. and Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897–899.

Eibe, F., Hall, M. A., and Witten, I. H. (2016). *The WEKA workbench. Online appendix for data mining: practical machine learning tools and techniques.* Morgan Kaufmann.

Miranda, E. E. (2018). Compare: ocupação e uso das terras no Brasil e nos EUA. *Agro DBO*, 14(96):38–39.

Witten, I. H., Frank, E., and Hall, M. A. (2005). Practical machine learning tools and techniques. In *Data Mining*, volume 2. Morgan Kaufmann, 3rd edition.