UNIVERSIDADE ESTADUAL DE PONTA GROSSA
SETOR DE CIÊNCIAS AGRÁRIAS E DE TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA

RONAN ASSUMPÇÃO SILVA

DAGER: UMA FERRAMENTA COMPUTACIONAL PARA AGRUPAMENTOS EM MINERAÇÃO DE DADOS AGRÍCOLAS GEORREFERENCIADOS

PONTA GROSSA
2012

RONAN ASSUMPÇÃO SILVA

DAGER: UMA FERRAMENTA COMPUTACIONAL PARA AGRUPAMENTOS EM MINERAÇÃO DE DADOS AGRÍCOLAS GEORREFERENCIADOS

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada, curso de Mestrado em Computação Aplicada da Universidade Estadual de Ponta Grossa, com requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Orientação: Prof Dr Petraq J. Papajorgji
Co-orientação: Profa Dra Alaine M. Guimarães

PONTA GROSSA
2012

Dados Internacionais de Catalogação na Publicação

S586 Silva, Ronan Assumpção

Dager: uma ferramenta computacional para agrupamentos em mineração de dados agrícolas georreferenciados. / Ronan Assumpção Silva. -- Ponta Grossa, 2012.
51 f. : il. ; 30 cm.

Orientador: Prof. Dr. Petraq J. Papajorgji
Co-orientadora: Profa. Dra. Alaine M. Guimarães

Dissertação (Mestrado em Computação Aplicada) - Programa de Pós-Graduação em Computação Aplicada. Universidade Estadual de Ponta Grossa. Ponta Grossa, 2012.

1. Agrupamento. 2. Dados Georreferenciados. 3. Agricultura de precisão. I. Papajorgji, Petraq J. II. Guimarães, Alaine M. III. Universidade Tecnológica Federal do Paraná, Campus Ponta Grossa. IV. Título.

CDD 004

TERMO DE APROVAÇÃO

RONAN ASSUMPÇÃO SILVA

"DAGER: UMA FERRAMENTA COMPUTACIONAL PARA AGRUPAMENTOS EM MINERAÇÃO DE DADOS AGRÍCOLAS GEORREFERNCIADOS."

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Orientador: Petraq J. Papajorgji
University of Tirana

José Carlos Ferreira da Rocha
UEPG

Gleifer Vaz Alves
UTFPR

Ponta Grossa, 17 de agosto de 2012.

Dedicado a família e especialmente minha esposa Louisi.

AGRADECIMENTOS

A Deus.

Ao Prof. Dr. Petraq J. Papajorgji, pela contribuição com seus conhecimentos e sugestões na orientação desta dissertação.

A Profa. Dra. Alaine M. Guimarães, pela contribuição com suas idéias e apoio na co-orientação desta pesquisa.

Ao Prof. Dr. José Carlos Ferreira da Rocha, pela colaboração com informações que auxiliaram na concretização deste estudo.

Ao Prof. Dr. Gleifer Vaz Alves por aceitar o convite em compor a banca e suas considerações para melhoria da dissertação.

Ao Prof. Dr. José Paulo Molin pela base de dados agrícola utilizada nos experimentos.

A todos os professores do curso de mestrado da UEPG.

Aos colegas do curso de mestrado.

Ao Prof. Dr. João Umberto Furquim de Souza (*in memorian*) pelo incentivo.

A Profa Dra Simone Nasser Matos e ao Prof Dr Lourival Aparecido De Gois pelo apoio.

Aos colegas de trabalho de todo o período em que estive no mestrado.

A minha família, em especial Louisi, minha querida esposa.

Quanto ao mais, irmãos, tudo o que é verdadeiro, tudo o que é honesto, tudo o que é justo, tudo o que é puro, tudo o que é amável, tudo o que é de boa fama, se há alguma virtude, e se há algum louvor, nisso pensai.

(Bíblia - Filipenses 4:8)

RESUMO

A agricultura demanda de soluções computacionais diversas, especialmente quando refere-se ao segmento da Agricultura de Precisão (AP). A Mineração de Dados é um dos recursos computacionais que pode beneficiar a análise de dados de AP, os quais normalmente são georreferenciados. Porém, há limitação de algoritmos e ferramentas computacionais quando se tem necessidade de agrupar características diversas considerando também sua posição geográfica. Nesse contexto, o objetivo desta dissertação foi desenvolver e implementar algoritmos de agrupamentos e visualização que considerem atributos georreferenciados juntamente com atributos diversos em base de dados agrícolas. Para tanto, foi criada uma nova ferramenta computacional para mineração de dados agrícolas georreferenciados, denominada DAGER. Os algoritmos PAM, CLARA e CLARANS foram implementados e, com base nesses, dois novos algoritmos, GCLARA e GCLARANS foram desenvolvidos e implementados na ferramenta. Além dos algoritmos, foi implementado um módulo de visualização gráfica dos agrupamentos. Para a realização dos experimentos, foi utilizada uma base de dados agrícola obtida por processos de Agricultura de Precisão e para a avaliação dos grupos foram empregados os métodos estatísticos ANOVA e o MANOVA. O resultado apresenta o mapeamento e visualização de regiões dentro de uma lavoura com características semelhantes, atendendo aos objetivos propostos.

Palavras-chave: Agrupamento, Dados Georreferenciados, Agricultura de Precisão.

ABSTRACT

Agriculture demands for various computing solutions, especially when refers to the Precision Agriculture (PA). Data Mining is one of the computing resources that can benefit the analysis of data from AP, which usually are georeferenced. However, there are limitations of algorithms and computational tools when it is need to group different characteristics also considering its geographical position. In this context, the aim of this work was to develop and implement algorithms that consider clustering and visualization of georeferenced attributes along with various attributes in the agricutural database. It was created a new computational tool for georeferenced agricultural data mining called Dager. The algorithms PAM, CLARA and CLARANS were implemented and, based on these two new algorithms, and GCLARA GCLARANS were developed and implemented in the tool. Besides the algorithms it was implemented a module for graphical visualization of clusters. For the experiments, a database obtained by Precision Farming and evaluation groups were employed statistical methods ANOVA and MANOVA. The result showed the mapping and visualization of regions within a field with similar characteristics, achieving the proposoal objectives.

Keywords: Clustering, Georeferenced Data, Precision Agriculture.

LISTA DE EQUAÇÕES

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

AP – Agricultura de Precisão

CLARA – *Clustering Large Applications*

CLARANS – *Clustering Large Applications based on Randomized Search*

DCBD – Descoberta de Conhecimento em Base de Dados

DCBDG – Descoberta de Conhecimento em Base de Dados Georreferenciados

EMPRAPA – Empresa Brasileira de Pesquisa Agropecuária

FORTRAN - *IBM Mathematical FORmula TRANslation System*

GCLARA – *Generalized* CLARA

GCLARANS – *Generalized* CLARANS

GPS – *Global Position System* (Sistema de Posicionamento Global)

KDD – *Knowledge Discovery in Databases*

MD – Mineração de Dados

MDG – Mineração de Dados Georreferenciados

PAM – *Partitioning Around Medoids*

SIG – Sistema de Informação Geográfica

UEPG – Universidade Estadual de Ponta Grossa

SUMÁRIO

1 INTRODUÇÃO

A agricultura demanda ferramentas computacionais que irão auxiliar nas tomadas de decisões. A interpretação dos dados, o entendimento do que acontece na plantação, onde acontece na plantação, a previsão da safra com base na gestão adotada são alguns dos exemplos de necessidades agrícolas em que ferramentas computacionais podem ser úteis.

A Agricultura de Precisão (AP) surge apontando técnicas e tecnologias para solucionar problemas na agricultura ao considerar que na lavoura existem diferenças de produtividade. Para Motomiya (2011), a Agricultura de Precisão é uma estratégia de manejo do solo e das culturas que busca fazer o melhor uso de insumos e tem importância para a preservação ambiental.

Uma das grandes preocupações da AP é o entendimento das variáveis relacionadas ao plantio. A partir da compreensão espacial do que acontece na lavoura, é possível avaliar correções a serem aplicadas no manejo do solo, fertilização das plantas, controle de pragas e até a previsão de colheita. Considerando que existem ferramentas para registrar os diversos atributos relacionados a agricultura, e conceitos inerentes a relação entre esses atributos, o estudo e a interpretação desses dados podem requerer ferramentas computacionais diferenciadas. Nesse panorama, uma solução em potencial é a Mineração de Dados (MD).

A Mineração de Dados atua na identificação de padrões ou regras que uma base de dados pode conter. A MD é um dos passos do Knowledge Discovery in Databases (KDD), que é a Descoberta de Conhecimento em Bases de Dados (DCBD). Deste modo, as regras encontradas deverão ser avaliadas por especialista para que dentre os padrões encontrados haja novo conhecimento.

Dentre as ramificações da MD está a Mineração de Dados Georreferenciados (MDG), que consiste em buscar padrões em bases de dados georreferenciados. Em função de suas características, a MDG pode ser adequada para o tratamento de dados obtidos por processos de agricultura de precisão. A MDG é complexa e apresenta desafios a serem vencidos para que possa ser efetivamente utilizada. Ferramentas e algoritmos foram propostos com o objetivo de superar esses desafios, como é o caso do GeoMiner (KOPERSKI; HAN, 1997), PADRAO, (SANTOS, 2001) e do Weka GDPM, que é uma extensão do Weka proposta por Bogorny et al. (2007). Porém, além da escassez de ferramentas, algumas complicações inviabilizam o uso das ferramentas, como a disponibilidade das ferramentas para uso ou a atualização da ferramenta para execução em ambiente operacional atual.

Face a estas dificuldades, esta dissertação visa implementar uma ferramenta própria, chamada DAGER, para realizar agrupamentos em Mineração de Dados Georreferenciados. Os algoritmos implementados são o PAM, CLARA, CLARANS, GCLARA e GCLARANS. Destes algoritmos, PAM, CLARA e CLARANS foram implementados seguindo o pseudo-código descritos por NG; HAN (2002). É proposto o algoritmo GCLARA como uma extensão de CLARA, para consideração de atributos de latitude e longitude juntamente com outros atributos para encontrar regiões de características semelhantes. Um outro algoritmo proposto é o GCLARANS, que é uma extensão baseada no algoritmo CLARANS para considerar atributos de latitude e longitude juntamente com outros atributos na busca de regiões semelhantes, com o diferencial de ser baseado em grafo. Também foi implementado em DAGER um módulo de visualização gráfica para sinalizar os grupos encontrados no talhão pelos algoritmos implementados.

Quanto a ordem, a presente dissertação foi organizada em capítulos estruturados sendo iniciada com introdução, objetivos, problema de pesquisa, hipótese e justificativa.

O capítulo 2 traz a revisão bibliográfica sobre os temas, para abordagem dos conceitos encontrados nesta pesquisa. É dado enfoque ao tratamento de Dados Georreferenciados, principalmente no agrupamento que é baseado na técnica de Mineração de Dados.

A Metodologia aplicada no decorrer da pesquisa é apresentada no capítulo 3. Neste capítulo são explanados os algoritmos criados e a ferramenta em que esses algoritmos tiveram funcionamento. Também é feita a apresentação da base utilizada neste estudo.

Todos os experimentos, resultados e discussão estão no capítulo 4. Além disso, todo o suporte estatístico que comprova os resultados e a validação dos métodos.

A conclusão da pesquisa se encontra no capítulo 5, seguida das referências bibliográficas que contribuíram para o progresso da pesquisa.

1.1 Objetivo Geral

Desenvolver uma nova ferramenta computacioanal para realizar agrupamento em bases de dados agrícolas georreferenciados e desenvolver um módulo para visualização dos grupos encontrados. Implementar os algoritmos PAM, CLARA e CLARANS. Realizar uma extensão de CLARA (GCLARA) e outra de CLARANS (GCLARANS) para manipular os atributos de latitude e longitude em conjunto com os outros atributos observados.

1.2 Objetivos Específicos

a) Desenvolver ferramenta computacional (DAGER) para execução de algoritmos de agrupamento em Mineração de Dados.
b) Acoplar à ferramenta DAGER um módulo de visualização do resultado dos agrupamentos.
c) Implementar algoritmos espaciais (georreferenciados): PAM, CLARA e CLARANS.
d) Propor e desenvolver os algoritmos GCLARA e GCLARANS considerando os atributos latitude e longitude em conjunto com os demais atributos observados.

1.3 Problema de Pesquisa

É possível agrupar objetos de forma significativa considerando no processo a posição geográfica das características analisadas?

1.4 Hipótese

É possível encontrar regiões semelhantes dentro de uma lavoura, levando em consideração o georreferenciamento dos dados utilizando agrupamento em Mineração de Dados agrícolas.

1.5 Justificativa

Estudos sobre Agricultura de Precisão (AP) são recentes no Brasil, segundo Souza et al. (2010). De acordo com ROLOFF; FOCHT (2006) a AP começou em 1997. A Agricultura de Precisão demanda de técnicas e ferramentas variadas na solução de problemas encontrados na lavoura. A interpretação de grande volume de dados é uma das abordagens da AP. Para Souza et al. (2010), ferramentas computacionais são requeridas em diferentes contextos agrícolas e são importantes na tomada de decisão para o correto tratamento das lavouras considerando a variabilidade espacial encontrada nos talhões.

Autores diversos trabalham em propostas para o entendimento dos dados por meio de técnicas de Mineração de Dados (BOGORNY, 2007; GUIMARÃES, 2003; HAN; KAMBER; PEI, 2011; WITTEN; FRANK, 2005). Neste sentido, técnicas que

considerem o georreferenciamento dos mesmos, para encontrar padrões geográficos em função da distribuição espacial dos dados, tona-se especialmente interessante. Porém, essas ferramentas são escassas, o que leva a ser considerada a possibilidade de desenvolvimento de uma nova.

Com a necessidade de identificar padrões que considerem a distribuição geográfica dos dados, novos algoritmos de mineração de dados contribuirão para o entendimento de problemas agrícolas como a administração de recursos numa determinada região e a previsão de safra relacionada ao local agrupado pela semelhança de produção. Assim, considerando que há diferentes características na lavoura dependendo da posição espacial, decisões poderão ser tomadas respeitando essas características. Então, o tratamento é diferenciado para o manejo do solo na lavoura nas diferentes áreas, assim como o controle de pragas, entre outros.

A interpretação e visualização das regras encontradas se mostra como um fator importante depois de aplicadas as técnicas de Mineração de Dados. No trabalho de Guimarães et al. (2003), as regras encontradas foram interpretadas e consideradas coerentes por parte dos profissionais da agronomia, que as julgaram interessantes pela forma de apresentação no formato SE-ENTÃO. Porém o trabalho não apresenta as regras em forma de mapa. Portanto, torna-se viável o desenvolvimento de um módulo de visualização na ferramenta DAGER na forma texto, que é a forma comum adotada pelas ferramentas de mineração de dados e uma visualização mapeada dos resultados obtidos.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo aborda as principais teorias utilizadas como base para andamento da pesquisa, como a Agricultura de Precisão, técnicas de Mineração de Dados, Visualização na Mineração de Dados Georreferenciados e embasamento estatístico.

2.1 Agricultura de precisão (AP)

A agricultura de precisão tem evoluído nas últimas três décadas no Brasil de forma a auxiliar a obtenção de maiores lucros e também o cuidado do meio ambiente, que é a matéria prima principal para a agricultura existir. Dessa forma, no Brasil, começa em 1997 a implantação da AP que se mostra solução relevante e adaptável aos problemas que a agricultura do país enfrenta (ROLOFF; FOCHT, 2006).

A agricultura de precisão é conceituada diferentemente pelos autores. Para alguns é uma filosofia, para outros, uma tecnologia que auxilia a gestão. Para Batchelor; Whigham; Dewitt (1997) agricultura de precisão é uma filosofia de manejo da fazenda na qual os produtores são capazes de identificar a variabilidade dentro de um campo, e então manejar aquela variabilidade para aumentar a produtividade e os lucros. Blackmore; Wheeler; Morris (1994) relatam que a agricultura de precisão é o termo que descreve a meta de aumentar a eficiência do manejo da agricultura, sendo uma tecnologia em desenvolvimento, que modifica técnicas existentes e incorpora novas ferramentas para o administrador utilizar. Alguns autores dizem respeito a relação espacial que envolve a agricultura de precisão para defini-la. É o caso de Corá et al. (2004), que afirma que a AP como sistema de produção deve conhecer a variabilidade espacial de atributos do solo que controlam a produtividade de uma determinada cultura, utilizando-se do manejo regionalizado de insumos e práticas agrícolas como estratégias de sustentabilidade.

As tecnologias que proporcionam a existência e o avanço da AP podem ser agrupadas em seis principais categorias (COELHO, 2005):

1. Computadores e programas: computadores com programas para manejo dos dados, formação de gráficos e mapas;
2. GPS – *Global Position System* (Sistema de Posicionamento Global): esta tecnologia permite saber a localização em qualquer parte do planeta;

3. SIG – Sistema de Informação Geográfica: conjunto integrado de programas, equipamentos, metodologias, dados e pessoas (usuários) para realizar a coleta, armazenamento, processamento e análise de dados georreferenciados;
4. Sensoriamento Remoto: aquisição de informações de objeto não presente fisicamente;
5. Sensores: instrumentos que geram impulsos elétricos em resposta a estímulos físicos como calor, luz, magnetismo, movimento, pressão e som;
6. Controladores Eletrônicos de Aplicação: componente de um sistema informatizado onde a informação tem o objetivo influenciar o sistema para realizar a aplicação localizada de insumos.

Por meio do emprego correto das tecnologias, a AP tem a importância na agricultura de fazer a aquisição e tratamento das informações, para posterior visualização por um agrônomo, que fará a tomada de decisão. Portanto, o sucesso da agricultura de precisão depende da decisão agronômica.

Dentre as vantagens que a AP oferece, Batchelor; Whigham; Dewitt (1997) citam a obtenção de informações para tomar decisões de manejo mais embasadas e o registro de fazenda mais detalhado e útil. Porém, a vantagem mais abordada pelos autores (BATCHELOR; WHIGHAM; DEWITT, 1997; GENTIL; FERREIRA, 1999) é a redução de risco da atividade agrícola, visando conter custos e aumentar o lucro.

Como limitações da AP, Molin (2004) comenta a dificuldade no diagnóstico e definição do manejo apropriado, coletar e gerenciar dados, a interpretação e explicação dos dados, custo ainda elevado das soluções e assistência técnica, dificuldade operacional, dificuldade no entendimento da variabilidade, entre outras.

Portanto, mesmo que propostas diversas ferramentas para a AP, grande parte delas é de alto custo e sua adoção pelo produtor acaba não sendo viável.

A base de dados utilizada nesta pesquisa foi obtida por meio de processos de AP. Segundo Guimarães (2005), a base foi fornecida pelo Professor Doutor José Paulo Molin, pesquisador da Escola Superior Agricultura Luis Queiroz da Universidade do Estado de São Paulo – ESALQ/ USP. A base contém dados de uma região de 22 ha localizada em Campos Novos Paulista – SP referente a safras de soja a partir do ano 1998 até 2003.

2.2 O processo de descoberta de conhecimento em dados

Também conhecido como KDD (Knowledge Discovery in Databases), a Descoberta de Conhecimento em Bases de Dados (DCBD) é apresentada por alguns autores como sinônimo de Mineração de Dados. Entretanto, diversos autores apresentam a MD como um dos passos do KDD.

O processo de descoberta de conhecimento é iterativo e interativo, e seus passos segundo Han; Kamber; Pei (2011) são:

1. Limpeza dos dados: remove ruído nos dados e inconsistência. Remove ou ajusta dados que no contexto não fazem nenhum sentido, como a digitação errada de registros ou a duplicação;
2. Integração dos dados: combinação de *data sources*, ou seja, combinação das fontes de dados. Na integração de diferentes fontes de dados é possível atributos virem com unidades de medidas diferentes. Neste caso, a padronização dos dados é necessária;
3. Seleção dos dados: dados relevantes a tarefa de análise e descoberta de conhecimento são escolhidos dentre os atributos apresentados pela base de dados. Informações relevantes são descartadas, limitando o espaço de busca;
4. Transformação dos dados: os dados são transformados e consolidados em formas apropriadas para mineração por realização de sumário ou operações de agregação;
5. Mineração dos dados: métodos inteligentes são aplicados para extrair padrões nos dados;
6. Validação dos padrões: identificar padrões de interesse representando conhecimento baseado em medidas de interessabilidade;
7. Representação do conhecimento: visualização e técnicas de representação do conhecimento usadas para apresentar conhecimento minerado aos usuários.

Por meio dos passos citados anteriormente, é dado suporte a tomada de decisão, visto que o conhecimento é extraído por meio de técnicas bem definidas as quais são revistas para aumentar a qualidade das informações obtidas. A precisão vem do processo iterativo e interativo de descoberta de conhecimento nos dados.

A presente proposta evidencia os passos 5 e 7 do KDD, que é a mineração de dados e a representação do conhecimento respectivamente, por ter como objetivo criar nova

ferramenta de mineração de dados agrícolas georreferenciados, incluindo um módulo para a visualização.

2.3 Mineração de Dados (MD)

Uma crescente quantidade de dados é disposta para ajudar humanos em diferentes contextos. O armazenamento e estudo dos dados podem ter vários objetivos, como entender um contexto, um problema, formular hipóteses diversas, entre outros. Por exemplo, saber condições para manifestação de um determinado evento. Com a mineração de dados é possível extrair padrões ou conhecimento, dependendo do método empregado (SFERRA; CORRÊA, 2003).

Segundo Han; Kamber; Pei (2011) a Mineração de Dados é um assunto interdisciplinar, que pode ser definido de maneiras diferentes. A MD é o processo de extrair informações úteis e importantes de conjuntos de dados, como uma base de dados. Os conjuntos dos dados podem ser: base de dados, dados transacionais, repositórios de dados diversos, entre outros.

Para Witten; Frank (2005), a Mineração de Dados refere-se a resolução de problemas por meio de análise, e é definida como o processo de descobrir padrões.

Existem várias ferramentas voltadas para a MD, como o WEKA, ILLIMINE, KDB2000, KXEN, TANAGRA, KNIME, entre outras. Alguns destes softwares são gratuitos. Entretanto, problemas de agricultura podem vir a requerer uma ferramenta de mineração de dados georreferenciados.

Santos (2001) evidencia que os algoritmos de MD disponíveis nas ferramentas de descoberta de conhecimento tradicionais não estavam preparados para análise dos dados georreferenciados, motivando criação de novos algoritmos, adaptação de algoritmos e utilização de sistemas gestores de dados espaciais. Hoje ainda é possível ver que esse problema persiste (BOGORNY, 2006; ESTER; KRIEGEL; SANDER, 2001; HAN; KAMBER; PEI, 2011; SOUZA et al., 2010).

É importante, no contexto dessa pesquisa, citar Mucherino; Papajorgji; Pardalos (2009) que afirmam que a Mineração de Dados é amplamente utilizada para resolver problemas relacionados a agricultura.

2.4 Mineração de Dados Georreferenciados (MDG)

Muitos trabalhos têm sido desenvolvidos considerando o tratamento dos dados georreferenciados (BOGORNY et al., 2007; MORAES et al., 2011). Além de Sistemas de Informação Geográfica (SIG) a quantidade de informações georreferenciadas geradas em AP demanda de outros recursos, como é o caso de Molin (2006), que determina parâmetros de desempenho na colheita mecanizada. Outra forma de abordar essa demanda por informações georreferenciadas é a utilização de técnicas de Mineração de Dados.

As ferramentas baseadas em MD considerando dados georreferenciados surgem como solução para os problemas que a mineração de dados clássica não consegue solucionar. Existe uma crescente quantidade de informações sendo gerada a cada segundo e que exige métodos específicos para análise. Nesse contexto, a MDG procura descobrir padrões em bases de dados com atributos espaciais que levem a um conhecimento localizado a partir de um mapa, foto de satélite, entre outros.

Por meio dos atributos espaciais é possível ter um entendimento dos padrões baseados em sua localização geográfica e a relação dos objetos entre seus vizinhos. Por isso, ferramentas e algoritmos para a mineração de dados foram criados ou foram adaptados (BOGORNY et al., 2007).

A MDG é um dos passos da Descoberta de Conhecimento em Bases de Dados Geográficos (DCBDG). A diferença entre DCBD e DCBDG está nos atributos espaciais da base, que irão permitir a verificação dos relacionamentos entre atributos espaciais e não espaciais (BIGOLIN; BOGORNY; ALVARES, 2003).

Segundo Faria (1998) a classificação espacial é feita da seguinte forma:

- Localização: relacionamentos de localização produzem a representação espacial dos objetos de forma geométrica.
- Orientação: Os relacionamentos de orientação verificam a existência de relacionamento de orientação entre objetos geométricos. Exemplos: acima, abaixo, esquerda, direita.
- Métrica: Os relacionamentos métricos são os relacionamentos de distância ou comprimento que descrevem o afastamento de um objeto em relação ao outro.

- Topológicos: relacionamentos topológicos entre objetos como um sobrepondo outro, um contendo outro, um está contido em outro, um objeto é disjunto a outro objeto, um objeto é igual a outro e um objeto é adjacente a outro objeto (vizinhos).

Considerando a complexidade temporal da implementação de uma ferramenta de Mineração de Dados Georreferenciados, uma alternativa é aplicar técnicas de Mineração de Dados isoladamente, em módulos, para posteriormente estar sendo considerada a ferramenta uma solução em Mineração de Dados. Mesmo assim, ferramentas com objetivos limitados produzem resultados significantes. É o caso de SIG, GPS, ferramentas para sobreposição de mapas, entre outros.

2.5 Técnicas de Mineração de Dados Georreferenciados

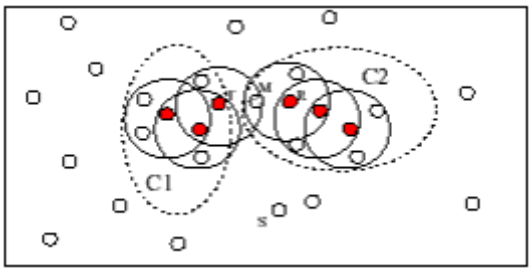
As técnicas de mineração de dados abordadas por Bogorny et al. (2007) e Ng; Han (2002) são a generalização, classificação, associação espacial e agrupamento. Para cada técnica foram criados algoritmos, ou algoritmos da MD foram adaptados para solucionar problemas de MDG. Devido o foco desta dissertação estar no agrupamento em Mineração de Dados, este tema é abordado a seguir.

2.5.1 Agrupamento

Os algoritmos de agrupamento têm um objetivo em comum, gerar partições de uma base de dados por meio da construção de uma nova coluna na Tabela objetos / atributos contendo a identificação do grupo ao qual esse objeto pertence. Para realizar essa tarefa os algoritmos utilizam diferentes heurísticas de particionamento que são auxiliadas pelo conceito de similaridade. Basicamente, um agrupamento deve formar grupos de alta similaridade entre as instâncias que compõe o grupo. De forma complementar, esse agrupamento deve ter na comparação entre grupos uma alta dissimilaridade entre eles, ou seja, uma instância de um grupo deve ter alta similaridade somente ao grupo a que pertence. Portanto, o agrupamento consiste em identificar conjuntos de objetos semelhantes entre si e diferentes de outros conjuntos. Uma das vantagens da técnica é que identifica estruturas diretamente dos dados sem o conhecimento prévio destes (MUCHERINO; PAPAJORGJI; PARDALOS, 2009). 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 (NEVES, 2002).

Figura 1 - Dois grupos (C1 e C2) encontrados pelo DBSCAN.



Fonte: Ester et al., 2001.

Ester; Kriegel; Sander (2001) traz a descrição do algoritmo de agrupamento baseado em densidade chamado de GDBSCAN, proposto por Sander et al. (1998). O GDBSCAN é um algoritmo generalizado do DBSCAN, com o qual é possível encontrar clusters espaciais como observado na Figura 1.

2.5.2 PAM

O algoritmo PAM (*Partitioning Around Medoids*) foi desenvolvido por Kaufman; Rousseeuw (1990). Para encontrar o agrupamento é determinado um objeto representativo para cada agrupamento, chamado de centróide. Cada centróide deve ser o objeto mais central perante seu agrupamento. Assim, cada objeto não centróide será associado ao centróide que mais se assemelha. É dito que um objeto Oj pertence a um grupo representado pelo centróide Om desde que a distância do objeto Oj a Om seja a menor perante todos os outros centróides. Esta afirmação é ilustrada na Equação 1:

Equação 1 - Menor distância de Oj perante centróides.

$$d(Oj,Om) = min_{Oe} d(Oj,Oe)$$

De acordo com a Equação 1, min_{Oe} significa o mínimo entre todos os centróides Oe. A diferença entre os objetos Oj e Om é representada por d(Oj, Om). A diferença pode ser uma medida de dissimilaridade como a Euclidiana ou a Manhatan.

O algoritmo PAM começa elegendo arbitrariamente k centróides, onde k é a quantidade de grupos e por consequência a quantidade de centróides que irão representar os grupos. Uma substituição é feita em cada passo, desde que implique em melhora.

Para o entendimento do algoritmo PAM, considera-se:

- Om é um centróide;
- Op é um possível novo centróide;
- Oj é um não centróide;
- Oj2 é um centróide próximo a Oj.

A substituição de Om por Op envolve um custo de substituição que determina se haverá ganho de qualidade ao agrupamento pela substituição. Este custo envolve quatro possíveis casos (NG; HAN, 2002):

Caso 1: Oj atualmente pertence ao cluster representado por Om. Oj é mais similar a Oj2 que Op, ou seja, a distância entre Oj e Op é maior ou igual a distância entre Oj e Oj2. Se Om for substituído por Op como centróide, Oj pertencerá ao cluster representado pelo centróide Oj2.

O custo desta troca é representada pela Equação 2 a seguir:

Equação 2 - Custo da troca quando acontecer o caso 1.

$$Csubst = d(Oj, Oj2) - d(Oj, Om)$$

Caso 2: Oj pertence ao agrupamento representado pelo centróide Om. Oj é menos similar a Oj2 que Op. Então, Om sendo substituído por Op, Oj irá pertencer ao agrupamento representado por Op.

O custo da troca do caso 2 é:

Equação 3 - Custo da troca do caso 2.

$$Csubst = d(Oj, Op) - d(Oj, Om)$$

Caso 3: Oj pertence a um agrupamento qualquer, ou seja, não pertence a Om. Oj2 é o centróide que representa o agrupamento o qual Oj está. Em relação a Oj, Oj2 é mais similar a que Op. Neste caso o custo é:

Equação 4 - Custo da troca referente ao caso 3.

$$Csubst = 0$$

Caso 4: Oj pertence ao agrupamento representado por Oj2. Oj é menos similar a Oj2 que Op. Substituindo Om por Op, implica em Oj pertencer a Op. O custo desta troca é:

Equação 5 - Custo da troca do caso 4.

$$Csubst = d(Oj, Op) - d(Oj, Oj2)$$

Combinando os quatro casos possíveis, é associado um custo total de substituir o Om por Op, que é:

Equação 6 - Custo total da substituição.

$$TCsubst = \sum_{j} Csubst$$

A partir do custo total da substituição, obtêm-se a resposta de qual o melhor centróide para representar aquele grupo.

Após a abordagem desses conceitos, é apresentado o algoritmo PAM na Figura 2.

Figura 2 - Algoritmo PAM

1. Selecionar k objetos arbitrariamente;
2. Calcular TCsubst para todos os pares de objetos Om, Op;
3. Selecionar o par Om, Op o qual corresponde ao $min_{Om,\ Op}$ TCsubst. Se TCsubst for negativo, substituir Om por Op e voltar ao passo 2.
4. Para todos os objetos não selecionados, encontrar o centroide mais similar.

Fonte: Ng; Han, 2002.

Segundo Kauffman; Russeeuw (1990), o algoritmo PAM é satisfatório para procura de grupos em pequenas bases (cerca de 100 instâncias e até 5 grupos). Esse algoritmo tem uma complexidade temporal por ser exaustivo, isto é, ele verifica qual é o melhor centróide comparando-se todas as instâncias da base. Estes autores ainda trazem implementação do algoritmo em FORTRAN - *IBM Mathematical FORmula TRANslation System*.

2.5.3 CLARA

Para reduzir a complexidade temporal de PAM, foi criado por Kaufman; Russeeuw (1990) o CLARA (*Clustering LARge Applications*). O algoritmo CLARA parte da premissa que os centróides podem ser encontrados a partir de uma amostra da base, ou que esses centróides da amostra sejam muito próximos dos centróides ideais. A partir dessa amostra, executa-se o PAM. Kauffman; Russeeuw (1990) afirmam que resultados satisfatórios foram encontrados com cinco amostras de tamanho 40+2k instâncias. Na Figura 3 é apresentado o algoritmo CLARA.

Figura 3 - Algoritmo CLARA

1. Para i=1 até 5, repita os seguintes passos:
2. Obtenha amostra de 40+2k instâncias da base. Execute PAM na amostra para encontrar os k centróides.
3. Para cada objeto Oj na base, determinar qual dos k centróides é o mais similar a ele.
4. Calcule a dissimilaridade média do agrupamento obtido no passo anterior. Se este valor for menor que o atual mínimo, utilize este valor como o atual mínimo e guarde os k centróides encontrados no passo 2 como os melhores centróides.
5. Retornar ao passo 1 seguindo com a próxima interação.

Fonte: Ng; Han, 2002.

Na Figura 3 é observado o funcionamento do algoritmo CLARA, que com base em amostras da base encontra aproximadamente os melhores centróides. Para isso, faz coleta da amostra e aplica o algoritmo PAM para encontrar os centróides. A busca pelos centróides em amostras acontece por 5 vezes, e o menor valor de TCmp de encontrado nessas repetições indicará os melhores centróides.

2.5.4 CLARANS

Com o objetivo de melhorar a eficácia de CLARA, Ng; Han (2002) propuseram o algoritmo CLARANS utilizando a teoria de grafos.

O grafo proposto por CLARANS determina que cada nó possua k elementos, onde k é o número de centróides, portanto, os elementos do nó são considerados possíveis centróides. Dois nós são vizinhos desde que apenas um desses k elementos seja diferente. Formalmente, a interseção de dois nós deve ter resultado k-1. Por exemplo, supondo que k seja 2, existe um nó com os objetos Obj1 e Obj2. Também existe outro nó com os objetos Obj1 e Obj3. Neste caso, esses dois nós são vizinhos devido a intersecção resultar em Obj1, ou seja, k-1=2.

Na Figura 4 é apresentado o algoritmo CLARANS.

Figura 4 - Algoritmo CLARANS.

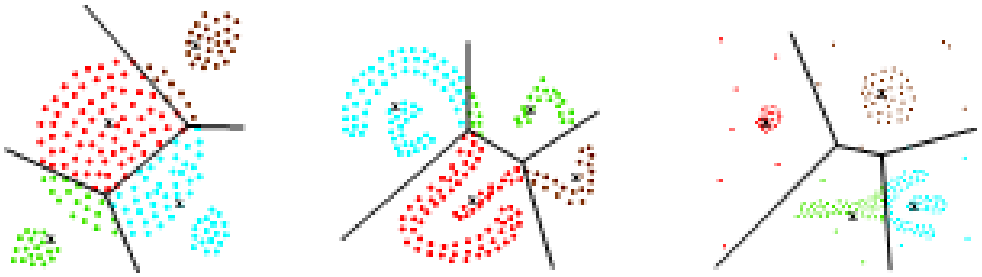
1. Insira os parâmetros numlocal e maxneighbor. Inicialize i com o valor 1. Coloque um valor alto em mincost.
2. Configure como current um nó qualquer no grafo.
3. A variável j deve receber o valor 1.
4. Considere um vizinho qualquer de current, denominado S. Baseado em TCsubst, enunciado pela Equação 6, estipule o custo diferencial dos dois nós.
5. Se S tem menor custo, current passa a ser S. Volte ao passo 3.
6. Caso contrário, incrementar j em 1. Se j<=maxneighbor, ir ao passo 4.
7. Caso j>maxneighbor, comparar o custo de current com mincost. Se current menor que mincost, current passa a ser mincost e bestnode passa a ser current.
8. Incrementar i em 1. Se i>numlocal, bestnode é a resposta e finaliza. Caso contrário, voltar ao passo 2.

Fonte: Ng; Han, 2002.

Na Figura 4, maxneighbor significa a quantidade máxima de vizinhos que serão comparados a partir de current. O atributo mincost é o menor custo. A repetição da busca por outro conjunto de menor custo até que numlocal seja atingido.

Na Figura 5 é representado grupos encontrados a partir de CLARANS.

Figura 5 - Três bases e seus grupos encontrados por CLARANS.



Base 1 **Base 2** **Base 3**

Fonte: Sander et al., 1998.

Como é possível observar na Figura 5 cada cor é um agrupamento. Foram utilizadas três diferentes bases e para cada base foram criados quatro grupos.

Quando analisados os algoritmos PAM, CLARA, CLARANS, é percebido que o funcionamento de PAM é aproveitado na construção de CLARA e CLARANS . A diferença principal é que o custo total de substituir os centróides é obtido verificando-se todas as instâncias no PAM. No CLARA, é a partir de amostra. No CLARANS as instâncias são dispostas na forma de grafos, onde cada nó é um grupo de centróides. Então é feita busca dos melhores centróides por meio da busca do nó com menor custo TCsubst. A busca não é feita em todos os nós, mas de forma aleatória a partir de um vizinho. Desta forma, não restringe a busca a uma amostra da base composta aleatoriamente, que é a proposta do CLARA.

O algoritmo tem implementação em MatLab, que é uma ferramenta paga. A implementação deste algoritmo nesta pesquisa, porém, foi baseada unicamente no pseudo-código (NG; HAN, 2002).

2.6 Visualização de resultados na MDG

A visualização de dados na MDG é de suma importância, pois é por meio da interpretação correta dos dados, informação e regras será agregado ao conhecimento a ser adquirido. Rabelo (2007) afirma que o mau emprego de técnicas de visualização pode comprometer o KDD.

Miller (2008) classifica a visualização de informação em:

1. Baseado em mapa: Permite o usuário interativamente representar dados georreferenciados em uma forma de gráfico;
2. Baseado em gráfico: representa os dados em grafo ou 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 a representação visual baseada em 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.

3 METODOLOGIA

Para formação de grupos baseados em vizinhança geográfica, foram implementados alguns algoritmos conhecidos. Baseados nestes algoritmos, foram implementadas modificações nestes algoritmos de agrupamento geográfico para que considerassem o georreferenciamento das instâncias em conjunto com as características não espaciais.

Para funcionamento dos algoritmos, foi criada uma ferramenta para consideração de dados georreferenciados quanto ao agrupamento e visualização, e quanto aos testes e validação dos algoritmos foi utilizada uma base de dados obtida por processos de AP referente ao levantamento de dados feito em campo, contendo características físico-químicas do solo em uma lavoura de grãos.

3.1 GCLARA

Como a pesquisa visava encontrar padrões úteis para o domínio agrícola, foi acrescentado ao algoritmo CLARA i a análise dos demais atributos e foram realizadas adaptações quanto a estrutura de dados e funcionamento. Portanto, foi adaptado o algoritmo CLARA para continuar considerando os atributos latitude e longitude somados a outros atributos responsáveis por características diversas, como produção ou matéria orgânica. A essa versão foi dado o nome de GCLARA (Generalized CLARA).

O funcionamento do GCLARA é semelhante ao CLARA. Para fazer os grupos, o algoritmo inicia coletando amostras aleatórias da base de dados. A partir daí, elege os melhores centróides. Cada uma das instâncias não centróides da base de dados, então, passam a fazer parte de um grupo cujo centróide ela mais se assemelha. A diferença está em que cada objeto Om, Oj, Oj2, Op tem os atributos latitude e longitude e também atributo(s) como índice de cone ou matéria orgânica. Desta forma, foi possível considerar separadamente atributos como produção de soja, matéria orgânica e valores referentes a índice de cone.

3.2 GCLARANS

CLARANS procura os grupos considerando somente os atributos georreferenciados. GCLARANS passa a considerar qualquer atributo não espacial juntamente com o espacial para buscar os grupos. O algoritmo foi proposto devido a necessidade de agrupar áreas na

lavoura com características comuns. Primeiramente foi considerada a produção como um fator determinante, mas após ajustes no algoritmo percebeu-se que o mesmo princípio poderia ser adotado para qualquer atributo não georreferenciado.

A diferença entre os algoritmos GCLARA e GCLARANS é em seu funcionamento, não na consideração dos atributos. Sendo assim, o algoritmo GCLARA coleta amostras aleatórias da base de dados e com repetição e comparação se chega a conclusão dos centróides ideais ou aproximadamente ideais. Já o GCLARANS, executa uma busca em grafo eme parte da base de dados. Com isso, o algoritmo pretende encontrar melhores centróides num espaço amostral sem estar confinando a base a uma amostra, mas confinar a busca em uma porção da base.

3.3 A ferramenta: DAGER

Para execução dos algoritmos, é foi proposta uma nova ferramenta computacional, denominada DAGER, que tem por objetivo fazer o agrupamento das instâncias por meios dos algoritmos: PAM, CLARA, CLARANS, GCLARA e GCLARANS. Estes algoritmos foram escolhidos a partir de pesquisa, considerando que haviam informações suficientes nas bibliografias por meio de pseudocódigo, o que facilita o entendimento e a implantação. Além disso, esses algoritmos foram testados e aprovados para solucionar problemas de natureza que a MDG pode resolver. A ferramenta implementada atendeu o objetivo de inserir novos algoritmos, tanto que até o término desta proposta faziam parte dela 5 algoritmos: PAM, CLARA, CLARANS, GCLARA E GCLARANS. Como visualização, traz na forma de gráfico o resultado dos grupos utilizando-se da técnica de pixel.

Para estudo dos dados georreferenciados a base de dados deve conter informações geográficas relacionadas aos atributos como latitude e longitude.

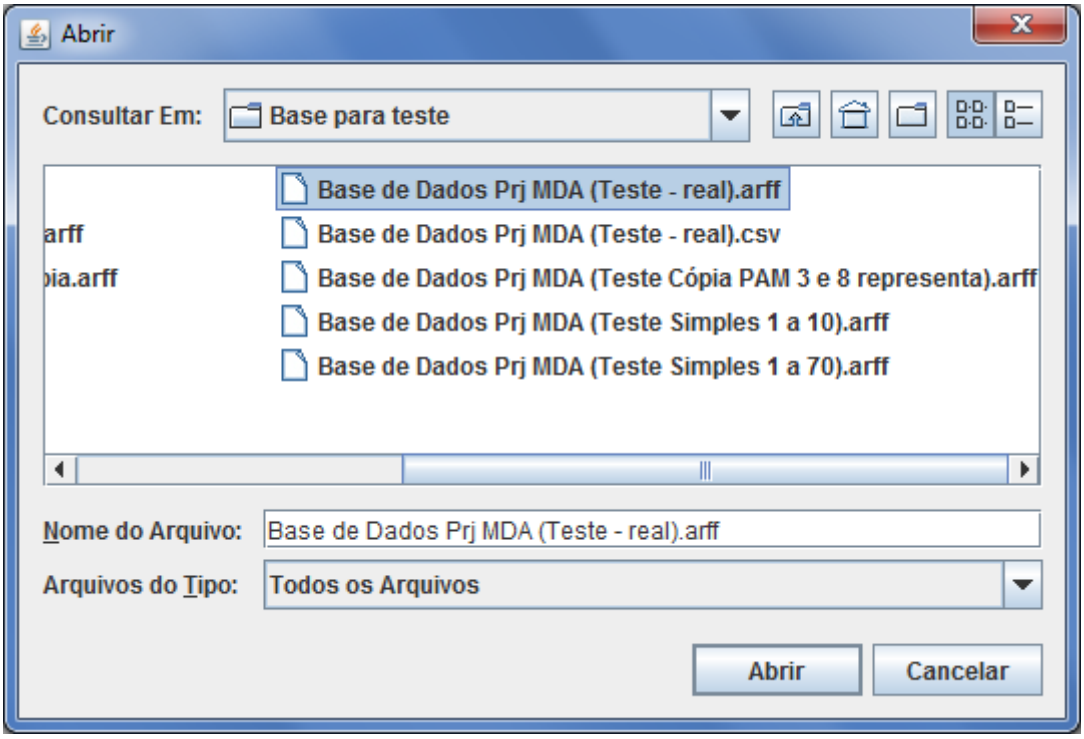
A ferramenta DAGER está organizada em módulos da seguinte maneira:

1. Ativação 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.

3.3.1 Ativação da base de dados

A base de dados deve ser informada pelo usuário, para que o sistema possa fazer a leitura. O sistema fará a leitura de arquivos ARFF desde que respeitada a estrutura como o sistema espera. Essa estrutura será semelhante a de outra ferramenta conhecida, o Weka.

Figura 6 - Escolha do arquivo ARFF que contém a base.



Fonte: O autor.

Para ativar a base de dados no sistema, é selecionado o arquivo como mostrado na Figura 6.

Figura 7 - Amostra de um arquivo ARFF referente a base usada.

```
@relation MDA_agri

@attribute 'X' real
@attribute 'Y' real
@attribute 'Soja (kg/ha)' real

@data
-49.98539534,-22.70066087,3149.30
```

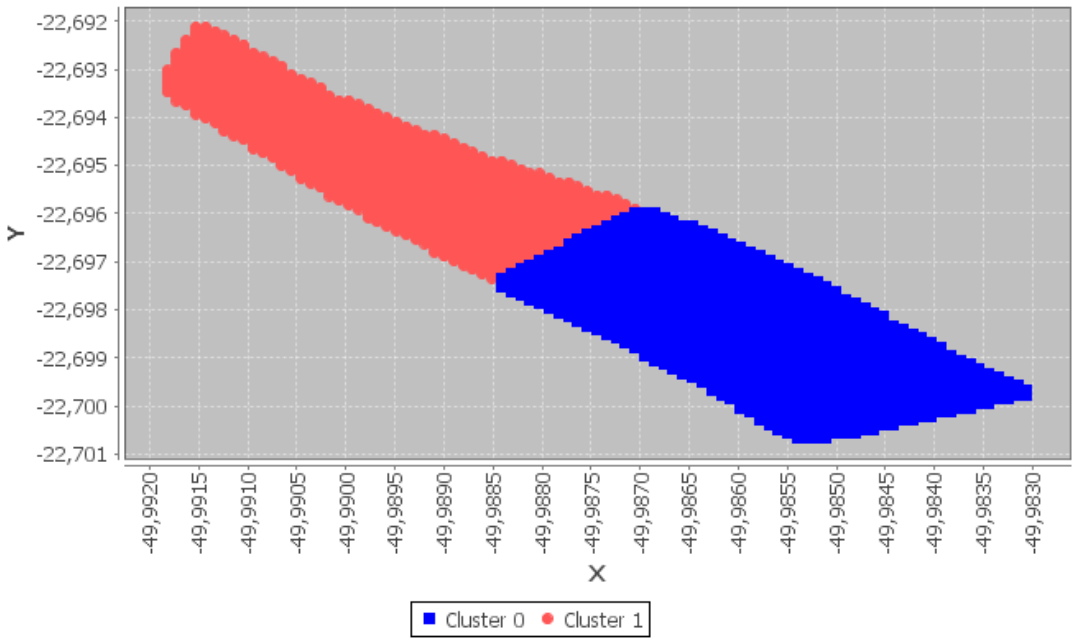
Fonte: O autor.

Na Figura 7 é possível visualizar uma amostra de um arquivo ARFF utilizado no decorrer do trabalho. Está descrito o @relation referente ao nome da base, o @attribute que é um atributo da base e o @data, que são os dados. Esses dados estão na ordem dos atributos.

3.3.2 Agrupamento

Após a ativação e leitura bem sucedida do arquivo ARFF, o sistema estará pronto para fazer o agrupamento. Os algoritmos PAM, CLARA e CLARANS fazem o agrupamento considerando somente os atributos latitude e longitude. Os algoritmos propostos GCLARA e GCLARANS consideram além dos atributos de latitude e longitude, outras características informadas pelo usuário.

Figura 8 - Dois grupos (cluster) representados na ferramenta.



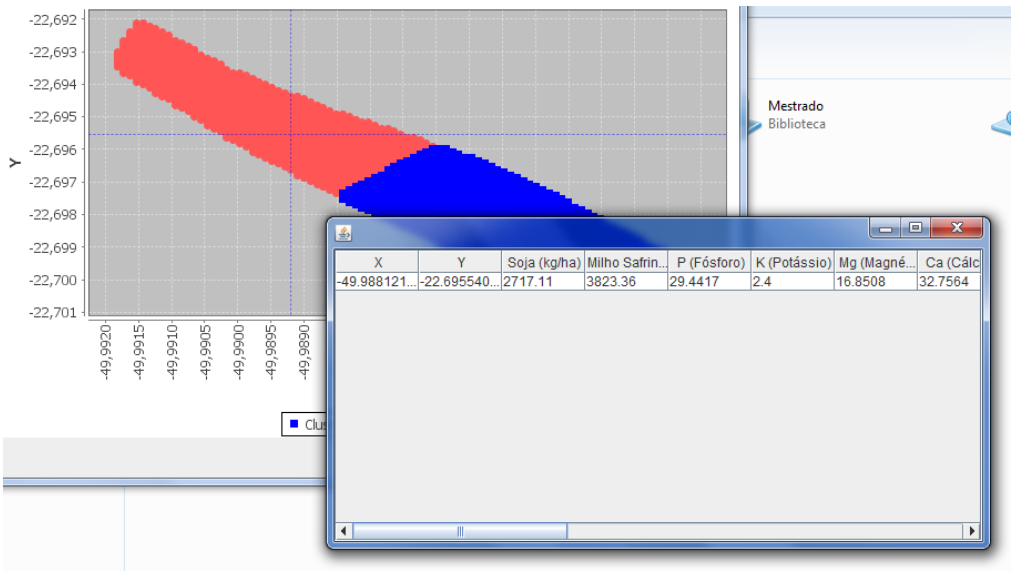
Fonte: O autor.

Na Figura 8 é visualizado o resultado de um experimento de agrupamento em que desejava-se obter dois grupos considerando somente os atributos latitude e longitude, com o objetivo de dividir a propriedade geograficamente. Cada ponto representado na ferramenta pode ser consultado, podendo-se então saber quais são os valores dos demais atributos daquela instância.

3.3.3 Visualização

Após os resultados serem encontrados, poderão ser escolhidas duas formas de representação: em texto e mapa. Ao escolher a representação de resultados em texto, é exibido na interface o conteúdo existente no arquivo gerado pelo agrupamento ou pela descoberta de conhecimento. Quando a visualização escolhida é na forma de mapa, as informações contidas no arquivo de resultados são interpretadas e são preenchidas na matriz de dispersão de forma a ter uma interpretação espacial de onde as regras encontradas estão sendo aplicadas. A organização de matriz se mostra eficiente para a representação das informações contidas nos resultados.

Figura 9 - Atributos da instância selecionada no gráfico.



Fonte: O autor.

Na Figura 9, é observado o valor dos atributos referentes a instância selecionada no gráfico. Devido a base possuir muitos atributos, para visualizar os demais atributos deve-se utilizar a barra de rolagem.

3.4 A base de dados

A base de dados foi obtida por meio de processos de AP na região de Campos Novos, no estado de São Paulo. Organizados numa planilha eletrônica, estes dados fornecem valores referentes a número da amostra, posição da amostra segundo georreferenciamento, nutrientes, matéria orgânica, acidez do solo, índice de cone, e a produção de soja, que é o atributo meta. Esta base possui 2416 registros. Nesta seção encontram-se informações pertinentes a cada um desses atributos.

ID – Identificação: Número sequencial para controle das coletas. Este atributo serve para organizar a coleta. Na mineração de dados, atributos com essa característica comumente são excluídos, por não representarem relação com o atributo meta, que nesse artigo é a produção de soja.

Atributos X e Y – Georregerenciamento: Estes atributos são referentes a longitude e a latitude, ambas medidas em graus. A longitude é feita a medida de uma distância de um ponto qualquer da terra a um meridiano (que se torna ponto inicial da medida). A latitude é medida de um ponto qualquer da terra até a linha do Equador.

Soja - Indica a produção de soja: A soja nesta base está representada em valores Kg/ha. Este atributo é o atributo considerado como meta. Por meio de observação, foi encontrada na base instâncias que registraram produção zero de soja. Estes registros foram descartados devido o os demais atributos possuírem valores. Desta forma, as poucas instâncias nessa situação não forçariam a uma conclusão errada, como a que uma determinada combinação de valores de atributos implicarem na falta de produção de soja.

Fósforo (P): O Fósforo é considerado um dos três nutrientes primários, juntamente com o Nitrogênio (N) e o Potássio (K). Seu papel é imprescindível na produção de energia (ATP) para o metabolismo das plantas, portanto, todos os demais nutrientes, direta ou indiretamente dependem dele.

Potássio (K): O Potássio é importante na fotossíntese, na formação de frutos, resistência ao frio e às doenças das plantas. O Potássio não é constituinte de nenhuma molécula orgânica no vegetal, entretanto, contribui em varias atividades bioquímicas, sendo um ativador de grande número de enzimas, regulador da pressão osmótica (entrada e saída de água da célula), abertura e fechamento dos estômatos.

Magnésio (Mg): O Magnésio é um constituinte da molécula de clorofila, necessário a várias reações enzimáticas. Por ser constituinte da clorofila sua deficiência aparece com um amarelecimento entre as nervuras das folhas mais velhas.

Cálcio (Ca): É um dos chamados macronutrientes secundários junto com o Magnésio (Mg) e o Enxofre (S). Os efeitos indiretos do cálcio são tão importantes quanto o seu papel como nutriente. O cálcio promove a redução da acidez do solo, melhora o crescimento das raízes, aumento da atividade microbiana, aumento da disponibilidade de molibdênio (Mo) e de outros nutrientes. Reduzindo a acidez do solo, diminui a toxidez do

alumínio (Al), cobre (Cu) e manganês (Mn). Plantas que apresentam altos teores de cálcio resistem melhor a toxidez destes elementos.

Matéria Orgânica (MO): A Matéria Orgânica é encontrada principalmente nas camadas superficiais do solo, suprindo os nutrientes aos vegetais e proporcionando propriedades físicas favoráveis ao crescimento das plantas. Métodos convencionais de preparo do solo e a baixa produção de resíduos são alguns dos fatores que reduzem a matéria orgânica. Para aumentar a matéria orgânica, deve ser adotado um sistema de rotação de culturas com inclusão de espécies de alta produção de fitomassa. Além disso, a redução do revolvimento do solo. Com a aplicação do carbono no solo, ocorre o aumento da concentração de biomassa, que é a matéria orgânica.

Hidrogênio e Alumínio (H+Al): Este atributo é determinante para saber se há acidez no solo. Na Tabela 8, é possível verificar o baixo desvio padrão para esse atributo. Além de determinar a acidez do solo, este atributo tem correlação com outro, o CTC (Capacidade de Troca de Cátions).

Acidez ativa (pH): São os íons H+ da solução do solo em concentrações baixas quando comparados com os íons H+ adsorvidos. É expressa pelo valor do pH do solo. O pH é determinado em água: usa-se uma relação solo:água de 1:2,5. Mede-se o pH através da imersão de um elétrodo de vidro ligando a um potenciômetro. Podem ser usadas, também, soluções salinas, como o KCl e CaCl2.

Soma das Bases (SB): É a soma das bases trocáveis K+, Ca2+, Mg2+ e às vezes Na+ quando este é significativo. Quanto maior a soma de bases, maior a fertilidade do solo. Em outras palavras, é a indicação de que o solo tem nutrientes disponíveis para as plantas.

Capacidade de Troca Catiônica (CTC): Obtida pela fórmula [K + Ca + Mg + (H+Al)], a CTC é a capacidade do solo segurar alguns nutrientes, no nível de acidez atual para cálcio, magnésio e potássio.

Saturação por Bases (V%): Obtida pela fórmula [(K + Ca + Mg)*100 / (K + Ca + Mg + (H+Al))] (Pauletti, 2004): é a porcentagem de bases trocáveis em relação à CTC. A

saturação por bases é um excelente indicativo das condições gerais de fertilidade do solo, sendo utilizada até como complemento na nomenclatura dos solos. Os solos podem ser divididos de acordo com a saturação por bases: solos eutróficos (férteis) = V% ≥ 50%; solos distróficos (pouco férteis) = V% < 50%. Quanto maior o valor de V%, mais nutrientes a planta tem para absorver.

Índice de Cone (IC): Este atributo é um índice, expresso em Pascal, para fornecer um método padrão de caracterização uniforme de resistência à penetração nos solos (American Society of Agricultural Engineers, 1999).

3.5 Linguagem de desenvolvimento e bibliotecas

Para desenvolvimento da proposta foi utilizada a linguagem de programação JAVA. Além de JAVA, bibliotecas escritas na linguagem foram utilizadas.

A biblioteca jfreechart 1.0.14 foi utilizada para auxiliar na exibição dos gráficos. Como dependência, esta biblioteca exige a biblioteca jcommons. Portanto, a versão jcommon usada foi a 1.0.17.

A escolha da linguagem de programação JAVA se justifica por ser gratuita para o uso acadêmico. Além disso, possui a vantagem de ser multiplataforma, ou seja, funciona em diversos Sistemas Operacionais como Windows e Linux. Porém, deve-se instalar o JRE (Java *Runtime Environment*) correspondente a cada Sistema Operacional.

4 RESULTADOS E DISCUSSÃO

Nesta seção é apresentado o resultado do funcionamento da ferramenta DAGER. Em seguida é abordado o resultado dos agrupamentos por meio de experimentos com os algoritmos GCLARA e GCLARANS.

4.1 A ferramenta para estudos de dados georreferenciados – DAGER

Esta dissertação traz um produto com o fim de contribuir para as pesquisas acadêmicas no auxílio ao desenvolvimento tecnológico e pesquisas para o desenvolvimento da área agrícola. A ferramenta DAGER se mostrou eficiente como sendo um ambiente para execução de algoritmos de agrupamento, bem como a visualização dos resultados, tanto pela forma textual como gráfica.

Em relação a visualização dos grupos, por meio de cores e símbolos representa com clareza o talhão de onde foram coletadas as amostras e também como foram agrupadas essas amostras. Este agrupamento foi realizado com os algoritmos presentes nesta proposta e suas respectivas saídas alimentavam o gráfico que a ferramenta apresenta.

4.2 GCLARA

No algoritmo GCLARA foi necessário ajustar a quantidade total de instâncias a serem analisadas. Com apenas 10% das instâncias analisadas, o algoritmo GCLARA apresentava resultados diferentes para o mesmo número de grupos. Para corrigir este problema, o algoritmo passou a considerar na soma total de amostras distribuídas nas repetições, cerca de 25% das instâncias para ter resultados semelhantes nas repetições dos experimentos com a mesma quantidade de grupos. É necessário que seja mencionado neste trecho que CLARA e GCLARA trabalham com amostra da base e essa amostra é composta aleatoriamente. Por esse fato, os grupos podem variar simplesmente pela escolha da amostra e determinação dos centróides. Entretanto, os centróides encontrados se aproximam ou até coincidem com os centróides ideais e as instâncias encontradas nesses grupos coincidem em pelo menos 83% das instâncias para 10 repetições utilizando a mesma quantidade de grupos. A diferença principal entre GCLARA e CLARA é que com uma amostragem de 10 % da base, CLARA foi capaz de produzir resultados satisfatórios na busca de grupos obtidos pela

proximidade espacial. O algoritmo GCLARA precisa analisar ao menos 25% da base para apresentar resultados satisfatórios na formação de grupos semelhantes considerando a influência da proximidade espacial.

4.3 GCLARANS

Este algoritmo acabou sendo pouco diferenciado em relação ao GCLARA. Isto se deu ao fato de que a amostra em GCLARA ter considerado pelo menos cerca de 25% das instâncias da base. O propósito de CLARANS e GCLARANS é buscar os melhores centróides partindo de uma representação em grafo. Quando GCLARA obteve melhores resultados aumentando o espaço amostral, o resultado foi significativamente próximo aos resultados encontrados por GCLARANS.

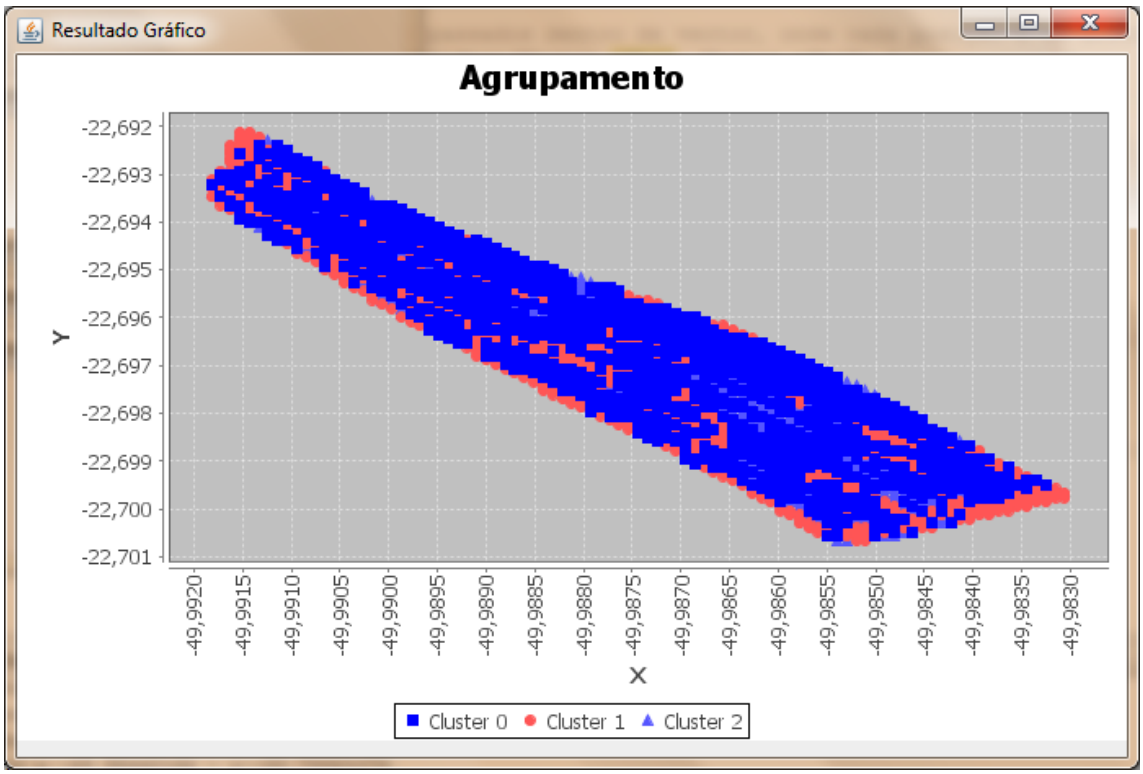
GCLARANS trouxe resultados muito próximos para cada repetição nos experimentos utilizando a base de dados agrícolas e comparando mesma quantidade de grupos. Portanto, não houve ajuste do algoritmo para se adaptar a necessidade de produzir resultados satisfatórios. Foi observado que o algoritmo conseguiu manipular atributos diversos juntamente de latitude e longitude produzindo resultados próximos nas repetições desde o começo. O resultado deste algoritmo, de forma semelhante a GCLARA, foi o encontro de regiões com características semelhantes.

Os experimentos foram realizados com ambos os algoritmos e devido a proximidade de resposta, foram considerados os agrupamentos obtidos na primeira execução de GCLARA. Além disso, para controle de condições no experimento, optou-se por realizar os experimentos sempre com 3 grupos.

4.4 Experimento 1: Produção.

Neste experimento o objetivo foi visualizar a distribuição da produção pelo talhão. O resultado foi exibido na Figura 10. Nela é possível ver 3 grupos, identificados pelas cores azul escuro (grupo 0), vermelho (grupo 1) e azul claro (grupo 2).

Figura 10 - Três grupos levando em consideração somente a produção.



Fonte: O autor.

Neste experimento, a proximidade entre pontos foi ignorada e somente foi comparado o valor da produção. Os grupos então apenas foram distribuídos conforme a produção de soja esperando-se três grupos de intervalos de produção: mínima - média, média, média - máxima. Entende-se por média apenas o valor numérico e não uma produção média da cultura soja.

Tabela 1 - Dados referentes ao experimento utilizando produção.

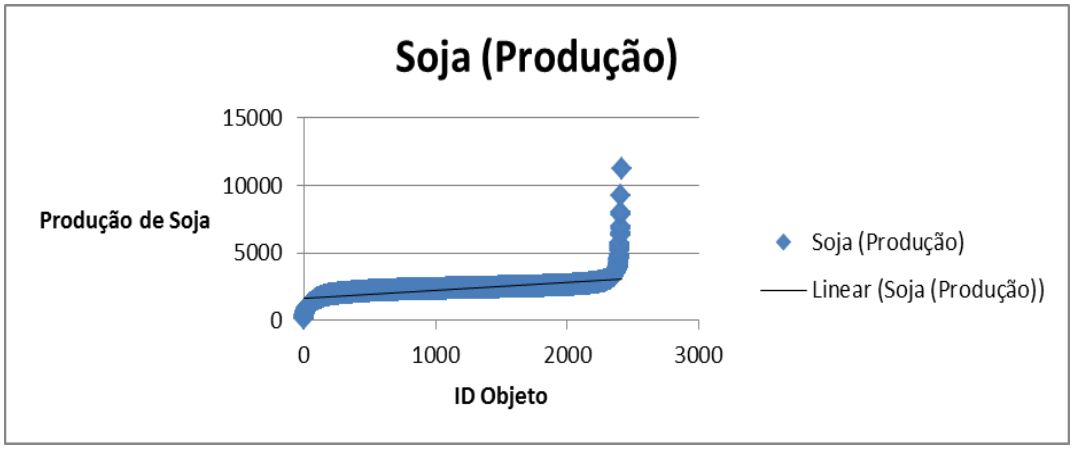
	Grupo 0	**Grupo 1**	**Grupo 2**
Produção de Soja mínima	2227,51	136,52	2664,99
Produção de Soja máxima	2662,42	2227,09	11196,29
Produção de Soja média	2423,37	1883,40	3144,65
Desvio padrão	112,94	391,24	936,77
Produção de soja registrada no centróide	2411,93	2042,9	2914,06
Quantidade de instâncias	1461	621	330
Correspondente em % das instâncias em relação ao total da base	60,57	25,74	13,68

Fonte: O autor.

Na Tabela 1 são distribuídas as informações referentes a cada agrupamento. É importante salientar que os grupos foram realizados considerando somente o atributo produção de soja. Cerca de 60% das instâncias compõe o grupo 0. Este agrupamento além de ser o maior, é o que possui menor desvio padrão. O grupo 1 foi responsável pela parcela de menor produção da base e o grupo 2 o maior. O grupo 1 tem cerca de 26% da base e possui desvio padrão menor que o grupo 2. O grupo 2 tem o maior desvio padrão e pega a parcela da produção que vai da média até a máxima produção registrada.

Para complementar as informações da Figura 10, foi feito um gráfico de dispersão utilizando toda a base (Figura 11). No eixo X está a identificação da instância (ordem crescente a partir de zero). No eixo Y está a produção referente a soja. Percebe-se na Figura 11 que há maior concentração de instâncias na faixa de produção de 2000 até 4000 Kg/Ha.

Figura 11 - Gráfico de dispersão sobre o agrupamento da Figura 10.



Fonte: O autor.

Observando o valor de produção de cada centróide fica mais clara a linha de tendência da Figura 11. Os grupos 0, 1 e 2 tem como valor de centroide para a produção de soja 2411,93; 2042,9 e 2914,06, respectivamente. O grupo 0 tem maior número de instâncias, seguido do grupo 1.

O objetivo deste experimento foi ter um panorama geral da produção para que então fosse usado como base para comparação com os demais experimentos. Assim, espera-se obter entendimento de como a lavoura está se comportando conforme são considerados os demais atributos.

Tabela 2 - ANOVA aplicada a produção.

Fonte da variação	SQ	Gl	MQ	F	valor-P	F crítico
Entre grupos	6845620119,83	1	6845620119,83	43889,57	0	3,84
Dentro dos grupos	752105305,74	4822	155973,73			
Total	7597725425,57	4823				

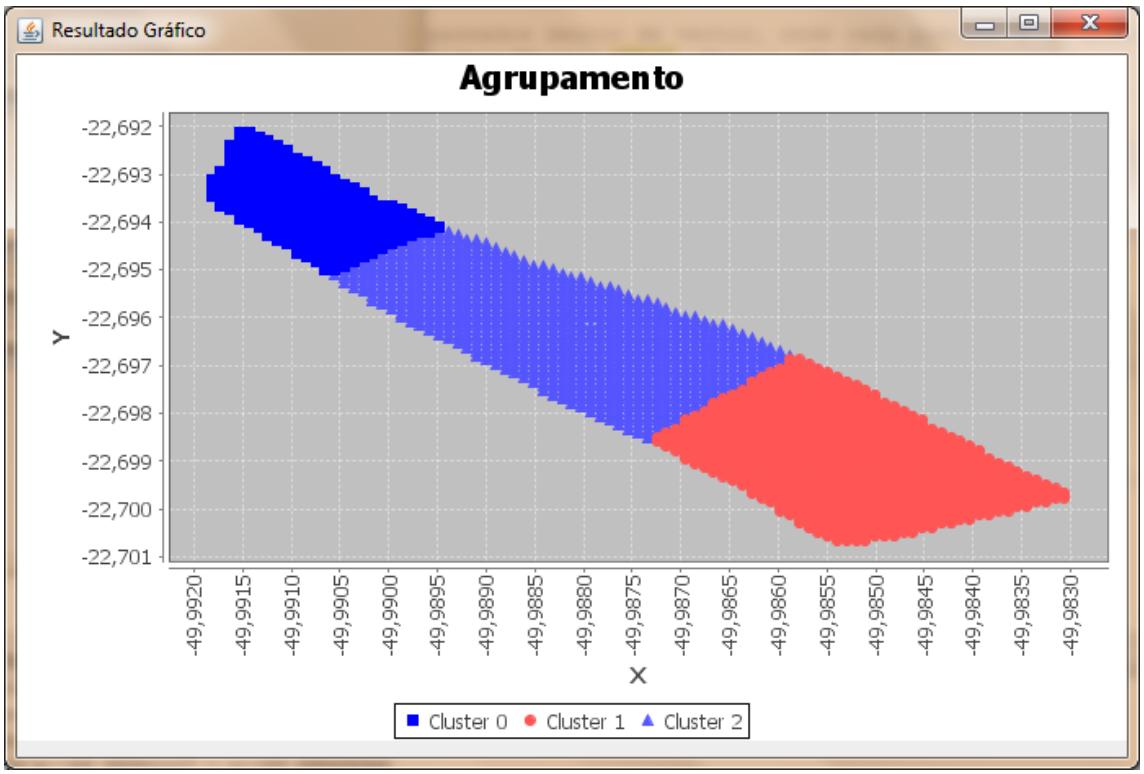
Fonte: O autor.

Na Tabela 2, com $\alpha = 5\%$, verificou-se por meio da ANOVA (Análise de Variância) o elevado relacionamento entre produção de soja e seu grupo e a diferença entre os grupos. Sendo observado na análise o valor de F que é maior que de F crítico, é dito que a produção tem diferença entre grupos. Além disso, o valor-P é menor que 5%, implicando em ignorar a igualdade de grupos.

4.5 Experimento 2: atributos de latitude e longitude (sem produção)

O objetivo deste experimento foi dividir a lavoura em três partes utilizando como base a vizinhança geográfica dos pontos em relação ao centróide de cada grupo. Portanto, o resultado deste experimento foi a divisão geográfica do talhão em 3 grupos e é exibido na Figura 12. Este experimento foi necessário para observar os grupos gerados sem a influência da produção de soja, pois noutro experimento realizado foi observada a vizinhança geográfica com a influência da produção de soja e é discutido posteriormente.

Figura 12 - Três grupos considerando somente atributos georreferenciados.



Fonte: O autor.

Considerando o resultado deste experimento, uma análise comparativa dos grupos é apresentada na Tabela 3.

Tabela 3 - Experimento de 3 grupos utilizando atributos georreferenciados.

	Grupo 0	**Grupo 1**	**Grupo 2**
Produção de Soja mínima	606,98	239,97	136,52
Produção de Soja máxima	5142,84	7961,58	11196,29
Produção de Soja média	2296,11	2388,05	2410,33
Desvio padrão	407,48	582,21	580,60
Produção de soja registrada no centróide	2255,28	2151,66	2443,05
Quantidade de instâncias	389	962	1061
Correspondente em % das instâncias em relação ao total da base	16,13	39,88	43,99

Fonte: O autor.

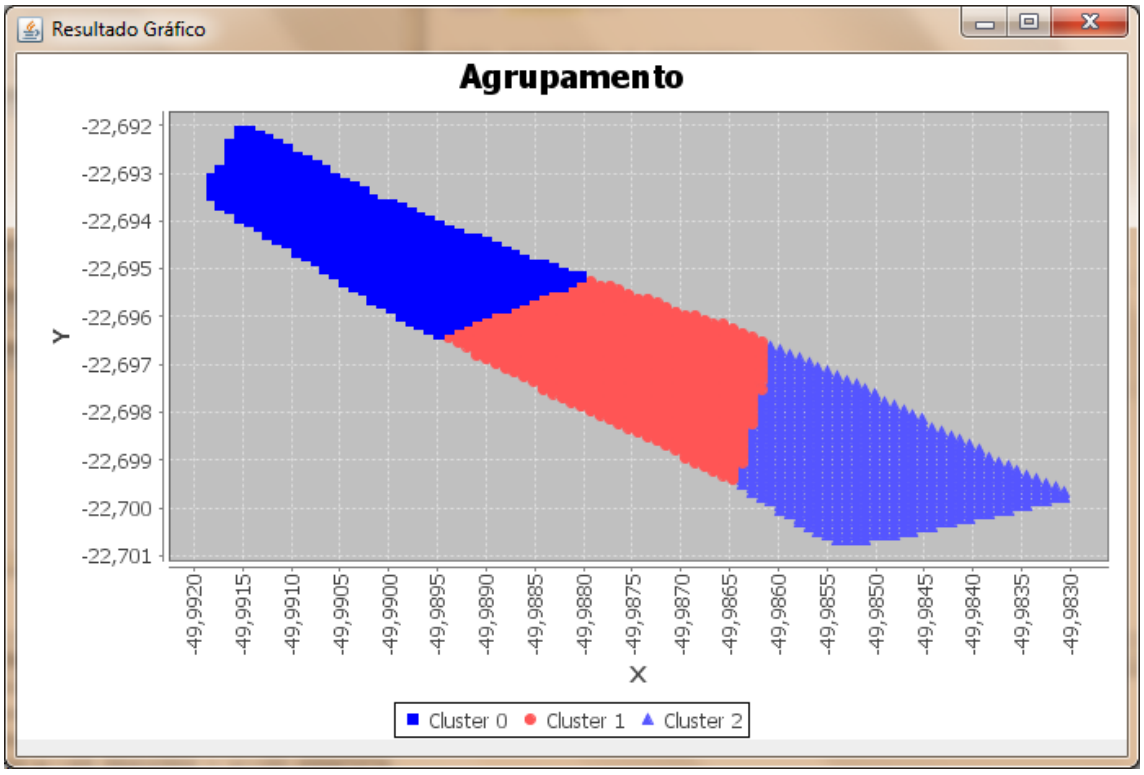
Observando a Tabela 3 é perceptível que dividir a propriedade em três grupos utilizando somente atributos georreferenciados não implica em encontrar uma produção

significativamente mais alta em um dos grupos. Analisando a produção mínima, a produção máxima, desvio padrão e média, percebe-se que a produção de cada pedaço é semelhante.

4.6 Experimento 3: atributos de latitude e longitude e a produção de soja

O experimento descrito deve verificar a influência geográfica na produção, considerando para tanto atributos georreferenciados e o atributo produção.

Figura 13 - Três grupos formados a partir de atributos georreferenciados e produção.



Fonte: O autor.

Para entendimento da Figura 13, é apresentada a Tabela 4 que possui informações de como os grupos foram formados considerando a produção e o local onde o ponto se encontra. Com a influência da produção de soja, os grupos ficaram melhor distribuídos em relação ao número de instâncias e em relação ao experimento que considerou somente os atributos de georreferenciamento. A produção de soja não possui diferença significativa entre os grupos, isto se deve ao fato do valor da produção “média” e “média a alta” ser distribuída em todo o talhão (Experimento 1).

Tabela 4 -Distribuição da produção quando considerada a posição geográfica e a produção.

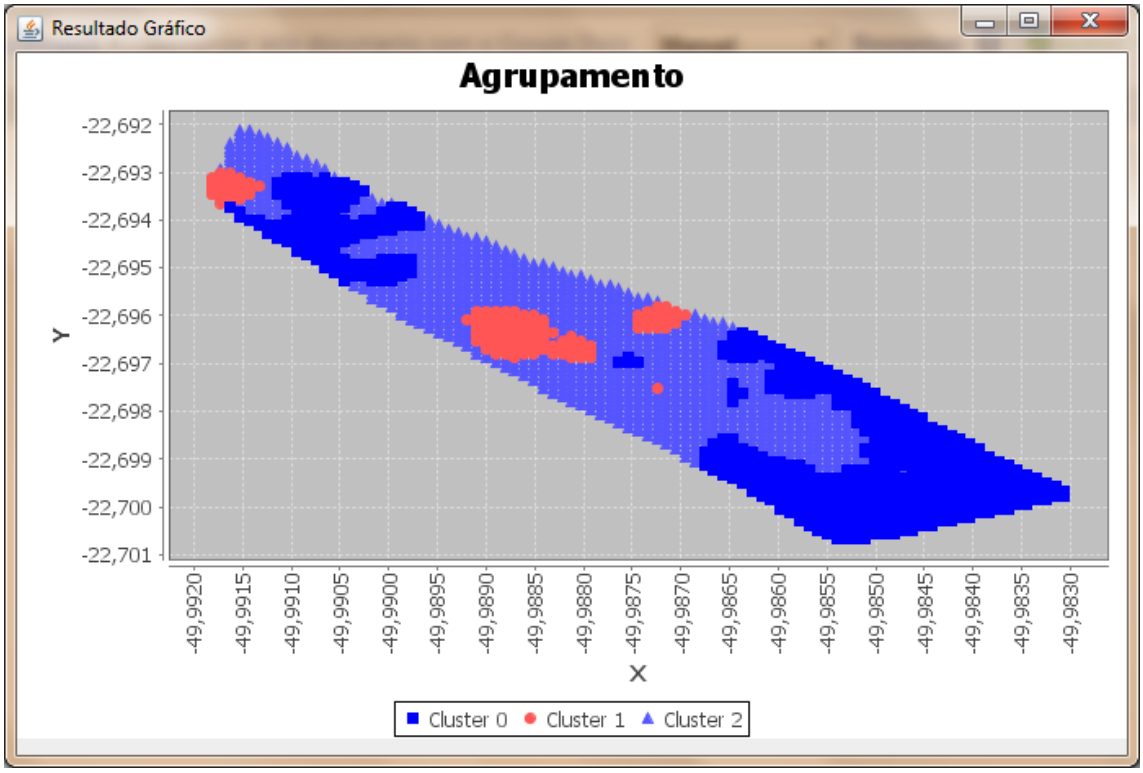
	Grupo 0	Grupo 1	Grupo 2
Produção de Soja mínima	604,52	136,52	239,97
Produção de Soja máxima	11196,29	7961,58	7817,78
Produção de Soja média	2368,33	2408,17	2370,66
Desvio padrão	617,08	516,11	543,99
Produção de soja registrada no centróide	2469,21	2440,03	2303,61
Quantidade de instâncias	735	841	836
Correspondente em % das instâncias em relação ao total da base	30,47	34,87	34,66

Fonte: O autor.

4.7 Experimento 4: características do solo (sem a produção de soja).

Para testar a similaridade dos atributos de características do solo, , sem considerar o georreferenciamento nem a produção, o experimento foi realizado para entendimento do comportamento da lavoura quanto às propriedades físico-químicas presentes na base.

Figura 14 - Agrupamentos considerando atributos com dados físico-químicos do solo.



Fonte: O autor.

Na Figura 14 é observada a similaridade dos grupos por meio da diversidade de cores. É percebida predominância da cor azul escuro na parte inferior direita da propriedade. Na região central, a predominância é da cor azul claro. Em blocos de menor concentração é percebida a cor vermelha.

Tabela 5 – Produção referente ao experimento com dados físico-químicos do solo.

	Grupo 0	**Grupo 1**	**Grupo 2**
Produção de Soja mínima	239,97	339,57	136,52
Produção de Soja máxima	7961,58	3503,96	11196,29
Produção de Soja média	2346,57	2238,97	2426,68
Desvio padrão	563,16	338,45	571,99
Produção de soja registrada no centróide	2250,63	2288,61	2404,99
Quantidade de instâncias	930	164	1318
Correspondente em % das instâncias em relação ao total da base	38,56	6,80	54,64

Fonte: O autor.

Na Tabela 5 é observada semelhança entre as produções encontradas nestes grupos. A diferença maior está na quantidade de instâncias que compõe esses grupos. O agrupamento que mais possui instâncias é o 2, seguido do grupo 0.

Para análise do agrupamento, foi utilizada a MANOVA (Análise de Variância Multivariada).

Tabela 6 - Resultado do Manova para o experimento 4.

	Test	**DF**			
Criterion	**Statistic**	**F**	**Num**	**Denom**	**P**
Wilk’s	0,29	255,12	16	4804	0,0
Lawley-Hotelling	2,20	330,15	16	4802	0,0
Pillai's	0,77	189,00	16	4806	0,0
Roy's	2,094				

Fonte: O autor.

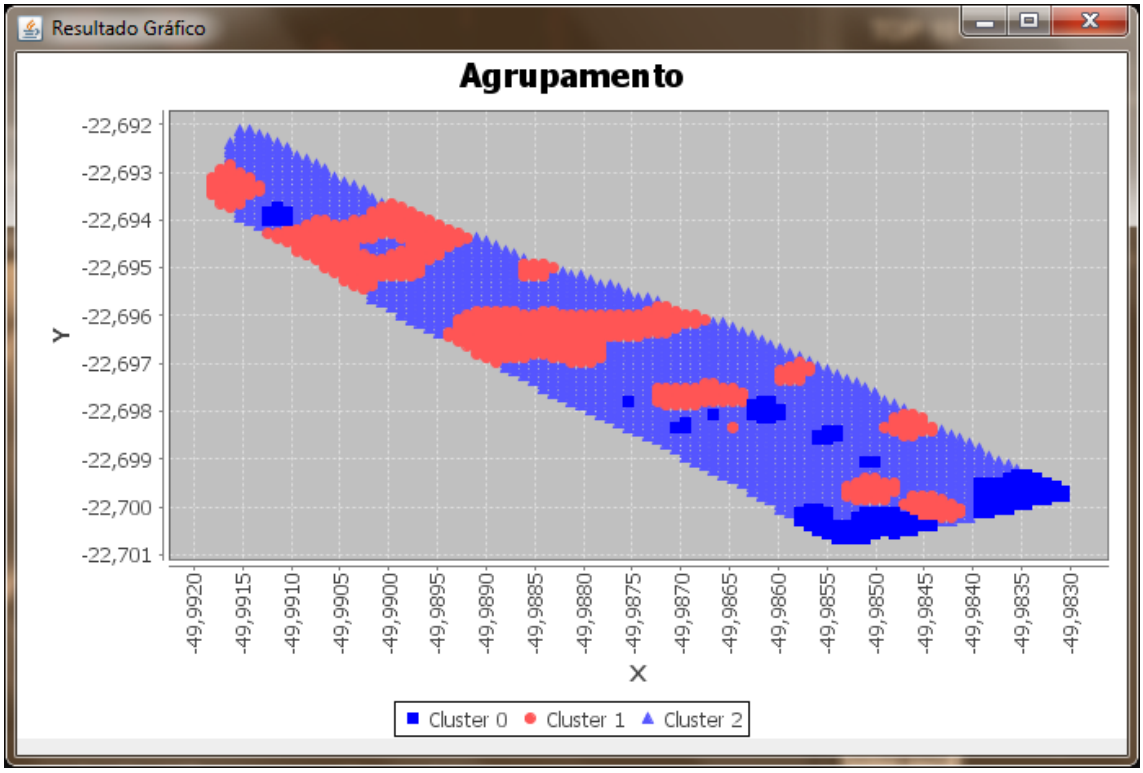
Na Tabela 6 estão os dados obtidos a partir de MANOVA para α=0,05. Neste teste estatístico, quanto menor o valor, maior a significância da diferença entre os grupos. Portanto, os valores deste teste indicam o quanto os grupos são diferentes, sendo compreendido o valor

desta diferença entre 0,0 e 1,0, sendo que um valor 0,0 significa diferença e para o valor 1,0 é tido como não tendo diferença entre os grupos. Então, é observado que no teste estatístico Wilk's o valor é de 0,29, o que implica em rejeitar a hipótese de média igual entre os grupos. Os três testes (Wilk's, Hotelling e Pillai's) indicam que o efeito multivariado é estatisticamente significante para o valor de P, que devem ser P's < ,001.

4.8 Experimento 5: três grupos com os índices de cone

Na base são encontrados os índices de cone medidas sempre de 5 em 5 centímetros: 0 a 5 cm, 5 a 10 cm, 10 a 15 cm, 15 a 20 cm, 20 a 25 cm, 25 a 30 cm, 30 a 35 cm e por último de 35 a 30 cm.

Figura 15 - Agrupamentos formados a partir dos atributos índice de cone.



Fonte: O autor.

Na Figura 15 é percebida a predominância da cor vermelha (grupo 1) e da cor azul claro (grupo 2). Azul (grupo 0) aparece numa parcela menor.

Tabela 7 - Dados referentes aos grupos formados a partir de índice de cone.

	Grupo 0	Grupo 1	Grupo 2
Produção de Soja mínima	239,97	136,52	604,52
Produção de Soja máxima	6473,36	7817,78	11196,29
Produção de Soja média	2155,07	2360,19	2416,27
Desvio padrão	747,04	488,23	555,08
Produção de soja registrada no centróide	3229,57	2337,25	2688,72
Quantidade de instâncias	170	638	1604
Correspondente em % das instâncias em relação ao total da base	7,05	26,45	66,50

Fonte: O autor.

A Tabela 7 mostra dados complementares ao experimento. É notável que a diferença na produção média é próxima, sendo que a produção média mais baixa é do grupo 0. Este mesmo agrupamento registra uma porcentagem baixa de instâncias e o maior desvio padrão.

Tabela 8 - MANOVA para o experimento com índices de cone.

	Test	DF			
Criterion	**Statistic**	**F**	**Num**	**Denom**	**P**
Wilk's	0,20	377,57	16	4804	0,0
Lawley-Hotelling	2,95	442,93	16	4802	0,0
Pillai's	1,03	317,94	16	4806	0,0
Roy's	2,49				

Fonte: O autor.

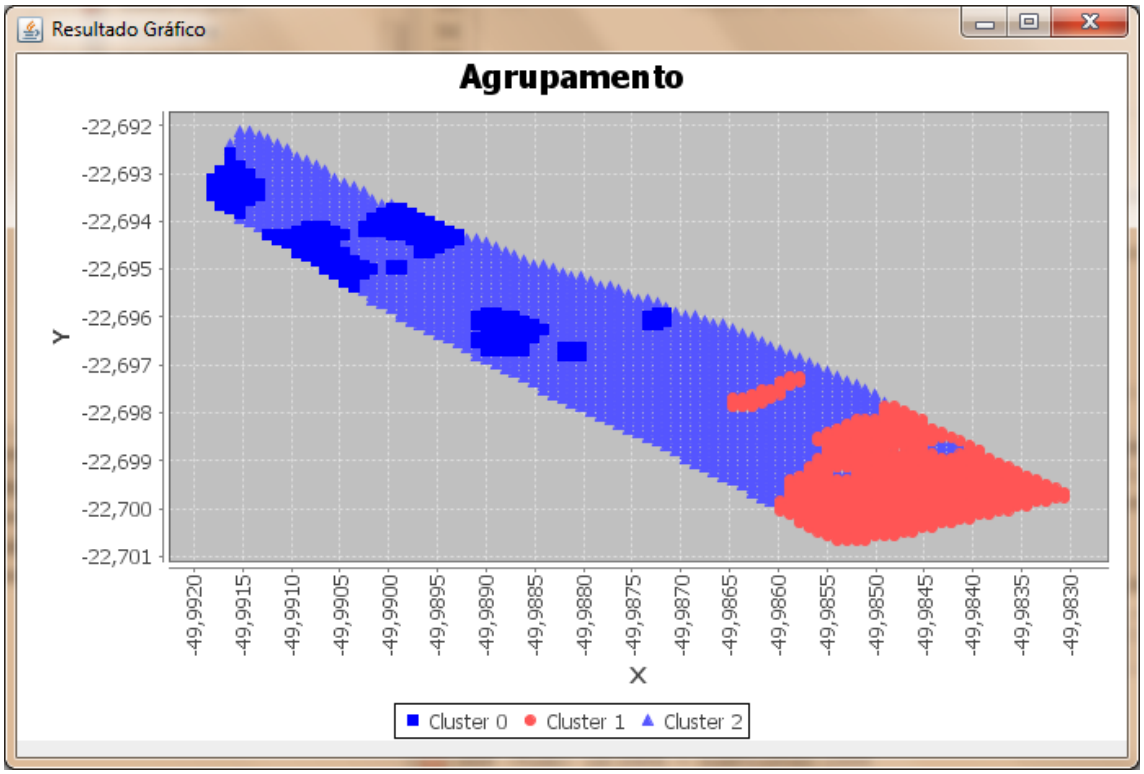
Na Tabela 8 é apresentado o resultado da aplicação de MANOVA para α=0,05. Por Wilk's é possível verificar que os grupos tem muita significância em seus grupos, pois 0,20 está mais para 0,0 apontando diferença do que 1,0 indicando igualdade entre grupos. O efeito multivariado é estatisticamente significante nos testes quando observado o valor de P sendo menor que 0,001.

4.9 Experimento 6: índices de cone e posição geográfica

O objetivo deste experimento é comparar se é possível obter produção diferenciada considerando a proximidade geográfica de pontos e o índice de cone. Também é objetivo

verificar uma estratégia para melhor conduzir a diferença entre os grupos obtidos pela influência do índice de cone na produção.

Figura 16 - Agrupamentos formados a partir de índice de cone e georreferenciamento.



Fonte: O autor.

Introduzindo dados georreferenciados no agrupamento com índice de cone força os grupos a serem formados também pela proximidade geográfica, como pode ser observado na Figura 16.

Quanto a produtividade desses grupos, foi percebida pouca diferença em relação a média e o desvio padrão. A maior diferença é a porcentagem de instâncias que acabou não refletindo numa qualidade de produção significativa.

Tabela 9 - Produção relativa ao experimento com índice de cone e posicionamento.

	Grupo 0	Grupo 1	Grupo 2
Produção de Soja mínima	648,68	239,97	136,52
Produção de Soja máxima	3985,45	7817,78	11196,29
Produção de Soja média	2266,22	2320,32	2424,89
Desvio padrão	386,78	618,04	555,04
Produção de soja registrada no centróide	2042,53	2334,95	2248,20
Quantidade de instâncias	262	568	1582
Correspondente em % das instâncias em relação ao total da base	10,86	23,55	65,59

Fonte: O autor.

Ainda sobre produtividade destes grupos, é apresenta a Tabela 9. Nesta Tabela percebe-se coerência entre a porcentagem de instâncias e porcentagem de produção.

O experimento mostrou que agrupar o índice de cone relacionado a proximidade espacial não produz diferença na produção. Também não é óbvia uma estratégia para se trabalhar diferença entre esses índices com relação ao local que eles se encontram.

Tabela 10 - MANOVA aplicada ao experimento com IC e georrefenciamento.

	Test	DF			
Criterion	**Statistic**	**F**	**Num**	**Denom**	**P**
Wilk's	0,38	186,58	16	4804	0,0
Lawley-Hotelling	1,27	190,12	16	4802	0,0
Pillai's	0,76	183,07	16	4806	0,0
Roy's	0,83				

Fonte: O autor.

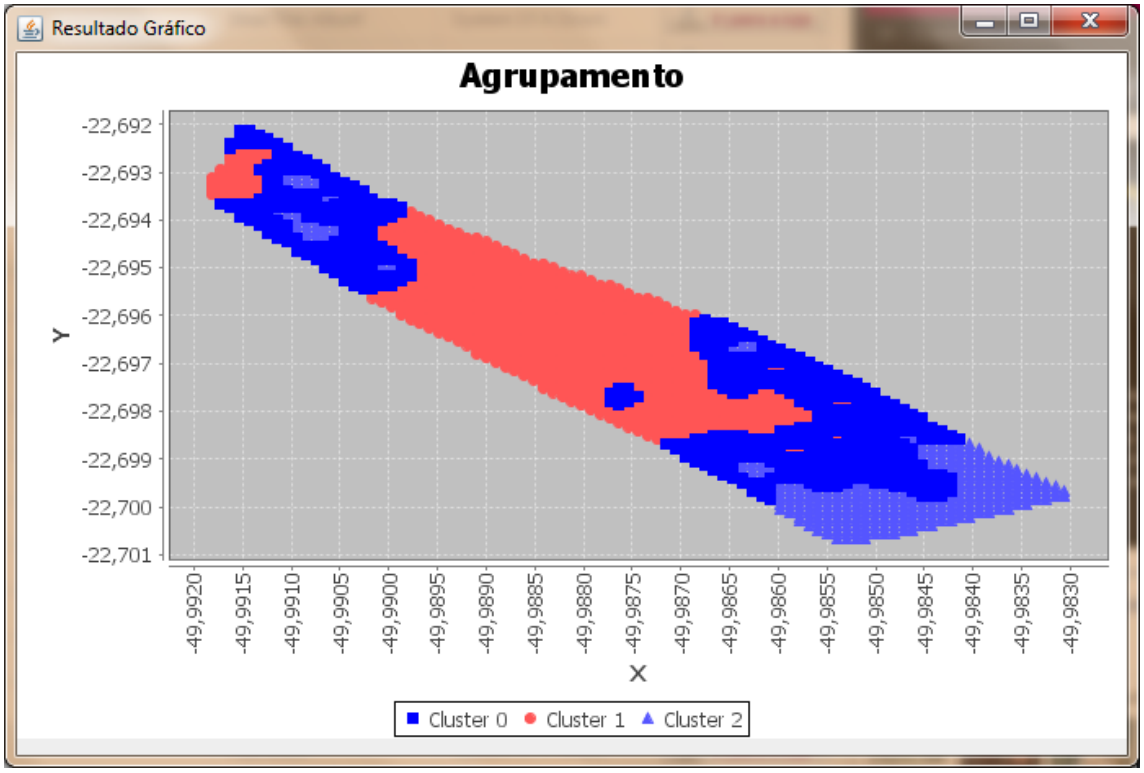
Na Tabela 10 é observado o resultado do teste estatístico MANOVA para α=0,05. Considerando que Wilk's deve ter um valor menor que um para ter diferença significante, os grupos tem diferença devido o valor ser 0,38. Comparando-se com o experimento anterior, percebe-se que a diferença diminuiu. Isto se deve ao fato de o experimento anterior não ter considerado os atributos latitude e longitude e, portanto, não utilizava-se da estratégia de agrupar características do IC considerando a proximidade geográfica. Os experimentos com IC apontaram que quando considerada uma região formada por características comuns, perde-se na diferença entre grupos. Quando considerado apenas o atributo sem atribuir peso a sua

posição geográfica há naturalmente grupos com maior diferença. Mesmo assim, os grupos tem significância segundo Wilk's e o valor de P sendo menor que 0,001.

4.10 Experimento 7: agrupamento de matéria orgânica

A matéria orgânica é um dos índices que medem a saúde do solo. O objetivo do experimento é verificar a saúde do solo baseado no valor da matéria orgânica e comparar com a produtividade encontrada nesses grupos.

Figura 17 - Resultado do agrupamento com MO sem georreferenciamento.



Fonte: O autor.

Na Figura 17 é visualizada a predominância do grupo 0 e do grupo 1, cores azul e vermelho respectivamente. Na Tabela 11, a produção média tem valor muito próximo comparando-se um agrupamento a outro.

Tabela 11 - Dados sobre produção considerando a MO somente.

	Grupo 0	Grupo 1	Grupo 2
Produção de Soja mínima	606,98	136,52	239,97
Produção de Soja máxima	7961,58	11196,29	7817,78
Produção de Soja média	2411,47	2397,81	2290,24
Desvio padrão	472,11	576,32	648,47
Produção de soja registrada no centróide	2425,66	2450,2	2253,41
Quantidade de instâncias	847	1126	439
Correspondente em % das instâncias em relação ao total da base	35,12	46,68	18,20

Fonte: O autor.

Tabela 12 - Dados referentes a MO nos grupos.

	Grupo 0	Grupo 1	Grupo 2
Valor mínimo	18,38	21,62	13,02
Valor máximo	21,61	37,20	18,37
Valor médio	20,02	24,18	17,24
Desvio padrão	0,96	2,28	0,93
Valor registrado no centróide	19,72	23,50	17,02

Fonte: O autor.

Na Tabela 12 é observada a quantidade de MO nos grupos. Em ordem crescente, a quantidade registrada no grupo 2,0 e 1. A maior quantidade de MO foi registrada no grupo 1, que possui cerca de 46% das instâncias da base. Mesmo assim, essa quantidade maior de MO não refletiu a maior produção. Estando próxima a produção, o grupo 0 registra a maior produção e o segundo maior valor de MO e é correspondente a 35% das instâncias.

Tabela 13 - ANOVA aplicado ao experimento com MO.

Fonte da variação	SQ	gl	MQ	F	valor-P	F crítico
Entre grupos	345324,29	1	345324,29	161384,10	0	3,84
Dentro dos grupos	10317,95	4822	2,14			
Total	355642,24	4823				

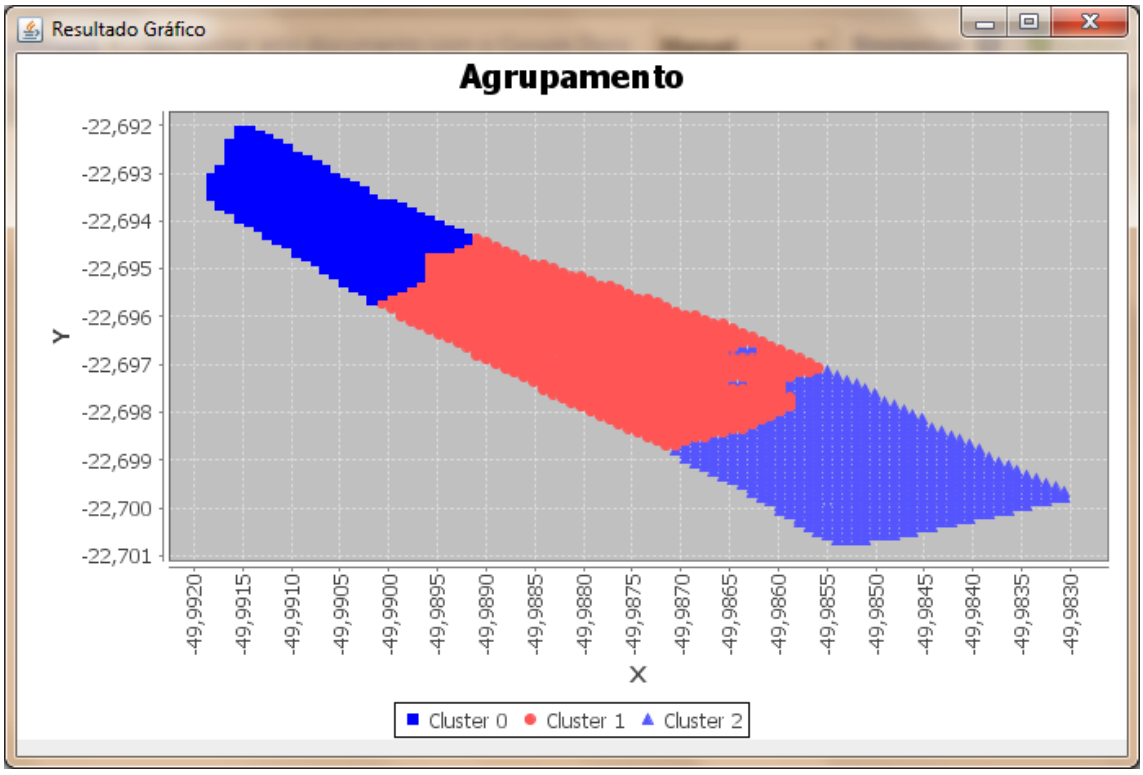
Fonte: O autor.

É possível visualizar o resultado de ANOVA para α = 5% na Tabela 13. Segundo a Tabela, MO tem diferença significante entre os grupos, pois o valor de F é maior que de F crítico. Há forte relação entre MO e seu grupo. O valor-P é menor que 5%, e assim é ignorada a hipótese de igualdade de grupos.

4.11 Experimento 8: agrupamento de MO e georreferenciamento

Com este experimento buscou-se agrupar a matéria orgânica pela proximidade geográfica. Apesar de MO não ser um valor mais alto em um trecho que outro, o valor médio ainda buscou separar a vizinhança em um valor menor, médio e maior.

Figura 18 - Resultado do agrupamento com MO e georreferenciamento.



Fonte: O autor.

A Figura 18 mostra uma separação onde o meio do talhão é predominantemente de uma faixa de MO. A separação utilizando os algoritmos propostos para esta base quando considerado o georreferenciamento, geralmente irá dividir o talhão em extremidades e parte central. A parcela de cada um se dá pelo atributo “produção”.

Tabela 14 - Produção considerando MO e georreferenciamento.

	Grupo 0	Grupo 1	Grupo 2
Produção de Soja mínima	604,52	136,52	239,97
Produção de Soja máxima	5142,84	11196,29	7961,58
Produção de Soja média	2307,27	2420,40	2379,75
Desvio padrão	427,61	566,86	607,88
Produção de soja registrada no centróide	2155,99	2206,84	2365,34
Quantidade de instâncias	492	1072	848
Correspondente em % das instâncias em relação ao total da base	20,40	44,44	35,16

Fonte: O autor.

Segundo a Tabela 14, a produção de soja não sofre influência significativa com essa distribuição geográfica da matéria orgânica.

Tabela 15 - Dados referentes a MO quando agrupados considerando o georreferenciamento.

	Grupo 0	Grupo 1	Grupo 2
Valor mínimo	16,07	17,07	13,02
Valor máximo	24,98	37,20	22,10
Valor médio	20,38	24,01	18,85
Desvio padrão	1,91	2,57	1,80
Valor registrado no centróide	20,39	25,11	16,36

Fonte: O autor.

Considerando o resultado da análise desse tipo de agrupamento na Tabela 15, é possível verificar que a vizinhança geográfica influenciou os valores de MO. Ainda assim, MO procurou agrupar entre valores mínimo-médio, médio e médio-máximo respectivamente nos grupos 2, 0 e 1. Nesta mesma ordem o valor médio registrado foi de 18,85; 20,38 e 24,01.

Tabela 16 - ANOVA para o experimento de MO e georreferenciamento.

Fonte da variação	SQ	gl	MQ	F	valor-P	F crítico
Entre grupos	332517,23	1	332517,23	154344,67	0,00	3,84
Dentro dos grupos	10388,43	4822	2,15			
Total	342905,65	4823				

Fonte: O autor.

Na Tabela 16 observamos o resultado de ANOVA para $\alpha = 5\%$ de significância. É observado que o valor de F é superior ao de F crítico. Há relação significativa entre MO e o grupo em que ele se encontra. Outra confirmação é o valor-P ser menor que 5%, implicando na diferença significativa.

4.12 Discussão

Os algoritmos GCLARA e GCLARANS tiveram resultados muito aproximados, devido a diferença de GCLARA trabalhar com cerca de 25%, e seu antecessor CLARA trabalhar com cerca de 10% das instâncias.

Quando analisada somente a distribuição da produção, sem o georreferenciamento, nota-se que a distribuição era uniforme no talhão. Exclusivamente as extremidades do talhão é que registraram valor reduzido.

Utilizando somente os atributos georreferenciados, o experimento mostrou que o algoritmo distribuiu a propriedade em três partes próximas do ideal que respeitavam a largura do talhão, com participação dos grupos de aproximadamente 16%, 40% e 44%. O resultado teve essa tendência devido ao algoritmo não considerar a quantidade de instâncias que devem compor cada agrupamento e sim buscar centróides mais representativos, ou seja, que sejam mais próximos do centro do grupo.

Como resultado da distribuição da produção de soja em três partes considerando a posição geográfica, o experimento mostrou que o experimento que o antecedeu teve uma mudança na distribuição geográfica. O grupo 0 teve um aumento na parcela do talhão passando a ter 30% de participação. O grupo 1 e 2 se aproximaram mais representando cada um 35%.

Considerando somente os atributos da base referentes as propriedades físico-químicas, portanto, sem considerar a produção nem a posição geográfica, o experimento mostrou semelhança nos atributos de forma diferente da registrada pelos experimentos que o antecederam. Considerando que a produção encontrada nesses grupos teve uma média aproximada. Concluiu se que não houve diferença significativa para a produção no agrupamento pelas propriedades físico-químicas.

No experimento utilizando índice de cone percebeu-se que o agrupamento que registrou a menor produção coincide com o experimento que considerou somente a produção. Este, de forma semelhante para a menor produção colocou agrupamento na extremidade do talhão. Observando os gráficos destes dois experimentos é possível verificar que na parte inferior direita registra grupos com produção de soja reduzida.

Considerando o georreferenciamento no agrupamento com índice de cone, o experimento 6 confirmou o observado no experimento 5, pois a menor produção média continuava sendo da extremidade inferior do talhão.

No experimento 7, com MO, o grupo com maior concentração era envolto pelo grupo com a segunda maior concentração de MO. A quantidade de instâncias nessas situações foi de respectivamente 47% e 35%. A menor concentração de MO está na extremidade inferior do talhão.

Comparando os experimentos 7 e 8, sem latitude e longitude e com latitude e longitude respectivamente, é percebida uma distribuição dos grupos diferentes. No experimento 8, o grupo responsável pela extremidade inferior do talhão faz interseção com os três agrupamentos obtidos no experimento 7. É percebido que o valor F é menor no experimento 8, indicando que houve perda de significância quando considerada a vizinhança geográfica ao agrupar o elemento MO.

Segundo os experimentos, a separação utilizando os algoritmos propostos para esta base quando considerado o georreferenciamento e três grupos, geralmente irá dividir o talhão em extremidades e parte central. A parcela de cada um se dá pelo(s) atributo(s) não georreferenciado(s). Porém, não é regra essa divisão, pois para atributos não georreferenciados que não tinham uma distribuição uniforme por todo o talhão acabaram sendo mapeados conforme sua semelhança e proximidade geográfica. A base estatística dá confiança aos experimentos, pois, aplicando ANOVA (experimentos com uma variável dependente) e MANOVA (experimentos com duas ou mais variáveis dependentes) confirmou-se diferença significativa entre os grupos.

5 CONCLUSÃO

A ferramenta DAGER permite o agrupamento de dados georreferenciados, a geração de resultados por texto e na forma de gráfico, conforme objetivo proposto. Os algoritmos PAM, CLARA e CLARANS foram implementados seguindo pseudo-código pré-existente. GCLARA e GCLARANS, que foram algoritmos propostos e implementados baseados em CLARA e CLARANS respectivamente, realizam a tarefa de agrupamento utilizando atributos diversos em conjunto com os atributos de determinação de posição geográfica (latitude e longitude) formando grupos de características próximas e geograficamente próximos.

Os experimentos mostraram que considerando os atributos com informações espaciais e atributos com características diversas ao realizar a tarefa de agrupamento, os grupos formados adquirem uma identidade geográfica incorporada a identidade formada pelos atributos de informação não geográfica. Em outras palavras, dentro de uma lavoura, os grupos encontrados são regiões geográficas com características semelhantes.

6 REFERÊNCIAS BIBLIOGRÁFICAS

BATCHELOR, B.; WHIGHAM, K.; DEWITT, J. **Precision agriculture: introduction to precision agriculture**. Iowa Cooperative Extension, 1997.

BIGOLIN, N. M. ; BOGORNY, V.; ALVARES, L. O. C.. Uma Linguagem de Consulta para Mineração de Dados em Banco de Dados Geográficos Orientado a Objetos. In: Conferencia Latinoamericana de Informatica, 29, 2003, La Paz. **Anais…** La Paz: CLEI, 2003. p. 23-35.

BLACKMORE, B. S.; WHEELER, P.N.; MORRIS, R.M. The role of precision farming in sustainable agriculture: a European perspective. In: International Conference on Site-Specific Management for Agricultural Systems, 2., 1994, Minneapolis, USA. **Anais…** Minneapolis, USA: American Soil Assoc, 1995.

BOGORNY, V. et al. Weka-GDPM: Integrating Classical Data Mining Toolkit to Geographic Information Systems. In: SBBD Workshop on Data Mining Algorithms and Applications (WAAMD'06), 2., 2006, Florianópolis, Brasil. **Anais…** Rio de Janeiro: SBC, 2006. p. 9-16.

BOGORNY, V. et al. Spatial Data Mining: From Theory to Practice with Free Software. In: WSL International Workshop on Free Software (WSL'07), 2007, Porto Alegre, Brasil. **Anais…** Porto Alegre: Nova Prova Gráfica e Editora, 2007. p. 205-212.

COELHO, A. M. Agricultura de Precisão: manejo da variabilidade espacial e temporal dos solos e culturas. **Documentos EMBRAPA Milho e Sorgo**, Sete Lagoas, 2005.

CORÁ, J. E et al. Variabilidade espacial de atributos do solo para adoção do sistema de agricultura de precisão na cultura de cana-de-açúcar. **Revista Brasileira de Ciência do Solo**, v.28, p.1013-1021, 2004.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J. Knowledge Discovery in Spatial Databases. In: German Conference on Artificial Intelligence, 23., 1999, Bonn, Germany. **Anais eletrônicos.** Bonn: ACM, 1999. Disponível em: <http://http://dl.acm.org/citation.cfm?id=731924&CFID=113916835&CFTOKEN=7251565>. Acesso em 12 mar. 2012.

ESTER, M. et al. Spatial data mining: a database approach. In: International Symposium on Large Spatial Databases (SSD ‘97), 5., 1997, Berlin, Germany. **Anais eletrônicos.** Berlin: Springer, 1997. p. 47-66.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J. Algorithms and Applications for Spatial Data Mining. **Geographic Data Mining and Knowledge Discovery**, Nova York: Taylor & Francis, 2001. p. 160-187.

FARIA, G. **Um Banco de dados espaço-temporal para desenvolvimento de aplicações em Sistemas de Informação Geográfica.** 1998. Dissertação (Mestrado em Ciência da Computação). Instituto de Computação, Universidade Estadual de Campinas, Campinas, São Paulo, 1998. Disponível em: < http://www.bibliotecadigital.unicamp.br/document/?code=000126613&opt=1>. Acesso em: 8 mai. 2012.

GENTIL, L.V.; FERREIRA, S.M. Agricultura de precisão: Prepare-se para o futuro, mas com os pés no chão. **Revista A Granja**, Porto Alegre, n. 610, p. 12-17, 1999.

GUIMARÃES, A. M. **Aplicação de Computação Evolucionária na Mineração de Dados físico-químicos da água e do Solo**. 2005. 127 f. Tese (Doutorado em Agronomia / Energia na Agricultura) - Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu, 2005.

GUIMARÃES, A. M. et al. Data mining na agricultura de precisão: uma abordagem comparativa. In: Congresso Brasileiro da Sociedade Brasileira de Informática Aplicada a Agropecuária e à Agricultura, 4., 2003, Porto Seguro, Brasil. **Anais eletrônicos.** Porto Seguro, 2003. Disponível em: <http://www.sbiagro.org.br/pdf/iv_congresso/art008.pdf>. Acesso em: 3 mar. 2012.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining:** Concepts and Techniques. 3. ed. Waltham: Elsevier, 2011. 744 p.

KAUFFMAN, L.; ROUSSEEUW, P.J. **Finding Groups in Data:** an Introduction to Cluster Analysis. 99. ed. Hoboken: John Wiley & Sons, 1990. 342 p.

KEIM, D. A.; KRIEGEL, H.-P. Visualization Techniques for Mining Large Databases: A Comparison. **IEEE Transactions on Knowledge and Data Engineering**, Los Alamitos, v. 8, n. 6, p. 923-938, dec. 1996.

KOPERSKI, K., HAN, J. GEOMINER: A System Prototype for Spatial Mining, In: ACM SIGMOD International Conference on Management of Data. 1997, Tucson, Arizona, USA. **Anais...** New York: ACM, 1997. p. 553-556.

MILLER, H. J. Geographic Data Mining and Knowledge Discovery. **The Handbook of Geographic Information Science**, Oxford, UK, 2008.

MOLIN, J. P. Tendências da Agricultura de Precisão no Brasil. In: Congresso Brasileiro de Agricultura de Precisão, 2004, São Paulo. **Anais eletrônicos.** Piracicaba: ESALQ/USP, 2004. Disponível em: <HTTP://www.leb.esalq.usp.br/download/TEC%202004.12.pdf>. Acesso em 30. mar. 2012.

MOLIN, J. P. et al. Utilização de dados georreferenciados na determinação de parâmetros de desempenho em colheita mecanizada. **Engenharia Agrícola**, Jaboticabal, v. 26, n. 3, p. 759-767, set./dez. 2006.

MORAES, A. C. et al. Sistema de Informações Geográficas: Uma Ferramenta para Gestão de Pesquisa Agrícola. In: Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 15, 2011, Curitiba, PR, Brasil. **Anais...** São José dos Campos: INPE, 2011. p. 8704-8708.

MOTOMIYA, A. V. DE A. et al. Variabilidade espacial de atributos químicos do solo e produtividade do algodoeiro. **Agrarian**, Dourados, v.4, n.11, p.01-09, 2011.

MUCHERINO, A.; PAPAJORGJI, P. J.; PARDALOS, P. M. **Data mining in agriculture**. New York: Springer. 2009. 274 p.

NEVES, M. C. **Procedimentos Eficientes Para Regionalização De Unidades Socioeconômicas Em Bancos De Dados Geográficos.** 2003. 160 f. Tese (Doutorado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais. 2002.

NG, R.; HAN, J. CLARANS: A Method for Clustering Objects for Spatial Data Mining, **IEEE Transactions on Knowledge and Data Engineering**. v. 14, n. 5, p. 1003-1016. set./out. 2002.

PIVATO, M. A. **Mineração de regras de associação em dados georreferenciados.** 2006. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo, São Carlos, 2006.

RABELO, E. **Avaliação de técnicas de visualização para mineração de dados**. Dissertação (Mestrado em Ciência da Computação). Universidade Estadual de Maringá, Maringá, 2007.

ROLOFF, G.; FOCHT, D. Brazil. **Handbook of precision agriculture: principles and applications**. Binghampton, p. 635-656, 2006.

SANDER J. et al. Density-Based Clustering in Spatial Databases: A New Algorithm and its Applications. **Data Mining and Knowledge Discovery,** Kluwer Academic Publishers, v.2, n. 2. 1998

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

SFERRA H. H.; CORRÊA A. M. C. J. Conceitos e Aplicações de Data Mining. **Revista de Ciência & Tecnologia.** v.11, n. 22, p.19-34, jul. 2003.

SOUZA, Z. M. et al.. Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. **Ciência Rural**, Santa Maria, v. 40, n. 4, abr. 2010.

WITTEN, I. H.; FRANK. E. **Data mining:** practical machine learning tools and techniques. 2. ed. Burlington: Morgan Kaufmann, 2005. 525 p.