# Estudo comparativo de algoritmos de agrupamento para a definição de Zonas de Manejo

**Murilo Camargo Nogueira[1], Raul Teruel dos Santos[1]**

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
Av. Fernando Corrêa da Costa, nº 2367 – Bairro Boa Esperança – Cuiabá – MT – Brasil

murilokmargo@gmail.com, raul@ic.ufmt.br

***Abstract.*** *- In recent years food demand has increased due to rapid population growth. This scenario creates the need for a form of cultivation associated with technology as a way to minimize this situation. In this sense, precision farming techniques can be employed to map management zones to assist in food production by correctly applying input quantities. For the generation of such maps we use techniques of data clustering on top of data collected from the crop. The data clustering techniques: K-means, BIRCH and CURE were applied to agricultural data in order to generate management maps and compare those algorithms through the results obtained from evaluative methods of such maps.*

***Resumo.*** *- Nos últimos anos a demanda de alimentos aumentou devido ao crescimento rápido da população. Este cenário cria a necessidade de uma forma de cultivo associado a tecnologia como forma de minimizar esta situação. Nesse sentido, técnicas de agricultura de precisão podem ser empregadas para a definição de mapas zonas de manejo com o objetivo de auxiliar na produção de alimentos aplicando quantidades de insumos de maneira correta. Para a geração de tais mapas se faz uso de técnicas de agrupamento de dados recolhidos do talhão. As técnicas de agrupamento de dados K-means, BIRCH e CURE foram aplicados em dados agrícolas com o intuito de gerar mapas de manejo e comparar tais algoritmos através dos resultados obtidos de métodos avaliativos de tais mapas.*

## 1. Introdução

Com o crescimento rápido da população, a demanda de alimentos também cresceu rapidamente, tornando-se crucial o maior aproveitamento de recursos agrícolas de uma forma sustentável [Janrao et al. 2015]. A associação da tecnologia a agricultura, que tem como um dos resultados a agricultura de precisão, tem sido cada vez mais relevante devido à necessidade de aumento de produtividade e rentabilidade, menor uso de agrotóxicos e redução do impacto ambiental em várias áreas rurais [Amaral et al. 2015].

Visando atender esta demanda, técnicas de plantio tradicionais já não são mais suficientes para satisfazer as demandas do mercado, visto isso a agricultura de precisão se faz necessária.

Diferente da agricultura convencional, a agricultura de precisão permite o manejo do talhão a fim de dividi-la em áreas específicas, conhecidas como zona de manejo, almejando o maior desempenho da plantação [Speranza et al. 2016].

O processo de geração de mapas de zonas de manejo consiste em vários passos, sendo um deles o de agrupamento de dados, que é responsável pela geração de mapas com os dados fornecidos. Para esta etapa é necessário o uso de algum método de avaliação das partições geradas, para a validação dos mapas gerados.

Neste sentido, o presente trabalho propõe a comparação de técnicas de agrupamento.

## 2. Materiais e métodos

Para a realização deste estudo, foi utilizado um conjunto de dados obtidos em uma fazenda em Diamantino-MT. Originadas de um talhão de 12,5 há, onde foram coletados 117 pares de amostras. As amostras foram normalizadas e interpoladas através de um processo de krigagem ordinária. Todos os dados coletados foram usados para a geração de todos os mapas abordados nesse trabalho.

Essa pesquisa fez uso de três métodos de agrupamento para a sua realização, considerando o uso de 2, 3 e 4 grupos para a geração de mapas. Dentre os métodos, o algoritmo de agrupamento particional K-means [MacQueen 1967], e dois algoritmos de agrupamento hierárquicos BIRCH [Zhang et al. 1997] e CURE [Guha et al. 1998]. Para a validação do resultado destes métodos foram usados os métodos estatísticos de avaliação Índice Silhueta [Rousseeuw 1987], Índice Davies-Bouldin [Davies and Bouldin 1979] e também uma análise nas variâncias contidas nos dados dos agrupamentos.

O k-means é um método particional que que tem como objetivo a otimização de uma função que pode ser descrita pela equação:

$$E = \sum_{i=1}^{C} \sum_{X \in C_i} d\left(x, m_i\right)$$

Nesta equação, $m_1$ é o centro do grupo $C_i$, já $d(x, m_i)$ é a distância euclidiana entre x e $m_i$. Então, a função critério E procura minimizar a distância entre o ponto e o centro do grupo ao qual o ponto pertence. O algoritmo inicia na criação de uma quantidade C de grupos estabelecida pelo usuário, conhecido como centroides, com o objetivo de medir a distância entre o mesmo e os dados. Após um dado medir a distância entre ele e a todos os centroides, o mesmo é atribuído ao grupo que o centroide com menor distância representa. A cada iteração os centroides são atualizados, assim como quais dados são atribuídos a cada centroide, podendo assim corrigir uma atribuição errônea feita no começo de sua execução. O algoritmo finaliza quando não há mais alterações ou um número máximo de iterações é atingida.

O algoritmo BIRCH é um método de agrupamento hierárquico projetado para trabalhar com conjuntos de dados numéricos grandes. Trabalha com a ideia de *clustering features* (*CF*) que é um vetor de três dimensões que contêm informações sobre os objetos de um grupo, sendo elas n, LS e SS, onde n é o número de objetos em um grupo, LS é a soma dos n objetos e SS é a soma quadrada dos n objetos. CFs contem a informação necessária para a tomada de decisões do algoritmo BIRCH, e são obtidos através de uma leitura do conjunto de dados. Através das CFs é montado uma CF tree inicial, uma árvore que guarda os CFs para um agrupamento hierárquico, que pode ser entendida como uma compressão dos dados tentando preservar suas características.

CURE também é um método de agrupamento hierárquico, porém possui uma abordagem diferente de BIRCH. CURE utiliza de um número C de pontos distribuídos em cada grupo de forma que descrevam o formato do mesmo. Em seguida, esses pontos são aproximados do centro do grupo por meio de um fator de encolhimento. Então, os pontos C são definidos como representantes do grupo, então a distância entre dois grupos é interpretada como sendo a distância dos dois representantes mais próximos. Esse algoritmo leva em conta apenas os C pontos representantes e não todos os pontos que um grupo tem.

Para a avaliação dos algoritmos de agrupamento foram utilizados os índices de Silhueta, Davies-Bouldin e a análise da variância.

O Índice de Silhueta define a qualidade dos agrupamentos com base na proximidade entre os objetos de um determinado grupo e na proximidade desses objeto ao grupo mais próximo. O resultado desse índice varia entre -1 e 1. Quanto mais próximo a 1 melhor a alocação do objeto no grupo, em contrapartida, quanto mais próximo de -1 pior é a alocação do objeto. Uma média dos valores desse índice consegue avaliar a qualidade de agrupamento.

O Índice Davies-Bouldin analisa o agrupamento avaliando a soma da dispersão interna dos grupos e a distância entre os grupos, retornando um valor menor quanto melhor for o resultado do agrupamento, sendo 0 o menor valor possível.

A variância é uma medida dispersão que mostra o quão distante cada valor de um conjunto está do valor central. Ela é analisada com o intuito de encontrar grupos com valores grandes de variância, o que pode ser prejudicial para a qualidade do mapa, já que buscamos grupos homogêneos entre os seus dados, e discrepantes em relação a outros grupos.

Com a finalidade da aplicação dos métodos, foi utilizado a linguagem de programação Python versão 3.7. 64 bits em conjunto as bibliotecas pyclustering e sklearn.

## 3. Resultados e discussão

Na Figura 1 podemos observar os mapas gerados pelos algoritmos de agrupamento quando considerado a divisão em 2 grupos. O mapa gerado por BIRCH a esquerda, K-Means no meio e a direita Cure.

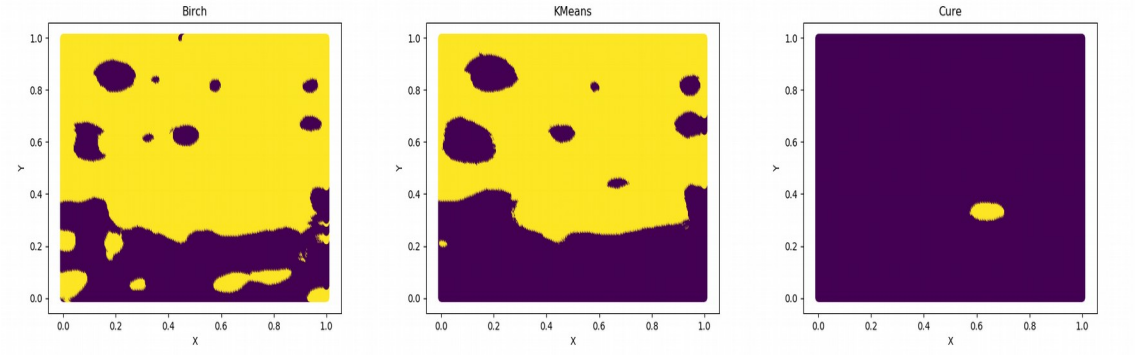


**Figura 1 - Mapas obtidos pelos algoritmos BIRCH, K-means e CURE respectivamente, considerando 2 grupos.**

Podemos observar na Figura 2, que quando considerado 3 grupos os algoritmos K-means e BIRCH sofrem uma mudança visual significativa se comparado a Figura 1, em contrapartida, CURE não apresenta mudanças substanciais.

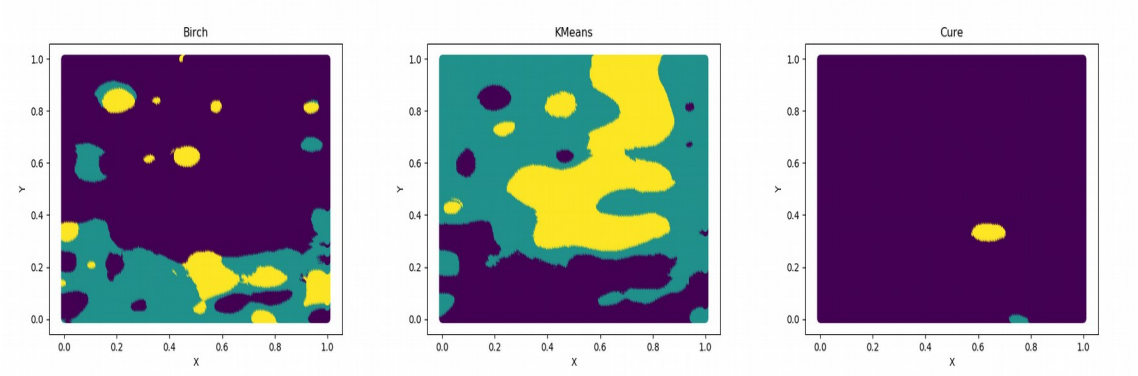


**Figura 2 - Mapas obtidos pelos algoritmos BIRCH, K-means e CURE respectivamente, quando considerado 3 grupos.**

Na Figura 3 são apresentados os mapas de zonas de manejo considerando 4 grupos, pode-se observar mudanças significativas nos mapas em comparação com a Figura 2 para os métodos K-meas e BIRCH, enquanto para CURE não houve mudanças significativas.

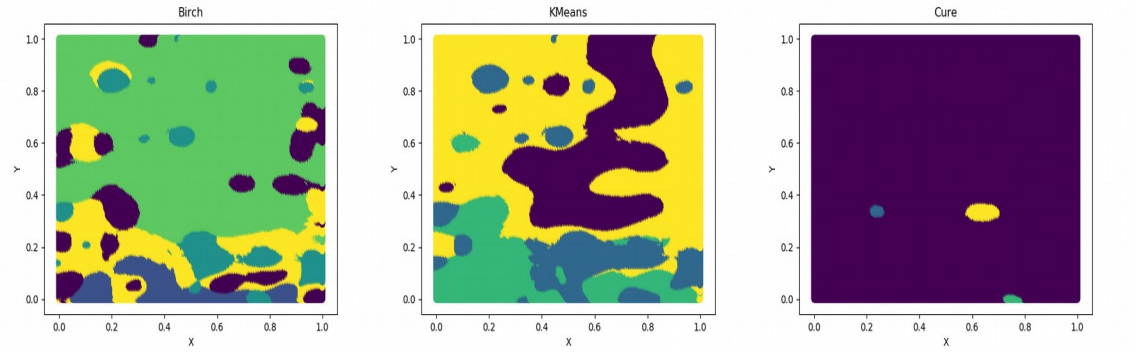


**Figura 3 - Mapas obtidos pelos algoritmos BIRCH, K-means e CURE respectivamente, quando considerado 4 grupos.**

Após a geração dos mapas, foram aplicados os índices para avaliar os mapas obtidos. O Índice Silhueta (SI) com o objetivo de auxiliar na escolha de qual quantidade de grupo era a mais apropriada para esse conjunto de dados.

**Tabela 1. Valores retornados pela aplicação dos índices**

| | K-means | BIRCH | CURE |
|---|---|---|---|
| Nº de grupos | SI | SI | SI |
| 2 | 0,405819 | 0,387496 | 0,466678 |
| 3 | 0,328491 | 0,348058 | 0,377509 |
| 4 | 0,356113 | 0,337854 | 0,280611 |

Pode-se observar na Tabela 1 que os melhores valores encontrados para o Índice de Silhueta é quando considera-se apenas 2 grupos. Pode-se observar também que os valores do índice reduziram de forma significativa quando considerado o aumento de

grupos. Por conta deste resultado foi considerado que a melhor divisão desse conjunto de dados é de 2 grupos.

O Índice Davies-Bouldin (DB) foi aplicado nos mapas gerados com 2 grupos pelos métodos K-means, BIRCH e CURE. O mapa gerado pelo método CURE obteve o melhor valor com o índice de 0,445676, seguido por K-means com 0,979984 e BIRCH com o valor de 0,996503.

Após a aplicação desses índices, foi realizada a análise variância nos grupos gerados pelos métodos, considerando os valores de Magnésio (Mn) e Fósforo (P), a fim de avaliar a variância nos grupos.

**Tabela 2. Valores de variância dos grupos gerados quando considerado 2 grupos**

| | K-means | | BIRCH | | CURE | |
|---|---|---|---|---|---|---|
| Grupo | Mn | P | Mn | P | Mn | P |
| 1 | 0,007332 | 0,009605 | 0,012842 | 0,011554 | 0,012077 | 0,024968 |
| 2 | 0,018396 | 0,010507 | 0,008448 | 0,014651 | 0,000786 | 0,001507 |

Considerando os valores da variância entre os grupos apresentados pela Tabela 2 é possível observar que a variância entre as amostras do grupo 1 e grupo 2 gerados pelo CURE são bem distintas, resultando no grupo 1 muito heterogêneo e o grupo 2 muito homogêneo. Isto demonstra que o agrupamento gerado por CURE não foi satisfatório nesse conjunto de dados. Em contrapartida, K-means e BIRCH apresentaram variâncias sem grandes diferenças, demonstrando agrupamentos satisfatórios.

Observa-se que o método DB utiliza a média das variâncias durante o cálculo do índice. Devido a este fato, temos uma atenuação dos valores reais da variância dos dois grupos gerados por CURE, levando a um resultado aparentemente melhor. Contudo, através da análise dos dados de variâncias individuais demonstradas pela Tabela 2, percebe-se que os dados do grupo 1 possuem uma variância alta, o que significa que, apesar de agrupados, os dados não possuem grande semelhança entre si.

O tempo de processamento fornecido pelos algoritmos para cada mapa usando os algoritmos K-means e BIRCH levou apenas um minuto, CURE leva em torno de 18 a 20 minutos para a mesma tarefa.

## 4. Conclusões

O algoritmo CURE apesar de ter obtido os melhores valores para os índices, se mostrou insatisfatória no conjunto de dados quando analisado a variância dos grupos gerados demonstrados na Tabela 2, apresentando uma variância muito alta em um de seus agrupamentos. Além disso, apresentou o tempo de processamento mais longo dentre os 3 algoritmos, levando cerca de 20 minutos para a geração dos mapas.

K-means obteve uma variância equilibrada e os melhores resultados em seus índices avaliativos após o CURE quando considerado 2 grupos. Em questão de tempo de processamento, K-means levou cerca de 1 minuto para a geração de mapas.

BIRCH se mostrou com uma variância equilibrada e valores ligeiramente inferiores ao k-means como resultado de seus índices. BIRCH também obteve rapidez no tempo de processamento.

Apesar dos valores dos índices apontarem o mapa gerado por CURE como o de melhor resultado, a análise de variância mostra que os grupos gerados apresentam muita discrepância entre os mesmos, apontando que não houve uma divisão significativa. Por isso, o mapa gerado por K-means, que obteve os melhores valores de índice após CURE e apresentou uma variância equilibrada, se mostra o mais adequado para a definição de zonas de manejo neste conjunto de dados dentre os três métodos utilizados.

O mapa gerado por Ramos et al. (2017) para o mesmo conjunto de dados é semelhante ao mapa gerado pelo K-means com 2 grupos, o que corrobora com os resultados apresentados neste artigo.

### 5. Agradecimentos

O presente trabalho foi realizado com o apoio da Fundação de Amparo à Pesquisa do Estado de Mato Grosso (FAPEMAT) e do Prof. Dr. Fabricio Tomaz Ramos fornecendo os dados usados para a realização deste estudo.

## 6. Referências

Amaral, L. R., Molin, J. P., Portz, G., Finazzi, F. B., & Cortinove, L. (2015). Comparison of crop canopy reflectance sensors used to identify sugarcane biomass and nitrogen status. Precision Agriculture, 16(1), 15-28.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, (2), 224-227.

Guha, S., Rastogi, R., & Shim, K. (1998, June). CURE: an efficient clustering algorithm for large databases. In ACM Sigmod Record (Vol. 27, No. 2, pp. 73-84). ACM.

Janrao, P., & Palivela, H. (2015, March). Management zone delineation in Precision agriculture using data mining: A review. In 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (pp. 1-7). IEEE.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Ramos, F. T., Santos, R. T., Campelo Junior, J. H., & MAIA, J. C. D. S. (2017). Defining management zones based on soil attributes and soybean productivity. Revista Caatinga, 30(2), 427-436.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Speranza, E. A., Ciferri, R. R., & CIFERRI, C. D. A. (2016). Clustering approaches and ensembles applied in the delineation of management classes in precision agriculture. In Embrapa Informática Agropecuária-Artigo em anais de congresso (ALICE). In: BRAZILIAN SYMPOSIUM ON GEOINFORMATICS, 17., 2016, Campos do Jordão. Proceedings... São José dos Campos: INPE, 2016..

Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. Data Mining and Knowledge Discovery, 1(2), 141-182.