Aplicação do Algoritmo de k-Means na Visualização de Mapas Digitais de Elevação de Solo na Região do Recôncavo Baiano

Gabrielle S. Pereira[1], Felipe Henriques[1], Renato Mauro[1], Diego Brandão[1], Marcos B. Ceddia[2]

[1]Programa de Pós-graduação em Ciência da Computação
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ
Rio de Janeiro, RJ, Brasil

[2]Instituto de Agronomia
Universidade Federal Rural do Rio de Janeiro
Seropédica, RJ, Brasil

gabrielle.pereira.1@aluno.cefet-rj.br, marcosceddia@gmail.com

{diego.brandao,felipe.henriques,renato.mauro}@cefet-rj.br

Resumo. *Os impactos econômicos devido à degradação do solo são cada vez mais claros. Dados da Organização das Nações Unidas (ONU) mostram que devido a compactação do solo, fenômenos como enchentes e secas são cada vez mais comuns no mundo. Dessa forma, a análise dos dados de solo é fundamental para o planejamento ambiental, produção agrícola, uso da terra e bem estar da população. O mapeamento digital de solos consiste em um conjunto de técnicas que facilitam tais análises, agregando informações oriundas dos experimentos laboratoriais com informações de imagens de satélite e drones, com técnicas matemáticas para gerar modelos descritivos do solo mais precisos. Neste contexto, este trabalho demonstra uma aplicação de mapeamento digital de solos, especificamente o algoritmo de k-Means é utilizado na construção de um modelo digital de elevação da região do Recôncavo Baiano.*

1. Introdução

A qualidade e a saúde do solo estão diretamente relacionadas à manutenção e melhoria global dos recursos do solo para produção de alimentos, contribuindo para a saúde humana, agricultura de precisão e sequestro de carbono, que é o processo de remoção do gás carbônico da atmosfera [Davies 2017]. A informação detalhada do solo é fundamental para estudar o impacto oriundo da sua degradação [Benedetti 2008].

A pedologia é a área da ciência responsável pelo processo de conhecimento do solo, de sua caracterização e mapeamento. Nos últimos anos essa tarefa de mapeamento do solo, como outras áreas da ciência, sofreu um processo de informatização intenso, a área de mapeamento digital dos solos surgiu agregando as análises tradicionais, informações oriundas de satélites, drones, sensores entre outros. Essa evolução foi ainda mais intensa com a utilização de técnicas de aprendizado de máquina, onde inúmeras pesquisas têm sido realizadas recentemente. Uma busca pelas palavras-chave "Digital mapping soil" e "Machine learning" na base de dados Google Scholar retorna pouco mais de 6 mil artigos no ano de 2020 enquanto que para 2021 até o mês de outubro esse número já ultrapassa a marca de 11 mil artigos.

Apesar de toda essa evolução do processo de mapeamento de solo, o processo de coleta de dados ainda é uma tarefa complexa e de alto custo, pois além de todas as dificuldades das análises laboratoriais, muitas vezes a área a ser analisada é bem extensa e/ou de difícil acesso, podendo tornar o estudo até inviável. Neste contexto, [Lagacherie et al. 1995] propõem uma técnica para determinar as chamadas áreas de referência, isto é, sub-regiões dentro de uma área que conseguem representar toda a área estudada. Para tanto, eles utilizam mapas e informações pedológicas, objetivando encontrar padrões entre áreas. Tais padrões podem ser analisados por meio da caracterização da cobertura de solo em pequenas áreas, realizando um levantamento detalhado delas. Com isso, essa Área de referência identifica todas as classes de solo de uma região maior, caracterizando-se como uma amostra da população (região).

A construção das áreas de referência exige o conhecimento de outras características do solo e que os dados estejam na mesma escala/ordem de grandeza. Assim, torna-se necessário a construção das bases de dados de cada uma dessas características. Uma dessas características é a determinação da altimetria, ou seja, a elevação da região em relação ao nível do mar.

Este trabalho tem como objetivo aplicar o algoritmo k-Means para determinar um mapa de elevação a partir de informações de solo da região do Recôncavo Baiano. Essa abordagem é um passo inicial na processo de determinação das sub-áreas representativas de uma região. Além desta seção, este artigo está organizado em outras quatro seções. Na Seção 2 é apresentado a fundamentação teórica, a Seção 3 apresenta o material e a metodologia que são utilizados, a Seção 4 apresenta os resultados das visualizações e a Seção 5 apresenta as considerações finais.

2. Fundamentação Teórica

Esta seção introduz os conceitos utilizados no desenvolvimento deste trabalho.

2.1. Mapeamento Digital de Solos

O Mapeamento Digital de Solos (MDS) consiste no desenvolvimento de um sistema de informações espaciais de solos, a partir de modelos numéricos. Esses possuem o objetivo de inferir as variações espaciais, temporais e propriedades de solos a partir de observações e dados de variáveis correlacionadas do ambiente [Lagacherie and Mcbratney 2007].

A abordagem proposta por [McBratney et al. 2000] consiste em técnicas quantitativas utilizadas em pedometria visando explicar as relações entre os fatores de formação do solo. Além disso, também buscou predizer as classes de solo ou as suas propriedades em uma determinada região, através da geoestatítica, lógica *Fuzzy*, redes neurais artificiais, regressões múltiplas e árvores de decisão [McBratney et al. 2003].

2.2. Modelo Digital de Elevação

Características do relevo são utilizadas em levantamentos de solo, pois estão relacionadas aos processos de formação do mesmo [Klingebiel et al. 1987]. Um MDE é definido por qualquer representação digital de uma variação contínua do relevo no espaço [Burrough 1986]. A representação usual desses dados é através de uma matriz, ou imagem *raster*, em que os elementos, ou *pixels*, tem como atributo o valor da elevação do terreno em relação à um dado referencial. As principais características de um relevo que

podem ser extraídas do MDE são as variáveis de declividade, elevação e curvas de nível [Collins 1981].

As formas tradicionais de representação do terreno são feitas através de curvas de níveis, cartas topográficas e pontos cortados. São feitas em termos qualitativos baseados nas descrições da fase de interpretação do terreno [Burrough et al. 2015]. O avanço da tecnologia contribuiu com a ampliação do uso de modelos digitais obtidos através do processamento de imagens de sensores de RADAR, fotografias aéreas e sensores ópticos, por exemplo, permitindo a modelagem e análises do aspecto físico de uma área [Burrough et al. 2015].

2.2.1. SRTM

A missão *Space Shuttle Topography Mission* (SRTM) foi realizada em conjunto *National Aeronautics and Space Administration* (NASA), *National Geospatial-Intellingence Agency* (NGA) e as agências espaciais da Alemanha e Itália no ano 2000, obtiveram dados topográficos de quase 90% da terra. Dados de radar foram coletados e disponibilizados com uma resolução espacial de 90 metros [FARR 2007].

No Brasil, o projeto TOPODATA do Instituto Nacional de Pesquisas Espaciais (INPE), disponibiliza o MDE e derivações básicas a nível nacional elaboradas a partir do SRTM. Esses arquivos foram disponibilizados corrigindo falhas e realizados refinamentos e pós-processamento. Esses dados foram refinados a partir de uma resolução espacial original de 3 arco-segundos, aproximadamente 90 metros, para 1 arco-segundo, aproximadamente 30 metros, por krigagem. Algoritmos de análise geomorfométricas foram aplicados sobre os dados refinados, para o cálculo das variáveis de declividade e curvatura vertical. [TOPODATA 2021].

2.3. K-Means

O algoritmo K-*means* é um método de agrupamento de dados, sendo relativamente simples de ser implementado [Han et al. 2011]. O Objetivo é buscar similaridade entre os registros, segmentando os P dados em K grupos, visando minimizar a distância entre os *clusters* [Likas et al. 2003]. Existem diversos métodos para o cálculo de distâncias entre os dados, para este trabalho, foi utilizada a distância Euclidiana como medida de dissimilaridade entre cada *pixel* de uma imagem e o centróide do *cluster* [Piloyan 2017].

Neste algoritmo, todos os *pixels* foram classificados baseados nas suas distâncias entre os *clusters*, tendo como input o números desejado de entrada de N *clusters*. A saída é uma imagem com cada *pixel* agrupado em um dos N *clusters* [Borra et al. 2019].

Com base na abordagem proposta por [Piloyan 2017], na qual utiliza um algoritmo de K-*means*, para classificação não supervisionada, de um *software* de análise espacial chamado *Whitebox* para um MDE de uma região da Armênia. É um processo iterativo de classificação de cada *pixel* de uma imagem, que encontra grupos estatisticamente semelhantes durante a análise. Os critérios de parada foram definidos em 7 classes, 25 iterações e 2% de tolerância para mudanças de classes de *pixels*. Os resultados mostram que é possível derivar

3. Material e métodos

Esta seção apresenta a origem dos dados utilizados, os softwares empregados, bem como uma breve introdução a técnica de k-Means.

3.1. Caracterização da área de estudo

A área de estudo escolhida situa-se Estado da Bahia, nordeste do Brasil, delimitada pela Bacia Sedimentar do Recôncavo Baiano, a área foi escolhida por ser uma região de grande importância para a produção e exploração de petróleo no Brasil. Encontra-se aproximadamente entre as latitudes 11º30'00"S e 13º30'00"S e as longitudes 36º30'00"W e 39º30'00"W. A área possui cerca de 12.000 km² e abrange 40 municípios incluindo a capital, Salvador.

A Figura 1 ilustra espacialmente a localização da área de estudo, bem como informações hidrográficas e seus principais trechos de rios e corpos d'água.

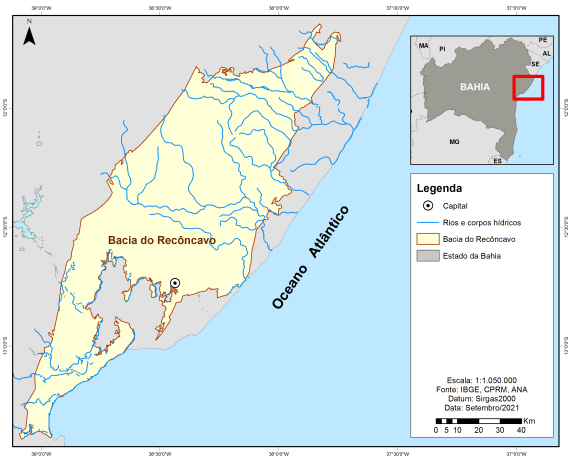

Figura 1. Localização Bacia do Recôncavo. Fonte: o autor

3.2. Material utilizado

Para a elaboração das visualizações dos mapas deste trabalho, foi necessário organizar as bases cartográficas em um único sistema de referência. As bases de dados foram transformadas em SIRGAS2000 (Sistema de Referência Geocêntrico para as Américas) quando não encontram-se neste sistema, pois este é o sistema de referência adotado oficialmente no Brasil em 25 de Fevereiro de 2005 [Fortes et al. 2007]. A seguir foram especificadas as bases utilizadas:

1. Delimitação dos Estados e Municípios do Brasil: Instituto Brasileiro de Geografia e Estatística (IBGE)
2. Delimitação das bacias sedimentares: Serviço Geológico do Brasil (CPRM)
3. Imagens SRTM: Topodata INPE

3.3. Software

Neste estudo, empregou-se o *software* de Sistema de Informação Geográfica (SIG) ArcGis 10.2 da Esri. A extração de parâmentros geomorfométricos das imagens SRTM, foram realizadas no *software open source* SAGA GIS 2.3.2

3.4. Metodologia

O Modelo Digital de Elevação (MDE) foi obtido através da classificação semi-automática das imagens SRTM referentes às folhas 11S39-ZN, 13S39-ZN, 13S405-ZN do projeto Topodata. Os parâmetros de terreno, como elevação e declividade, foram usados em uma única imagem e foram processadas no *software* SAGA.

Implementou-se a análise através do módulo K-*Means Clustering for Grids* do SAGA, onde os critérios de parada são o número máximo de iterações e o número mínimo de *clusters*. Neste *software*, o algoritmo começa localizando aleatoriamente os K *clusters* no espaço espectral. Cada *pixel* no grupo de imagem de entrada é atribuído ao centroide do *clusters* mas próximo. Essa classificação é repetida até a condição de parada ser alcançada [Piloyan 2017].

Neste trabalho os parâmetros utilizados foram determinados por 10 *clusters* e 30 iterações. Os parâmetros podem ser modificados e definidos com base mas características dos objetos de estudos, assim como a qualidade do resultado do processamento e o limite de processamento do computador utilizado. Foram realizados testes com diferentes combinações entre números de *clusters* e iterações.

4. Resultados

Este trabalho foi realizado em uma máquina com processador Intel(R) Core(TM) i7-4770, memória RAM 8GB, placa de vídeo NVIDIA GeForce GT 635 1GB, sistema operacional Windows versão 8.1 Pro.

4.1. Classificação de MDE

A tabela 1 é a análise da quantidade de elementos de cada *clusters* após o processamento da imagem MDE através da metodologia de classificação não supervisionada k-*means* da região da bacia do Recôncavo com o *software* SAGA. Nota-se que o *clusters* 1 é o que possui o maior número de elementos. Observando-se o mapa da Figura 2, este *clusters* representa altitudes entre 131 e 160 metros. O *clusters* com o menor número de elementos é o de número 10, que contém altitudes entre 201 e 245 metros. O Mapa da Figura 2 reflete as classes altimétricas existentes, aponta altitudes mais baixas próximas ao nível do mar, o que era esperado. Em vermelho nota-se os *clusters* que representam as maiores altitudes.

O Mapa da Figura 2 reflete as classes altimétricas existentes, aponta altitudes mais baixas próximas ao nível do mar, o que era esperado. Em vermelho nota-se os *clusters* que representam as maiores altitudes.

Tabela 1. Análise resultado processamento k-*means* (fonte: o autor)

Cluster	Elementos
1	1900230
2	1652634
3	1624068
4	1166132
5	859370
6	578875
7	397822
8	339789
9	1065049
10	209376

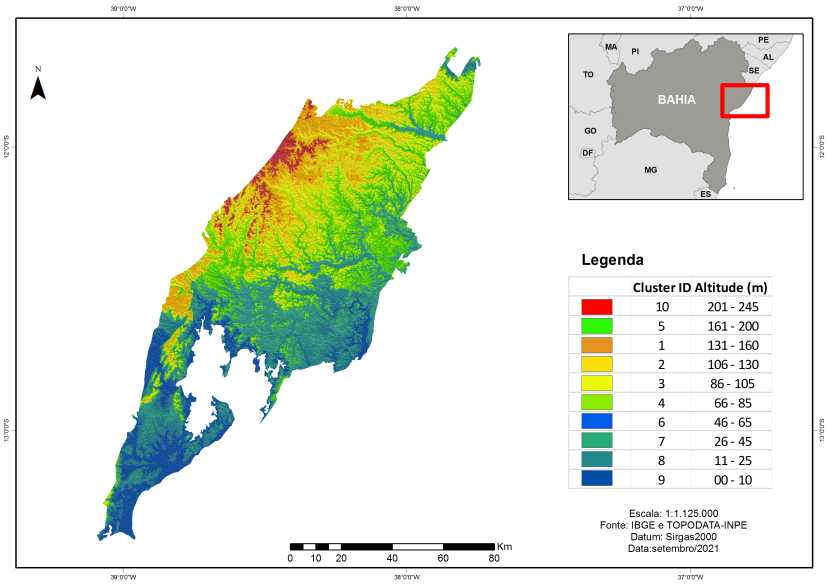

Figura 2. Classificação de MDE. Fonte: o autor

5. Conclusão

O presente trabalho teve como objetivo apresentar um estudo inicial sobre o mapeamento digital de solos com a utilização de algoritmos de aprendizado de máquina na região do Recôncavo Baiano. O mapa de elevação da região foi apresentado. A imagem SRTM originou o MDE, com 10 *clusters* que representam as altitudes da região. Sugestão para trabalhos futuros são: a caracterização do meio físico da área de estudo, classificação da vegetação, relevo, hidrografia e geomorfologia, um comparativo com os mapeamentos convencionais de solo e os digitais e as análises dos mapeamentos pedológicos existentes.

É válido ressaltar que resultados de MDE são difíceis de serem verificados quantitativamente devido à ausência de dados coletados em campo de características geomorfológicas e de altitudes neste trabalho. A vantagem do método aqui utilizado é que requer um número pequeno de parâmetros, além disso, foi possível a utilização de dados públicos

e processa-los em um *software open source*.

Referências

[Benedetti 2008] Benedetti, Marcelo, e. a. (2008). Representatividade e potencial de utilização de um banco de dados de solos do brasil. *Revista Brasileira de Ciência do Solo*, 32:2591–2600.

[Borra et al. 2019] Borra, S., Thanki, R., and Dey, N. (2019). *Satellite image analysis: clustering and classification*. Springer.

[Burrough 1986] Burrough, P. A. (1986). Principles of geographical information systems for land resources assessment. clarendon.

[Burrough et al. 2015] Burrough, P. A., McDonnell, R. A., and Lloyd, C. D. (2015). *Principles of geographical information systems*. Oxford university press.

[Collins 1981] Collins, H. S. (1981). Algorithms for dense digital terrain models. *PHOTOCRAMMETERNIGCI NEERINANGD REMOTES ENSING*, 47:71–76.

[Davies 2017] Davies, J. (2017). The business case for soil. *Nature*, 543:309–311.

[FARR 2007] FARR, T. G., e. a. (2007). The shuttle radar topography mission. *Reviews of Geophysics,*, 45.

[Fortes et al. 2007] Fortes, L., Costa, S., Lima, M., Fazan, J., and Santos, M. (2007). Accessing the new sirgas2000 reference frame through a modernized brazilian active control network. *Dynamic Planet. International Association of Geodesy Symposia*, 130.

[Han et al. 2011] Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[Klingebiel et al. 1987] Klingebiel, A., Horvath, E., Moore, D., and Reybold, W. (1987). Use of slope, aspect, and elevation maps derived from digital elevation model data in making soil surveys. *Soil survey techniques*, 20:77–90.

[Lagacherie et al. 1995] Lagacherie, P., LEGROS, J. P., and BURROUGH, P. (1995). Soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. *Geoderma*, pages 283–301.

[Lagacherie and Mcbratney 2007] Lagacherie, P. and Mcbratney, A. B. (2007). Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. in: Lagacherie, p. et al. digital soil mapping: an introductory perspective. amsterdam:. *Elsevier*, pages 3–22.

[Likas et al. 2003] Likas, A., Vlassis, N., and J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461. Biometrics.

[McBratney et al. 2003] McBratney, A., Mendonça Santos, M., and Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117(1):3–52.

[McBratney et al. 2000] McBratney, A. B., Odeh, I. O., Bishop, T. F., Dunbar, M. S., and Shatar, T. M. (2000). An overview of pedometric techniques for use in soil survey. *Geoderma*, 97(3):293–327.

[Piloyan 2017] Piloyan, A. (2017). Semi-automated classification of landform elements in armenia based on srtm dem using k-means unsupervised classification. *Quaestiones Geographicae*, 36:93–103.

[TOPODATA 2021] TOPODATA (2021). Banco de dados geomorfométricos do brasil | topodata. Disponível em: http://www.dsr.inpe.br/topodata/dados.php. Acesso em: 04 set 2021.