Comparativo entre os algoritmos K-Means e ckMeans para mapeamento automatizado de uso do solo

Sidnei Luís Bohn Gass[1]
Cristiano Galafassi[1]
Rogério Rodrigues de Vargas[1]

[1] Universidade Federal do Pampa – UNIPAMPA - Campus Itaqui
LabSIM – Laboratório de Sistemas Inteligentes e Modelagem
Rua Luís Joaquin de Sá Brito, sn, CEP 97650-000, Itaqui, RS, Brasil
sidneigass@unipampa.edu.br
cristianogalafassi@unipampa.edu.br
rogeriovargas@unipampa.edu.br

Abstract. Remote sensing allow us to acquire information about an object or phenomenon without the need to make physical contact with the object, which turn it usable in many fields, like hydrology, ecology, oceanography, glaciology, geology. Remote Sensing generally refers to the use of satellite-based (or aircraft) sensor technologies to detect and classify objects on Earth. Classification is the process of extracting information in images (or data) to recognize patterns and homogeneous objects and are used in remote sensing to map areas of the earth's surface. This article makes a comparison between two algorithms used in image classification applied to remote sensing. The first one is the well-known K-Means, that has the characteristic to be fast and its modeling is relatively simple, and the second is the fuzzy ckMeans algorithm that allows to model inaccurate data according to their membership degree. The ckMeans algorithm proved to be a good alternative in the image segmentation process. To validate the work we compared the classification of an image, obtained by a satellite, of the western border of the state of Rio Grande do Sul and defined a priori four clusters. Then, the classification between K-Means and ckMeans algorithms was performed. Finally, a domain knowledge specialist discussed the resultant classification obtained by these algorithms.

Palavras-chave: image processing, data grouping, land cover, processamento de imagens, agrupamento de dados, cobertura da terra.

1. Introdução

O sensoriamento remoto, para Jensen (2007), enquanto ciência, é uma ferramenta similar à matemática, pois a extração de informações a partir dos dados armazenados em imagens com o uso de algoritmos e procedimentos estatísticos é uma atividade científica. Por sua vez, pode ainda ser considerada uma arte, pois o processo de interpretação visual de uma imagem exige não apenas conhecimento científico, mas conhecimento acumulado ao longo da vida sobre os elementos que compõem determinados espaços.

Dentre os procedimentos de classificação de uma imagem de satélite, podem ser citados os processos de agrupamento, ou seja, aqueles que buscam agrupar os dados a partir de um determinado grau de similaridade, usando técnicas de distâncias mínimas (Meneses & Almeida, 2012). Os trabalhos de Bandyopadhyay & Maulik (2002), Rebouças, et al. (2015), Maciel, Vinhas & Câmara (2015) e Augusto-Silva, et al. (2013), demonstraram, ao tratar de temáticas como a análise de diferentes classificadores para o mapeamento de uso e cobertura do solo, algoritmos de agrupamento para o tratamento de dados de sensoriamento remoto para a separação de culturas agrícolas e tipos de uso do solo, a importância do tratamento matemático para a otimização de resultados. Além disso, é perceptível a importância da adaptação e complementação de alguns algoritmos já consolidados, com o intuito de melhorar o resultado gerado com a sua aplicação.

O algoritmo K-Means (Macqueen, 1967) está entre os algoritmos mais conhecidos e utilizados na tarefa de classificação pois a sua implementação é relativamente simples e eficaz. O objetivo deste método é particionar n observações em K grupos (centroides) onde cada observação pertence ao grupo mais próximo da média do centroide. Em outras palavras,

decide-se a qual grupo o dado pertence através da menor distância entre ele e o centro dos grupos.

Nos algoritmos utilizados para classificação tipo *fuzzy* os elementos são associados a mais de um grupo e assim, é estabelecido um grau de pertinência a cada dado. Dessa forma, cada dado pertence com um certo grau de verdade a cada grupo, mas com diferentes graus de pertinência (Nascimento, Mirkin, 2000). O algoritmo de agrupamento *fuzzy* mais popular é o Fuzzy C-Means, proposto por Dunn (1973) e estendido por Bezdek (1981), o qual possui a ideia de funcionamento análogo ao algoritmo K-Means. Este algoritmo, no processo de segmentação de imagens, mede a similaridade entre cada pixel e cada padrão existente na imagem. Quanto mais próximo do valor 1 maior é o indicativo de que o pixel pertence ao grupo em questão. Se for mais próximo de zero, significa que são pequenas as chances do pixel pertencer ao agrupamento analisado.

Em Vargas & Bedregal (2010) foi proposto um algoritmo *fuzzy* denominado ckMeans. Este algoritmo é uma extensão do algoritmo Fuzzy C-Means que utiliza a ideia do algoritmo K-Means na forma de calcular os centroides. Este algoritmo demonstrou ser mais rápido e classificou melhor os dados em alguns experimentos conforme pode ser visto em Vargas & Bedegral (2010) e Vargas et al. (2011).

Nos trabalhos de Vargas et al. (2014) e Vargas et al. (2016) os autores propuseram a aplicação do algoritmo ckMeans no processo de segmentação de imagens. Entretanto, não fizeram uma comparação qualitativa com outro algoritmo de agrupamento. Este trabalho tem por objetivo apresentar um comparativo entre o agrupamento obtido pelos algoritmos K-Means e ckMeans sobre uma imagem de satélite para fins de classificação não supervisionada de uso do solo.

O trabalho está organizado em quatro seções. A seção 2 apresenta uma descrição acerca da área de estudo, bem como as etapas de desenvolvimento do experimento. Na seção 3 são discutidos os resultados obtidos através da comparação dos dois métodos de classificação e, por fim, as considerações finais são apresentadas na seção 4.

2. Materiais e métodos

2.1 A área de estudo

Para a aplicação do presente estudo foi selecionado um fragmento da cena 163_133 da imagem do satélite CBERS-4, instrumento imageador PAN, de 23 de outubro de 2015, localizada na fronteira oeste do Rio Grande do Sul, entre os municípios de Itaqui, Uruguaiana e Alegrete, como pode ser verificado nas Figuras 1 e 2. A área selecionada caracteriza-se pelo seu relevo plano, com a presença de rios de grande porte (rios Uruguai e Ibicuí) e pelo predomínio das áreas de cultivo de arroz e pastagem para a pecuária.

Em função das características morfometrias das áreas a montante da confluência entre os rios Uruguai e Ibicuí, este ponto recebe uma grande quantidade de água quando ocorrem eventos de precipitação extrema nas áreas mais próximas as nascentes. Como na região da foz do rio Ibicuí o relevo é mais plano, o acúmulo de água abrange áreas maiores, provocando processos de inundação mais abrangentes, como já demonstrado nos estudos de Gass (2015) e Ogassawara (2015).

Na área selecionada para o estudo, as declividades variam entre 0 e 3º, representando ambientes de formação de relevo de planícies fluviais ou flúvio-lacustres. De acordo com Viero & Silva (2010), a região em questão é formada por “ambiente de planície aluvionar recente” com a presença de “material inconsolidado e de espessura variável que da base para o topo é formado por cascalho, areia e argila”.

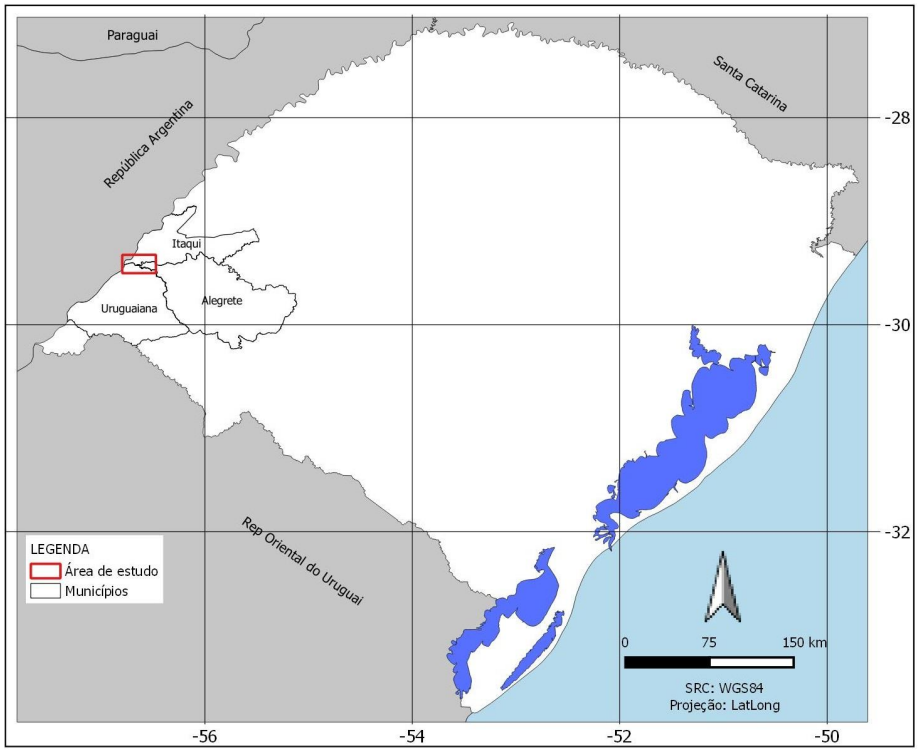

Figura 1. Localização da área de estudo.

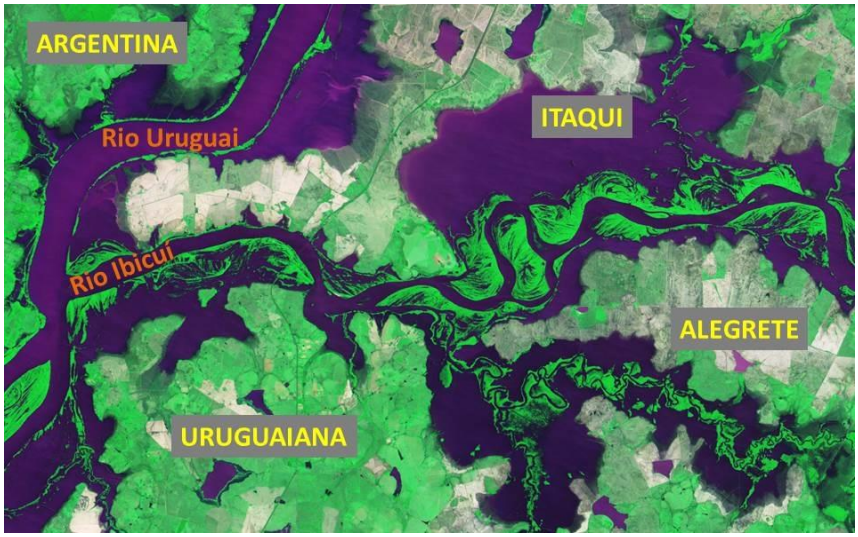

Figura 2. Composição colorida 3-4-2 em RGB.

As características do solo predominante na região, corroboram com os dados apresentados, por serem Plintossolos Argilúvicos eutrófico petroplinticos, com lençol freático alto (próximo a superfície), drenagem imperfeita e textura baseada em argilas expansíveis, o que aumenta a capacidade e retenção da água, como pode ser verificado em Embrapa (2013).

2.2 Materiais utilizados

Para a aplicação dos métodos de agrupamento, foi utilizado um fragmento da cena 163_133 da imagem do satélite CBERS-4, instrumento imageador PAN10, de 23 de outubro de 2015, com 10 metros de resolução espacial e 8 bits de resolução radiométrica. A resolução espectral das bandas utilizadas para os processamentos é: banda 2 (0,52 - 0,59 µm), banda 3

(0,63 - 0,69 µm) e banda 4 (0,77 - 0,89 µm), que correspondem, respectivamente, aos comprimentos de onda do verde, do vermelho e do infravermelho próximo.

Cabe ressaltar que as motivações que levaram a equipe a optar por imagens do satélite CBERS-4 foram as seguintes: 1) o satélite foi desenvolvido com parcela de investimentos públicos brasileiros; 2) é o primeiro satélite da série com imagens com 10 metros de resolução espacial; 3) as imagens são de uso gratuito, e, 4) a pouca disponibilidade, até o momento, de estudos que utilizaram tais imagens.

A aplicação do algoritmo K-Means foi efetuado a partir do software TerrSet 18.21 (Eastman, 2016) em um notebook com sistema operacional Windows®, versão 10, com 8 GB de memória RAM e processador Core i7 de 4ª geração. Por sua vez, o algoritmo ckMeans foi desenvolvido em linguagem C/C++ e executado em um computador similar ao utilizado para o algoritmo K-Means.

2.3 As etapas de trabalho

A primeira etapa do trabalho consistiu na escolha da imagem a ser utilizada e sua aquisição através do catálogo de imagens do INPE, disponível no endereço http://www2.dgi.inpe.br/CDSR. Após o download das bandas da cena selecionada, as mesmas foram importadas no TerrSet, analisadas e foi realizado o recorte para a área de interesse, a qual está representada na Figura 2, já apresentada anteriormente. O recorte resultou numa imagem com dimensões de 3155x1963x3 pixels, sendo as coordenadas do seu centroide 29° 24' 44,21087"S e 56° 38' 01,14281"W, recobrindo uma área de 61932 hectares.

A composição colorida 3-4-2 em RGB e as bandas individualizadas foram exportadas para o formato JPG para serem utilizadas como base para a aplicação do algoritmo ckMenas, como descrito no item materiais utilizados. No TerrSet, rodou-se o algoritmo K-Means, utilizando como arquivos de entrada, os recortes das três bandas individuais da imagem selecionada para o estudo.

Os parâmetros de entrada para o algoritmo K-Means foram: número máximo de clusters = 4, método de inicialização = eixo diagonal, porcentagem de migração de pixels inferior a 5%, número máximo de iterações = 50, número mínimo de pixels por cluster = 5%. Já os parâmetros de inicialização do algoritmo ckMeans foram 4 clusters, a inicialização deu-se de forma aleatória e épsilon em 0,01 como critério de parada.

3. Resultados e discussões

A partir da composição colorida 3-4-2 em RGB (Figura 2), é possível identificar os leitos dos rios Uruguai e Ibicuí, além de áreas com solo exposto, vegetação ciliar e áreas com cultivos agrícolas. Em função do grande acúmulo de chuvas ocorrido no período, verifica-se a presença de áreas inundadas que são o resultado do processo natural do mútuo represamento dos rios mencionados e do aumento dos reservatórios de água para a irrigação da cultura do arroz e sua interligação com o extravasamento fluvial. Estes processos estão diretamente associados às características da região, mencionadas no item 2.1 (a área de estudo) do presente trabalho.

As imagens resultantes da aplicação dos algoritmos K-Means e ckMeans (Figuras 3 e 4, respectivamente), representam quatro classes de usos do solo (grupos), que podem ser assim definidas: massas de água, solo exposto, uso agrícola 1 e uso agrícola 2. A classes massas de água representa os rios Uruguai e Ibicuí, os reservatórios para fins de irrigação, as áreas inundadas e em alguns casos os pixels nos quais há uma significativa presença de umidade, mesmo que seja possível identificar seu uso como agrícola. Morfologicamente, estas áreas estão associadas as planícies fluviais.

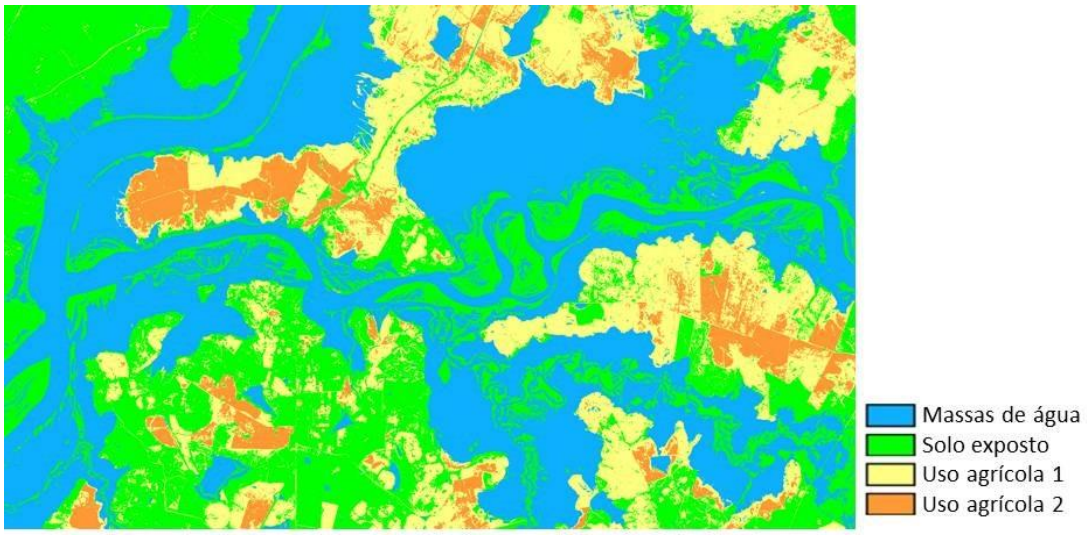

Figura 3. Resultado do processamento com o algoritmo K-Means.

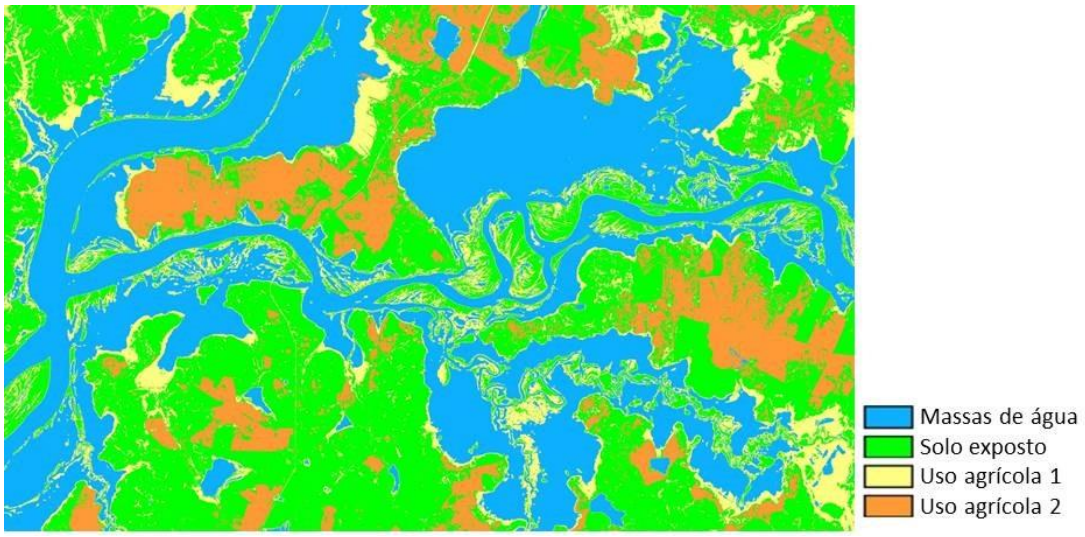

Figura 4. Resultado do processamento com o algoritmo ckMeans.

O algoritmo ckMeans, no caso da classe em questão, conseguiu agrupar de maneira mais precisa os pixels que representam efetivamente água, incluindo em outras classes aquelas áreas que demonstram presença de umidade, mas que é perceptível a presença de algum tipo de vegetação. Desta forma, considerando a possibilidade do uso do algoritmo para a identificação de áreas inundadas, sua aplicação gera resultados satisfatórios.

A classe solo exposto, na imagem resultante do algoritmo K-Means, apresentou resultados melhores nas áreas em que a reflectância dos alvos era maior. Nas áreas com presença de umidade, ocorreu a redução da reflectância, gerando ruídos e tendendo assim a incorporar tais pixels predominantemente na classe uso agrícola 1. Este fato se deve ainda em função da estreita correlação que pode ser estabelecida com as áreas em que os cultivos agrícolas estavam iniciando seu processo de germinação, permitindo a identificação das plantas em conjunto com o solo, em função do baixo índice de vegetação associado ao processo de desenvolvimento das plantas. Por sua vez, o resultado do algoritmo ckMeans, agrupou as áreas de solo exposto com as áreas em processo inicial de germinação além de incluir áreas já em processo mais avançado de desenvolvimento, o que gera certa inconsistência na sua avaliação.

Para a classe uso agrícola 1, é possível verificar que o algoritmo K-Means associou a ela os pixels das áreas em processo mediano de desenvolvimento, ou seja, aquelas em que a vegetação minimamente recobre o solo, permitindo assim a sua dissociação das áreas de solo exposto e das áreas de desenvolvimento fenológico mais avançado, quando se percebe o maior índice de vegetação. O algoritmo ckMeans, para esta classe, associou basicamente os pixels que podem ser caracterizados como sendo de alta presença de umidade mas ainda com possibilidade de identificação da vegetação. Isto nos leva a inferir que, comparando as imagens resultantes dos dois algoritmos, é nesta classe que o ckMeans colocou ou pixels que permitiram refinar o resultado da classe massas de água.

A última classe definida para o experimento é a denominada de uso agrícola 2. Nesta classe devem estar presentes os pixels que representam aquelas áreas nas quais há uma maior presença de áreas agrícolas já mais desenvolvidas e as áreas de vegetação ciliar e eventuais áreas florestadas. O algoritmo ckMeans apresentou um resultado com alto grau de generalização, agrupando, em sua grande maioria, os pixels que representam as características aqui definidas para as duas últimas classes.

4. Considerações finais

O presente trabalho fez uma análise comparativa do resultado de dois algoritmos para a classificação do processo de segmentação de imagens de satélite. Para testes, utilizou-se uma imagem localizada nos municípios de Itaqui, Uruguaiana e Alegrete, na região oeste do estado do Rio Grande do Sul, abrangendo parte dos rios Uruguai e Ibicuí.

Tanto o algoritmo K-Means quanto o algoritmo ckMeans se mostraram flexíveis e habilitados para auxiliar o sensoriamento remoto nos seus processos de tratamento de imagens, pois cada imagem é processada sem depender de resultados anteriores. Na imagem discutida, onde se percebe uma variação maior na intensidade e na forma dos elementos, o algoritmo K-Means é mais eficaz em casos onde a imagem tem agrupamentos bem definidos. Por outro lado, o algoritmo ckMeans permite classificar a imagem onde os grupos tendem a se sobrepor em função de suas características.

É possível inferir que os resultados apresentados pelo algoritmo ckMeans são mais precisos nos casos em que as imagens originais apresentam um conjunto de pixels mais homogêneo, permitindo assim a extração de áreas core, como foi possível verificar com as massas de água. Por sua vez, para a caracterização mais acurada do uso do solo, o algoritmo K-Means se mostrou mais eficiente.

Dada a característica do algoritmo ckMeans em lidar com graus de pertinência, como trabalho futuro, pretende-se mapear áreas de incertezas. Uma vez que cada pixel é classificado em todos os grupos, pretende-se criar um limiar e destacar os pixels com baixo grau de pertinência, apoiando-se no conhecimento do especialista para a tomada de decisões.

Referências

Augusto-Silva, P. B.; Valério, L. P.; Santos, T. B. dos; Alcântara, E. H. de; Stech, J. L. Análise de classificadores para mapeamento de uso e cobertura do solo. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 16. (SBSR), 2013, Foz do Iguaçu. **Anais...** São José dos Campos: INPE, 2013. p. 2424-2430.

Bandyopadhyay, S.; Maulik, U. An evolutionary technique based on K-Means algorithm for optimal clustering in R. **Information Sciences**, 146 (2002) 221–237.

Bezdek, J. **Pattern Recognition with Fuzzy Objective Function Algorithms.** Norwell, MA, USA: Kluwer Academic Publishers, 1981.

Dunn, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. **Journal of Cybernetics**, Taylor & Francis, v. 3, n. 3, p. 32–57, 1973.

Eastman, R. **TerrSet Geospatial Monitoring and Modeling software.** Worcester: ClarkLabs, 2016.

Embrapa. **Sistema Brasileiro de Classificação de Solos.** 3ª ed. rev. ampl. Brasília: Embrapa, 2013.

Gass, S. L. B. Enchente na fronteira oeste do Rio Grande do Sul em dezembro de 2015: subsídios para uma agenda de pesquisa aplicada. **Confins** [online], 25, 2015.

Jensen, J. R. **Remote sensing of the environment: an earth resource perspective.** 2nd edition. Prentice Hall, 2007.

Maciel, A. M.; Vinhas, L.; Câmara, G. Algoritmos de clustering para separação de culturas agrícolas e tipos de uso e cobertura da Terra utilizando dados de sensoriamento remoto. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 17. (SBSR), 2015, João Pessoa. **Anais...** São José dos Campos: INPE, 2015. p. 4620-4627.

Macqueen, J. Some methods for classification and analysis of multivariate ob- servations. In: **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**. Berkeley, CA: University of California Press, 1967. p. 281-297.

Menezes, P. R. Almeida, T. de. **Introdução ao processamento de imagens de sensoriamento remoto.** Brasília: UNB, CNPq, 2012.

Nascimento, S.; Mirkin, B. A fuzzy clustering model of data and fuzzy c-means. In: **The ninth IEEE International Conference on Fuzzy Systems:** Soft Computing in the Information Age (FUZZ-IEEE 2000), IEEE Neural Networks Council. San Antonio, USA,: [s.n.], 2000. p. 302–307.

Ogassawara, J. F. **Análise morfométrica dos afluentes principais da bacia hidrográfica do rio Uruguai e sua influência nas enchentes na cidade de Itaqui ï RS.** Trabalho de conclusão de curso. Universidade Federal do Pampa. Campus Itaqui. Itaqui, 2015.

Rebouças, R. A.; Santos, R. D. C. dos; Hanermann, M.; Shiguemori, E. H. Uso do algoritmo divisão K-médias adaptado para definição de background em imagens do Landsat 8. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 17. (SBSR), 2015, João Pessoa. **Anais...** São José dos Campos: INPE, 2015. p. 5538-5545.

Vargas, R. R. de; Bedregal, B. R. C. A comparative study between fuzzy c-means and ckmeans algorithms. In **Proc. Conf. North American Fuzzy Information Processing Society** (NAFIPS 2010). Toronto, Canada, 2010.

Vargas, R. R. de; Bedregal, B. R. C.; Dimuro, G. Using ckMeans algorithm in image segmentation process: Preliminary results on mammography analysis. In **Anais do Congresso Nacional de Matemática Aplicada e Computacional** (XXXV CNMAC), Natal, Rio Grande do Norte, Brasil, 2014.

Vargas, R. R. de; Galafassi, C.; Amorim, N. C.; Freddo, R. Algoritmo ckMeans Aplicado ao Sensoriamento Remoto. In: XXXVI Congresso Nacional de Matemática Aplicada e Computacional, 2016, Gramado. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics** - CNMAC 2016. São Carlos, 2016. p. 1-6.

Vargas, R.; Bedregal, B. Uma Nova Forma de Calcular o Centro dos Clusters no Algoritmo Fuzzy C-Means. In: XXXIII Congresso Nacional de Matemática Aplicada e Computacional, 2010, Águas de Lindóia. **Anais do CNMAC.** São Paulo: SBMAC, 2010. v. 3. p. 486-492.

Vargas, R.; Bedregal, B.; Palmeira, E. A Comparison between K-Means, FCM and ckMeans Algorithms. In: Cavalheiro, S. A. da C.; Foss, L.; Aguiar, M. S. de; Dimuro, G. P.; Costa, A. C. da R. (Org.). **Post-Proceedings of the Workshop-School on Theoretical Computer Science**. Los Alamitos: IEEE, 2011, v. 1, p. 32-38.

Viero, A. C.; Silva, D. R. A. da (org.) **Geodiversidade do Estado do Rio Grande do Sul.** Porto Alegre: CPRM, 2010.