# DAGER: UMA FERRAMENTA PARA MINERAÇÃO DE DADOS GEORREFERENCIADOS

Ronan Assumpção Silva (UEPG) E-mail: ronan.ras@gmail.com
Prof Dr Petraq J. Papajorgji (UEPG) E-mail: petraq@gmail.com
Profa Dra Alaine M. Guimarães (UEPG) E-mail: alainemg@hotmail.com
Prof Dr José Carlos Ferreira da Rocha (UEPG) E-mail: jrocha@uepg.br

**Resumo**: O artigo apresenta a ferramenta DAGER, criada para realizar mineração de dados em contextos agrícolas. A ferramenta traz implementada os algoritmos PAM, CLARA, CLARANS, GCLARA e GCLARANS. Parte destes algoritmos trata somente de atributos de latitude e longitude e a outra parte trata de atributos diversos com latitude e longitude. Também faz parte da ferramenta a visualização dos agrupamentos gerados por meio da técnica de pixel.

**Palavras-chave**: Mineração de dados, Agricultura de Precisão, Dados Georreferenciados, DAGER.

# DAGER: A TOOL FOR GEOREFERENCED DATA MINING

**Abstract**: This work presents DAGER, a tool created to perform data mining on agricultural contexts. The tool brings implemented the PAM, CLARA, CLARANS, GCLARA and GCLARANS algorithms. Some of these algorithms involve only attributes related to latitude and longitude of each observed point. The others algorithms deal with several attributes about and also the latitude and longitude. Another feature of the developed tool is presenting the visualization of clusters generated by the use of the pixel technique.

**Keywords**: Data Mining, Precision Agriculture, Georeferenced Data, DAGER.

## 1. INTRODUÇÃO

A descoberta de conhecimento a partir do processamento de bases de dados com atributos georreferenciados tem grande importância para a pesquisa e desenvolvimento de atividades na agricultura (SANTOS, 2001). Existem várias ferramentas para realizar tal tarefa como GeoMiner (Koperski & Han 1997), PADRAO, (Santos, 2001) e Weka GDPM, uma extensão do Weka proposta por Bogorny et al (2007). Porém, dentre as opções de ferramentas disponíveis, nem todas consideram o tratamento de atributos georreferenciados tão importantes quanto os demais atributos. Também nas ferramentas estudadas não foi visto estudos voltados para a solução de problemas agrícolas.

O estudo de dados de domínio agrícola tem importância tanto para a economia quanto para os pesquisadores. Considerando isto, este trabalho apresenta uma ferramenta de software denorminada DAGER, que permite a execução de algoritmos que exploram de forma combinada atributos georreferenciados e não georreferenciados. Atualmente apresenta os algoritmos de agrupamento: PAM, CLARA, CLARANS, GCLARAN e GCLARANS. Além disso, a ferramenta DAGER possibilita a visualização de resultados em formato de texto e gráfico. O gráfico tem o objetivo de possibilitar a visualização da distribuição espacial dos agrupamentos encontrados.

Este trabalho está organizado da seguinte maneira. A Seção 2 apresenta a motivação do trabalho, assim como as pesquisas relacionadas a mineração de dados georreferenciados. A seção 3 traz uma breve explanação sobre mineração de dados e algoritmos de agrupamento, assim como os algoritmos implementados na ferramenta. A visualização é abordada na seção 4 trazendo exemplo de visualização pela ferramenta DAGER. É descrito brevemente como é o funcionamento da ferramenta DAGER na seção 5. Em seguida, a conclusão.

## 2. MOTIVAÇÃO

A agricultura de precisão (AP) objetiva o desenvolvimento de métodos e tecnologias para solucionar problemas na agricultura ao considerar que na lavoura existem diferenças de produtividade. Uma situação que serve como exemplo é a de Oliveira (2007), que verificou alta variabilidade espacial na produtividade, tamanho de frutos e concentração de nutrientes no solo e na planta num talhão. Para Motomiya (2011), a agricultura de precisão é uma estratégia de manejo do solo e de culturas que busca fazer o melhor uso de insumos e tem importância para a preservação ambiental. Assim, por meio da agricultura de precisão os produtores devem ser capazes de identificar a variabilidade dentro de um campo, tratando-a para aumentar a produtividade e os lucros.

Ferramentas de software para Mineração de dados agrícolas tem se mostrado úteis para a AP (SOUZA et al, 2010). A Mineração de Dados é uma das etapas da Descoberta de Conhecimento em Base de Dados (DCBD). Com a Mineração de Dados busca-se obter regras ou padrões por meio do estudo de uma base de dados (SOUZA et al., 2010). Existem várias ferramentas que realizam a Mineração de Dados como Weka (Hall et al., 2012), GeoMiner (KOPERSKI & HAN, 1997), PADRAO (SANTOS, 2001).

As ferramentas de Mineração de Dados que manipulam dados georreferenciados comumente tem propósitos gerais. Como a Agricultura de Precisão necessita do georreferenciamento dos dados, uma ferramenta de Mineração de Dados que busque resolver problemas agrícolas é necessária.

Muitos trabalhos tem sido desenvolvidos por meio do uso de ferramentas de mineração de dados. Entre os algoritmos que os pesquisadores do domínio agrícola tem comumente usado está o de árvore de decisão (MARTINS & FONSECA, 2009), que gera regras do tipo SE ENTÃO, cuja interpretação é facilitada pela forma de apresentação.

A ferramenta computacional DAGER inicialmente propõe algoritmos de agrupamento considerando atributos georreferenciados. Em alguns desses algoritmos são considerados atributos georreferenciados e não georreferenciados. Algoritmos de agrupamento visam formar agrupamentos de elementos de características semelhantes. Algoritmos de agrupamento que considerem somente dados georreferenciados podem separar grupos pela sua distância geográfica, por exemplo. Já os que consideram somente os atributos não georreferenciados, podem agrupar os dados de uma lavoura pelo tipo da cultura.

Uma forma de apresentação dos agrupamentos é por meio de gráficos. O gráfico é uma das formas de representação de dados que visa o fácil entendimento e interpretação, para que então produza informação por quem o visualiza. Por meio da técnica de mapa de pixel, DAGER possibilita visualizar os agrupamentos gerados após o processo de mineração dos dados. Além disso, a ferramenta permite que o usuário saiba todas as informações da instância que compõe determinado ponto no gráfico.

## 3. MINERAÇÃO DE DADOS E AGRUPAMENTO

A ferramenta DAGER parte da necessidade de soluções na Agricultura, como tipos de visualização diferenciados. Porém, para chegar nos tipos de visualização, foi considerado que sendo ferramenta de Mineração de Dados, deveria executar um mínimo de algoritmos para que então pudesse ser tratada a parte de visualização. Desta forma, começando por algoritmos de agrupamento, foram considerados algoritmos que trabalhassem bases de dados com atributos georreferenciados, tanto de forma isolada como considerando também os atributos não georreferenciados. Além disso, para o desenvolvimento desta ferramenta foi considerada

uma base de dados agrícola obtida por processos de Agricultura de Precisão. Portanto, o desenvolvimento da ferramenta surgiu pela observação da demanda de soluções na Agricultura, especificamente da Agricultura de Precisão.

O agrupamento consiste em identificar conjuntos de objetos semelhantes entre si e diferentes de outros agrupamentos (Kaufman & Rousseeuw, 1990). Uma das vantagens da técnica é que identifica estruturas diretamente dos dados sem o conhecimento prévio destes dados. O agrupamento dos dados pode ser baseado em funções de distância. Os algoritmos de agrupamento podem ser classificados nas seguintes categorias: hierárquicos, de particionamento ou realocação interativa, de densidade e de restrições de contiguidade.

Considerando que para a realização da tarefa de agrupamento devem ser considerados atributos georreferenciados e não georreferenciados foi feito levantamento bibliográfico sobre possíveis algoritmos a serem implementados. Dentre as opções foram considerados os algoritmos GDBScan (SANDER et al, 1998) e CLARANS (Ng & Han, 1994). O primeiro, uma modificicação do DBScan, é baseado em densidade. Traz vantagem de propor organização diferenciada para bases de dados georreferenciadas quanto ao gerenciador de banco de dados. Já o segundo, baseado em particionamento, vem de uma evolução de vários algoritmos findando em uma estratégia em que os atributos georreferenciados são dominantes ou não são dominantes. Neste sentido, o CLARANS tem duas variantes: o atributo georreferenciado como dominante (CLARANS SD) e o atributo georreferenciado como não dominante (CLARANS NSD).

Para o desenvolvimento da ferramenta, foi proposta modificação de dois algoritmos de agrupamento: CLARA e CLARANS. GCLARA (Generalized CLARA) é baseado em partição e tem como objetivo construir agrupamentos sem explorara a base completamente. Usa estratégia de amostra para chegar ou se aproximar de agrupamentos ideais. GCLARANS (Generalized CLARANS) tem o mesmo objetivo de encontrar os agrupamentos, mas usa estratégia de grafo para não precisar percorrer a base por completo.

## 4. VISUALIZAÇÃO

A visualização de dados na MDG é de muita importância, pois auxilia na interpretação dos mesmos. Rabelo (2007) afirma que o mau emprego de técnicas de visualização pode comprometer o processo de DCBD.

Miller (2007) classifica a visualização de informação em:

1. Baseada em mapa: Permite o usuário interativamente representar dados georreferenciados em uma forma de gráfico;
2. Baseada em gráfico: representa os dados em gráfico como gráficos de dispersão e gráficos de pizza;
3. Projeção: usa transformações estatísticas para representar dados em espaços alternativos não Euclidianos;
4. Pixel: mapeia os valores dos dados em pixels individuais que estão em alguma ordem de significância ou posição na tela, como por exemplo, a ordem temporal;
5. Iconográfico: utiliza símbolos para representar o sentido do todo ou descartar pequenas diferenças entre os dados;
6. Técnicas de rede: organiza representação visual baseada em específicas estruturas lógicas como árvores.

A técnica de pixel é utilizada para representar um panorama de bases de dados com muitos elementos. Como a presente proposta visa representar o conhecimento de uma lavoura como um todo, esta técnica se mostra eficiente para solucionar o problema da representação de uma base de dados georreferenciados.

Para a visualização dois tipos são apresentados: arquivo de texto e mapa de pixel.

O arquivo de texto deve ser organizado a fim de apresentar o resultado do agrupamento. Além disso, foi planejado para ser lido e apresentado também na visualização do tipo mapa de pixel. O mapa de pixel tem por finalidade apresentar na forma de gráfico os grupos resultantes do processo de agrupamento.

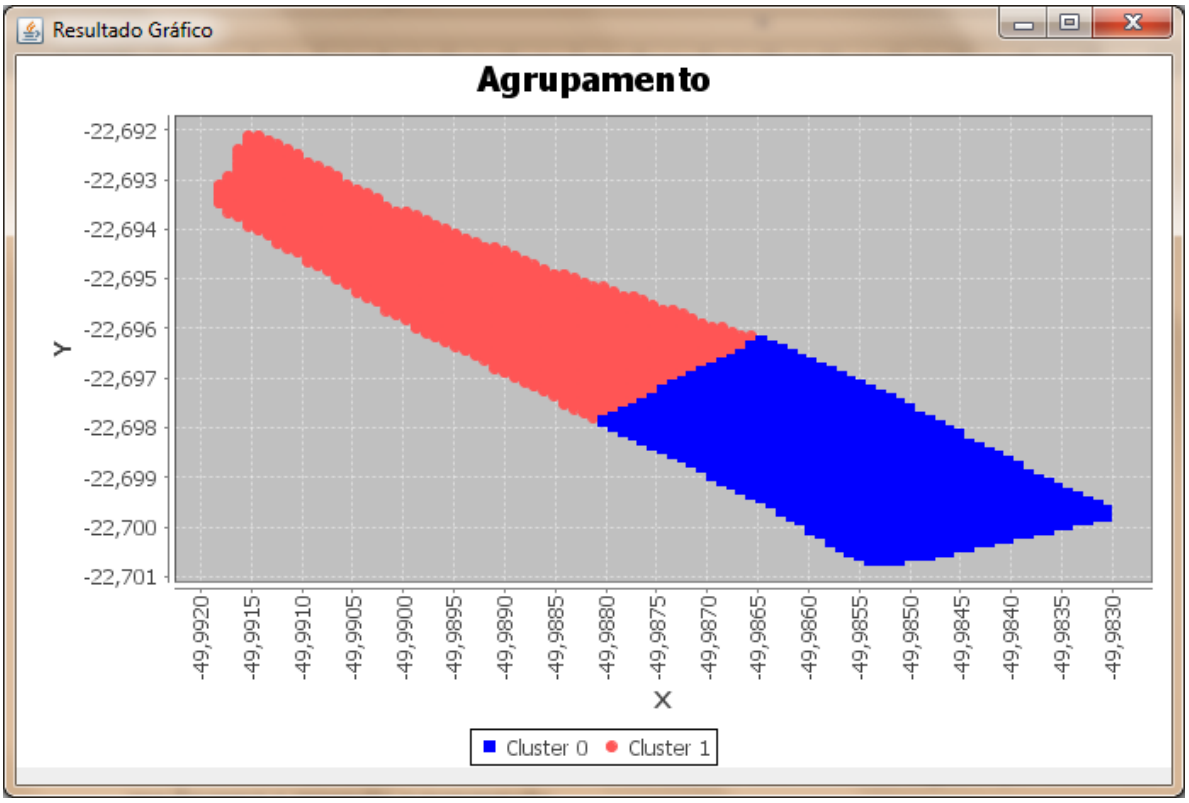


Figura 1 - Dois agrupamentos encontrados pelo algoritmo GCLARA e sua representação visual na ferramenta DAGER.

Na Figura 1 é visualizado o talhão representado pela base de dados. As cores azul e vermelha mostram o resultado de dois agrupamentos obtidos por meio do algoritmo GCLARA. Esta base de dados agrícola possui atributos georreferenciados e não georreferenciados. Também foi utilizado como parâmetro encontrar dois agrupamentos, utilizando atributos georreferenciados e a produção de Soja.

## 5. A FERRAMENTA DAGER

A ferramenta é dividida em vários módulos, divididos da seguinte maneira:

1. Carregamento da base de dados: O usuário irá apontar onde está a base de dados que o sistema irá trabalhar;
2. Agrupamento: Com base na distância entre uma amostra e outra, o sistema pode

agrupar características comuns dos atributos levando em consideração uma distância mínima e máxima;

3. Visualização: as regras geradas serão visualizadas de forma georreferenciada.

Para minerar dados georreferenciados a base de dados deve conter informações geográficas relacionadas aos atributos como latitude e longitude. Além disso, a ferramenta segue o padrão de arquivo ARFF para leitura da base. Este mesmo padrão é utilizado na ferramenta mencionada Weka.

## 5. CONCLUSÕES

A ferramenta DAGER busca encontrar agrupamentos pela similaridade de dados georreferenciados e/ou dados não georreferenciados. Permite que os algoritmos apresentem seus resultados na forma de gráfico pela técnica de pixel.

Trabalhos futuros envolvem a inserção de novos algoritmos na ferramenta, não só de agrupamento, mas também de classificação ou quaisquer outros necessários para a realização de mineração de dados, principalmente atuando em contextos de domínio agrícola.

## REFERÊNCIAS

**BOGORNY, V.; PALMA, A.; KUIJPERS, B.; ALVARES, L. O.** Spatial Data Mining: From Theory to Practice with Free Software. In: Proc. of WSL International Workshop on Free Software (WSL'07). Porto Alegre, 2007.

**HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B. REUTEMANN, P. WITTEN, I. H.** The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1. 2009.

**KOPERSKI, K., HAN, J.** GEOMINER: A System Prototype for Spatial Mining. In: Proceedings ACM-SIGMOD. Arizona, 1997.

**MARTINS, V. A.; FONSECA, L. M. G.** Classificação de uso de solo baseada na análise orientada a objeto e mineração de dados utilizando imagens SPOT/HRG-5. In: Anais XIV Simpósio Brasileiro de Sensoriamento Remoto. Natal, 2009.

**MILLER, H. J.** Geographic Data Mining and Knowledge Discovery. In: Wilson, J., Fotheringham, A.S. (eds.) The Handbook of Geographic Information Science, Wiley-Blackwell. 2007.

**MOTOMIYA, A. V. DE A.; MOTOMIYA, W. R.; MOLIN, J. P. LIRA, A.; OLIVEIRA, DI OLIVEIRA, J. R. G.; BISCARO, G. A.** Variabilidade espacial de atributos químicos do solo e produtividade do algodoeiro. In: Agrarian. 2011. Disponível em: http://www.periodicos.ufgd.edu.br/index.php/agrarian/article/view/1118/670. Consulta em 07 de Setembro de 2011.

**NG, R. T.; HAN, J.** Efficient and effective clustering methods for spatial data mining. In: Proceedings of the 20th VLDB Conference. Santiago - Chile,1994.

**OLIVEIRA, P. C. G.** Variabilidade espacial de macronutrientes correlacionados com a produtividade em pomares cítricos do município de Capitão Poço - PA. Dissertação - Universidade Federal Rural da Amazônia. 2007.

**RABELO, E.** Avaliação de técnicas de visualização para mineração de dados. Dissertação - Universidade Estadual de Maringá. Programa de Pós-graduação em Ciência da Computação, 2007.

**SANTOS, M. Y.** Padrão: um sistema de descoberta de conhecimento em bases de dados geo-referenciadas. Tese – Universidade do Minho. 2001.

**SANDER J., ESTER M., KRIEGEL H.P.; XU, X.** Density-Based Clustering in Spatial Databases: A New Algorithm and its Applications. In: Data Mining and Knowledge Discovery, an International Journal, Kluwer Academic Publishers, Vol.2, No. 2. 1998.

**SOUZA, Z. M.; CERRI, D. G. P.; COLET, M. J.; RODRIGUES, L. H. A.; MAGALHÃES, P. S. G.; MANDONI, R. J. A.** Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. Cienc. Rural [online]. 2010, vol.40, n.4. 2010.