UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE DADOS ESPACIAIS APLICADA NO DELINEAMENTO DE UNIDADES DE GESTÃO DIFERENCIADA EM AGRICULTURA DE PRECISÃO

EDUARDO ANTONIO SPERANZA

ORIENTADOR: PROF. DR. RICARDO RODRIGUES CIFERRI

São Carlos – SP

Setembro/2017

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE DADOS ESPACIAIS APLICADA NO DELINEAMENTO DE UNIDADES DE GESTÃO DIFERENCIADA EM AGRICULTURA DE PRECISÃO

EDUARDO ANTONIO SPERANZA

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação, área de concentração: Engenharia de Software, Banco de Dados e Interação Humano-Computador.
Orientador: Prof. Dr. Ricardo Rodrigues Ciferri

São Carlos – SP

Setembro/2017

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato Eduardo Antonio Speranza, realizada em 12/09/2017.

Prof. Dr. Ricardo Rodrigues Ciferri
UFSCar

Prof. Dr. Ricardo Cerri
UFSCar

****

Prof. Dr. Estevam Rafael Hruschka Junior
UFSCar

Prof. Dr. Ricardo Yassushi Inamasu
EMBRAPA

Prof. Dr. José Paulo Molin
USP

Certifico que a defesa realizou-se com a participação à distância do membro Estevam Rafael Hruschka Junior, depois das arguições e deliberações realizadas, o participante à distância está de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Prof. Dr. Ricardo Rodrigues Ciferri (presidente)

Para Tati, companheira para toda a vida.

Para Pedro, que tornou a vida do papai "mais azul".

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus por me dar saúde e por guiar todas as decisões tomadas em minha vida.

Agradeço as seguintes instituições envolvidas no desenvolvimento desta tese, das quais o suporte administrativo, financeiro e técnico foi essencial para que o trabalho pudesse ser realizado: Empresa Brasileira de Pesquisa Agropecuária (Embrapa), pela oportunidade de participar do programa interno de pós-graduação, possibilitando a minha dedicação exclusiva ao curso de doutorado; e Universidade Federal de São Carlos (UFSCar), pelo apoio administrativo e técnico proporcionado por seus funcionários e docentes.

Sou extremante grato ao prof. Ricardo Ciferri, por aceitar o desafio de orientar um tese em Ciência da Computação aplicada em uma área tão complexa como a Agricultura; e à profa. Cristina Ciferri, pelo auxílio na orientação e pelas oportunidades de divulgação do tema Agricultura de Precisão no meio acadêmico.

Agradeço também a todos os colegas envolvidos na Rede AP, que de alguma maneira auxiliaram neste trabalho: Célia, Cristina, Vicente, Bassoi, Álvaro, Marina, Luchiari, João Camargo, Alberto, Ricardo, prof. Molin e André Torre, que me "apresentou"a Embrapa e o tema Agricultura de Precisão; aos colegas do LabGeo da Embrapa Informática, pelo incentivo; e em especial, aos colegas do GBD/UFSCar, com quem compartilhei os desafios durante os últimos anos.

Finalizo agradecendo imensamente a toda minha família. Aos meus pais, Toninho e Maria, pela educação sólida na infância e adolescência, contribuindo para a formação de uma pessoa pró-ativa tanto na vida profissional quanto pessoal, sem deixar de lado a humildade e a honestidade. Agradeço, em especial, à minha amada esposa Tatiane, pela paciência e dedicação dispensada a mim durante todos esses anos, e ao nosso amado filho Pedro Henrique, enviado por Deus para que eu tivesse a oportunidade de ser pai.

RESUMO

A Agricultura de Precisão (AP) é uma estratégia de cultivo agrícola que se utiliza de tecnologias e princípios para gerenciar a variabilidade espacial e temporal relacionada a todos os aspectos que envolvem uma lavoura, com o objetivo de aumentar a produtividade de maneira sustentável, possibilitando tanto a redução dos impactos ao meio ambiente quanto o aumento do retorno econômico. Um dos processos utilizados por essa estratégia para atingir esse objetivo é o delineamento da área de cultivo em parcelas menores com características similares, conhecidas como unidades de gestão diferenciada (UGDs). Para que esse processo possa ser executado com êxito, as dissimilaridades entre as UGDs devem ser corretamente identificadas a partir de dados espaciais. Desse modo, o delineamento de UGDs pode ser considerado como um processo orientado ao agrupamento de dados espaciais, no qual uma solução de agrupamento corresponde a um mapa de UGDs. As abordagens computacionais existentes na literatura para viabilizar a automatização desse processo, normalmente baseadas em algoritmos de agrupamento *fuzzy*, não consideram, em sua maior parte, as coordenadas geográficas que compõem as amostras coletadas durante a execução dos seus métodos, o que pode tornar os mapas de UGDs excessivamente estratificados. Assim, é possível observar a falta de uma abordagem de consenso que permita a obtenção de mapas de UGDs com variabilidade interna mínima e que, ao mesmo tempo, sejam facilmente interpretados pelos usuários especialistas. Diante do exposto, o processo para delineamento de UGDs em AP é abordado nesta tese, cujas principais contribuições são: (i) o critério de validação interna *SD-Spatial*, que considera questões relativas à coesão e separação de grupos tanto no espaço de atributos quanto no espaço de coordenadas; (ii) a abordagem de agrupamento espacial *SWMU Clustering*, que explora de maneira ponderada as dissimilaridades dos atributos convencionais e espaciais, utilizando-se de parâmetros fornecidos pelo usuário especialista sem que o determinismo das soluções seja prejudicado; e (iii) a abordagem complementar *SWMU Polygon*, que permite representar os mapas de UGDs em formato poligonal. Com base nos experimentos, a abordagem *SWMU Clustering* apresentou ganhos médios de 31,94% em medida de validação considerando tanto o espaço de atributos quanto o espaço de coordenadas, com relação a abordagens que utilizam agrupamento *fuzzy*; e a abordagem complementar *SWMU Polygon* proporcionou ganhos médios de desempenho de 61,14% na recuperação de mapas de UGDs armazenados em bancos de dados espaciais.

Palavras-chave: Mineração de Dados Espaciais, Agrupamento Espacial, Agricultura de Precisão, Unidades de Gestão Diferenciada, Zonas de Manejo

ABSTRACT

Precision Agriculture (PA) is an agricultural cultivation strategy which uses technologies and principles to manage the spatial and temporal variability related to all aspects that surround a crop, in order to increase yield in a sustainable way, enabling both the reduction of environmental risks and the increase of profits. One of the processes used by this strategy to achieve this goal is the delineation of the crop area in smaller plots with similar characteristics, known as differentiated management units (DMUs). In order to achieve success in this process, dissimilarities between the DMUs must be properly identified from spatial data collected through field or remote sensors. Therefore, the delineation of DMUs may be considered as a spatial data clustering oriented process, in which a clustering solution corresponds to a map of DMUs. The computational approaches found in literature in order to assist in automating this process, usually based in fuzzy clustering algorithms, do not consider, for the most part, the geographic coordinates which compose the collected samples during their methods execution, which can make the DMU maps overly stratified. Therefore, it is possible to observe the lack of a consensus approach in the literature which may allow expert users to obtain DMU maps with minimum internal variability and which, at the same time, are easily interpreted by expert users. Given the above, the process for the delineation of DMUs in PA is discussed in this thesis, whose main contributions are: (i) the SD-Spatial internal validation criteria, which considers issues related to clusters cohesion and separation both in the attribute space and in the coordinate space; (ii) the SWMU Clustering spatial clustering approach, which explores in a weighted way the dissimilarities of the conventional and spatial attributes, using parameters provided by the expert user without the determinism of the solutions being impaired; and (iii) the complementary approach SWMU Polygon, which allows to represent the DMU maps in polygonal shape. Based on the experiments, the SWMU Clustering approach presented average gains of 31.94% in the validation measure considering both the attribute space and the coordinate space, in comparison to approaches using fuzzy clustering; and the complementary approach SWMU Polygon provided average performance gains of 61.14% in the retrieval of DMU maps stored in spatial databases.

Keywords: Spatial Data Mining, Spatial Clustering, Precision Agriculture, Differentiated Management Units, Management Zones

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Contextualização

Dados espaciais são utilizados para descrever fenômenos que possuem uma dimensão espacial associada, representada por um sistema de coordenadas (ARONOFF, 1989). Ou seja, um dado espacial possui uma localização no espaço, um formato espacial bem definido e fronteiras que permitem a distinção do seu interior e do seu exterior. Esses dados também são conhecidos como dados espaciais *crisp*. Quando a dimensão espacial se refere ao posicionamento de um fenômeno que ocorre no globo terrestre e no seu espaço próximo, esse conceito é particularizado e conhecido como dado geoespacial, dado georreferenciado ou dado geográfico (LONGLEY, 2005). Segundo Ciferri (1995), um dado espacial é constituído por um conjunto de coordenadas que descreve a geometria de um objeto espacial. Essas coordenadas se referem a um sistema de projeção cartográfica, e podem representar objetos espaciais com forma de ponto, linha ou polígono, ou seja, podem representar objetos no formato vetorial. Para facilitar o entendimento e a utilização dos termos, dados geoespaciais, dados georreferenciados e dados geográficos serão tratados nesta tese, de uma maneira mais geral, como sinônimos de dados espaciais, assumindo que estes podem possuir coordenadas obtidas a partir de qualquer plano do espaço Euclidiano bidimensional.

Nos anos 90, com o surgimento da internet, a disponibilização gratuita do sinal de GPS[1] e a emergência da sociedade da informação, a importância da ciência geográfica passou a ser amplamente revelada, criando-se novas perspectivas para o seu desenvolvimento (JULIÃO, 1999). A partir desse momento, iniciou-se um apoio cada vez maior da Tecnologia da Informação (TI) na produção de bases de dados georreferenciadas, permitindo que estas bases fossem visualizadas e interpretadas por Sistemas de Informações Geográficas (SIGs) (BURROUGH, 1986;

[1] Do inglês *Global Positioning System*

ARONOFF, 1989).

Devido à grande quantidade de dados georreferenciados que vem sendo produzida ao longo dos anos, seja por empresas privadas, órgãos governamentais, universidades ou instituições de pesquisa, é natural o surgimento de ferramentas computacionais capazes de manipulá-los. Além dos SIGs e das extensões espaciais para os já conhecidos Sistemas Gerenciadores de Bancos de Dados (SGBDs) (QUEIROZ; FERREIRA, 2005) - que proporcionam armazenamento, recuperação e manipulação eficiente de dados espaciais - surgiram também modelos específicos para bancos de dados que tratam esse tipo de dado, baseados no formalismo de entidade e relacionamento ou na orientação a objetos (RIBEIRO-JÚNIOR, 2007). Mais adiante, com o objetivo de proporcionar a integração e a organização dos dados espaciais produzidos por diferentes entidades, surgiram os padrões e perfis de metadados geoespaciais (FGDC, 1998; ISO, 2003; BRASIL, 2009).

Acompanhando a evolução desse cenário, diversas aplicações nas mais diferentes áreas surgiram ao longo do tempo, com o objetivo de aproveitar a grande massa de dados espaciais produzida. Dentre essas áreas, destaca-se a Agricultura de Precisão (AP). A AP é definida como um sistema de gerenciamento agrícola baseado na variabilidade espacial das propriedades do solo e das plantas de uma lavoura, objetivando o aumento da produtividade e a economia de insumos agrícolas a partir de ações sustentáveis de proteção ao meio ambiente (BERNARDI, 2014; MOLIN; AMARAL; COLAÇO, 2015). Para que esses objetivos sejam atingidos, é necessário o uso de um conjunto de tecnologias e procedimentos relacionados à otimização das culturas com base na variabilidade espacial de sua produção (MOLIN, 2003). Enquanto na agricultura tradicional a aplicação de fertilizantes e corretivos em uma cultura é realizada de maneira equivalente para toda a área de cultivo - podendo causar o desperdício ou a escassez desses insumos em alguns locais, a falta de incentivo para o aumento da produtividade em áreas com alto potencial e a depreciação desnecessária do meio ambiente -, a AP permite ao seu usuário final gerenciar racionalmente os insumos, aplicando-os de maneira espacialmente diferenciada. Para tanto, esse sistema possui um ciclo de vida muito bem definido no que diz respeito às questões agronômicas, onde cada safra deve possuir as fases de preparação do solo, plantio, acompanhamento da lavoura e colheita. Os principais conceitos e aplicações da AP utilizados no âmbito desta tese estão descritos no Capítulo 2.

A exemplo do que ocorre em outras áreas, a AP evoluiu muito com relação ao desenvolvimento de tecnologias de informação e comunicação (TICs) capazes de coletar, compartilhar e armazenar dados espaciais com alta acurácia. Dentre elas, destacam-se os sensores para medição do comportamento do solo e das plantas em uma lavoura, e câmeras para obtenção de imagens a partir de satélites ou aeronaves em diversas bandas representativas com resolução

espacial na casa dos centímetros. Desse modo, novos conceitos relacionados às TICs, como a Internet das Coisas (*IoT*[2]), *Big Data*, computação em nuvem e mineração de dados têm entrado no escopo das pesquisas mais recentes em AP (LI; CHUNG, 2015). Consequentemente, novos avanços são necessários para a construção de ferramentas que permitam o armazenamento, a recuperação e, principalmente, o uso eficaz dos dados levando em consideração esses novos conceitos. Nesse contexto, alguns processos específicos, como a subdivisão de um talhão[3] em parcelas menores com variabilidade interna suficientemente pequena, conhecidas como unidades de gestão diferenciada (UGDs), podem ser muito facilitadas com a construção de ferramentas computacionais capazes de obter o conhecimento intrínseco aos dados de maneira eficaz, considerando os devidos tratamentos que devem ser realizados para dados de natureza complexa, como os dados espaciais.

1.2 Objetivo

Esta tese de doutorado propõe a criação de uma abordagem computacional capaz de auxiliar na obtenção do conhecimento intrínseco aos dados espaciais obtidos em campo agrícola, por meio de técnicas eficazes de mineração de dados espaciais. Seu principal objetivo foi produzir uma nova abordagem para a tarefa de agrupamento de dados espaciais no formado vetorial, levando em consideração as dissimilaridades entre amostras e considerando tanto o espaço de atributos quanto o espaço de coordenadas (dimensão espacial dos dados), para que sejam obtidos mapas de UGDs menos estratificados e de mais fácil interpretação por parte do usuário final. Resumidamente, os objetivos específicos a serem atingidos com esta tese foram:

- O tratamento de questões relacionadas à influência do usuário final na determinação de parâmetros, permitindo: a identificação dos diferentes atributos de solo e planta a serem utilizados nessa tarefa; a validação qualitativa para escolha da melhor solução dentre um conjunto de soluções disponíveis; e a redução da estratificação dos mapas de UGDs, para que possam ser melhor interpretados.

- O desenvolvimento de soluções computacionais que permitam o tratamento diferenciado do espaço de atributos com relação ao espaço de coordenadas, a partir de ações como: a utilização de restrições espaciais; representação de mapas de UGDs em formato poligonal; e a criação de um sistema protótipo para facilitar o uso de algoritmos e abordagens

[2] Do inglês *Internet of Things*

[3] Representação da divisão real ou imaginária de uma propriedade rural, utilizada para facilitar o gerenciamento da produção.

utilizadas durante todas as etapas do processo de descoberta de conhecimento em bancos de dados espaciais que suportam a tarefa de delineamento de UGDs em AP.

1.3 Motivação

O grande volume de dados espaciais que vêm sendo coletado nos últimos anos, seja a partir de sensores providos de GPS, imagens de sensoriamento remoto, ou até mesmo a partir de informações voluntárias disponíveis na internet (SILVA, 2014) tende a aumentar a dificuldade dos usuários em analisá-los, mesmo que estes sejam especialistas do domínio em questão.

As técnicas tradicionais de mineração de dados utilizadas na tarefa de agrupamento de amostras sem classificação prévia não foram desenvolvidas considerando as peculiaridades de dados com natureza complexa, como é o caso dos dados espaciais. Essas técnicas visam minimizar o erro a partir de funções objetivo que consideram medidas de distância métrica para verificar a dissimilaridade entre as amostras, calculadas com base em todos os atributos que as compõem, buscando sempre a obtenção de grupos coesos e bem separados. No entanto, quando o foco é o agrupamento de dados espaciais, os atributos referentes às coordenadas devem ser tratados de maneira diferenciada com relação aos atributos convencionais (e.g., inteiro, decimal), evitando, por exemplo, que apenas amostras próximas espacialmente sejam associadas a um mesmo grupo, ou que os grupos obtidos se tornem demasiadamente esparsos e sem significado útil ao usuário final quando visualizados em um SIG. Assim, além da distância métrica, relacionamentos puramente espaciais como adjacência, união, subtração e intersecção devem ter influência na dissimilaridade entre amostras de dados espaciais.

Diversas pesquisas em agrupamento de dados vêm sendo realizadas na tentativa de se obter soluções capazes de tratar o espaço de atributos de maneira distinta com relação ao espaço de coordenadas (ANANDHI; SUBRAMANYAM, 2009; FAN, 2009; WANG; BU, 2010; JAYAKUMARAN; KARUPPANAN, 2013). Algumas soluções propostas realizam esse tratamento gerando agrupamentos com base em densidade de pontos (ZHONG; LIU; LI, 2010; WANG; WANG, 2011), e adicionalmente encontram grupos de maneira hierárquica (GUO; PEUQUET; GAHEGAN, 2003; CHOO, 2007) ou os representam em forma de polígonos (AKDAG; EICK; CHEN, 2014). Por fim, existem também algumas soluções para agrupamento de dados espaciais com representação poligonal nativa (ASSUNCÃO, 2006; GUO, 2008; JOSHI, 2011). Entretanto, essas abordagens, de uma maneira geral, são baseadas na densidade amostral no espaço de coordenadas, o que inviabiliza a sua utilização para agrupar amostras que são interpoladas em grade regular, como normalmente acontece com os conjuntos de dados obtidos em AP.

Por outro lado, as abordagens mais utilizadas para o processo de delineamento de UGDs simplesmente não consideram o espaço de coordenadas para o agrupamento dos dados (KITCHEN, 2005; LI, 2007; MORARI; CASTRIGNANÒ; PAGLIARIN, 2009; SONG, 2009; XIN-ZHONG, 2009). Apesar disso, essas abordagens utilizam algoritmos *fuzzy* capazes de agrupar dados considerando incertezas, permitindo que uma amostra possa pertencer a diferentes grupos com certo grau de pertinência (BEZDEK; EHRLICH; FULL, 1984). A justificativa para o uso desse tipo de algoritmo está relacionada com a vagueza que é natural em dados coletados em campo, por conta da imprecisão que pode ocorrer tanto nas medidas obtidas pelos sensores, quanto nas coordenadas geográficas capturadas. Além disso, esses dados devem ter sua continuidade considerada, pois é fato que mudanças bruscas nos atributos do solo e da cultura não ocorrem com frequência em pequenos espaços da lavoura. De maneira complementar, a utilização de algoritmos de agrupamento *fuzzy* também pode ser justificada pela dificuldade em se conhecer, *a priori*, a distribuição probabilística dos atributos utilizados, sejam eles espaciais ou convencionais. Desse modo, somos induzidos a pensar que os limites das UGDs não devem ser obtidos de maneira rígida, o que resultaria na associação de cada amostra a diferentes unidades, considerando certo grau de pertinência para cada uma delas. Entretanto, se analisarmos em termos práticos, a definição precisa de onde termina uma UGD e onde começa outra é necessária para dar suporte às operações de campo e, consequentemente, proporcionar uma gestão mais apropriada da lavoura. Nesse sentido, todas essas abordagens realizam um procedimento conhecido como *defuzzificação*, fazendo com que cada amostra seja associada de fato a uma única UGD. Mesmo que esse processo seja necessário, essas abordagens acabam não utilizando por completo os benefícios oferecidos pela matriz de pertinências fornecida por algoritmos de agrupamentos *fuzzy*.

Em trabalhos mais recentes, a preocupação com a utilização correta dos atributos relacionados à localização espacial tem se tornado mais evidente. O trabalho desenvolvido por Ruß e Kruse (2011) propõe uma abordagem hierárquica aglomerativa para agrupar dados espaciais obtidos em agricultura de precisão com a ideia básica de manter, o quanto for possível, a contiguidade espacial do agrupamento gerado, permitindo, por meio de um parâmetro fornecido pelo usuário final, que apenas ao final de sua execução sejam geradas UGDs extremamente similares considerando o espaço de atributos, mas não adjacentes com relação ao espaço de coordenadas. Já as soluções propostas por Córdoba (2013) e Peeters (2015), realizam o tratamento da informação espacial vinculada aos atributos convencionais antes da execução do algoritmo de agrupamento, utilizando, respectivamente, análise de componentes principais e índices de correlação espacial.

Outra questão importante discutida na literatura é relacionada à quantidade ótima de UGDs

para uma determinada área agrícola, que não deve ser demasiadamente dividida de forma a prejudicar determinadas atividades em campo (MOLIN; AMARAL; COLAÇO, 2015). Nesse momento, é importante salientar que estamos tratando UGDs como áreas espacialmente contíguas no espaço de coordenadas. Desse modo, podem existir UGDs separadas espacialmente, mas onde os atributos utilizados para identificá-las são tão similares a ponto de que as mesmas sejam geridas de maneira idêntica, constituindo o que chamamos de classes de gestão diferenciada (CGDs). Assim, uma CGD pode ser constituída por uma ou mais UGDs, onde cada UGD é espacialmente disjunta. Estima-se que mesmo para áreas maiores que 100 hectares, de 2 a 3 CGDs são suficientes para que a utilização desse conceito proporcione benefícios (TAYLOR; MCBRATNEY; WHELAN, 2007). No entanto, em uma rápida análise das abordagens da literatura, é possível verificar a utilização de até 5 CGDs para diferenciar espacialmente uma área. Abordagens que determinam automaticamente esse número normalmente utilizam as medidas *Fuzzy Performance Index* (FPI) e *Normalized Classification Entropy* (NCE) (ODEH; CHITTLEBOROUGH; MCBRATNEY, 1992). No entanto, essas medidas consideram questões particulares de algoritmos *fuzzy*, limitando o seu uso a esse tipo de algoritmo. Além disso, a possibilidade de não convergência dos valores fornecidos por essas medidas em alguns casos fez com que surgissem novas soluções, baseadas, por exemplo, na verificação da redução da variabilidade intraclasse à medida que o número de classes aumenta (ZHANG, 2010).

Em resumo, é possível identificar que, apesar de existirem as mais variadas abordagens para auxílio ao delineamento de UGDs em AP, não existe na literatura um consenso a respeito de qual delas é a mais apropriada. Além disso, essas soluções normalmente são validadas em situações muito particulares, no que diz respeito ao clima, solo, tipo de cultura e localização espacial. Assim, a busca por uma solução que fornecesse aos usuários finais de AP mapas de UGDs menos estratificados, com o intuito de facilitar a sua interpretação, visualização e uso prático efetivo, independentemente da localização, condições climáticas e tipos de solo e cultura presentes na lavoura, foi a principal motivação para o desenvolvimento desta tese.

1.4 Exemplo de Aplicação Alvo em Agricultura de Precisão

Suponha que um produtor rural possua dados espaciais de sua lavoura, obtidos por meio de sensores de campo ou imagens, representando o comportamento do solo e das plantas com relação a diversos fatores, além da produtividade obtida ao longo dos anos. Todos esses dados, por sua vez, apesar de já serem considerados complexos por possuírem coordenadas espaciais associadas a variáveis representativas para a cultura em questão, são originários de diferentes fontes e, portanto, podem estar armazenados em diferentes formatos de arquivos. Entretanto,

a adoção da AP em uma lavoura não está vinculada apenas a obtenção dos dados, mas sim em armazená-los de maneira organizada, com o objetivo de facilitar a sua recuperação e utilização futura em análises úteis para tomadas de decisão em nível gerencial. Um dos processos utilizados em AP para auxiliar na gestão da lavoura de uma propriedade rural é a divisão de cada talhão em UGDs. As UGDs permitem ao produtor iniciar a gestão de sua propriedade de maneira espacialmente diferenciada, realizando tratamentos específicos considerando as características de cada uma delas e sugerindo, por exemplo, maiores investimentos em unidades com potencial produtivo maior, proporcionando a economia de insumos e a preservação ao meio ambiente. Diante desse contexto, são evidentes as necessidades dos usuários finais de AP com relação a soluções que possam viabilizar o gerenciamento das informações de maneira organizada, com o objetivo de facilitar a sua utilização em ações gerenciais. Desse modo, devem ser criadas novas soluções em mineração de dados espaciais que permitam o descobrimento semi-automático do conhecimento intrínseco nos dados, sem deixar de considerar o conhecimento do próprio usuário final e outros fatores que podem influenciar em processos como o delineamento de UGDs.

1.5 Organização da Tese

Esta tese está organizada da seguinte forma:

- O Capítulo 2 aborda a fundamentação teórica relacionada à Agricultura de Precisão (AP), necessária para a compreensão dos conceitos utilizados nesta tese, e conceitos de Ciência da Computação relacionados com essa área de aplicação.
- O Capítulo 3 aborda a fundamentação teórica relacionada a dados espaciais, necessária para a compreensão das diferenças relacionadas à manipulação desse tipo de dado complexo com relação a dados convencionais. Mais especificamente, são explicados conceitos sobre bancos de dados espaciais, padrões OGC para interoperabilidade de dados espaciais via *Web* e SIGs para a *Web*.
- O Capítulo 4 aborda a fundamentação teórica relacionada à descoberta de conhecimento em bancos de dados espaciais, necessária para a compreensão das diferenças que devem ser consideradas em tarefas de agrupamento desse tipo de dado com relação a dados convencionais.
- O Capítulo 5 é constituído por uma revisão da literatura, contendo a sumarização de trabalhos correlatos a esta tese, considerando o agrupamento de dados espaciais e modelos

para manipulação de dados em AP. De maneira complementar, são apresentados experimentos que analisam as vantagens e desvantagens do uso das abordagens que constituem o estado da arte no delineamento de UGDs em AP.

- O Capítulo 6 descreve os objetivos gerais e específicos desta tese, bem como a definição formal do problema alvo de pesquisa, hipóteses investigadas e a metodologia utilizada.

- O Capítulo 7 descreve, de maneira detalhada, o critério de validação interna *SD-Spatial*, e apresenta experimentos comparativos utilizando critérios de validação interna e abordagens de agrupamento tradicionais.

- O Capítulo 8 descreve, de maneira detalhada, a abordagem de agrupamento espacial *SWMU Clustering*, e apresenta experimentos que comparam a sua eficácia com relação às abordagens que constituem o estado da arte no delineamento de UGDs em AP.

- O Capítulo 9 descreve, de maneira detalhada, a abordagem complementar *SWMU Polygon*, e apresenta experimentos que mostram o seu desempenho com relação ao armazenamento e recuperação de mapas de UGDs em bancos de dados espaciais.

- O Capítulo 10 descreve o sistema protótipo desenvolvido para a realização dos experimentos no âmbito desta tese, bem como uma aplicação prática de mapas de UGDs em operações agrícolas.

- O Capítulo 11 conclui os resultados e contribuições obtidos por esta tese, considerando propostas de trabalhos futuros para continuidade da pesquisa.

- O Apêndice A complementa a revisão da literatura realizada no Capítulo 5, onde são descritos os trabalhos relacionados à aplicação de técnicas de mineração de dados já existentes para auxílio ao delineamento de UGDs em AP.

- O Apêndice B transcreve o capítulo intitulado *Análise das possibilidades e tendências do uso de tecnologias da informação e comunicação em Agricultura de Precisão*, publicado em 2014 como parte integrante do livro técnico-científico *Agricultura de Precisão: resultados de um novo olhar*.

- O Apêndice C transcreve o artigo científico intitulado *A Cluster-Based Approach to Support the Delineation of Management Zones in Precision Agriculture*, apresentado oralmente e publicado nos anais da *IEEE International Conference on eScience 2014*.

- O Apêndice D transcreve o artigo científico intitulado *Utilização de Dados Espaciais na Tarefa de Agrupamento de Dados Aplicada na Gestão Agrícola*, aceito para publicação no periódico *Revista T.I.S. - Tecnologias, Infraestrutura e Software*.

- O Apêndice E transcreve o artigo científico intitulado *Clustering Approaches and Ensembles Applied in the Delineation of Management Classes in Precision Agriculture*, apresentado oralmente e publicado nos anais do *XVII Brazilian Symposium on GeoInformatics 2016*.

- O Apêndice F transcreve o artigo científico intitulado *Utilizando Ensembles com Abordagens de Agrupamento Espacial para o Delineamento de Classes de Manejo em Agricultura de Precisão*, aceito para publicação no periódico *Revista Brasileira de Cartografia*.

- O Apêndice G transcreve o artigo científico intitulado *Integração de Ferramentas de SIG e Mineração de Dados para Utilização em Atividades de Gestão Espacialmente Diferenciada Aplicada na Agricultura de Precisão*, aceito para publicação e apresentação oral no *XI Congresso Brasileiro de Agroinformática*.

- O Apêndice H exibe os procedimentos de preparação e execução da revisão sistemática realizada no Capítulo 5, com o auxílio da ferramenta StArt.

- O Anexo A exibe a estatística descritiva dos atributos utilizados nos experimentos desta tese, bem como índices de correlação existentes entre eles.

Capítulo 2

AGRICULTURA DE PRECISÃO

2.1 Considerações Iniciais

O primeiro registro de utilização experimental dos conceitos de AP foi realizado por Linsley e Bauer (1929), com a geração de mapas de acidez do solo para aplicação de calcário de forma espacialmente diferenciada. Entretanto, com a dificuldade de utilização dessa técnica na prática, que foi retomada apenas com o desenvolvimento da automação agrícola e a disponibilização do sinal de GPS na década de 90, muitas definições sobre AP surgiram a partir desse período, porém com objetivos gerais equivalentes. Pierce e Nowak (1999) definem AP como a aplicação de tecnologias e princípios para gerenciamento de variabilidades espaciais e temporais relacionadas a todos os aspectos da produção agrícola. Já Srinivasan (2006) apontam AP como uma estratégia amigável e holística para o meio ambiente, na qual podem ser utilizadas diferentes variáveis de entrada e métodos de cultivo para que sejam encontradas variações de solo e condições de cultura no campo. No Brasil, a AP foi recentemente definida como um sistema de gerenciamento agrícola baseado na variabilidade espacial das propriedades do solo e das plantas de uma lavoura, objetivando o aumento da produtividade e a economia de insumos agrícolas a partir de ações sustentáveis de proteção ao meio ambiente (BERNARDI, 2014; MOLIN; AMARAL; COLAÇO, 2015). Com o objetivo de inserir a AP no âmbito da computação, Ruß (2012) a definiu como um campo de pesquisa baseado em grandes coletas de dados para tomadas de decisão a respeito de operações agrícolas, com foco no gerenciamento sítio específico. Considerando os grandes conjuntos de dados que são coletados e utilizados por tecnologias da informação e comunicação (TICs), um modelo para o ciclo da AP baseado nesse contexto pode ser proposto, conforme exibido na Figura 2.1.

Figura 2.1: Ciclo da AP considerando a utilização de TICs.

Fonte: Elaborado pelo autor.

A partir do ciclo da Figura 2.1, pode ser observado o uso de sensores e equipamentos móveis na fase de coleta de dados; o uso de algoritmos para mineração de dados e geoprocessamento na fase de interpretação; e o uso de eletrônica embarcada para as intervenções realizadas na fase de aplicação. Além disso, deve-se observar a conexão em rede que já é possível de ser realizada entre todos os dispositivos envolvidos no processo.

Independentemente da disponibilidade das diferentes TICs que podem auxiliar a adoção da AP, em termos práticos, o principal motivo que deve levar os produtores rurais a optarem por esse sistema é a possibilidade de melhorar a gestão da sua lavoura como um todo. Os principais objetivos que devem ser atingidos com essa melhoria na gestão estão relacionados ao aumento do lucro obtido com a produção, diminuindo os prejuízos causados ao meio ambiente e melhorando a precisão de suas intervenções ao considerar os fatores de tempo e espaço nos fenômenos que ocorrem com o solo e a cultura durante o ciclo produtivo. Entretanto, para que a AP seja adotada com sucesso, é desejável a presença de três elementos críticos ao processo: informação, que está relacionada à disponibilidade de dados; tecnologia, que está relacionada às TICs envolvidas no processo; e tomada de decisão, que está relacionada à gestão da lavoura como um todo, englobando as análises realizadas com os dados obtidos e as recomendações de intervenção geradas a partir delas.

Neste capítulo, são apresentados os conceitos de AP e contribuições na área de TICs que podem ser aplicadas em AP, auxiliando na fundamentação teórica utilizada nesta tese. O restante do capítulo está organizado da seguinte forma:

- A Seção 2.2 apresenta as diferentes estratégias e abordagens que podem ser utilizadas em AP.
- A Seção 2.3 apresenta conceitos relacionados à amostragem de solo e intervenções à taxa variada utilizados em AP.
- A Seção 2.4 descreve os principais tipos de dados que são coletados em AP.
- A Seção 2.5 apresenta como representar os tipos de dados obtidos em AP em forma de mapa.
- A Seção 2.6 apresenta e discute a evolução e utilização do conceito de UGDs, utilizado como tema de aplicação desta tese.
- A Seção 2.7 faz um levantamento das contribuições em Ciência da Computação que vem sendo aplicadas em AP, relatando as principais TICs utilizadas na coleta e compartilhamento dos dados e novos conceitos que vem sendo utilizados para absorver o grande volume e variedade de dados espaciais que podem ser gerados com a sua utilização.

2.2 Estratégias e Abordagens para AP

Tradicionalmente, a AP pode seguir duas estratégias distintas de utilização. A primeira, chamada de iniciação, é mais simples e realiza a aplicação de fertilizantes e corretivos à taxa variada de acordo com a localização, com base apenas em dados obtidos por análise laboratorial de amostras do solo coletadas na área da lavoura. A segunda, mais abrangente, é chamada de especializada e considera, além de atributos do solo, informações sobre a cultura e dados de produtividade de anos anteriores para realização das intervenções. A partir dessa estratégia, pode ser construída uma base de dados mais consistente, fazendo com que a análise e interpretação dos fatores que implicam em diferentes valores de produtividade e, consequentemente, em diferentes necessidades de insumos para cada porção da lavoura, se torne uma tarefa bastante complexa (MOLIN, 2004). A escolha da estratégia a ser adotada depende muito dos objetivos que o usuário final deseja atingir com a utilização da AP. Se o foco está voltado apenas para a economia de insumos, a estratégia de iniciação já é suficiente, pois diferentemente das intervenções realizadas na agricultura tradicional, essa estratégia realiza a aplicação

estritamente necessária dos insumos para cada localização, evitando excessos desnecessários e contribuindo para a preservação do meio ambiente. No entanto, se existir a necessidade adicional do aumento da produtividade da lavoura, a estratégia especializada deve ser utilizada. Para tanto, investimentos em dados referentes ao comportamento do solo e das plantas e em mapas de produtividade devem ser realizados, para que, ao longo do tempo, o equilíbrio da lavoura com relação à produção e qualidade do produto final seja atingido (MOLIN; AMARAL; COLAÇO, 2015).

No Brasil, a predominância do uso de AP ainda está focada na estratégia de iniciação. No entanto, devido ao crescimento da disponibilidade de ferramentas e serviços relacionados à obtenção de dados que vem ocorrendo nos últimos anos, essa técnica deve ser encarada como um sistema de gestão mais amplo, que considera a variabilidade espacial relacionada à produtividade, solo, planta, infestações, doenças e pragas (INAMASU; BERNARDI, 2014). Assim, pesquisas multidisciplinares relacionadas ao auxílio na tomada de decisão em tarefas de AP tendem a aumentar. No contexto da Ciência da Computação, é esperado o surgimento de novas soluções relacionadas ao armazenamento, tratamento e uso correto de dados espaciais originários de diferentes fontes por meio do uso de TICs (QUEIRÓS, 2014).

2.3 Amostragem de Solo em Grade e Aplicação à Taxa Variada

A técnica para geração de mapas individuais para cada indicador de fertilidade do solo por meio de amostras coletadas em campo - conhecida como amostragem de solo em grade - é bastante popular em AP. Nessa técnica, utilizada para determinar as necessidades do solo de uma lavoura de maneira detalhada, um talhão é dividido em quadrículas imaginárias de onde são retiradas amostras de solo. Quando se tem um conhecimento prévio da variabilidade da área, obtido, por exemplo, a partir do delineamento de UGDs, deve-se utilizar a amostragem por ponto. Nesse tipo de amostragem, em cada quadrícula, são retiradas para análise laboratorial uma amostra normalmente central e algumas subamostras, localizadas em um raio suficientemente pequeno com relação à amostra central para não prejudicar a variabilidade espacial da área de estudo. No caso em que não existe um conhecimento prévio da variabilidade espacial da área, além da amostragem por pontos, também pode ser utilizada a amostragem por células, onde a subamostragem pode ser realizada a partir de diferentes percursos dentro de cada célula (MOLIN; AMARAL; COLAÇO, 2015).

Apesar de bastante difundida em AP, a execução dessa técnica pode ser limitada devido ao

seu alto custo operacional. Desse modo, um número mínimo de amostras de solo a serem retiradas pode ser determinado por especialistas na área de cultivo em questão. Entretanto, quando se utiliza amostragem por pontos, a grade obtida pode ter sua resolução espacial aumentada com a utilização de técnicas de interpolação espacial, como o inverso da potência das distâncias (SHEPARD, 1968) e a *krigagem* (MATHERON, 1969).

A partir dos dados espaciais obtidos por meio das amostras e de equações previamente definidas, podem ser gerados mapas de recomendação de aplicação individual de insumos à taxa variada, normalmente realizadas por máquinas agrícolas automatizadas contendo eletrônica embarcada capaz de interpretar esses mapas (SAWYER, 1994). As equações mais simples e mais utilizadas nesse tipo de intervenção estão relacionadas à aplicação de corretivos como calcário e gesso, que utilizam os atributos de saturação por bases (V%), capacidade de troca de cátions (CTC) e textura do solo (MOLIN; AMARAL; COLAÇO, 2015).

Além de informações obtidas a partir de amostragem do solo, outros atributos relacionados ao solo e também a cultura podem ser utilizados para a geração de um mapa de recomendação. O importante é levar em consideração que, para cada valor obtido em campo associado a uma localização espacial, deverá ser obtido um valor de recomendação para um determinado insumo agrícola para a mesma localização. Assim, um mapa de recomendação de aplicação à taxa variada deverá ser criado com base na mesma grade de amostras utilizada para os dados coletados.

2.4 Tipos de Dados em AP

A variedade de equipamentos embarcados e sensores capazes de produzir informações úteis para AP tem crescido muito ao longo dos anos. No entanto, algumas características comuns a todos esses dados devem ser observadas durante o desenvolvimento de soluções capazes de agrupá-los para uma finalidade específica. Primeiramente, os atributos presentes em conjuntos de dados utilizados em AP estão normalmente associados a pontos bidimensionais representados por coordenadas de posicionamento de latitude e longitude do globo terrestre, fazendo com que sejam considerados como dados geoespaciais, dados georreferenciados, dados geográficos ou ainda, de uma maneira mais geral, como dados espaciais. Essa característica traz a necessidade de se tratar cada amostra considerando dois espaços multidimensionais distintos: o espaço de atributos, que contém os valores medidos para uma variável em uma determinada localização; e o espaço de coordenadas, que contém a localização espacial da amostra em si.

Outra característica importante e que deve ser considerada é com relação à densidade amos-

tral. A utilização de técnicas tradicionais de amostragem de solo, ou até mesmo de sensores móveis que percorrem a lavoura a uma velocidade normalmente constante, nos permite considerar que são obtidas quantidades similares de amostras por hectare. Por outro lado, novas técnicas, como a desenvolvida por Tangerino (2014), permitem amostrar dados em campo considerando a variabilidade e dependência espacial de um determinado atributo, fazendo com que seja obtida apenas a quantidade necessária de amostras para representá-lo. A partir dessa metodologia, em áreas com maior variabilidade são retiradas maiores quantidades de amostras do que em áreas com menor variabilidade. Entretanto, devido à necessidade de se acomodar os valores obtidos para diferentes atributos em uma grade regular única para possibilitar, por exemplo, a sua utilização pelos algoritmos de agrupamento utilizados nas diferentes abordagens para delineamento de UGDs em AP, os dados utilizados para validação dos resultados desta tese sempre possuem densidade constante, de acordo com a resolução espacial definida pelo usuário final.

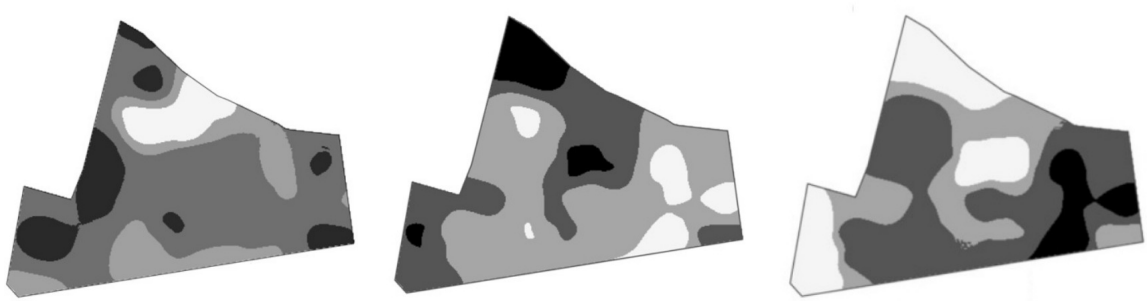
Do ponto de vista de bancos de dados espaciais, os dados brutos produzidos em AP podem ser de natureza vetorial ou matricial (CASANOVA, 2005). Dentre os vetoriais, destacam-se aqueles coletados por sensores de campo relacionados ao solo e as plantas, inspeções para busca de pragas e doenças, amostras de solo georreferenciadas e dados de produtividade. Com relação aos matriciais, destacam-se as imagens de sensoriamento remoto e aéreas, cuja informação é armazenada em imagens compostas por pixels contendo informações em diferentes bandas espectrais capazes de identificar características tanto do solo quanto da cultura. Na Seção 2.7, são exemplificadas diferentes maneiras de se obter esses dados, bem como novas formas de armazenamento, recuperação e manipulação dos dados espaciais para viabilizar o seu uso nas atividades de AP.

2.5 Geração de Mapas em AP

De posse dos dados espaciais obtidos para uma determinada lavoura, o usuário final de AP poderá criar, a partir de um SIG, diferentes mapas temáticos relacionados às variáveis coletadas. Cada mapa corresponde a um gráfico que exibe as amostras representadas por pontos em um sistema cartesiano, utilizando o eixo x para os valores de longitude e o eixo y para os valores de latitude presentes no espaço de coordenadas. Os valores de cada variável que compõem o espaço de atributos podem ser graduados em diferentes intervalos para uma melhor visualização e utilização do mapa. Dados relacionados a atributos do solo obtidos a partir de amostragem e interpolação espacial, e suas correspondentes recomendações para aplicação individual de insumos, conforme descrito na Seção 2.3, são exemplos de mapas temáticos que podem ser gerados em AP. A Figura 2.2 ilustra um exemplo de mapa de recomendação graduado para

aplicação de calcário, gerado a partir de mapas de atributos do solo.

Figura 2.2: Exemplo de mapas graduados em 4 intervalos com a utilização de SIG: (a) Mapa de capacidade de troca de cátions; (b) Mapa de saturação por bases; (c) Mapa de recomendação para aplicação de calcário.

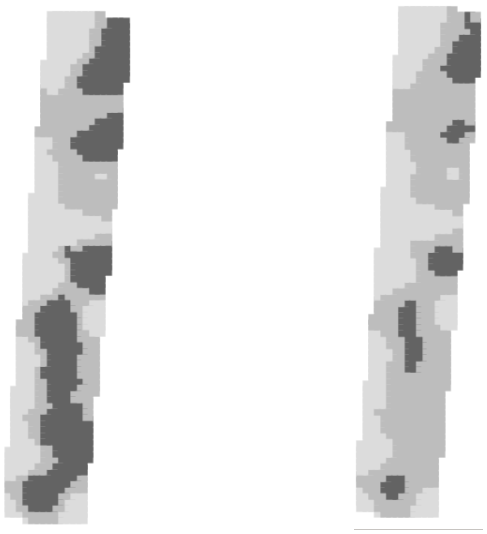


(a) (b) (c)

Fonte: Adaptada de Bernardi (2016).

Além do processo de interpolação espacial, a fase de pré-processamento dos dados para a geração de mapas temáticos pode conter os processos de identificação, caracterização e remoção de erros (MENEGATTI, 2003), com o intuito de melhorar a qualidade final dos mapas. Para que um mapa temático seja interpretado de maneira eficaz pelos usuários finais, é muito importante a manipulação correta de alguns parâmetros em sua construção, pois a atribuição de intervalos para uma determinada variável sem a utilização de um critério adequado poderá limitar as possibilidades de análise das condições de uma lavoura. A Figura 2.3 exibe um exemplo de dois mapas temáticos construídos a partir do mesmo conjunto de dados espaciais de um atributo do solo, porém graduado utilizando-se metodologias distintas.

Figura 2.3: Exemplo de mapas temáticos gerados a partir do mesmo conjunto de dados espaciais, graduados considerando (a) a metodologia de quantis; e (b) a metodologia de intervalos iguais.



(a) (b)

Fonte: Elaborada pelo autor.

A partir da Figura 2.3 é possível observar, em ambos os mapas, três diferentes tons de cinza que caracterizam três intervalos distintos de graduação para o atributo considerado. Entretanto, a metodologia de quantis (Figura 2.3 (a)), que procura gerar intervalos de graduação que permitam dispor quantidades similares de amostras em cada intervalo, fornece um mapa temático diferente do que é obtido utilizando-se a metodologia de intervalos iguais (Figura 2.3 (b)), definidos a partir do intervalo de valores obtidos para o atributo. Essas diferenças estão relacionadas, por exemplo, a subconjuntos de amostras que ficaram dispostas na mesma faixa de graduação considerando a metodologia de quantis; e em faixas de graduação distintas, considerando a metodologia de intervalos iguais. Assim, a graduação utilizando metodologias distintas pode gerar interpretações diferentes do comportamento da lavoura com relação ao atributo estudado e, consequentemente, proporcionar a obtenção de mapas de recomendação distintos. A estatística descritiva dos dados é uma das ferramentas que pode ajudar o usuário final a escolher qual a melhor graduação para visualização dos conjuntos de dados espaciais de cada atributo em forma de mapas temáticos. Entretanto, essa questão não será discutida no âmbito desta tese, uma vez que os mapas temáticos a serem gerados para o usuário final são compostos por UGDs previamente rotuladas por abordagens de agrupamento espacial.

Uma alternativa aos mapas de recomendação para aplicação à taxa variada, principalmente quando não se tem a disponibilidade do serviço de automação para a realização dessa tarefa, é a geração de mapas de regiões delimitadas conhecidas comercialmente como zonas de aplicação

(BRASIL, 2013). Esses mapas são aproximações dos intervalos gerados em mapas tradicionais de recomendação graduados, fixando valores médios para aplicação de insumos agrícolas em regiões delimitadas, permitindo intervenções a partir de máquinas convencionais, ou seja, sem eletrônica embarcada. Apesar de viabilizarem, mesmo que de uma maneira aproximada, intervenções com taxas diferenciadas dentro de um talhão, na prática essas intervenções podem se tornar um tanto quanto complexas, por conta de regulagens e manobras que devem ser realizadas quando um maquinário desprovido de automação é utilizado.

2.6 Unidades de Gestão Diferenciada (UGDs)

A essência da AP está no tratamento da lavoura de maneira espacialmente diferenciada, independentemente da resolução espacial utilizada nesse processo. Dependendo do tipo de solo ou da cultura, a variabilidade espacial pode ser percebida tanto em poucos metros quadrados de área, quanto em hectares que transcendem a área de um talhão. Assim, diversas estratégias para o tratamento localizado podem ser utilizadas, tais como: aplicações em tempo real por meio de sensores de solo ou planta; aplicações à taxa variada baseadas em mapas de recomendação; e aplicações baseadas em unidades de gestão diferenciada (UGDs) (MOLIN; AMARAL; COLAÇO, 2015). UGDs são mais comumente definidas na literatura como zonas de manejo (ZMs)[1].

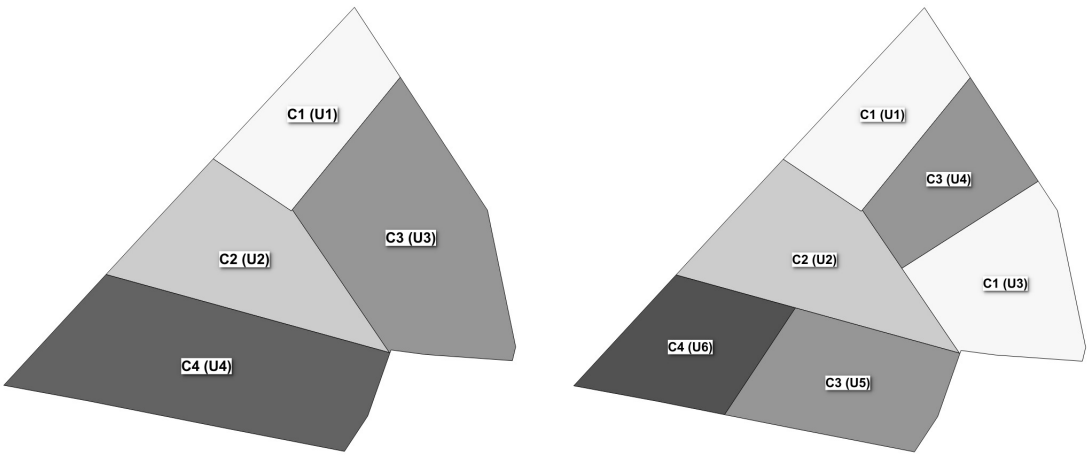
Considerando o contexto da agricultura, as UGDs são regiões espaciais internas a um talhão com características e potencial produtivo distintos, expressados por uma combinação de fatores limitantes à produtividade que impossibilitam o seu desempenho vegetal homogêneo (DOERGE, 1999; SALVADOR; ANTUNIASSI, 2011; SCHWALBERT, 2014). Já considerando o contexto de dados espaciais, Taylor, McBratney e Whelan (2007), Pedroso (2010), Córdoba (2013) definem UGD como uma área espacialmente contígua em que um tratamento particular deve ser aplicado, introduzindo mais dois conceitos relacionados: classe de manejo (CM), composta por um conjunto de regiões disjuntas (ou seja, por UGDs), para as quais o tratamento aplicado deve ser o mesmo; e unidade de manejo (UM), termo genérico utilizado para se referir tanto a CMs quanto a UGDs. Finalmente, sob a perspectiva de mineração de dados espaciais, Ruß (2012) define a tarefa de delineamento de UGDs como a transformação de coleções de dados espaciais pontuais obtidos de um talhão agrícola em agrupamentos para um fim específico.

Devido ao atual momento da AP, onde pode ser observado o aumento do uso de TICs e, consequentemente, a recuperação dos conceitos que deram origem a esse sistema de gerenciamento agrícola, o termo UGDs tem sido muito utilizado para identificar as ZMs. Segundo

[1] Da tradução literal do inglês *management zones*

Molin, Amaral e Colaço (2015), as UGDs são regiões delimitadas em uma área de produção agrícola com variabilidade interna desprezível, devendo ser consistentes ao longo do tempo de forma a caracterizar o potencial de resposta dessa área com relação à produtividade. Desse modo, UGDs que possuem as mesmas características são agrupadas em uma mesma CGD. Assim, uma CGD pode ser composta por uma ou mais UGDs, onde cada UGD que compõe a CGD é disjunta com relação às outras UGDs, fazendo com que essa CGD não seja contígua espacialmente. Além disso, uma CGD deve ser composta por pelo menos uma UGD. A Figura 2.4 ilustra dois exemplos de subdivisão de uma área hipotética em UGDs, considerando CGDs contíguas e não contíguas espacialmente.

Figura 2.4: Mapas contendo possíveis delineamentos de UGDs para uma área hipotética, com (a) CGDs exclusivamente contíguas; e (b) CGDs contíguas e não contíguas espacialmente.



(a) (b)

Fonte: Elaborada pelo autor.

Na Figura 2.4 (a), é possível observar que as 4 CGDs, identificadas nos mapas por diferentes tons de cinza e pelos rótulos C1, C2, C3 e C4 são exclusivamente contíguas, ou seja, são compostas por UGDs únicas e identificadas, respectivamente, pelos rótulos U1, U2, U3 e U4. Por outro lado, na Figura 2.4 (b), é possível observar a existência de duas CGDs não contíguas, identificadas pelos rótulos C1 e C3, e que são compostas, respectivamente, pelos conjuntos de UGDs (U1,U3) e (U4,U5). Adicionalmente, as CGDs C2 e C4 são compostas por uma única UGD cada uma, identificadas, respectivamente, pelos rótulos U2 e U6.

Com o intuito de consolidar as definições acima citadas, nesta tese o termo UGD será utilizado para identificar uma área espacialmente contígua com gestão agrícola independente. Além disso, será utilizado o termo CGD para identificar a composição de uma ou mais UGDs espacialmente não adjacentes, mas cujos procedimentos de gestão agrícola devem ser os mesmos. A

partir dessas definições, dentro dos limites de uma UGD podem ser consideradas intervenções homogêneas relacionadas à aplicação de insumos agrícolas, tornando-as unidades de gestão agrícola independentes. Para que se tornem permanentes, tais regiões devem ser obtidas a partir de conjuntos de dados espaciais de qualidade, de forma a representar a variabilidade espacial intrínseca da área que, por sua vez, deve ser independente de fatores que podem ser afetados pela interferência humana, como intervenções agronômicas utilizando produtos químicos. Desse modo, esses conjuntos de dados espaciais devem ser compostos por atributos associados à natureza do solo, como caraterísticas físicas, de textura e capacidade de retenção de água; e por atributos relacionados à cultura, como produtividade histórica e índices de biomassa.

Considerando as estratégias e abordagens de AP definidas na Seção 2.2, as UGDs estão mais adequadas para a estratégia especializada. A utilização de UGDs bem definidas em uma lavoura viabiliza ao usuário final de AP um nível de gerenciamento que permite o estabelecimento de um padrão de tratamento específico para cada uma delas, fazendo com que as atividades de investigação e coleta de dados demandem uma gestão individual personalizada (MOLIN; AMARAL; COLAÇO, 2015). Assim, podem ser utilizadas estratégias de economia de insumos agrícolas em UGDs com baixo potencial de resposta, ou seja, onde as características intrínsecas de solo e cultura não permitem um aumento da produtividade, independentemente da quantidade de insumos aplicada; e estratégias para intensificar a utilização desses mesmos insumos visando o aumento da produtividade em UGDs onde o solo e a cultura respondem melhor a esses tratamentos.

As UGDs podem ser utilizadas em diversas aplicações que compõem as etapas da gestão de uma área de cultivo agrícola. Novas amostragens de solo podem ser realizadas considerando uma única amostra por UGD ao invés da utilização de grades regulares, reduzindo substancialmente o custo operacional dessa técnica. Nas etapas relacionadas à aplicação de insumos, podem ser recomendadas doses uniformes para intervenções dentro dos limites de cada UGD. Já em aplicações relacionadas ao plantio, as UGDs podem ser utilizadas para diferenciar a população de plantas e, consequentemente, reduzir o custo com sementes em regiões de baixo potencial produtivo (MOLIN; AMARAL; COLAÇO, 2015). Em aplicações mais específicas, como a irrigação de precisão, as UGDs podem ser utilizadas para diferenciar espacialmente a lâmina de água para suprir as necessidades hídricas em localizações distintas da lavoura (CID-GARCIA; BRAVO-LOZANO; RIOS-SOLIS, 2014; OLDONI; BASSOI, 2016). No Brasil, o percentual de adoção da AP ainda é baixo, fazendo com que ainda sejam coletados dados com resolução espacial mais baixa e as escalas em nível de pixel sejam utilizadas apenas em aplicações mais específicas, relacionadas à diferenciação de plantas. Desse modo, as UGDs podem ser utilizadas para auxiliar na simplificação do manejo agrícola como um todo, porém já priorizando um

tratamento espacialmente diferenciado dentro de cada talhão.

Complementarmente, conceitos relacionados à estratificação dos mapas de UGDs serão definidos a seguir no âmbito desta tese, com o intuito de proporcionar um embasamento teórico para as contribuições obtidas. A Figura 2.5 exibe dois exemplos de mapas de UGDs delineados a partir das mesmas amostras de dados espaciais coletados em campo, porém utilizando duas abordagens distintas de agrupamento de dados espaciais.

Figura 2.5: Exemplos de mapas de UGDs delineados por duas abordagens distintas de agrupamento de dados espaciais: (a) abordagem A; (b) abordagem B.

(a) (b)

Fonte: Elaborada pelo autor.

Na Figura 2.5, ambos os mapas foram delineados utilizando 4 CGDs categorizadas em diferentes tons de cinza, de acordo com parâmetros definidos pelo usuário final. Analisando visualmente esses mapas, é fácil perceber que o mapa de UGDs disponibilizado pela abordagem B (Figura 2.5 (b)) fornece uma solução clara contendo 3 CGDs compostas por uma única UGD cada, e 1 CGD composta por 2 UGDs, exibidas no tom mais claro, na parte superior do mapa. Entretanto, se considerarmos o mapa disponibilizado pela abordagem A (Figura 2.5 (b)), percebe-se uma quantidade muito grande de UGDs, em geral com áreas muito pequenas, compondo cada uma das CGDs, principalmente considerando as duas CGDs identificadas pelos tons mais claros, na parte superior do mapa. A partir dessa análise, é possível verificar que a abordagem A agrupou os dados utilizando apenas os atributos convencionais do conjunto de dados, enquanto que a abordagem B se utilizou de restrições aplicadas no espaço de coordenadas para obter CGDs mais contíguas e fáceis de interpretar, mesmo que em um dos casos uma CGD acabou sendo formada por mais de uma UGD.

Considerando essas questões, definimos no âmbito desta tese que as UGDs que possuírem um percentual de área relativa com relação à área total da unidade produtiva estudada menor do que um percentual definido pelo usuário final, gerando perda de contiguidade espacial e, consequentemente, uma estratificação acentuada e visualmente perceptível no mapa de UGDs,

serão denominadas estratos. Esse percentual de área deverá ser, no mínimo, a área ocupada por uma única amostra do conjunto de dados utilizado, ou seja, correspondente à resolução da grade regular espacial utilizada no conjunto de dados. Da mesma maneira, também serão considerados estratos os buracos formados por essas UGDs em CGDs adjacentes a elas. Levando-se em consideração essas definições, a Figura 2.5 (a) exibe um exemplo de mapa de UGDs excessivamente estratificado, ou seja, formado por diversas UGDs e buracos compostos pela área ocupada por uma única amostra. Por outro lado, o mapa de UGDs exibido na Figura 2.5 (b), apesar de apresentar perda de contiguidade espacial em uma das CGDs, pode ser considerado como válido, caso a área das UGDs que compõem essa CGD possuam razão com relação à área total da unidade produtiva maior do que o percentual mínimo definido pelo usuário final.

Um dos objetivos desta tese foi o desenvolvimento de uma abordagem de agrupamento espacial que auxilie no delineamento de CGDs formadas por uma ou mais regiões disjuntas, obtendo-se mapas de UGDs com pouco ou nenhum nível de estratificação, sem que as características relacionadas à variabilidade espacial dos atributos de solo e cultura fossem prejudicadas. Assim, a definição clara dos conceitos relacionados à UGDs e CGDs foi de extrema importância para a obtenção de uma solução útil e de fácil utilização pelo usuário final de AP.

2.7 Contribuições da Ciência da Computação para AP

A caracterização da AP como um sistema de gerenciamento agrícola baseado na coleta e no tratamento de dados espaciais oriundos de diversas fontes, obtidos nas diferentes fases de crescimento e acompanhamento das lavouras agrícolas, tem tornado essa técnica cada vez mais próxima e dependente de avanços tecnológicos vinculados à Ciência da Computação. Desse modo, conceitos como as TICs, Mineração de Dados, *IoT* e *Big Data* tem aparecido frequentemente em pesquisas aplicadas em AP (LI; CHUNG, 2015). A seguir, são descritos conceitos e ferramentas que envolvem a aplicação da Ciência da Computação na AP, e que de alguma maneira foram utilizados durante o desenvolvimento desta tese.

2.7.1 Tecnologias da Informação e Comunicação

As TICs são ferramentas computacionais utilizadas por pessoas e organizações para processamento de suas informações e propósitos de comunicação (ZHANG; AIKMAN; SUN, 2008). No contexto da agricultura, as TICs podem ser utilizadas nas diversas fases da cadeia de produção agrícola, que vão desde a produção de sementes, passando pelo plantio e colheita, até a distribuição dos alimentos processados para o consumo (MASSRUHÁ, 2014). O grande desafio das

TICs na agricultura, e em especial na AP, está relacionado à vulnerabilidade a que estão submetidos os processos agrícolas, devido às variações climáticas, infestações e outras possíveis forças externas que contribuem para que o rendimento da agricultura seja mais baixo do que nas outras indústrias (TING, 2011). No contexto da AP, a utilização de TICs está centrada em atividades que envolvem geoprocessamento, com a utilização de SIGs e algoritmos de mineração de dados; e nas atividades de coleta de dados, com a utilização de sensoriamento remoto e *software* embarcado.

2.7.2 Utilização de TICs na Coleta de Dados Espaciais para AP

Conforme já descrito na Seção 2.3, a amostragem de solo georreferenciada, mesmo sendo em geral muito dispendiosa, ainda é uma das técnicas mais utilizadas para se obter dados confiáveis capazes de identificar a variabilidade espacial de uma área de cultivo. Desse modo, trabalhos como o desenvolvido por Morari, Castrignanò e Pagliarin (2009), onde foram utilizados teores de areia, argila e cascalho para a determinação de UGDs, são muito comuns na literatura. Uma alternativa que vem sendo amplamente utilizada em conjunto ou até mesmo em substituição a algumas informações extraídas da amostragem de solo são as medidas de condutividade elétrica (CE) do solo. Essas medidas são baseadas na capacidade do solo em variar espacialmente a condução de corrente elétrica conforme suas propriedades físicas e químicas, com destaque para a salinidade, porcentagem de saturação, densidade volumétrica e umidade (RABELLO, 2009). Normalmente, os dados experimentais de CE são coletados em diferentes profundidades em cada área agrícola, a partir de diferentes equipamentos baseados em princípios distintos. Dentre os mais utilizados, destacam-se os que utilizam medidas de resistividade elétrica por diferença de potencial entre dois pares de eletrodos; e os que utilizam medidas de indução eletromagnética, obtidas a partir de bobinas de transmissão indutoras de corrente elétrica (RABELLO, 2011).

Os dados altimétricos e mapas de declividade, obtidos por meio de modelos digitais de elevação oriundos de imagens de satélite, ou até mesmo aproveitando-se a dimensão de altitude fornecida pelo GPS acoplado a outros sensores, também são largamente utilizados em AP. Segundo Krummel e Su (1996), os dados altimétricos podem influenciar na produtividade de uma cultura, pois podem estar relacionados, por exemplo, com a retenção de água no solo em alguns locais.

Em pesquisas mais recentes, redes de sensores fixados em campo (capazes de capturar informações importantes como a umidade e temperatura do solo) tem sido utilizadas em conjunto com estações climatológicas para auxiliar na tomada de decisão em atividades baseadas

em AP, como a irrigação de precisão (TORRE-NETO, 2005). Dentre esses novos equipamentos, destacam-se também os sensores que medem a refletância no dossel das plantas, obtida a partir de informações radiométricas nos comprimentos de onda do espectro eletro magnético do verde, vermelho e azul (KITCHEN; GOULDING; SHANAHAN, 2008). Para que possam ser utilizadas no contexto agrícola, essas informações são normalmente transformadas em índices de vegetação como o NDVI (*Normalized Difference Vegetation Index*), capazes de estimar a biomassa de uma cultura (ROSA, 2015). Ainda com relação às plantas, mais especificamente em culturas normalmente produzidas em áreas pequenas, é possível obter informações relacionadas ao teor de clorofila das folhas por meio de clorofilômetros portáteis, de baixo custo e fáceis de operar (NASCIMENTO, 2013).

Devido à grande oferta de dispositivos móveis disponíveis no mercado com receptor de GPS integrado, surgiram também aplicativos capazes de auxiliar na obtenção de dados em campo a partir de inspeções visuais para identificação de doenças ou pragas que atingem as lavouras (JORGE, 2011). Com a aumento do uso dos *smartphones* e a rápida adoção do sistema operacional móvel Android (ANDROID, 2017), diversas aplicações comerciais tem surgido, incluindo diferentes funcionalidades capazes de auxiliar o usuário final de AP durante o ciclo produtivo, servindo para estimular o produtor rural a adotar técnicas para o gerenciamento espacialmente diferenciado da lavoura.

Outra tipo de informação espacial extremamente importante para a AP são os dados históricos de produtividade. Segundo Molin, Amaral e Colaço (2015), os mapas de produtividade, também chamados de mapas de colheita, representam a informação mais completa para a visualização da variabilidade espacial de uma lavoura, trazendo a resposta mais exata da cultura no que diz respeito ao processo de colheita realizado, considerando as tecnologias atuais existentes para a sua mensuração. Normalmente, esses valores são medidos em toneladas por hectare e obtidos por sensores de fluxo e GPS acoplados a uma colhedora automatizada, como acontece nas culturas de grãos e cana-de-açúcar. Mesmo nos casos em que a maioria das colheitas são realizadas manualmente, é possível que estes valores sejam obtidos por meio de outros mecanismos, como contagem de caixas de produto produzidas por planta, por exemplo. Uma alternativa consiste em estimar dados de produtividade a partir de outras variáveis relacionadas à cultura. Scarpari, Gomes e Beauclair (2009) desenvolveram um modelo para estimar a produção de cana-de-açúcar sob diferentes condições climáticas, por meio de dados como maturação, idade do canavial e propriedades físico-químicas do solo. Zhang (2010) utilizaram índices de vegetação produzidos por imagens de satélite para prever a produção de uma cultura de trigo. Com foco mais voltado para a construção de algoritmos, Ruß e Kruse (2010) utilizaram modelos de regressão espacial aplicados a dados de CE, índices de vegetação e quantidade aplicada

de fertilizantes para predição da produtividade em uma cultura de trigo.

Nos últimos anos, as pesquisas em AP têm sido impulsionadas por dados matriciais obtidos por sensores capazes de capturar diferentes características relacionadas ao solo e as plantas em diferentes bandas espectrais, proporcionando o cálculo de índices importantes para mensuração da biomassa das culturas, como o NDVI. A principal fonte desse tipo de dado é o sensoriamento remoto, que no contexto da AP deve considerar preferencialmente sensores que fornecem resolução espacial de poucos metros ou centímetros. Imagens originárias de satélites como GeoEye, WorldView, QuickBird e IKONOS[2] são as mais utilizadas em AP, porém são comercializadas por empresas privadas, o que pode tornar o seu custo de aquisição alto. Como alternativa, podem ser utilizadas imagens de satélite gratuitas, com as obtidas pela missão indiana IRS-P6, direcionada para estudos de vegetação e culturas; e pela missão europeia Sentinel-2, direcionada para monitoramento agrícola, detecção de desastres, dentre outros. As imagens geradas por essas missões possuem boa resolução espacial e temporal e cobertura mundial, sendo distribuídas, respectivamente, pelos catálogos de imagens do INPE[3] (Instituto Nacional de Pesquisas Espaciais) e EarthExplorer[4]. Nos últimos anos, uma técnica que vem sendo muito utilizada para obtenção de dados matriciais em campo, com alta resolução espacial, é o imageamento aéreo, principalmente após o surgimento das aeronaves remotamente pilotadas (ARPs), popularmente conhecidas como *drones*. Com um custo operacional que vem sendo bastante reduzido ao longo do tempo, esses veículos são capazes de realizar voos programados com câmeras multi e hiperespectrais a bordo, capazes de capturar imagens com cerca de 300 bandas distintas contendo informações úteis relacionadas à vegetação, solo e água (JORGE; INAMASU; CARMO, 2011).

Em resumo, todos os tipos de dados obtidos pelas TICs citadas nesta seção podem ser utilizados para auxiliar no delineamento de UGDs para AP, desde que estejam relacionados a características intrínsecas do solo e da cultura e não sejam fruto de ações humanas, como teores de elementos químicos presentes em insumos e corretivos agrícolas. Entretanto, para que sejam geradas UGDs confiáveis, invariantes no tempo e que auxiliem efetivamente em uma gestão espacialmente diferenciada da lavoura ao longo dos anos, esses dados devem possuir uma qualidade no mínimo aceitável. Essa qualidade pode estar relacionada a diversos fatores, tais como o tipo de equipamento utilizado na coleta, precisão das medidas e do sensor de GPS utilizado, remoção de leituras erradas com valores muito distantes da média, dentre outros. Para que esses fatores possam ser minimizados, é muito importante que esses dados sejam catalogados e sigam padrões comuns de armazenamento e recuperação.

[2]Digital Globe Inc., Longmont, Colorado, USA, http://www.digitalglobe.com
[3]Catálogo de Imagens do INPE, São José dos Campos, SP, http://www.dgi.inpe.br/CDSR/
[4]USGS EarthExplorer, Reston, VA, USA, http://earthexplorer.usgs.gov

2.7.3 Metadados e Fluxo da Informação em AP

Na agricultura, em especial nas atividades relacionadas à AP, a questão da padronização do fluxo da informação já vem sendo discutida há muito tempo, principalmente devido à utilização de componentes de eletrônica embarcada em máquinas agrícolas oriundos de diferentes fabricantes, dificultando a comunicação e troca de informações entre eles. No final da década de 80, foram realizados esforços entre grupos de pesquisa, associações e a ISO (*International Organization for Standardization*) para implementação da norma internacional conhecida como ISOBUS. O propósito da ISOBUS é prover um padrão aberto para interconexão de sistemas eletrônicos embarcáveis por meio de um barramento, permitindo a interconexão de dispositivos e a comunicação de dados entre eles, mesmo que sejam produzidos por fabricantes distintos (BRASIL, 2011). No Brasil, um grupo de pesquisadores e representantes da indústria criou a Força Tarefa ISOBUS-Brasil, promovendo a orientação a grupos interessados em desenvolver sistemas de controle e automação agrícola baseados nesses padrões (SOUSA, 2011).

Entretanto, quando o contexto é a organização das informações espaciais já armazenadas em bancos de dados, os esforços relacionados à padronização são mais recentes. Além dos padrões de interoperabilidade desenvolvidos pela OGC (Open Geospatial Consortium), já utilizados em pesquisas relacionadas à criação de modelos para desenvolvimento de sistemas para suporte à AP (MURAKAMI, 2007; RIBEIRO-JÚNIOR, 2007), também existe a preocupação em se criar perfis de metadados específicos, a partir de padrões ou perfis já existentes. Com base nessa prerrogativa, foi desenvolvido pela Embrapa um perfil de metadados adaptado às necessidades de AP no Brasil, tomando como base o perfil MGB e o padrão ISO 19115:2003 (SPERANZA; QUEIRÓS, 2010; QUEIRÓS, 2011). A partir da utilização de padrões e perfis como este, é esperado que tarefas de AP que dependem de dados obtidos com qualidade, como o delineamento de UGDs, sejam facilitadas no sentido de auxiliar o usuário final em quais conjuntos de dados devem ser utilizados.

2.7.4 Novas Tecnologias para Manipulação de Dados Espaciais em AP

A crescente disponibilidade de dispositivos capazes de adquirir grande quantidade e variedade de dados espaciais, além de meios de comunicação entre esses dispositivos para que os dados sejam compartilhados e utilizados para diferentes finalidades, tem impulsionado a utilização de novos conceitos envolvendo armazenamento, recuperação, compartilhamento e uso dos dados em AP.

A grande variedade de atributos de solo e da cultura que pode ser adquirida com a utili-

zação desses dispositivos, além do aumento considerável da quantidade de dados disponíveis, se levarmos em consideração as altas resoluções espaciais alcançadas pelas imagens aéreas e de sensoriamento remoto, fazem com que o conceito de *Big Data* (HAN, 2011) esteja presente em pesquisas mais recentes de AP. Complementarmente, as questões relacionadas às conexões de rede disponíveis para esses dispositivos, habilitando interfaces de coleta e leitura de dados para se comunicarem entre si, de forma a permitir processamento em tempo real e cada vez mais preciso, fazem com que o conceito de Internet das Coisas (IoT) (KOPETZ, 2011) também seja atualmente muito pesquisado em AP. Considerando essas questões e o desafio de prover funcionalidades que proporcionem escalabilidade, elasticidade e tolerância a falhas para grandes quantidades de dados que devem estar sempre disponíveis para trafegar entre os diversos dispositivos conectados em uma rede, as plataformas de computação em nuvem (ABADI, 2009) surgem como uma solução promissora para o contexto da AP.

Uma maneira de utilizar esses novos conceitos de forma prática na agricultura é por meio dos Sistemas de Informações para Gestão Agrícola (FMIS[5]) (SØRENSEN, 2011). No contexto da AP, um FMIS é definido como um sistema dedicado capaz de lidar com grandes quantidades de dados históricos que devem ser utilizados de maneira sítio-específica, ou seja, com foco no armazenamento e transferência de dados espaço-temporais (FOUNTAS, 2015). Em outras palavras, um FMIS aplicado em AP deve ser um *software* personalizado para as atividades diárias de uma propriedade rural, com foco na variabilidade espaço-temporal verificada nos dados coletados em uma lavoura para análises subsequentes, gerando conhecimento para suporte à tomada de decisão pelo usuário final. Para que esse conhecimento seja gerado de maneira correta, proporcionando informações úteis e que agreguem valor à tomada de decisão do usuário com relação ao gerenciamento da lavoura, processos que permitem análises mais complexas dos dados, como a descoberta de conhecimento em bancos de dados espaciais, devem ser levados em consideração. Processos desse tipo são subdivididos em diversas etapas, dentre elas a mineração de dados espaciais, onde estão inseridas as principais contribuições desta tese.

2.8 Considerações Finais

Neste capítulo, foram descritos os principais conceitos e definições a respeito de AP no contexto da agricultura, bem como as contribuições da área de Ciência da Computação que vem sendo aplicadas nesse tema, além de novos conceitos relacionados ao armazenamento e à recuperação dos dados que podem ser capturados, manipulados, distribuídos e compartilhados por meio de TICs. Os capítulos 3 e 4, a seguir, abordam a fundamentação teórica relacionada a

[5]FMIS: Do inglês *Farm Management Information Systems*

bancos de dados e mineração de dados espaciais, cujos conceitos e definições foram utilizados ao longo do desenvolvimento desta tese.

Capítulo 3

DADOS ESPACIAIS

3.1 Considerações Iniciais

Neste capítulo, são apresentadas as definições necessárias para a compreensão dos conceitos relacionados a dados espaciais presentes nesta tese. O capítulo está organizado da seguinte forma:

- A Seção 3.2 descreve as definições básicas a respeito de bancos de dados espaciais e suas implementações em Sistemas Gerenciadores de Bancos de Dados Espaciais (SGBDEs).
- A Seção 3.3 descreve as abordagens já estabelecidas para a criação de modelos conceituais específicos para dados espaciais.
- A Seção 3.4 apresenta os principais conceitos relacionados a Sistemas de Informações Geográficas (SIGs).
- A Seção 3.5 descreve os padrões conceituais estabelecidos pela OGC para armazenamento, recuperação, utilização e compartilhamento de informações espaciais, bem como informações a respeito de *frameworks*, aplicativos e servidores de mapas disponíveis para auxílio à construção de SIGs para a *Web*.

3.2 Bancos de Dados Espaciais

Os Sistemas Gerenciadores de Bancos de Dados (SGBDs) são programas de computador de propósito geral que auxiliam os processos de modelagem, carga, construção, manipulação, busca e compartilhamento de bancos de dados entre diversos usuários e aplicações (ELMASRI;

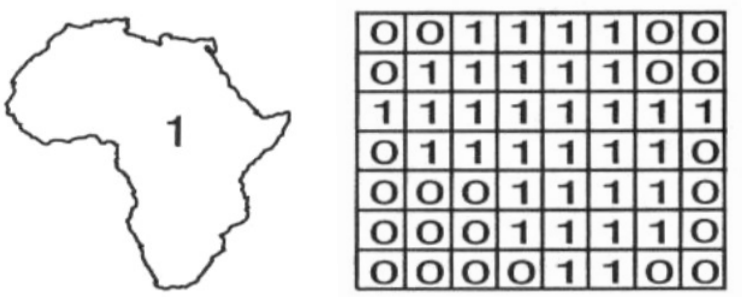
NAVATHE, 2011). Dentre as implementações de SGBDs mais conhecidas e utilizadas, destacam-se o PostgreSQL (POSTGRESQL, 2017) e o Oracle (ORACLE, 2017), que, dentre outras características, possuem mecanismos que possibilitam o desenvolvimento de extensões geográficas capazes de armazenar, recuperar e analisar dados espaciais de maneira diferenciada (CIFERRI, 1995; CASANOVA, 2005). A presença desses mecanismos nos permite classificá-los como SGBDs Espaciais (SGBDEs).

Os SGBDEs permitem a representação de dados espaciais no espaço bi ou tridimensional, com o intuito de manipular informações referentes a fenômenos do mundo real que possuem uma localização espacial vinculada. Em AP, os dados espaciais normalmente utilizam objetos bidimensionais no espaço $\mathbb{R}^2$, que é composto pelas coordenadas espaciais de latitude e longitude. Em alguns casos, dados de altitude podem ser coletados aproveitando-se o receptor de GPS acoplado a algum sensor de campo que realiza coletas de dados, fazendo com que os dados espaciais passem a utilizar objetos tridimensionais no espaço $\mathbb{R}^3$. Entretanto, mesmo nesses casos, a abordagem mais utilizada é a mesma aplicada a todas as outra variáveis obtidas em AP, onde os valores medidos são representados no banco de dados como componente do espaço de atributos. Consequentemente, podemos considerar qualquer SGBDE que possua a característica de armazenar dados espaciais, utilizando objetos bidimensionais no espaço $\mathbb{R}^2$ para representar a disposição espacial de um ou mais fenômenos naturais que ocorrem em uma lavoura, como apto a ser utilizado para armazenar dados obtidos em AP. Esses fenômenos, por sua vez, devem ter seus valores medidos armazenados em campos representados por tipos de dados convencionais (e.g., inteiro, decimal) vinculados a essa disposição espacial.

Os SGBDEs normalmente representam os dados espaciais no modelo vetorial por meio de pontos, linhas e polígonos (GÜTING, 1994; CLEMENTINI; DI FELICE, 1996; CIFERRI, 2002; SCHNEIDER; BEHR, 2006). Esses tipos de dados são definidos de maneira precisa no espaço com relação à sua localização, fronteira, interior e exterior (CARNIEL; CIFERRI; CIFERRI, 2016). Além disso, podem ser representados também em grupos de objetos, conhecidos como complexos, onde cada objeto pode representar um conjunto homogêneo de pontos, linhas e polígonos. No âmbito da AP, a representação por pontos é a mais utilizada, devido ao fato de a maioria dos dados coletados estarem associados a um único par de coordenadas em $\mathbb{R}^2$. Essa característica pode ser estendida, inclusive, aos dados matriciais, onde o ponto localizado no centro de cada pixel pode ser utilizado para representar espacialmente essas informações. As linhas normalmente são utilizadas para representar objetos ou recursos naturais presentes em uma propriedade, como canais de irrigação e rios. Já os polígonos, normalmente são utilizados para representar objetos espaciais como a área de uma propriedade e de seus talhões, UGDs, edificações, lagos ou até mesmo rios. Independentemente disso, um conjunto de dados espaciais

originalmente representado no formato vetorial pode possuir sua representação equivalente no formato matricial, e vice-versa. A Figura 3.1 exibe um conjunto de dados espaciais inicialmente representado por um polígono; e sua representação equivalente no formato matricial, onde os pixels que correspondem à borda e ao interior do polígono possuem o valor 1 associado, e os pixels externos ao polígono possuem o valor 0 associado .

Figura 3.1: Representação de um conjunto de dados espaciais no formato vetorial (a); e sua representação equivalente no formato matricial (b).



(a) (b)

Fonte: Adaptada de Davis (2001).

Devido à sua complexidade, os dados espaciais não possuem relação de ordem total, ou seja, não é possível ordenar um conjunto completo de dados espaciais em um espaço Euclidiano bidimensional ou tridimensional. Por conta disso, os dados espaciais não podem ser relacionados com a utilização de operadores comumente utilizados em dados convencionais, como $<, >, \leq$ e $\geq$. De uma maneira geral, esses dados são manipulados por operadores de conjunto (e.g. união, interseção e diferença), topológicos (e.g. igual, disjunto, toca, contém, cobre, intersecta, está contido e coberto por) e geométricos (e.g. distância e área), também providos pelos SGBDEs. Os predicados topológicos são amplamente investigados na literatura, pois são muito utilizados em consultas espaciais para representar intersecções entre bordas, interiores e exteriores de dois objetos espaciais (CLEMENTINI; SHARMA; EGENHOFER, 1994; CLEMENTINI; DI FELICE, 1996; SCHNEIDER; BEHR, 2006). Em AP, esses predicados podem ser utilizados, por exemplo, em consultas espaciais para encontrar pontos de coleta que determinaram o delineamento de uma UGD.

3.3 Modelagem de Dados Espaciais

A modelagem conceitual é de extrema importância para o sucesso no desenvolvimento de uma aplicação que manipula informações armazenadas em um banco de dados (ELMASRI; NA-

VATHE, 2011). O Modelo Entidade Relacionamento (MER) (CHEN, 1976) é o mais conhecido e utilizado para modelagens desse tipo. Sua notação gráfica, conhecida como Diagrama Entidade Relacionamento (DER), é adotada por diversos aplicativos específicos para modelagem conceitual em bancos de dados. Metodologias baseadas em Orientação a Objetos (OO), como a UML (*Unified Modeling Language*) (BOOCH; RUMBAUGH; JACOBSON, 1997), são também muito populares, tanto para a modelagem de bancos de dados, quanto para a modelagem de software. Algumas alternativas baseadas no MER ou em OO podem ser encontradas na literatura, com o intuito de adaptar o uso dessas metodologias para bancos de dados espaciais (SHEKHAR, 1997; HADZILACOS; TRYFONA, 1997; BÉDARD, 1996, 1999; KOSTERS; PAGEL; SIX, 1996; PARENT, 1998; LISBOA-FILHO; IOCHPE, 1999; BORGES; DAVIS; LAENDER, 2001).

O Modelo de Dados Orientado a Objetos para Aplicações Geográficas (OMT-G) (BORGES; DAVIS; LAENDER, 2001), é uma especialização da abordagem de modelagem de software conhecida como OMT (*Object-Modeling Technique*) (RUMBAUGH, 1991). Essa solução, concebida com o objetivo de aproximar o modelo real ao de representação efetivo para dados espaciais, utiliza primitivas geográficas que permitem a modelagem da geometria e da topologia dos dados. Para atender a esse objetivo, o OMT-G permite a criação de três tipos de diagramas: diagrama de classes, que define a representação gráfica para a estrutura de classes, contendo objetos espaciais e convencionais e relacionamentos entre eles; diagrama de transformação, que define as operações de transformações espaciais que podem ocorrer entre os objetos definidos no diagrama de classes; e diagrama de apresentação, que permite a explanação das necessidades do usuário com relação às alternativas de exibição de cada objeto espacial definido.

3.4 Sistemas de Informações Geográficas (SIGs)

Os Sistemas de Informações Geográficas (SIGs) são programas de computador utilizados para manipular dados que representam objetos e fenômenos onde a localização geográfica é uma característica intrínseca indispensável à análise da informação (ARONOFF, 1989). Os SIGs são compostos por diferentes módulos integrados que permitem a centralização de dados espaciais oriundos de fontes heterogêneas de forma transparente ao usuário final, podendo ser utilizados nas mais variadas áreas de aplicação. As principais funcionalidades de um SIG estão relacionadas à entrada, integração, processamento, visualização, plotagem, armazenamento e recuperação de dados espaciais (CIFERRI, 1995; CÂMARA, 1996), atuando de forma integrada aos SGBDEs.

Devido ao aumento da disponibilidade de dados espaciais, proporcionado principalmente

pela disponibilização do sinal de GPS em meados dos anos 90 e pela grande quantidade de dispositivos capazes de gerá-los, o mercado para o desenvolvimento de SIGs também emergiu, agregando ao longo dos anos diferentes funcionalidades de geoprocessamento para as mais variadas aplicações.

Dentre os SIGs disponíveis no mercado de *software* proprietário, a plataforma ArcGIS (ARCGIS, 2017) deve ser destacada. Criada em 2001 pela empresa ESRI[1] e amplamente disseminada nas mais variadas comunidades de aplicações que utilizam geoprocessamento, essa plataforma integra diferentes extensões que permitem aos seus usuários realizar análises tridimensionais e geoestatísticas em dados espaciais, além das diversas funções de geoprocessamento comumente disponíveis em SIGs. Suas versões mais atuais já envolvem os novos conceitos para manipulação de dados espaciais disponibilizados em nuvem.

Apesar das diversas funcionalidades disponibilizadas pela plataforma ArcGIS, o alto custo para aquisição das licenças acaba limitando o seu uso. Quando existe essa limitação por parte do usuário final, o mercado de *software* livre e de distribuição gratuita deve ser considerado. Nesse contexto, o SIG QuantumGIS (QGIS) (QGIS, 2017) surge como a ferramenta mais promissora, devido à diversidade de complementos disponíveis para as mais variadas atividades de geoprocessamento. Por meio desses complementos é possível criar, por exemplo, fluxos de execução para processos que envolvem mineração de dados considerando diferentes algoritmos externos, fazendo com que o controle das operações seja todo centralizado na interface do SIG. Por ser um *software* de código aberto, o QGIS permite que os seus próprios usuários, com um conhecimento básico em linguagens de programação de computadores, implementem as suas próprias funcionalidades e algoritmos. Além disso, pode ser integrado com outros SIGs e outras bibliotecas que disponibilizam funcionalidades relacionadas à geoestatística e mineração de dados espaciais, como o ambiente *R* (R, 2017).

No contexto da AP, por se tratar de um sistema baseado em análise de dados espaciais, os SIGs são extremamente importantes. A principal dificuldade dos usuários finais de AP em utilizar SIGs genéricos como o ArcGIS e o QGIS está na conversão dos dados originários de campo, muitas vezes em formato proprietário do fabricante do sensor utilizado, dificuldade esta que também é observada na geração de mapas de recomendação que possam ser interpretados por equipamentos de automação agrícola de diferentes fabricantes (MOLIN; AMARAL; COLAÇO, 2015). Neste caso, existem SIGs específicos para AP e que já possuem diversas interfaces de conversão desses dados para os mais variados fabricantes. Entretanto, com a crescente disseminação dos dados espaciais na *Web* e o surgimento de iniciativas que permitem a interoperabili-

[1] Environmental Systems Research Institute, Redlands, Calinfórnia, EUA, http://www.esri.com

dade entre sistemas, como o padrão ISOBUS e os padrões OGC, essas dificuldades tendem a diminuir.

3.5 Dados Espaciais via Web

A disponibilização dos SIGs e de serviços distribuídos de informações espaciais na *Web* impulsionou o desenvolvimento de novas tecnologias para compartilhamento de informações espaciais em forma de mapa, permitindo a criação de aplicações interativas com diversas funcionalidades para o usuário final (MITCHELL, 2005). Essas novas tecnologias impulsionaram o uso dos SIGs por agregarem novos requisitos, como a disseminação, o acesso, a exploração, a geovisualização e o processamento de dados espaciais (DRAGIĆEVIĆ, 2004).

Um dos objetivos desta tese foi a construção de um sistema protótipo, baseado em SIG, padrões OGC e outros softwares livres e de distribuição gratuita, para facilitar a execução e visualização dos resultados fornecidos pelas soluções desenvolvidas por parte dos usuários especialistas.

3.5.1 Open Geospatial Consortium (OGC)

A OGC (Open Geospatial Consortium) é uma organização voluntária internacional sem fins lucrativos, criada para o desenvolvimento de padrões de consenso para dados geoespaciais. Possui diversos membros distribuídos entre a indústria, governo, institutos de pesquisa e universidades, com a missão de evidenciar o desenvolvimento, a promoção e a harmonização de padrões geoespaciais abertos (OGC, 2017). A seguir, são descritas as especificações da OGC utilizadas durante o desenvolvimento do sistema protótipo desta tese.

3.5.2 Armazenamento e Manipulação de Dados Espaciais via *Web*: GML e o tipo de dados *Geometry*

A GML (*Geography Markup Language*) é uma gramática XML (*eXtensible Markup Language*) de modelagem para sistemas geográficos que atua no transporte e armazenamento de informações e no intercâmbio de transações geográficas na *Web* (PERCIVALL, 2003; COX, 2002). Sua principal função é descrever abstrações de fenômenos do mundo real associados com uma localização relativa ao globo terrestre, definidas pela OGC como *features* geográficas (PERCIVALL, 2003). As características de portabilidade, sintaxe simplificada, extensibilidade e flexibilidade (PERCIVALL, 2003; COX, 2002) podem ser consideradas como fatores determinantes

para que a GML seja utilizada como base para a criação de linguagens de sistemas geográficos para aplicações específicas. A Figura 3.2 exibe representações em GML para armazenamento e transação de *features* geográficas.

Figura 3.2: Representações em GML para (a) armazenamento e (b) transação de *features* geográficas.

```
<gml:featureMember>
  <ogr:gml fid="gml.0">
    <ogr:geometryProperty>
       <gml:Point srsName="EPSG:32723">
              <gml:coordinates>
                     299559.9861,7505209.25499
              </gml:coordinates>
      </gml:Point>
    </ogr:geometryProperty>
    <ogr:ce30>2.2</ogr:ce30>
    <ogr:ce90>1.5</ogr:ce90>
  </ogr:gml>
</gml:featureMember>
```

(a)

```
<wfs:Transaction service="WFS" version="1.0.0"
  <wfs:Delete typeName="municipios">
    <ogc:Filter>
      <ogc:PropertyIsEqualTo>
        <ogc:PropertyName>nome</ogc:PropertyName>
        <ogc:Literal>Campinas</ogc:Literal>
      </ogc:PropertyIsEqualTo>
    </ogc:Filter>
  </wfs:Delete>
</wfs:Transaction>
```

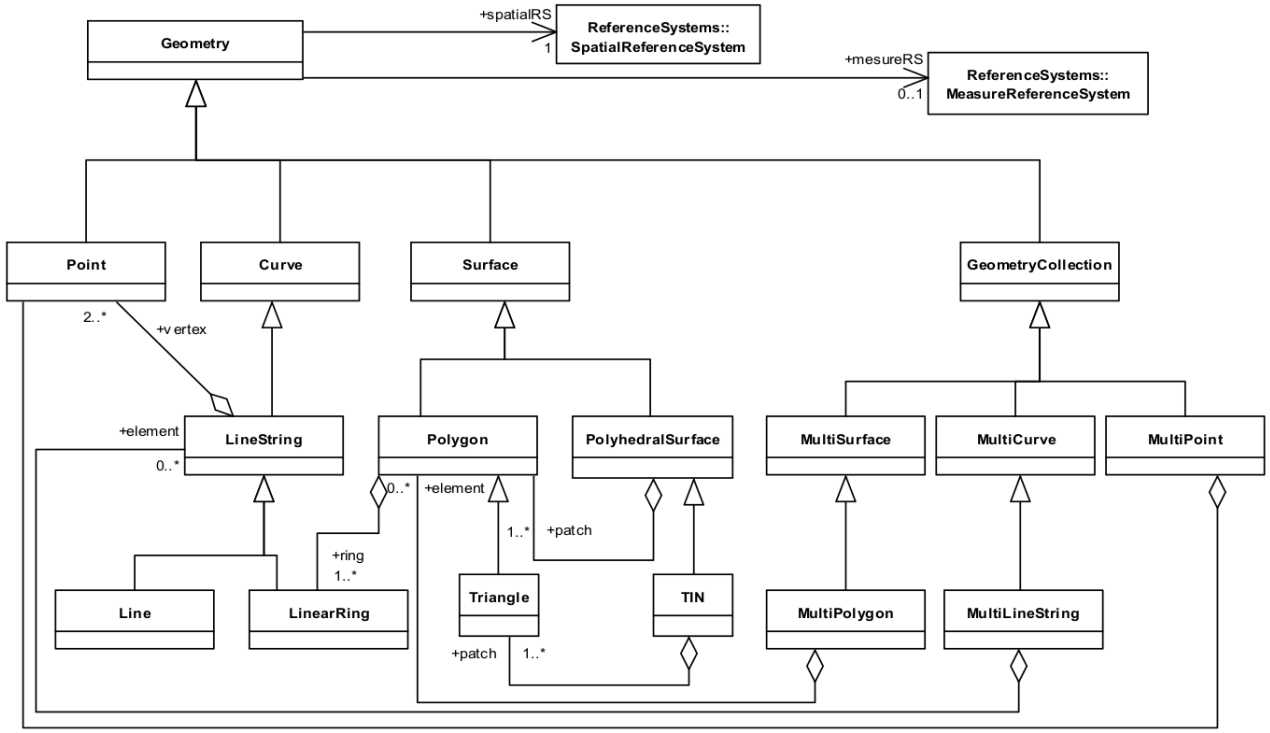
(b)

Fonte: Elaborada pelo autor.

Um atributo importante a ser considerado pela GML para descrição de uma *feature* geográfica é o Sistema de Referência Espacial (SRE). Um SRE define uma projeção específica de mapeamento para um conjunto de dados no globo terrestre, bem como modelos de transformação de coordenadas espaciais para outros SREs. Cada SRE é composto por um sistema de coordenadas e por um modelo matemático para representação na superfície terrestre ao nível do mar, conhecido como *datum* (PERCIVALL, 2003; LOTT, 2004). A definição correta do SRE a ser utilizado é de extrema importância nas coletas de dados espaciais em AP. Se forem coletados, por exemplo, dois conjuntos de dados espaciais para uma mesma área utilizando SREs distintos, será necessário que esses dados sejam transformados em um único SRE, permitindo a geração de uma grade única de coordenadas espaciais para posterior análise dos dados.

Em se tratando do armazenamento em bancos de dados, as *features* geográficas também podem seguir as especificações criadas pela OGC com base no tipo de dados *Geometry* (HERRING, 2001). Essas especificações provêm os meios para descrições quantitativas - a partir de coordenadas geográficas e funções matemáticas que levam em consideração o SRE utilizado - das características espaciais das *features*, como dimensão, posição, tamanho, forma e orientação. Além disso, podem ser implementadas pelos SGBDEs em conjunto com os tipos de dados nativos, possibilitando que as *features* geográficas armazenadas sejam manipuladas pelas especificações da GML. A Figura 3.3 exibe a hierarquia de classes para o tipo de dados *Geometry*.

Figura 3.3: Hierarquia de classes para o tipo de dados Geometry.



Fonte: Herring (2001).

Por meio da hierarquia exibida na Figura 3.3, é possível observar que cada dado espacial pertencente à classe *Geometry* deve estar obrigatoriamente associado a um único SRE, que nesse caso está representado como um objeto da subclasse *ReferenceSystems::SpatialReferenceSystem*. A hierarquia ainda propõe especializações e agregações realizadas a partir da classe *Geometry* para obtenção das subclasses que serão utilizadas no âmbito desta tese, tais como: pontos, representados por objetos da subclasse *Point*; linhas, representadas por objetos da subclasse *LineString*; polígonos, representados por objetos da subclasse *Polygon*; e coleções de objetos, representados pela subclasse *GeometryCollection*. As subclasses *MultiPolygon* e *MultiPoint*, que se constituem de especializações da subclasse *GeometryCollection*, podem ser utilizadas, por exemplo, para representar uma CGD formada por mais de uma UGD utilizando um único objeto espacial.

3.5.3 Especificações OGC para Serviços *Web*

A especificação OGC *Web Map Service* (WMS) provê um serviço que produz mapas espacialmente referenciados dinamicamente a partir de informações geográficas (PERCIVALL, 2003; BEAUJARDIERE, 2006). Essa especificação disponibiliza operações que permitem a recuperação

e descrição dos mapas em si e das *features* geográficas exibidas nos mesmos. Na interface cliente de requisição da especificação WMS, devem ser definidas as *layers* correspondentes aos mapas e a área de interesse a ser processada. Como resposta, o servidor deverá retornar uma ou mais imagens desses mapas em formatos como JPEG e PNG, permitindo a sua exibição por meio de *browsers*.

A especificação OGC *Web Feature Service* (WFS) provê serviços para acesso e edição de dados geoespaciais, definidos por um conjunto de operações que permitem aos usuários clientes manipularem *features* geográficas disponibilizadas por servidores *Web* (PERCIVALL, 2003; VRETANOS, 2005). Ao contrário da especificação WMS, que retorna apenas imagens estáticas dos mapas solicitados, a especificação WFS oferece acesso refinado à informação geográfica em nível de *feature* e de suas propriedades. Além disso, os usuários podem manipular apenas os dados de seu interesse, evitando assim maior tráfego na rede. Nas operações de busca e recuperação de dados, devem ser definidas as *layers*, área de interesse, filtros que devem ser aplicados no processamento e o formato de saída dos dados. Como resposta, o servidor deverá retornar os dados no formato especificado. Já nas operações transacionais, devem ser incluídas as informações relacionadas ao espaço de atributos e de coordenadas da nova *feature* geográfica, no caso de inclusão; e o identificador da *feature* geográfica a ser excluída ou alterada, nos casos de exclusão ou alteração, respectivamente. Nesse caso a resposta do servidor será um documento indicando o sucesso ou o erro ocorrido durante o processamento da requisição.

A especificação OGC *Web Processing Service* (WPS) provê uma interface padronizada para facilitar a publicação de processos de geoprocessamento, tais como algoritmos, cálculos ou modelos que operam sobre os dados, tornando-os disponíveis ao usuário final (PERCIVALL, 2003; SCHUT; WHITESIDE, 2007). Os processos devem ser executados por requisições ao servidor, fornecendo-se os parâmetros necessários obtidos a partir de uma requisição para sua descrição. Após executar um processo, a interface de resposta da WPS fornecerá um documento indicando o sucesso ou o erro ocorrido durante o processamento da requisição.

Finalizando, a especificação da arquitetura de serviços da OGC permite a definição de cadeias de serviços em séries dependentes, com o intuito de se resolver tarefas mais longas e permitir a combinação de dados e serviços em formatos que não foram predefinidos pelos provedores (PERCIVALL, 2002, 2003). Com isso, podem ser criadas ferramentas importantes para tarefas de apoio à tomada de decisão, como mapas com valor agregado.

3.5.4 Sistemas de Informações Geográficas para a *Web*

Os sítios *Web Mapping* ou *WebGIS* têm se tornado cada vez mais populares, permitindo a criação de aplicações interativas que proporcionam ao usuário executar operações básicas nos mapas, como deslocamento, aproximação e medidas de área e distância (MITCHELL, 2005). Essa popularidade fez com que surgissem diversas plataformas e bibliotecas para auxiliar em sua construção, bem como servidores de mapas capazes de acessar, manipular e entregar dados espaciais seguindo os padrões da OGC.

Com relação aos servidores de mapas, diversas implementações de software livre e de distribuição gratuita estão disponíveis, como o MapServer (MAPSERVER, 2017), GeoServer (GEOSERVER, 2017) e Deegree (DEEGREE, 2017). Já para a construção das aplicações, destaca-se a biblioteca OpenLayers (OPENLAYERS, 2017), também de código aberto e disponibilizada gratuitamente. Entretanto, para soluções mais elaboradas, é recomendada a utilização de plataformas que disponibilizam os diversos serviços e operações realizadas nos dados espaciais em diferentes camadas internamente acopladas, como a OpenGeo(OPENGEO, 2017). Essa plataforma, que possui versões de distribuição gratuita, fornece uma solução integrada que define camadas de banco de dados, servidores de mapas (incluindo operações de *cache*), bibliotecas para *WebGIS* e clientes *Web* e *desktop* para visualização e interatividade.

3.6 Considerações Finais

As definições e especificações apresentadas neste capítulo foram de grande importância para o desenvolvimento das atividades relacionadas à modelagem, armazenamento e recuperação de dados espaciais realizadas pelo sistema protótipo. O Capítulo 4, a seguir, é composto por definições relacionadas à teoria da descoberta de conhecimento em bancos de dados espaciais, auxiliando diretamente na obtenção das principais contribuições científicas desta tese.

Capítulo 4

DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS ESPACIAIS

4.1 Considerações Iniciais

Neste capítulo, são apresentadas as definições necessárias para a compreensão dos conceitos relacionados ao processo de descoberta de conhecimento em bancos de dados utilizados nesta tese, retratando as diferenças e adaptações necessárias que devem ser realizadas com relação a esse mesmo processo, quando são utilizados dados espaciais. O capítulo está organizado da seguinte forma:

- A Seção 4.2 descreve os conceitos relacionados à descoberta de conhecimento em bancos de dados convencionais, envolvendo as principais abordagens e algoritmos utilizados na tarefa de agrupamento de dados convencionais, além das principais medidas utilizadas para avaliar a qualidade dos resultados obtidos.
- A Seção 4.3 descreve os conceitos específicos relacionados à descoberta de conhecimento em bancos de dados espaciais, retratando as dificuldades encontradas e adaptações que devem ser realizadas quando são utilizados dados espaciais, em especial àquelas que estão relacionadas com o processo de delineamento de UGDs em AP.

4.2 Descoberta de Conhecimento em Bancos de Dados Convencionais

A quantidade de dados produzida no mundo tem crescido substancialmente ano após ano. Diante desse cenário, o que pode ser notado é que, à medida que o volume de dados aumenta,

diminui a capacidade das pessoas em gerenciá-los e interpretá-los, fazendo com que as técnicas de mineração de dados se tornem cada vez mais importantes e necessárias (WITTEN; FRANK; HALL, 2011). Essas técnicas permitem a resolução de problemas analisando dados armazenados em banco de dados, e são essencialmente definidas como processos de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis intrínsecos aos dados (FAYYAD, 1996). Esses processos podem ser executados de maneira automática ou semiautomática, onde os padrões descobertos devem ser significativos na resolução do problema em questão.

A mineração de dados compõe uma das etapas do processo de descoberta de conhecimento em bancos de dados, tradicionalmente conhecido como KDD[1] (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; WEISS; INDURKHYA, 1998). De uma maneira geral, a etapa de mineração de dados deve ser executada entre outras duas importantes etapas do KDD: o pré-processamento, onde os dados são preparados e selecionados; e o pós-processamento, onde os padrões obtidos devem ser avaliados e validados. Duas etapas, relacionadas à identificação do problema e seleção dos dados utilizando o conhecimento do especialista podem ser consideradas antes do pré-processamento; e outra etapa, relacionada à utilização do conhecimento, pode ser considerada após o pós-processamento, tornando o processo de KDD mais completo (REZENDE, 2003).

A etapa de identificação do problema está relacionada ao conhecimento do domínio da aplicação. No âmbito da AP, é de extrema importância a participação nessa etapa de profissionais especialistas da área agrícola que sejam conhecedores da área de cultivo em questão. Em seguida, uma etapa de seleção de dados pode ser considerada, onde são selecionadas pelo usuário especialista quais serão as fontes de dados brutos que proverão as informações a serem utilizadas nas etapas posteriores. A etapa de pré-processamento compreende as transformações que devem ser realizadas nos dados brutos. Atividades como padronização e limpeza dos dados, seleção e redução de atributos podem ser realizadas nessa etapa.

A etapa de mineração de dados deverá possibilitar ao usuário atingir os objetivos definidos na etapa de identificação do problema. Essa etapa compreende um processo iterativo, onde podem ser experimentadas diferentes soluções para possibilitar uma maior eficácia no cumprimento desses objetivos. Para tanto, a escolha da tarefa de mineração de dados a ser utilizada faz-se necessária. Tarefas de classificação, que realizam a predição das classes de novas amostras a partir de um conjunto de amostras pré-classificadas; ou de agrupamento, que tem o objetivo de agrupar amostras que não possuem classificação prévia, como é o caso do delineamento de UGDs em AP, são exemplos de tarefas de mineração usadas nessa etapa. De-

[1] Do inglês *Knowledge Discovery in Databases*

pendendo da tarefa escolhida, deverão ser selecionados algoritmos que utilizam técnicas de aprendizado de máquina para viabilizar a sua execução. Essas técnicas compõem uma área de estudo em inteligência artificial que tem como objetivo criar métodos computacionais capazes de resolver problemas de descrição e predição, com base em regras e padrões adquiridos a partir dos dados (MITCHELL, 1997; JAIN; DUBES et al., 1988). Assim, tarefas de classificação normalmente utilizam métodos classificados como aprendizado de máquina supervisionado; e tarefas de agrupamento, métodos classificados como aprendizado de máquina não supervisionado. Existe também a classe dos métodos semissupervisionados, que podem ser utilizados quando o conjunto de dados inicial possui apenas algumas amostras classificadas.

O pós-processamento tem o objetivo principal de diminuir a quantidade de padrões descobertos na etapa anterior, de forma a encontrar apenas aqueles que são relevantes para as análises que serão realizadas pelos especialistas do domínio (REZENDE, 2003). Considerando o contexto do delineamento de UGDs em AP, essa etapa pode estar associada, por exemplo, à utilização de métricas de qualidade que auxiliem na escolha pelo especialista de quais mapas de UGDs podem ser utilizados, dentre as diversas soluções que podem ser disponibilizadas pela etapa de mineração de dados.

Finalmente, na etapa de utilização do conhecimento, ainda podem ser utilizadas as métricas da etapa anterior para avaliar, de maneira relativa, soluções obtidas variando-se os parâmetros de uma abordagem de agrupamento, ou até mesmo comparar as soluções obtidas utilizando-se diferentes abordagens. Além disso, podem ser utilizados métodos que permitem comparar, por exemplo, uma solução obtida a partir de uma abordagem de agrupamento com relação a uma solução de referência. Em AP, essa solução de referência pode ser, por exemplo, um mapa de UGDs delineado manualmente pelo usuário final (ZHANG, 2010).

O delineamento de UGDs em AP pode ser considerado como um processo completo de KDD, desde que sejam considerados alguns requisitos relacionados a dados espaciais. De uma maneira geral, a principal etapa a ser considerada nesse contexto por esse processo é a mineração de dados, onde foi verificada a necessidade de desenvolvimento de novas abordagens específicas, capazes de extrair padrões de conhecimento a partir de dados espaciais com atributos oriundos de diferentes fontes e sem classificação prévia. Consequentemente, nas etapas de pós-processamento e utilização do conhecimento, também foi verificada a necessidade de desenvolvimento de novas métricas capazes de avaliar a qualidade dos agrupamentos obtidos na etapa de mineração de dados, com o intuito de auxiliar na tomada de decisão do usuário considerando de maneira diferenciada os espaços de atributos e de coordenadas que compõem os conjuntos de dados espaciais utilizados. Essas necessidades deram origem às principais con-

tribuições obtidas nesta tese.

Na sequência, são apresentados conceitos gerais e abordagens que envolvem o processo de KDD em bancos de dados convencionais, ou seja, que possuem apenas atributos contendo tipos primitivos (e.g., inteiro, decimal), com ênfase para técnicas bastante utilizadas no âmbito desta tese, tais como: seleção de atributos e redução de dimensionalidade, na etapa de pré-processamento; algoritmos de agrupamento de dados, na etapa de mineração de dados; e métricas utilizadas para validar e comparar agrupamentos obtidos na etapa anterior, conhecidas como critérios de validação. Em geral, esses conceitos e abordagens são a base para a construção de técnicas específicas para dados espaciais.

4.2.1 Seleção de Atributos Utilizando o Coeficiente de Correlação de Pearson

Conforme já descrito no Capítulo 2, a utilização de atributos obtidos com qualidade em campo, capazes de retratar da maneira mais fiel possível as caraterísticas do solo e da lavoura, são muito importantes para que sejam obtidos mapas de UGDs confiáveis e consistentes ao longo do tempo. Nesse contexto, além da seleção natural de atributos que não são influenciados pela ação humana, técnicas capazes de verificar a correlação existente entre variáveis podem auxiliar o usuário final em selecionar atributos relevantes para a resolução do seu problema, com o intuito de proporcionar agrupamentos de qualidade e úteis para a aplicação.

Uma das técnicas mais conhecidas e utilizadas para verificar a correlação existente entre dois atributos é o coeficiente de correlação de *Pearson* (PEARSON, 1895; BENESTY, 2009). Esse coeficiente verifica a compatibilidade linear entre dois vetores de dados representando dois atributos distintos, tentando identificar tendências sem considerar os fatores de dimensão ou escala, desprezando-se a média e a variância (Equação 4.1).

$$\rho(p,q) = p' \bullet q' \tag{4.1a}$$

$$p'_k = (p_k - \mu_p)/\sigma_p \tag{4.1b}$$

$$q'_k = (q_k - \mu_q)/\sigma_q \tag{4.1c}$$

A Equação 4.1a define a correlação de Pearson (ρ) entre dois atributos p e q como sendo o produto vetorial entre os seus respectivos vetores de amostras p' e q'. Em ambos, para cada amostra k, devem ser desprezados os valores da média (μ) e variância (σ), conforme exibido nas equações 4.1b e 4.1c. A correlação de Pearson fornece valores entre -1 e 1 que indicam o grau

de correlação entre dois atributos, interpretados da seguinte maneira: valores de ρ próximos de 0 indicam uma correlação neutra ou não existente; valores de ρ próximos de 1 indicam alta correlação linear; e valores de ρ próximos de -1 também indicam alta correlação linear, porém em sentidos opostos.

No contexto do delineamento de UGDs em AP, diversos atributos capazes de identificar diferentes características do solo e da cultura podem estar disponíveis ao usuário final, conforme já exemplificado no Capítulo 2. Entretanto, em diversas situações, é desejável que apenas alguns deles sejam selecionados para essa tarefa, por conta de diversos motivos, tais como: a redução do custo computacional para execução das abordagens de agrupamento; a exclusão de atributos que não são relevantes com relação à variabilidade espacial; e até mesmo a ausência de qualidade por conta de medidas não confiáveis obtidas em campo. Nesses casos, o coeficiente de correlação de Pearson pode ajudar a privilegiar a seleção de atributos que se correlacionam bem com a maioria dos outros atributos do conjunto de dados.

Se considerarmos uma matriz de correlações C com dimensões $n \times n$, onde n é a quantidade total de atributos, cada célula dessa matriz deve conter o valor de ρ calculado entre cada par atributos, segundo a equação 4.1a. A Tabela 4.1 mostra um exemplo de uma matriz de correlações C, em formato de tabela, para um conjunto hipotético de 4 atributos.

Tabela 4.1: Exemplo de matriz de correlações C para 4 atributos hipotéticos.

-	a_1	a_2	a_3	a_4
a_1	1	0,8	0,3	-0,6
a_2	0,8	1	-0,1	0,9
a_3	0,3	-0,1	1	0,1
a_4	-0,6	0,9	0,1	1

De acordo com a Tabela 4.1, células contendo valores próximos a 1 e -1 mostram que o par de atributos correspondente a esse valor possui boa correlação linear considerando, respectivamente, o mesmo sentido ou sentidos opostos. Desse modo, pode-se considerar, por exemplo, que valores acima de 0,5 e abaixo de -0,5 indicam atributos bem correlacionados. Levando-se em consideração essas definições, no caso do exemplo da Tabela 4.1, é fácil verificar que os atributos hipotéticos a_1, a_2 e a_4 se correlacionam bem entre si; e o atributo a_3 não se correlaciona bem com nenhum dos outros atributos. Desse modo, para esse exemplo, em uma necessidade de seleção de um conjunto reduzido de atributos, o atributo a_3 poderia ser desconsiderado. Ainda vale ressaltar que na matriz C representada pela Tabela 4.1, os valores da diagonal principal são sempre 1, pois consideram a correlação de um atributo com ele mesmo. Além disso, essa

mesma matriz poderia ser escrita em formato triangular superior ou inferior, pois a ordem dos fatores não altera os valores obtidos no cálculo de ρ entre dois atributos.

Uma metodologia similar à descrita no exemplo citado acima foi utilizada para a seleção de atributos para conjuntos de dados utilizados nos experimentos desta tese. Entretanto, se o conjunto de dados possuir uma quantidade muito grande de atributos bem correlacionados entre si, pode-se optar por utilizar critérios mais rígidos para seleção dos atributos, permitindo, por exemplo, a utilização de atributos que se correlacionam com a maioria dos outros atributos do conjunto de dados com valores de ρ acima de 0,8 ou abaixo de -0,8.

4.2.2 Redução de Dimensionalidade Utilizando a Análise de Componentes Principais (PCA)

Além da seleção de atributos utilizando o coeficiente de correlação de Pearson, outra maneira de reduzir a quantidade de atributos na etapa de pré-processamento de um processo de KDD é por meio de técnicas que proporcionam a redução da dimensionalidade. Uma técnica estatística bastante conhecida e utilizada para esse fim é a PCA (*Principal Component Analysis*) (HOTELLING, 1933). Essa técnica utiliza transformações ortogonais para converter conjuntos de amostras de atributos que são potencialmente correlacionados em um novo conjunto de atributos que não são linearmente correlacionados, chamados de componentes principais. Nesse contexto, o primeiro componente principal está associado ao maior sinal de variabilidade considerando o conjunto de dados completo e o último componente em geral indica ruídos ou variabilidade falsa. A quantidade de componentes principais é geralmente menor do que a quantidade de atributos do conjunto de dados original, o que proporciona uma real redução de dimensionalidade.

Porém, a redução de dimensionalidade geralmente proporciona uma perda da qualidade da informação. Neste sentido, a técnica da PCA procura diminuir os efeitos dessa perda, separando informações realmente relevantes e que contribuem para a variabilidade do conjunto de dados das informações que podem ser consideradas ruidosas. Apesar de a PCA ser muito utilizada em AP, questões relacionadas com a diferenciação entre os espaços de atributos e coordenadas presentes nos conjuntos de dados espaciais precisam ser tratadas, para que a redução de dimensionalidade não prejudique a visualização desses dados em forma de mapa, como é o caso do delineamento de UGDs. Desse modo, trabalhos como os desenvolvidos por Córdoba (2013) e Peeters (2015), que se utilizam, respectivamente, de técnicas como a MULTISPATI-PCA e a estatística espacial Gi* (ORD; GETIS, 1995), realizam etapas de pré-processamento, como a seleção de atributos, já considerando as relações espaciais presentes entre eles.

4.2.3 Agrupamento de Dados Convencionais

O principal objetivo das abordagens de agrupamento é agrupar amostras de maneira natural, fazendo com que aquelas menos dissimilares com relação às características impostas pelo domínio em questão sejam alocadas em um mesmo grupo, e aquelas mais dissimilares sejam alocadas em grupos distintos. Medidas algébricas conhecidas, como as distâncias Euclidiana, Diagonal e Mahalanobis para dados numéricos; e abordagens de correspondência simples para dados nominais, são normalmente utilizadas para determinar a dissimilaridade entre as amostras. Existem diferentes abordagens que podem ser utilizadas para o agrupamento dados e a sua escolha está fortemente relacionada às particularidades dos dados da aplicação e a disponibilidade de cada uma delas para o usuário final (WITTEN; FRANK; HALL, 2011).

Quando os dados referentes ao problema a ser resolvido são de fácil separação em grupos distintos, possibilitando o uso de abordagens conhecidas como de particionamento sem sobreposição, o algoritmo mais usado é o *k-means* (MACQUEEN et al., 1967). Considerado um dos dez algoritmos mais influentes em mineração de dados (WU, 2007), o *k-means* é um método iterativo bastante simples, e que deve ser iniciado com a escolha, pelo usuário final, do valor para o parâmetro que define a quantidade de grupos desejada (parâmetro k). Dessa forma, k amostras são escolhidas, normalmente de maneira aleatória dentro do conjunto de dados, para representar os centroides iniciais representantes de cada grupo. Em cada passo do algoritmo, cada amostra é associada ao grupo menos dissimilar, considerando a distância da amostra com relação ao centroide do grupo. Ao final de cada iteração, a média das amostras de um grupo determina o seu novo centroide. Se os centroides não são alterados com relação à iteração anterior, o algoritmo converge e o agrupamento final é obtido. Desse modo, o *k-means* possui como objetivo principal minimizar a soma das distâncias ao quadrado de cada amostra com relação ao centroide do grupo ao qual foi associada, também conhecida como soma dos erros quadráticos ou função J (Equação 4.2).

$$J = \sum_{i=1}^{k} \sum_{x_j \in G_i} d\left(x_j, \bar{x_i}\right)^2 \tag{4.2}$$

Na Equação 4.2, k define a quantidade total de grupos do agrupamento, x_j é uma amostra qualquer pertencente ao grupo G_i e $\bar{x_i}$ corresponde ao centroide representante desse mesmo grupo.

Apesar de sua simplicidade, eficiência e eficácia, os resultados obtidos pelo *k-means* são altamente influenciados pela escolha aleatória dos centroides iniciais dos grupos, não fornecendo garantias em se atingir um resultado que possa ser considerado como ótimo global. De qualquer

maneira, os resultados obtidos podem se tornar mais confiáveis se escolhermos, por exemplo, os centroides iniciais com base em distribuições de probabilidade uniformes (WITTEN; FRANK; HALL, 2011). Variações do *k-means* (KAUFMAN; ROUSSEEUW, 1987) podem melhorar a eficácia do resultado final, evitando, por exemplo, a influência de valores extremos, conhecidos como *outliers*. Apesar de possuir uma complexidade computacional baixa ($\mathcal{O}(n)$), o *k-means* também possui variações desenvolvidas para melhorar a eficiência na sua execução e que podem ser utilizadas, por exemplo, quando o conjunto de dados possui grandes quantidades de amostras e atributos (LLOYD, 1982; ALSABTI; RANKA; SINGH, 1997; KANUNGO, 2002).

Já as abordagens hierárquicas permitem a obtenção recursiva de possíveis agrupamentos particionais aninhados com diferentes quantidades de grupos para a resolução de um problema (JAIN; DUBES et al., 1988). Essa hierarquia pode ser visualizada por meio de uma árvore binária, conhecida como dendrograma. Para essa abordagem, duas estratégias distintas podem ser utilizadas: a aglomerativa (*bottom-up*), que considera inicialmente cada amostra do conjunto de dados como um grupo, fundindo pares de grupos em cada nível da hierarquia; e a divisiva (*top-down*), que considera inicialmente todos as amostras pertencentes a um único grupo, dividindo-as em grupos menores em cada nível da hierarquia (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A estratégia aglomerativa é a mais utilizada e explorada na literatura, principalmente por conta da dificuldade em se encontrar os pontos de divisão dos grupos exigidos na estratégia divisiva, podendo inviabilizar o seu uso em aplicações como o delineamento de UGDs em AP (RUSS, 2012). Considerando a estratégia aglomerativa, em cada nível do dendrograma devem ser fundidos sempre os dois *grupos* menos dissimilares, até que se atinja o nível final com um único grupo. Como cada grupo é composto por uma ou mais amostras, a medida de dissimilaridade a ser utilizada em cada iteração, normalmente baseada na distância Euclidiana, é o principal diferencial entre os algoritmos hierárquicos aglomerativos tradicionais disponíveis na literatura, descritos a seguir.

O algoritmo *single-linkage* (FLOREK, 1951; SNEATH, 1957) calcula a dissimilaridade entre dois grupos como sendo a distância mínima entre pares de amostras compostos por uma amostra de cada grupo. Apesar de ser capaz de fornecer agrupamentos que não são exclusivamente globulares no espaço de atributos, o *single-linkage* é muito sensível a ruídos e *outliers*. De maneira similar, o algoritmo *complete-linkage* (SØRENSEN, 1948) calcula essa dissimilaridade como sendo a distância máxima entre pares de amostras compostos por uma amostra de cada grupo. Apesar de ser menos suscetível a ruídos e *outliers*, o algoritmo *complete-linkage* tende a dar preferência para a subdivisão de grupos com maior cardinalidade, enviesando a obtenção de agrupamentos globulares. Já o algoritmo *average-linkage* (SOKAL, 1958) calcula a dissimilaridade entre dois grupos como sendo a distância média entre pares de amostras compostos

por uma amostra de cada grupo. Apesar de também favorecer grupos globulares, o *average-linkage* é muito menos suscetível a ruídos e *outliers* do que o *complete-linkage*, o que o torna um algoritmo mais robusto e capaz de fornecer resultados mais consistentes.

Diferentemente dos algoritmos supracitados, o algoritmo de Ward (WARD JR, 1963) calcula a dissimilaridade entre grupos com base na somatória dos erros quadráticos (J) do agrupamento, conforme definido na Equação 4.2. Assim, em cada passo da sua hierarquia, são considerados menos dissimilares os grupos para os quais uma suposta fusão proporcionará o menor aumento possível de J, visando minimizar o crescimento das variâncias intragrupos. Como isso, o algoritmo de Ward pode ser considerado como análogo ao *k-means*, só que em uma abordagem hierárquica. Apesar de proporcionar resultados que o tornam similar ao *average-linkage* com relação à robustez, diversos estudos indicam que o algoritmo de Ward possui eficácia superior aos outros algoritmos hierárquicos (GROSS, 1972; KUIPER; FISHER, 1975; MOJENA, 1977; BAYNE, 1980; GOLDEN; MEEHL, 1980; MILLIGAN; SCHILLING, 1985).

Em geral, é importante ressaltar que, independentemente do algoritmo hierárquico utilizado, se o conjunto de dados permitir a criação de grupos compactos e bem separados, todos eles deverão produzir dendrogramas semelhantes (WITTEN; FRANK; HALL, 2011). Por outro lado, a geração de agrupamentos aninhados impossibilita que uma amostra seja associada a um novo grupo em um determinado nível, impedindo a correção de erros executados em passos anteriores (XU; WUNSCH-II, 2005; EVERITT, 2011). Outra questão relevante em alguns casos é com relação à complexidade computacional, que para algoritmos de agrupamento hierárquico aglomerativo é, em geral, de ordem quadrática ($\mathcal{O}(n^2)$). Quando essa questão é relevante para o contexto da aplicação, um esquema de parametrização pode ser utilizado para reduzir o cálculo de distâncias e assim amenizar o custo computacional desses algoritmos, desde que estes sejam utilizados em sua forma tradicional (LANCE; WILLIAMS, 1967). Por outro lado, a capacidade de exploração do dendrograma pelo usuário, permitindo definir a quantidade e disposição dos grupos que melhor atende as necessidades da aplicação, pode ser considerada como a característica mais positiva das abordagens de agrupamento hierárquicas.

Tanto a abordagem por particionamento quanto a hierárquica - no caso da segunda, considerando cada um dos níveis do dendrograma - permitem que cada amostra seja associada a apenas um único grupo. Por conta disso, são classificadas como técnicas de agrupamento sem sobreposição. Entretanto, em grande parte dos problemas do mundo real, separar certas amostras do conjunto de dados não é uma tarefa simples, levando a necessidade de considerar a associação de cada amostra a mais de um grupo. Para tanto, as abordagens capazes de produzir grupos cujas amostras são associadas a eles considerando uma probabilidade ou grau de perti-

nência, conhecidas como técnicas de agrupamento com sobreposição, devem ser consideradas. Um dos principais algoritmos existentes para atender a essas abordagens é o EM (*Expectation-Maximization*) (DEMPSTER; LAIRD; RUBIN, 1977), que foi concebido originalmente como método de aprendizado de máquina semissupervisionado, mas que pode ser utilizado como técnica de agrupamento com sobreposição por se basear em cálculos probabilísticos. Devido às suas características, a utilização do *k-means* como técnica agrupamento sem sobreposição é altamente correspondente com a utilização do EM como técnica de agrupamento com sobreposição.

Outro algoritmo de agrupamento com sobreposição bastante similar ao *k-means*, mas que considera a pertinência das amostras ao invés de probabilidades com relação aos grupos é o *Fuzzy c-Means* (FCM) (BEZDEK; EHRLICH; FULL, 1984). Também inicializado com a escolha aleatória de k centroides, em cada iteração do FCM é calculado o grau de pertinência de cada amostra com relação a cada um dos k grupos (ω_k), considerando a distância (d) da mesma com relação ao centroide desses grupos, além de um parâmetro de *fuzzificação* (m) . A Equação 4.3 exibe esse cálculo, considerando j como sendo a quantidade de grupos desejada.

$$\omega_k = \frac{1}{\sum_j \left(\frac{d(centroide(k),x)}{d(centroide(j),x)} \right)^{\frac{2}{(m-1)}}} \tag{4.3}$$

A partir da Equação 4.3, pode ser verificado que quanto maiores são os valores utilizados pelo parâmetro m, maior será a vagueza associada às fronteiras de separação entre os grupos. No limite $m=1$, o grau de pertinência deve convergir para 0 ou 1, implicando na obtenção de grupos sem sobreposição semelhantes aos obtidos pelo *k-means*. Ao final de cada iteração os centroides são recalculados, levando-se em consideração todas as amostras do conjunto de dados e seu grau de pertinência com relação a cada grupo (HU; MENG; SHI, 2008). O FCM converge quando os valores de pertinência não sofrem alterações acima de um limiar, que deve ser informado como parâmetro para o algoritmo. Diferentemente do *k-means*, que fornece como resultado a associação de cada amostra a um único grupo, o FCM retorna uma matriz de dados contendo o grau de pertinência - com valores entre 0 e 1 - de cada amostra com relação a cada grupo. Entretanto, se for necessária a obtenção de agrupamentos sem sobreposição para o resultado final, uma operação de *defuzzificação* pode ser realizada. Normalmente, essa operação é realizada associando cada amostra ao grupo cujo grau de pertinência seja o maior, ou considerando outras formas que levam em consideração vizinhanças, localização espacial ou modelos probabilísticos (LE; ALTMAN; GARDINER, 2012). O FCM possui como principal vantagem a capacidade de fornecer melhores resultados em conjuntos de dados onde ocorrem muitas sobreposições. A principal desvantagem está associada à dificuldade na escolha do limiar adequado para a con-

vergência do algoritmo. Quanto menor esse valor, o que normalmente proporciona resultados mais precisos, maior é o número de iterações que devem ser executadas pelo algoritmo, aumentando significativamente o seu custo computacional (CAI; CHEN; ZHANG, 2007). Por outro lado, um limiar que proporcione a rápida convergência do algoritmo também pode resultar em um mínimo local ao invés de global (HATHAWAY; BEZDEK, 1986).

De uma maneira geral, apesar de a abordagem hierárquica levar vantagem por não ser necessário informar a quantidade de grupos desejada e os centroides iniciais, a mesma não consegue fornecer bons resultados quando os dados são complexos e difíceis de separar. No entanto, essa característica é bastante explorada nas abordagens de agrupamento com sobreposição, que normalmente fornecem melhores resultados para esse tipo de dado. Desse modo, reforça-se a ideia de que a escolha da abordagem correta a ser utilizada depende muito da aplicação e das características relacionadas aos dados utilizados. Considerado a tarefa delineamento de UGDs em AP, as abordagens utilizadas devem considerar a complexidade dos conjuntos de dados espaciais, onde o espaço de coordenadas deve ser tratado de maneira diferenciada com relação ao espaço de atributos, de tal forma que sejam obtidos agrupamentos coesos e bem separados considerando esses dois espaços de maneira equilibrada. Para tanto, algoritmos tradicionais de agrupamento, como os que foram descritos nesta seção, necessitam ser modificados para que possam atender a esses requisitos.

4.2.4 *Ensembles* de Agrupamentos

A utilização de diferentes tipos de algoritmos de agrupamento, utilizando variações nos conjuntos de dados e valores parametrizados para resolver tarefas em aplicações específicas, como o delineamento de UGDs em AP, pode proporcionar a obtenção de resultados com algumas diferenças que geram dúvidas ao usuário final sobre qual é a melhor solução a ser utilizada. Por conta disso, tem surgido na literatura diversas abordagens que permitem a obtenção de agrupamentos consensuais e, consequentemente, mais robustos, a partir de dois ou mais agrupamentos gerados para um mesmo objetivo (STREHL; GHOSH, 2002; FRED; JAIN, 2005; NALDI; CARVALHO; CAMPELLO, 2013; JASKOWIAK, 2015). Essas abordagens são conhecidas como *ensembles* de agrupamentos e podem ser obtidas de diferentes maneiras, tais como: utilizando agrupamentos obtidos a partir de diferentes algoritmos; variando valores de um mesmo parâmetro de um único algoritmo; ou selecionando subconjuntos de atributos de um mesmo conjunto de dados. Nesses casos, deve ser sempre utilizada a mesma quantidade de amostras nos agrupamentos originais (GHOSH; ACHARYA, 2011). Duas dessas abordagens, tradicionais na literatura, foram utilizadas nesta tese e estão descritas a seguir.

4.2.4.1 *Ensembles* de agrupamentos baseados em grafos

O principal objetivo das três abordagens de *ensembles* desenvolvidas por Strehl e Ghosh (2002) é a obtenção de soluções de consenso que busquem a maior quantidade possível de informação mútua compartilhada proveniente dos agrupamentos originais. A abordagem mais simples, conhecida como CSPA (*Cluster-based Similarity Partitioning Algorithm*), é baseada na prerrogativa de que duas amostras devem ser similares se pertencem ao mesmo grupo; e dissimilares se não pertencem. Desse modo, deve ser criada uma matriz binária de dimensão $n \times n$, onde n é a quantidade de amostras, para representar essa similaridade em cada agrupamento original. Para reagrupar essas amostras, o CSPA utiliza um algoritmo baseado em particionamento de grafos (KARYPIS; KUMAR, 1998). Um pouco mais sofisticada, a abordagem conhecida como HGPA (*HyperGraph Partitioning Algorithm*) considera um *ensemble* de agrupamentos como um problema de particionamento de hipergrafos, onde as hiperarestas representam os agrupamentos originais e os seus relacionamentos. Para reagrupar as amostras, é utilizado um algoritmo de particionamento de hipergrafos que permite a poda de uma quantidade mínima de hiperarestas (HAN, 1997). A principal diferença entre o CSPA e o HGPA é que enquanto o primeiro considera apenas relacionamentos entre pares de amostras, o segundo considera também os relacionamentos entre os agrupamentos em si. Finalmente, a abordagem MCLA (*Meta-Clustering Algorithm*) permite a representação de cada grupo, considerando todos os agrupamentos envolvidos, como uma hiperaresta, associando cada amostra com a hiperaresta na qual ela participa de maneira mais ativa. Ao final da execução, essas hiperarestas são reagrupadas, identificando um agrupamento consolidado.

Enquanto a abordagem CSPA possui custo computacional quadrático, as abordagens HGPA e MCLA podem ser executadas em tempo linear. Além disso, de acordo com Strehl e Ghosh (2002), a abordagem MCLA tende a fornecer melhores resultados com relação ao compartilhamento de informação mútua do que as abordagens CSPA e HGPA, quando os agrupamentos originais possuem baixas taxas de ruído e diversidade. Apesar de terem sido utilizadas em experimentos ao longo desta tese, essas abordagens de *ensembles* para agrupamentos envolvem conceitos extras de particionamento de grafos e hipergrafos, que podem dificultar o seu entendimento por parte do usuário final.

4.2.4.2 *Ensembles* de agrupamentos baseados em acúmulo de evidência

O conceito de acúmulo de evidências permite que cada agrupamento original seja visto como uma evidência independente de organização dos dados. Na abordagem desenvolvida por Fred e Jain (2005), esses agrupamentos são inicialmente combinados por um mecanismo

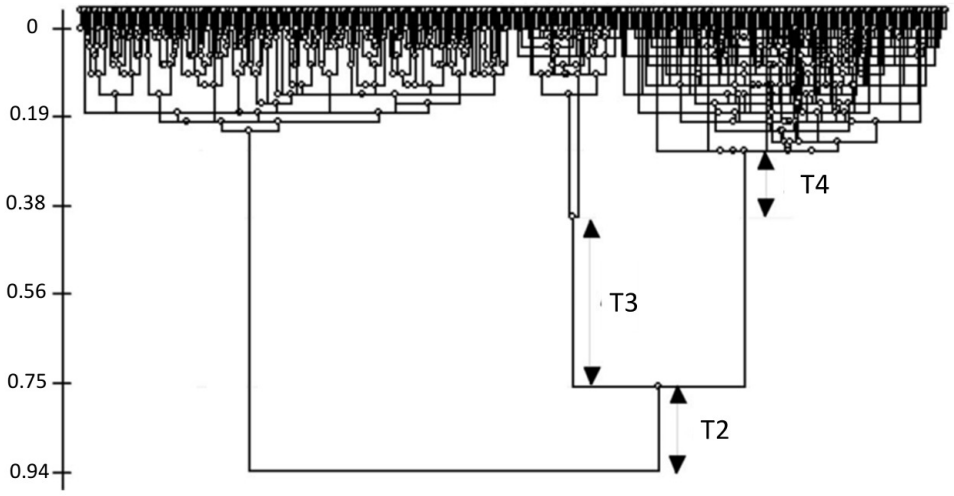
de votos para gerar uma nova matriz de similaridades $n \times n$, conhecida como matriz de co-associação. A Equação 4.4 define o cálculo que deve ser realizado para a criação de uma matriz de co-associação M, considerando todos os pares de amostras.

$$M(i,j) = \frac{n_{ij}}{N} \tag{4.4}$$

Na Equação 4.4, n_{ij} representa o número de vezes que o par composto pelas amostras i e j está associado a um mesmo grupo, considerando os N agrupamentos originais.

A segunda etapa dessa abordagem utiliza a matriz de co-associação M como insumo para um algoritmo hierárquico aglomerativo, que realiza o reagrupamento das amostras objetivando a obtenção de um resultado mais robusto. A partir do dendrograma obtido por esse algoritmo, o agrupamento final escolhido deverá ser aquele com o maior tempo de vida dentro da hierarquia, conforme exemplificado na Figura 4.1.

Figura 4.1: Dendrograma produzido por algoritmo de agrupamento hierárquico utilizando uma matriz de co-associação.



Fonte: Adaptada de Fred e Jain (2005).

De acordo com a Figura 4.1, os valores representados por T2, T3 e T4 correspondem, respectivamente, à distância utilizada como medida de similaridade para fundir dois grupos em um, ou tempo de vida para o agrupamento que contém dois grupos; três grupos em dois, ou tempo de vida para o agrupamento que contém três grupos; e quatro grupos em três, ou tempo de vida para o agrupamento que contém quatro grupos. Assim, observa-se um maior tempo de vida para o agrupamento que contém três grupos e que, de acordo com essa abordagem, deve ser considerado como a solução ideal e mais robusta para a resolução do problema.

Para essa segunda etapa da abordagem, podem ser utilizados os algoritmos *single-linkage* e *average-linkage*, com melhores resultados obtidos pelo segundo a partir de experimentos realizados por Fred e Jain (2005). Os autores também realizaram experimentos comparativos com as abordagens propostas por Strehl e Ghosh (2002), concluindo que a abordagem de acúmulo de evidências possui um desempenho em geral melhor, principalmente em reconhecer grupos com formas arbitrárias ou não-circulares no espaço de atributos, viabilizando o seu uso em conjuntos de dados com boa correlação. Com relação à complexidade computacional, essa abordagem possui custo quadrático, com tempo de execução muito próximo a um algoritmo de agrupamento hierárquico aglomerativo.

Devido ao seu melhor desempenho com relação a outras abordagens, e pelo fato de ser baseada única e exclusivamente em algoritmos de agrupamento para reagrupar as amostras, a abordagem de acúmulo de evidências foi utilizada na abordagem de agrupamento de dados espaciais desenvolvida nesta tese, com o intuito de proporcionar a obtenção de agrupamentos mais robustos a partir de resultados individuais obtidos por meio de variações nos parâmetros da abordagem de agrupamento espacial proposta.

4.2.5 Critérios de Validação para Agrupamentos Convencionais

Os métodos de aprendizado de máquina supervisionado dispõem de diversas alternativas para avaliar a qualidade dos modelos gerados (POWERS, 2011). Já para os métodos não supervisionados, baseados em agrupamento, existe uma maior dificuldade nessa avaliação, pois os grupos são obtidos sem o auxílio de um conjunto de amostras previamente rotuladas. Entretanto, avaliar os agrupamentos obtidos de maneira objetiva e qualitativa, considerando diversos fatores, tais como a utilização de diferentes algoritmos de agrupamento, quantidades de grupos e parametrizações, é uma atividade importante para auxiliar o usuário final em decidir quais soluções são úteis para a sua aplicação (JAIN; DUBES et al., 1988). Na literatura, podem ser encontradas diversas métricas capazes de realizar esse tipo de avaliação, conhecidas como critérios de validação. Essas métricas procuram avaliar, em um primeiro momento, a qualidade dos grupos encontrados e determinar, de maneira relativa, qual a quantidade mais apropriada de grupos em que deve ser dividido um conjunto de dados. Posteriormente, os grupos encontrados podem ser validados estatisticamente com relação à aleatoriedade, para verificar se são realmente naturais ou se foram obtidos ao acaso. Os critérios utilizados para validação de agrupamentos são normalmente divididos em dois tipos: externos, onde é verificada a correspondência do agrupamento gerado com relação a uma solução de referência; e internos, onde é verificada a qualidade dos agrupamentos obtidos a partir dos dados utilizados para sua obtenção.

4.2.5.1 Critérios de Validação Externa

Os critérios de validação externa são utilizados para medir o grau de correspondência entre um agrupamento gerado por uma determinada abordagem com relação a um agrupamento de referência previamente conhecido. O agrupamento de referência é normalmente gerado a partir do conhecimento do especialista do domínio, mas pode ter sido obtido por outra abordagem, ou por variações nos parâmetros e dados de entrada. Na literatura, a maioria dos critérios de validação externa são baseados na verificação de verdadeiros e falsos positivos e negativos, o que pode proporcionar algumas limitações e desempenho ruim em alguns casos (JACCARD, 1901; RAND, 1971). Por outro lado, alguns coeficientes de correlação podem atuar como critérios externos, e proporcionar comparações mais precisas e justas entre dois agrupamentos.

No contexto da AP, o coeficiente de correlação estatístico Kappa (COHEN, 1960) tem sido muito utilizado para diversos objetivos, principalmente para verificar a correlação de mapas de UGDs com relação aos atributos utilizados pelo algoritmo de agrupamento para sua obtenção (KHOSLA, 2008; SONG, 2009; KHOSLA, 2010); e como critério de validação externa, comparando, por exemplo mapas de UGDs gerados por um mesmo algoritmo de agrupamento, mas utilizando conjuntos de atributos distintos (KITCHEN, 2005). No caso da validação externa, o coeficiente Kappa utiliza uma tabela de contingência cruzada para correlacionar amostras rotuladas com a mesma localização geográfica nos dois mapas que estão sendo comparados.

Outro coeficiente de correlação que pode ser utilizado como critério de validação externa de agrupamentos é a estatística Γ de Hubert (HUBERT; ARABIE, 1985). Mesmo sendo baseado na verificação de falsos e verdadeiros positivos e negativos, esse critério considera, adicionalmente, uma interpretação probabilística que permite que sejam obtidos resultados mais confiáveis. A Equação 4.5 define, de maneira simplificada, o cálculo realizado para obtenção do índice de correlação entre duas partições seguindo a estatística Γ de Hubert.

$$\Gamma = \frac{N_p \times yy - (yy+yn)(yy+ny)}{\sqrt{(yy+yn)(yy+ny)(nn+yn)(nn+ny)}} \tag{4.5}$$

De acordo com a Equação 4.5, e considerando $A1$ e $A2$ como os dois agrupamentos a serem comparados: N_p é a quantidade total de pares de amostras que deverão ser comparadas; *yy* representa a quantidade de pares de amostras que aparecem no mesmo grupo em $A1$ e $A2$; *yn* representa a quantidade de amostras que aparecem no mesmo grupo em $A1$ mas em grupos distintos em $A2$; *ny* representa a quantidade de amostras que aparecem em grupos distintos em $A1$ mas no mesmo grupo em $A2$; e *nn* representa a quantidade de amostras que aparecem em grupos distintos tanto em $A1$ quanto em $A2$.

Assim como o critério de correlação de Pearson, tanto o coeficiente *Kappa* quando a estatística Γ de Hubert podem fornecer índices que variam de -1 até 1, onde os valores próximos de 0 devem indicar ausência de correlação, e os valores próximos a 1 e -1 devem indicar, respectivamente, alta correção no mesmo sentido e alta correlação em sentidos opostos dos agrupamentos verificados. Como na maioria das comparações entre agrupamentos são utilizadas partições de referência únicas, contendo uma quantidade definida de grupos, agrupamentos obtidos a partir de abordagens hierárquicas devem ser comparados com partições de referência realizando-se cortes no dendrograma obtido, considerando a quantidade de grupos desejada.

4.2.5.2 Critérios de Validação Interna

Considerando que, em termos práticos, a necessidade de utilizar uma abordagem de agrupamento para particionar os dados vem justamente do fato de não existir um agrupamento ou classificação de referência para efeito de comparação, os critérios de validação interna surgem como uma maneira de avaliar a qualidade de um agrupamento obtido a partir dos dados utilizados para sua obtenção (JAIN; DUBES et al., 1988; VENDRAMIN; CAMPELLO; HRUSCHKA, 2010). Para medir essa qualidade, normalmente são utilizadas medidas ponderadas de coesão intergrupos e de separação intragrupos (DAVIES; BOULDIN, 1979; TRAUWAERT, 1988). Os critérios de validação interna são considerados relativos quando possuem a capacidade de avaliar qual agrupamento é melhor em um conjunto de dois ou mais agrupamentos. Desse modo, um critério de validação interna relativo pode servir para identificar, por exemplo, qual o melhor agrupamento para uma determinada quantidade de grupos dentre diversos agrupamentos gerados por abordagens diferentes ou por diferentes parâmetros de uma mesma abordagem; ou identificar qual a quantidade de grupos mais adequada considerando agrupamentos gerados com diferentes quantidades de grupos para uma única abordagem (JAIN; DUBES et al., 1988).

Na literatura, diversos critérios de validação interna podem ser encontrados, porém poucos estudos foram realizados para verificar o seu desempenho em diferentes cenários. O trabalho de Milligan e Cooper (1985) foi pioneiro nesse sentido, e utilizado como referência para Vendramin, Campello e Hruschka (2010) realizarem experimentos comparativos entre os critérios de validação interna mais conhecidos considerando diferentes quantidades de grupos e de atributos, em diferentes situações. A partir desses experimentos, o critério da largura de silhueta (ROUSSEEUW, 1987) foi considerado, de uma maneira geral, como o mais robusto dentre os critérios estudados. A Equação 4.6 exibe o cálculo da largura de silhueta para cada amostra utilizada na geração de um agrupamento.

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{4.6}$$

De acordo com a Equação 4.6, $a(i)$ é a dissimilaridade média da i-ésima amostra com relação às outras amostras associadas ao mesmo grupo (coesão); e $b(i)$ é a dissimilaridade média da i-ésima amostra com relação às amostras associadas aos outros grupos (separação). Essa dissimilaridade é calculada utilizando-se uma medida conhecida, como a distância Euclidiana. A partir do cálculo de s para cada amostra i, a silhueta s_k de um grupo k é calculada como sendo o valor médio de s de todas as amostras pertencentes a esse grupo; e a silhueta global do agrupamento (S) é calculada como sendo o valor médio de s_k, para k variando de 1 até K, sendo K a quantidade total de grupos do agrupamento avaliado. Os índices obtidos para S podem variar entre -1 e 1, onde valores mais altos indicam maior qualidade nos agrupamentos. Desse modo, o critério da largura de silhueta é conhecido como critério de maximização.

Outro critério de validação interna que vem sendo muito utilizado na literatura, principalmente para avaliar a qualidade de agrupamentos obtidos a partir de dados espaciais, é o critério SD (HALKIDI; VAZIRGIANNIS; BATISTAKIS, 2000). Esse critério utiliza vetores de variâncias para determinar a coesão intragrupos (S) e dissimilaridade entre os centroides representantes dos grupos para determinar a separação intergrupos (D) de um agrupamento. A Equação 4.7 exibe os passos para o cálculo do critério SD para um agrupamento contendo K grupos.

$$S = \frac{\frac{1}{K}\sum_{k=1}^{K}\left\|V^k\right\|}{\|V\|} \tag{4.7a}$$

$$D = \frac{D_{max}}{D_{min}}\sum_{k=1}^{K}\frac{1}{\sum_{k'=1;k'\neq k}^{K}\left\|G^k - G^{k'}\right\|} \tag{4.7b}$$

$$SD = \alpha * S + D \tag{4.7c}$$

Na Equação 4.7a, $\|V\|$ é o vetor de variâncias dos atributos considerando todas as amostras originais; e $\left\|V^k\right\|$ é esse mesmo vetor, porém considerando apenas as amostras pertencentes ao grupo k, para k variando entre 1 e K. Na equação 4.7b, D_{max} e D_{min} correspondem, respectivamente, à maior e menor dissimilaridade entre os centroides representantes dos grupos, normalmente obtida a partir da distância Euclidiana; e G^k e $G^{k'}$ representam, respectivamente, os centroides representantes dos grupos k e k', para k diferente de k'. Finalmente, a Equação 4.7c calcula o valor final do critério SD para o agrupamento, somando-se o valor da separação (D) com a coesão (S). O valor de S ainda é ponderado por um valor α correspondente ao valor de D para o agrupamento com a maior quantidade de grupos a ser comparado. Se o critério for utilizado para calcular apenas o valor de SD para um único agrupamento, o valor de α será igual

a D. Entretanto, essa cálculo nos permite classificar o critério SD como de validação interna relativo, pois o cálculo de α sugere a comparação de agrupamentos com diferentes quantidades de grupos. No caso do critério SD, índices próximos de 0 indicam agrupamentos com melhor qualidade, não existindo limite para valores máximos. Assim, o critério SD é conhecido como critério de minimização.

No contexto da AP, os critérios de validação interna são importantes para auxiliar o usuário final em avaliar a qualidade dos agrupamentos que representam as UGDs, bem como de verificar qual deve ser a quantidade ideal de CGDs considerando os diferentes atributos, abordagens e parâmetros que podem ser utilizados para o seu delineamento.

4.2.5.3 Validação Interna Utilizando Métodos Estatísticos

Apesar da grande disponibilidade de critérios de validação interna presentes na literatura, sugerindo que a proposta de novos critérios para diferentes situações e tipos de dados seja algo factível, a avaliação qualitativa de agrupamentos considerando apenas os dados utilizados para gerá-los pode se tornar uma tarefa bastante complexa. A principal dificuldade para o usuário final na utilização desses critérios está relacionada com o estabelecimento de intervalos para os índices fornecidos que possam identificar, por exemplo, agrupamentos de alta ou baixa qualidade. Para que essa identificação possa ser realizada, podem ser utilizadas ferramentas estatísticas baseadas em testes de significância sob hipóteses predeterminadas (FISHER, 1925). No contexto dos critérios de validação interna para agrupamentos, esse tipo de teste tenta identificar faixas de valores nos índices que identificam agrupamentos válidos, ou seja, gerados a partir de uma subdivisão natural dos dados; e faixas de valores nos índices que identificam agrupamentos gerados de maneira aleatória (JAIN; DUBES et al., 1988).

Para que um teste de significância seja utilizado nesse contexto, é necessário que seja definida uma medida estatística (T) e uma hipótese nula (H_0) a ser comprovada. Se rejeitada, H_0 favorecerá a comprovação de uma hipótese alternativa H_1 onde se procuram evidências, de acordo com os resultados obtidos. Assim, podemos definir um teste de significância utilizando a teoria das probabilidades (Equação 4.8).

$$P(T \geq t_\alpha | H_0) = \alpha \tag{4.8}$$

Segundo a Equação 4.8, se assumirmos que o valor de T obtido por um experimento é t^*, H_0 deverá ser rejeitada ao nível de significância α se $t^* \geq t_\alpha$, fazendo com que H_1 seja aceita (JAIN; DUBES et al., 1988).

No contexto da análise qualitativa de agrupamentos, a estatística T pode ser constituída, por exemplo, por uma série de índices fornecidos por critérios de validação interna para avaliar diferentes agrupamentos gerados, com valores de α variando normalmente entre 1% e 5%. A Equação 4.8 em sua forma original é válida apenas para critérios de validação classificados como de maximização, como a largura de silhueta. Para critérios de minimização, como o SD, o operador de comparação deverá ter o sinal invertido na equação.

A H_0 mais utilizada em análises estatísticas de agrupamento é a de posições aleatórias equiprováveis (JAIN; DUBES et al., 1988). Segundo essa hipótese, espera-se que muitos conjuntos de dados, representados dentro de um hipercubo n-dimensional considerando n atributos, possuam valores usuais ou frequentes para T, e que apenas alguns desses conjuntos possuam valores não usuais ou pouco frequentes para T. Desse modo, testar H_0 para posições aleatórias equiprováveis é o mesmo que avaliar a probabilidade da distribuição de T ser gerada a partir de dados aleatórios e que não podem ser agrupados de maneira natural.

Uma maneira bastante utilizada de executar computacionalmente os testes de significância é por meio de análises de Monte Carlo (METROPOLIS; ULAM, 1949). Essas análises são executadas a partir de métodos que estimam parâmetros e probabilidades por meio de amostragens e simulações computacionais (JAIN; DUBES et al., 1988). Nesse contexto, e considerando a H_0 de posições aleatórias equiprováveis, podem ser obtidos, por exemplo, $N-1$ conjuntos de dados sintéticos, com valores para os n atributos gerados aleatoriamente a partir da média e variância do conjunto de dados original, caracterizando efetivamente as posições aleatórias de H_0. Esses $N-1$ conjuntos de dados são unidos ao conjunto de dados original, possibilitando a geração de N agrupamentos ($C_1..C_N$) para uma quantidade k de grupos predefinida. Esses agrupamentos, por sua vez, podem ser avaliados a partir de um critério de validação interna de minimização S que fornece N valores distintos ($S1..SN$). Considerando a Equação 4.8 com sinal invertido, conforme relatado acima, onde $S = T$, e $S1 = t^*$ para o conjunto de dados original, se o valor de $S1$ estiver entre os $\alpha\%$ menores valores de T, H_0 deverá ser rejeitada e a evidência H_1 encontrada será de que o agrupamento gerado com dados reais é de fato uma subdivisão natural do conjunto de dados, com um nível de significância de $\alpha\%$. Caso contrário, esse agrupamento será considerado aleatório, pois o valor de T para esse agrupamento é frequentemente obtido, mesmo que os dados utilizados para a obtenção do agrupamento sejam sintéticos.

A aleatoriedade identificada por um teste de significância para um agrupamento obtido a partir de dados originais pode possibilitar diversas constatações, tais como: o conjunto de dados reais utilizado é muito complexo e difícil de separar, provavelmente com muitos atributos que não seguem uma distribuição normal ou gaussiana; a abordagem de agrupamento ou até

mesmo os parâmetros utilizados para subdividir os dados não são ou não foram determinados de maneira eficaz; ou o critério de validação interna utilizado como estatística não é uma medida válida para avaliar a qualidade dos agrupamentos obtidos no contexto da aplicação.

4.3 Descoberta de Conhecimento em Bancos de Dados Espaciais

Quando o processo de KDD envolve a utilização de dados espaciais, novos conceitos e características relacionados à complexidade desse tipo de dado devem ser considerados, tornando-o um processo de descoberta de conhecimento em bancos de dados espaciais, ou KDSD[2](ESTER; KRIEGEL; SANDER, 1999). Essa complexidade está vinculada ao fato dos dados espaciais considerarem as características de localização espacial - descritas pelo espaço de coordenadas - de maneira distinta com relação aos eventos que estão atrelados a essa localização - descritos pelo espaço de atributos.

Considerando essas questões, a mineração de dados, principal etapa e responsável pela extração de padrões de conhecimento em um KDD, deve ser adaptada para a utilização de dados espaciais em um KDSD. Nesse sentido, Ng e Han (1994) definiram a mineração de dados espaciais como sendo a descoberta de relacionamentos e características específicas intrínsecas em bancos de dados espaciais. Segundo Ester (2000), a principal diferença entre a mineração de dados espacial com relação a convencional está ligada ao fato de que os valores dos atributos de uma amostra podem ser influenciados pelos valores dos atributos dos seus vizinhos espaciais. Devido à natureza desses dados, esses vizinhos são facilmente identificáveis por meio de primitivas de bancos de dados, baseadas em relacionamentos de vizinhança espacial. A maneira explícita de expressar a localização e a extensão de objetos espaciais utilizada pelos SGBDEs permite a definição desses relacionamentos, que podem ser utilizados por algoritmos de mineração específicos para dados espaciais (ESTER; KRIEGEL; SANDER, 1999).

De uma maneira prática e objetiva, Yang, Bai e Gong (2009) definiram as principais características da mineração de dados espaciais e suas diferenças com relação ao processo convencional. Nesse estudo, os autores constataram que os métodos tradicionalmente utilizados para as tarefas de mineração de dados convencionais, tais como classificação, agrupamento, regressão e regras de associação, são também utilizados na mineração de dados espaciais. Entretanto, a natureza dos dados espaciais, suas aplicações e formas de representação do conhecimento um tanto quanto distintas com relação aos dados convencionais, sugerem a criação de novas técni-

[2]Do inglês *Knowledge Discovery in Spatial Databases*

cas para melhorar a eficiência e eficácia na identificação dos padrões espaciais desejados. Essas novas técnicas devem tratar características mais específicas dos dados espaciais, conforme definido por Jin e Miao (2010):

- A quantidade massiva de dados espaciais sugere a criação de algoritmos mais eficientes para tratá-los;
- O relacionamento não linear que ocorre entre os atributos espaciais;
- A possibilidade de utilização de dados espaciais em diferentes escalas, partindo de informações mais generalizadas até as mais refinadas;
- O incremento da dimensão espacial ou quantidade de bandas disponíveis em imagens aéreas ou de sensoriamento remoto;
- A ambiguidade da informação espacial, presente na localização e correlação espacial, bem como nos valores de atributos *fuzzy*;
- A ausência de dados, que vem impulsionando pesquisas em como recuperar dados perdidos e estimar os seus parâmetros de distribuição.

Na sequência desta seção, serão discutidas as etapas do KDD convencional que foram adaptadas para dados espaciais no âmbito desta tese, permitindo definir o delineamento de UGDs em AP como um processo de KDSD. Nesse contexto, serão consideradas as seguintes tarefas: interpolação de dados espaciais e análise espacial de componentes principais, realizada na etapa de pré-processamento; agrupamento de dados espaciais, realizada na etapa de mineração de dados; e validação dos agrupamentos ou mapas de UGDs obtidos, fornecendo suporte à tomada de decisão ao usuário final em verificar as melhores soluções para o problema considerando os dados originais.

4.3.1 Interpolação de Dados Espaciais

Assim com acontece na mineração de dados convencionais, a etapa de pré-processamento no contexto espacial também deve executar tarefas como a normalização dos atributos, limpeza, seleção e redução dos dados, levando em consideração tanto o espaço de atributos quanto o espaço de coordenadas. No entanto, devido ao fato dos conjuntos de dados de AP serem obtidos por diferentes equipamentos em intervalos de tempo e espaço distintos, a tarefa de interpolação espacial torna-se extremamente importante para viabilizar a sua distribuição em uma grade

espacial única. Essa distribuição é um pré-requisito para que os algoritmos de agrupamento possam encontrar padrões em dados espaciais que estão associados a um conjunto de atributos. Apesar de estimar valores em locais normalmente não amostrados (o que faz com que os dados reais obtidos em campo nem sempre sejam utilizados) sem a execução de uma interpolação espacial em grade única, a tarefa de agrupamento teria que lidar com diferentes quantidades e localizações de amostras para cada um dos atributos, aumentado a sua complexidade. Para compensar as perdas que podem ocorrer com a execução dessa tarefa, é necessário avaliar algumas situações particulares de cada atributo para a escolha do algoritmo correto. Independentemente disso, o objetivo final a ser atingido é estimar valores individuais para todas as variáveis que irão compor o espaço de atributos, em todas as amostras espaciais de uma grade única estabelecida. Para tanto, devem ser considerados os valores dessas mesmas variáveis em pontos localizados em sua vizinhança espacial no conjunto de dados original (BOHLING, 2005).

As abordagens menos sofisticadas para interpolação espacial utilizam métodos determinísticos. O algoritmo do vizinho mais próximo, ou NN (*Nearest Neighbor*), é o mais simples deles, onde para cada amostra da nova grade é associado o valor do atributo da amostra espacialmente mais próxima presente no conjunto de dados original. Esse algoritmo é o único que garante que não serão gerados novos valores, pois o valor interpolado seguramente será um dos pertencentes ao conjunto de dados original, provavelmente com uma localização distinta (FRANKE, 1982). Já o algoritmo da média dos k vizinhos mais próximos, ou *k-NN* (*k-Nearest Neighbors*) (ALTMAN, 1992), é utilizado como abordagem de interpolação espacial regressiva, onde o valor do atributo a ser atribuído a cada amostra da nova grade é a média dos seus k vizinhos espacialmente mais próximos pertencentes ao conjunto de dados original. Os métodos determinísticos mais elaborados utilizam a soma ponderada dos valores em locais próximos para a execução da interpolação espacial. O algoritmo do inverso da potência das distâncias, ou IDW (*Inverse Distance Weighted*) (SHEPARD, 1968), estipula pesos inversamente proporcionais à distância elevada a uma potência da amostra cujo valor do atributo será estimado, com relação a cada uma das amostras do conjunto de dados original. A forma mais utilizada dessa algoritmo, também conhecida como inverso do quadrado das distâncias (ISAAKS; SRIVASTAVA, 1989), define um valor fixo de potência quadrática. Além disso, algumas implementações desse método permitem também a definição de quais vizinhos mais próximos serão utilizados no cálculo.

As abordagens mais sofisticadas utilizam métodos geoestatísticos, onde a *krigagem* (KRIGE, 1952; MATHERON, 1963, 1969) destaca-se como principal algoritmo. Esse interpolador é baseado em regressão espacial, com o objetivo de estimar valores de um atributo em uma localização não amostrada a partir de valores associados às amostras de sua vizinhança espacial. Essas estimativas são ponderadas de acordo com valores de covariância espacial, obtidos a par-

tir da construção de uma função conhecida como semivariograma (BOHLING, 2005). Como é praticamente impossível a determinação exata de que tipo de equação matemática descreve a variabilidade espacial dos dados obtidos em AP, o semivariograma se torna uma solução interessante por permitir a interpretação física do fenômeno em questão (VIEIRA, 2000b). Após a sua construção, é necessário que o mesmo seja ajustado a um modelo teórico aproximado (e.g., linear, esférico, exponencial, gaussiano), para que possa ser utilizado como fator de ponderação pelo algoritmo interpolador da *krigagem*.

A *krigagem* possui diversas formas de implementação, sendo que a mais utilizada em AP é a ordinária, devido à sua flexibilidade (MOLIN; AMARAL; COLAÇO, 2015). Essa implementação satisfaz a maioria dos problemas de estimativa, pois considera as relações espaciais obtidas a partir do semivariograma ajustado e assume que a média é constante na vizinhança local de cada ponto estimado (VIEIRA, 2000a). O passo principal desse algoritmo realiza uma combinação linear entre os valores dos atributos dos pontos amostrados para estimar o valor desejado (Equação 4.9).

$$Z^*(x_0) = \sum_{i=1}^{N} \lambda_i Z(x_i) \tag{4.9}$$

Na equação 4.9, $Z^*(x_0)$ é o valor do atributo a ser estimado para o ponto não amostrado (x_0), e $Z(x_i)$ é o valor do atributo no ponto x_i, ponderado por um peso λ_i, obtido a partir do semivariograma ajustado. Para que a estimativa feita pela *krigagem* não seja tendenciosa, é necessário que a soma dos pesos λ_i seja igual a 1. Essa característica faz com que esse algoritmo forneça resultados com variância mínima e seja considerado o melhor interpolador linear não enviesado (VIEIRA, 2000b).

Os métodos originais da *krigagem* têm sido amplamente estudados e sugestões de melhorias e modificações podem ser encontradas na literatura. Em algumas dessas pesquisas, estão sendo tratadas questões relacionadas à utilização de parâmetros incertos para o semivariograma e de dados vagos para os valores dos atributos das amostras (BARDOSSY; BOGARDI; KELLY, 1988, 1990; DIAMOND, 1989; LOQUIN; DUBOIS, 2010, 2012); e em como melhorar a eficiência e eficácia do interpolador, diminuindo a influência de *outliers* (LIU; CHEN; LU, 2012).

Alguns fatores relacionados à densidade e distribuição espacial dos dados originais podem contribuir para a escolha correta do algoritmo de interpolação a ser utilizado em uma determinada situação. Normalmente, quando o conjunto de dados espaciais é bastante denso, ou seja, quando possui, em geral, uma grande quantidade de amostras nas proximidades das amostras da nova grade espacial que terão os seus valores estimados, os algoritmos determinísticos, como o

k-NN e o *IDW*, são suficientes para que sejam obtidas estimativas confiáveis. Entretanto, para conjuntos de dados espaciais menos densos, a identificação da dependência espacial entre pares de amostras, como é realizado com a construção de semivariogramas, é importante para que se tenha êxito nas estimativas, devido a uma quantidade menor de amostras vizinhas espaciais que, consequentemente, diminuem as evidências para estimativas mais precisas. Nesses casos, a utilização de algoritmos geoestatísticos, como a *krigagem*, passa a ser necessária, desde que se tenha uma quantidade mínima de amostras originais para que sejam obtidas estimativas com o menor erro possível. Independentemente disso, se os dados são densos e uniformemente distribuídos na área de estudo, serão obtidas boas estimativas para o valor do atributo em qualquer ponto não amostrado. Já se os dados estiverem distribuídos espacialmente em grupos com largos espaços entre eles, as estimativas obtidas não serão confiáveis. A maioria dos algoritmos de interpolação subestimam valores altos e superestimam valores baixos, pois são inerentes à média (ISAAKS; SRIVASTAVA, 1989).

4.3.2 Análise Espacial de Componentes Principais

Conforme já mencionado, a Análise de Componentes Principais (PCA) é uma das maneiras de se reduzir a dimensionalidade do espaço de atributos em um processo de KDD convencional. Entretanto, em um KDSD, a correlação espacial existente entre as diferentes localizações que compõem o conjunto de dados espaciais também deve ser verificada, fazendo com que a PCA aplicada apenas em atributos convencionais nem sempre seja suficiente.

Em termos práticos, se considerarmos uma matriz $X_{(n\times p)}$, onde n é a quantidade de amostras e p é o número de variáveis no espaço de atributos, a análise realizada pela PCA retorna, a partir dessa matriz, os componentes principais obtidos a partir de combinações de variáveis que visam à maximização da variabilidade de todo o conjunto, diminuindo a sua dimensionalidade.

Para estender essa análise considerando conjuntos de dados espaciais, Dray, Chessel e Thioulouse (2003) desenvolveram a MULTISPATI-PCA. Essa análise gera adicionalmente uma matriz de pesos $W_{(n\times n)}$, que identifica as conexões espaciais existentes entre as amostras e permite a determinação de uma matriz de atrasos $\bar{X}_{(n\times p)} = W \times X$. Nesse caso, os componentes principais são obtidos por meio da maximização do produto escalar das combinações lineares obtidas a partir de X, com as combinações obtidas a partir de $\bar{X}$.

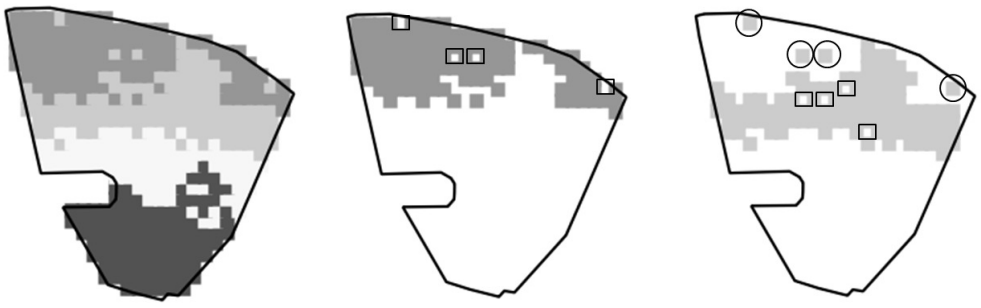
Ao tratar a redução da dimensionalidade do espaço de atributos considerando também os relacionamentos existentes entre as amostras no espaço de coordenadas, entende-se que a MULTISPATI-PCA é uma técnica que objetiva melhores resultados em um processo de KDSD do que a PCA convencional.

4.3.3 Agrupamento de Dados Espaciais

A presença do espaço de coordenadas associado ao espaço de atributos em cada amostra de um conjunto de dados espaciais faz com que os métodos convencionais de agrupamento, em sua essência, não sejam os mais adequados para a extração de padrões espaciais. Devido à sua importância em estabelecer relacionamentos espaciais entre as amostras, o espaço de coordenadas não deve ser ignorado durante essa tarefa, tampouco ser tratado como componente do espaço de atributos. Por conta disso, adaptações vêm sendo realizadas nesses métodos ao longo dos anos, com o intuito de aproveitar as relações espaciais existentes entre as amostras para a obtenção de agrupamentos mais fáceis de entender e úteis na prática.

No contexto da AP, a maioria das abordagens de agrupamento utilizadas para o delineamento de UGDs presentes na literatura não realiza o tratamento diferenciado do espaço de atributos, ou apenas utiliza o espaço de coordenadas para exibir mapas de UGDs em um SIG. Desse modo, pode-se notar que, em grande parte dos casos, as CGDs exibidas no mapa possuem uma perda significativa de contiguidade espacial, tornando os mapas extremamente estratificados e de difícil interpretação pelo usuário final. Esse efeito pode ser verificado quando uma ou mais CGDs são subdivididas em muitas UGDs, que normalmente ocupam áreas muito pequenas. Conforme já mencionado no Capítulo 2, essas pequenas UGDs, bem como os buracos gerados por elas em áreas pertencentes à CGDs vizinhas, serão definidos e tratados no contexto desta tese como estratos. A Figura 4.2 exibe um exemplo de mapa de UGDs gerado a partir do algoritmo de agrupamento FCM, utilizando apenas dados do espaço de atributos como entrada, e CGDs estratificadas em destaque.

Figura 4.2: (a) Mapa de UGDs gerado pelo algoritmo FCM utilizando apenas o espaço de atributos; (b) CGD destacada contendo buracos considerados como estratos, marcados por retângulos; (c) CGD destacada contendo buracos, marcados por retângulos, e UGDs considerados como estratos, marcadas por círculos.



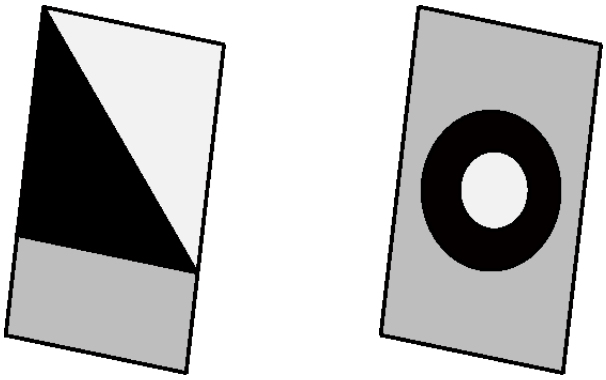
(a) (b) (c)

Fonte: Elaborada pelo autor.

O mapa de UGDs exibido na Figura 4.2 (a) mostra um nível considerável de estratificação, que pode causar dúvidas ao usuário final em como utilizá-lo para a tomada de decisão. Na Figura 4.2 (b), é exibida em destaque uma CGD contendo diversos buracos que podem ser considerados como estratos, marcados por retângulos. Esses buracos foram gerados pelas UGDs consideradas como estratos presentes na CGD destacada na Figura 4.2 (c), marcadas por círculos. Vale destacar também que alguns buracos presentes nessa CGD, também marcados por retângulos, foram gerados pela CGD destacada na Figura 4.2 (b), e outros não.

Uma maneira simples e intuitiva de tentar reduzir o efeito da estratificação nas CGDs poderia ser a inclusão dos atributos de localização latitude e longitude, presentes no espaço de coordenadas, diretamente no espaço de atributos correspondente ao conjunto de dados que será utilizado por um algoritmo de agrupamento. No estudo realizado por Santos, Saraiva e Molin (2012), o uso desse artifício permitiu que fossem eliminadas UGDs pequenas, porém os autores concluíram que a homogeneidade dos grupos considerando apenas o espaço de atributos foi prejudicada. Essa questão também foi estudada em um experimento realizado no âmbito desta tese, descrito no artigo aceito para publicação transcrito no Apêndice D, onde foi verificada uma melhoria na acuidade visual dos mapas de UGDs. Porém, notou-se que a coesão e a separação dos agrupamentos no espaço de atributos, medidas a partir dos critérios de validação interna SD e largura de silhueta, foi em geral prejudicada. Adicionalmente, o trabalho de Ruß (2012) constatou que essa abordagem é uma maneira ingênua de agrupar dados espaciais, pois pode enviesar a obtenção de mapas considerando muito fortemente as dissimilaridades no espaço de coordenadas com UGDs em formatos exclusivamente convexos, principalmente quando são utilizados algoritmos de agrupamento particionais. A Figura 4.3 ilustra uma exemplo prático simplificado dessa constatação no contexto de delineamento de UGDs em AP.

Figura 4.3: Exemplos de diferentes mapas de UGDs para uma mesma área hipotética: (a) Utilização de coordenadas geográficas como variáveis do espaço de atributos; (b) Configuração desejada com atributo de altitude sendo determinante no resultado. Em ambos os mapas, cada tom de cinza representa uma CGD distinta, cada uma delas composta por uma única UGD.



(a) (b)

Fonte: Elaborada pelo autor.

Por meio da Figura 4.3 (a), é possível observar a forte influência das coordenadas geográficas na obtenção das UGDs, gerando grupos estritamente convexos quando visualizados em forma de mapa. A Figura 4.3 (b) retrata o que seria a configuração ideal de UGDs para a mesma área hipotética, considerando que essa configuração tenha sido obtida apenas por dados de altitude e sem a utilização das coordenadas geográficas como variáveis do espaço de atributos. Nesse caso, estamos supondo que essa mesma área possua valores mais elevados de altitude na região central, que vão decaindo até as extremidades, formando uma espécie de cone quando esta é visualizada em três dimensões.

Portanto, por meio deste exemplo prático e das constatações em trabalhos e experimentos supracitados, podemos concluir que a simples utilização das coordenadas geográficas como componentes do espaço de atributos não é ideal para uma aplicação como o delineamento de UGDs em AP, pois essa abordagem acaba forçando a convexidade dos grupos quando exibidos em forma de mapa, podendo provocar alterações significativas no resultado que seria o esperado, conforme exibido na Figura 4.3.

Assim, abordagens como a hierárquica aglomerativa, que pode realizar o tratamento do espaço de coordenadas de maneira diferenciada aproveitando as relações de vizinhança espacial, são bastante indicadas para atender a esse tipo de demanda. Nesse caso, como em cada passo do algoritmo é realizada a fusão dos dois grupos mais similares com relação ao espaço de atributos, essas relações podem servir para restringir o espaço de busca (PERRUCHET, 1983), fundindo-se apenas grupos espacialmente adjacentes e, consequentemente, espacialmente contíguos em

todos os níveis do dendrograma (PERRUCHET, 1983; DAVIDSON; RAVI, 2005; MORALES; MENDIZABAL, 2010). Essas restrições ainda podem ser amenizadas, sendo aplicáveis apenas até um limiar de heterogeneidade dos grupos com relação ao espaço de atributos (MARGULES; FAITH; BELBIN, 1985; RUSS; SCHNEIDER; KRUSE, 2010; RUSS, 2012). No caso dessas abordagens, uma subdivisão inicial a partir de algoritmos de particionamento utilizando exclusivamente o espaço de coordenadas, gerando mapas iniciais similares a uma tesselação de Voronoi[3], pode ser utilizada para inicializar o dendrograma, agrupando inicialmente amostras espacialmente mais próximas e reduzindo a sua quantidade de passos. Ao contrário da utilização das coordenadas diretamente como variáveis do espaço de atributos, esse artifício, também conhecido como tesselação inicial, não limita a obtenção da configuração final de UGDs em formatos convexos (CHOO, 2007; RUSS, 2012). Entretanto, por conta de o valor de k ser um parâmetro normalmente empírico e determinado pelo usuário, deve ser utilizado com cautela, para que resultados como o da Figura 4.3 (a) sejam evitados.

Com relação às abordagens de agrupamento com sobreposição, estas podem ser úteis para dados espaciais se imaginarmos, por exemplo, que as incertezas tratadas por elas estão intrínsecas na precisão maior ou menor do dispositivo de captura de coordenadas utilizado na obtenção dos dados. Em se tratando de UGDs em AP, a incerteza pode estar associada também aos limites de separação entre uma UGD e outra por conta de não ocorrerem, por exemplo, mudanças bruscas relacionadas ao solo e as plantas em intervalos espaciais muito pequenos (KITCHEN, 2005; LI, 2007; MORARI; CASTRIGNANÒ; PAGLIARIN, 2009; SONG, 2009; XIN-ZHONG, 2009). Algumas extensões dos algoritmos de agrupamento com sobreposição podem trazer melhores resultados para dados espaciais, a partir da utilização de distâncias adaptáveis a distância Euclidiana (GUSTAFSON; KESSEL, 1978) e estimativas de máxima verossimilhança (GATH; GEVA, 1989), com o intuito de detectar grupos com diferentes formas geométricas, tamanhos e densidades no espaço de atributos, que podem contribuir para a redução da estratificação no espaço de coordenadas.

As medidas de dissimilaridade utilizadas pelos algoritmos de agrupamento também podem influenciar nos resultados obtidos, especialmente quando existe correlação entre as variáveis do espaço de atributos e do espaço de coordenadas, como é o caso dos dados espaciais produzidos em AP. A distância Euclidiana, em sua forma original, requer que as variáveis utilizadas possuam variância semelhante entre si, fato que raramente acontece com esses dados, pois são capturados por diferentes tipos de sensores, constituindo diferentes intervalos e escalas de medidas (ZHANG, 2010). Para compensar essa variância desigual, é importante que os dados de

[3]Tipo especial de decomposição de um espaço, determinado pela distância para uma determinada família de objetos nesse espaço.

entrada referentes aos atributos convencionais sejam normalizados em uma escala única, possuindo, por exemplo, valores entre 0 e 1 ou entre -1 e 1. Essa normalização deve ser realizada em uma etapa de transformação que pode ser considerada anteriormente à mineração de dados, tanto no KDD quanto no KDSD, e que em algumas abordagens pode ser inserida como uma atividade da etapa de pré-processamento. A partir da normalização dos dados de entrada, é garantido que nenhum atributo convencional influenciará mais do que outro no resultado final fornecido pelo algoritmo de agrupamento.

A partir dessa análise, entende-se que atividades que implicam em tomadas de decisão a partir de conjuntos de dados espaciais, como o delineamento de UGDs em AP, devem utilizar abordagens que realizam o tratamento da complexidade desse tipo de dado considerando o espaço de coordenadas de maneira diferenciada com relação ao espaço de atributos, com o intuito de se obter resultados mais satisfatórios e que permitam um melhor entendimento por parte do usuário final. Segundo Estivill-Castro e Lee (2000, 2002), o agrupamento de dados espaciais mostra um conflito de interesses entre a comunidade de SIG, mais interessada na qualidade dos resultados obtidos; e a comunidade de mineração de dados, mais interessada na eficiência computacional. Dessa forma, métodos que permitem análises exploratórias por parte dos usuários podem atender as necessidades dessas duas comunidades, incorporando características como a proximidade espacial ou adjacência que fazem com que sejam obtidos agrupamentos de qualidade e de fácil interpretação. No âmbito do delineamento de UGDs em AP, algumas características desejáveis adicionais podem ser consideradas, tais como: considerar que os dados de entrada possuem densidade espacial constante por conta da interpolação espacial que deve ser realizada previamente; e a manutenção da contiguidade espacial das CGDs o tanto quando for possível, para facilitar o seu entendimento por parte do usuário final (RUSS, 2012).

Por conta da utilização do espaço de coordenadas para exibição, em forma de mapa, de agrupamentos obtidos a partir de conjuntos de dados espaciais, pode ser interessante, em alguns casos, que os grupos formados (ou UGDs, no contexto desta tese) sejam convertidos em polígonos. Assim, técnicas que permitem delinear polígonos a partir de amostras representadas por pontos, como o algoritmo do caixeiro viajante (GUTIN; PUNNEN, 2006) (*TSP*[4]), devem ser considerados. O algoritmo TSP é baseado em uma heurística que tem o objetivo de encontrar o menor caminho, a partir de um ponto qualquer, passando por todos os pontos uma única vez e regressando ao ponto inicial. As arestas obtidas a partir da construção desse caminho podem ser utilizadas então para o delineamento de polígonos.

O Capítulo 5 descreve resumidamente diversas abordagens e algoritmos específicos para

[4]Do inglês *Traveling Salesman Problem*

agrupamento de dados espaciais encontrados na literatura, e quais de suas características foram extraídas para compor o desenvolvimento da abordagem de agrupamento espacial para delineamento de UGDs em AP desenvolvida nesta tese.

4.3.4 Métodos para Validação de Mapas de UGDs

A validação de agrupamentos considerando as características do conjunto de dados utilizado e da aplicação em si, com o intuito de se escolher o melhor agrupamento dentre diversas soluções disponíveis não é tarefa simples. No âmbito do delineamento de UGDs em AP, onde essa escolha significa também definir a quantidade de CGDs em que será dividida uma área de cultivo, a metodologia mais convencional necessita do conhecimento do usuário final. Em diversas abordagens da literatura, o usuário determina um intervalo possível de quantidades de CGDs, normalmente entre 2 e 5, seja de maneira subjetiva ou a partir de algum conhecimento sobre a área em questão, e calcula individualmente a acurácia de cada solução a partir de experimentos comparativos dos mapas finais de UGDs com relação a variáveis isoladas (JAYNES; COLVIN; KASPAR, 2005; KITCHEN, 2005; KHOSLA, 2010; PEDROSO, 2010; MORAL; TERRÓN; REBOLLO, 2011). Esses experimentos podem ser realizados utilizando coeficientes de correlação como Pearson, já descrito na Subseção 4.2.1 com a função de viabilizar a seleção de atributos em um conjunto de dados; e Kappa, já descrito na Subseção 4.2.5.1 com a função de verificar a correlação existente entre duas soluções de agrupamento. Outras abordagens definem essa quantidade considerando restrições para que as UGDs definidas sejam úteis na prática, como o número máximo possível de porções distintas que podem ser configuradas em um equipamento convencional para aplicação de insumos agrícolas (ORTEGA; SANTIBÁÑEZ, 2007).

Em se tratando de abordagens para o delineamento de UGDs baseadas em algoritmos de agrupamento com sobreposição, a quantidade considerada ideal de CGDs é normalmente encontrada a partir da convergência dos valores fornecidos pelos índices NCE (*Normalized Classification Entropy*) e FPI (*Fuzziness Performance Index*). O NCE (BEZDEK, 1981) é uma medida para estimar o grau de desorganização criado por um determinado número de grupos em um agrupamento. Já o FPI (ODEH; CHITTLEBOROUGH; MCBRATNEY, 1992; BOYDELL; MCBRATNEY, 2002) estima o grau de incerteza na definição de uma determinada quantidade de grupos de um agrupamento, a partir do nível de compartilhamento do grau de pertinência entre eles. Apesar de amplamente utilizados, em determinados conjuntos de dados de AP esses índices podem não convergir com relação à quantidade ideal de CGDs a ser utilizada, causando dúvidas ao usuário final sobre qual estratégia adotar (BROCK, 2005).

Diante dessas dificuldades, Zhang (2010) desenvolveram um novo método para determina-

ção da quantidade ideal de CGDs, que leva em consideração a variância média das variáveis presentes no espaço de atributos. A partir de estudos de conjuntos de dados de AP, os autores notaram que a variância entre as amostras associadas a cada CGD diminui consideravelmente conforme a quantidade de classes aumenta, com tendência a se estabilizar a partir de certa quantidade de classes estabelecida. Desse modo, foi criado um algoritmo iterativo, que determina o número ideal de CGDs quando duas condições são satisfeitas: a redução total da variância das amostras associadas a cada CGD for de 50%, a partir da variância obtida no passo inicial; e a redução consecutiva dessa variância, ou seja, a redução entre os passos k e $k+1$ do algoritmo, for inferior a 20%. A segunda condição também pode ser determinada por uma quebra de tendência, ou seja, se a variância aumenta ao invés de diminuir.

Apesar de essas metodologias poderem ser consideradas como critérios de validação interna relativos e, portanto, guiarem o usuário na escolha de um mapa ideal de UGDs dentre diversas soluções disponíveis, alguns fatores devem ser levados em consideração. Além da possibilidade de não convergência dos valores obtidos pelos índices NCE e FPI, essas medidas só podem ser obtidas a partir de mapas de UGDs gerados utilizando algoritmos de agrupamento com sobreposição, que nem sempre são utilizados para este fim. Já na abordagem desenvolvida por Zhang (2010), apesar dos autores não reportarem a sua eficácia de maneira detalhada, é importante notar que apenas uma medida de coesão interna de agrupamentos foi utilizada, sem considerar o importante fator de separação entre os grupos.

Considerando essas questões, foi constatada a necessidade de se encontrar um critério de validação interna relativo que avalie de maneira equilibrada tanto a coesão quanto à separação existente em mapas de UGDs com variância mínima e capazes de proporcionar bons resultados na prática. Dentre os critérios de validação interna para agrupamentos de dados convencionais, entende-se que o critério SD pode ser adequado para o contexto do delineamento de UGDs em AP, pois procura encontrar agrupamentos com mínima variância interna e bem separados com relação aos centroides no espaço de atributos, características desejáveis para UGDs em AP. Mesmo que o critério SD ou outros similares possam ser técnicas adequadas para avaliar a eficácia dos mapas de UGDs, nenhum deles se preocupa em avaliar o arranjo das CGDs no espaço de coordenadas, ou seja, se as CGDs delineadas possuem um nível de estratificação excessivo que pode prejudicar a análise visual do usuário. Assim, foi verificada a necessidade de desenvolvimento de novos critérios ou de modificações nos critérios já existentes, para que os agrupamentos obtidos a partir de dados espaciais possam ser avaliados de uma maneira equilibrada com relação à complexidade desse tipo de dado. Essa necessidade proporcionou o desenvolvimento de uma extensão para o critério SD para análise de agrupamentos espaciais, e que se constituiu em uma das principais contribuições desta tese.

4.4 Considerações Finais

As definições teóricas apresentadas neste capítulo são de grande importância e serviram como motivação para o desenvolvimento das duas atividades mais importante e que constituem as principais contribuições desta tese: o desenvolvimento de uma nova abordagem para agrupamento de dados espaciais pontuais e com densidade constante, obtida a partir de interpolação espacial realizada na etapa de pré-processamento do KDSD; e a criação de um critério de validação interna relativo que permite medir o nível de estratificação de agrupamentos de dados espaciais quando exibidos em forma de mapa.

O Capítulo 5, a seguir, apresenta uma revisão dos trabalhos presentes na literatura e que estão relacionados às contribuições desenvolvidas no âmbito desta tese.

Capítulo 5

REVISÃO DA LITERATURA

5.1 Considerações Iniciais

Neste capítulo, são apresentados os trabalhos correlatos (ou relacionados) estudados e avaliados durante a revisão da literatura. As constatações obtidas a partir dessa revisão, nas quais foram verificadas as limitações dos trabalhos existentes e as possíveis novas contribuições para a área de mineração de dados espaciais aplicada ao contexto da AP, são parte da motivação para a pesquisa realizada ao longo do desenvolvimento desta tese. O capítulo está organizado da seguinte forma:

- A Seção 5.2 descreve os trabalhos relacionados à tarefa de agrupamento de dados espaciais, divididos em três diferentes abordagens: baseadas em densidade; com características aplicáveis ao contexto desta tese; e desenvolvidas especificamente para a tarefa de delineamento de unidades de gestão diferenciada (UGDs) em AP.
- A Seção 5.3 descreve trabalhos relacionados a modelos, processos, fluxos de dados e infraestrutura computacional desenvolvidos para auxiliar em tarefas relacionadas à AP. Algumas características desses trabalhos foram utilizadas durante o desenvolvimento do sistema protótipo.
- A Seção 5.4 descreve experimentos preliminares utilizando dados reais e abordagens de agrupamento que podem ser consideradas, a partir da revisão da literatura realizada, como sendo o estado da arte no delineamento de UGDs em AP.
- A Seção 5.5 finaliza o capítulo com as considerações finais.

Os trabalhos descritos nas Subseções 5.2.2 e 5.2.3 foram selecionados com o auxílio da

técnica de revisão sistemática, aplicada em bases de pesquisa disponíveis na *Web* com a utilização da ferramenta StArt (ZAMBONI, 2010). Os procedimentos realizados nessa revisão estão descritos detalhadamente no Apêndice H desta tese.

5.2 Abordagens para Agrupamento de Dados Espaciais

Nesta seção, estão descritos os trabalhos relacionados à tarefa de agrupamento de dados espaciais estudados e analisados durante a revisão da literatura. Inicialmente, são exploradas abordagens mais tradicionais de agrupamento de dados espaciais, baseadas em densidade. Na sequência, são exploradas abordagens que possuem características aplicáveis aos conceitos desta tese. Finalizando, são exploradas abordagens desenvolvidas especificamente para o contexto do delineamento de UGDs em AP.

5.2.1 Abordagens Baseadas em Densidade

As abordagens baseadas em densidade são as mais tradicionalmente utilizadas para agrupamentos de dados espaciais. O algoritmo DBSCAN (ESTER, 1996; XU, 1998) foi um dos primeiros a ser desenvolvido, servindo como base para a maioria das soluções subsequentes. Em sua forma original, esse algoritmo tenta identificar grupos em conjuntos de dados em um determinado espaço considerando a ideia de que, para cada amostra pertencente a um grupo, a sua vizinhança, considerando um raio ε, deve conter um número mínimo de amostras (*MinPts*), onde ε e *MinPts* devem ser informados pelo usuário final. Com isso, espera-se que as amostras centrais de um grupo possuam uma vizinhança bastante densa, quando comparadas com as amostras de borda desse mesmo grupo, ou a eventuais ruídos presentes no conjunto de dados.

Tomando como base essa ideia central, diversas abordagens surgiram com o intuito de identificar grupos e isolar conjuntos de dados ruidosos com base na densidade amostral dos dados no espaço. Boa parte dessas abordagens utiliza-se de estruturas hierárquicas, que podem ser baseadas em técnicas como: particionamento da área espacial em células retangulares (WANG; YANG; MUNTZ, 1997) ou em subespaços (AGRAWAL, 1998); ordenação de grupos para determinação de parâmetros para grupos multinível (ANKERST, 1999); triangulação de Delaunay[1] (DELAUNAY, 1934; ESTIVILL-CASTRO; LEE, 2000; GUO; PEUQUET; GAHEGAN, 2003; CHOO, 2007); e árvores ou grafos para estabelecimento de hierarquias e verificação de similaridade (GUO; PEUQUET; GAHEGAN, 2003; ASSUNCÃO, 2006; CHOO, 2007; KHANI, 2013). Outras utilizam estratégias

[1] Triangulação DT(P) para um conjunto de pontos P onde nenhum ponto em P é interno à circunferência formada por qualquer triângulo presente em DT(P).

mais específicas, tais como: cadeias de Markov com saltos reversíveis (PEI, 2009); gravidade entre amostras (ZHONG; LIU; LI, 2010); entropia espacial (WANG; WANG, 2011); e representação final dos grupos na forma de polígonos (AKDAG; EICK; CHEN, 2014).

Apesar das amplas possibilidades proporcionadas pelas abordagens acima, o fato de serem baseadas na densidade amostral inviabiliza o seu uso na tarefa de delineamento de UGDs em AP, cujos conjuntos de dados espaciais possuem densidade constante no espaço de coordenadas após a interpolação em uma grade espacial única. Além disso, outras características de algumas dessas abordagens também dificultam a sua utilização nessa aplicação, tais como: considerar a verificação de relacionamentos entre as variáveis apenas no espaço de atributos (AGRAWAL, 1998; ANKERST, 1999); gerar grupos considerando apenas o espaço de coordenadas (ESTER, 1996; XU, 1998; WANG; YANG; MUNTZ, 1997; ESTIVILL-CASTRO; LEE, 2000; PEI, 2009; KHANI, 2013); e não ser facilmente extensível ao agrupamento de conjuntos de dados contendo mais de um atributo (WANG; YANG; MUNTZ, 1997). Por conta dessas características, os algoritmos desenvolvidos para atender a essas abordagens podem produzir agrupamentos não significativos para o delineamento de UGDs em AP.

5.2.2 Abordagens com Características Aplicáveis à Tarefa de Delineamento de UGDs em AP

As abordagens descritas nesta seção possuem características interessantes e com potencial para aplicação no contexto do delineamento de UGDs em AP. Entretanto, da maneira como essas características foram desenvolvidas ou estão sendo utilizadas por essas abordagens, seriam necessárias adaptações para que pudessem ser utilizadas diretamente nesse contexto. Essas abordagens estão descritas a seguir.

5.2.2.1 MOSAIC

O algoritmo MOSAIC (CHOO, 2007) constitui-se da implementação de uma abordagem de agrupamento hierárquica aglomerativa, desenvolvido com o objetivo principal de superar a limitação das abordagens tradicionais em encontrar apenas grupos com formas convexas considerando um espaço euclidiano bidimensional. Esse algoritmo é iniciado com uma etapa para criação de pequenos grupos convexos a partir de algoritmos de particionamento tradicionais, chamada de tesselação inicial. Na etapa seguinte, esses pequenos grupos são fundidos a partir de um grafo de proximidade, criado com base na triangulação de Delaunay e na tesselação de Voronoi, permitindo a formação de grupos não convexos.

A principal limitação do MOSAIC, que impede a sua aplicação direta no delineamento de UGDs em AP, é a não diferenciação do espaço de atributos com relação ao espaço de coordenadas em nenhuma de suas etapas, cujos efeitos já foram relatados na Subseção 4.3.3 do Capítulo 4 e no artigo aceito para publicação, transcrito no Apêndice D desta tese. Assim, ao utilizar as variáveis do espaço de coordenadas como atributos convencionais, os agrupamentos obtidos por essa abordagem para representar as UGDs podem se tornar estritamente convexos quando exibidos em forma de mapa (ou seja, utilizando o espaço bidimensional que representa os atributos de latitude e longitude), fazendo com que o principal objetivo do algoritmo MOSAIC não possa ser atingido.

Entretanto, a ideia da tesselação inicial pode ser válida para o contexto do delineamento de UGDs em AP. Nessa etapa, pode ser gerada uma grande quantidade de pequenos grupos iniciais a partir da utilização apenas das variáveis do espaço de coordenadas, levando-se em consideração os princípios básicos da geoestatística de que amostras muito próximas no espaço de coordenadas possuem valores muito similares no espaço de atributos. Mesmo que os grupos gerados na tesselação inicial possuam formas convexas (por terem sido gerados a partir do particionamento de uma grade regular espacial), a utilização dos atributos convencionais em passos seguintes, onde o objetivo final é gerar agrupamentos com poucos grupos e, consequentemente, mapas com poucas UGDs, possibilita que sejam obtidas UGDs com formas não convexas capazes de retratar com fidelidade as características da lavoura. Além disso, a criação de k grupos iniciais, com valores de k menores do que n, com n sendo a quantidade total de amostras do conjunto de dados, é importante para reduzir a quantidade de passos do dendrograma e, consequentemente, o custo computacional de abordagens hierárquicas. Apesar dessas vantagens, determinar o valor de k de maneira empírica, ou seja, sem considerar a variabilidade do conjunto de dados no espaço de atributos, pode prejudicar o resultado final, evidenciando a necessidade de se tratar essa questão no desenvolvimento de novas abordagens de agrupamento para o delineamento de UGDs em AP.

5.2.2.2 REDCAP

Outra abordagem utilizando agrupamento hierárquico aglomerativo é encontrada no algoritmo REDCAP (GUO, 2008). Nessa abordagem, são incluídas estratégias de contiguidade espacial total e de primeira ordem a partir dos algoritmos *single-linkage*, *average-linkage* e *complete-linkage*, gerando seis combinações distintas de algoritmos de agrupamento. A estratégia de ordem total exige a utilização das distâncias no espaço de atributos de todas as amostras de um grupo A com relação a todas as amostras de um grupo B, durante o cálculo de similari-

dade entre eles. A estratégia de primeira ordem, por sua vez, define que sejam utilizadas, para a realização desse mesmo cálculo, apenas as distâncias no espaço de atributos das amostras de um grupo A espacialmente vizinhas às amostras de um grupo B.

Uma característica relevante do REDCAP é a de permitir a fusão apenas de grupos espacialmente contíguos, identificados por meio de uma árvore de contiguidade espacial atualizada em cada passo de sua execução. No entanto, essa restrição rígida pode não ser desejável em todos os casos de delineamento de UGDs em AP. A própria definição de CGDs, abordada no Capítulo 2, sugere a existência de casos onde o tratamento da lavoura deve ser o mesmo em diferentes regiões de um talhão. De qualquer maneira, avaliações realizadas por Guo (2008) em bancos de dados reais mostram que estratégias mais sofisticadas - como as que utilizam a restrição espacial de ordem total em conjunto com os algoritmos *complete-linkage* e *average-linkage* - tendem a tratar melhor a heterogeneidade do espaço de atributos na formação dos grupos, quando comparadas a estratégias mais simples - como a que utiliza a restrição espacial de primeira ordem em conjunto com o algoritmo *single-linkage*. Essas constatações são importantes para mostrar que estratégias mais simples, apesar de mais eficientes por serem, em geral, menos complexas computacionalmente, não teriam a mesma adequação das mais sofisticadas no âmbito do delineamento de UGDs em AP.

5.2.2.3 Linha de Varredura

Ainda considerando abordagens hierárquicas aglomerativas, Zalik e Zalik (2009) desenvolveram um algoritmo baseado em linha de varredura para agrupar dados espaciais. Os grupos são gerados conectando-se pontos suficientemente próximos a partir de um limiar fornecido pelo usuário, considerando a distância Euclidiana no espaço de atributos como medida de similaridade. Esse processo é realizado em passos a partir de duas linhas de varredura imaginárias, traçadas horizontalmente na área a ser avaliada, com uma distância de separação também fornecida pelo usuário.

Os experimentos realizados pelos autores não deixam claro se foram utilizados dados espaciais com densidade constante, o que não nos permite afirmar que este algoritmo possa ser diretamente utilizado para o delineamento de UGDs em AP. Além disso, o espaço de coordenadas é utilizado de maneira limitada, considerando apenas relações espaciais de vizinhança horizontais. Entretanto, o resultado final do agrupamento por linha de varredura permite a obtenção de grupos não convexos, que é uma característica desejável na formação dos mapas de UGDs.

5.2.2.4 Agrupamento Espacial Híbrido

O conceito de agrupamento espacial híbrido é introduzido pela abordagem criada por Fan (2009). Desenvolvida para atender uma aplicação de seleção de localizações para centrais de serviços ao consumidor, essa abordagem propõe inicialmente a criação de um armazém de dados espaciais contendo informações dos clientes a serem atendidos. Algoritmos tradicionais, como *k-means* e recozimento simulado (KIRKPATRICK; GELATT; VECCHI, 1983; CERNY, 1985) são utilizados, respectivamente, na tarefa para redução do espaço amostral e como estratégia de busca por soluções ótimas. Além disso, são utilizadas funções e relacionamentos espaciais para análise de obstáculos críticos ao agrupamento espacial, como restrições ambientais, de terreno e de tráfego.

Essa abordagem é divida em três algoritmos principais. O primeiro algoritmo é responsável por gerar predicados que indicam relações espaciais entre a localização dos consumidores e diversos elementos que devem ser levados em consideração para restringir o acesso a centrais de serviços, como estradas de ferro e colinas. O segundo algoritmo é responsável por calcular as distâncias entre pontos de consumidores, considerando os obstáculos entre eles (e.g. rios, montanhas). O terceiro algoritmo, por sua vez, é dividido em uma série de algoritmos componentes, responsáveis por: adicionar peso aos pontos de consumo, relacionados ao nível de serviço exigido; obter centroides de grupos candidatos a centrais de serviços; eliminar candidatos devido a restrições espaciais; e buscar pela solução ótima considerando os candidatos restantes.

Apesar de essa abordagem ser uma solução bastante completa para atender às necessidades da localização de centrais de serviços, a sua adaptação por completo para aplicações como o delineamento de UGDs em AP pode ser bastante complexa. Pelo fato de realizar o agrupamento considerando unicamente o espaço de coordenadas, essa abordagem gera grupos estritamente contíguos, o que não é uma característica desejável para a formação das UGDs. No entanto, o tratamento de obstáculos no cálculo da distância entre duas amostras, realizado pelo segundo algoritmo, pode ser útil para a aplicação desta tese. Esse algoritmo leva em consideração o conceito de distância obstruída entre dois objetos, definido como sendo a menor distância Euclidiana entre eles sem que um obstáculo seja cruzado (ZAIANE; LEE, 2002). A partir dessa definição, utiliza o menor caminho espacial para encontrar a distância entre duas amostras com um obstáculo localizado entre elas. Apesar da distância calculada por esse algoritmo também considerar apenas o espaço de coordenadas, nada impede que o espaço de atributos seja utilizado. Para o caso da AP, a distância no espaço de atributos entre duas amostras separadas por um obstáculo pode ser calculada com a utilização de pontos espacialmente intermediários entre elas.

5.2.2.5 FCM com Informação Espacial

A abordagem de Wang e Bu (2010), foi criada com o intuito de adaptar o algoritmo FCM para o tratamento de distribuições espaciais de pixels aplicado à segmentação de imagens. A principal alteração com relação ao FCM original está na função que calcula a pertinência de um pixel com relação a um grupo. Para tanto, é necessário que seja calculada uma função de características locais de cada pixel, com base no pixel central de uma janela com dimensões predefinidas pelo usuário. Essa função é composta por uma correlação espacial do espaço de coordenadas, baseada na distância espacial do pixel com relação ao pixel central dessa janela; e por um cálculo realizado no espaço de atributos, que verifica a proximidade dos valores de níveis de cinza do pixel com relação ao pixel central dessa janela. Os valores da função de características locais atribuídos a cada pixel são utilizados como fator de ponderação para uma função de pertinência modificada, utilizada pelo FCM durante o agrupamento. Com essas modificações, as regiões homogêneas terão sua pertinência com relação a um grupo reforçadas. Por outro lado, se forem encontrados pixels ruidosos, os pesos terão valores menores e classificações incorretas poderão ser corrigidas.

Como essa abordagem foi desenvolvida com o intuito de resolver exclusivamente problemas de segmentação de imagens, tentando separar, por exemplo, objetos presentes na imagem de componentes de fundo e ruídos, a sua aplicação por completo no delineamento de UGDs em AP não deve ser considerada. No entanto, a distribuição espacial dos pixels, realizada em grade regular, segue os mesmos princípios da distribuição espacial de pontos com densidade constante utilizada em dados interpolados de AP. Além disso, essa abordagem considera separadamente o espaço de coordenadas, representado pelo plano cartesiano bidimensional, do espaço de atributos, representado pelos níveis de cinza, em sua função para cálculo de características locais. A utilização dessa função durante o processo de agrupamento do FCM mostrou-se uma alternativa interessante para que essa abordagem considere as diferenças existentes entre o espaço de coordenadas e o espaço de atributos.

5.2.2.6 Agrupamento de Polígonos Considerando Restrições

As abordagens descritas anteriormente tratam exclusivamente do agrupamento espacial de pontos ou pixels. A abordagem CPSC (*Constrained Polygon Spatial Clustering*) (JOSHI; SOH; SAMAL, 2012), foi desenvolvida com o intuito de resolver as dificuldades encontradas com a utilização de algoritmos de agrupamento espacial específicos para pontos, em abordagens para agrupamento de polígonos. Essas dificuldades estão relacionadas à necessidade de manutenção da contiguidade espacial e às características topológicas mais elaboradas que os polígonos

possuem, quando comparados aos pontos.

A abordagem CPSC é baseada em um algoritmo de busca que utiliza as funções heurísticas restritivas F e F', utilizadas para selecionar, respectivamente, o melhor grupo para crescimento e o melhor polígono a ser associado a esse grupo em cada iteração. Essas funções são compostas por parâmetros que consideram restrições espaciais, como a manutenção da contiguidade espacial e de compactação em nível de grupo, bem como restrições no espaço de atributos. Além disso, permitem verificar o efeito de crescimento de um determinado grupo no conjunto todo. A quantidade de grupos desejada (k) deve ser informada como parâmetro, servindo para determinar, aleatoriamente ou com base em alguma restrição, os k polígonos individuais considerados como sementes para os k grupos iniciais. Os polígonos restantes devem ser associados aos seus respectivos grupos seguindo os valores determinados por F e F'. Inicialmente, é escolhido o melhor grupo (BC) para crescimento, com base na função F calculada para cada um dos grupos pertencentes ao conjunto. Com a escolha de BC, é determinado então um subconjunto de polígonos candidatos que podem ser associados a ele, levando em consideração relações de adjacência espacial. Na sequência, deve ser selecionado o melhor polígono (BP) desse subconjunto, para que seja associado à BC. A abordagem pode entrar em estado de *deadlock*, caso o polígono selecionado já tenha sido associado a um grupo C_d em uma iteração anterior. Nesse caso, BC e C_d são retirados da lista de grupos para crescimento, e um novo BC é determinado, com base nos valores de F dos grupos remanescentes nessa lista.

Uma característica importante dessa abordagem é o tratamento realizado com relação aos valores de F e F'. Enquanto BC é escolhido considerando o valor máximo de F, a escolha de BP considerada o valor mínimo de F'. Esse tratamento faz com que os grupos cresçam simultaneamente sem interferir nas restrições espaciais, que já são satisfeitas a partir da escolha dos polígonos candidatos. Se BP fosse considerado, por exemplo, como o polígono com maior valor de F' entre os candidatos, a sua associação a BC manteria esse grupo bastante compacto e com grandes possibilidades de ser escolhido novamente para crescimento na próxima iteração. Considerando o valor mínimo de F', BC é penalizado de certa forma, possibilitando que outro grupo seja o escolhido. Entretanto, essas características fazem como que a abordagem CPSC não garanta convergência, ou seja, em alguns casos pode acontecer que nem todo polígono seja associado a um grupo. Para resolver esse problema, foi criada a extensão CPSC*, permitindo certo relaxamento das restrições, o que faz com que os polígonos restantes não associados possam se tornar potenciais BCs. Devido a esse relaxamento, pode ocorrer o fato do agrupamento final não satisfazer todas as restrições em nível de grupo. Desse modo, a extensão PS* foi criada, com o intuito de melhorar os resultados fornecidos pela extensão CPSC*, assumindo que grupos podem ser divididos. Essa extensão seleciona o grupo com menor valor de F para

remoção, dividindo-o em dois polígonos menores que são então associados a outros grupos já existentes. Esse processo é repetido até que todas as restrições sejam satisfeitas.

A abordagem CPSC foi avaliada em aplicações relacionadas à redistribuição distrital para eleições nos Estados Unidos, levando em consideração restrições como o tamanho da população. Para esse tipo de aplicação, essa abordagem se mostrou mais adequada em comparação com outras técnicas, como particionamento de grafos, recozimento simulado e algoritmos genéticos. A extensão PS* foi avaliada utilizando apenas uma base de dados sintética, enquanto que a extensão CPSC* foi avaliada em um problema para obtenção de distritos escolares.

Concluindo, para que essa abordagem pudesse ser aplicada na tarefa de delineamento de UGDs em AP, seriam necessárias adaptações que permitissem a transformação prévia dos conjuntos de dados (originalmente associados a amostras representadas por pontos no espaço de coordenadas) em polígonos, utilizando, por exemplo, artifícios como a poligonização de grupos formados em uma tesselação inicial. Entretanto, a fusão de polígonos durante a construção do dendrograma de uma abordagem hierárquica pode prejudicar muito o custo computacional, uma vez que em cada fusão devem ser executadas funções de geoprocessamento para determinação de novas bordas. Por outro lado, a utilização de funções heurísticas restritivas com o intuito de manter, o quanto for possível, a contiguidade espacial dos grupos formados, é uma característica desejável para que sejam obtidos mapas de UGDs pouco estratificados e de fácil interpretação pelo usuário final.

5.2.3 Abordagens Específicas para Auxílio ao Delineamento de UGDs em AP

Algumas abordagens para agrupamentos de dados espaciais aplicadas especificamente no auxílio ao delineamento de UGDs em AP tem surgido nos últimos anos. Três dessas abordagens, consideradas de extrema importância para fundamentar as contribuições obtidas nesta tese, estão resumidamente descritas a seguir.

5.2.3.1 *HACC-Spatial*

O desenvolvimento da abordagem de agrupamento hierárquica aglomerativa *HACC-Spatial* (*Hierarchical Agglomerative Clustering with Spatial Constraint*) (RUSS KRUSE, 2011; RUSS, 2012) teve como principal motivação a necessidade de atender a requisitos específicos relacionados à tarefa de delineamento de UGDs em AP. Dentre esses requisitos, destacam-se: o tratamento diferenciado do espaço de coordenadas com relação ao espaço de atributos; a pre-

servação, o tanto quanto for possível, da contiguidade espacial dos grupos que representarão as CGDs; e permitir ao usuário uma análise exploratória por meio de árvores hierárquicas, eliminando a necessidade de fixação da quantidade de CGDs desejada.

Segundo Ruß (2012), somente as abordagens hierárquicas aglomerativas específicas para agrupamento de dados espaciais são capazes de atender a todos os requisitos para o delineamento de UGDs em AP. A preservação da contiguidade espacial das CGDs é uma característica desejável para os usuários de AP, pois facilita a interpretação do comportamento da lavoura e, consequentemente, a aplicação espacialmente diferenciada de insumos agrícolas e corretivos. Por outro lado, uma solução que trate essa restrição de maneira rígida, pode proporcionar a formação de CGDs que deixam de exibir alguns fatores importantes para a tomada de decisão. Para tentar atender a esse importante requisito, a abordagem *HACC-Spatial* utiliza-se de um parâmetro para a restrição de contiguidade espacial (*cp*), a ser informado pelo usuário.

A abordagem *HACC-Spatial* pode ser divida em três etapas, e deve ser executada a partir de um conjunto de dados espaciais comumente utilizados em AP, ou seja, com amostras contendo variáveis relacionadas ao comportamento do solo e da cultura no espaço de atributos e associadas a uma grade de pontos regularmente distribuída no espaço de coordenadas, ou seja, com densidade amostral constante. Na primeira etapa, não obrigatória e conhecida como pré-tesselação ou tesselação inicial, as amostras são agrupadas pelo algoritmo *k-means*, considerando apenas o espaço de coordenadas. Segundo Ruß e Kruse (2011), Ruß (2012), essa etapa é importante para que sejam obtidos agrupamentos iniciais contendo amostras espacialmente similares já agrupadas, proporcionando a economia de esforço computacional na etapa principal. Para que a mesma seja executada, o usuário deve informar um valor k menor do que n, onde k corresponde à quantidade inicial de grupos a ser gerada, e n é a quantidade total de amostras do conjunto de dados. Se essa etapa não for executada, a abordagem *HACC-Spatial* deverá ser iniciada da maneira tradicional, ou seja, com n grupos, onde cada grupo é representado por uma única amostra.

Com o agrupamento inicial obtido, as outras duas etapas são realizadas a partir da execução do algoritmo de agrupamento hierárquico *average-linkage*, neste caso considerando a restrição de contiguidade espacial para a fusão de dois grupos proposta com a utilização do parâmetro cp. Em cada passo de construção do dendrograma, são calculadas as matrizes de dissimilaridades, no espaço de atributos, entre pares de grupos espacialmente adjacentes e não adjacentes. A partir dessas matrizes, são selecionados os grupos mais similares nesses dois conjuntos, e realizado o cálculo do critério de contiguidade espacial. Esse cálculo corresponde à razão das distâncias médias entre os grupos espacialmente adjacentes com relação às distâncias médias entre os gru-

pos espacialmente não adjacentes e, enquanto essa razão não superar o limiar cp, informado pelo usuário final, apenas os dois grupos mais similares e espacialmente adjacentes poderão ser fundidos. Os passos de construção do dendrograma realizados até esse momento constituem então a segunda etapa da abordagem. A partir do momento em que o limiar é atingido, a restrição espacial é desabilitada e serão fundidos os dois grupos mais similares, independentemente de estes serem adjacentes ou não, constituindo a terceira etapa da abordagem, que termina quando um único grupo é obtido e a construção do dendrograma é finalizada. Assim, a partir dessa etapa, começam a ser obtidas CGDs que podem ser constituídas por UGDs espacialmente disjuntas, conforme definido na Seção 2.6 do Capítulo 2. Como resultado, é fornecido ao usuário final um conjunto de k agrupamentos, permitindo uma análise exploratória de diferentes mapas de UGDs.

A razão de distâncias utilizada como critério de contiguidade espacial foi estabelecida pelos autores com base na prerrogativa de que as distâncias entre grupos adjacentes e não adjacentes aumenta durante a construção do dendrograma previsto pela abordagem hierárquica aglomerativa. Além disso, a utilização de algoritmos sofisticados para verificação de similaridade entre grupos, como o *average-linkage*, faz com que as distâncias médias entre os grupos adjacentes e não adjacentes aumentem seguindo um mesmo padrão. Desse modo, a razão entre essas distâncias pode ser considerada como um critério estável para determinar o momento correto em que a restrição de contiguidade espacial deve ser desabilitada.

A abordagem *HACC-Spatial* pode apresentar resultados expressivos para a tarefa de delineamento de UGDs em AP, se considerarmos a necessidade dos usuários em obter mapas mais fáceis de interpretar e, consequentemente, com maior aplicabilidade. A característica mais importante dessa abordagem é o tratamento diferenciado do espaço de coordenadas com relação ao espaço de atributos, sem negligenciar a relação existente entre eles a partir da utilização da razão de contiguidade espacial. Entretanto, a determinação, por parte do usuário final, do valor correto para o parâmetro cp, pode ser considerada uma tarefa bastante trabalhosa. Muitas vezes essa tarefa pode ser baseada em tentativa e erro, fazendo com que o valor ideal desse parâmetro para um determinado conjunto de dados não seja facilmente obtido. A característica exploratória do resultado fornecido pela abordagem também pode ser muito útil aos usuários finais, possibilitando a liberdade de análise e escolha do mapa de UGDs mais conveniente para determinadas situações. Por outro lado, essa abordagem faz com que exista a necessidade de análises adicionais a um custo computacional alto para a construção do dendrograma completo. Mesmo que seja realizada a etapa de tesselação inicial, pode ainda existir o desperdício de esforço para o delineamento de mapas de UGDs que podem não ser relevantes à aplicação. Além disso, a tesselação inicial realizada determinando-se a quantidade k de grupos de maneira empírica, e

principalmente, a inicialização aleatória dos centroides desses k grupos, pode proporcionar a obtenção de soluções não determinísticas por parte da abordagem *HACC-Spatial*, gerando dúvidas ao usuário final sobre qual mapa de UGDs utilizar. Essas questões são comprovadas em experimentos realizados no âmbito desta tese, descritos na Seção 5.4.

5.2.3.2 KM-sPC

A abordagem desenvolvida por Córdoba (2013) tem como principal objetivo possibilitar ao usuário de AP a obtenção de mapas de UGDs com estratificação reduzida. Nessa abordagem, o diferencial está relacionado ao tratamento das relações espaciais existentes entre os dados em uma etapa de pré-processamento. Segundo Córdoba (2013), algoritmos de agrupamento como o FCM podem agrupar dados considerando variáveis sintetizadas, como as que são obtidas com a aplicação da redução da dimensionalidade. Entretanto, conforme já descrito no Capítulo 4, abordagens de agrupamento baseadas nesse tipo de algoritmo não incluem em sua essência o tratamento da correlação espacial existente em dados espaciais. A abordagem desenvolvida por Córdoba (2013), conhecida como KM-sPC, não modifica o FCM para tratar a questão da correlação espacial entre as variáveis. Entretanto, utiliza-se da técnica MULTISPATI-PCA para obtenção de componentes principais que não maximizam apenas a variância total entre as amostras, mas também a correlação espacial existente entre as diferentes localizações. Essas componentes, por sua vez, são posteriormente utilizadas pelo FCM para agrupamento dos dados utilizando apenas o espaço de atributos. A partir dessa configuração, é esperada a geração de mapas de UGDs com melhor nível de contiguidade espacial e menor quantidade de ruídos, se comparados aos que seriam produzidos por uma aplicação direta do FCM no conjunto de variáveis do espaço de atributos.

A abordagem KM-sPC pode ser divida em três fases, compostas por diferentes algoritmos e funcionalidades. A fase de pré-processamento realiza operações como a remoção de *outliers*, mapeamento para grade regular e conversão de coordenadas geográficas (em graus) para cartesianas (em metros), com o intuito de facilitar a interpretação visual dos dados. Em seguida, é executada a fase de determinação dos componentes principais considerando as correlações espaciais, a partir da técnica MULTISPATI-PCA. Finalmente, a fase final dessa abordagem realiza o agrupamento dos dados considerando como entrada uma matriz $X_{(n\times a)}$, onde n é a quantidade de amostras e a é o número de componentes principais obtidos na fase anterior, considerando $a < p$, onde p é a quantidade de atributos convencionais. Além disso, deve ser fornecido um valor k para a quantidade de CGDs desejada, obtido por meio da convergência dos valores fornecidos pelos índices FPI (*Fuzzy Performance Index*) e NCE (*Normalized Classification Entropy*).

Como saída, é fornecida uma matriz $U_{(n\times k)}$, contendo os valores de pertinência das n amostras com relação aos grupos.

Essa abordagem foi avaliada pelos autores em três áreas de cultivo de trigo e soja, a partir da utilização de diferentes atributos relacionados ao solo e às plantas. Como medida de comparação para verificar o desempenho da KM-sPC, foram utilizadas configurações alternativas considerando apenas variáveis de solo (KM-SV) e com a redução de dimensionalidade realizada pela PCA clássica (KM-PC). Os resultados obtidos mostraram que a redução de dimensionalidade do conjunto de atributos considerando as correlações espaciais existentes entre eles pode melhorar o resultado final obtido por abordagens de agrupamento que consideram apenas o espaço de atributos. Entretanto, a realização desse tratamento apenas na fase de pré-processamento deixa de considerar questões importantes durante a formação dos grupos, como a adjacência e o tamanho mínimo para a área de uma UGD, que poderiam ser impostos a partir de restrições espaciais e proporcionar uma redução mais efetiva da estratificação. Por conta disso, os mapas de UGDs obtidos podem não atingir um nível de contiguidade espacial que facilite a interpretação visual.

5.2.3.3 *K-Means* Utilizando a Estatística Espacial Gi*

A abordagem desenvolvida por Peeters (2015) possui o mesmo objetivo principal das abordagens *HACC-Spatial* e KM-sPC de reduzir a estratificação dos mapas de UGDs obtidos a partir de conjuntos de dados espaciais. Apesar de ter sido inicialmente proposta e avaliada com a utilização de dados oriundos de pomares de citros, onde as amostras são referenciadas por árvores igualmente espaçadas, essa abordagem também pode ser estendida para outras culturas, se considerarmos a utilização de conjuntos de atributos dispostos em grades espaciais regulares. De maneira similar ao que acontece na KM-sPC, essa abordagem utiliza um algoritmo tradicional de particionamento, nesse caso o *k-means*, para agrupar as amostras de acordo com uma quantidade de grupos k determinada pelo usuário final. Entretanto, ao invés de utilizar componentes principais obtidos a partir do conjunto de dados original, essa abordagem utiliza uma estatística indicativa de autocorrelação espacial entre as amostras, conhecida como Gi* (ORD; GETIS, 1995). Essa estatística é calculada para cada um dos atributos, considerando os valores originais e uma matriz de pesos gerada a partir de uma distância d, determinada pelo usuário final e considerada como limiar de vizinhança para verificação da correlação espacial entre as amostras.

Desse modo, a abordagem possui duas etapas de agrupamento consideradas pelos autores. A primeira consiste na aplicação da estatística Gi* para todos os atributos, onde é fornecido um

valor para cada amostra em cada um deles, conhecido como *z-score*. Por meio do *z-score*, já é possível identificar grupos com autocorrelação espacial para cada atributo, conhecidos como *hot-spots* e *cold-spots*, e grupos espacialmente aleatórios. Em uma segunda etapa, os valores de *z-score* para cada atributo são utilizados como atributos de entrada para a execução do algoritmo *k-means*, considerando uma quantidade k de CGDs determinada pelo usuário final. Os resultados obtidos mostraram mapas de UGDs espacialmente mais contíguos do que se fossem utilizados os atributos convencionais aplicados diretamente no algoritmo *k-means*. Entretanto, assim como ocorre para a abordagem KM-sPC, a realização do tratamento da correlação espacial apenas em uma etapa anterior à realizada pela abordagem de agrupamento particional pode deixar de considerar questões durante a formação dos grupos, que poderiam ser impostas por restrições espaciais e proporcionar uma redução mais efetiva da estratificação.

5.2.4 Sumário das Abordagens para Agrupamento de Dados Espaciais

As abordagens relacionadas ao agrupamento específico para dados espaciais, descritas nesta seção, podem ser utilizadas por diversas aplicações onde o objetivo principal seja agrupar dados não convencionais associados a coordenadas espaciais. Na Tabela 5.1 estão elencadas as principais características que uma abordagem para agrupamento de dados espaciais aplicada no delineamento de UGDs em AP deve levar em consideração, e quais delas são tratadas (marcadas com ✓) ou não (marcadas com ✗) pelas abordagens estudadas nesta seção. Por conta das características já listadas na Subseção 5.2.1, as abordagens baseadas em densidade não serão consideradas nessa tabela.

Tabela 5.1: Abordagens de agrupamento espacial e características desejáveis para aplicação no delineamento de UGDs em AP.

Abordagem	Densidade Constante (Coordenadas)	Separação de Espaços	Contiguidade Espacial	Formas Não Convexas	Obstáculos	Formas Poligonais
Choo (2007)	✗	✗	✗	✓	✗	✗
Guo (2008)	✓	✓	✓	✗	✗	✗
Zalik e Zalik (2009)	✗	✓	✓	✓	✗	✗
Fan (2009)	✗	✓	✓	✗	✓	✗
Wang e Bu (2010)	✓	✓	✓	✗	✗	✗
Joshi, Soh e Samal (2012)	✗	✓	✓	✗	✗	✓
Ruß e Kruse (2011)	✓	✓	✓	✓	✗	✗
Córdoba (2013)	✓	✓	✓	✓	✗	✗
Peeters (2015)	✓	✓	✓	✓	✗	✗

Considerando o contexto da AP, a característica de densidade constante dos dados no espaço de coordenadas após a interpolação espacial nos permite verificar que, abordagens baseadas em densidades distintas para formação de grupos - como é o caso das abordagens descritas na Subseção 5.2.1 e da abordagem de Choo (2007) - não são em geral recomendadas para a aplicação de delineamento de UGDs, que costuma se utilizar de conjuntos de dados espaciais dispostos em uma grade espacial regular, principalmente após o pré-processamento utilizando interpolação espacial. Por outro lado, características das abordagens hierárquicas aglomerativas, como o tratamento da correlação espacial das amostras durante a formação dos grupos, permitindo

a geração de grupos com diferentes formas geométricas não convexas que privilegiam a contiguidade espacial o tanto quanto for possível, são desejáveis e importantes nesse contexto. No entanto, abordagens como a desenvolvida por Guo (2008), que permite fundir exclusivamente grupos espacialmente contíguos, estão limitadas a fornecer mapas onde cada CGD é composta por uma única UGD, o que não é desejável para esse contexto. Outras características, como o tratamento de obstáculos espaciais (FAN, 2009) e a possibilidade de representação das UGDs em estruturas espaciais mais simples e de fácil entendimento - como os polígonos (JOSHI; SOH; SAMAL, 2012) -, são úteis para permitir o delineamento de UGDs mais eficazes para serem utilizadas na prática. Além disso, a representação de UGDs em formato poligonal pode permitir, por exemplo, a redução do custo computacional necessário para a recuperação de informações relacionadas ao histórico de aplicação de insumos agrícolas em cada uma delas. Finalizando, apesar das abordagens de agrupamento espacial específicas para atender à aplicação desta tese, como as desenvolvidas por Ruß e Kruse (2011), Córdoba (2013) e Peeters (2015), serem bastante recentes e com eficácia comprovada na prática pelos autores, ainda não podem ser consideradas soluções de consenso capazes de influenciar diretamente na gestão de uma lavoura.

Concluindo, a partir das vantagens e limitações identificadas nas abordagens de agrupamento espacial descritas nesta seção, as abordagens desenvolvidas por Ruß e Kruse (2011), Córdoba (2013) e Peeters (2015) serão comparadas com a abordagem de agrupamento espacial desenvolvida nesta tese, por conta de terem sido desenvolvidas especificamente para o delineamento de UGDs em AP e realizarem, em uma ou mais etapas ou fases de sua execução, o tratamento diferenciado do espaço de atributos com relação ao espaço de coordenadas. Além disso, o uso tradicional do algoritmo FCM, considerando apenas as variáveis do espaço de atributos, também será comparado com a nova abordagem, já que a maioria das abordagens para delineamento de UGDs em AP que se utilizam de técnicas de mineração de dados já existentes são baseadas nesse algoritmo, conforme descrito no Apêndice A complementar a este capítulo. Assim, essas abordagens podem ser consideradas como componentes do estado da arte para o delineamento de UGDs em AP, no que diz respeito à etapa de mineração de dados presente no processo de KDSD que envolve essa aplicação. Complementarmente, nos experimentos realizados no âmbito desta tese, o algoritmo FCM será identificado como abordagem FCM, representando como um todo as abordagens de delineamento de UGDs que o utilizam.

A validação qualitativa dos resultados obtidos pelas diferentes abordagens de agrupamento descritas nesta seção pode ser realizada por critérios de validação interna, conforme já descrito no Capítulo 4. No caso do delineamento de UGDs em AP, essa validação está totalmente relacionada à obtenção de mapas de UGDs coesos e bem separados, possibilitando, inclusive, a determinação da quantidade ideal de CGDs para cada situação. Algumas das abordagens des-

critas nesta seção utilizam-se de métodos específicos para realizar essa validação. A abordagem de Choo (2007) utilizou o critério da largura de silhueta para verificar as diferenças de coesão e separação com relação aos agrupamentos obtidos por outras abordagens. Já a abordagem de Guo (2008), realizou esse mesmo tipo de comparação utilizando medidas de heterogeneidade global, tamanho de região, variações de intervalos e preservação da distribuição dos dados especificadas por Assuncão (2006). Por se tratarem de aplicações específicas, a abordagem de Fan (2009) realizou a comparação com outras abordagens por meio do cálculo da distância média para centros de atendimento ao consumidor, enquanto que a abordagem de Wang e Bu (2010) utilizou medidas de qualidade específicas para segmentação de imagens. A abordagem de agrupamento poligonal de Joshi, Soh e Samal (2012) realizou a verificação do grau de compactação dos grupos obtidos por meio de um critério específico desenvolvido por Schwartzberg (1965). Finalizando, com relação às aplicações em AP, nas abordagens desenvolvidas por Ruß e Kruse (2011) e Córdoba (2013), a qualidade dos mapas de UGDs obtidos foi avaliada visualmente em experimentos realizados pelos autores. Já na abordagem desenvolvida por Peeters (2015), o critério de validação interna desenvolvido por Caliński e Harabasz (1974) foi utilizado para verificação da quantidade ideal de CGDs utilizada em cada experimento.

De forma a verificar a quantidade ideal de CGDs a ser utilizada em uma área de cultivo, as abordagens para o delineamento de UGDs em AP baseadas em técnicas já existentes, descritas no Apêndice A, utilizam-se de critérios ou técnicas de validação já descritos anteriormente. Especificamente com relação à utilização de critérios, os mais utilizados são os índices FPI e NCE, que podem fornecer soluções divergentes que geram dúvidas ao usuário final. Outras abordagens analisam apenas de maneira visual os dados obtidos, ou seja, consideram como o mapa ideal de UGDs aquele que fornece subdivisões da área de cultivo bem distribuídas considerando apenas o espaço de coordenadas. Desse modo, novas metodologias, como a desenvolvida por Zhang (2010), são válidas para a realização de análises mais precisas. Entretanto, essas novas metodologias devem ter a capacidade de avaliar, de maneira equilibrada, a coesão e a separação dos agrupamentos gerados considerando tanto o espaço de atributos quanto o espaço de coordenadas.

5.3 Modelos para Manipulação de Dados Espaciais em AP

Nesta seção, estão descritos trabalhos relacionados a modelos desenvolvidos para auxiliar especificamente a manipulação de dados em AP. Esses modelos não estão relacionados à principal contribuição desta tese, pois não tratam especificamente de ferramentas para mineração de dados espaciais. Entretanto, algumas de suas características foram úteis ao desenvolvimento

do sistema protótipo utilizado nos experimentos executados no âmbito desta tese.

5.3.1 Infraestrutura para Desenvolvimento de Sistemas de Informação em AP

O trabalho desenvolvido por Nash, Korduan e Bill (2009) teve como principal objetivo a aplicação de serviços OGC para auxiliar a automatização do processamento de dados espaciais agrícolas. Nesse trabalho, os autores utilizaram AP como estudo de caso, devido à grande quantidade e variedade de dados espaciais que podem ser coletados, armazenados, compartilhados e analisados. Uma importante contribuição dessa pesquisa foi mostrar como as cadeias de serviços OGC podem ser úteis na execução de processos paralelos em servidores distintos de maneira transparente aos usuários, para atender tarefas de suporte à tomada de decisão.

Já o trabalho de Murakami (2007) propôs uma infraestrutura completa para desenvolvimento de sistemas de informação para AP, orientada a serviços e baseada em plataformas abertas e padrões para comunicação de dados e interoperabilidade entre sistemas. A principal contribuição desse trabalho foi a criação da PAML (*Precision Agriculture Markup Language*), uma extensão da GML que permite a especialização de dados geográficos em objetos relacionados a dados agrícolas, como talhão, amostras de solo e de produtividade. Dando continuidade a esse trabalho, Ribeiro-Júnior (2007) propuseram novos serviços e funcionalidades a essa estrutura, criando um portal *Web* de serviços agrícolas.

5.3.2 Processos para Delineamento de UGDs em AP

O trabalho desenvolvido por Santos e Saraiva (2015) propôs um modelo de referência para o delineamento de UGDs em AP. A principal contribuição desse trabalho foi a padronização dos processos que devem ser realizados na composição dessa tarefa, permitindo a escolha dos dados, ferramentas e métodos corretos para melhoria da qualidade dos mapas finais. Dentro desse contexto foi definido um processo geral, que contém diversos passos a serem realizados de maneira sequencial para a obtenção desses mapas, tais como: coleta de dados, filtragem de dados, seleção de dados, agrupamento e avaliação dos mapas.

Na coleta de dados, devem ser definidos os atributos que serão utilizados, bem como escolhidas as estratégias e metodologias para amostragem georreferenciada. Em seguida, no passo de filtragem dos dados, devem ser eliminadas amostras que não estão totalmente contidas na área em questão, com erros de posicionamento, valores extremos ou inválidos devido a calibrações equivocadas de sensores. No passo de seleção de dados, devem ser eliminadas variáveis

inconsistentes, irrelevantes ou redundantes do espaço de atributos, com o intuito de não prejudicar a qualidade dos mapas de UGDs que serão gerados. Além disso, nesse passo são realizadas as interpolações espaciais necessárias para que todas as variáveis do espaço de atributos sejam distribuídas em uma mesma grade espacial. No passo de agrupamento, deve ser escolhido um algoritmo que permita a extração de padrões a partir de amostras não classificadas, para geração de agrupamentos que formarão o delineamento das UGDs. Finalmente, no passo de avaliação dos mapas, devem ser selecionados os métodos que permitirão avaliar os mapas finais obtidos no passo anterior.

A especificação desse modelo mostra diversas etapas que fazem parte do processo geral de KDSD. Entretanto, as especificidades propostas relacionadas ao tratamento de dados no âmbito da AP foram observadas e consideradas durante os experimentos realizados no desenvolvimento desta tese.

5.4 Experimentos Preliminares

Levando-se em consideração a revisão da literatura descrita neste capítulo, é possível verificar uma grande incidência no uso da abordagem particional FCM para o delineamento de UGDs em AP. Entretanto, conforme já mencionado anteriormente, essa abordagem não foi originalmente concebida para tratar a complexidade dos dados espaciais. Se por um lado a utilização das coordenadas geográficas apenas para visualização dos agrupamentos também pode proporcionar a obtenção de mapas de UGDs extremamente estratificados, por outro lado o seu uso como variáveis pertencentes ao espaço de atributos pode proporcionar resultados indesejáveis. A abordagem *HACC-Spatial*, por sua vez, trata o espaço de coordenadas de maneira diferenciada com relação ao espaço de atributos, além de proporcionar ao usuário final a possibilidade de explorar soluções obtidas utilizando diversas quantidades de CGDs. Com o intuito de identificar as vantagens e desvantagens da utilização das abordagens FCM e *HACC-Spatial* para o delineamento de UGDs, foram realizados diversos experimentos preliminares, descritos a seguir, utilizando variações nos parâmetros e conjuntos de atributos utilizados por essas abordagens. Além disso, foram realizados experimentos comparativos com a abordagem desenvolvida por Peeters (2015), com o intuito de verificar o efeito das variações do parâmetro relacionado à distância de vizinhos no espaço de coordenadas na etapa de pré-processamento. Esses experimentos complementaram a revisão da literatura, e serviram de base para o desenvolvimento de uma nova abordagem de agrupamento espacial no âmbito desta tese.

Conforme já descrito no Capítulo 4, a utilização das coordenadas geográficas diretamente

como variáveis pertencentes ao espaço de atributos pode causar efeitos indesejáveis nos mapas de UGDs gerados por abordagens de agrupamento convencionais, como a FCM. De maneira a verificar esses efeitos, foram realizados experimentos que estão relatados em artigo específico aceito para publicação em periódico, transcrito no Apêndice D desta tese.

Com relação à abordagem FCM, experimentos preliminares com variações dessa abordagem, já relatando algumas de suas vantagens e limitações, estão descritos no artigo publicado em evento, transcrito no Apêndice C (SPERANZA, 2014) desta tese. Vale ressaltar que, para cada experimento realizado com a abordagem FCM no âmbito desta tese, o algoritmo particional foi executado dez vezes, sendo utilizado, em cada um deles, o resultado contendo a menor soma dos erros quadráticos. Essa é uma prática comum em abordagens que utilizam esse tipo de algoritmo, com o intuito de minimizar os efeitos proporcionados pela inicialização aleatória dos centroides que pode resultar na obtenção de uma solução ótima local ao invés de uma solução ótima global.

Na sequência, foram realizados novos experimentos para verificar o efeito produzido pela variação do parâmetro de *fuzzificação m*, na tentativa de reduzir a estratificação final de mapas de UGDs gerados a partir de certas condições. As figuras 5.1 e 5.2 mostram, respectivamente, mapas contendo 3 CGDs gerados pela abordagem FCM a partir dos conjuntos de atributos UP-CA1 e UP-CA2. Esses dois conjuntos de atributos também foram utilizados em experimentos para validação das abordagens desenvolvidas nesta tese, e estão descritos na Seção 6.5 do Capítulo 6.

Figura 5.1: Mapas contendo 3 CGDs delineadas utilizando a abordagem FCM e variações nos valores do parâmetro *m*, utilizando o conjunto de atributos UP-CA1, onde (a) *m=2*; (b) *m=1,9*; (c) *m=1,7*; (d) *m=1,5*; (e) *m=1,3*; e (f) *m=1,1*.

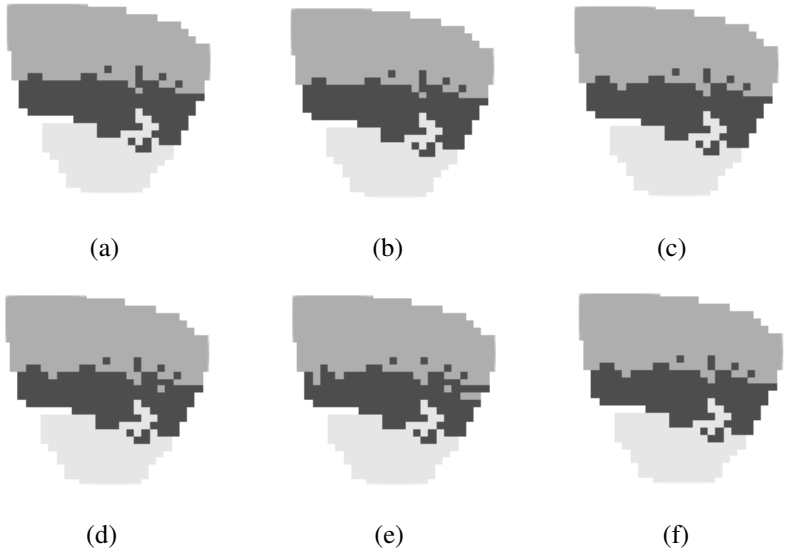
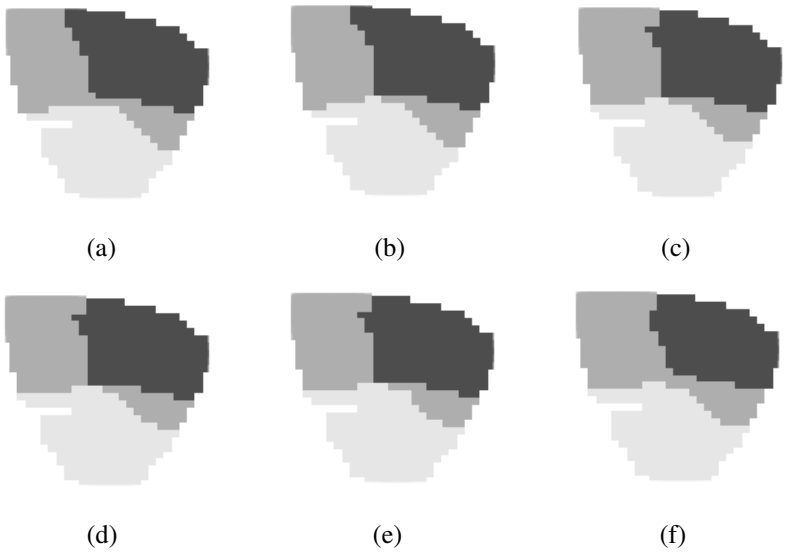
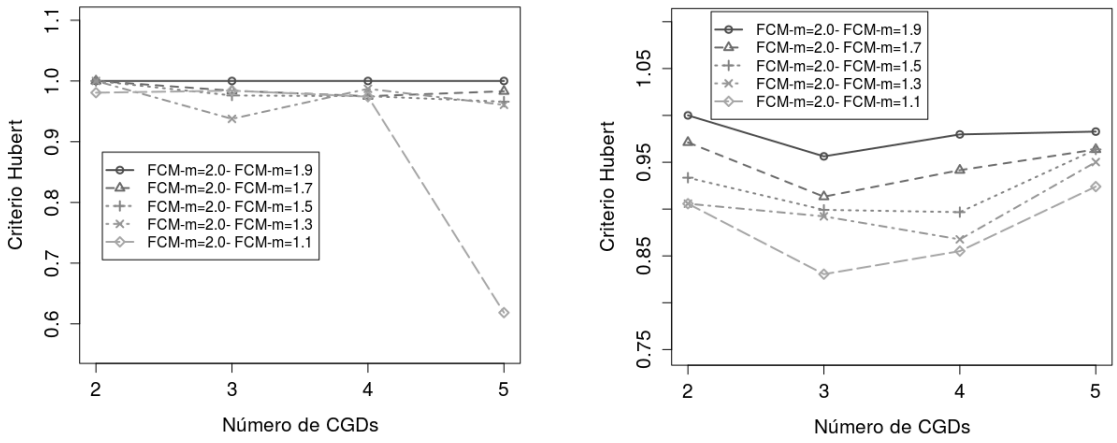


Figura 5.2: Mapas contendo 3 CGDs delineadas utilizando a abordagem FCM e variações nos valores do parâmetro *m*, utilizando o conjunto de atributos UP-CA2, onde (a) *m=2*; (b) *m=1,9*; e (c) *m=1,7*; (d) *m=1,5*; (e) *m=1,3*; e (f) *m=1,1*.



A Figura 5.1 mostra que a variação no parâmetro *m* praticamente não influencia na diminuição ou aumento da estratificação gerada pelos resultados obtidos pela abordagem FCM nos mapas finais de UGDs. Para resultados que não geram mapas estratificados, como os exibidos na Figura 5.2, a variação do parâmetro *m* exerce influência, embora pequena, na definição dos limites das CGDs geradas. Essas pequenas modificações nos mapas podem ser comprovadas a partir dos índices obtidos com a utilização da estatística Γ de Hubert para comparar os mapas obtidos pela abordagem FCM nas figuras 5.1 e 5.2, utilizando como agrupamento de referência o valor padrão de *m* igual a 2 (Figura 5.3).

Figura 5.3: Índices obtidos pela estatística Γ de Hubert comparando uma solução de referência fornecida pela abordagem FCM (*m=2*), com soluções obtidas variando-se o parâmetro *m* e utilizando os conjunto de atributos (a) UP-CA1; e (b) UP-CA2.

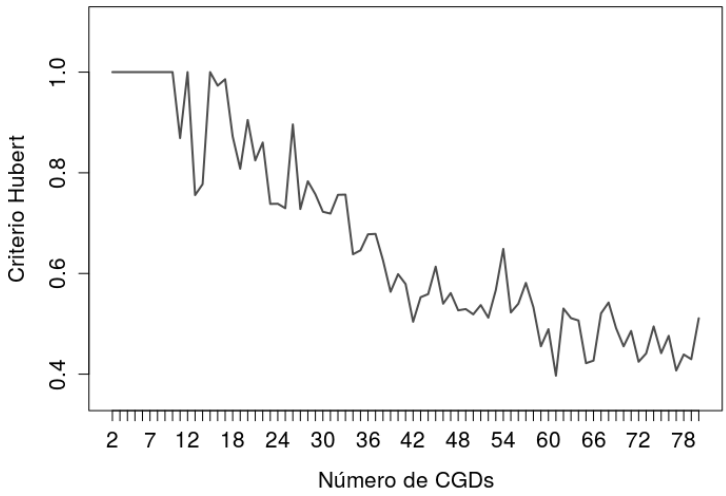


(a) (b)

Os gráficos da Figura 5.3 mostram variações mínimas ou, em alguns casos, a ausência de variação nos mapas finais de UGDs obtidos pela abordagem FCM utilizando variações do parâmetro *m* para ambos os conjuntos de atributos UP-CA1 e UP-CA2. Desse modo, pode-se considerar que o parâmetro *m* proporciona apenas pequenos ajustes no resultado final, sendo que a escolha de um ou outro mapa de UGDs dependerá das necessidades do usuário final. Por outro lado, o fato do grau de pertinência provido pela abordagem FCM ser utilizado na tarefa de delineamento de UGDs apenas para indicar em qual grupo uma amostra será associada (nesse caso, ao grupo com maior pertinência), faz com que sejam obtidos mapas finais de UGDs muito próximos aos mapas fornecidos pelo algoritmo *k-means*, que possui execução similar à abordagem FCM no limite onde $m = 1$. Assim, a importante informação do grau de pertinência fornecida pela abordagem FCM, que é a sua principal vantagem com relação ao *k-means*, poderia ser melhor utilizada pelas abordagens para delineamento de UGDs em AP, como por exemplo, em auxiliar na determinação dos limites apropriados para divisão entre duas UGDs.

Outra questão que deve ser levada em consideração ao se avaliar os resultados fornecidos pelo FCM é com relação ao efeitos causados pela inicialização aleatória dos centroides que irão representar inicialmente os grupos. A Figura 5.4 exibe um gráfico contendo valores obtidos pela estatística Γ de Hubert comparando mapas de UGDs gerados por duas execuções distintas da abordagem FCM utilizando os mesmos parâmetros e atributos de entrada, para quantidades de CGDs entre 2 e 80.

Figura 5.4: Índices obtidos pela estatística Γ de Hubert comparando duas execuções distintas da abordagem FCM, utilizando os mesmos parâmetros e atributos de entrada, para quantidades de CGDs variando entre 2 e 80.

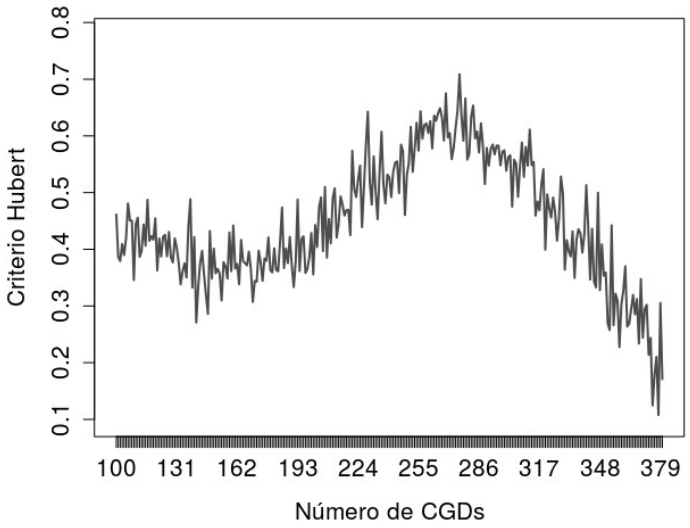


Por meio do gráfico da Figura 5.4, pode-se observar que, para a faixa de interesse normalmente utilizada no delineamento de UGDs em AP (entre 2 e 5 CGDs), duas execuções distintas da abordagem FCM utilizando os mesmos parâmetros e atributos tendem a fornecer resultados idênticos, que quando comparados apresentam valor máximo (1,0) para a estatística Γ de Hubert. Entretanto, é fácil perceber que conforme a quantidade de CGDs aumenta, as diferenças entre as duas soluções também aumenta. Esse efeito ocorre muito devido a aleatoriedade aplicada aos centroides iniciais, pois uma maior quantidade de centroides aleatórios pode aumentar a tendência em se obter soluções de agrupamento identificadas por mínimos locais. Desse modo, podemos considerar que a abordagem FCM é não determinística, pois para um mesmo conjunto de atributos de entrada e de valores para os seus parâmetros, não existe a garantia de obtenção de resultados idênticos. No contexto do delineamento de UGDs em AP, essa característica da abordagem FCM pode proporcionar dúvidas ao usuário final sobre qual solução utilizar, pois para casos onde é necessária uma subdivisão em uma quantidade maior de CGDs do que o que é habitualmente utilizado, essa abordagem pode fornecer soluções muito distintas. Essa diferença também está relacionada com a presença cada vez mais acentuada da estratificação, à medida que a quantidade pretendida de CGDs aumenta.

A questão do não determinismo também pode ser estendida para a abordagem *HACC-Spatial*, pois a tesselação inicial utilizando as variáveis do espaço de coordenadas que normalmente é realizada por essa abordagem é executada pelo algoritmo *k-means*. A Figura 5.5 exibe um gráfico similar ao anterior, nesse caso mostrando os índices obtidos pela estatística Γ de Hubert para agrupamentos que podem ser considerados como tesselação inicial para a execução

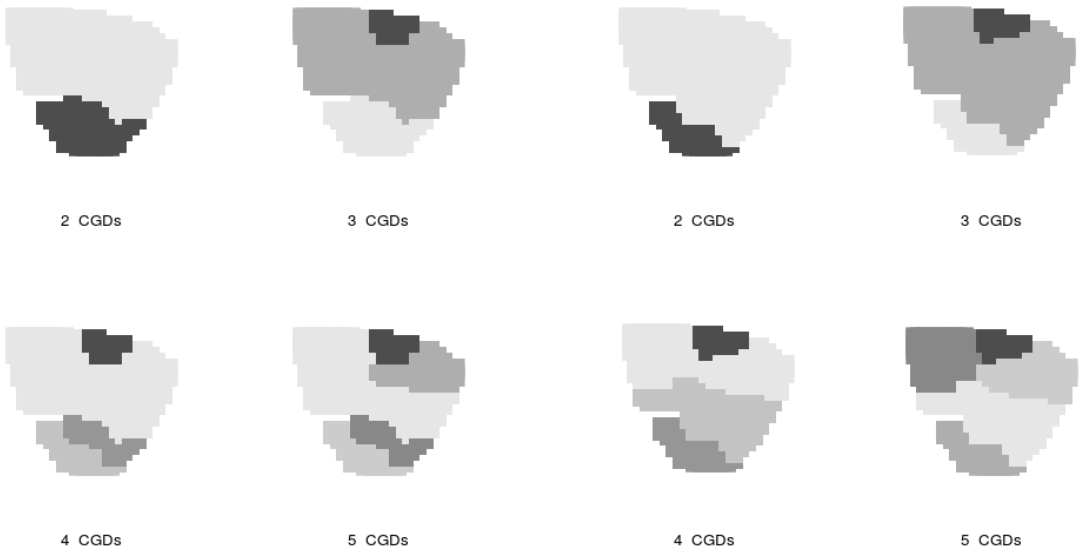
da abordagem *HACC-Spatial*, em um conjunto de dados contendo 415 amostras. Considerando essa quantidade de amostras, e por se tratar de uma tesselação inicial que visa à obtenção de grupos pequenos, foram utilizadas quantidades de grupos variando entre 100 e 380.

Figura 5.5: Índices obtidos pela estatística Γ de Hubert comparando duas tesselações iniciais distintas da abordagem *HACC-Spatial*, utilizando os mesmos parâmetros e atributos de entrada, para quantidades de grupos variando entre 100 e 380.

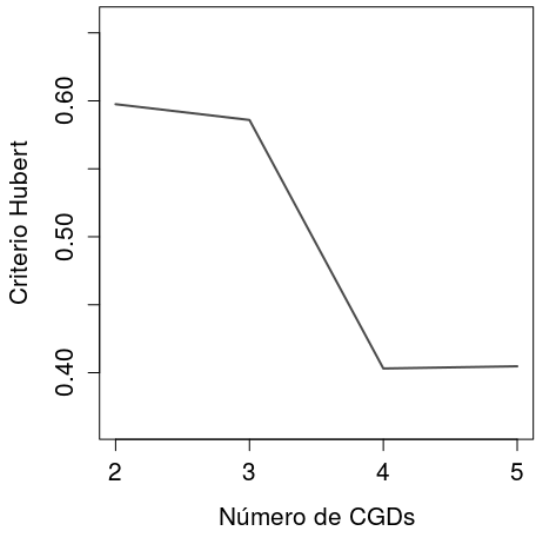


Por meio do gráfico da Figura 5.5, se observa que em nenhum momento as soluções fornecidas por duas execuções distintas da tesselação inicial, contendo a mesma quantidade de grupos, foram totalmente compatíveis. Desse modo, diferentemente da abordagem FCM, onde ao menos os resultados finais que em geral atendem às necessidades dos usuários de AP (entre 2 e 5 CGDs) são normalmente compatíveis, no caso da abordagem *HACC-Spatial*, o não determinismo obtido durante a tesselação inicial pode proporcionar resultados bem diferentes, mesmo que sejam utilizados os mesmos atributos e parâmetros de entrada. A Figura 5.6 mostra mapas de UGDs obtidos por meio de duas execuções distintas da abordagem *HACC-Spatial*, considerando em ambos os casos uma tesselação inicial de k=100 grupos, o valor de 0,5 para o parâmetro cp e o conjunto de atributos UP-CA2; e uma análise de correlação entre os mapas obtidos utilizando a estatística Γ de Hubert.

Figura 5.6: Mapas de UGDs obtidos por meio de duas execuções distintas - (a) e (b) - da abordagem *HACC-Spatial*, considerando os mesmos parâmetros, conjunto de atributos de entrada e cortes no dendrograma para valores entre 2 e 5 grupos; e (c) índice de correlação entre os resultados obtidos em (a) e (b) utilizando a estatística Γ de Hubert.



(a) (b)

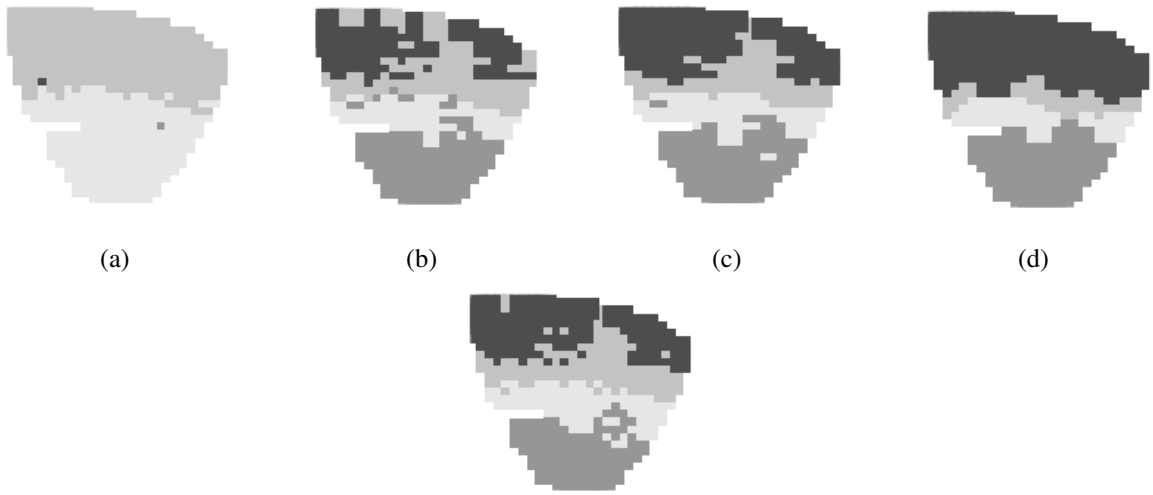


(c)

Por meio das figuras 5.6 (a) e 5.6 (b), já é possível observar algumas diferenças visuais significativas nos mapas para 2 e 3 CGDs, e que aumentam consideravelmente a partir de 4 CGDs. Essas diferenças são confirmadas estatisticamente pelo gráfico de correlação (5.6 (c)). Desse modo, diferentemente da FCM, a abordagem *HACC-Spatial* acaba fornecendo resultados não determinísticos também para mapas contendo entre 2 e 5 CGDs.

Os experimentos relacionados à variação de parâmetros aplicados na abordagem FCM também foram estendidos para a abordagem *HACC-Spatial*. Inicialmente, foram utilizadas variações do parâmetro referente à tesselação inicial (k), com valores de k iguais a n (sem tesselação), 75% de n, 50% de n e 25% de n, onde n é a quantidade total de amostras do conjunto de dados. Apesar do conjunto de dados utilizado influenciar bastante nas questões de contiguidade espacial, os melhores resultados da abordagem *HACC-Spatial* relatados na literatura utilizaram o valor de 0,5 para o critério cp. Desse modo, esse valor foi fixado para os experimentos realizados com variações na tesselação inicial considerando os conjuntos de dados utilizados nesta tese. A Figura 5.7 mostra mapas com 4 CGDs gerados pela abordagem *HACC-Spatial*, a partir dessas definições, em conjunto com um mapa de 4 CGDs gerado pela abordagem FCM utilizando o mesmo conjunto de dados e o valor padrão 2 para o parâmetro m.

Figura 5.7: Mapas contendo 4 CGDs delineadas utilizando o conjunto de dados UP-CA1 a partir das abordagens (a) *HACC-Spatial* com *cp=0,5* e *k=n*; (b) *HACC-Spatial* com *cp=0,5* e *k=75%* de *n*; (c) *HACC-Spatial* com *cp=0,5* e *k=50%* de *n*; (d) *HACC-Spatial* com *cp=0,5* e *k=25%* de *n*; (e) FCM com *m=2*.



(a) (b) (c) (d)

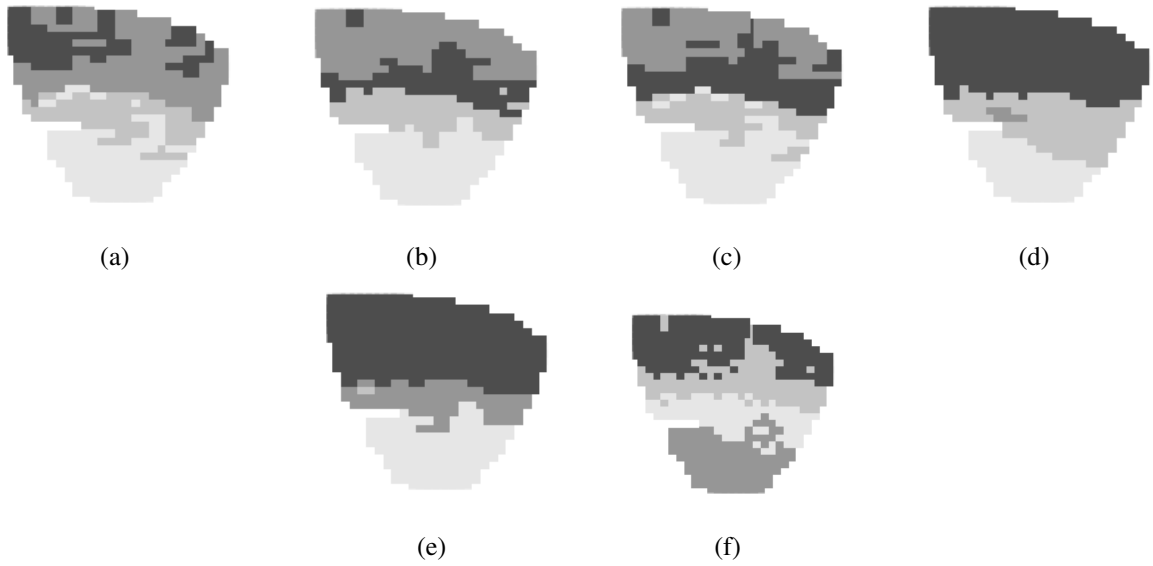
(e)

Por meio da Figura 5.7, é possível verificar a importância da utilização da tesselação inicial para a obtenção de resultados satisfatórios utilizando a abordagem *HACC-Spatial*, pois a não utilização desse artifício pode proporcionar a obtenção de CGDs extremamente pequenas (Figura 5.7 (a)) e que resultam em mapas muito diferentes dos que foram obtidos pela abordagem FCM (Figura 5.7 (e)). Por outro lado, é possível verificar que, na medida em que a quantidade de grupos da tesselação inicial diminui, a sensação de estratificação também diminui, proporcionando resultados visuais melhores e ainda próximos aos resultados obtidos pela abordagem FCM. Entretanto, o parâmetro k deve ser utilizado com cautela, pois pode forçar uma conti-

guidade espacial maior que pode proporcionar diferenças que vão além da simples redução da estratificação do mapa final, modificando bastante o formato das UGDs obtidas (Figura 5.7 (d)).

Na sequência, foram utilizadas variações de valores para o parâmetro cp. Para tanto, foi utilizado um valor de k igual a 50% de n, que no exemplo anterior proporcionou uma visível redução de estratificação do mapa de UGDs, sem que o formato das UGDs com relação ao resultado obtido pela abordagem FCM fosse prejudicado (Figura 5.7 (c)). Considerando o valor de 0,5 como padrão, foram utilizadas variações entre 0 e 1 para o parâmetro cp, com o intuito de verificar os efeitos de se desabilitar de maneira muito precoce ou muito tardia a restrição de contiguidade espacial na construção do dendrograma e, consequentemente, dos mapas finais de UGDs obtidos pela abordagem *HACC-Spatial*. A Figura 5.8 mostra os resultados considerando essas variações, para mapas contendo 4 CGDs.

Figura 5.8: Mapas contendo 4 CGDs delineadas utilizando o conjunto de dados UP-CA1, e as abordagens (a) *HACC-Spatial* com *cp=0,1* e *k=50%* de *n*; (b) *HACC-Spatial* com *cp=0,3* e *k=50%* de *n*; (c) *HACC-Spatial* com *cp=0,5* e *k=50%* de *n*; (d) *HACC-Spatial* com *cp=0,8* e *k=50%* de *n*; (e) *HACC-Spatial* com *cp=1* e *k=50%* de *n*; e (f) FCM com *m=2*.

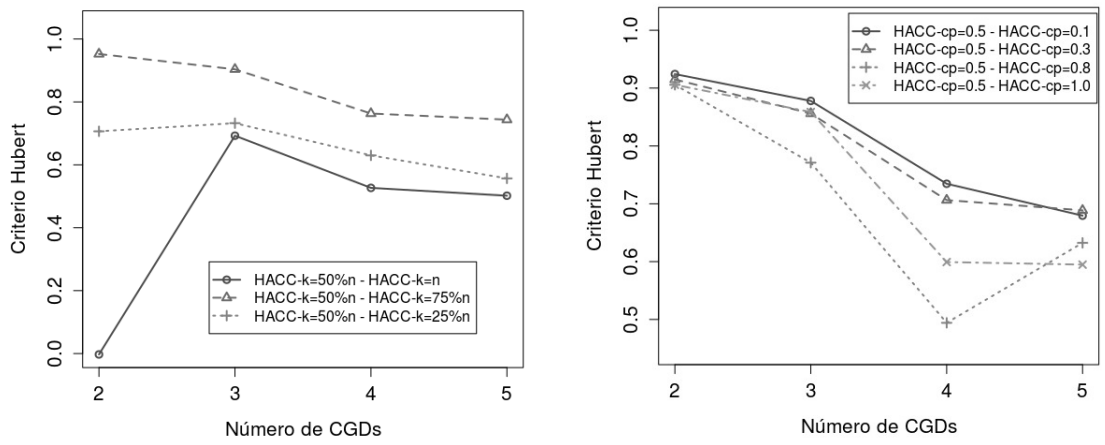


Os resultados exibidos pela Figura 5.8 mostram o efeito e as diferenças que ocorrem nos mapas finais de UGDs, à medida em que a restrição de contiguidade é desabilitada em diferentes passos da construção do dendrograma da abordagem *HACC-Spatial*. Para valores mais baixos de cp, como no caso dos itens (a) e (b), pode-se verificar que o limiar é atingido já nos primeiros passos da construção do dendrograma. Com isso, poucos grupos espacialmente adjacentes são fundidos, fazendo com que surjam algumas estruturas contíguas, mas também algumas UGDs muito pequenas. Essas UGDs provavelmente ficaram isoladas por serem unidas a outras UGDs que compõem a mesma CGD após a restrição de contiguidade ser desabilitada. Um detalhe

importante pode ser verificado quando é utilizado 0,5 para o valor de cp (Figura 5.8 (c)), onde a restrição de adjacência é desabilitada em passos posteriores do dendrograma. Nesse caso, algumas UGDs aumentam de tamanho, por conta da manutenção da restrição de contiguidade por mais tempo. Porém, a quantidade de UGDs pequenas e isoladas também aumenta. Vale ressaltar que esse mapa foi obtido utilizando-se os mesmos parâmetros e atributos de entrada do mapa exibido na Figura 5.7 (c) e, portanto, seria desejável que as duas execuções do algoritmo fornecessem resultados idênticos. Entretanto, é possível verificar, mesmo que visualmente, a existência de muitas diferenças entre os dois mapas, confirmando novamente a característica de não determinismo da abordagem *HACC-Spatial*. Finalmente, os mapas das figuras 5.8 (d) e 5.8 (e) foram gerados sem que o limiar de contiguidade fosse atingido durante a construção do dendrograma, fazendo com que a restrição de fundir apenas grupos espacialmente contíguos não fosse desabilitada. Em ambos os casos, apesar dos mapas de CGD apresentarem estruturas com formas bem definidas, uma das CGDs possui tamanho muito menor do que as outras, muito provavelmente formada já durante a tesselação inicial ou nos primeiros passos do dendrograma. Esse desbalanceamento no tamanho das CGDs também proporciona um efeito de estratificação, mostrando ao usuário final que essas CGDs poderiam ser fundidas a uma CGD maior. Porém, tanto para a abordagem *HACC-Spatial*, quanto para a abordagem FCM, não é possível ao usuário final restringir, por exemplo, a área mínima considerada como satisfatória para uma UGD.

A Figura 5.9 exibe dois gráficos que avaliam, por meio da estatística Γ de Hubert, as diferenças proporcionadas pela variação dos parâmetros k e cp da abordagem *HACC-Spatial* utilizando o mesmo conjunto de atributos (UP-CA1), para mapas variando de 2 a 5 CGDs. Para os dois casos, foram utilizados como referência os mapas obtidos utilizando valores de k=50% de n, e cp=0,5.

Figura 5.9: Índices obtidos pela estatística Γ de Hubert comparando-se mapas de UGDs obtidos com a utilização do conjunto de atributos UP-CA1 e variações nos parâmetros (a) *k*; e (b) *cp* da abordagem *HACC-Spatial*.

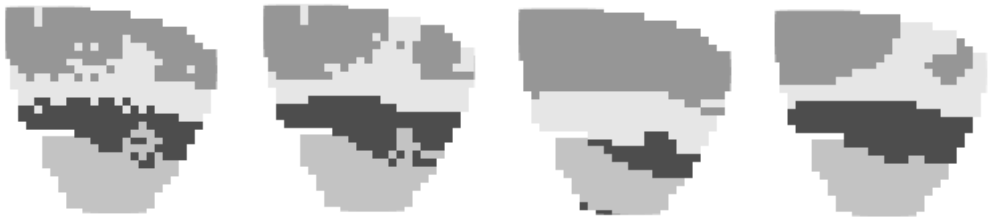


(a) (b)

Por meio da Figura 5.9, observa-se que, tanto as variações do parâmetro k, quanto às variações do parâmetro cp, proporcionam a obtenção de mapas de UGDs com diferenças significativas com relação aos agrupamentos considerados como referência. Na maioria dos casos, pode ser observada uma tendência na diminuição da correlação entre mapas, conforme a quantidade de CGDs aumenta. Desse modo, pode-se concluir que, diferentemente do que acontece para a abordagem FCM, os parâmetros da abordagem *HACC-Spatial* não proporcionam apenas ajustes no resultado final obtido, mas mudanças significativas que podem comprometer as análises realizadas pelo usuário final.

Finalmente, com o intuito de verificar as questões de redução de estratificação e também do não determinismo proporcionado pela abordagem desenvolvida por Peeters (2015), experimentos utilizando variações no parâmetro de distância d considerado para verificação da autocorrelação espacial entre as amostras foram realizados. Da mesma maneira que foi realizado para os experimentos com a abordagem particional FCM, o algoritmo *k-means* foi executado dez vezes para cada experimento realizado com essa abordagem. A Figura 5.10 exibe mapas contendo 4 CGDs, obtidos a partir da utilização do conjunto de atributos UP-CA1 para três diferentes valores do parâmetro d, em comparação com mapas obtidos pela abordagem FCM utilizando parametrização padrão e o mesmo conjunto de atributos.

Figura 5.10: Mapas contendo 4 CGDs delineadas utilizando o conjunto de atributos UP-CA1, e (a) a abordagem FCM com parametrização padrão (*m=2*); e a abordagem desenvolvida por Peeters (2015), considerando diferentes valores para o parâmetro *d*: (b) $20 \times \sqrt{2}$**, (c)** $30 \times \sqrt{2}$ **e (d)** $40 \times \sqrt{2}$**.**



(a) (b) (c) (d)

Para os resultados da Figura 5.10 (b)-(d), os valores de d, em metros, são multiplicados por $\sqrt{2}$ devido à grade regular utilizada, fazendo com que sejam considerados os vizinhos na horizontal, vertical e diagonal. Por meio dos mapas exibidos na Figura 5.10, é possível verificar a redução da estratificação proporcionada pela abordagem desenvolvida por Peeters (2015) em comparação ao resultado de uma abordagem tradicional que utiliza o algoritmo FCM. Entretanto, as variações utilizadas para parâmetro d podem gerar novas UGDs que podem ser consideradas como estratos, como pode ser visualizado no mapa da Figura 5.10 (c).

Alguns casos de não determinismo também podem ser observados com a utilização da abordagem desenvolvida por Peeters (2015), quando são utilizados os mesmos parâmetros e atributos em execuções distintas. A Figura 5.11 exibe mapas contendo 5 CGDs obtidos por essa abordagem, utilizando o conjunto de atributos UP-CA1 e o valor de d igual a $40 \times \sqrt{2}$, em três execuções distintas.

Figura 5.11: Mapas não determinísticos contendo 5 CGDs obtidos pela abordagem desenvolvida por Peeters (2015), utilizando o conjunto de atributos UP-CA1 e *d*= $40 \times \sqrt{2}$**, em três execuções distintas.**

(a) (b) (c)

Por meio da Figura 5.11, é possível observar diferenças na forma das UGDs obtidas pelas três soluções, além de uma UGD que poderia ser considerada como estrato na Figura 5.11 (b). Essas diferenças podem gerar dúvidas ao usuário final sobre qual solução utilizar em uma

aplicação prática.

A partir dos experimentos descritos nesta seção, diversas vantagens e desvantagens com relação à utilização das abordagens existentes para a tarefa de delineamento de UGDs em AP puderam ser verificadas. Com relação à abordagem FCM, a estratificação acentuada observada em alguns casos pode prejudicar a análise do usuário final, que pode preferir mapas mais contíguos espacialmente e, portanto, mais fáceis de interpretar. Entretanto, mapas estratificados não devem ser considerados uma exclusividade das abordagens que utilizam o algoritmo FCM, mas sim de todas as abordagens que utilizam funções objetivo que não consideram a complexidade do espaço de coordenadas. Essa estratificação pode ser reduzida com a utilização da abordagem *HACC-Spatial*, que permite a obtenção de mapas de UGDs com estruturas espaciais mais bem definidas a partir da utilização de restrições espaciais. Entretanto, os parâmetros que deveriam proporcionar ao usuário final a possibilidade de determinar qual seria o nível aceitável de estratificação para um mapa de UGDs podem ser definidos de maneira empírica e, em muitos casos, proporcionar resultados muito distintos entre si. A abordagem desenvolvida por Peeters (2015) também pode auxiliar na redução dessa estratificação. Porém, podem existir dificuldades por parte do usuário final em determinar qual o valor correto para o parâmetro d, além da definição de qual solução deveria ser utilizada dentre os diferentes mapas que podem ser apresentados devido ao seu não determinismo. Em resumo, pode-se dizer que a parametrização de abordagens de agrupamento espacial para o delineamento de UGDs é necessária para que o resultado final obtido seja mais compatível com o que é esperado pelo usuário final, mesmo que gere dúvidas sobre a definição dos valores para os parâmetros. Essas dúvidas podem ser amenizadas com a utilização de parâmetros menos empíricos e que considerem mais diretamente as características dos dados utilizados. Por outro lado, abordagens não determinísticas que fornecem resultados distintos a partir de um mesmo conjunto de atributos e valores para os seus parâmetros devem ser evitadas, pois podem ser obtidas infinitas soluções para a resolução de um mesmo problema, o que por muitas vezes pode inviabilizar as análises por parte do usuário final.

Essas constatações, bem como outros experimentos similares e revisões da literatura realizados ao longo do desenvolvimento desta tese, forneceram subsídios para que uma nova abordagem de agrupamento espacial, contendo parâmetros mais simples de determinar e que se baseiam nos próprios dados coletados em campo, fosse proposta e desenvolvida, a qual é descrita no Capítulo 8.

5.5 Considerações Finais

O desenvolvimento de abordagens e algoritmos de agrupamento específicos para dados espaciais tem crescido ao longo dos anos, principalmente devido ao aumento na produção e utilização desse tipo de dado nas mais variadas áreas de aplicação. No contexto da AP, a diversidade de atributos relacionados ao solo e à cultura que podem ser obtidos por diferentes técnicas e equipamentos em campo, faz com que sejam gerados conjuntos de dados nas mais variadas densidades amostrais. Entretanto, para que esses atributos possam ser utilizados em conjunto por algoritmos de agrupamento, normalmente os dados originais são pré-processados em uma mesma grade espacial regular, a partir de algoritmos de interpolação espacial. Desse modo, constatou-se que as abordagens de agrupamentos baseadas em densidade não são uma alternativa factível para o delineamento de UGDs considerando tanto o espaço de atributos quanto o espaço de coordenadas. A base para essa constatação está relacionada ao fato de que, apesar de o espaço de atributos desses conjuntos de dados permitir a identificação de grupos com base na densidade amostral a partir da utilização, por exemplo, de hipercubos n-dimensionais, quando essa análise passa a ser realizada no espaço de coordenadas, onde a densidade se torna constante após a interpolação espacial, a separação das amostras é muito dificultada. Essa dificuldade se deve ao fato de não existir concentração maior de amostras em diferentes regiões do plano cartesiano bidimensional determinado pelas coordenadas de latitude e longitude quando as amostras são interpoladas em grade regular.

Os trabalhos correlatos apresentados neste capítulo foram de grande importância para verificação do atual estado da arte da área de pesquisa para qual esta tese foi direcionada. Com relação ao agrupamento de dados espaciais de uma maneira geral, pode ser observado que os primeiros trabalhos são mais focados em identificar grupos exatamente a partir da densidade amostral. No entanto, também se observa uma evolução para abordagens hierárquicas com e sem restrições, até chegar ao agrupamento de dados poligonais. Diante dessa diversidade de possibilidades e soluções apresentadas nas abordagens estudadas, bem como dos experimentos preliminares realizados, alguns fatores positivos e negativos presentes nas mesmas puderam ser analisados. A partir dessas análises, foi possível elaborar as principais contribuições desta tese, descritas nos capítulos seguintes.

Capítulo 6

OBJETIVOS, HIPÓTESES E METODOLOGIA

6.1 Considerações Iniciais

Neste capítulo são apresentados os objetivos atingidos, as hipóteses investigadas e a metodologia utilizada nesta tese. O capítulo está organizado da seguinte forma:

- A Seção 6.2 descreve o objetivo geral e os objetivos específicos e secundários alcançados com o desenvolvimento desta tese.
- A Seção 6.3 apresenta formalmente o problema-alvo da tese proposta, bem como as hipóteses que foram investigadas durante a sua execução.
- A Seção 6.4 descreve a metodologia utilizada para a realização dos experimentos a partir de dados reais obtidos de diferentes culturas.
- A Seção 6.5 descreve os dados reais utilizados nos experimentos.
- A Seção 6.6 finaliza o capítulo com as considerações finais.

6.2 Objetivos Atingidos

Esta tese tem como objetivo geral fornecer contribuições originais para viabilizar melhorias na aplicação da tarefa de delineamento de UGDs em AP. Tais contribuições impactam diretamente no processo de definição dos métodos adequados para o agrupamento de dados espaciais aplicáveis a essa tarefa, e estão relacionadas ao desenvolvimento de novas soluções para a tarefa de agrupamento de dados espaciais no formato vetorial. Para tanto, foi utilizada como estudo de caso a aplicação de delineamento de UGDs, na qual o conhecimento adquirido a partir dos

dados deve ser representado de forma simplificada ao usuário final, com o intuito de facilitar o seu entendimento e a sua interpretação. Para que esses objetivos fossem atingidos, as abordagens desenvolvidas nesta tese realizam o tratamento diferenciado do espaço de coordenadas com relação ao espaço de atributos que compõem os dados coletados em AP, visando reduzir os efeitos da estratificação gerada nos mapas finais de UGDs que proporcionam dificuldades de interpretação por parte do usuário final. A partir dessas considerações, diversos objetivos específicos anteriormente previstos foram atingidos durante o desenvolvimento desta tese, conforme relatado a seguir.

Com relação à influência dos usuários especialistas na determinação de parâmetros para o delineamento de UGDs, os objetivos específicos atingidos por esta tese foram:

- A identificação dos diferentes atributos de solo e planta não manipuláveis por ações humanas e que possuem variabilidade no espaço de atributos altamente influenciada pela localização geográfica e, portanto, úteis para o delineamento de UGDs em AP.
- O desenvolvimento de uma nova abordagem para um critério de validação interna que permite ao usuário final verificar a qualidade dos mapas de UGDs obtidos considerando tanto o espaço de atributos quanto o espaço de coordenadas, auxiliando na escolha de qual mapa com uma determinada quantidade de CGDs deve ser utilizado para intervenções em uma determinada área de cultivo.
- O desenvolvimento de uma abordagem de agrupamento espacial que permite ao usuário final influenciar na redução da estratificação que pode ser gerada em mapas de UGDs, definindo a variabilidade mínima no espaço de atributos e a área mínima de uma UGD por meio de parâmetros não empíricos, de fácil interpretação e que proporcionam apenas ajustes no resultado final do agrupamento.

Já com relação à abordagem de agrupamento de dados espaciais desenvolvida para auxílio efetivo ao delineamento de UGDs, os objetivos específicos atingidos por esta tese foram:

- A utilização de restrições relacionadas à contiguidade espacial das CGDs, considerando questões relacionadas à vizinhança espacial, coesão e separação de atributos e coordenadas, além da utilização de dados de declividade como obstáculos espaciais, contribuindo para a redução da estratificação quando estas são exibidas em forma de mapa.
- O desenvolvimento de uma abordagem para agrupamento espacial de amostras pontuais considerando os espaços de atributos e de coordenadas de maneira diferenciada, permitindo a obtenção de mapas de UGDs mais eficazes e de mais fácil interpretação por parte

do usuário final, quando comparada com as abordagens que constituem o estado da arte no tema de pesquisa por meio de experimentos considerando variações nos conjuntos de dados de entrada e atributos.

- O desenvolvimento de uma abordagem complementar, que permite representar os mapas de UGDs em formato poligonal e, consequentemente, melhorar a eficiência de armazenamento e recuperação desses mapas em SGBDEs.
- O desenvolvimento de um sistema protótipo, baseado em *software* livre e nos padrões da OGC, que permite ao usuário final realizar todo o processo de KDSD utilizado para o delineamento de UGDs em AP a partir da interface de um SIG.

Finalmente, ainda foram atingidos os seguintes objetivos, relacionados à produção científica e divulgação dos resultados:

- Publicações de artigos científicos em periódicos e eventos nacionais e internacionais.
- Palestras e apresentações do tema de pesquisa em seminários internos e externos de graduação e pós-graduação.
- Produção de biblioteca baseada no ambiente *R* contendo os códigos-fonte das abordagens desenvolvidas, que deverá ser registrada junto ao Instituto Nacional da Propriedade Industrial (INPI).

6.3 Hipóteses da Tese

Os levantamentos de fundamentação teórica realizados nos capítulos 3 e 4 e, principalmente, a revisão da literatura realizada no Capítulo 5, mostraram a existência de diversas abordagens que podem auxiliar na tarefa de delineamento de UGDs em AP. Entretanto, essa notável diversidade pode indicar a ausência de um consenso com relação ao seu uso específico para essa tarefa. Além disso, as abordagens desenvolvidas especificamente para esse contexto ainda apresentam deficiências que podem impedir a sua utilização prática. Diante disso, concluiu-se que as abordagens existentes não são capazes de fornecer, considerando tanto o espaço de atributos quanto o espaço de coordenadas, mapas de UGDs eficazes de forma a auxiliar efetivamente na gestão localizada de uma área de cultivo agrícola.

A partir dessas considerações, foi possível o estabelecimento da seguinte tese:

T: *Desenvolver uma abordagem para o delineamento de unidades de gestão diferenciada em agricultura de precisão a partir da tarefa de agrupamento específica para dados espaciais em uma região previamente delimitada melhora a eficácia desse processo.*

Para a validação desta tese, as seguintes hipóteses foram investigadas:

H_1: O tratamento do espaço de coordenadas de maneira diferenciada com relação ao espaço de atributos das amostras fornece resultados mais eficazes do que os que são obtidos pelas abordagens que negligenciam o seu uso, ou simplesmente consideram o espaço de coordenadas como parte do espaço de atributos.

H_2: Os parâmetros a serem fornecidos pelo usuário para delineamento de unidades de gestão diferenciada de uma área de cultivo devem, preferencialmente, ser determinados de maneira não empírica, a fim de privilegiar a redução da variabilidade no espaço de atributos e da estratificação do espaço de coordenadas, proporcionando apenas ajustes no resultado final sem prejudicar o determinismo da solução.

H_3: A utilização de informações relacionadas à declividade, utilizadas como obstáculos espaciais para aumentar a dissimilaridade entre amostras no espaço de atributos, pode melhorar a eficácia das unidades de gestão diferenciada que deverão ser delineadas.

H_4: A representação de unidades de gestão diferenciada em forma de polígonos torna mais eficiente o armazenamento e a recuperação das suas informações anexadas em bancos de dados espaciais.

6.4 Metodologia

As práticas descritas nesta seção foram adotadas para que os objetivos desta tese fossem alcançados, permitindo a comprovação das hipóteses por meio da realização de experimentos que confrontaram as abordagens desenvolvidas com aquelas que constituem o estado da arte no delineamento de UGDs em AP.

Durante todo o desenvolvimento desta tese, foram realizadas frequentes revisões que permitiram a manutenção dos conceitos relacionados à fundamentação teórica e aos trabalhos correlatos ao tema de pesquisa em questão. Além disso, foram realizadas reuniões periódicas com o orientador e com membros do grupo de pesquisa e firmadas parcerias para a obtenção de novos conhecimentos e dados reais que pudessem ser utilizados nos experimentos, descritos na Seção 6.5.

Devido à diversidade das abordagens utilizadas nos experimentos, foi necessária a criação

de uma biblioteca que reunisse todas elas e fosse acessível ao usuário final por meio da interface de um SIG. Essa biblioteca foi desenvolvida utilizando o ambiente *R* (R, 2017), sendo parte integrante do sistema protótipo desenvolvido nesta tese, descrito no Capítulo 10.

6.5 Utilização de Dados Reais

Esta seção descreve os dados reais utilizados nos experimentos realizados durante o desenvolvimento desta tese. Esses dados foram coletados em campo utilizando diferentes técnicas e sensores, conforme descrito no Capítulo 2. As áreas de cultivo provedoras dos dados constituem-se de talhões experimentais de empresas particulares em diferentes culturas e com tamanho de área distintos. Essas áreas são tratadas como unidades piloto (UPs) da Rede de Agricultura de Precisão liderada pela Embrapa, e foram estabelecidas por meio de parcerias. A seguir, são descritas algumas características dessas UPs, e os atributos disponibilizados pelas mesmas para os experimentos realizados no âmbito desta tese. Esses atributos, por terem sido coletados por diferentes equipamentos e com diferentes densidades amostrais, foram interpolados em grades espaciais únicas para que pudessem ser utilizados em conjunto por abordagens de agrupamento de dados. Para atributos com uma maior quantidade de amostras, foi utilizado o algoritmo determinístico IDW com valor fixo de potência quadrática. Para atributos com uma quantidade menor de amostras (no mínimo 80), foi utilizado o algoritmo geoestatístico da krigagem.

6.5.1 UP em Cultura de Cana-de-açúcar

A UP em cultura de cana-de-açúcar é constituída por um talhão com aproximadamente 17 hectares de área, pertencente à Fazenda Aparecida, localizada no município de Mogi-Mirim, estado de São Paulo, com coordenadas geográficas centrais em 7505136 ao norte e 299621 ao leste, considerando o SRE WGS84 UTM Zona 23S. Nessa área, o cultivo de cana-de-açúcar (*Saccharum officinarum L.*) da variedade RB-867515 tem sido realizado por aproximadamente 17 anos, no período de 1999 a 2016. O conjunto de dados disponibilizado é referente a coletas realizadas em campo, entre os anos de 2010 e 2013, dos seguintes atributos:

- Condutividade elétrica (CE) do solo, nas profundidades de 15, 30 e 90 cm.
- Textura do solo, contendo dados referentes a teores de areia e argila nas profundidades de 15 e 30 cm.
- Outros atributos do solo, como umidade e densidade, nas profundidades de 15 e 30 cm.

- Produtividade histórica da cultura, medida em toneladas por hectare e contagens de colmos por metro quadrado.
- Índice vegetativo indicador de biomassa (NDVI) obtido a partir de imagens de sensoriamento remoto, em três datas distintas.
- Dados de cota altimétrica e de declividade, obtidos a partir de imagens de sensoriamento remoto.

Para os experimentos realizados no âmbito desta tese, a UP de cana-de-açúcar será chamada de UP-CA. Para essa UP, três configurações distintas referentes ao conjunto de atributos convencionais disponíveis serão consideradas, conforme a listagem a seguir:

- UP-CA1: Utiliza apenas os atributos de CE do solo à 30 e 90 cm de profundidade, e a cota altimétrica. Esses atributos foram selecionados por possuírem correlação de Pearson acima de 0,5 entre si, e por poderem ser obtidos a partir de um único equipamento.
- UP-CA2: Utiliza 13 atributos dos 20 disponíveis no conjunto de dados original, selecionados a partir da aplicação do critério de correlação de Pearson entre pares de atributos. Foram selecionados os atributos que se correlacionam bem, ou seja, que possuem correlação de Pearson acima de 0,5 para pelo menos outros 4 atributos.
- UP-CA3: Utiliza todos os 20 atributos do conjunto de dados original.

Após a interpolação espacial aplicada aos atributos, foi gerada uma grade única contendo 415 amostras com resolução espacial de 20 metros, cada uma delas contendo medidas para todos os 20 atributos do conjunto de dados original. A tabela contendo dados estatísticos desses atributos, bem como os resultados da correlação de Pearson utilizados para selecioná-los estão disponíveis no Anexo A.

A partir de 2016, essa área passou por uma reforma da produção, e teve seu primeiro plantio utilizando a variedade CTC-20 realizado em março desse mesmo ano. Novos dados oriundos de aplicações à taxa variada de insumos agrícolas, obtidos em setembro de 2016, e dados de produtividade obtidos durante a colheita, realizada em julho de 2017, foram utilizados em uma aplicação prática dos conceitos desta tese, descrita no Capítulo 10.

6.5.2 UP em Cultura de Uva de Mesa

A UP em cultura de uva de mesa é constituída por um parreiral de aproximadamente 1 hectare de área, pertencente à Fazenda Capellaro, localizada no Perímetro Irrigado Senador Nilo

Coelho, Núcleo 8, município de Petrolina, estado de Pernambuco, com coordenadas geográficas centrais em 8972641 ao norte e 339309 ao leste, considerando o SRE WGS84 UTM Zona 24S. Nessa área é realizado o cultivo de uva de mesa (*Vitis vinifera*) da variedade Red Globe IAC-313. O conjunto de dados disponibilizado é referente a coletas realizadas em campo no ano de 2012, considerando os atributos de condutividade elétrica (CE) do solo a 20 e 40 cm de profundidade; e índices de clorofila A, B e total.

Para os experimentos realizados no âmbito desta tese, a UP de uva de mesa será chamada de UP-UV. Para essa UP, foram utilizadas três configurações distintas referentes ao conjunto de atributos convencionais disponíveis, conforme a listagem a seguir:

- UP-UV1: Utiliza apenas os atributos de CE do solo a 20 e 40 cm de profundidade. Esses atributos foram selecionados por possuírem correlação de Pearson acima de 0,5 entre si, e por poderem ser obtidos a partir de um único equipamento.
- UP-UV2: Utiliza os atributos de índices de clorofila A e B, também selecionados por possuírem correlação de Pearson acima de 0,5 entre si e poderem ser obtidos a partir de um único equipamento.
- UP-UV3: Utiliza os atributos CE do solo a 20 e 40 cm, e índice de clorofila total, por possuírem correlação de Pearson acima de 0,5 entre si.

Após a interpolação espacial aplicada aos atributos, foi gerada uma grade única contendo 1101 amostras com resolução espacial de 3 metros, cada uma delas contendo medidas para todos os 5 atributos do conjunto de dados original. A tabela contendo dados estatísticos desses atributos, bem como os resultados da correlação de Pearson utilizados para a selecioná-los estão disponíveis no Anexo A.

6.5.3 UP em Cultura de Grãos

A UP em cultura de grãos é constituída por parte de um talhão com aproximadamente 52 hectares de área, pertencente à Fazenda Lambary, localizada no município de Planaltina de Goiás, estado de Goiás, com coordenadas geográficas centrais em 8303834 ao norte e 217742 ao leste, considerando o SRE WGS84 UTM Zona 23S. Nessa área é realizado o cultivo de milho (*Zea mays*) da variedade híbrida simples Pioneer P3862H, em rotação com o cultivo de soja (*Glycine max*). O conjunto de dados disponibilizado é referente a coletas realizadas em campo entre os anos de 2010 e 2012, considerando os atributos de condutividade elétrica (CE) do solo a 30 e 90 cm de profundidade; produtividade de uma safra de milho e soja, medidos em

toneladas por hectare; e altimetria e declividade obtida a partir de imagens de sensoriamento remoto.

Para os experimentos realizados no âmbito desta tese, a UP de grãos será chamada de UP-G. Para essa UP, também foram utilizadas três configurações distintas referentes ao conjunto de atributos convencionais disponíveis, conforme a listagem a seguir:

- UP-G1: Utiliza apenas os atributos de CE do solo a 30 e 90 cm de profundidade e a altitude. Esses atributos foram selecionados por possuírem em geral uma boa correlação, e por poderem ser obtidos a partir de um único equipamento.
- UP-G2: Utiliza os atributos de produtividade de uma safra de soja e milho, selecionados por se tratarem de dados referentes a duas culturas similares, porém distintas, obtidos em uma mesma área em períodos próximos.
- UP-G3: Utiliza os atributos CE do solo à 90 cm de profundidade, produtividade de uma safra de soja e altitude, por possuírem correlação de Pearson acima de 0,5 entre si.

Após a interpolação espacial aplicada aos atributos, foi gerada uma grade única contendo 1291 amostras com resolução espacial de 20 metros, cada uma delas contendo medidas para todos os atributos do conjunto de dados original. As tabelas contendo análises estatísticas desses atributos, bem como os resultados da correlação de Pearson utilizados para a seleção de atributos estão disponíveis no Anexo A.

6.6 Considerações Finais

Neste capítulo foram apresentados os objetivos atingidos com o desenvolvimento desta tese, bem como as hipóteses que foram investigadas e a metodologia e dados reais utilizados nos experimentos que viabilizaram tal investigação. Os próximos capítulos descrevem, de maneira detalhada, as principais contribuições resultantes dessa investigação: o critério de validação interna *SD-Spatial* (Capítulo 7), a abordagem de agrupamento espacial *SWMU Clustering* (Capítulo 8) e a abordagem complementar *SWMU Polygon* (Capítulo 9).

Capítulo 7

CRITÉRIO DE VALIDAÇÃO INTERNA SD-SPATIAL

7.1 Considerações Iniciais

Neste capítulo é apresentado o critério de validação interna *SD-Spatial*, uma nova abordagem para o critério SD aplicada no contexto de dados espaciais. O capítulo está organizado da seguinte forma:

- A Seção 7.2 reforça os conceitos relacionados a critérios de validação interna e a necessidade de avaliar de maneira qualitativa os agrupamentos gerados a partir de dados espaciais, considerando tanto o espaço de atributos quanto o espaço de coordenadas.
- A Seção 7.3 descreve a proposta de uma nova abordagem para o critério de validação interna SD, denominado *SD-Spatial*.
- A Seção 7.4 descreve os resultados experimentais que comprovam a eficácia do critério *SD-Spatial* utilizando dados reais.
- A Seção 7.5 finaliza o capítulo com as considerações finais.

7.2 Introdução

A avaliação qualitativa dos agrupamentos gerados a partir de abordagens de agrupamento considerando o conjunto de dados original não é tarefa fácil. Entretanto, diversas medidas que podem ser utilizadas para esse tipo de avaliação, conhecidas como critérios de validação interna, estão disponíveis na literatura, conforme descrito no Capítulo 4. Considerando o contexto do delineamento de UGDs em AP, essas medidas são utilizadas não só para avaliar a qualidade dos agrupamentos obtidos, mas também para verificar, de uma maneira relativa, qual a quantidade

ideal de CGDs em que deve ser subdivida a área de cultivo em estudo. Apesar de a literatura mostrar algumas abordagens específicas para avaliar de maneira qualitativa os mapas de UGDs gerados, essas abordagens possuem algumas limitações, também discutidas no Capítulo 4, que podem dificultar a análise e interpretação correta dos resultados por parte do usuário final.

Levando-se em consideração essas questões, verificou-se que para o contexto do delineamento de UGDs em AP existe a necessidade de se buscar na literatura critérios de validação interna capazes de avaliar tanto a coesão quanto a separação das CGDs, considerando as questões de variabilidade espacial mínima o suficiente para justificar o seu uso. Assim, por se basear na variância interna como medida de coesão, e na dissimilaridade entre centroides como medida de separação, o critério SD, descrito no Capítulo 4, mostra-se promissor para ser utilizado para essa tarefa. Entretanto, tanto o critério SD, quando os outros critérios levantados na literatura que são utilizados para avaliar agrupamentos que geram mapas de UGDs, foram concebidos para avaliar a eficácia desses mapas considerando apenas o espaço de atributos. Desse modo, questões relacionadas ao arranjo espacial das CGDs, que permitem identificar estratificações excessivas nos mapas, capazes de prejudicar a análise e interpretação por parte do usuário final, não são consideradas por essas medidas. Assim, com o intuito de proporcionar ao usuário final uma maneira de avaliar os mapas de UGDs com relação aos dados originais, considerando tanto o espaço de coordenadas quanto o espaço de atributos, uma das contribuições obtidas nesta tese foi a criação de uma nova abordagem para o critério de validação interna SD, denominado critério *SD-Spatial*.

7.3 Cálculo do Critério *SD-Spatial* para Avaliação Qualitativa e Relativa de Agrupamentos Espaciais

Baseado no critério SD, o *SD-Spatial* utiliza os relacionamentos espaciais de adjacência, subtração e cálculo de área no espaço de coordenadas como fatores de ponderação para as medidas de variância do espaço de atributos. Da mesma maneira que acontece com o seu precursor, o *SD-Spatial* é um critério de minimização, ou seja, valores menores indicam melhores resultados. O critério *SD-Spatial* deve ser calculado considerando um determinado agrupamento e seu conjunto de dados original em três passos: espalhamento médio (coesão intragrupos), considerando fatores de ponderação espaciais ($SD-Spatial_{Scat}$); separação total intergrupos, considerando tanto o espaço de atributos quanto o espaço de coordenadas ($SD-Spatial_{Dis}$); e o valor final, considerando as medidas obtidas nos dois primeiros passos. Desse modo, o cálculo do espalhamento médio é definido segundo a Equação 7.1a.

$$SD-Spatial_{Scat}(K) = \frac{\frac{1}{K}\sum_{k=1}^{K}\left[\left\|V^{\{k\}}\right\| * (1+strat) * (1+nSH) * (2-rDMU) * (2-rDMC)\right]}{\|V\|} \tag{7.1a}$$

$$V^k = (Var(V_1^k), ..., Var(V_p^k)) \tag{7.1b}$$

$$V = (Var(V_1), ..., Var(V_p)) \tag{7.1c}$$

$$strat = \frac{nSDMU}{nDMU} \tag{7.1d}$$

$$rDMU = \frac{min(area(DMU))}{area(k)} \tag{7.1e}$$

$$rDMC = \frac{area(k)}{totalArea} \tag{7.1f}$$

Na Equação 7.1a, o valor de k representa o rótulo do grupo (ou CGD, no contexto desta tese), variando de 1 até K, onde K é a quantidade de grupos do agrupamento avaliado. Na Equação 7.1b, V^k é o vetor de variâncias correspondente aos p atributos convencionais das amostras que foram associadas ao grupo k. Segundo a equação 7.1a, a norma desse vetor deve ser calculada, e dividida pela norma do vetor de variâncias dos p atributos considerando todas as amostras, representado pela Equação 7.1c. Até esse ponto, o cálculo da coesão interna para o critério *SD-Spatial* considera apenas o espaço de atributos, e é realizado da mesma maneira que no critério SD.

Devido à complexidade dos dados espaciais, e com o intuito de incluir, de maneira diferenciada, o espaço de coordenadas no cálculo da coesão interna para agrupamentos gerados por esse tipo de dado, quatro fatores de ponderação foram adicionados ao cálculo original utilizado pelo critério SD, levando em consideração relacionamentos espaciais de adjacência, subtração e cálculo de área. Alguns desses fatores utilizam o parâmetro $porcDMU$, que deve ser informado pelo usuário final no momento da execução do critério *SD-Spatial*. Esse parâmetro determina a porcentagem mínima de área que uma UGD ou um buraco de uma CGD deve possuir, com relação à área total representada pelos dados espaciais agrupados, para que não seja considerado um estrato, ou seja, para que não ocasione estratificação.

O primeiro fator (Equação 7.1d) representa a proporção de UGDs consideradas como estratos ($nSDMU$) com relação à quantidade total de UGDs em que foi subdividida a CGD k. Por conveniência matemática, será sempre adicionado o valor 1 ao fator $strat$, para evitar que o valor $SD-Spatial_{Scat}$ seja igual a 0 quando não existirem UGDs consideradas como estratos na composição de uma CGD k. Essa conveniência matemática também é adotada para o segundo fator, que representa a quantidade de buracos (nSH) considerados como estratos presentes na

CGD k. Entretanto, diferentemente do primeiro, esse fator considera a quantidade total de buracos e não uma razão. Assim, o critério penaliza mais CGDs com buracos considerados como estratos do que CGDs com UGDs consideradas como estratos. A ideia dessa penalização diferenciada é pela observação de que os buracos de uma CGD k são sempre produzidos por UGDs consideradas como estratos em uma CGD $k^{'}$, ou são a causa para a produção dessas UGDs na CGD $k^{'}$, para $k^{'}$ diferente de k. Por outro lado, UGDs consideradas como estratos nem sempre produzem buracos em outras CGDs. Um exemplo desse tipo acontece quando são produzidas UGDs consideradas como estratos adjacentes à borda do polígono representativo da área de estudo, e que não interceptam a área de outra CGD (Figura 7.1).

Figura 7.1: Recorte de um mapa de CGDs contendo buracos e UGDs consideradas como estratos.

Fonte: Elaborada pelo autor.

A Figura 7.1 mostra um recorte de um mapa de CGDs em geral adjacentes, mas que possuem UGDs e buracos considerados como estratos que fazem com que a sensação de descontinuidade e estratificação esteja presente. Nesse exemplo, as UGDs marcadas com um círculo, pertencentes à CGD representada pelo tom de cinza mais claro, não formam buracos na CGD representada pelo tom de cinza mais escuro. Entretanto, as UGDs marcadas com um retângulo, por não estarem localizadas na borda da área e, além disso, ocuparem um percentual da área total menor do que o mínimo aceitável pelo usuário final, são definidas como buracos considerados como estratos da CGD representada pelo tom de cinza mais escuro.

O terceiro fator (Equação 7.1e) representa a razão entre a menor UGD que compõe a CGD k com relação à área total da CGD k; e o quarto fator (Equação 7.1f) representa a razão da área total da CGD k com relação à área total representada pelos dados espaciais agrupados. O objetivo desses dois fatores é a penalização de CGDs contendo UGDs muito pequenas, bem como CGDs muito pequenas considerando o mapa todo, de forma a também auxiliar em prevenir a obtenção de mapas de UGDs extremamente estratificados. Para esses fatores, também por conveniência matemática, seus valores são decrementados de 2. Nesses casos, não se pode utilizar apenas o valor 1 para essa conveniência, porque podem existir CGDs compostas por uma única UGD, ou agrupamentos compostos por uma única CGD, que podem fazer com que o valor de

$SD-Spatial_{Scat}$ seja igual a 0. Desse modo, valores altos para essas razões penalizarão, respectivamente, UGDs e CGDs com áreas muito pequenas. Entretanto, para esses dois casos, o parâmetro *porcDMU* não é utilizado, por não se tratar de uma questão relacionada a áreas consideradas como estratos, mas sim de verificação da presença de CGDs mais balanceadas com relação à área total.

Na medida em que o valor calculado pela Equação 7.1a significa para o critério *SD-Spatial* o quão coesas estão as CGDs ou grupos formados, considerando de maneira diferenciada os espaços de atributos e de coordenadas, a separação entre esses grupos também deve levar em consideração essa diferenciação. Entretanto, ao contrário do que acontece com o espalhamento dos dados, a dissimilaridade entre os grupos pode ser calculada da mesma maneira para ambos os espaços, conforme a Equação 7.2a.

$$SD-Spatial_{Dis}(K,S) = \frac{D_{max}}{D_{min}} \sum_{k=1}^{K} \frac{1}{\sum_{k'=1}^{K} \left\| G(S)^{(k)} - G(S)^{(k')} \right\|} \tag{7.2a}$$

$$D_{max}(K,S) = max_{k \neq k'} \left\| G(S)^{k} - G(S)^{k'} \right\| \tag{7.2b}$$

$$D_{min}(K,S) = min_{k \neq k'} \left\| G(S)^{k} - G(S)^{k'} \right\| \tag{7.2c}$$

Na Equação 7.2a, $G(S)^k$ é o centroide do grupo k, para k diferente de $k^{'}$, no plano cartesiano que representa o espaço S. Assim, considerando conjuntos de dados espaciais, esse centroide deve ser calculado a partir dos atributos convencionais, se considerarmos o espaço de atributos *AS*; e a partir dos atributos espaciais (coordenadas longitude e latitude), se considerarmos os espaço de coordenadas *CS*. O valor de $SD-Spatial_{Dis}$ para um agrupamento com K grupos é então calculado com sendo a média aritmética entre $SD-Spatial_{Dis}(AS)$ e $SD-Spatial_{Dis}(CS)$ (Equação 7.3).

$$SD-Spatial_{Dis}(K) = \frac{SD-Spatial_{Dis}(K,AS) + SD-Spatial_{Dis}(K,CS)}{2} \tag{7.3}$$

No passo final, o valor de *SD-Spatial* para K CGDs ou grupos deve ser calculado considerando as medidas de coesão ($SD-Spatial_{Scat}$) e separação ($SD-Spatial_{Dis}$) obtidas respectivamente a partir dos espaços de atributos e de coordenadas, conforme a Equação 7.4.

$$SD-Spatial(K) = \alpha * SD-Spatial_{Scat} + \beta * SD-Spatial_{Dis} \tag{7.4}$$

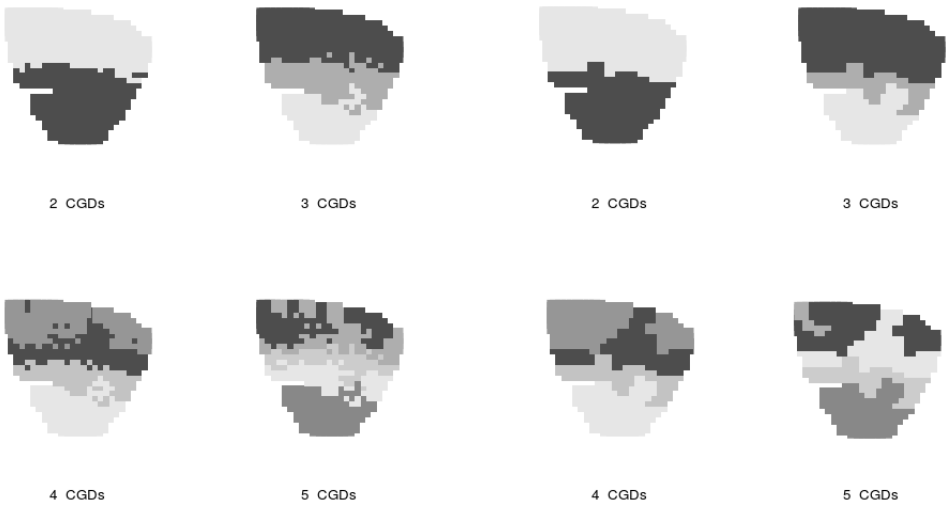
Na Equação 7.4, α é um fator de ponderação definido a partir do cálculo de $SD-Spatial_{Dis}$ para *maxK*; e β é um fator de ponderação definido a partir do cálculo de $SD-Spatial_{Scat}$

para *maxK*. Nesse contexto, *maxK* é a quantidade máxima de grupos utilizada pelo usuário final para a geração dos agrupamentos, permitindo que o critério seja utilizado de maneira relativa comparando agrupamentos com diferentes quantidades de grupos. Diferentemente do que acontece no critério SD, onde apenas o fator de ponderação α é utilizado, o uso do fator de ponderação β pelo critério *SD-Spatial* permite, de certa forma, atenuar a tendência de queda no valor da medida de separação dos grupos considerando o espaço de atributos quando o espaço de coordenadas é considerado pela abordagem de agrupamento.

7.4 Experimentos

Com o intuito de verificar a eficácia do critério *SD-Spatial* em validar agrupamentos de dados espaciais considerando tanto o espaço de atributos quanto o espaço de coordenadas, experimentos com dados reais foram realizados confrontando os resultados obtidos pelo novo critério *SD-Spatial* com relação aos critérios SD e largura de silhueta. Nesse sentido, foram utilizados mapas de UGDs gerados pelas abordagens FCM e *HACC-Spatial* e dados espaciais coletados a partir das UPs referenciadas na Seção 6.5 do Capítulo 6. Além disso, para verificar a eficácia dos métodos em analisar agrupamentos de maneira relativa, foram utilizados mapas contendo de 2 a 5 CGDs, que é o intervalo normalmente utilizado na prática. Nesses mapas, cada CGD é representada graficamente por um tom de cinza. A Figura 7.2 mostra exemplos de mapas de CGDs gerados pelos algoritmos FMC e *HACC-Spatial*, considerando como entrada o conjunto de atributos UP-CA1.

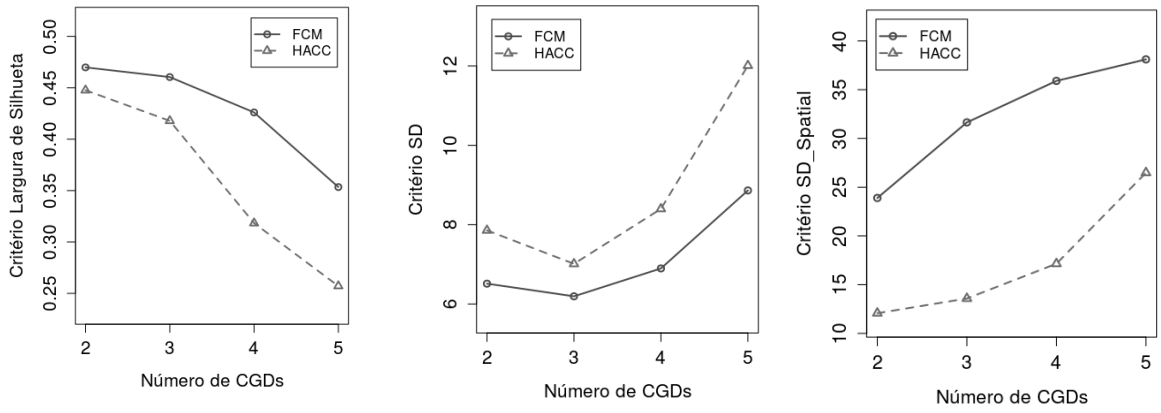
Figura 7.2: Mapas de CGDs gerados utilizando o conjunto de atributos UP-CA1 e as abordagens (a) FCM; e (b) *HACC-Spatial*.



(a) (b)

A Figura 7.2 nos permite verificar claramente as restrições espaciais impostas pela abordagem *HACC-Spatial* para redução da estratificação dos mapas de UGDs obtidos, característica esta que não é tratada pela abordagem FCM. Os gráficos da Figura 7.3 mostram o resultado da aplicação dos critérios de validação da largura de silhueta, SD e *SD-Spatial* nos mapas de UGDs da Figura 7.2, considerando o parâmetro *porcDMU* com valor igual a 5% para o último critério.

Figura 7.3: Critérios de validação interna (a) largura de silhueta; (b) SD; e (c) SD-Spatial com *porcDMU=5%*, utilizados de maneira relativa para comparação de mapas de CGDs gerados pelas abordagens FCM e *HACC-Spatial* e pelo conjunto de atributos UP-CA1.

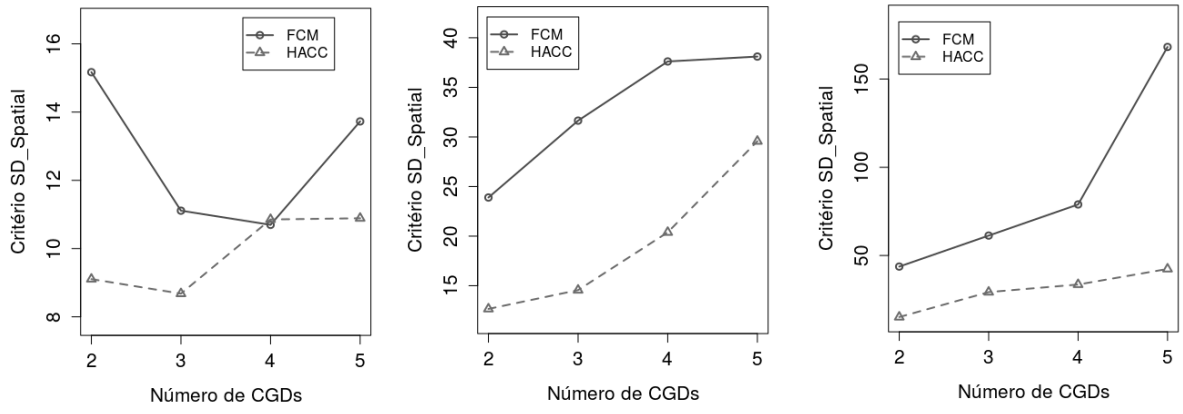


(a) (b) (c)

Tanto o gráfico referente ao critério da largura de silhueta (Figura 7.3 (a)), onde se espera índices maiores para agrupamentos melhores, quanto o gráfico referente ao critério SD (Figura 7.3 (b)), onde se espera índices menores para agrupamentos melhores, mostram uma tendência parecida de queda na qualidade dos agrupamentos, conforme a quantidade de grupos aumenta. Apesar da diferença pequena entre os resultados obtidos, a abordagem FCM apresentou um melhor desempenho em uma avaliação considerando apenas o espaço de atributos, obtido muito em função das restrições espaciais impostas pela abordagem *HACC-Spatial*. Essas restrições prejudicam as medidas de coesão e separação obtidas pelos resultados proporcionados por essa abordagem no espaço de atributos, porém permitem a obtenção de resultados visuais melhores, quando os agrupamentos representam mapas de UGDs. Como essa questão só é tratada pelo *SD-Spatial*, o gráfico dos resultados obtidos com a utilização desse critério (Figura 7.3 (c)), onde também se espera índices menores para agrupamentos melhores, mostra um desempenho muito superior da abordagem *HACC-Spatial* com relação à abordagem FCM.

Apesar de proporcionar resultados que privilegiam questões relacionadas à visualização dos mapas de UGDs, o parâmetro *porcDMU* deve ser utilizado com cautela pelo usuário final, pois está relacionado diretamente à sua percepção em identificar se determinadas UGDs e buracos devem ser considerados como estratos ou não devem ser considerados como estratos. A Figura 7.4 exibe gráficos dos resultados da aplicação do critério *SD-Spatial* nos mesmos mapas da Figura 7.2, considerando três valores distintos para o parâmetro *porcDMU*.

Figura 7.4: Critério de validação interna SD-Spatial com variações do parâmetro *porcDMU* de (a) 0% ; (b) 10% ; e (c) 20% .

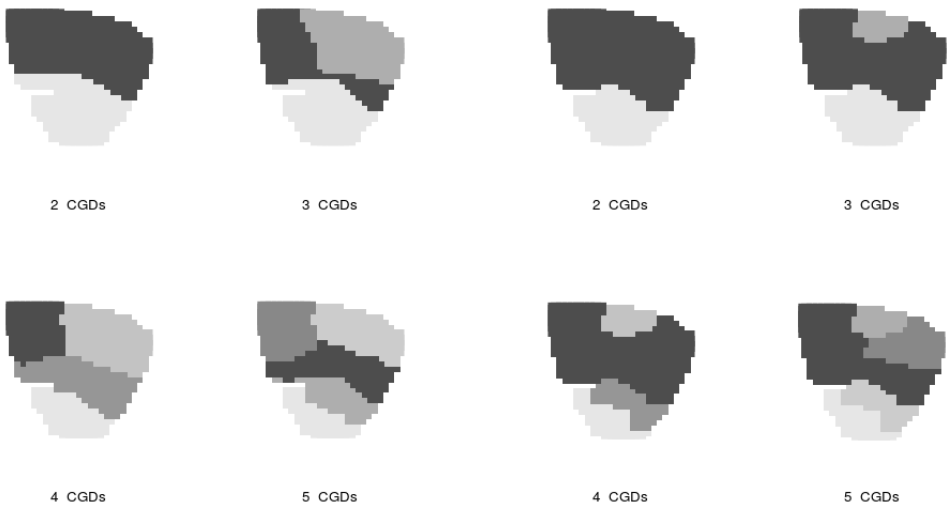


Por meio da Figura 7.4, é possível verificar que a não identificação de UGDs e buracos como estratos por parte do usuário final, que ocorre quando o valor de 0% para o parâmetro *porcDMU* é utilizado (Figura 7.4 (a)), permite que a abordagem FCM alcance resultados

melhores em alguns casos. Esses resultados ocorrem devido à maior coesão obtida pela abordagem FCM com relação à abordagem *HACC-Spatial* considerando apenas o espaço de atributos. Nesse caso, o primeiro e segundo fatores de ponderação do cálculo de $SD - Spatial_{Scat}$ se tornam neutros, fazendo com que a variância no espaço de atributos prevaleça. Por outro lado, a utilização de valores mais altos para *porcDMU*, entre 10% (Figura 7.4 (b)) e 20% (Figura 7.4 (c)), por exemplo, mostram uma tendência geral de aumento no índice alcançado pelo critério conforme a quantidade de grupos aumenta. Nesses casos, algumas UGDs obtidas tanto pela abordagem FCM quanto pela *HACC-Spatial*, principalmente nos mapas com 4 e 5 CGDs, passam a ser consideradas como estratos, fazendo com que os índices obtidos para o critério *SD-Spatial* aumentem. Apesar dos valores atribuídos para o parâmetro *porcDMU* dependerem exclusivamente do usuário final, experimentos como os descritos nesta seção mostram que valores entre 5% e 10% para esse parâmetro produzem resultados dentro do esperado para o critério *SD-Spatial*.

Outro experimento foi realizado utilizando uma quantidade maior de atributos. A Figura 7.5 mostra exemplos de mapas de CGDs gerados pelas abordagens FMC e *HACC-Spatial*, considerando agora como entrada o conjunto de atributos UP-CA2.

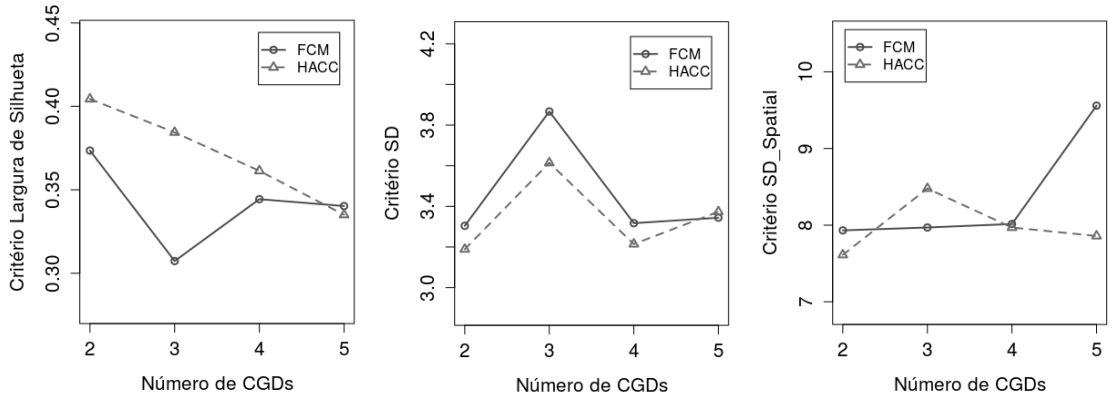
Figura 7.5: Mapas de CGDs gerados utilizando o conjunto de atributos UP-CA2 e as abordagens (a) FCM; e (b) *HACC-Spatial* .



Por meio dos mapas da Figura 7.5, é possível perceber que a estratificação praticamente não existe e foi reduzida drasticamente, principalmente considerando os resultados da abordagem FCM, quando comparados aos mapas exibidos na Figura 7.2. Nesse caso, a utilização de

atributos de qualidade e com boa correlação permitiram a obtenção de mapas de UGDs mais uniformes e contínuos. Os gráficos da Figura 7.6 mostram o resultado da aplicação dos critérios de validação da largura de silhueta, SD e *SD-Spatial* nos mapas de UGDs da Figura 7.5, considerando novamente o parâmetro *porcDMU* com valor igual a 5% para o último critério.

Figura 7.6: Critérios de validação interna (a) largura de silhueta; (b) SD; e (c) SD-Spatial com *porcDMU=5%*, utilizados de maneira relativa para comparação de mapas de CGDs gerados pelas abordagens FCM e *HACC-Spatial* e o conjunto de atributos UP-CA2.



(a) (b) (c)

Os gráficos das figuras 7.6 (a) e 7.6 (b) mostram índices muito próximos obtidos pelas abordagens FCM e *HACC-Spatial* para os critérios da largura de silhueta e SD. Entretanto, segundo esses critérios, a abordagem *HACC-Spatial* obteve sempre resultados melhores, exceto para 5 grupos. Já os resultados obtidos pelo critério *SD-Spatial* (Figura 7.6 (b)) exigem uma análise mais elaborada. Para 3 grupos, podemos verificar ausência de estratificação em ambos os mapas, porém um maior equilíbrio no tamanho das CGDs que fez com que o resultado fornecido pela abordagem FCM fosse considerado melhor pelo critério. Esse desbalanceamento no tamanho das CGDs obtidas pela abordagem *HACC-Spatial* foi provocado pela fusão dos dois grupos localizados na parte inferior do mapa de 4 CGDs, fazendo com que uma das 3 CGDs ficasse com tamanho em área desbalanceado com relação às outras. Para os mapas de 5 CGDs, o critério *SD-Spatial* mostra uma penalização grande para a estratificação mínima gerada por uma das CGDs do mapa de UGDs obtido pela abordagem FCM. Essa CGD foi dividida em duas UGDs, sendo uma delas muito pequena, o que causou uma diferença considerável no índice obtido pelo critério SD-Spatial em favor da abordagem *HACC-Spatial*.

A partir dos experimentos descritos nesta seção, foi possível verificar a eficácia do critério *SD-Spatial* em identificar agrupamentos espaciais de qualidade, considerando não só o espaço de atributos como também as questões relativas ao espaço de coordenadas, que no contexto desta tese estão totalmente relacionadas à visualização dos mapas de UGDs. Desse modo, esse

critério foi utilizado para avaliar os mapas de UGDs obtidos nos experimentos subsequentes desta tese.

7.5 Considerações Finais

Neste capítulo foi apresentado, de maneira detalhada, o critério de validação interna *SD-Spatial*, uma nova abordagem para o critério SD que avalia, de maneira qualitativa, os agrupamentos obtidos a partir de dados espaciais considerando tanto o espaço de atributos quanto o espaço de coordenadas. Considerando a aplicação de delineamento de UGDs em AP, esse novo critério permite ao usuário final definir uma porcentagem de área mínima para uma UGD com relação à área total, possibilitando a identificação de UGDs e buracos que podem contribuir para a obtenção de mapas estratificados e de difícil interpretação.

Capítulo 8

SWMU CLUSTERING: UMA ABORDAGEM DE AGRUPAMENTO ESPACIAL

8.1 Considerações Iniciais

Neste capítulo é apresentada a abordagem de agrupamento espacial *SWMU Clustering*, composta por restrições espaciais para auxílio no delineamento de UGDs em AP. O capítulo está organizado da seguinte forma:

- A Seção 8.2 apresenta detalhadamente a proposta de uma nova abordagem de agrupamento para dados espaciais chamada *SWMU Clustering*.
- A Seção 8.3 apresenta e discute experimentos realizados para comparar a abordagem proposta *SWMU Clustering* com as abordagens que constituem o estado da arte no delineamento de UGDs em AP.
- A Seção 8.4 apresenta um teste estatístico de significância que permite verificar a não aleatoriedade dos agrupamentos obtidos pela abordagem *SWMU Clustering*, confrontando-os com agrupamentos obtidos a partir de dados sintéticos.
- A Seção 8.5 descreve uma análise preliminar de complexidade computacional para a abordagem *SWMU Clustering*.
- A Seção 8.6 finaliza o capítulo com as considerações finais.

8.2 A Abordagem *SWMU Clustering*

Nesse seção é descrita a proposta de uma nova abordagem de agrupamento espacial, denominada *Spatial Ward's Management Units Clustering* (*SWMU Clustering*), inicialmente voltada para o contexto do delineamento de UGDs em AP. Desse modo, a necessidade de se obter CGDs com variância interna desprezível e, ao mesmo tempo, pouco estratificadas e que permitam a obtenção de mapas de UGDs facilmente interpretáveis pelo usuário final, foi de extrema importância para direcionar o desenvolvimento dessa nova abordagem.

Na tarefa de agrupamento de dados espaciais no contexto do delineamento de UGDs em AP, o espaço de coordenadas deve ser utilizado para restringir ou permitir o agrupamento entre amostras realizado de acordo com o espaço de atributos. Desse modo, as abordagens hierárquicas, apesar de computacionalmente mais complexas, proporcionam ao usuário final a possibilidade de explorar melhor os resultados obtidos, pois não exigem que seja predefinida uma quantidade fixa de grupos. Além disso, abordagens hierárquicas são independentes de convergência para uma solução ótima, o que faz com que sua execução termine quando a construção do dendrograma é finalizada. Essas características permitem um maior controle das restrições espaciais que podem ser impostas e verificadas passo à passo.

8.2.1 Restrição do Centroide Espacial

Dentre as abordagens de agrupamento hierárquicas tradicionais, a que mais se assemelha aos objetivos que se deseja atingir com o delineamento de UGDs em AP é a abordagem de Ward, que calcula a dissimilaridade entre os grupos com base no aumento da variância interna com relação ao centroide, ou soma dos erros quadráticos (J), conforme descrito na Equação 4.2 do Capítulo 4. Desse modo, definiu-se que, para a abordagem *SWMU Clustering*, a dissimilaridade entre os grupos deveria ser calculada de maneira similar ao que já é realizado pela abordagem de Ward, porém considerando restrições relacionadas ao espaço de coordenadas para definir uma amostra representante do centroide de cada grupo em cada passo da construção do dendrograma. As equações 8.1a e 8.1b definem um novo cálculo para a soma dos erros quadráticos de cada grupo k em cada passo de construção do dendrograma, agora utilizando restrições espaciais para atender a abordagem *SWMU Clustering*.

$$\bar{sx}_k = x_j \in G_k \mid D(\bar{x}_k[c_x, c_y], x_j[c_x, c_y]) = min \left\| \bar{x}_k[c_x, c_y] - x_j[c_x, c_y] \right\| \tag{8.1a}$$

$$SJ_k = \sum_{x_j \in G_k} d\left(x_j[a_1, ..., a_p], \bar{sx}_k[a_1, ..., a_p]\right)^2 \tag{8.1b}$$

Segundo a Equação 8.1a, o centroide espacial $\bar{sx}_k$ de um grupo G_k deve ser a amostra x_j pertencente a esse grupo mais similar ao seu centroide, quando este é obtido considerando apenas as variáveis de latitude e longitude presentes no espaço de coordenadas bidimensional representado pelo plano cartesiano (c_x, c_y). Desse modo, a soma dos erros quadráticos *SJ* de um grupo G_k (Equação 8.1b) deve ser calculada somando-se as distâncias Euclidianas quadráticas, considerando um espaço de atributos contendo p variáveis, de cada amostra x_j pertencente a esse grupo para o seu centroide espacial $\bar{sx}_k$, obtido a partir da Equação 8.1a.

Essa alteração no cálculo para obtenção do centroide, e consequentemente, na soma dos erros quadráticos de cada grupo, visa a diminuição da estratificação dos mapas de UGDs, proporcionando que a soma geral dos erros quadráticos dos grupos sofra um incremento menor quando são fundidos grupos espacialmente mais próximos. Desse modo, esse novo cálculo proporciona uma restrição espacial na construção do dendrograma da abordagem *SWMU Clustering* utilizando apenas o conjunto de dados de entrada e sem a necessidade de parâmetros informados pelo usuário final. A Figura 8.1 exibe mapas contendo 3 CGDs obtidos a partir da abordagem de agrupamento de Ward tradicional, em comparação com a inclusão da restrição do centroide espacial proposta pela abordagem *SWMU Clustering*.

Figura 8.1: Mapas contendo 3 CGDs obtidas (a) a partir da abordagem de Ward tradicional; e (b) a partir da utilização de centroide espacial definido pela abordagem *SWMU Clustering* para o cálculo do erro quadrático.

(a) (b)

Por meio da Figura 8.1 (b), é possível verificar alguns efeitos causados com as mudanças no cálculo do erro quadrático, que proporcionou a manutenção da contiguidade espacial para a CGD representada pelo tom de cinza mais escuro. A utilização da abordagem de Ward tradicional proporcionou uma estratificação acentuada dessa mesma CGD, que acabou sendo representada por uma UGD de maior área e 3 UGDs muito pequenas (Figura 8.1 (a)). Por outro lado, o novo cálculo também acabou gerando uma UGD muito pequena, na parte inferior do mapa, e adjacente à CGD representada pelo tom de cinza mais escuro (Figura 8.1 (b)). Essa UGD foi gerada no primeiro passo da construção do dendrograma, onde em abordagens hierárquicas aglomerativas cada grupo é associado a uma única amostra, e não foi fundida com

nenhum outro grupo durante o restante da execução do algoritmo, mantendo ainda um efeito de estratificação no mapa de UGDs.

8.2.2 Restrições Espaciais Parametrizadas

Apesar de aparentemente proporcionar uma redução na estratificação dos mapas de UGDs fornecidos, a restrição imposta pelo cálculo do centroide espacial pode ser insuficiente e proporcionar resultados ainda indesejados, onde ainda são obtidos mapas de UGDs estratificados. Desse modo, levando-se em consideração características identificadas em abordagens da literatura, outras três restrições espaciais opcionais, determinadas a partir de parâmetros fornecidos pelo usuário final, foram incluídas na abordagem *SWMU Clustering*, resumidamente descritas a seguir:

- O particionamento inicial do conjunto de dados considerando as dissimilaridades no espaço de coordenadas, chamado de tesselação inicial. Isto é obtido considerando o limiar de variância média das amostras no espaço de atributos determinado a partir do parâmetro *varTess*, com *varTess* maior do que 0.

- A determinação de um tamanho mínimo para a área de uma UGD com relação à área total, por meio do parâmetro *porcDMU*, com *porcDMU* maior do que 0. Isto é obtido a partir da quantidade máxima de CGDs desejada pelo usuário final, determinada pelo parâmetro *maxDMC*. Essa restrição permite que UGDs menores que o tamanho mínimo determinado sejam fundidas a UGDs mais próximas, considerando o espaço de coordenadas.

- A utilização de dados que permitem identificar obstáculos espaciais, como os mapas de declividade, influenciando no cálculo das dissimilaridades entre amostras espacialmente separadas por esses obstáculos. Isto é realizado por meio da identificação de polígonos representando barreiras espaciais. No caso de utilização dessa restrição, determinada a partir do parâmetro *usaObst*, o cálculo da dissimilaridade entre as amostras deverá considerar a distância obstruída, conforme descrito na Equação 8.2.

O Algoritmo 1 descreve, em alto nível de abstração, os principais passos de execução da abordagem de agrupamento hierárquica aglomerativa *SWMU Clustering*, considerando as restrições espaciais descritas anteriormente.

Algoritmo 1: *SWMU Clustering*.

Entrada: V = conjunto de vetores de dados espaciais com n amostras nos espaços de atributos (a) e de coordenadas (s); $contorno$ = polígono do contorno da área; $varTess$ = porcentagem de variância média mínima dos grupos para tesselação inicial; $porcDMU$ = porcentagem mínima de área de uma UGD com relação à área total; $minDCM$ = quantidade mínima de CGDs; $maxDMC$ = quantidade máxima de CGDs; $usaObst$ = determina a utilização ou não de obstáculos espaciais; B = polígonos representando as barreiras espaciais consideradas como obstáculos; $dist$ = Matriz de dissimilaridades entre os grupos.

Saída: M = Conjunto de mapas de UGDs obtidos a partir do agrupamento hierárquico.

```
1  início
2      se varTess > 0 então
3          varA = 0 ;
4          k ← n ;                                  // k =quantidade inicial de grupos
           // Cálculo da tesselação inicial
5          repita
6              k ← k − 1;
7              M ← k-means(V(s),k) ;
8              varA ← S(M,V(a)) ;
9          até (varA ≥ VarTess) ou (k = maxDMC);
10     senão
11         M ← V;
12     fim se
13     numDMCs = k ;
       // Cálculo da matriz de dissimilaridades inicial
14     para cada par de grupos (Mi,Mj) ∈ M faça
15         dist[Mi,Mj] ← errQuad([Mi,Mj],usaObst,B);
16     fim para
       // Construção do dendrograma hierárquico
17     repita
18         se numDMCs = maxDMC então
19             M(numDMCs) ← fundeUGDsPequenas(M(numDMCs),porcDMU,contorno);
20         fim se
21         parFus ← (Mi,Mj) ∈ M | dist(Mi,Mj) = min(dist);
22         M ← M \ parFus;
23         M ← M ∪ (parFus[1] ∪ parFus[2]);
24         atualiza(dist,usaObst,B);
25         numDMCs ← numDMCs − 1;
26     até numDMCs = minDMC;
27     retorna M[minDMC:maxDMC];
28 fim
```

O algoritmo é iniciado com a atribuição da restrição espacial proposta pela tesselação inicial (linhas 2 à 9), caso o usuário final tenha optado por realizá-la determinando um valor maior do que 0 para o parâmetro *varTess*. Desse modo, são realizadas diversas execuções do algoritmo *k-means* utilizando os atributos do espaço de coordenadas ($V(s)$), até que a variância média considerando o espaço de atributos $V(a)$, calculada de maneira similar à coesão intragrupos S definida pelo critério de validação interna SD (Equação 4.7a), atinja o valor determinado pelo parâmetro *varTess*. Um detalhe importante nesse passo é que, se o valor atribuído ao parâmetro *varTess* não for atingido até que o valor de k seja igual a *maxDMC*, serão considerados *maxDMC* grupos para esse passo, fazendo com que a execução do algoritmo não seja interrompida. Assim, é de extrema importância uma definição não empírica do valor do parâmetro *varTess* para que a abordagem *SWMU Clustering* forneça resultados satisfatórios. A linha 11 define o mapa inicial de UGDs com cada amostra representando um único grupo, caso o usuário final tenha optado por não executar a tesselação inicial. Na sequência, as linhas de 14 à 16 calculam a matriz de dissimilaridades (*dist*) entre os pares de grupos iniciais, por meio da função *errQuad*, descrita em alto nível de abstração pelo Algoritmo 2. Esse matriz possui dimensão inicial $k \times k$, onde cada célula representa a soma dos erros quadráticos de um grupo formado pela fusão de um determinado par de grupos (M_i, Mj).

Algoritmo 2: errQuad.

Entrada: par = Par de grupos para o qual se deseja calcular a dissimilaridade; $usaObst$ = determina a utilização ou não de obstáculos espaciais (valor verdadeiro indica o uso de obstáculos); B = polígonos representando as barreiras espaciais consideradas como obstáculos;

Saída: $diss$ = Dissimilaridade entre os dois grupos.

```
1 início
2     se usaObst = verdadeiro então
3         diss ← SJO(B)_{par[1]∪par[2]};
4     senão
5         diss ← SJ_{par[1]∪par[2]};
6     fim se
7     retorna diss;
8 fim
```

A função *errQuad* calcula a dissimilaridade direta entre os grupos seguindo diretamente a Equação 8.1b, caso não seja considerada a restrição espacial de obstáculos pelo usuário final (linha 5). Caso contrário, será utilizada a distância obstruída prevista na Equação 8.2, descrita mais adiante, para o cálculo do erro quadrático pela Equação 8.1b (linha 3).

Com a matriz de dissimilaridades iniciais (*dist*) calculada, a construção do dendrograma da abordagem hierárquica é então executada (linhas 17 à 26 do Algoritmo 1). O par de CGDs mais similar é fundido em uma única CGD em cada iteração, até que se atinja a quantidade de CGDs determinada pelo parâmetro *minDMC*. Nesse caso, as CGDs i e j, para i diferente de j, serão consideradas as mais similares se a dissimilaridade armazenada na posição (i,j) da matriz *dist* for a menor entre todas as dissimilaridades. Ainda em cada iteração, a matriz *dist* é atualizada e a sua dimensão é sempre reduzida com a inclusão de novas entradas com valores de dissimilaridade do novo grupo formado com relação aos outros grupos remanescentes, e a exclusão das entradas contendo os valores de dissimilaridades dos grupos Mi e Mj com relação aos outros grupos remanescentes, seguindo cálculos similares aos realizados pela função *errQuad*.

Durante a construção do dendrograma, especificamente na iteração onde é realizado o agrupamento utilizando a quantidade de CGDs definidas pelo parâmetro *maxDMC*, as UGDs com porcentagem relativa de área (com relação a área total) com valor menor ou igual ao valor definido pelo parâmetro *porcDMU* são fundidas utilizando a função $fundeUGDsPequenas$, definida em alto nível de abstração pelo Algoritmo 3.

Algoritmo 3: fundeUGDsPequenas.

Entrada: $MAtual$ =Agrupamento obtido para *maxDMC* CGDs; $porcDMU$ = porcentagem mínima de área de uma UGD com relação à área total; $contorno$ =Polígono contendo o contorno da área total.

Saída: $MAtual$ =Agrupamento de saída ajustado.

```
1  início
2      para cada UGDi ∈ MAtual faça
3          se area(UGDi) ≤ (area(MAtual) × porcDMU) então
4              proxUGD ← UGDMaisProxima(UGDi);
5              MAtual ← MAtual \ UGDi;
6              MAtual ← MAtual \ proxUGD;
7              MAtual ← MAtual ∪ (UGDi ∪ proxUGD);
8          fim se
9      fim para
10     retorna MAtual;
11 fim
```

O Algoritmo 3 verifica a área de todas as UGDs pertencentes às CGDs do agrupamento *MAtual* com relação à área total, a partir do parâmetro *porcDMU*. Se determinada UGD i for considerada pequena, então a mesma é fundida à sua UGD k espacialmente mais próxima. Nesse caso, k será a UGD cujo centroide espacial é o mais próximo do centroide espacial de

i, considerando o grupo de *K* UGDs espacialmente adjacentes a *i*. Essa nova UGD formada é incluída no agrupamento, e as UGDs *i* e *k* são retiradas (linhas 4 à 7). Esse tipo de fusão pode fazer com que o agrupamento *MAtual* fique com uma quantidade menor de CGDs do que antes da execução do Algoritmo 3. Finalizando a execução do Algoritmo 1, apenas os agrupamentos obtidos utilizando entre *minDMC* e *maxDMC* grupos são disponibilizados ao usuário final (linha 28).

A seguir, é realizado um detalhamento, de maneira incremental, dos efeitos causados pelo uso de cada uma das restrições espaciais parametrizadas para a abordagem *SWMU Clustering* em conjunto com a restrição do centroide espacial. Esse detalhamento permite uma análise mais aprofundada dos principais passos dos algoritmos 1, 2 e 3.

8.2.2.1 Tesselação Inicial

A tesselação inicial é uma estratégia usada para evitar a presença de UGDs determinadas no início da construção do dendrograma e que acabam se tornando isoladas e consideradas como estratos. Entretanto, diferentemente do que acontece na abordagem *HACC-Spatial*, onde a quantidade de grupos da tesselação inicial é determinada de maneira empírica por meio do parâmetro *k*, a tesselação inicial proposta pela abordagem *SWMU Clustering* proporciona ao usuário final a determinação dessa quantidade por meio de um parâmetro relacionado aos dados de entrada, denominado *varTess*. Por meio desse parâmetro, o usuário final define qual o percentual de variância média permitido para o agrupamento que iniciará a construção do dendrograma, considerando o espaço de atributos. Desse modo, uma variância média menor do que a determinada pelo parâmetro *varTess* pode ser considerada como desprezível pelo usuário final, caracterizando uma área mínima inicial para cada CGD, já que a tesselação é realizada particionando-se as amostras a partir do espaço de coordenadas.

A partir da definição do parâmetro *varTess*, a abordagem *SWMU Clustering* executa a tesselação inicial agrupando as amostras considerando o espaço de coordenadas *m* vezes, para uma quantidade de grupos variando de *n* até $n-m$, onde *n* é a quantidade total de amostras do conjunto de dados, até que o limiar *varTess* seja atingido ou $n-m$ seja igual ao valor definido para o parâmetro *maxDMC*. Para verificação desse limiar em cada um dos *m* agrupamentos realizados, a coesão interna média dos grupos é calculada utilizando a medida de dispersão do critério SD (*S*) considerando as variáveis do espaço de atributos, conforme a Equação 4.7a descrita no Capítulo 4. Assim, a *m*-ésima e última partição gerada para a tesselação inicial é atingida quando o valor de *S* para o agrupamento é maior do que *varTess*. Consequentemente, a partição gerada no passo anterior, contendo $m-n-1$ grupos, deve ser considerada como a tesselação

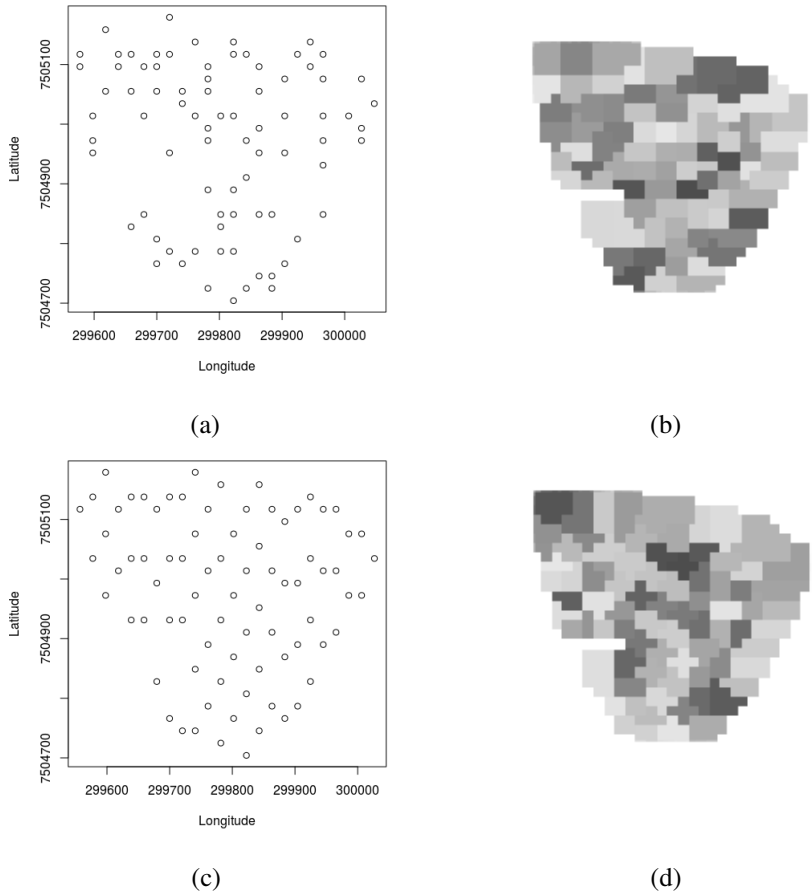
inicial que representará o passo inicial para a construção do dendrograma da abordagem *SWMU Clustering*.

Para essa subdivisão inicial considerando o espaço de coordenadas, a abordagem *SWMU Clustering* também utiliza, assim como a abordagem *HACC-Spatial*, o algoritmo de particionamento *k-means*. Essa escolha se baseou no fato de que necessariamente cada amostra deve ser associada a um único grupo nessa subdivisão. Além disso, o algoritmo *k-means*, nesse contexto, fornece resultados muito similares aos que são obtidos pelo FCM, porém com um custo computacional menor (GHOSH; DUBEY, 2013). No entanto, para essa nova abordagem, é utilizada uma versão modificada do *k-means* desenvolvida por Lloyd (1982), onde são utilizados diagramas de Voronoi contínuos ao invés da determinação do centroide mais próximo de cada amostra discreta. Além disso, com o intuito de evitar o não determinismo do resultado final, uma estratégia para determinação não aleatória dos centroides foi utilizada na abordagem *SWMU Clustering*, levando-se em consideração apenas os atributos de longitude (x) e latitude (y) do espaço de coordenadas, a partir dos seguintes passos:

- Ordena-se as n amostras do conjunto de dados primeiramente pela longitude (x) e depois pela latitude (y), atribuindo-se à elas índices de 1 à n.
- Encontra-se a quantidade média de amostras previstas por grupo ($tamGrupo$), considerando que os k grupos que serão obtidos devem possuir tamanhos similares. Assim, $tamGrupo$ deve ser o número inteiro mais próximo da divisão de n por k.
- O primeiro centroide será a amostra com índice igual a 1; os centroides seguintes serão as amostras com índice igual ao do centroide anterior, somando-se o valor de $tamGrupo$.

Assim, tomando como exemplo um conjunto de dados com n igual a 35 amostras, se desejarmos particioná-lo em k igual a 8 grupos, deveremos considerar o valor 4 para $tamGrupo$. Desse modo, seriam escolhidos como centroides iniciais representantes dos grupos as amostras com índices 1, 5, 9, 13, 17, 21, 25 e 29. A Figura 8.2 mostra um exemplo de distribuição dos centroides da maneira tradicional (aleatória) em comparação com a metodologia proposta, bem como o resultado final da tesselação inicial utilizando-se ambas as metodologias.

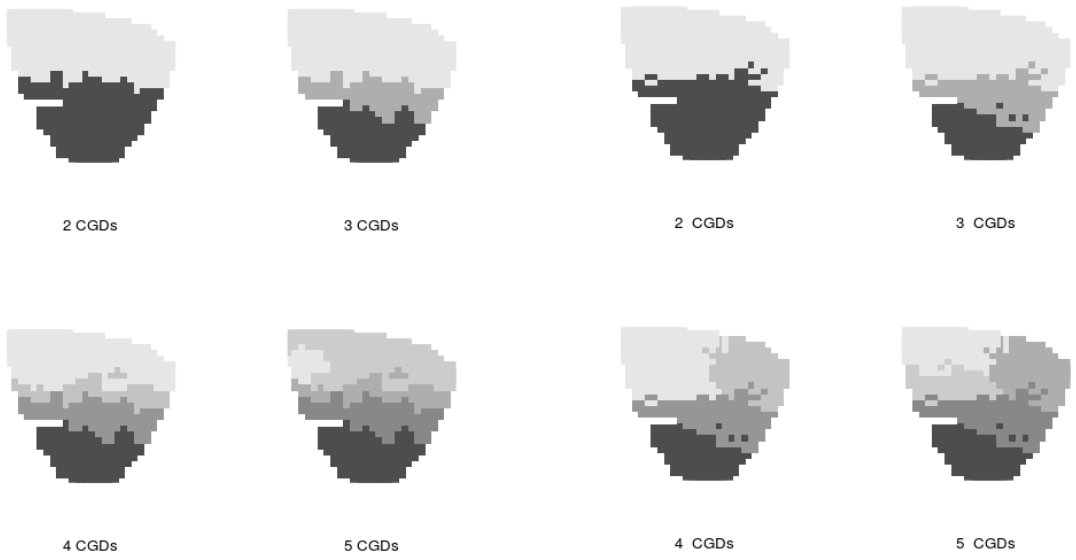
Figura 8.2: (a) Centroides obtidos de maneira aleatória; e (b) sua correspondente tesselação inicial, em comparação com (c) centroides obtidos pela metodologia proposta; e (d) sua correspondente tesselação inicial.



Por meio da Figura 8.2, é possível verificar uma distribuição mais regular dos centroides iniciais utilizando a metodologia proposta (Figura 8.2 (c)) com relação à distribuição aleatória (Figura 8.2 (a)). Apesar de proporcionar um leve aumento na variância interna média dos grupos gerados considerando as variáveis do espaço de atributos, quando comparada com a solução aleatória, a tesselação realizada utilizando a metodologia proposta permite a obtenção de grupos iniciais com formas e tamanhos mais regulares. Além disso, a metodologia criada para a inicialização dos centroides iniciais faz com que estes sejam sempre os mesmos para um mesmo conjunto de dados espaciais, diminuindo a probabilidade de obtenção de tesselações não determinísticas que poderiam gerar diferenças nos mapas de UGDs em execuções distintas do algoritmo hierárquico, considerando os mesmos parâmetros e atributos.

A Figura 8.3 exibe mapas contendo entre 2 e 5 DMCs, gerados a partir da utilização das novas metodologias de cálculo do aumento do erro quadrático para construção do dendrograma (*SJ*) e tesselação inicial desenvolvidas para a abordagem *SWMU Clustering*; e os mesmos mapas gerados a partir dos mesmos dados, porém utilizando a abordagem de Ward tradicional.

Figura 8.3: Mapas contendo entre 2 e 5 CGDs delineadas utilizando (a) o cálculo de *SJ* e a tesselação inicial da abordagem *SWMU Clustering*; e (b) a abordagem de Ward tradicional.



(a) (b)

Por meio da Figura 8.3 (a), é possível verificar que os novos procedimentos propostos pela abordagem *SWMU Clustering* proporcionam uma diminuição considerável da estratificação visualizada nos mapas, quando comparados com as soluções obtidas a partir da abordagem de Ward tradicional (Figura 8.3 (b)). Entretanto, em alguns casos, como nos mapas contendo 4 e 5 CGDs, pode ser observada a presença de UGDs muito pequenas, e que ainda proporcionam um efeito de estratificação. Assim, o parâmetro *porcDMU*, utilizado no desenvolvimento do critério *SD-Spatial*, também foi incluído na abordagem *SWMU Clustering*. Com isso, o usuário final pode definir um tamanho mínimo desejado para a área de UGDs que podem compor uma CGD.

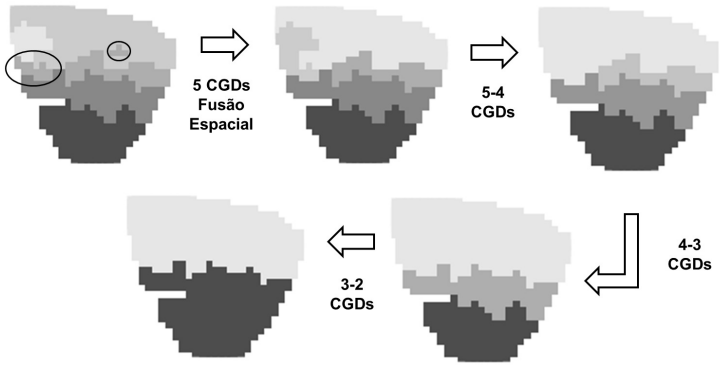
8.2.2.2 Área Mínima para UGDs

A utilização do parâmetro opcional *porcDMU* na execução da abordagem *SWMU Clustering*, que determina o tamanho de área mínimo para uma UGD, bem como a possibilidade de determinar uma quantidade mínima e máxima de CGDs que podem ser utilizadas na prática pela aplicação, sugeriu a criação de outros dois parâmetros adicionais: *minDMC* e *maxDMC*, que determinam, respectivamente, a quantidade mínima e máxima de CGDs desejadas pelo usuário final. Apesar de a abordagem hierárquica possuir a característica exploratória, na prática a divisão de uma área em uma quantidade muito alta de CGDs faz com que a usabilidade do conceito não seja completamente atingida, o que leva os usuários finais a normalmente subdividir uma

área de cultivo em até 5 CGDs. Considerando os parâmetros *porcDMU* e *maxDMC*, a abordagem *SWMU Clustering* realiza uma checagem no tamanho em área de todas as UGDs, quando o passo da hierarquia que gera a quantidade de CGDs informada pelo parâmetro *maxDMC* for atingido. Se, nesse passo, existirem UGDs consideradas como estratos, ou seja, com área relativa à área total menor do que o valor informado pelo usuário final para o parâmetro *porcDMU*, essas UGDs serão fundidas à UGDs adjacentes e espacialmente mais próximas, considerando como medida de proximidade a distância entre centroides no espaço de coordenadas. Essa checagem é realizada de maneira recursiva, até que não existam mais UGDs consideradas como estratos e a hierarquia possa ser finalizada.

Assim, se o usuário escolher, por exemplo, um valor de 5% para o parâmetro *porcDMU*, e uma quantidade máxima de 5 CGDs por meio do parâmetro *maxDMC*, a estratificação exibida na Figura 8.3 (a), presente nos mapas contendo 4 e 5 CGDs, será suprimida (Figura 8.4).

Figura 8.4: Passos finais da hierarquia de construção do dendrograma da abordagem *SWMU Clustering*, a partir da utilização dos parâmetros *porcDMU* e *maxDMC*.



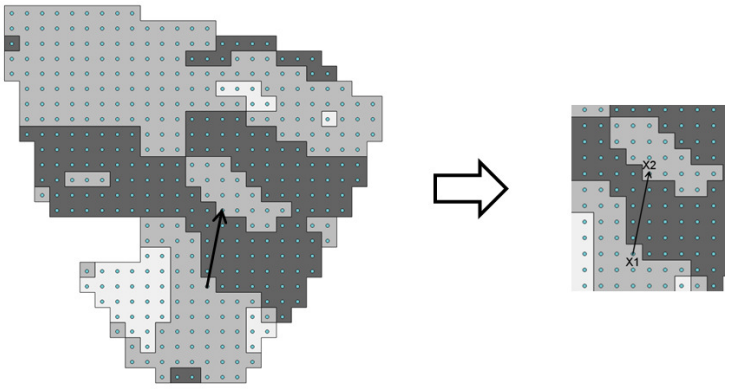
Por meio da Figura 8.4, é possível verificar a identificação de 2 UGDs, marcadas no mapa original de 5 CGDs, com tamanho em área menor do que 5% da área total. Nesse instante, a abordagem *SWMU Clustering* funde essas UGDs com a sua UGD mais próxima, considerando o espaço de coordenadas. Nesse caso, por essa fusão ocorrer sempre entre duas UGDs adjacentes, é natural que a UGD antes considerada como estrato seja associada a uma nova CGD.

8.2.2.3 Obstáculos Espaciais

A última restrição espacial opcional incluída no desenvolvimento da abordagem *SWMU Clustering* constitui-se da verificação de obstáculos espaciais presentes entre as amostras como fator de ponderação para os cálculos de dissimilaridade. Para que essa restrição seja utilizada, é necessária a presença de um conjunto de dados espaciais que identifique os obstáculos e permita verificar se existe obstrução em um suposto caminho direto entre duas amostras, considerando o

espaço de coordenadas. A Figura 8.5 exibe um mapa com a distribuição das amostras representadas por pontos em uma grade regular, em conjunto com uma camada poligonal que define a porcentagem de declividade, obtida a partir de imagens SRTM com resolução espacial original de 30 metros.

Figura 8.5: Mapa contendo amostras distribuídas em grade regular e camada poligonal representando classes de declividade, com caminho entre duas amostras representado no detalhe.



Na Figura 8.5, as diferentes tonalidades de cinza da camada poligonal identificam regiões com diferentes porcentagens de declividade, segundo a classificação proposta em Embrapa (1979). Nessa representação, as regiões com tons de cinza mais escuros possuem porcentagem de declividade maior do que as regiões representadas em tons de cinza mais claros. No detalhe, está representado qual seria o caminho direto, considerando o espaço de coordenadas, entre as amostras X1 e X2. Pode-se observar também, que apesar das amostras X1 e X2 estarem localizadas em áreas pertencentes à mesma classe de declividade, existe uma região com declividade mais alta que intercepta esse caminho. Na abordagem *SWMU Clustering*, esse tipo de região é considerada como um obstáculo espacial ou barreira, proporcionando um incremento no cálculo da dissimilaridade realizada durante a construção da hierarquia, caso uma dessas amostras seja considerada, em algum momento, o centroide de um grupo. Assim, a dissimilaridade entre duas amostras, considerando a existência de uma barreira entre elas, é calculada pela abordagem *SWMU Clustering* segundo a Equação 8.2.

$$d(X1,X2;B) = \frac{\sum_{i=1}^{nB} d(X1,x_i)}{nB} + \frac{\sum_{i=1}^{nB} d(X2,x_i)}{nB} \tag{8.2}$$

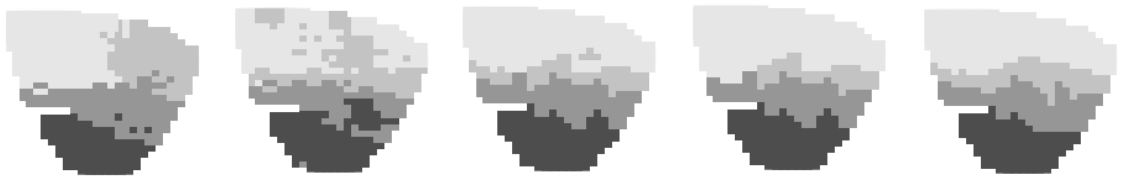
Na equação 8.2, B representa a barreira poligonal presente entre X1 e X2, onde estão espacialmente contidas nB amostras. A dissimilaridade entre as duas amostras X1 e X2 é então calculada somando-se a dissimilaridade média entre X1 e todas as nB amostras que estão contidas na barreira B com a dissimilaridade média entre X2 para as mesmas amostras contidas nessa barreira. Essa dissimilaridade é calculada considerando apenas as variáveis do espaço

de atributos e utilizando a distância Euclidiana como medida, assim como ocorre em todo o processo hierárquico da abordagem *SWMU Clustering*.

8.2.2.4 Visão Geral e Conclusões

A Figura 8.6 exibe os efeitos causados pela inclusão de cada uma das restrições espaciais desenvolvidas na abordagem *SWMU Clustering* em mapas contendo 4 CGDs, comparando-os com os resultados obtidos pela abordagem de Ward, utilizando sempre o mesmo conjunto de atributos.

Figura 8.6: Mapas com 4 CGDs delineadas (a) com a utilização da abordagem de Ward; e com a utilização da abordagem *SWMU Clustering* incluindo, de maneira incremental: (b) a restrição do centroide espacial no cálculo do erro quadrático; (c) a tesselação inicial; (d) o tamanho mínimo para UGDs; e (e) os obstáculos espaciais.

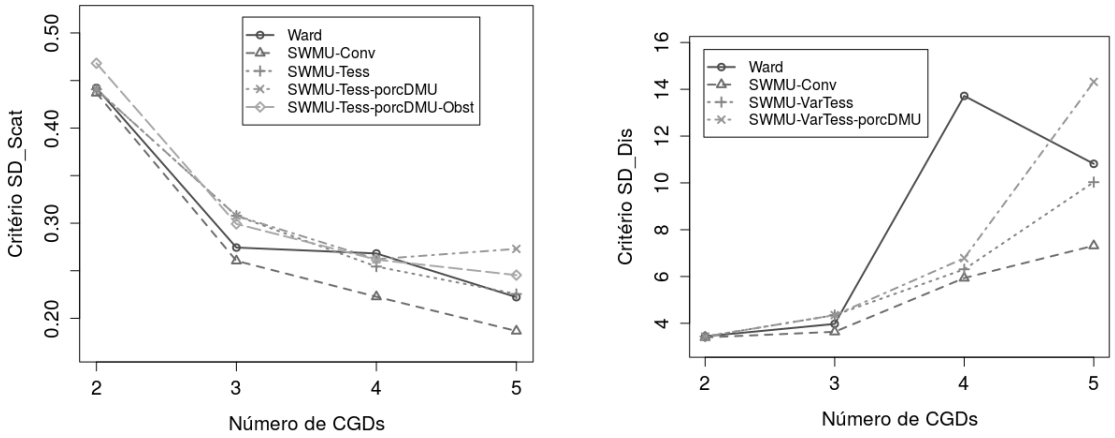


(a) (b) (c) (d) (e)

Em resumo, a partir de uma análise visual dos resultados exibidos na Figura 8.6, algumas conclusões podem ser obtidas. A simples obtenção do centroide de um grupo considerando o espaço de coordenadas reduz a estratificação em alguns pontos do mapa, porém pode gerar um aumento da mesma em outras regiões que anteriormente eram mais contínuas (Figura 8.6 (b)). A inclusão da tesselação inicial proporciona uma redução considerável da estratificação, porém UGDs muito pequenas ainda podem estar presentes (Figura 8.6 (c)), mas que podem ser suprimidas com a utilização do parâmetro *porcDMU* pelo usuário final (Figura 8.6 (d)). Finalmente, a utilização de obstáculos proporciona modificações no mapa final porque possui o intuito de dificultar o agrupamento, o quanto for possível, de amostras separadas por barreiras geográficas, como exemplo o uso de diferentes níveis de declividade (Figura 8.6 (e)).

Adicionalmente, é importante verificar o efeito do acúmulo das restrições impostas pela abordagem *SWMU Clustering* nos índices de coesão e separação do espaço de atributos obtidos pelos agrupamentos gerados. A Figura 8.7 exibe dois gráficos independentes contendo, respectivamente, os índices de coesão e separação para o critério SD obtidos pela abordagem tradicional de Ward, em comparação com a abordagem *SWMU Clustering* acumulando de maneira incremental as restrições espaciais. Os resultados foram obtidos utilizando-se como entrada o mesmo conjunto de atributos.

Figura 8.7: Índices de (a) coesão e (b) separação do critério SD obtidos pela abordagem tradicional Ward e pela abordagem *SWMU Clustering* considerando diferentes restrições espaciais.

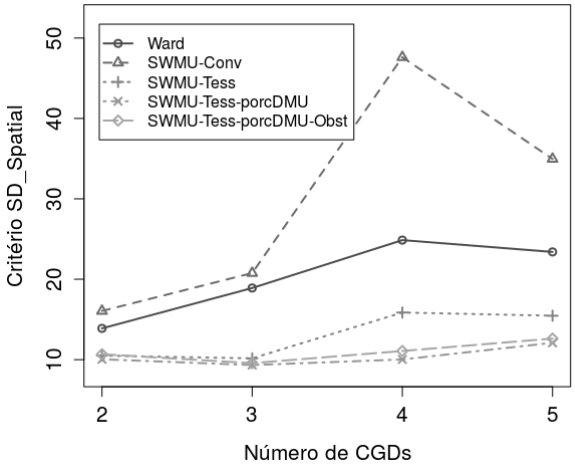


(a) (b)

Por meio dos gráficos da Figura 8.7, onde valores menores indicam grupos mais coesos e bem separados, e tomando como solução de referência os resultados obtidos pela abordagem tradicional de Ward, podem ser verificados ganhos na eficácia dos resultados, em alguns casos bem significativos, da abordagem *SWMU Clustering* utilizando apenas o cálculo diferenciado do erro quadrático a partir do centroide espacial (SWMU-Conv). Em média, com o uso dessa abordagem o ganho foi de 9,7% para a coesão, e de 24,7% para a separação, em relação a abordagem tradicional de Ward. À medida que as restrições espaciais são adicionadas à abordagem *SWMU Clustering*, o ganho com a coesão e a separação no espaço de atributos com relação aos resultados obtidos pela abordagem de Ward tende a diminuir, passando em alguns casos a proporcionar pequenas perdas, principalmente para 4 e 5 CGDs. Considerando o incremento da tesselação inicial (SWMU-Tess), ocorreu uma perda média de 2,1% para a coesão, porém ainda foi observado um ganho de 13,0% na separação. Já a inclusão da restrição do tamanho mínimo para uma UGD (SWMU-Tess-PorcDMU), proporcionou uma perda média de 8,22% com relação à coesão, mas ainda um ganho médio de 2,24% na separação entre os grupos. Finalizando, a inclusão dos obstáculos espaciais (SWMU-Tess-PorcDMU-Obst) proporcionou uma perda média de 5,79% para a coesão, mas um ganho médio ainda de 1,87% na separação. Apesar de proporcionar uma diminuição na coesão dos grupos formados no espaço de atributos, em comparação com o que pode ser obtido com a utilização da abordagem tradicional de Ward, as restrições espaciais propostas na abordagem *SWMU Clustering* proporcionam, em geral, uma diminuição significativa da estratificação dos mapas de UGDs, cuja qualidade considerando tanto o espaço de atributos quanto o espaço de coordenadas pode ser medida a partir do critério *SD-Spatial*. A Figura 8.8 mostra um gráfico contendo os índices de coesão e separação

obtidos com a utilização do critério *SD-Spatial*, comparando os mesmos resultados obtidos pela abordagem de Ward e as variações de restrição espacial da abordagem *SWMU Clustering*.

Figura 8.8: Índices para o critério *SD-Spatial* obtidos pela abordagem tradicional Ward e pela abordagem *SWMU Clustering* considerando diferentes restrições espaciais.



A partir do gráfico da Figura 8.8, pode-se verificar a importância da utilização da tesselação inicial, que proporcionou a obtenção de mapas mais contíguos e impediu a estratificação de regiões anteriormente contínuas, permitindo a obtenção de índices para o *SD-Spatial* bem melhores em comparação com as abordagens que não utilizaram esse artifício (Ward e SWMU-Conv). De uma maneira geral, a utilização da abordagem *SWMU Clustering* utilizando apenas o cálculo diferenciado do erro quadrático a partir do centroide espacial proporcionou uma perda média de coesão e separação, considerando tanto o espaço de atributos quando o espaço de coordenadas, de 41,6% com relação à abordagem de Ward. Entretanto, foram obtidos ganhos médios 35,2% utilizando a tesselação inicial (SWMU-Tess), 46,6% incrementando-se o tamanho mínimo para uma UGD (SWMU-Tess-porcDMU), e 43,5% incrementando-se os obstáculos espaciais (SWMU-Tess-porcDMUObst).

A partir dessas análises, pode-se concluir que as restrições espaciais incluídas na abordagem *SWMU Clustering* podem proporcionar, em alguns casos, pequenas prejuízos com relação à coesão dos agrupamentos e, consequentemente, dos mapas de UGDs gerados. Entretanto, esses perdas podem ser consideradas não significativas, se comparadas com os ganhos obtidos por meio da obtenção de mapas menos estratificados e de mais fácil interpretação e, consequentemente, mais coesos e bem separados considerando tanto o espaço de atributos quanto o espaço de coordenadas.

A seguir, são descritos experimentos executados variando-se os conjuntos de atributos e parâmetros que permitem comparar a abordagem proposta *SWMU Clustering* com as abordagens

FCM e *HACC-Spatial*. Esses experimentos tiveram como objetivo principal possibilitar a verificação das situações em que o uso dessa nova abordagem é mais vantajoso e pode proporcionar melhores resultados ao usuário final.

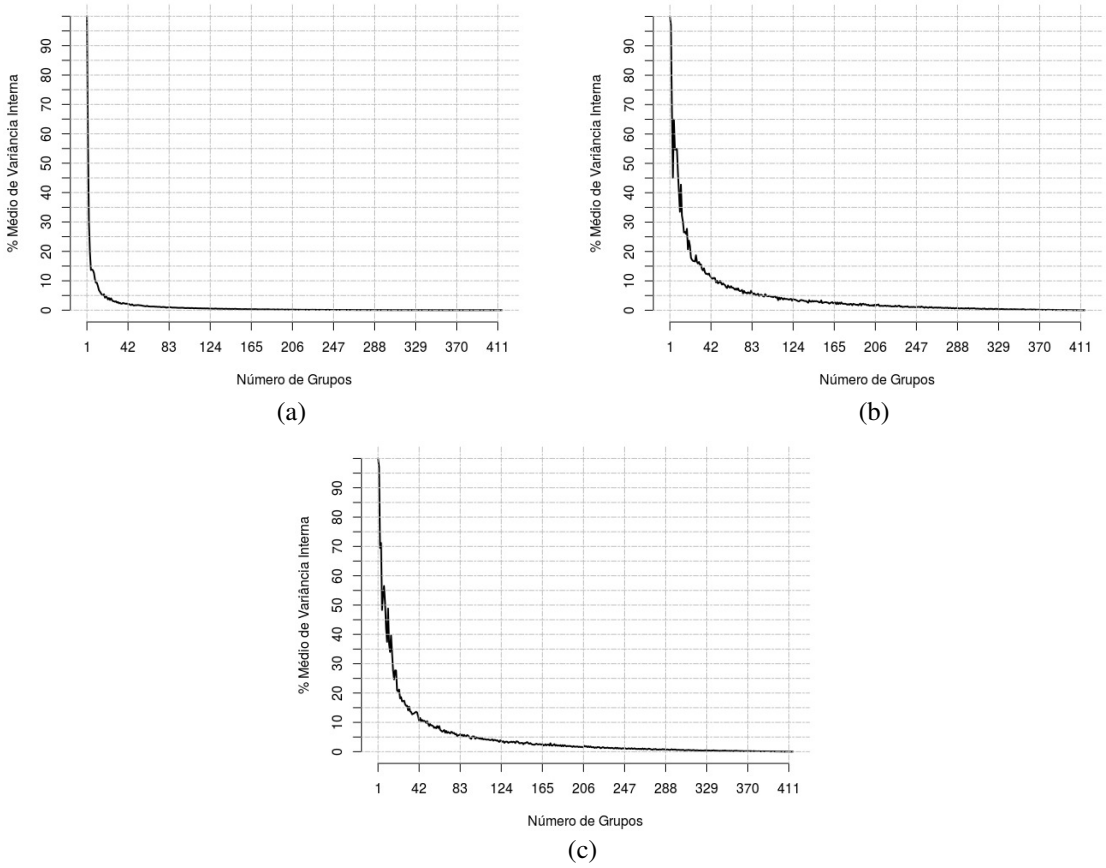
8.3 Análise Comparativa da Abordagem *SWMU Clustering* com Relação ao Estado da Arte

Os experimentos descritos nesta seção foram realizados com o intuito de analisar, de maneira relativa e considerando a coesão e a separação dos agrupamentos gerados tanto no espaço de atributos quanto no espaço de coordenadas, os resultados obtidos pela abordagem *SWMU Clustering* com relação às abordagens FCM e *HACC-Spatial*, indicando as vantagens e desvantagens de sua utilização em diferentes cenários. Para a realização desses experimentos, foram utilizados os conjuntos de atributos obtidos a partir das UPs descritas na Seção 6.5 do Capítulo 6.

8.3.1 Experimentos Utilizando Atributos da UP-CA

Inicialmente, foram realizados experimentos utilizando dados oriundos da UP-CA, respeitando os conjuntos predefinidos de atributos UP-CA1, UP-CA2 e UP-CA3. Com o intuito de guiar a escolha de valores para os parâmetros *varTess* e *k*, que devem ser fornecidos pelo usuário final, respectivamente, para as abordagens *SWMU Clustering* e *HACC-Spatial* (no caso de utilização da restrição da tesselação inicial), foram gerados gráficos contendo uma possível divisão de grupos utilizando o algoritmo *k-means* tradicional considerando o espaço de coordenadas e a sua respectiva coesão interna no espaço de atributos calculada pelo critério SD (Equação 4.7a). A partir desse momento, essa coesão interna será identificada como percentual médio de variância interna (MVI) (Figura 8.9).

Figura 8.9: Percentual médio de variância interna (MVI) para dados oriundos da UP-CA, considerando os conjuntos de atributos (a) UP-CA1; (b) UP-CA2; e (c) UP-CA3.

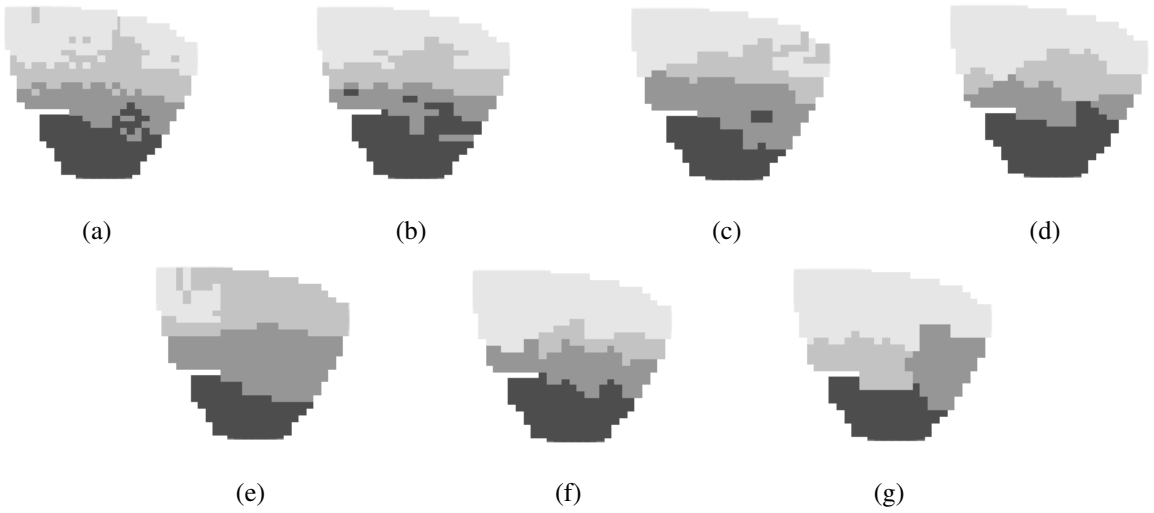


Para todos os gráficos da Figura 8.9, é possível definir três tipos de tesselação para cada conjunto de atributos: utilizando grandes quantidades de grupos, onde o percentual MVI é praticamente nulo; considerando quantidades de grupos menores que a metade da quantidade de amostras do conjunto de dados, onde o percentual MVI possui uma tendência geral de aumento; e considerando uma quantidade menor de grupos, onde ocorre em geral um aumento mais brusco no percentual MVI. Essas constatações permitiram a realização de experimentos determinando-se três valores distintos para os parâmetros *varTess* e *k*, utilizados, respectivamente, pelas abordagens *SWMU Clustering* e *HACC-Spatial* para execução da tesselação inicial. Com relação à abordagem FCM, para todos os experimentos foi utilizado o valor padrão igual a 2 para o parâmetro *m*, levando-se em consideração os experimentos preliminares descritos na Seção 5.4 do Capítulo 5.

8.3.1.1 Conjunto de Atributos UP-CA1 e Variações de Parâmetros

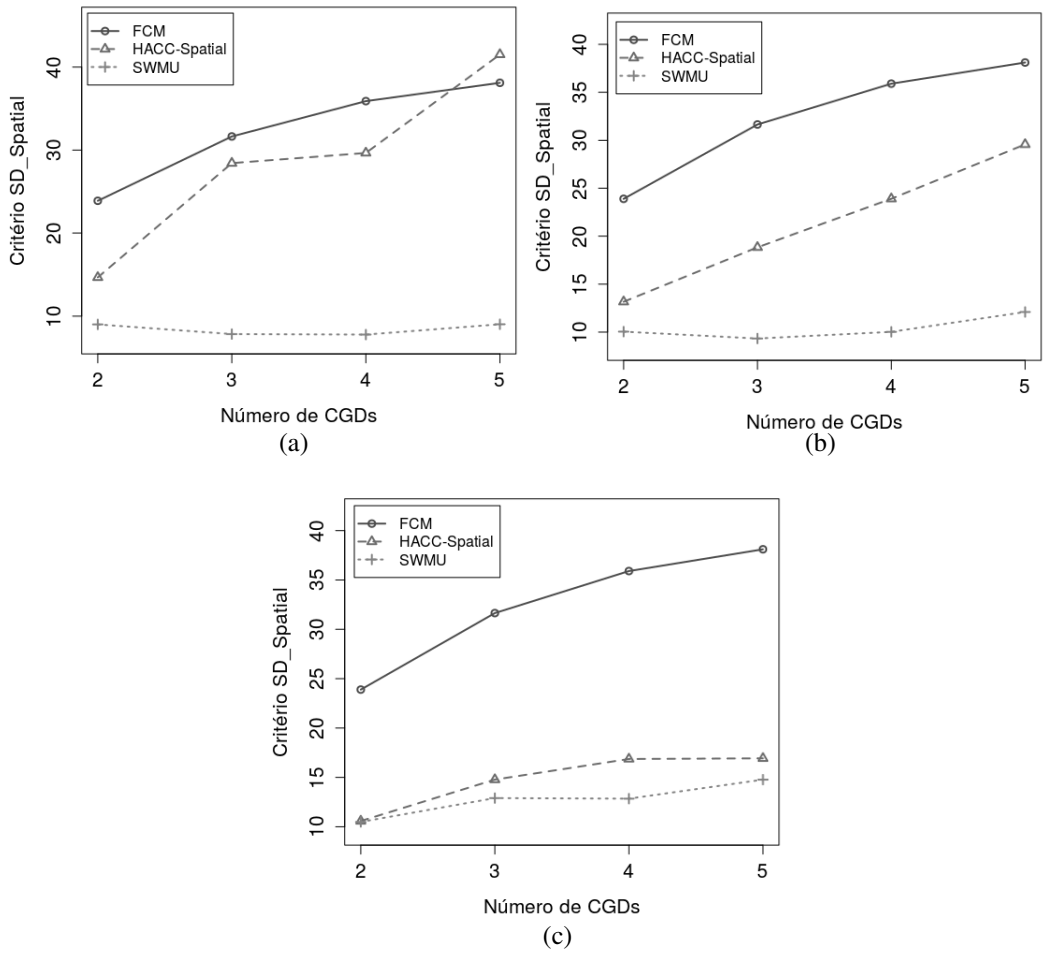
Inicialmente, considerando o conjunto de atributos UP-CA1 e o gráfico da Figura 8.9 (a), foram definidos os valores de 0,3%; 1%; e 2,5% para o parâmetro *varTess*, da abordagem *SWMU Clustering*, e, de maneira similar e correspondente, os valores de 207, 88 e 32 para o parâmetro *k*, da abordagem *HACC-Spatial*. Ainda, para a abordagem *SWMU Clustering*, foi fixado o valor de 5% para o parâmetro *porcDMU* e não foi utilizada a restrição de obstáculos espaciais; e para a abordagem *HACC-Spatial*, foi fixado o valor de 0,5 para o parâmetro cp. A Figura 8.10 exibe mapas contendo 4 CGDs delineadas considerando essas variações.

Figura 8.10: Mapas contendo 4 CGDs delineadas utilizando as abordagens (a) FCM, com *m=2*; *HACC-Spatial*, com *cp=0,5* e *k* igual a (b) 207, (c) 88, e (d) 32; e *SWMU Clustering*, com *porcDMU=5%* e *varTess* igual a (e) 0,3%, (f) 1%, e (g) 2,5%, sem a restrição de obstáculos espaciais.



De uma maneira geral, pode-se observar uma menor estratificação presente nos mapas obtidos utilizando a abordagem *SWMU Clustering* (Figura 8.10 (e)-(g)) do que nos mapas obtidos pela abordagem *HACC-Spatial* (Figura 8.10 (b)-(d)) e pelo algoritmo FCM (Figura 8.10 (a)). De maneira complementar, os índices alcançados pelo critério *SD-Spatial* (Figura 8.11) permitem uma análise qualitativa mais elaborada, considerando tanto o espaço de atributos quanto o espaço de coordenadas, fornecendo subsídios ao usuário para uma tomada de decisão mais efetiva.

Figura 8.11: Índices alcançados pelo critério *SD-Spatial*, considerando agrupamentos gerados pelo FCM com *m=2* e (a) *HACC-Spatial* com *k=207* e *SWMU Clustering* com *varTess=0,3%*; (b) *HACC-Spatial* com *k=88* e *SWMU Clustering* com *varTess=1%*; e (c) *HACC-Spatial* com *k=32* e *SWMU Clustering* com *varTess=2,5%*.

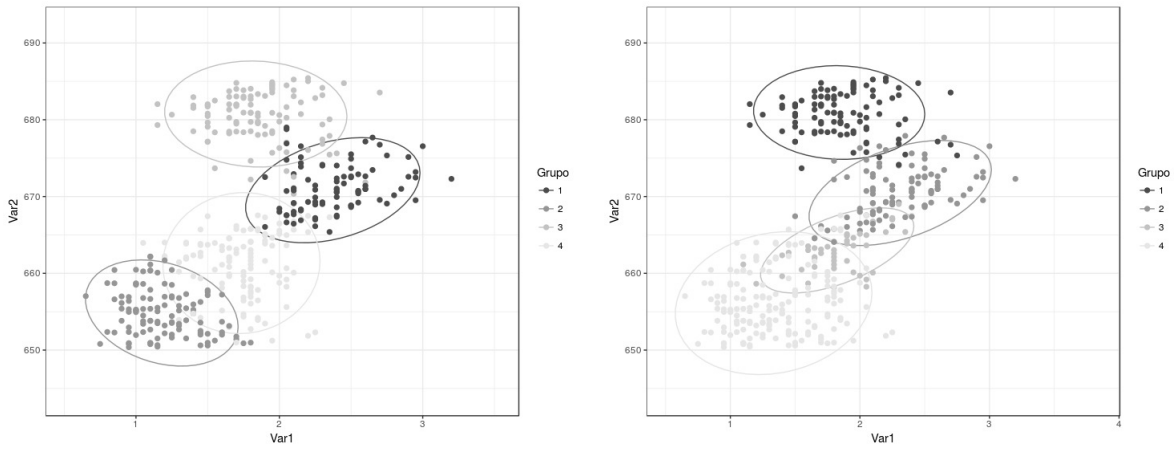


Por meio dos gráficos da Figura 8.11 é possível observar, em geral, um ganho das abordagens hierárquicas com relação à abordagem FCM, considerando que para o critério *SD-Spatial* valores mais baixos indicam melhores resultados. Na maioria dos casos, a abordagem *SWMU Clustering* apresentou melhores resultados com relação a abordagem *HACC-Spatial*. Além da utilização do mesmo valor para o parâmetro *porcDMU* pela abordagem *SWMU Clustering* e pelo critério *SD-Spatial* (5%), contribuiu também para esse resultado a estratificação ainda observada em diversos mapas obtidos utilizando a abordagem *HACC-Spatial*, principalmente quando a tesselação inicial é realizada considerando uma maior quantidade de grupos. Essa estratificação só é efetivamente reduzida quando é utilizado o valor de $k = 32$, proporcionando para a abordagem *HACC-Spatial* índices para o critério *SD-Spatial* mais próximos dos obtidos pela abordagem *SWMU Clustering* (Figura 8.11 (c)).

Um dos motivos que fazem com que as abordagens de agrupamento tradicionais, como

a FCM, proporcionem a obtenção de mapas de UGDs extremamente estratificados, como o que é exibido na Figura 8.10 (a), é o fato de essas abordagens possuírem a tendência de obter grupos com formas normalmente esféricas considerando o espaço de atributos, pois utilizam-se do centroide obtido a partir de atributos convencionais para alocar as amostras nos seus respectivos grupos. Por outro lado, as restrições espaciais impostas por abordagens como a *SWMU Clustering*, fazem com que essa estratificação seja reduzida, assim como a tendência de obtenção de grupos esféricos no espaço de atributos. A Figura 8.12 exibe os agrupamentos formados, considerando um espaço bidimensional a partir de duas variáveis convencionais do conjunto de atributos UP-CA1, para os mapas de UGDs exibidos nas figuras 8.10 (a) e 8.10 (g).

Figura 8.12: Configurações de agrupamentos visualizadas a partir de um espaço bidimensional formado por variáveis do espaço de atributos, obtidas por meio das abordagens (a) FCM; e (b) *SWMU Clustering*.

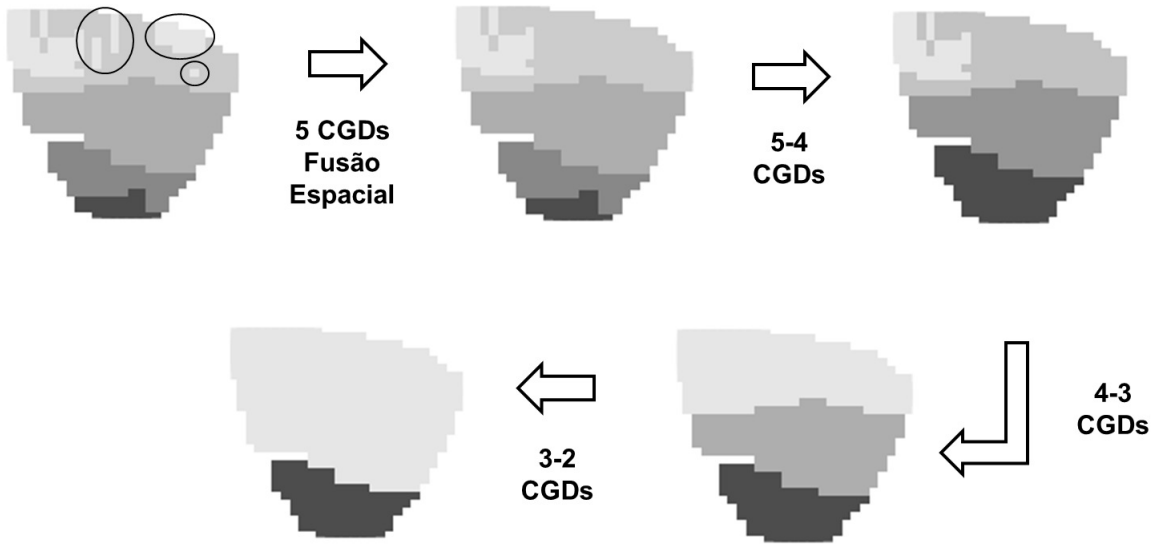


Por meio do gráfico da Figura 8.12 (a), é possível verificar a tendência de formação de grupos esféricos em torno dos centroides do espaço de atributos quando é utilizada uma abordagem de agrupamento convencional, como a FCM. Uma alternativa para reduzir essa tendência poderia ser a utilização de medidas de dissimilaridades alternativas, como a distância de Mahalanobis, que se baseia em correlações entre as variáveis envolvidas. Entretanto, em experimentos preliminares realizados, verificou-se que o uso dessa distância pouco contribui para reduzir a estratificação dos mapas de UGDs gerados pela abordagem FCM. Por outro lado, ao utilizarmos abordagens com restrições espaciais, neste caso considerando o centroide espacial, a tesselação inicial e o tamanho mínimo para a área de uma UGD presentes na abordagem *SWMU Clustering*, pode ser observada uma dificuldade maior em se delinear esferas no espaço de atributos contendo amostras pertencentes a um único grupo (Figura 8.12 (b)). Com isso, as restrições espaciais da abordagem *SWMU Clustering* diminuem a chance de serem obtidos agrupamentos

contendo grupos exclusivamente esféricos considerando o espaço de atributos.

Além dos efeitos causados por variações no parâmetro *varTess*, que permite a comparação direta das abordagens *SWMU Clustering* e *HACC-Spatial* por ambas possuírem uma etapa de tesselação inicial, também foram realizados experimentos variando-se o parâmetro *porcDMU* da abordagem *SWMU Clustering*, que permite ao usuário final determinar uma restrição espacial de tamanho mínimo permitido para uma UGD. Nesses experimentos, verificou-se que, ao aumentarmos o valor desse parâmetro, ocorrem situações em que a abordagem *SWMU Clustering* não consegue fornecer a quantidade máxima de CGDs pretendida pelo usuário final, definida pelo parâmetro *maxDMC*. Essa questão é observada principalmente em casos onde uma CGD é formada por uma única UGD que, ao não atingir a área mínima definida pelo usuário final, deve ser associada a outra CGD, diminuindo assim a quantidade máxima de CGDs possíveis definida pelo parâmetro *maxDMC*. Em um primeiro experimento, foram utilizadas variações do parâmetro *porcDMU* para os valores de 5%, 10% e 15%, utilizando valores fixos dos parâmetros *varTess* igual 0,3%, *minDMC* igual a 2 e *maxDMC* igual a 5, e a não utilização da restrição de obstáculos espaciais. A Figura 8.13 exibe os mapas de 2 a 5 CGDs obtidos utilizando *porcDMU* igual a 5%, bem como o mapa contendo 5 CGDs que sofreu os efeitos da restrição espacial imposta por esse parâmetro.

Figura 8.13: Mapas contendo de 2 a 5 CGDs geradas utilizando a restrição de tamanho mínimo de UGDs imposta pela utilização do parâmetro *porcDMU=5%* pela abordagem *SWMU Clustering*.



Por meio da Figura 8.13, é possível observar o processo realizado pela abordagem *SWMU Clustering* quando é atingido o limite máximo de CGDs desejado pelo usuário, que nesse caso é 5. Similar ao que aconteceu no exemplo da Figura 8.4, as UGDs com tamanho de área menor ou igual à 5% da área total, destacadas no primeiro mapa, são fundidas às UGDs espacialmente

mais próximas, forçando a obtenção de um mapa com 5 CGDs que garante a restrição imposta pelo parâmetro *porcDMU*. Na sequência, a abordagem segue os passos normais de construção do dendrograma, até que seja obtido o mapa com 2 CGDs. Ao aumentarmos o valor do parâmetro *porcDMU* para 10%, a fusão das UGDs pequenas em uma mapa contendo 5 CGDs não foi suficiente, fazendo com que fossem fornecidos como resultados apenas os mapas contendo entre 2 e 4 CGDs, idênticos aos exibidos na Figura 8.13. Nesse caso, a única UGD da CGD representada pelo tom de cinza mais escuro, localizada na extremidade inferior do mapa de 5 CGDs, possuía tamanho de área menor ou igual a 10% da área total, e teve que ser fundida com a UGD mais próxima, que também era representante única da sua CGD, fazendo com que a quantidade máxima de CGDs diminuísse. Esse processo também foi observado quando aumentou-se o parâmetro *porcDMU* para 15%, onde inicialmente aconteceram as mesmas fusões do caso anterior. Entretanto, no mapa contendo 4 CGDs, a UGD representada pelo tom de cinza mais claro, e que também era a única representante da CGD à qual pertencia, possuía tamanho de área menor ou igual a 15% da área total, sendo também fundida com a UGD mais próxima, também representante única de sua CGD. Assim, com esse novo aumento no valor do parâmetro *porcDMU*, a abordagem *SWMU Clustering* conseguiu fornecer ao usuário apenas os dois mapas contendo, respectivamente, 2 e 3 CGDs.

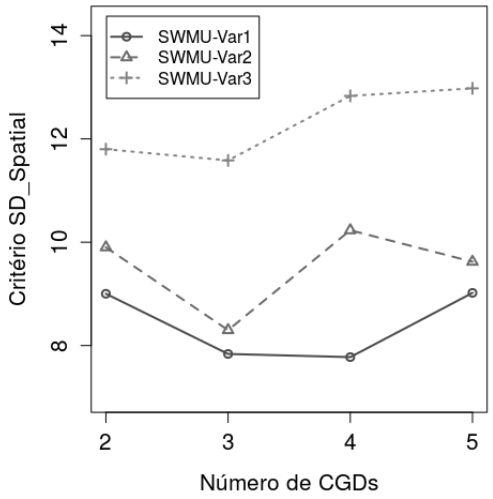
De maneira similar, foram realizados novos experimentos, porém variando-se o valor de *maxDMC* para 10. Nesse caso, utilizando-se o valor de 5% para o parâmetro *porcDMU*, a abordagem *SWMU Clustering* conseguiu fornecer como resultado mapas contendo de 2 à 9 CGDs; e utilizando-se o valor de 10% para esse mesmo parâmetro, foi possível a obtenção de mapas contendo de 2 à 6 CGDs. Para o valor de 15%, a quantidade de mapas possíveis foi bastante reduzida, fazendo com que a abordagem fornecesse, da mesma forma que no experimento anterior, apenas mapas contendo 2 e 3 CGDs. É importante destacar que, da mesma forma que acontece com as variações do parâmetro *varTess*, os diferentes valores utilizados para os parâmetros *porcDMU* e *maxDMC*, que estão totalmente relacionados no processo de agrupamento determinado pela abordagem *SWMU Clustering*, podem proporcionar resultados diferentes. A Figura 8.14 exibe mapas contendo 5 CGDs, utilizando variações desses parâmetros descritas nos dois últimos experimentos.

Figura 8.14: Mapas contendo 5 CGDs obtidos com a utilização da abordagem *SWMU Clustering* sem considerar a restrição de obstáculos espaciais, com *varTess=0,3%*, *minDMC=2* e (a) *porcDMU=5%* e *maxDMC=5*; (b) *porcDMU=5%* e *maxDMC=10*; e (c) *porcDMU=10%* e *maxDMC=10*.

(a) (b) (c)

Por meio da Figura 8.14, podem ser observadas apenas algumas diferenças no mapa final de 5 UGDs quando se aumenta apenas a quantidade máxima de CGDs desejada. Entretanto, a variação do parâmetro *porcDMU* proporciona alterações consideráveis, fazendo com que algumas CGDs sejam reposicionadas e tenham o seu tamanho em área bastante alterado. A Figura 8.15 exibe gráficos com os índices alcançados pelo critério *SD-Spatial* para agrupamentos contendo de 2 a 5 mapas de CGDs obtidos pelas mesmas variações de parâmetros utilizadas nos mapas da Figura 8.14.

Figura 8.15: Índices do critério *SD-Spatial* (com *porcDMU=5%*), para agrupamentos gerados a partir do conjunto de atributos UP-CA1 e valores de *porcDMU=5%* e *maxDMC=5* (SWMU-Var1); *porcDMU=5%* e *maxDMC=10* (SWMU-Var2); e *porcDMU=10%* e *maxDMC=10* (SWMU-Var3).



A partir do gráfico da Figura 8.15, é possível verificar que as diferenças exibidas no mapa da Figura 8.14 (c) com relação aos outros resultados proporciona uma perda significativa também nos índices de coesão e separação medidos pelo critério *SD-Spatial*. Desse modo, pode-se constatar que uma restrição de tamanho mínimo para uma UGD maior do que 5% da área total,

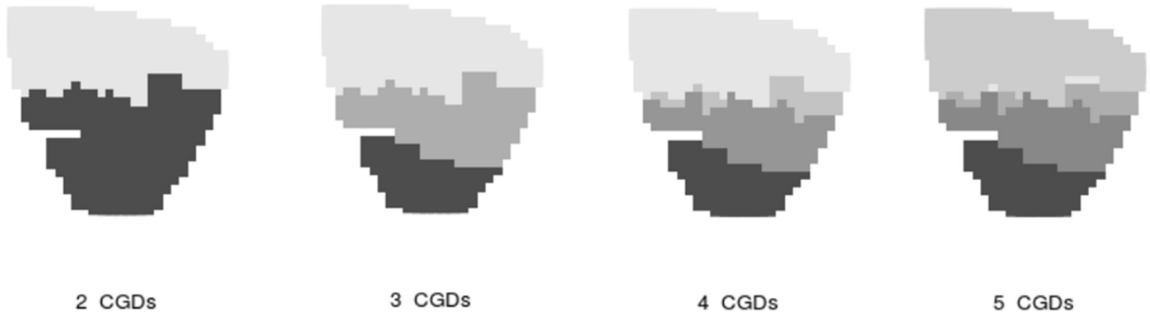
para a quantidade padrão entre 2 e 5 CGDs que normalmente é utilizada, pode influenciar demasiadamente no resultado final obtido pela abordagem *SWMU Clustering*. Além da obtenção de mapas de UGDs mais distantes do objetivo principal da restrição imposta pelo parâmetro *porcDMU* de apenas realizar ajustes para evitar a estratificação, essa restrição também pode impossibilitar a abordagem *SWMU Clustering* de fornecer mapas com a quantidade máxima de CGDs pretendida pelo usuário, prejudicando também o resultado final. Desse modo, para os próximos experimentos, o parâmetro *porcDMU* será fixado em 5%, tanto para abordagem *SWMU Clustering*, quanto para o critério de validação *SD-Spatial*. Entretanto, nada impede que o usuário utilize valores menores para esse parâmetro, tendo a consciência de que áreas muito pequenas não serão consideradas como estratos; ou valores maiores, quando se tem a restrição, por exemplo, da largura de plataforma e, consequentemente, da área mínima ocupada pelo implemento agrícola que realizará as intervenções em campo utilizando mapas de UGDs.

8.3.1.2 Conjunto de Atributos UP-CA1 e *Ensembles* de Agrupamentos

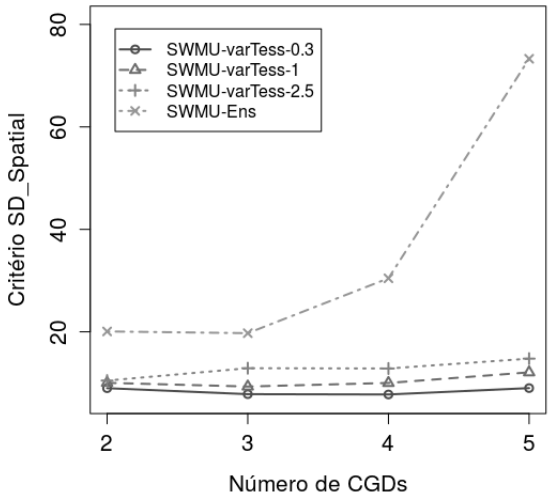
Após diversos experimentos utilizando-se variações de parâmetros da abordagem *SWMU Clustering* e comparações com relação as abordagens FCM e *HACC-Spatial*, a definição correta dos parâmetros ainda pode gerar dúvidas ao usuário final, principalmente com relação à tesselação inicial. Desse modo, os *ensembles* de agrupamentos, descritos no Capítulo 4, podem ser utilizados para que sejam obtidos mapas de UGDs mais robustos e eficazes. Em um experimento preliminar, relatado em publicação transcrita no Apêndice E (SPERANZA; CIFERRI; CIFERRI, 2016), as abordagens de *ensembles* desenvolvidas por Strehl e Ghosh (2002) foram utilizadas com o intuito de se obter soluções mais robustas a partir de agrupamentos individuais obtidos com o particionamento do conjunto de atributos e a utilização das abordagens de agrupamento FCM e *HACC-Spatial*. Nesse experimento, verificou-se que as abordagens de *ensembles* utilizadas acabaram proporcionando o aumento da estratificação dos mapas finais de UGDs. Complementarmente, um novo experimento para obtenção de agrupamentos de consenso foi realizado utilizando-se os mesmos agrupamentos originais, porém considerando a abordagem de *ensembles* por acúmulo de evidências (FRED; JAIN, 2005). Nesse novo experimento, relatado em artigo aceito para publicação transcrito no apêndice F, verificou-se a maior capacidade da abordagem de *ensembles* utilizando acúmulo de evidências em lidar com questões relacionadas à estratificação. Assim, realizou-se novo experimento utilizando essa abordagem, considerando agora como agrupamentos individuais os resultados alcançados pela abordagem *SWMU Clustering* descritos anteriormente para o conjunto de atributos UP-CA1, todos eles sem a utilização da restrição de obstáculos espaciais e com valores fixos para *minDMC* igual a 2, *maxDMC* igual a 5 e *porcDMU* igual a 5%; e variações do parâmetro *varTess* para 0,3%, 1% e 2,5%.

A Figura 8.16 mostra os agrupamentos de consenso obtidos nesse experimento, bem como um gráfico contendo os índices alcançados pelo critério SD-Spatial para comparação relativa com os agrupamentos individuais.

Figura 8.16: (a) Mapas contendo de 2 a 5 CGDs obtidos a partir da utilização de *ensembles* de agrupamentos por acúmulo de evidência; e (b) índices alcançados pelo critério *SD-Spatial* para os agrupamentos envolvidos no experimento.



(a)



(b)

Por meio da Figura 8.16 (a), é possível verificar a estratificação gerada pelo agrupamento de consenso nos mapas contendo 4 e 5 CGDs. Mesmo com os agrupamentos originais sendo bastante compatíveis com relação ao espaço de atributos, a solução de consenso não conseguiu fornecer resultados melhores em nenhum dos três agrupamentos originais, considerando os valores obtidos com a utilização do critério *SD-Spatial*.

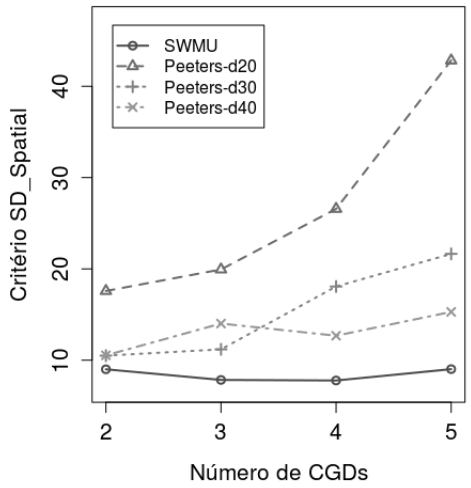
Os experimentos realizados com a utilização de *ensembles* no âmbito desta tese mostraram que as questões relacionadas à distribuição das UGDs no espaço de coordenadas prejudicam muito a principal função dessa técnica, que é fornecer soluções para reagrupar de maneira mais robusta diversos agrupamentos obtidos a partir de variações de abordagens, parâmetros e atri-

butos. Desse modo, para proporcionar a obtenção de mapas de UGDs mais robustos e úteis na prática, as abordagens de *ensembles* de agrupamentos tradicionais precisam ser adaptadas para considerar as peculiaridades existentes nos conjuntos de dados espaciais.

8.3.1.3 Conjunto de Atributos UP-CA1 e Abordagem de Peeters (2015)

Com o intuito de comparar os resultados obtidos pela abordagem *SWMU Clustering* com relação à abordagem desenvolvida por Peeters (2015), foi realizado um último experimento considerando o conjunto de atributos UP-CA1 e o critério de validação interna *SD-Spatial*. A Figura 8.17 mostra os resultados alcançados, considerando-se mapas contendo entre 2 e 5 CGDs obtidos por essas duas abordagens.

Figura 8.17: Índices alcançados pelos critério *SD-Spatial*, considerando mapas contendo entre 2 e 5 CGDs obtidos pela abordagem *SWMU Clustering* sem a utilização da restrição de obstáculos espaciais, e com *varTess=0,3%* e *porDMU=5%*; e pela abordagem desenvolvida por Peeters (2015) considerando *d*=$20 \times \sqrt{2}$ (Peeters-d20), *d*=$30 \times \sqrt{2}$ (Peeters-d30) e *d*=$40 \times \sqrt{2}$ (Peeters-d40).



A partir do gráfico da Figura 8.17, é possível observar que, considerando o critério *SD-Spatial*, alguns mapas fornecidos pela abordagem de Peeters (2015) ainda com certo grau de estratificação proporcionaram à abordagem *SWMU Clustering* resultados sempre melhores. Esse resultado evidencia a importância de se tratar de maneira diferenciada o espaço de coordenadas durante a execução da abordagem de agrupamento, e não somente em uma etapa de pré-processamento.

8.3.1.4 Conjunto de Atributos UP-CA2

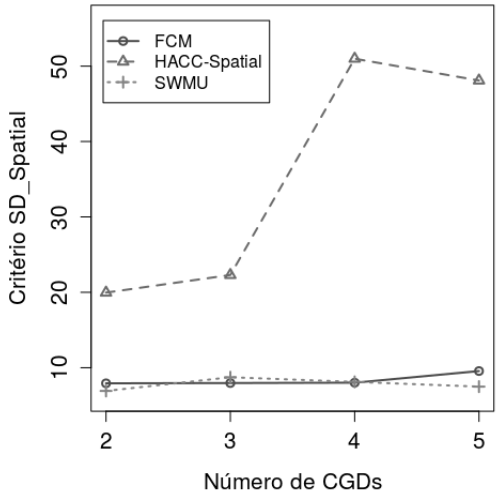
Os experimentos seguintes foram realizados utilizando-se o conjunto de atributos UP-CA2. Nesse caso, levando-se em consideração o gráfico da Figura 8.9 (b), verifica-se um maior aumento na porcentagem de MVI a partir de 5%, valor este definido para o parâmetro *varTess* utilizado pela abordagem *SWMU Clustering*. Nesses experimentos, a abordagem *SWMU Clustering* utilizou também a restrição de obstáculos espaciais proporcionada pelo mapa de classes de declividade. Consequentemente, para manter as comparações e compatibilidade com os resultados obtidos pela abordagem *HACC-Spatial*, foi utilizado o valor de k igual a 100, mantendo-se o valor padrão do parâmetro cp igual a 0,5. Com a utilização de um conjunto maior de atributos bem correlacionados, é possível verificar uma redução grande na estratificação dos mapas gerados, independentemente da abordagem utilizada. A Figura 8.18 exibe os mapas contendo 5 CGDs obtidos a partir dos atributos UP-CA2, utilizando as abordagens FCM, *HACC-Spatial* e *SWMU Clustering*.

Figura 8.18: Mapas contendo 5 CGDs obtidos a partir dos atributos UP-CA2, utilizando as abordagens (a) FCM, com *m=2*; (b) *HACC-Spatial*, com *k=100* e *cp=0,5*; e (c) *SWMU Clustering*, com *porcDMU=5%*, *varTess=5%* e a utilização da restrição de obstáculos espaciais.

(a) (b) (c)

Analisando-se os mapas da Figura 8.18, é possível observar a ausência de estratificação no mapa fornecido pela abordagem *SWMU Clustering*, e a presença de uma única UGD que pode ser considerada como estrato, nos mapas fornecidos pelas abordagens FCM e *HACC-Spatial*. Esse tipo de estratificação, mesmo influenciando muito pouco na visualização dos mapas gerados, pode auxiliar na validação dos agrupamentos formados quanto à coesão e a separação considerando tanto o espaço de atributos quanto o espaço de coordenadas. A Figura 8.19 exibe gráficos contendo os índices alcançados pelo critério *SD-Spatial* para mapas contendo de 2 a 5 CGDs, obtidos pelas abordagens FCM, *HACC-Spatial* e *SWMU Clustering* nesse mesmo experimento.

Figura 8.19: Índices obtidos pelo critério *SD-Spatial* para mapas contendo entre 2 e 5 CGDs, obtidas utilizando-se o conjunto de atributos UP-CA2 e as abordagens FCM, *HACC-Spatial* e *SWMU Clustering*.



Por meio do gráfico da Figura 8.19, é possível verificar que a estratificação gerada pela abordagem *HACC-Spatial*, onde foi obtida uma CGD representada por uma única UGD relativamente pequena nos mapas contendo 4 e 5 CGDs, é muito mais penalizada do que a estratificação gerada pela abordagem FCM, onde apenas para o mapa contendo 5 CGDs foi gerada uma UGD muito pequena, porém não única de uma das CGDs. Com isso, o desempenho das abordagens FCM e *SWMU Clustering* foi muito parecido considerando o critério *SD-Spatial*, prejudicando um pouco o índice atingido pela primeira abordagem para 5 CGDs, onde essa pequena estratificação foi observada.

8.3.1.5 Conjunto de Atributos UP-CA3

Em seguida, também foram realizados experimentos utilizando todos os atributos disponibilizados pela UP-CA, identificados pelo conjunto UP-CA3. Levando-se em consideração o gráfico da Figura 8.9 (c), verifica-se, assim como ocorreu para o conjunto de atributos UP-CA2, um maior aumento na porcentagem de MVI a partir de 5% para o conjunto de atributos UP-CA3, valor este então definido para o parâmetro *varTess* utilizado pela abordagem *SWMU Clustering*, proporcionando um valor de k igual a 100 para a abordagem *HACC-Spatial*. Nesses experimentos, a abordagem *SWMU Clustering* não utilizou a restrição de obstáculos espaciais. Entretanto, os valores padrão utilizados para os parâmetros cp igual a 0,5, para a abordagem *HACC-Spatial*, e *porcDMU* igual a 5%, para a abordagem *SWMU Clustering*, foram mantidos. A Figura 8.20 exibe mapas contendo 4 CGDs obtidos para esse experimento pelas abordagens

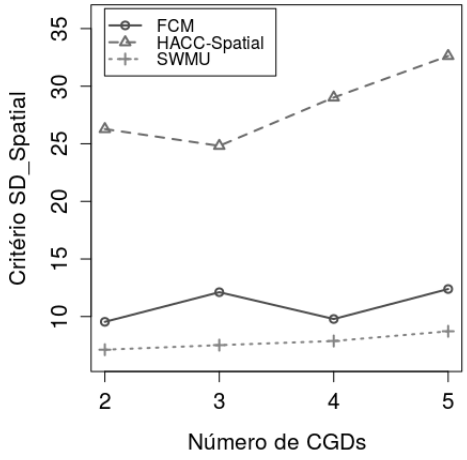
FCM, *HACC-Spatial* e *SWMU Clustering* considerando esses parâmetros.

Figura 8.20: Mapas contendo 4 CGDs obtidos com a utilização do conjunto de atributos UP-CA3 e as abordagens (a) FCM; (b) *HACC-Spatial* e (c) *SWMU Clustering*.

(a) (b) (c)

Por meio da Figura 8.20, observa-se uma redução muito grande na estratificação dos mapas de CGDs utilizando-se todos os atributos disponíveis para a UP-CA. Entretanto, ainda é verificada a presença de uma CGD de tamanho proporcionalmente bem menor que as outras para o resultado da abordagem *HACC-Spatial*. Esse resultado é refletido nos índices alcançados pelo critério *SD-Spatial* para essa abordagem, fazendo com que as abordagens *SWMU Clustering* e FCM obtivessem resultados muito melhores (Figura 8.21).

Figura 8.21: Índices alcançados pelo critério *SD-Spatial* para mapas contendo entre 2 e 5 CGDs obtidos utilizando-se as abordagens FCM, *HACC-Spatial* e *SWMU Clustering*.



Devido à grande quantidade de atributos disponibilizada pelo conjunto UP-CA3, foi realizado um experimento de redução de dimensionalidade por meio da técnica MULTISPATI-PCA. Assim como no trabalho de Córdoba (2013), foram selecionadas 3 componentes principais para serem utilizadas como dados de entrada para as abordagens *SWMU Clustering*, *HACC-Spatial* e FCM. Entretanto, ao utilizarmos apenas essas componentes como atributos de entrada, o valor de 5% atribuído para o parâmetro *varTess* gerou uma tesselação inicial para a abordagem

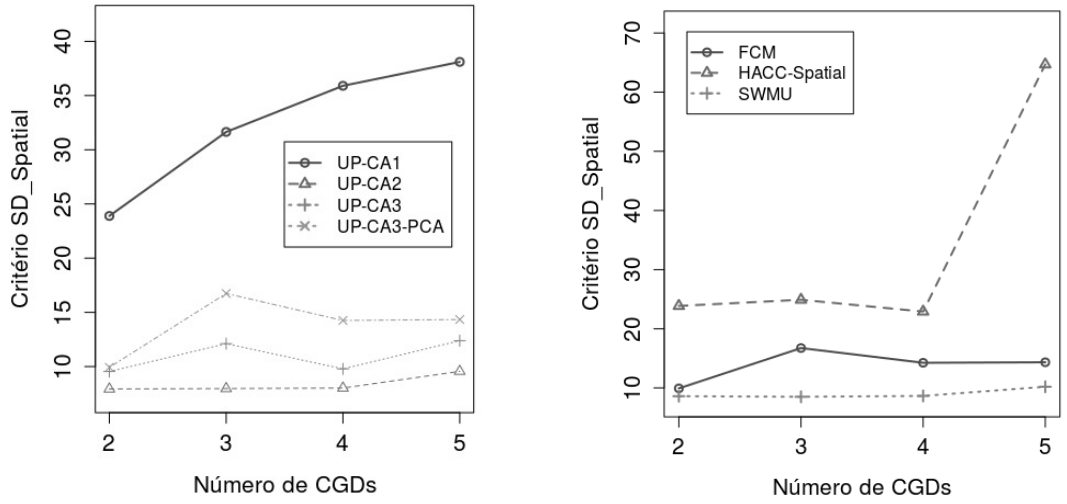
SWMU Clustering contendo 104 grupos, utilizado como valor de k para a abordagem *HACC-Spatial*. A Figura 8.22 exibe mapas contendo 3 CGDs obtidos pelas três abordagens utilizadas no experimento.

Figura 8.22: Mapas contendo 3 CGDs obtidos com a utilização das 3 componentes principais do conjunto de atributos UP-CA3, determinadas pela técnica MULTISPATI-PCA e as abordagens (a) FCM; (b) *HACC-Spatial* e (c) *SWMU Clustering*.

(a) (b) (c)

Por meio da Figura 8.22, é possível verificar que a redução da dimensionalidade provoca um aumento na estratificação do mapa obtido pela abordagem FCM, se considerarmos os resultados obtidos com a utilização dos conjuntos de atributos UP-CA2 e UP-CA3. Além disso, as três abordagens apresentam resultados bastante distintos visualmente, mesmo considerando apenas 3 CGDs. A Figura 8.23 exibe os índices alcançados por essas abordagens para o critério *SD-Spatial*, bem como os índices alcançados por esse mesmo critério considerando execuções da abordagem FCM utilizando a redução de dimensionalidade (UP-CA3-PCA) e os três conjuntos de atributos da UP-CA (UP-CA1, UP-CA2 e UP-CA3). Em ambos os gráficos, mesmo quando se tratam de agrupamentos gerados a partir de componentes principais obtidas do conjunto de atributos UP-CA3, o critério foi executado considerando os atributos originais, para que o efeito da redução de dimensionalidade pudesse ser efetivamente verificado.

Figura 8.23: Índices alcançados pelo critério *SD-Spatial* para mapas contendo entre 2 e 5 CGDs obtidos utilizando-se (a) a abordagem FCM e os conjuntos de atributos UP-CA1, UP-CA2, UP-CA3 e as 3 componentes principais (UP-CA3-PCA) originárias do conjunto de atributos UP-CA3; e (b) as abordagens FCM, *HACC-Spatial* e *SWMU Clustering* e as 3 componentes principais UP-CA3-PCA.



Por meio da Figura 8.23 (b), observa-se que a abordagem FCM conseguiu resultados muito próximos aos obtidos pela abordagem *SWMU Clustering* quando foram utilizadas as componentes principais. Nesse caso, apesar da abordagem FCM ainda ter obtido alguns resultados com certo nível de estratificação, como no mapa exibido na Figura 8.22 (a), a redução de dimensionalidade considerando o espaço de coordenadas proporcionou a obtenção de resultados melhores com relação aos que foram alcançados com o conjunto de atributos UP-CA1. No entanto, a seleção de atributos por meio da correlação de Pearson (UP-CA2), bem como a utilização de todos os atributos (UP-CA3), proporcionou resultados melhores do que a redução de dimensionalidade, conforme mostra o gráfico da Figura 8.23 (a). Além disso, para as abordagens *HACC-Spatial* e *SWMU Clustering*, a utilização dessa técnica também não proporcionou ganhos significativos nos índices alcançados pelo critério *SD-Spatial*, em comparação com os resultados obtidos por essas abordagens com os outros conjuntos de atributos da UP-CA. Para a abordagem *HACC-Spatial*, a obtenção de uma CGD muito pequena com relação às outras, para o mapa contendo 5 CGDs, proporcionou um desempenho muito inferior com relação às outras abordagens. Em geral, a abordagem *SWMU Clustering* obteve melhores resultados do que a abordagem FCM, que nesse experimento foi utilizada da mesma maneira que na abordagem desenvolvida por Córdoba (2013) quando utilizou-se de componentes principais.

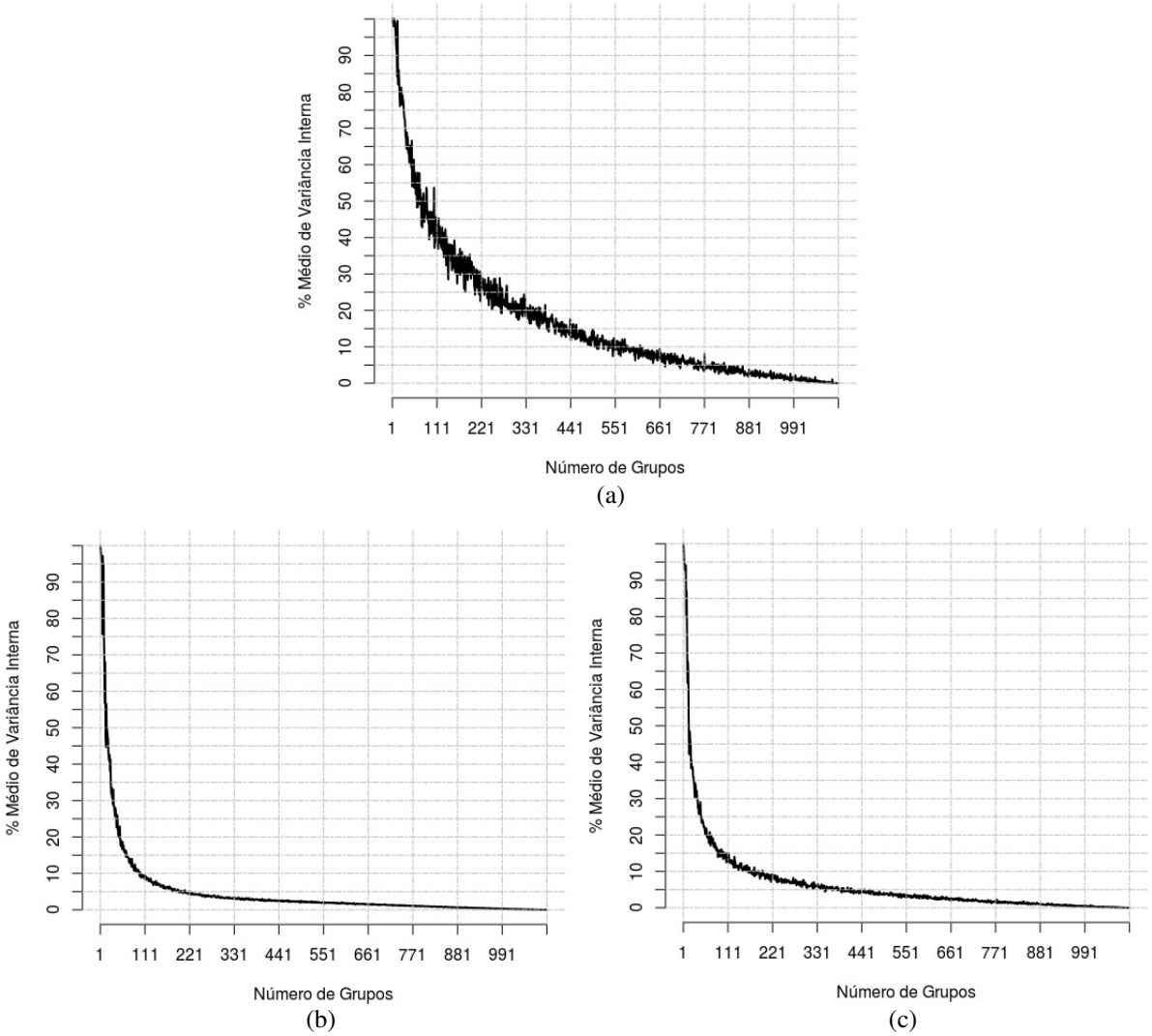
Com isso, pode-se reforçar a ideia de que, apesar do tratamento diferenciado do espaço de atributos com relação ao espaço de coordenadas durante o pré-processamento ser importante,

como foi realizado pelas abordagens desenvolvidas por Córdoba (2013) e Peeters (2015), procedimento similar deve ser realizado também durante o agrupamento das amostras para que melhores resultados possam ser obtidos.

8.3.2 Experimentos Utilizando Atributos da UP-UV

Os experimentos a seguir retratam o desempenho obtido pelas abordagens FCM, *HACC-Spatial* e *SWMU Clustering* utilizando atributos oriundos da UP-UV. Da mesma maneira que foi realizado para a UP-CA, a Figura 8.24 exibe gráficos contendo os valores de MVI para possíveis divisões dos dados realizadas pelo algoritmo *k-means* para a tesselação inicial, considerando-se os conjuntos de atributos UP-UV1, UP-UV2 e UP-UV3.

Figura 8.24: Percentual médio de variância interna (MVI) para dados oriundos da UP-UV, considerando os conjuntos de atributos (a) UP-UV1; (b) UP-UV2; e (c) UP-UV3.



A partir dos gráficos da Figura 8.24, é possível verificar que o valor de MVI aumenta mais rapidamente a partir de 10%, para o conjunto de atributos UP-UV1; e a partir de 5% para os conjunto de atributos UP-UV2 e UP-UV3. Desse modo, esses valores foram definidos para o parâmetro *varTess* utilizado pela abordagem *SWMU Clustering* nos experimentos em que cada um dos conjuntos de atributos da UP-UV foi utilizado. Para os outros parâmetros, foram fixados os valores mais utilizados nos experimentos com dados da UP-CA, com *minDMC* igual a 2, *maxDMC* igual a 5 e *porcDMU* igual a 5%. Além disso, não foram obtidos mapas de declividade para a UP-UV e, portanto, a restrição de obstáculos espaciais não foi utilizada em nenhum experimento.

8.3.2.1 Conjunto de Atributos UP-UV1

Inicialmente, foram realizados experimentos utilizando o conjunto de atributos UP-UV1, com o intuito de gerar mapas de UGDs exclusivamente a partir de parâmetros do solo. Para manter a compatibilidade com a tesselação inicial utilizada na abordagem *SWMU Clustering*, com *varTess* igual a 10%, foi definido o valor de 656 para o parâmetro *k* da abordagem *HACC-Spatial*. A Figura 8.25 exibe mapas contendo 4 CGDs obtidos pelas abordagens FCM, *HACC-Spatial* e *SWMU Clustering* para esse experimento.

Figura 8.25: Mapas contendo 4 CGDs obtidos a partir da utilização das abordagens (a) FCM; (b) *HACC-Spatial*; e (c) *SWMU Clustering*.

(a) (b) (c)

Por meio da Figura 8.25, é possível observar resultados compatíveis fornecidos pela abordagem *SWMU Clustering* com relação à abordagem FCM, sendo que a utilização de restrições espaciais impostas pela nova abordagem fizeram com que fosse obtido uma mapa de UGDs menos estratificado. A abordagem *HACC-Spatial* apresentou um resultado bem diferente e com CGDs e UGDs desbalanceadas com relação ao tamanho da área. Nesse caso, a parametrização utilizada fez com que a restrição espacial de contiguidade fosse desabilitada já nos

primeiros passos da construção do dendrograma, fazendo com que surgissem UGDs muito pequenas e que poderiam ser consideradas como estratos. Com relação à validação qualitativa, o resultado extremamente estratificado obtido pela abordagem *HACC-Spatial* foi retratado pelo critério *SD-Spatial*, onde verificou-se uma compatibilidade maior entre os resultados obtidos pelas abordagens FCM e *SWMU Clustering*, conforme exibido nos mapas da Figura 8.25.

8.3.2.2 Conjunto de Atributos UP-UV2

Em seguida, foram realizados experimentos utilizando-se o conjunto de atributos UP-UV2, com o intuito de se obter mapas de UGDs baseados apenas em atributos relacionados à cultura. A Figura 8.25 exibe novamente mapas contendo 4 CGDs obtidos pelas abordagens FCM, *HACC-Spatial* e *SWMU Clustering*.

Figura 8.26: Mapas contendo 4 CGDs obtidos pelas abordagens (a) FCM; (b) *HACC-Spatial*; e (c) *SWMU Clustering*.

(a) (b) (c)

Nesse caso, ao contrário do que foi observado no experimento anterior, a Figura 8.26 exibe mapas de UGDs muito parecidos para as três abordagens, sendo que em nenhum dos casos verifica-se um nível de estratificação capaz de prejudicar a análise do usuário final. Entretanto, pode-se observar transições mais suaves entre as UGDs no mapa obtido pela abordagem FCM, fazendo com que esta apresentasse melhores índices para o critério *SD-Spatial*, considerando mapas contendo entre 2 e 5 CGDs.

8.3.2.3 Conjunto de Atributos UP-UV3

Finalizando os experimentos realizados com a UP-UV, foram gerados mapas de CGDs utilizando o conjunto de atributos UP-UV3, selecionados pela correlação de Pearson e compondo informações relacionadas tanto ao solo quanto à cultura. A Figura 8.27 exibe mapas contendo

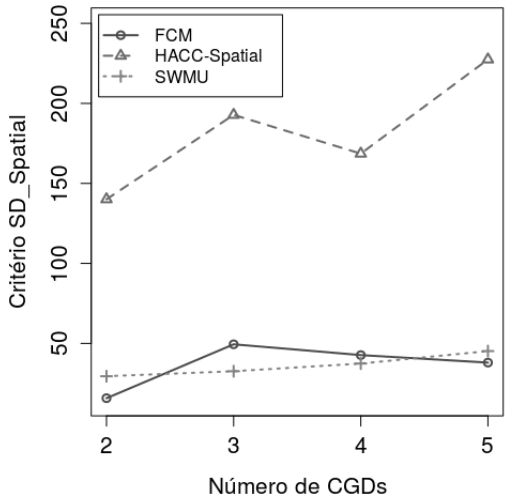
3 CGDs obtidos pelas abordagens FCM, *HACC-Spatial* e *SWMU Clustering*.

Figura 8.27: Mapas contendo 3 CGDs obtidos pelas abordagens (a) FCM; (b) *HACC-Spatial*; e (c) *SWMU Clustering*.

(a) (b) (c)

A partir da Figura 8.27, é possível observar que a inclusão de atributos de solo proporcionou aos resultados obtidos pelo *HACC-Spatial* os mesmos problemas já relatados no experimento utilizando o conjunto de atributos UP-UV1. Da mesma forma, as soluções obtidas pelas abordagens FCM e *SWMU Clustering* são bastante compatíveis, sendo que a segunda privilegia a formação de UGDs maiores e espacialmente mais contíguas por conta das restrições espaciais impostas. A Figura 8.28 exibe um gráfico contendo os índices alcançados pelas três abordagens considerando o critério *SD-Spatial*.

Figura 8.28: Índices obtidos pelo critério *SD-Spatial* para as abordagens de agrupamento FCM, *HACC-Spatial* e *SWMU Clustering*.



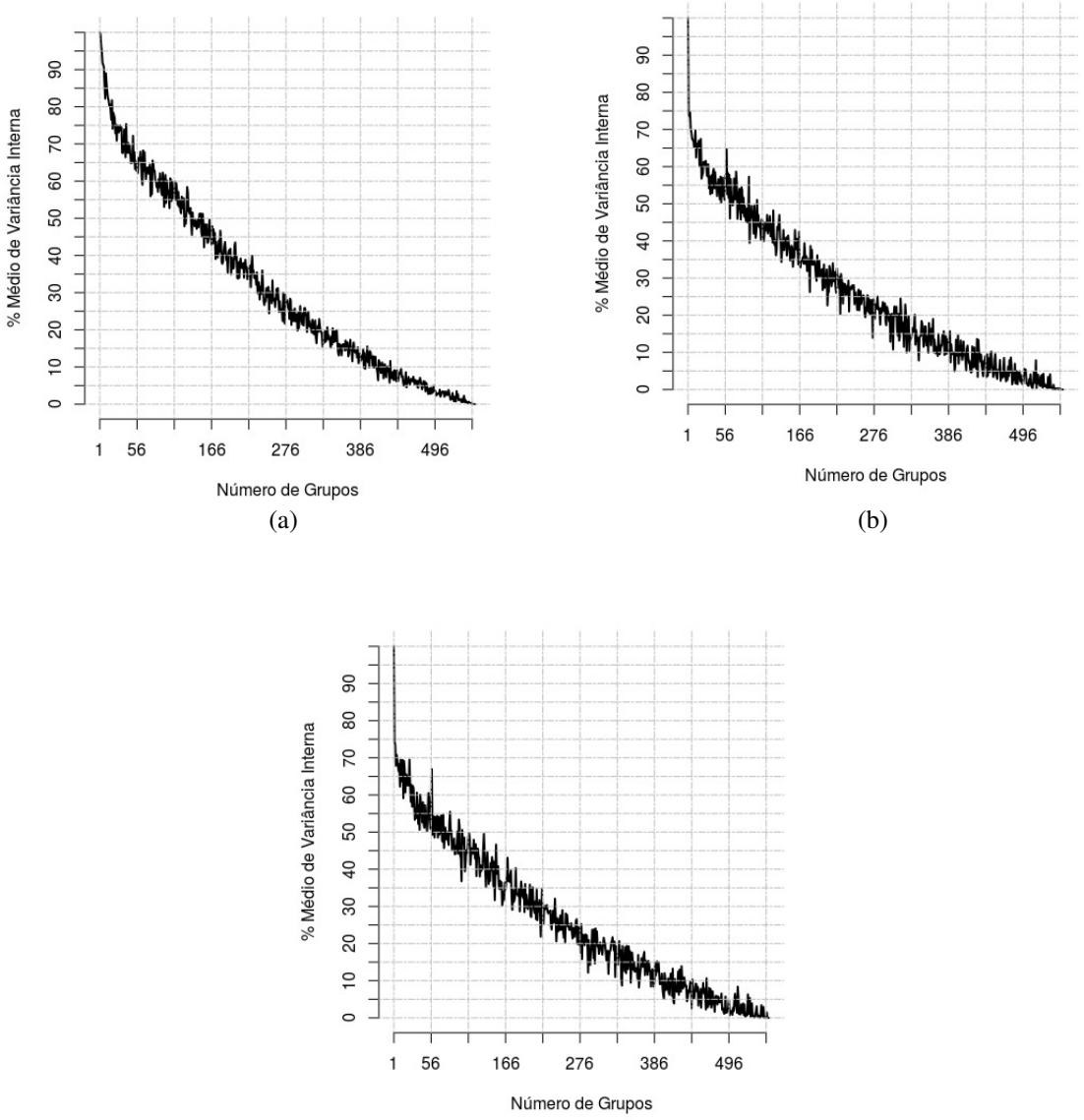
A partir da Figura 8.28, verifica-se que os resultados fornecidos pelo *SD-Spatial* são capazes de retratar de maneira fiel a coesão e separação obtida pelas abordagens tanto no espaço de

atributos quanto no espaço de coordenadas, onde os melhores resultados entre 2 e 5 CGDs variam entre as abordagens FCM e *SWMU Clustering*.

8.3.3 Experimentos Utilizando Atributos da UP-G

Os experimentos a seguir retratam o desempenho obtido pelas abordagens FCM, *HACC-Spatial* e *SWMU Clustering* utilizando dados da UP-G. Da mesma maneira que foi realizado para a UP-CA e a UP-UV, a Figura 8.29 exibe gráficos contendo os valores de MVI para possíveis divisões dos dados realizadas pelo algoritmo *k-means* para a tesselação inicial, considerando-se os conjuntos de atributos UP-G1, UP-G2 e UP-G3.

Figura 8.29: Percentual médio de variância interna (MVI) para dados oriundos da UP-G, considerando os conjuntos de atributos (a) UP-G1; (b) UP-G2; e (c) UP-G3.



No caso dos dados da UP-G é possível observar, a partir dos gráficos da Figura 8.29, um crescimento contínuo da porcentagem de MVI dos agrupamentos, conforme a quantidade de grupos diminui. Entretanto, em muitos pontos dos gráficos, verificam-se muitos picos de aumento e redução desse valor, que proporcionam maiores dificuldades ao usuário final em definir qual seria o valor ideal para o parâmetro *varTess* em todos os conjuntos de atributos. Desse modo, definiu-se um valor fixo de 25% para esse parâmetro em todos os casos, proporcionando uma tesselação inicial contendo uma quantidade de grupos próxima da metade da quantidade de amostras. Além disso, também foram fixados os valores dos parâmetros *porcDMU* para 5%,

minDMC para 2 e *maxDMC* para 5, conforme já realizado em experimentos anteriores.

8.3.3.1 Conjunto de Atributos UP-G1

O primeiro experimento utilizou o conjunto de atributos UP-G1, contendo apenas informações referentes ao solo. A Figura 8.30 exibe os mapas contendo 4 CGDs gerados para esse experimento pelas abordagens FCM, *HACC-Spatial* e *SWMU Clustering* sem a utilização da restrição de obstáculos espaciais.

Figura 8.30: Mapas contendo 4 CGDs obtidos a partir das abordagens (a) FCM; (b) *HACC-Spatial*; e (c) *SWMU Clustering* sem a utilização da restrição de obstáculos espaciais.

(a) (b) (c)

Com os parâmetros determinados nesse experimento, a abordagem *SWMU Clustering* conseguiu gerar apenas mapas contendo entre 2 e 4 CGDs, o que já retrata a dificuldade encontrada na definição da tesselação inicial e também durante a obtenção dos mapas pelas abordagens. O resultado obtido pela abordagem FCM exibe uma estratificação acentuada, onde é possível a identificação apenas de uma UGD na parte superior do mapa. O resultado obtido pela abordagem *HACC-Spatial* exibe uma tentativa de redução da estratificação, onde uma UGD representando unicamente uma CGD é também caracterizada na parte superior do mapa. Entretanto, as outras CGDs continuam bastante estratificadas, gerando inclusive buracos em uma CGD maior. Já a abordagem *SWMU Clustering* procurou gerar um mapa mais contínuo, preservando a contiguidade espacial das CGDs o quanto possível, de forma a entregar ao usuário final um mapa de UGDs passível de análise e utilização. No que diz respeito aos índices alcançados pelo critério *SD-Spatial*, a abordagem *SWMU Clustering* obteve resultados muito superiores com relação às outras abordagens, mostrando a eficiência na utilização dos seus parâmetros para restrições espaciais.

8.3.3.2 Conjunto de Atributos UP-G2

O experimento seguinte foi realizado com a utilização do conjunto de atributos UP-G2, contendo apenas dados de produtividade histórica. Devido ao fato desse conjunto de atributos possuir dados de apenas dois anos de produtividade e para culturas diferentes (milho e soja), foram encontradas muitas dificuldades por parte das abordagens na obtenção de mapas de CGDs consistentes. A abordagem *SWMU Clustering* conseguiu obter, a partir dos parâmetros determinados, apenas mapas contendo 2 e 3 CGDs, enquanto que as abordagens *HACC-Spatial* e FCM, assim como no experimento anterior, obtiveram mapas muito estratificados e que dificilmente seriam utilizados na prática pelo usuário final, conforme mostra a Figura 8.31.

Figura 8.31: Mapas contendo 3 CGDs obtidos a partir das abordagens (a) FCM; (b) *HACC-Spatial*; e (c) *SWMU Clustering* sem a utilização da restrição de obstáculos espaciais.

(a) (b) (c)

Apesar da necessidade de fundir diversas UGDs consideradas como estratos a partir da restrição que impõe um tamanho mínimo para a área de uma UGD, fazendo com que a separação dos grupos no espaço de atributos fosse bastante prejudicada, a abordagem *SWMU Clustering* alcançou, nesse experimento, índices melhores para o critério *SD-Spatial* do que as abordagens FCM e *HACC-Spatial*.

8.3.3.3 Conjunto de Atributos UP-G3

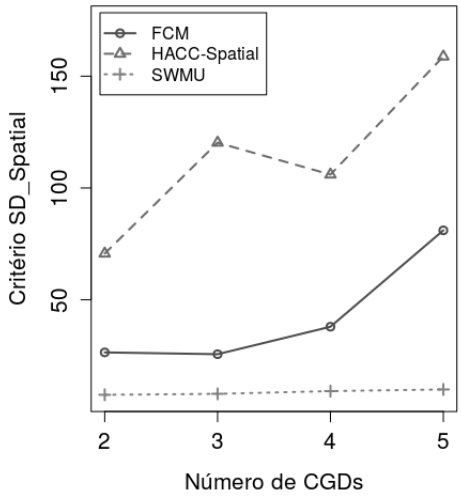
Finalizando, um último experimento foi realizado com a utilização do conjunto de atributos UP-G3, selecionados a partir da alta correlação existente entre eles. Para esse experimento, o mapa de declividade da área foi utilizado como obstáculo pela abordagem *SWMU Clustering*, na tentativa de reduzir a dependência do parâmetro *porcDMU* no tratamento da estratificação ao final da construção do dendrograma. Como resultado, foi possível a obtenção de mapas contendo 5 CGDs utilizando todas as abordagens, exibidos na Figura 8.32.

Figura 8.32: Mapas contendo 5 CGDs obtidos pelas abordagens (a) FCM; (b) *HACC-Spatial*; e (c) *SWMU Clustering* considerando a restrição de obstáculos espaciais.

(a) (b) (c)

Por meio da Figura 8.32, é possível verificar novamente a dificuldade da abordagem *HACC-Spatial* em obter mapas de CGDs que podem ser utilizados na prática, principalmente por conta da atribuição dos valores para os seus parâmetros. Nesse caso, uma tesselação inicial contendo uma quantidade de grupos em torno da metade da quantidade de amostras, aliado ao desligamento da restrição imposta pelo parâmetro cp logo nos primeiros passos da construção do dendrograma, fez com que fossem identificadas UGDs e CGDs muito pequenas e que tornaram-se buracos considerados como estratos em uma CGD maior que ocupa quase toda área de estudo. Com relação à abordagem FCM, a utilização de parâmetros bem correlacionados proporcionou uma diminuição considerável na estratificação do mapa final em comparação com os experimentos utilizando os conjuntos de atributos UP-G1 e UP-G2, permitindo uma identificação mais clara das CGDs. Já o resultado obtido pela abordagem *SWMU Clustering* mostra a eficácia do uso dos obstáculos espaciais, fazendo com que o parâmetro *porcDMU* influenciasse menos no resultado final e possibilitasse a obtenção de um mapa contendo 5 CGDs. A Figura 8.33 exibe os índices alcançados pelo critério *SD-Spatial* para as abordagens FCM, *HACC-Spatial* e *SWMU Clustering* na obtenção de mapas contendo de 2 a 5 CGDs.

Figura 8.33: Índices alcançados pelo critério *SD-Spatial* para mapas contendo entre 2 e 5 CGDs obtidos pelas abordagens FCM, *HACC-Spatial* e *SWMU Clustering*.



O gráfico da Figura 8.33 mostra que a estratificação identificada nos resultados da abordagem FCM proporciona um efeito negativo nos índices do critério *SD-Spatial*, onde os resultados obtidos pela abordagem *SWMU Clustering* foram sempre melhores. Com relação à abordagem *HACC-Spatial*, a obtenção de muitas UGDs consideradas como estratos proporcionou uma diferença significativa nos índices obtidos, em comparação com as outras abordagens.

8.3.4 Consolidação dos Resultados

Os experimentos realizados considerando diferentes abordagens e conjuntos de atributos foram consolidados, de forma a resumir os resultados obtidos. A Tabela 8.1 resume os resultados obtidos pelas abordagens FCM, *HACC-Spatial* e *SWMU Clustering* nos experimentos onde foi possível comparar o desempenho das mesmas, considerando o critério de validação interna *SD-Spatial*.

Tabela 8.1: Índices médios alcançados pelas abordagens FCM, *HACC-Spatial* e *SWMU Clustering* considerando diferentes conjuntos de atributos e o critério de validação interna *SD-Spatial*; e porcentagem de variação dos índices obtidos pelas abordagens hierárquicas com relação à abordagem FCM.

Atributos	VarTess / k	FCM	HACC	%Var.HACC	SWMU	%Var.SWMU
UP-CA1	0,3 / 207	32,39	28,57	-14,31	8,41	-73,06
UP-CA1	1,0 / 88	32,39	21,37	-35,11	10,37	-67,20
UP-CA1	2,5 / 32	32,39	14,79	-54,40	12,75	-60,20
UP-CA2	5,0 / 100	8,37	35,33	317,66	7,82	-5,84
UP-CA3	5,0 / 100	10,96	28,18	160,05	7,81	-28,11
UP-CA3-PCA	5,0 / 104	13,81	34,09	150,24	8,99	-32,67
UP-UV1	10,0 / 656	56,94	178,33	215,64	48,26	-13,04
UP-UV2	5,0 / 208	8,05	11,33	42,37	14,92	85,99
UP-UV3	5,0 / 460	36,47	182,23	468,23	36,17	14,94
UP-G1	25,0 / 318	426,23	131,25	-58,16	99,31	-67,65
UP-G2	25,0 / 307	386,47	215,96	-17,91	110,26	-60,36
UP-G3	25,0 / 307	42,78	113,93	202,71	8,63	-76,08
Média		**90,60**	**82,95**	**114,75**	**31,14**	**-31,94**

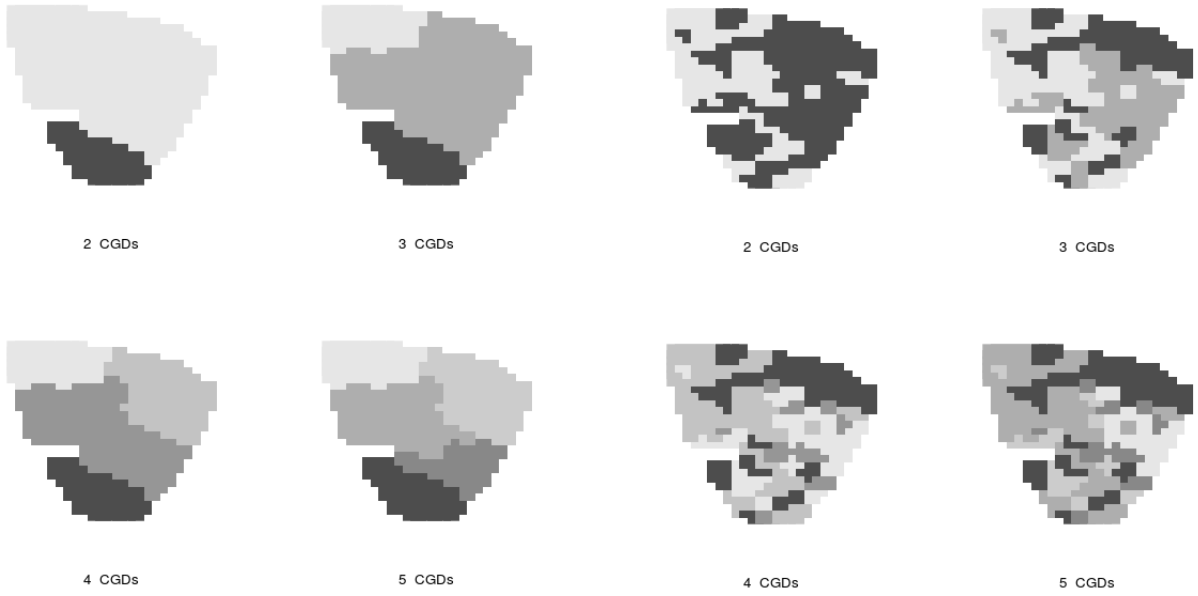
A partir dos resultados apresentados na Tabela 8.1 e levando em consideração que valores menores para o critério *SD-Spatial* representam melhores agrupamentos, é possível verificar que, apesar de apresentar valores médios inferiores aos obtidos pela abordagem FCM, a abordagem *HACC-Spatial* obteve um percentual de variação com relação a ela extremamente alto (114,75%). Contribuíram bastante para que esse valor fosse atingido alguns experimentos onde a abordagem *HACC-Spatial* acabou gerando mapas com CGDs bastante estratificadas e desbalanceadas entre elas com relação ao tamanho de sua área. Já com relação à abordagem *SWMU Clustering*, as restrições espaciais impostas permitiram a obtenção de agrupamentos com ganhos de coesão e separação, considerando tanto o espaço de atributos quanto no espaço de coordenadas, de 5,84% até 76,08%, com ganho médio de 31,94%, com relação aos agrupamentos obtidos pela abordagem FCM. Desse modo, pode-se concluir, a partir dos experimentos comparativos realizados, que a abordagem *SWMU Clustering* é capaz de fornecer agrupamentos mais coesos e bem separados considerando tanto o espaço de atributos quanto o espaço de coordenadas, em comparação aos agrupamentos fornecidos pelas abordagens FCM e *HACC-Spatial*, proporcionando ao usuário final mapas de UGDs mais fáceis de interpretar e que podem ser úteis na prática.

8.4 Teste Estatístico de Significância

Os critérios de validação interna permitem avaliar de forma qualitativa diferentes particionamentos de um conjunto de dados a partir da utilização de diferentes abordagens de agrupamento, conforme realizado nos experimentos descritos até aqui. Entretanto, para que essa avaliação seja mais precisa, é necessário verificar se a abordagem utilizada produz agrupamentos que constituem particionamentos reais dos dados, ou se os mesmos são obtidos de maneira aleatória. No contexto do delineamento de UGDs em AP, essa questão está focada em verificar se os mapas de UGDs gerados por determinada abordagem de agrupamento constituem uma divisão natural da área de cultivo com relação às características do solo e da cultura, ou se essa divisão foi gerada de maneira aleatória e capaz de ser obtida a partir de um conjunto de dados sintético. Ao analisarmos a Tabela 8.1, é possível verificar que o conjunto de atributos UP-CA3 proporcionou os melhores resultados obtidos pela abordagem *SWMU Clustering* considerando o critério *SD-Spatial*. Desse modo, esse conjunto de atributos foi selecionado para a realização de um teste estatístico de significância.

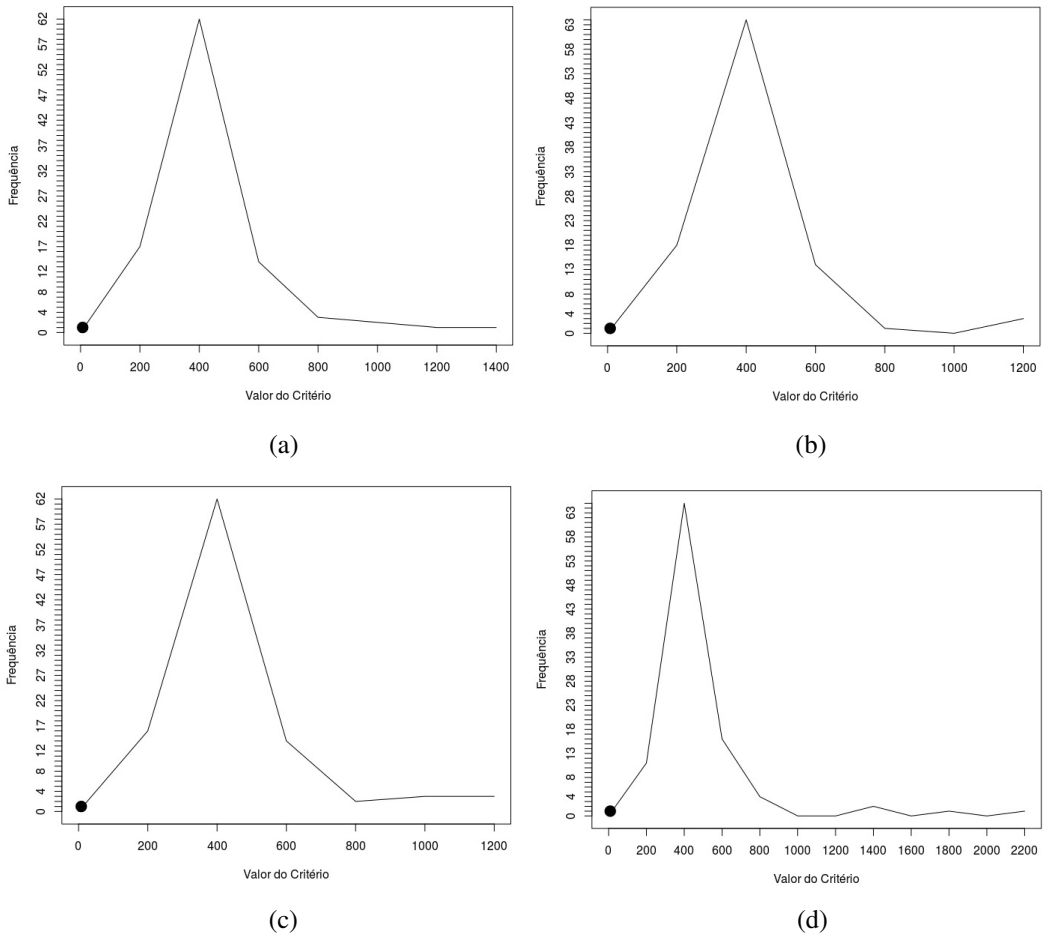
Para esse teste foi utilizada a hipótese nula (H_0) de posições aleatórias equiprováveis com nível de significância α de 5%. Além dos dados originais que compõem o conjunto de atributos UP-CA3, também foram gerados outros 99 conjuntos de dados sintéticos contendo amostras com as mesmas coordenadas geográficas do conjunto de dados original, mas com valores obtidos de maneira aleatória para as 20 dimensões pertencentes ao espaço de atributos, considerando a média e a variância do conjunto de dados original. Com isso, foi considerado um conjunto de 100 soluções de agrupamento $S = (S_1, S_2, ..., S_{100})$, onde S_1 possui mapas contendo de 2 a 5 CGDs obtidos a partir do conjunto de dados original; e as outras 99 soluções possuem mapas de 2 a 5 CGDs obtidos a partir dos conjuntos de dados sintéticos. Complementarmente, para as 100 soluções, a restrição de área mínima para uma UGD, determinada pelo parâmetro *porcDMU*, não foi utilizada; e a tesselação inicial utilizada para as 99 soluções sintéticas foi a mesma da solução original. Com isso, o teste de significância realizado permitiu a verificação da não aleatoriedade da abordagem *SWMU Clustering* apenas considerando o seu fluxo principal de construção do dendrograma. A Figura 8.34 exibe mapas contendo entre 2 e 5 CGDs geradas a partir do conjunto de atributos UP-CA3 original, e uma das soluções obtidas utilizando dados sintéticos.

Figura 8.34: Mapas contendo de 2 a 5 CGDs obtidos a partir da abordagem *SWMU Clustering* utilizando (a) o conjunto de atributos UP-CA3 (S_1); e (b) um conjunto de dados sintéticos obtidos aleatoriamente a partir da média e variância do conjunto de atributos UP-CA3.



Por meio da Figura 8.34, verifica-se que os mapas obtidos em S_1 permitem uma fácil identificação visual de divisão das CGDs, compostas em todos os mapas por uma única UGD. Entretanto, os mapas de UGDs gerados aleatoriamente a partir de dados sintéticos não apresentam uma subdivisão clara, tornando-se extremamente estratificados. Consequentemente, o resultado obtido em S_1 está entre os 5% melhores, considerando o critério *SD-Spatial* para todas as quantidades de CGDs, conforme exibido nos histogramas da Figura 8.35.

Figura 8.35: Histogramas contendo a frequência de índices do critério *SD-Spatial* obtidos pelo conjunto S de soluções de agrupamento gerados para o teste de significância, considerando (a) 2 CGDs; (b) 3 CGDs; (c) 4 CGDs; e (d) 5 CGDs.



Nos histogramas da Figura 8.35, os valores de *SD-Spatial* alcançados por $S1$ estão identificados em destaque por um ponto. Em valores absolutos, as soluções de agrupamento geradas em S_1 para 2, 3, 4 e 5 CGDs alcançaram, respectivamente, os índices de 7,12; 7,52; 7,88; e 8,72, enquanto que os índices médios para os mesmos agrupamentos, considerando as soluções obtidas a partir de dados sintéticos, foram de 341,48; 328,11; 348,45; e 372,55. Por terem obtido índices entre os 5% menores fornecidos pelo *SD-Spatial*, bem abaixo dos valores frequentemente obtidos pelas soluções utilizando dados sintéticos, os agrupamentos gerados pela abordagem *SWMU Clustering* em S_1 podem ser considerados, segundo esse critério, como subdivisões naturais do conjunto de dados, a um nível de significância de 5%.

8.5 Análise Preliminar de Complexidade Computacional

O desenvolvimento de abordagens de agrupamento para a tarefa de delineamento de UGDs em AP possui como objetivo principal fornecer ao usuário final mapas de UGDs eficazes, fáceis de interpretar e úteis para a tomada de decisão a respeito de operações agrícolas sítio-específicas. Desse modo, o desenvolvimento da abordagem *SWMU Clustering*, no âmbito desta tese, priorizou a criação de uma solução capaz de fornecer mapas de UGDs representados por agrupamentos altamente coesos e bem separados, levando-se em consideração tanto o espaço de atributos quanto o espaço de coordenadas. Entretanto, mesmo que de maneira secundária, questões relacionadas com a eficiência e a complexidade computacional foram levadas em consideração no processo de desenvolvimento. Desse modo, uma primeira análise da complexidade computacional da abordagem *SWMU Clustering* foi realizada e está descrita a seguir.

Se considerarmos as abordagens utilizadas pelo estado da arte, a complexidade computacional da abordagem *SWMU Clustering* é bastante próxima da *HACC-Spatial*, por ambas se tratarem de abordagens de agrupamento hierárquico aglomerativo. Em geral, para essas abordagens, os melhores casos proporcionam complexidade computacional da $\mathcal{O}(n^2)$, e os piores casos da $\mathcal{O}(n^3)$, com n sendo a quantidade total de amostras agrupadas. A inclusão das restrições espaciais na abordagem *SWMU Clustering* pode proporcionar a redução ou o aumento dessa complexidade, dependendo dos parâmetros fornecidos pelo usuário.

Levando-se em consideração a utilização da tesselação inicial, principalmente quando o parâmetro *varTess* indica uma subdivisão em uma quantidade de grupos muito menor do que a quantidade de amostras, a complexidade computacional pode ser bastante reduzida, por conta da utilização do algoritmo *k-means*, com complexidade da $\mathcal{O}(n)$, em substituição a diversos passos de construção do dendrograma. Considerando o melhor caso, a tesselação inicial poderia proporcionar a redução da complexidade da abordagem *SWMU Clustering* para $\mathcal{O}((k-1)^2) + (n-k)$), com k sendo a quantidade de grupos delimitada pela tesselação inicial, sempre menor que a quantidade de amostras n. Com isso, sendo utilizada de maneira ponderada, a tesselação inicial pode auxiliar tanto na redução da estratificação dos mapas de UGDs, quando na redução da complexidade computacional da abordagem *SWMU Clustering*.

Por outro lado, a busca pela amostra mais próxima do centroide espacial, apesar de também auxiliar na redução da estratificação dos mapas de UGDs, gera um custo extra da $\mathcal{O}(N_k)$ em cada passo do dendrograma da abordagem *SWMU Clustering*, com N_k sendo a quantidade de amostras pertencentes a um grupo k. Um custo extra também é gerado quando é incluída a restrição espacial relacionada aos obstáculos, onde deve ser verificada a existência ou não dos

obstáculos entre duas amostras e, em caso positivo, realizado o cálculo de distância cumulativa, conforme descrito no Algoritmo 1. Finalmente, a verificação de UGDs consideradas como estratos envolve operações de geoprocessamento relacionadas à verificação de tamanho de área e união de polígonos, que geralmente possuem complexidade computacional elevada. Porém, essas operações são executadas em um único passo do dendrograma, quando a quantidade de grupos atingida é igual ao valor fornecido para o parâmetro *maxDMC*, não proporcionando grande impacto no tempo final de execução da abordagem *SWMU Clustering*.

Se tomarmos como exemplo mapas de UGDs obtidos com a utilização do conjunto de atributos UP-CA2, que contém 415 amostras e 13 atributos, em experimentos realizados utilizando um computador pessoal com processador modelo Intel(R) Core (TM) i5-3337U, de 1.8 GHz, 6 GB de memória principal (RAM), sistema operacional Ubuntu de 64 bits, versão 16.04, e algoritmos de agrupamento implementados no ambiente *R*, a abordagem FCM consumiu, em média, 0,4 segundos para fornecer ao usuário final mapas contendo entre 2 e 5 CGDs. Considerando essas mesmas amostras, atributos e configurações de *hardware* e *software*, além de todas as restrições espaciais possíveis de serem utilizadas, a abordagem *SWMU Clustering* consumiu, em média, 85 segundos para fornecer ao usuário final esses mesmos resultados. Apesar dessa grande diferença no tempo de execução, o que pode-se concluir, a partir dessa análise preliminar, é que as restrições espaciais impostas pela abordagem *SWMU Clustering* não são capazes de prejudicar a sua complexidade computacional a ponto de impedir que os mapas de UGDs solicitados pelo usuário não sejam gerados em um período de tempo de execução aceitável. Entretanto, por se tratar de uma abordagem hierárquica aglomerativa, um conjunto de dados de entrada com quantidade elevada de amostras pode proporcionar um aumento considerável desse período de tempo, sugerindo modificações em sua forma de execução que podem considerar, por exemplo, conceitos de processamento paralelo e distribuído.

8.6 Considerações Finais

Neste capítulo foi apresentada, de maneira detalhada, a abordagem de agrupamento espacial *SWMU Clustering*, composta em sua essência por uma restrição espacial relacionada à identificação do centroide imposta à tradicional abordagem de agrupamento hierárquica de Ward; e por três outras restrições espaciais opcionalmente impostas a partir de parâmetros definidos pelo usuário final: a tesselação inicial considerando o espaço de coordenadas e a variabilidade no espaço de atributos; a utilização de obstáculos espaciais no cálculo da dissimilaridade entre amostras; e o tamanho mínimo desejado para uma UGD. Diferentes configurações de parâmetros e conjuntos de atributos de entrada foram utilizados em experimentos que permitiram com-

parar a abordagem *SWMU Clustering* com outras abordagens que constituem o estado da arte no delineamento de UGDs em AP. Os resultados obtidos mostraram que a abordagem *SWMU Clustering* é capaz de proporcionar a obtenção de mapas de UGDs válidos, menos estratificados e mais fáceis de interpretar por parte do usuário final.

Capítulo 9

SWMU Polygon: Uma Abordagem Para Delineamento de UGDs Poligonais

9.1 Considerações Iniciais

Neste capítulo é apresentada a abordagem complementar *SWMU Polygon*, que permite representar os mapas de UGDs em formato poligonal. O capítulo está organizado da seguinte forma:

- A Seção 9.2 descreve as vantagens de se utilizar mapas de UGDs com representações poligonais ao invés de mapas de UGDs representados por pontos.
- A Seção 9.3 descreve a abordagem complementar *SWMU Polygon*, desenvolvida para possibilitar a representação dos mapas de UGDs gerados a partir da abordagem *SWMU Clustering* em formato poligonal.
- A Seção 9.4 descreve experimentos realizados com dados reais que mostram as vantagens para o armazenamento e a recuperação de mapas de UGDs em formato poligonal em bancos de dados espaciais.
- A Seção 9.5 finaliza o capítulo com as considerações finais.

9.2 Representação de Mapas de UGDs em Formato Poligonal

A representação de mapas de UGDs em formato poligonal tem o objetivo principal de reduzir a quantidade de dados espaciais a serem armazenados, pois em geral é utilizada uma quantidade menor de coordenadas para representar uma área por meio de polígonos do que para

representar essa mesma área por meio de pontos, principalmente para os conjuntos de dados com altas resoluções espaciais coletados em AP. A Figura 9.1 mostra um exemplo de uma mesma UGD representada por pontos e por polígonos. Nesse caso, a representação por pontos pode ser considerada como uma forma de representação matricial, onde cada ponto representa a coordenada central de um pixel.

Figura 9.1: Exemplo de UGD representada por (a) geometrias de pontos; e (b) por um único polígono.

(a) (b)

Considerando as UGDs exibidas na Figura 9.1, foram necessárias 509 amostras contendo coordenadas de latitude e longitude para representar a UGD exibida no item (a); e apenas 136 amostras para representar a UGD exibida no item (b), a partir de uma simplificação utilizando-se uma forma geométrica poligonal composta por essas amostras. Em alguns casos, essa diferença pode diminuir e até mesmo fazer com que a representação por pontos utilize menos geometrias do que a representação poligonal, principalmente nos casos em que a UGD possui uma área muito pequena de forma a ser considerada como estrato. Entretanto, se considerarmos os resultados obtidos pela abordagem *SWMU Clustering*, que prioriza, por meio de suas restrições espaciais, a obtenção de mapas com UGDs de tamanho balanceado e pouco estratificadas, a redução do custo de armazenamento e recuperação convertendo esses mapas em formas geométricas poligonais será em geral similar ao que ocorre no exemplo da Figura 9.1.

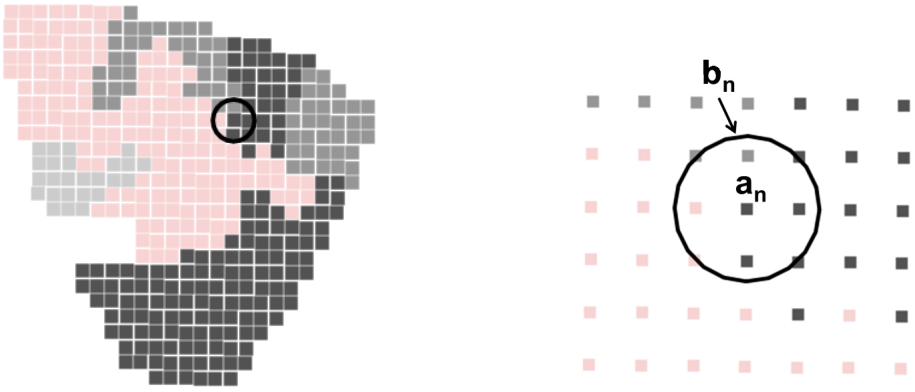
9.3 A Abordagem *SWMU Polygon*

Com o intuito de permitir a representação dos mapas de UGDs em formato de pontos delineados pela abordagem *SWMU Clustering* em formas poligonais mais sofisticadas, foi desenvolvida uma abordagem complementar, denominada *SWMU Polygon*. Os principais passos e algoritmos utilizados na execução dessa abordagem estão descritos a seguir.

Além do mapa de UGDs gerado pela abordagem *SWMU Clustering* e o respectivo conjunto de dados que proporcionou a geração desse mapa, com as amostras dispostas obrigatoriamente em uma grade espacial regular, a abordagem complementar *SWMU Polygon* considera como entrada o polígono *PC* representativo da área de estudo, para a realização de ajustes na forma das UGDs poligonais geradas ao final de sua execução. A ideia principal da abordagem *SWMU Polygon* é fazer com que as bordas das UGDs sejam constituídas pelas próprias amostras que geraram o mapa de UGDs obtido pela abordagem *SWMU Clustering*.

Primeiramente, com o intuito de se obter o seu conjunto de amostras vizinhas mais próximas, é gerado, para cada amostra a_n de um total de N amostras pertencente ao mapa original de UGDs, um *buffer* circular b_n que abrange um raio de tamanho $r \times \sqrt{2}$, onde r é a resolução espacial da grade regular em que o mapa foi delineado. A Figura 9.2 mostra um exemplo de *buffer* gerado seguindo esse procedimento.

Figura 9.2: Exemplo de *buffer* circular b_n com raio de tamanho $r \times \sqrt{2}$ gerado para uma amostra a_n, visualizado (a) em um mapa de UGDs como um todo; e (b) em uma escala aproximada.



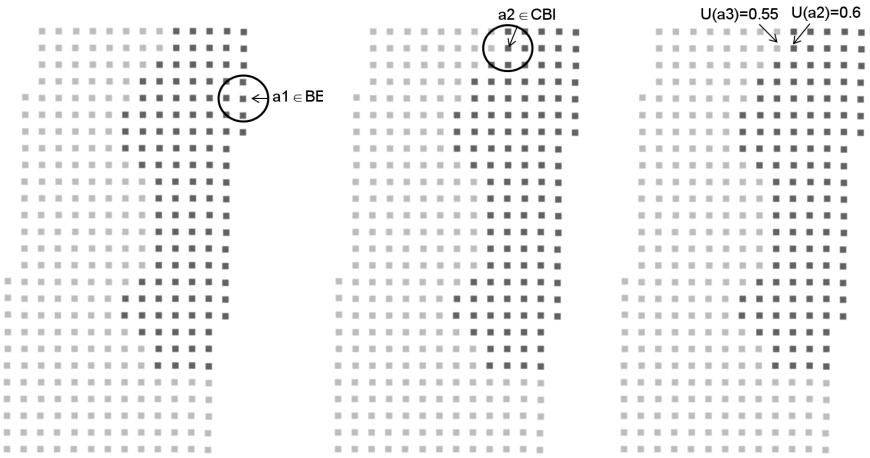
(a) (b)

A partir da Figura 9.2 (b), é possível identificar a presença de outras oito amostras que interceptam o *buffer* circular b_n e, portanto, são consideradas vizinhas da amostra a_n. Nesse exemplo, também é possível identificar que quatro dessas amostras vizinhas pertencem à mesma CGD da amostra a_n, pois são exibidas no mapa com o mesmo tom de cinza; e as outras quatro amostras vizinhas pertencem a outras duas CGDs. A partir dos N *buffers* gerados, são identificados dois tipos de amostras que poderão fazer parte das bordas dos polígonos das UGDs. Inicialmente, as amostras que possuem no máximo 5 amostras vizinhas espaciais, considerando o seu *buffer* circular, são marcadas como pertencentes à borda externa, ou seja, serão anexadas à borda do contorno da área de estudo, constituindo o conjunto *BE*. Em seguida, também considerando o *buffer* circular, são identificadas as amostras candidatas à borda interna. Nesse caso,

serão marcadas como candidatas as amostras que possuírem ao menos uma amostra vizinha interna a seu *buffer* b_n associada a uma CGD diferente da sua. Essas amostras serão identificadas pelo conjunto *CBI*.

A próxima etapa de execução da abordagem complementar *SWMU Polygon* considera uma matriz U de dimensões $N \times K$, contendo os graus de pertinência de cada amostra n do mapa com relação a cada uma das K CGDs, gerada a partir da execução da abordagem de agrupamento FCM em sua forma original, ou seja, considerando apenas o espaço de atributos do conjunto de dados utilizado para o delineamento das UGDs. Nessa etapa, para cada amostra do conjunto *CBI*, deve ser encontrada, ainda considerando o seu *buffer* circular, a amostra vizinha espacialmente mais próxima pertencente à uma CGD distinta da qual essa amostra foi associada pela abordagem *SWMU Clustering*, formando pares de amostras candidatas. Para cada um dos pares, é escolhida a amostra com menor grau de pertinência para a CGD à qual foi associada, segundo a matriz U, caracterizando a existência de uma dúvida maior sobre a associação dessa amostra com sua respectiva CGD. Em termos práticos, entende-se que essa amostra pertence menos à CGD a qual foi associada do que o seu par para a sua respectiva CGD, fazendo com que se torne mais apta a fazer parte de uma borda do que da própria CGD em si. As amostras selecionadas são agora identificadas pelo conjunto *BI*. A Figura 9.3 exibe graficamente um exemplo do processo executado para a seleção de amostras pertencentes aos conjuntos *BI* e *BE*.

Figura 9.3: Exemplo gráfico para definição dos pontos de borda pela abordagem *SWMU Polygon*, onde são selecionadas (a) uma amostra a_1 pertencente à borda externa; (b) uma amostra a_2 candidata à borda interna; e (c) um par de amostras (a_2 e a_3) candidatas à borda interna, onde a seleção final ocorre a partir do menor valor de pertinência da matriz U.



A partir da Figura 9.3, é possível verificar graficamente a sequência para seleção das amostras que devem ser consideradas como borda. Nos recortes de mapa exibidos nessa figura,

diferentes tons de cinza atribuídos às amostras indicam a associação à diferentes CGDs no agrupamento gerado pela abordagem *SWMU Clustering*, e os círculos indicam os *buffers* circulares das amostras citadas. Para efeitos desse exemplo, será rotulada como *G1* a CGD identificada pelo tom de cinza mais escuro; e como *G2* a CGD identificada pelo tom de cinza mais claro. No item (a), a amostra *a1* é selecionada para o conjunto *BE* porque possui apenas cinco amostras vizinhas espaciais, considerando o seu *buffer* b_1. Já no item (b), a amostra *a2* é selecionada para o conjunto *CBI* porque, além de possuir mais de cinco amostras vizinhas considerando o seu *buffer* b_2, duas delas pertencem à CGD *G2*, ou seja, diferente da CGD à qual a própria amostra foi associada (*G1*), considerando o agrupamento gerado pela abordagem *SWMU Clustering*. Finalmente, no item (c), a amostra *a3* é identificada como a vizinha mais próxima de *a2* e pertencente a outra CGD. Considerando o par de amostras *(a2,a3)*, a amostra *a3* deve ser selecionada para compor *BI*, uma vez que essa amostra possui, segundo a matriz *U* gerada a partir da execução da abordagem FCM, um grau de pertinência para a CGD *G2* menor do que o grau de pertinência da amostra *a2* para a CGD *G1*. Uma vez que uma amostra é incluída no conjunto *BI*, esta já é dada como selecionada, e não pode mais ser considerada como vizinha em processo posterior de seleção de pares de amostras, como executado na Figura 9.3 (c). A partir da definição das amostras do conjunto *BI*, é gerado o conjunto de adjacências de cada uma delas, ou seja, são identificadas quais as UGDs que irão compartilhar essas amostras em suas bordas, com o auxílio de uma triangulação de Delaunay. Esse conjunto de adjacências é identificado como *AI*.

Utilizando-se dos conjuntos *BI* e *AI*, a etapa seguinte consiste na execução do algoritmo TSP para delineamento dos polígonos representantes das UGDs. A partir desse algoritmo, as amostras são ligadas por linhas, considerando a sua heurística de procurar sempre o menor caminho, até que a última amostra seja ligada à primeira para formar um polígono. Entretanto, como esses polígonos são gerados de maneira independente entre si, podem ocorrer intersecções entre as suas áreas, que são identificadas e tratadas pela abordagem *SWMU Polygon*. Esse tratamento é realizado da seguinte maneira: para cada intersecção entre duas ou mais UGDs, a geometria gerada será unida ao grupo de polígonos que representam as UGDs pertencentes à CGD que rotula a maioria das amostras internas à ela. Consequentemente, essa mesma geometria será subtraída das geometrias correspondentes às outras UGDs que proporcionaram a intersecção gerada. No caso de uma CGD possuir mais de uma UGD poligonal, a sua representação será realizada por meio de uma coleção de polígonos.

Finalizando, para cada amostra pertencente ao conjunto *BI*, é verificado o seu ponto mais próximo pertencente ao polígono *PC* que irá substituí-la, a fim de realizar ajustes no mapa de UGDs com relação ao contorno da área de estudo. O Algoritmo 4 descreve, em alto nível, os

principais passos de execução da abordagem *SWMU Polygon*.

Algoritmo 4: *SWMU Polygon*.

Entrada: V = conjunto de vetores de dados espaciais com n amostras nos espaços de atributos e coordenadas; PC = polígono do contorno da área; $VClust$ = mapa contendo o agrupamento gerado pela abordagem *SWMU Clustering*.

Saída: M = Mapa de UGDs em formato poligonal.

1 BE,CBI e $BI \leftarrow$ vetores de amostras;
2 $buffer \leftarrow$ vetor de buffers;
3 $U \leftarrow$ FCM (V,numCGD(VClust));
4 **início**
    // Geração de *buffer* e seleção de amostras candidatas à borda
5 **para** *cada amostra* $a_n \in VClust$ **faça**
6 $buffer_n$ = buffer(a_n, res($VClust$));
7 **se** $a_n \notin BE$ *e* $a_n \notin BI$ **então**
8 **se** $nVizinhos(a_n) \leq 5$ **então**
9 $BE \leftarrow BE \cup a_n$;
10 **senão**
11 **se** $\exists a_v \in viz(a_n) \mid CGD(a_v) \neq CGD(a_n)$ **então**
12 $CBI \leftarrow CBI \cup a_n$;
13 **fim se**
14 **fim se**
15 **fim se**
16 **fim para**
    // Seleção de amostras para a borda interna a partir do grau de pertinência.
17 **para** *cada amostra* $a_n \in CBI$ **faça**
18 $a_{vp} \leftarrow$ maisProximo (a_n,viz(a_n) | CGD(a_v) $\neq$ CGD(a_n));
19 **se** $U(a_{vp}) \geq U(a_n)$ **então**
20 **se** $a_n \notin BI$ **então**
21 $BI \leftarrow BI \cup a_n$;
22 **fim se**
23 **senão**
24 **se** $a_{vp} \notin BI$ **então**
25 $BI \leftarrow BI \cup a_{vp}$;
26 **fim se**
27 **fim se**
28 **fim para**
29 $AI \leftarrow delaunayTri(BI)$; // Encontra adjacências pela triangulação de Delaunay
30 $MParc \leftarrow TSP(BI,AI)$; // Gera os polígonos utilizando o algoritmo TSP
31 $MParc \leftarrow ajustaInter(MParc)$; // Ajusta intersecções dos polígonos
32 $M \leftarrow ajustaPolArea(MParc,PC)$; // Ajusta polígonos das UGDs com a borda externa
33 **retorna** M;
34 **fim**

A partir do Algoritmo 4 podem ser identificados, de maneira resumida, todos os passos de execução da abordagem *SWMU Polygon*. As linhas numeradas de 5 à 16 resumem as operações

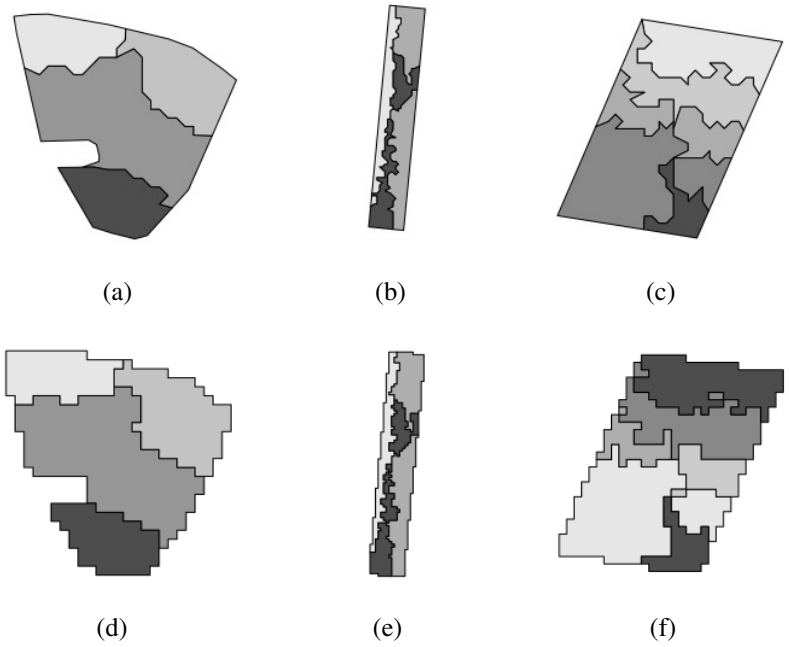
que verificam se uma amostra a_n deve ser identificada como borda externa ou candidata à borda interna. As linhas numeradas de 17 à 28 resumem as operações que definem as amostras que irão compor a borda interna, a partir do conjunto de amostras candidatas. Finalmente, as linhas de 29 à 32 resumem as operações realizadas para delineamento dos polígonos, como a triangulação de Delaunay (*delaunayTri*), algoritmo *TSP*, ajuste das intersecções (*ajustaInter*) e ajuste das UGDs à borda externa da área de estudo (*ajustaPolArea*).

9.4 Experimentos Comparativos

9.4.1 Abordagem *SWMU Polygon* e Algoritmo *Polygonize*

Com o intuito de verificar a eficácia da abordagem *SWMU Clustering* na obtenção de UGDs em formato poligonal, alguns resultados obtidos a partir dessa abordagem foram comparados com os resultados fornecidos pelo algoritmo *Polygonize*, disponível nas ferramentas de geo-processamento de diversos SIGs e também na biblioteca GDAL (*Geospacial Data Abstraction Library*) (GDAL, 2017). Esse algoritmo permite transformar regiões conectadas por pixels pertencentes a uma imagem raster e que compartilham um mesmo valor de pixel em polígonos vetoriais, onde cada polígono resultante recebe o valor atribuído a esses pixels. Devido ao fato de os mapas de UGDs gerados pela abordagem *SWMU Clustering* serem disponibilizados em pontos vetoriais, uma etapa de pré-processamento para transformar esses dados em pixels de uma imagem raster faz-se necessária antes da execução do algoritmo *Polygonize*. A Figura 9.4 exibe exemplos de mapas de UGDs poligonais obtidos para as UPs descritas na Seção 6.5 do Capítulo 6, considerando tanto a abordagem *SWMU Polygon*, quanto o algoritmo *Polygonize*. No caso do algoritmo *Polygonize*, foi sempre utilizada, para conversão dos pontos no formato vetorial para pixels no formato raster, a mesma resolução r da grade regular atribuída aos pontos no formato vetorial, fazendo com que cada pixel ocupe uma área de $r \times r$.

Figura 9.4: Mapas de UGDs poligonais gerados pela abordagem *SWMU Polygon* para as áreas (a) UP-CA, (b) UP-UV e (c) UP-G; e pelo algoritmo *Polygonize* para as áreas (d) UP-CA, (e) UP-UV e (f) UP-G.



Conforme pode ser visualizado na Figura 9.4, a abordagem *SWMU Polygon* gerou mapas poligonais de UGDs com bordas mais suaves do que os mapas gerados pelo algoritmo *polygonize*, onde a forma dos polígonos que representam as UGDs é guiada pela união de quadrados de resolução predefinida que representam os pixels. A Tabela 9.1 resume a quantidade de amostras original dos conjuntos de dados referentes à UP-CA, UP-UV e UP-G, bem como a quantidade de amostras ou geometrias de pontos utilizados pela abordagem *SWMU Polygon* e pelo algoritmo *Polygonize* para representação de mapas de UGDs poligonais da Figura 9.4.

Tabela 9.1: Quantidade de geometrias de pontos utilizadas pelas abordagens *SWMU Clustering*, *Polygonize* e *SWMU Polygon* para representação dos mapas de UGDs da Figura 9.4.

Mapa de UGDs	*SWMU Clustering*	*Polygonize*	*SWMU Polygon*
UP-CA-4CGDs	415	250	298
UP-UV-3CGDs	1101	718	878
UP-G-5CGDs	555	516	552

Conforme descrito na Tabela 9.1, tanto a abordagem *SWMU Polygon*, quanto o algoritmo *Polygonize*, proporcionaram uma diminuição na quantidade de pontos necessários para representar mapas de UGDs com relação à abordagem *SWMU Clustering*, indicando uma possível redução no custo de armazenamento dos mapas de UGDs a partir da utilização de ambas. Tam-

bém é necessário destacar que, devido ao formato simplificado dos polígonos proporcionado pela utilização do *Polygonize*, os mapas fornecidos por esse algoritmo tendem a utilizar menos amostras do que os fornecidos pela abordagem *SWMU Polygon*.

9.4.2 Desempenho de Bancos de Dados Espaciais na Recuperação de Mapas de UGDs Poligonais

Com o intuito de analisar o custo computacional para recuperação de mapas de UGDs poligonais gerados pela abordagem complementar *SWMU Polygon* com relação aos mapas de pontos gerados pela abordagem *SWMU Clustering*, foram realizados experimentos que simulam o desempenho na recuperação desses mapas a partir de operações disponíveis em um SIG conectado a um SGBDE. Para a realização dos experimentos, 97 mapas de UGDs contendo entre 2 e 5 CGDs, gerados a partir dos conjuntos de atributos oriundos das UPs descritas na Seção 6.5 do Capítulo 6 e da abordagem *SWMU Clustering*, foram armazenados em uma tabela única de um SGBDE relacional com extensão espacial PostGIS, denominada *unidade_gestao_diferenciada_ponto*. Cada CGD presente nos 97 mapas foi armazenada em um registro diferente nessa tabela, e representada por uma geometria do tipo *MULTIPOINT*, derivada do tipo abstrato de dados *Geometry* (HERRING, 2001), totalizando 335 registros. Da mesma maneira, esses 97 mapas foram transformados para o formato poligonal pela abordagem *SWMU Polygon*, e armazenados no mesmo SGBD, em uma tabela denominada *unidade_gestao_diferenciada_pol*. Nesse caso, cada CGD foi representada por uma geometria do tipo *MULTIPOLYGON*, também derivada do tipo de dados *Geometry*, fazendo com que fosse gerado também um total de 335 registros.

Com os dados armazenados, foi definido um experimento que simula o desempenho de um SIG em operações de visualização dos mapas oriundos de um SGBDE, utilizando quatro diferentes escalas de *zoom*: retângulo envolvente mínimo (REM) da área de estudo; e subretângulos contendo 50%, 25% e 12,5% da área do REM original, que identificaremos, respectivamente, como REM-2, REM-4 e REM-8. Para a simulação dessas operações, foram gerados *scripts* SQL contendo predicados capazes de recuperar tanto os dados armazenados na tabela *unidade_gestao_diferenciada_ponto* quanto na tabela *unidade_gestao_diferenciada_pol*, considerando as escalas de *zoom* REM, REM-2, REM-4 e REM-8. A ferramenta *pgbench*[1], disponível no SGBDE utilizado neste experimento, permite simular o seu desempenho a partir da execução de *scripts* SQL personalizados, determinando, por exemplo, a quantidade de usuários simultâneos e um período de tempo de execução de uma mesma operação. Como resultado, essa

[1] https://www.postgresql.org/docs/devel/static/pgbench.html

ferramenta fornece a quantidade de transações por segundo (TPS) possíveis de serem executadas pelo SGBDE, considerando os parâmetros fornecidos. Como o objetivo deste experimento foi simular as operações de um SIG, podemos considerar a visualização de cada mapa como uma transação, transformando a medida TPS em uma suposta quantidade de mapas por segundo (MPS) que podem ser fornecidos pelo SGBDE para visualização simultânea (VIASIG, 2017). Devido à todas as possibilidades de parametrização e simulação, permitindo verificar o desempenho do SGBDE em recuperar dados a partir de *scripts* SQL específicos, a ferramenta *pgbench* foi selecionada para executar os experimentos de desempenho especificados.

Para a realização dos experimentos de desempenho, as tabelas supracitadas foram armazenadas em um SGBDE PostgreSQL, versão 9.5, contendo a extensão espacial PostGIS, versão 2.2, instalados em um computador pessoal com processador modelo Intel(R) Core (TM) i5-3337U, de 1.8 GHz, 6 GB de memória principal (RAM), e sistema operacional Ubuntu de 64 bits, versão 16.04. Considerando os 97 mapas disponíveis em forma de pontos e os seus correspondentes em forma de polígonos, totalizando 194 mapas, além das quatro possíveis escalas de *zoom*, foram executadas 776 operações de recuperação de mapas, considerando para todas elas acessos simultâneos de 100 usuários. A Tabela 9.2 exibe um resultado consolidado da execução das operações, medido em MPS.

Tabela 9.2: Resumo do desempenho de simulações de operações de SIG com acessos ao SGBDE, medido em mapas por segundo (MPS), utilizando os mesmos mapas de UGDs em forma de pontos e de polígonos para diferentes escalas de *zoom*.

Escala de Zoom	MPS (Pontos)	MPS (Polígonos)	% Variação
REM	240,73	401,66	66,85
REM-2	241,72	283,77	17,39
REM-4	225,34	395,50	75,47
REM-8	211,93	391,72	84,83
Média	**229,93**	**368,14**	**61,14**

Na Tabela 9.2, a primeira coluna indica qual a escala de *zoom* utilizada nos experimentos; a segunda e terceira colunas indicam, respectivamente, os valores médios de MPS obtidos na operação de recuperação de mapas de CGDs formados por pontos (*SWMU Clustering*) e por polígonos (*SWMU Polygon*), nas quatro possíveis escalas de *zoom* definidas; e a quarta coluna indica o percentual de variação nos valores médios de MPS obtidos quando se utiliza os mapas formados por polígonos ao invés de pontos. Desse modo, experimentos com valores de MPS maiores indicam uma maior quantidade de mapas possível de ser exibida e, portanto, um melhor desempenho. Assim, é possível verificar, a partir da Tabela 9.2, que em relação

à recuperação dos dados armazenados em um SGBE para operações comumente encontradas em um SIG, o armazenamento utilizando mapas de UGDs em forma de polígonos apresentou ganhos de desempenho de 17,39% até 84,83%, com ganho médio de 61,14%, com relação ao armazenamento dos mesmos mapas em forma de pontos.

Alguns detalhes com relação ao aumento ou diminuição dos valores de MPS obtidos também devem ser destacados. A recuperação dos mapas utilizando a escala de *zoom* considerando o REM de cada mapa retorna sempre a totalidade dos dados do mapa, ou seja, todos os pontos ou polígonos são recuperados. Entretanto, mesmo obtendo-se uma quantidade maior de dados do que nas outras escalas, essa escala de *zoom* não exige a utilização de predicados espaciais de intersecção por parte do SGBDE, o que faz com que operações desse tipo sejam em geral menos custosas computacionalmente. Especificamente para a escala de *zoom* REM-2, é possível observar um ganho de desempenho menor dos mapas em forma de polígonos com relação aos mapas de pontos, se compararmos com os resultados obtidos nas outras escalas. Esse fato provavelmente indica que na escala REM-2 existe uma maior incidência de polígonos pequenos do que nas outras escalas, fazendo com que, em alguns casos, seja necessário mais pontos para representar uma UGD ou CGD em forma de polígonos do que em forma de pontos.

9.5 Considerações Finais

Neste capítulo foi apresentada, de maneira detalhada, a abordagem complementar *SWMU Polygon*, que permite representar os mapas de UGDs obtidos pela abordagem *SWMU Clustering* em formas poligonais sofisticadas. Os experimentos realizados mostraram as vantagens de se utilizar a abordagem *SWMU Polygon* ao invés do algoritmo *Polygonize* para este fim, bem como as vantagens da representação das UGDs em formato poligonal ao invés do formato de pontos em questões relacionadas ao armazenamento e recuperação dos mapas de UGDs em bancos de dados espaciais.

Capítulo 10

SISTEMA PROTÓTIPO E APLICAÇÃO PRÁTICA

10.1 Considerações Iniciais

Neste capítulo, são apresentados o sistema protótipo e uma aplicação prática de UGDs em campo agrícola desenvolvidos nesta tese. O capítulo está organizado da seguinte forma:

- A Seção 10.2 descreve o desenvolvimento de um sistema protótipo para auxiliar nas etapas de KDSD necessárias para o processo de delineamento de UGDs em AP.
- A Seção 10.3 descreve uma aplicação prática realizada em campo agrícola, para regulagem de doses de aplicação de insumos agrícolas a partir do delineamento de UGDs.
- A Seção 10.4 finaliza o capítulo com as considerações finais.

10.2 Desenvolvimento do Sistema Protótipo

Com o intuito de proporcionar a integração das ferramentas utilizadas para o desenvolvimento e realização de experimentos considerando o estado da arte no delineamento de UGDs em AP, bem como o critério de validação *SD-Spatial* e as abordagens de agrupamento desenvolvidas, foi proposta uma arquitetura computacional baseada em *software* livre e de distribuição gratuita. Essa arquitetura proporcionou o desenvolvimento de um sistema protótipo, utilizado para a obtenção dos resultados descritos nesta tese.

Levando-se em consideração o modelo de referência para o delineamento de UGDs em AP proposto por Santos e Saraiva (2015), o sistema protótipo proporciona ao usuário final a realização de diversas etapas do processo de KDSD que devem ser executadas para essa tarefa, tais como: selecionar os atributos a serem utilizados; realizar procedimentos de pré-processamento,

como filtragem e interpolação espacial; executar abordagens de agrupamento para sugerir mapas de UGDs; e avaliar os mapas obtidos por meio dos critérios de validação. Todos as funcionalidades utilizadas nessas etapas foram desenvolvidas e incluídas em uma biblioteca para o ambiente *R* (R, 2017), denominada *SWMUClustering*.

Com o intuito de proporcionar uma interface amigável e facilitar o acesso às ferramentas e aos dados, bem como a publicação de resultados finais, o SIG livre e de distribuição gratuita QuantumGIS foi utilizado como base para a arquitetura do sistema protótipo desenvolvido. Para tanto, os dados a serem utilizados devem ser armazenados no SGBDE PostgreSQL com extensão espacial PostGIS, e publicados em uma instância do servidor de mapas *GeoServer*, proporcionando um acesso simplificado e de alto nível à cada conjunto de dados espacial disponível. Desse modo, o usuário final pode acessar e utilizar cada conjunto de dados de AP em um formato de camadas tradicionalmente utilizado em SIGs onde, além dos dados espaciais, podem ser adicionados metadados que indicam a origem e a qualidade das informações que serão utilizadas. A partir da seleção dos dados, o usuário final poderá acessar as funcionalidades disponibilizadas pela biblioteca *SWMUClustering* por meio de um complemento presente no QuantumGIS que proporciona diversas facilidades, tais como: a criação de interfaces para parametrização; visualização dos resultados em forma de gráficos e em novas camadas geradas para o SIG; e a criação de modelos para execução em lote de diversas funcionalidades com diferentes entradas e saídas.

A arquitetura proposta e o sistema protótipo desenvolvido, bem como melhorias planejadas para novas versões, estão descritas com mais detalhes no artigo aceito para publicação, disponibilizado no Apêndice G desta tese.

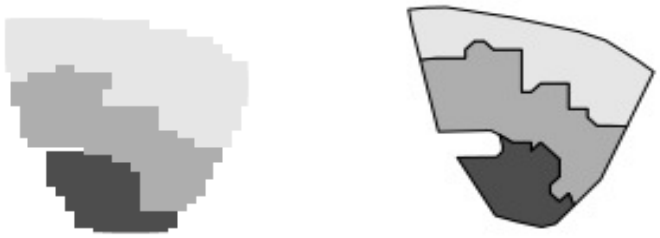
10.3 Aplicação Prática em Campo Agrícola

Com o intuito de verificar a eficácia dos mapas de UGDs gerados pelas abordagens desenvolvidas nesta tese, uma aplicação prática na área experimental da UP-CA foi proposta. Esse experimento teve como principal objetivo aplicar e validar essas abordagens utilizando dados reais obtidos de um campo agrícola, e trata-se de uma contribuição secundária desta tese. Uma reforma do talhão realizada em 2016, conforme mencionado na seção 6.5, proporcionou a elaboração de um projeto de aplicação prática com o intuito de regular doses de insumos agrícolas para intervenções futuras considerando mapas de UGDs. Por outro lado, a ausência de máquinas agrícolas automatizadas limitou a realização de experimentos mais elaborados.

10.3.1 Definição do Mapa de UGDs

Os dados reais da UP-CA foram utilizados em trabalhos preliminares contendo análises geoestatísticas que permitiram verificar a variabilidade espacial presente na área de cultivo considerando tanto as características do solo, por meio de medidas de condutividade elétrica do solo (GREGO, 2011b), quanto as características da cultura, por meio de medidas de produtividade histórica (GREGO, 2011a). Para essa aplicação, foi utilizado como base o mapa contendo 3 CGDs delineado pela abordagem *SWMU Clustering* e transformado em um mapa de polígonos pela abordagem *SWMU Polygon* a partir do conjunto de atributos UP-CA2 (Figura 10.1).

Figura 10.1: (a) Mapas contendo 3 CGDs utilizados como base na aplicação prática, obtidos a partir do conjunto de atributos UP-CA2 e a abordagem *SWMU Clustering*; e (b) transformado em mapa de polígonos pela abordagem *SWMU Polygon*.



(a) (b)

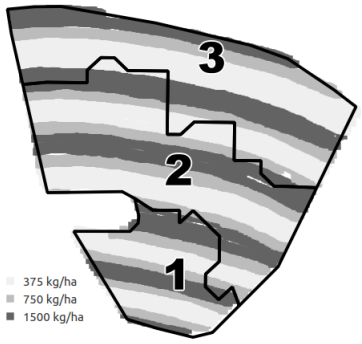
Os mapas da Figura 10.1 foram escolhidos para a aplicação prática porque, além de terem sido obtidos a partir de um conjunto de atributos com variáveis muito bem correlacionadas (UP-CA2), também foram utilizadas todas as restrições impostas pela abordagem *SWMU Clustering*, definidas pelos parâmetros *varTess*, *porcDMU*, *minDMC*, além da utilização de mapas de declividade como obstáculos. Além disso, devido à disponibilidade apenas de equipamentos com regulagem manual para aplicação em campo, ou seja, desprovidos de automação agrícola, a disposição espacial das CGDs nesse mapa favoreceu a aplicação das diferentes doses em faixas contínuas que atravessam a área toda, seguindo as demarcações prévias de curvas de nível.

10.3.2 Metodologia Para Aplicação à Taxa Variada

A aplicação considerando taxas variadas foi realizada com base em uma mistura que já vinha sendo utilizada na UP-CA, considerando um terço de calcário, um terço de gesso e um terço de cloreto de potássio, em uma aplicação uniforme de 1500 kg/ha. Como o objetivo, nesse momento, era não modificar demasiadamente as operações e tipos de insumos normalmente utilizados na área experimental, outras duas doses, contendo 750 kg/ha e 375 kg/ha dessa mesma

mistura, foram definidas para essa aplicação de regulagem. A partir do mapa de UGDs da Figura 10.1 (b), foram definidas então faixas que atravessam a área seguindo as curvas de nível, da esquerda para à direita, de forma que existissem, em cada CGD, regiões contendo a aplicação das três diferentes doses propostas para a regulagem. A Figura 10.2 apresenta novamente o mapa com as 3 CGDs, agora rotuladas (1, 2 e 3), além das amostras coletadas correspondentes às três doses de aplicação de insumos, identificadas em diferentes tons de cinza e conforme a legenda que acompanha o mapa.

Figura 10.2: Mapa de aplicação de mistura de insumos à taxa variada, considerando três diferentes doses para cada uma das CGDs representadas por faixas com três diferentes tons de cinza.

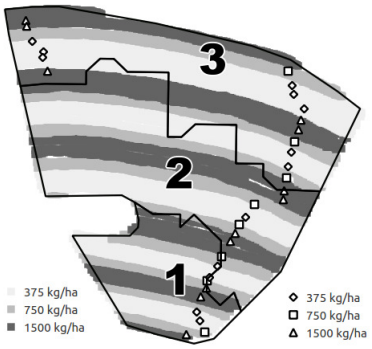


A aplicação de insumos à taxa variada, ilustrada na Figura 10.2, foi realizada nos dias 26 e 27 de setembro de 2016, cerca de 6 meses após o plantio da variedade de cana-de-açúcar CTC-20, que possui ciclo produtivo de um ano e meio. Os implementos agrícolas utilizados para esse trabalho foram uma calcareadeira modelo MasterFlow 8000, da Nogueira, de regulagem manual; e um trator Massey Ferguson, modelo 275, para conduzi-la. Por conta da utilização desse modelo de calcareadeira, a regulagem das doses foi realizada apenas durante as manobras do trator para mudança de uma faixa de aplicação para a outra. As doses foram alteradas a cada duas ou quatro faixas de aplicação, sendo que cada uma delas ocupava cerca de 5 linhas de plantio, considerando o alcance da calcareadeira. Assim, garantiu-se uma distribuição espacial similar no valor das doses para as 3 CGDs, fazendo com que fossem obtidas amostras suficientes para posterior avaliação com relação à produção. Devido às dificuldades em se acomodar equipamentos eletrônicos na cabina do trator utilizado, os dados foram coletados - coordenadas GPS e anotações das doses aplicadas - utilizando-se um *smartphone* Motorola, modelo Razr D1, e a versão gratuita do aplicativo PDFMaps, para o sistema operacional Android, suficientes para a realização do experimento.

10.3.3 Avaliação dos Resultados Considerando Dados de Produtividade

Com as diferentes doses de insumos aplicadas nas 3 CGDs presentes no mapa de UGDs inicialmente delineado, o objetivo principal desta aplicação prática foi avaliar, a partir de um mapa da colheita prevista para o início de julho de 2017, para qual dose seria obtida uma melhor resposta da cultura, ou seja, um maior retorno com relação à produtividade, em cada uma das CGDs delineadas. Com as doses reguladas, a aplicação de insumos para o próximo ciclo produtivo já poderia ser realizada com valores fixos para cada uma das 3 CGDs. Os procedimentos prévios para a realização do mapa de colheita, com pesagens do produto cana-de-açúcar em uma grade contendo 54 pontos igualmente distribuídos nas áreas correspondentes à cada dose de aplicação foram previamente planejados. Porém, problemas operacionais fizeram com que a colheita tivesse que ser antecipada para o dia 20 de junho de 2017 e, consequentemente, as amostras tiveram que ser coletadas em apenas alguns pontos onde foi possível realizar a pesagem da cana-de-açúcar, ou seja, de maneira bem mais restrita do que havia sido planejado. A Figura 10.3 exibe novamente o mapa de CGDs contendo as faixas de aplicação, e a inclusão desses pontos onde foi possível realizar a medição da produtividade da cana-de-açúcar.

Figura 10.3: Mapa de aplicação de mistura de insumos à taxa variada, considerando três diferentes doses para cada uma das CGDs representadas por faixas com três diferentes tons de cinza; e pontos onde foram obtidos dados de produtividade.



Apesar de não representar uma grade ideal para a geração de um mapa de produtividade, as amostras coletadas, representadas por losangos, quadrados e retângulos presentes na Figura 10.3, permitiram uma análise preliminar de dados médios de produtividade para tentativa de regulagem da dose para cada uma das 3 CGDs. Assim, foi possível verificar que, em algumas regiões do talhão, a quantidade de insumos normalmente aplicada (1500 kg/ha) pode ser reduzida, sem que a produtividade da área seja prejudicada. Para uma análise mais elaborada, é desejável a utilização de índices como o NVDI, obtidos a partir de imagens de sensoriamento com alta resolução espacial, com o intuito proporcionar ao usuário final uma estimativa melhor sobre a regulagem de dose de aplicação de insumos para cada uma das CGDs.

Apesar das dificuldades encontradas, a aplicação prática em campo agrícola foi útil para verificar o auxílio que os mapas de UGDs gerados pelas abordagens *SWMU Clustering* e *SWMU Polygon* podem proporcionar para um manejo agrícola que vise, inicialmente, a economia de insumos. Além disso, foi possível verificar que esse tipo de aplicação pode ser realizado mesmo com equipamentos mais simples, fazendo com que o produtor se sinta motivado em investir em novas tecnologias que possam de fato auxiliar no aumento da lucratividade de sua produção.

10.4 Considerações Finais

Neste capítulo foi apresentado, de maneira resumida, o sistema protótipo desenvolvido a partir de ferramentas livres e de distribuição gratuita, utilizado para a realização dos experimentos no âmbito desta tese. Além disso, foi apresentada uma aplicação prática realizada em campo agrícola, que possibilitou a verificação do auxílio que pode ser proporcionado por mapas de UGDs obtidos a partir das abordagens desenvolvidas nesta tese em operações de gerenciamento agrícola.

Capítulo 11

CONCLUSÕES E TRABALHOS FUTUROS

11.1 Considerações Iniciais

Neste capítulo, são apresentadas as conclusões e sugestões de trabalhos futuros para continuidade da pesquisa científica desenvolvida nesta tese. O capítulo está organizado da seguinte forma:

- A Seção 11.2 resume as principais contribuições obtidas no desenvolvimento desta tese a partir da investigação das hipóteses propostas.
- A Seção 11.3 conclui o trabalho desenvolvido, destacando novamente e de maneira resumida as contribuições obtidas e suas limitações.
- A Seção 11.4 sugere pesquisas e trabalhos futuros que podem ser realizados a partir das contribuições obtidas nesta tese.
- A Seção 11.5 finaliza o capítulo com as considerações finais.

11.2 Contribuições

Os avanços científicos obtidos com o desenvolvimento desta tese envolvem a sua aplicação em uma área de caráter multidisciplinar, que é o caso da AP. Em resumo, as suas principais contribuições são o desenvolvimento do critério *SD-Spatial* e das abordagens *SWMU Clustering* e *SWMU Polygon*, permitindo a investigação das hipóteses propostas, descrita a seguir.

A partir dos experimentos realizados no Capítulo 8, foi possível verificar que o espaço de coordenadas presente nos conjuntos de dados obtidos em AP deve ser tratado de maneira diferenciada com relação ao espaço de atributos pelas abordagens de agrupamento utilizadas na

tarefa de delineamento de UGDs. Como o objetivo final dessa tarefa é sempre a geração de um mapa, as amostras espaciais do conjunto de dados devem ser agrupadas de forma a proporcionar ao usuário final resultados de fácil interpretação e passíveis de serem utilizados na prática na lavoura. Entretanto, utilizar o espaço de coordenadas apenas para exibir agrupamentos que materializam as UGDs em forma de mapa pode proporcionar resultados não desejáveis ao usuário final, com excessiva estratificação que muitas vezes é causada pelo viés das abordagens de agrupamento tradicionais, como a FCM, em obter grupos extremamente coesos e bem separados considerando apenas o espaço de atributos.

Devido à sua complexidade e pelo fato de não possuírem relação de ordem total, os dados espaciais, neste caso representados pelas variáveis do espaço de coordenadas, não devem ser comparados entre si pelos mesmos operadores utilizados pelas abordagens de agrupamento tradicionais para verificar a dissimilaridade entre atributos convencionais. No âmbito desta tese, onde as amostras que representam os conjuntos de atributos estão distribuídas em uma grade espacial regular, o fato das variáveis do espaço de coordenadas serem simplesmente consideradas como atributos convencionais pode influenciar a formação de UGDs com formas exclusivamente convexas, tornando o mapa obtido muito parecido com um diagrama de Voronoi e podendo, em muitos casos, inibir características do solo e da cultura que poderiam ser determinantes para a separação ou união de duas ou mais regiões. Essa constatação foi obtida a partir da análise dos resultados do artigo aceito para publicação transcrito no Apêndice D, onde verificou-se que, mesmo que proporcione a obtenção de mapas de UGDs menos estratificados, o fato de considerarmos as coordenadas geográficas como atributos convencionais pode prejudicar a coesão e a separação dos agrupamentos obtidos no espaço de atributos. Para que esses prejuízos não ocorram de maneira significativa, a ponto de se obter mapas de UGDs totalmente contrários às características do solo e da cultura da área em questão, as coordenadas geográficas devem ser utilizadas apenas de maneira restritiva pelas abordagens de agrupamento. Assim, as restrições espaciais impostas pela abordagem *SWMU Clustering*, tais como: a tesselação inicial; a utilização do centroide e obstáculos espaciais; e a fusão de UGDs e CGDs consideradas como estratos, estão adequadas à verdadeira função restritiva do espaço de coordenadas nas abordagens de agrupamento. Mesmo assim, ainda são observadas perdas no espaço de atributos com relação à coesão e separação, porém com variações muito menores do que os ganhos obtidos no espaço de coordenadas, que puderam ser verificados a partir dos resultados obtidos pelo critério *SD-Spatial*. Sendo assim, o desenvolvimento e os resultados obtidos pelo critério *SD-Spatial* e pela abordagem *SWMU Clustering*, considerando de maneira equilibrada e diferenciada a coesão e a separação tanto do espaço de atributos quanto do espaço de coordenadas, contribuíram para que a hipótese $\mathbf{H}_1$ investigada nesta tese pudesse ser comprovada.

O estudo da parametrização das abordagens tradicionais de agrupamento e o desenvolvimento de uma nova abordagem com parâmetros não empíricos, determinados levando-se em consideração os próprios conjuntos de dados a serem agrupados, também geraram contribuições importantes para esta tese. Com relação à abordagem FCM, ficou constado que o parâmetro de *fuzzificação m* pouco contribui para melhorar a coesão e a separação dos agrupamentos no espaço de atributos, ou para reduzir a estratificação dos mapas de UGDs gerados em diversos cenários. Entretanto, esse parâmetro cumpre bem o seu papel de realizar apenas ajustes no resultado final.

Já com relação à abordagem *HACC-Spatial*, os parâmetros k e cp, por serem determinados de maneira bastante empírica, podem proporcionar a obtenção de resultados não desejados por parte do usuário final. Utilizar um valor de k muito alto, fazendo com que a tesselação inicial seja realizada considerando uma quantidade muito grande de grupos, aliado a um valor de cp que faça com que a restrição espacial de fundir apenas grupos adjacentes seja desabilitada logo no início da construção do dendrograma, pode fazer com que surjam estratificações indesejadas no mapa de UGDs, dificultando a análise e interpretação por parte do usuário final. Por outro lado, um valor de k muito pequeno, proporcionando uma tesselação inicial com poucos grupos, pode forçar a obtenção de UGDs com formas exclusivamente convexas e que proporcionam resultados parecidos com os que são obtidos quando as coordenadas geográficas são tratadas como atributos convencionais. Esse efeito também pode ser observado quando é escolhido um valor para o parâmetro cp de tal forma que a restrição espacial não é desabilitada durante a construção do dendrograma, fazendo com que sejam fundidos apenas grupos espacialmente adjacentes. Ainda com relação ao parâmetro cp, apesar de este ser um limiar totalmente relacionado ao comportamento dos dados, a sua determinação não é trivial, uma vez que o usuário final precisa visualizar uma razão de distâncias ótima entre grupos adjacentes e não adjacentes. Em resumo, os parâmetros utilizados pela abordagem *HACC-Spatial*, além de complexos de se determinar, também proporcionam muitas variações nos mapas de UGDs gerados, conforme verificado em experimentos do Capítulo 8.

Além das dificuldades encontradas na determinação do valor dos parâmetros, a abordagem *HACC-Spatial* utiliza a forma convencional de inicialização aleatória dos centroides do algoritmo *k-means* na etapa de tesselação inicial. Apesar do *k-means* proporcionar, em geral, resultados idênticos ou muito similares quando executado mais de uma vez para o mesmo conjunto de dados e considerando os mesmos parâmetros, a inicialização aleatória dos centroides faz com que esse algoritmo seja classificado como não determinístico. Nesse caso, quando é utilizada uma quantidade maior de grupos, como normalmente acontece com a tesselação inicial, a possibilidade de se obter uma solução ótima local ao invés de global aumenta, fazendo com

que os efeitos do não determinismo do *k-means* sejam refletidos no resultado final fornecido pelo *HACC-Spatial*.

Considerando todas as dificuldades de parametrização relacionadas às abordagens do estado da arte, a abordagem *SWMU Clustering* foi desenvolvida considerando parâmetros determinados com o auxílio dos próprios conjuntos de dados, a fim de reduzir o empirismo em sua determinação. Além da distribuição inicial dos centroides, que possibilitam uma divisão mais equilibrada e determinística do espaço de coordenadas, o fato do parâmetro *varTess* considerar um limiar de variância média interna do espaço de atributos dos dados para dividir as amostras pelo espaço de coordenadas, faz com que a tesselação inicial priorize as dissimilaridades existentes entre os dados para que se chegue a uma quantidade de grupos satisfatória. Apesar do valor de *varTess* também ser determinado pelo usuário final, a variância interna média de um agrupamento é obtida a partir dos próprios dados, ou seja, diferente da determinação normalmente empírica de uma quantidade fixa de grupos. Os parâmetros *porcDMU* e *maxDMC* auxiliam o usuário a determinar um tamanho mínimo para uma UGD, com o intuito de não limitar certas intervenções que poderão ser realizadas em campo e dependem, por exemplo, da área ocupada pela plataforma do equipamento agrícola que será utilizado. Em resumo, o fato dos parâmetros da abordagem *SWMU Clustering* serem determinados de maneira não empírica, considerando as particularidades do conjunto de dados e restrições impostas pelo usuário final, faz com que os mesmos proporcionem apenas ajustes necessários aos mapas de UGDs gerados. Sendo assim, a verificação da eficácia da abordagem *SWMU Clustering* ao utilizar esse tipo de parâmetro contribuiu para que a hipótese $\mathbf{H}_2$ investigada nesta tese pudesse ser comprovada.

A utilização de obstáculos espaciais como forma de aumentar a dissimilaridade no espaço de atributos entre duas amostras também proporcionou a obtenção de mapas de UGDs mais coesos e bem separados, considerando tanto o espaço de atributos quanto o espaço de coordenadas, a partir dos índices obtidos pelo critério *SD-Spatial*. Os efeitos da utilização de obstáculos puderam ser notados com maior destaque para os resultados obtidos utilizando-se os conjuntos de dados da UP-G, onde a variabilidade do espaço de atributos é bastante grande, proporcionando dificuldades no delineamento das UGDs. A utilização dos mapas de declividade como obstáculos, nesse caso, proporcionou a obtenção de resultados melhores segundo o critérios *SD-Spatial*. Esses resultados contribuíram para que a hipótese $\mathbf{H}_3$ investigada nesta tese pudesse ser comprovada.

Finalmente, os experimentos realizados no Capítulo 9 serviram para comprovar os ganhos no custo computacional com relação ao armazenamento e recuperação de mapas de UGDs, quando estes são transformados em mapas poligonais pela abordagem *SWMU Polygon*. A si-

mulação de recuperação e visualização desses mapas a partir de um SIG, considerando diferentes escalas de *zoom*, mostrou que a quantidade de geometrias necessárias para armazenar UGDs representadas por polígonos, em geral menor do que para armazenar UGDs representadas por pontos, influencia diretamente no tempo de resposta, e, consequentemente, no custo computacional para recuperação desses mapas por parte do usuário final. Desse modo, os resultados desses experimentos contribuíram para que a hipótese $\mathbf{H}_4$ investigada nesta tese pudesse ser comprovada.

11.3 Conclusões

A gestão sítio-específica por meio de unidades de gestão diferenciada (UGDs) é uma das maneiras práticas de se adotar os conceitos de Agricultura de Precisão (AP), possibilitando o aumento da produtividade a partir de estratégias que evitam o desperdício de insumos e ao mesmo tempo contribuem para a redução dos impactos ao meio ambiente. Por ser muito dependente de atributos espaciais relacionados ao solo e à cultura, o delineamento de UGDs deve considerar a aplicação dos conceitos de descoberta de conhecimento em bancos de dados espaciais (KDSD) para que possa ser uma ferramenta útil para a tomada de decisão por parte do usuário final. Nesse contexto, diversas etapas, relacionadas à seleção de atributos, pré-processamento, mineração de dados, pós-processamento e utilização do conhecimento, todas elas contando com a participação do usuário final, devem ser executadas para que os atributos espaciais coletados em campo sejam transformados em mapas de UGDs. Entretanto, uma vez que a maioria das abordagens para delineamento de UGDs em AP não considera a complexidade do espaço de coordenadas na principal etapa do KDSD, relacionada à mineração de dados e extração de padrões, tampouco nas etapas subsequentes que tem o objetivo de avaliar e validar as soluções obtidas, o usuário final acaba, por diversas vezes, se deparando com mapas de UGDs extremamente estratificados e de difícil interpretação.

Em contrapartida, nesta tese, a complexidade do espaço de coordenadas foi considerada para essas etapas do KDSD aplicadas no delineamento de UGDs em AP, proporcionando a obtenção de três contribuições principais. Primeiramente, o desenvolvimento do critério *SD-Spatial* proporcionou a possibilidade de avaliação e validação de mapas considerando não só a coesão e a separação das UGDs no espaço de atributos, mas verificando também questões relacionadas à estratificação obtida quando esses mapas são exibidos em Sistemas de Informações Geográficas (SIGs), penalizando a formação de buracos e a ocorrência de estratos. A segunda contribuição está relacionada ao desenvolvimento da abordagem de agrupamento *SWMU Clustering*, onde as dissimilaridades entre os atributos espaciais são exploradas de ma-

neira ponderada, considerando tanto o espaço de atributos quanto o espaço de coordenadas, além de parâmetros fornecidos pelo usuário especialista de maneira não empírica e evitando o não determinismo das soluções. Com isso, são obtidos mapas de UGDs que privilegiam a contiguidade espacial, reduzindo os efeitos da estratificação e facilitando a interpretação por parte do usuário final. Finalmente, a abordagem complementar *SWMU Polygon* possibilita que os mapas de UGDs sejam representados em formato poligonal, visando melhorar a sua eficiência de armazenamento e recuperação em bancos de dados espaciais, e, consequentemente, de informações históricas de intervenções que podem ser realizadas em diversos ciclos produtivos.

A principal característica do critério *SD-Spatial* está relacionada aos fatores de ponderação incluídos no cálculo de coesão intragrupos já presente no critério SD. Além do cálculo da variância média realizado pelo seu precursor, o critério *SD-Spatial*, por meio de fatores de ponderação, penaliza mapas contendo UGDs muito pequenas e buracos a ponto de serem considerados como estratos, além de UGDs relativamente pequenas considerando a CGD à qual pertence, e CGDs relativamente pequenas considerando a área total, tudo isso levando em consideração o parâmetro *porcDMU* fornecido pelo usuário final. O espaço de atributos também é tratado no cálculo de separação entre as UGDs, tornando o *SD-Spatial* um critério que permite avaliar, de maneira ponderada, questões relacionadas ao espaço de atributos e de coordenadas. Com isso, possibilita ao usuário final selecionar mapas de UGDs fáceis de interpretar quando visualizados em um SIG, sem deixar de considerar as questões relacionadas aos atributos do solo e da cultura. Pelo fato de proporcionar avaliações mais equilibradas com relação aos mapas de UGDs, o tratamento diferenciado realizado pelo critério *SD-Spatial* para o espaço de coordenadas a partir de funcionalidades de geoprocessamento proporciona um aumento do custo computacional de sua execução, identificado como a principal limitação desse novo critério, já que os critérios existentes realizam apenas cálculos aritméticos que incluem operações simples como somatórias e médias.

Já a abordagem *SWMU Clustering* possui como principal característica a utilização de restrições espaciais em diferentes fases da construção do dendrograma. A tesselação inicial permite agrupar amostras espacialmente próximas a partir de um algoritmo de particionamento tradicional, considerando um parâmetro relacionado à variância dos atributos do solo e da cultura e a inicialização não aleatória dos centroides. Com isso, é obtido um agrupamento inicial determinístico, capaz de eliminar passos considerados desnecessários na construção do hierarquia de soluções, proporcionando uma redução do custo computacional sem prejuízos significativos para a coesão e separação no espaço de atributos. Com esse mesmo objetivo, a utilização da restrição do centroide espacial permite a verificação da dissimilaridade entre amostras considerando tanto o espaço de atributos quanto o espaço de coordenadas, assim como a utilização

de obstáculos espaciais. Por fim, a restrição para verificação de UGDs consideradas como estratos, embora prejudique a coesão e a separação do espaço de atributos, permite que sejam eliminadas áreas que podem contribuir para dificultar a interpretação por parte do usuário final. Em experimentos realizados, a abordagem *SWMU Clustering* apresentou ganhos de 5,84% até 76,08%, com ganho médio de 31,94%, na coesão e separação com relação à abordagem FCM, considerando tanto o espaço de atributos quanto o espaço de coordenadas.

Apesar do melhor desempenho apresentado com relação às abordagens tradicionais, a abordagem *SWMU Clustering* ainda possui algumas limitações que necessitam ser superadas. A primeira delas está relacionada ao custo computacional elevado quando comparada, por exemplo, com a abordagem FCM, por conta da utilização de uma abordagem de agrupamento hierárquica com algumas restrições espaciais que podem contribuir para aumentar ainda mais esse custo. Apesar dessa questão não ser a mais relevante para a área de aplicação desta tese, conforme já relatado na Seção 8.5 do Capítulo 8, o crescente aumento da resolução espacial e da variedade dos atributos relacionados ao solo e à cultura que vem sendo coletados, principalmente por meio de imagens aéreas, pode fazer com que a eficiência de execução de abordagens de agrupamento hierárquicas como a *SWMU Clustering* necessite ser melhorada para que possa continuar a ser utilizada. Outra limitação, relatada nos experimentos, está relacionada à algumas configurações de parâmetros utilizadas em determinados conjuntos de atributos e que, em alguns casos, impossibilitam a abordagem *SWMU Clustering* de fornecer ao usuário final todos os mapas de UGDs desejados. Assim, questões como a fusão de UGDs pequenas considerando apenas o espaço de coordenadas podem ser revistas. Ainda, com relação à utilização de *ensembles*, as abordagens tradicionais não se mostraram eficientes a ponto de proporcionarem a obtenção de agrupamentos mais robustos do que os fornecidos pelas abordagens individuais, mostrando ser necessária a sua adaptação para encontrar soluções de consenso considerando tanto o espaço de atributos quanto o espaço de coordenadas.

Finalizando, a transformação dos mapas de UGDs gerados pela abordagem *SWMU Clustering* em representações poligonais, a partir da abordagem complementar *SWMU Polygon*, é baseada na identificação de amostras que tendem a compor a borda dos polígonos que representam as UGDs. Essas amostras são identificadas a partir do grau de pertinência obtido pela abordagem FCM utilizando os atributos convencionais e a mesma quantidade de CGDs do mapa original. Complementarmente, é realizado o delineamento dos polígonos que compõem cada uma das UGDs a partir de heurísticas conhecidas, acompanhado de operações de intersecção, união e subtração para finalização do mapa. Em experimentos realizados, a abordagem complementar *SWMU Polygon* apresentou ganhos de desempenho de 17,39% até 84,83%, com ganho médio de 61,14%, com relação à abordagem *SWMU Clustering*, no que diz respeito à recupe-

ração de mapas de UGDs armazenados em bancos de dados por meio de operações comumente realizadas em SIGs. Com relação às limitações, em alguns casos a abordagem *SWMU Polygon* encontrou dificuldades com relação à obtenção de polígonos topologicamente válidos para as UGDs, ou seja, que respeitam o tipo de dado geométrico utilizado. Desse modo, alternativas à triangulação de Delaunay e ao algoritmo TSP, utilizados para delinear os polígonos e suas respectivas adjacências, devem ser investigadas para que sejam obtidos resultados ainda mais eficazes.

11.4 Trabalhos Futuros

A partir das contribuições supracitadas, pode-se concluir que as abordagens desenvolvidas nesta tese podem auxiliar efetivamente o usuário final de AP em obter mapas de UGDs menos estratificados e mais fáceis de interpretar, podendo proporcionar uma gestão eficaz dos insumos e corretivos que devem ser aplicados nos ciclos produtivos da lavoura. Entretanto, com o intuito de buscar novas contribuições originais para a aplicação de delineamento de UGDs em AP que possibilitem superar as limitações das abordagens desenvolvidas, esta tese também motivou a proposta de trabalhos futuros. Sendo assim, questões relacionadas à redução do custo computacional do critério *SD-Spatial* e da abordagem *SWMU Clustering*, parametrização da abordagem *SWMU Clustering* e erros de topologia encontrados nos polígonos obtidos pela abordagem *SWMU Polygon* devem ser consideradas em trabalhos futuros complementares à esta tese. Além disso, a evolução dos conceitos e ferramentas disponíveis para AP possibilitou o levantamento de novas possibilidades de pesquisa na área de Ciência da Computação aplicada à AP, descritas a seguir.

Os experimentos realizados no âmbito desta tese proporcionaram a obtenção e validação de mapas de UGDs a partir dos mais variados atributos relacionados ao solo e à cultura, em três diferentes tipos de cultura. Entretanto, com o crescente desenvolvimento e disponibilidade de novos sensores, capazes de medir novos atributos e gerar novos índices que podem indicar a variabilidade espacial de uma lavoura, novos experimentos com dados obtidos a partir dessas novas tecnologias, incluindo também outras culturas, devem ser realizados constantemente, com o intuito de consolidar as abordagens desenvolvidas na comunidade de usuários de AP. Atributos relacionados à qualidade podem auxiliar, por exemplo, na separação de uma área de cultivo em UGDs contendo frutos específicos para serem utilizados como matéria prima; e UGDs contendo frutos para serem consumidos como produto final.

O aumento da variedade de atributos que podem ser obtidos em campo também vêm acom-

panhado de outras questões. Considerando principalmente as altas resoluções espaciais encontradas em imagens obtidas a partir de sensoriamento remoto e ARPs, o volume de dados a ser analisado e, consequentemente, agrupado, tende a crescer consideravelmente. Além disso, a crescente utilização da internet como meio de comunicação para integrar diferentes bancos de dados espaciais, distribuídos em servidores localizados em pontos distintos, faz com que as questões relacionadas ao desempenho e a velocidade com que os dados trafegam na rede também seja cada vez mais estudada. Todas essas questões remetem à construção de novas abordagens para o processo de KDSD, considerando conceitos de *Big Data* e computação em nuvem. A partir desses novos conceitos, todas as abordagens utilizadas nas diferentes etapas do KDSD devem ser adaptadas, com o intuito de proporcionar ao usuário final, tomando como exemplo o delineamento de UGDs em AP, mapas que agregam cada vez mais o valor de todo o trabalho de análise realizado para a sua obtenção.

Outra questão que deve ser considerada é com relação à expectativa do usuário final a respeito da utilização das UGDs. Para que sejam obtidos mapas de UGDs menos estratificados a partir das abordagens desenvolvidas nesta tese, é necessária uma perda da coesão e separação no espaço de atributos que pode proporcionar, em alguns casos, o aumento do erro proporcionado pelas intervenções espacialmente diferenciadas a partir de UGDs. Desse modo, a utilização de critérios como o índice de oportunidade para adoção de AP (PRINGLE, 2003; OLIVEIRA, 2014), que quantifica a variabilidade espacial de uma área de cultivo em função da viabilidade operacional das tecnologias disponíveis para o gerenciamento sítio-específico, pode ser uma alternativa para avaliar se as diferenças obtidas entre mapas de UGDs menos estratificados com relação a mapas de UGDs com menor erro podem compensar a sua adoção para determinadas aplicações.

Finalizando, para que abordagens relacionadas à construção e validação de agrupamentos, como as que foram desenvolvidas nesta tese, sejam inseridas e utilizadas nas atividades diárias e corriqueiras de uma propriedade rural que utiliza ou pretende utilizar os conceitos de AP, é necessário que as mesmas sejam integradas em pesquisas relacionadas aos FMIS. Nesse caso, é necessária uma maior integração das abordagens desenvolvidas com os SIGs, proporcionando ao usuário final realizar alterações, por exemplo, nos limites das UGDs de acordo com sua conveniência, e verificar automaticamente qual o impacto econômico e ambiental que essa alteração pode causar.

11.5 Considerações Finais

Neste capítulo foram apresentas as contribuições científicas obtidas com o desenvolvimento desta tese, geradas a partir da comprovação das hipóteses propostas. Além disso, foram apresentadas as conclusões obtidas a partir dessas contribuições, as limitações encontradas e propostas de trabalhos futuros para continuidade da pesquisa científica.

REFERÊNCIAS

ABADI, D. J. Data management in the cloud: limitations and opportunities. *IEEE Bulletin of the Technical Committee on Data Engineering*, v. 32, n. 1, p. 5–12, 2009.

AGGELOPOOULOU, K. et al. Delineation of management zones in an apple orchard in Greece using a multivariate approach. *Computers and Electronics in Agriculture*, Elsevier B.V., v. 90, p. 119–130, 2013. ISSN 01681699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2012.09.009>.

AGRAWAL, R. et al. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 27, n. 2, p. 94–105, jun. 1998. ISSN 0163-5808. Disponível em: <http://dx.doi.org/10.1145/276305.276314>.

AKDAG, F.; EICK, C.; CHEN, G. Creating Polygon Models for Spatial Clusters. In: ANDREASEN, T. et al. (Ed.). *Foundations of Intelligent Systems*. Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8502). p. 493–499. ISBN 978-3-319-08325-4. Disponível em: <http://dx.doi.org/10.1007/978-3-319-08326-1_50>.

ALSABTI, K.; RANKA, S.; SINGH, V. An efficient k-means clustering algorithm. *Electrical engineering and Computer Science*, 1997.

ALTMAN, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, v. 46, n. 3, p. 175–185, 1992. Disponível em: <http://dx.doi.org/10.1080/00031305.1992.10475879>.

ANANDHI, R. J.; SUBRAMANYAM, N. Efficient fusion of cluster ensembles using inherent voting. In: *Intelligent Agent Multi-Agent Systems, 2009. IAMA 2009. International Conference on*. [s.n.], 2009. p. 1–5. Disponível em: <http://dx.doi.org/10.1109/IAMA.2009.5228053>.

ANDROID. 2017. Disponível em: <http://www.android.com>. Acesso em: 16 mai. 2017.

ANKERST, M. et al. OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 28, n. 2, p. 49–60, jun. 1999. ISSN 0163-5808. Disponível em: <http://dx.doi.org/10.1145/304181.304187>.

ARCGIS: ESRI ArcGIS. 2017. Disponível em: <http://www.arcgis.com/>. Acesso em: 02 mai. 2017.

ARONOFF, S. Geographic information systems: A management perspective. *Geocarto International*, v. 4, n. 4, p. 58, 1989. Disponível em: <http://dx.doi.org/10.1080/10106048909354237>.

ASSUNCÃO, R. M. et al. Efficient regionalization techniques for socioeconomic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, Taylor & Francis, v. 20, n. 7, p. 797–811, ago. 2006. ISSN 1365-8816. Disponível em: <http://dx.doi.org/10.1080/13658810600665111>.

BARDOSSY, A.; BOGARDI, I.; KELLY, W. Imprecise (fuzzy) information in geostatistics. *Mathematical Geology*, Kluwer Academic Publishers-Plenum Publishers, v. 20, n. 4, p. 287–311, 1988. ISSN 0882-8121. Disponível em: <http://dx.doi.org/10.1007/BF00892981>.

BARDOSSY, A.; BOGARDI, I.; KELLY, W. Kriging with imprecise (fuzzy) variograms. I: Theory. *Mathematical Geology*, Kluwer Academic Publishers-Plenum Publishers, v. 22, n. 1, p. 63–79, 1990. ISSN 0882-8121. Disponível em: <http://dx.doi.org/10.1007/BF00890297>.

BAYNE, C. K. et al. Monte Carlo comparisons of selected clustering procedures. *Pattern Recognition*, Elsevier, v. 12, n. 2, p. 51–62, 1980. ISSN 0031-3203. Disponível em: <http://dx.doi.org/10.1016/0031-3203(80)90002-3>.

BAZZI, C. L. *Software para definição e avaliação de unidades de manejo em agricultura de precisão.* 111 p. Tese (Doutorado) — Universidade Estadual do Oeste do Paraná, Cascavel, 2011.

BAZZI, C. L. et al. Software Para Definição e Avaliação de Unidades de Manejo em Agricultura de Precisão. In: *Anais do Congresso Brasileiro de Agricultura de Precisão - COnBAP*. Ribeirão Preto, SP: SBEA, 2012.

BAZZI, C. L. et al. Management zones definition using soil chemical and physical attributes in a soybean area. *Engenharia Agrícola*, Scielo, v. 33, p. 952 – 964, 10 2013. ISSN 0100-6916. Disponível em: <http://dx.doi.org/10.1590/S0100-69162013000500007>.

BEAUJARDIERE, J. d. L. OpenGIS Web Map Server Implementation Specification. *Open Geospatial Consortium Inc., OGC*, p. 6–42, 2006. Disponível em: <http://portal.opengeospatial.org/files/?artifact_id=14416>.

BÉDARD, Y. Visual modelling of spatial databases: towards spatial PVL and UML. *Geomatica*, v. 53, n. 2, p. 169–186, 1999. ISSN 1195-1036.

BÉDARD, Y. et al. Adapting Data Models For The Desing Of Spatio-Temporal Databases. *Computer, Environment and Urban Systems an International Journal*, v. 20, n. 1, p. 19–41, 1996. ISSN 0198-9715. Disponível em: <http://dx.doi.org/10.1016/S0198-9715(96)00008-7>.

BENESTY, J. et al. Pearson Correlation Coefficient. In: *Noise Reduction in Speech Processing SE - 5*. Springer Berlin Heidelberg, 2009, (Springer Topics in Signal Processing, v. 2). p. 1–4. ISBN 978-3-642-00295-3. Disponível em: <http://dx.doi.org/10.1007/978-3-642-00296-0_5>.

BERNARDI, A. C. C. et al. Spatial variability of soil properties and yield of a grazed alfalfa pasture in Brazil. *Precision Agriculture*, Springer US, v. 17, n. 6, p. 737–752, 2016. ISSN 1385-2256. Disponível em: <https://dx.doi.org/10.1007/s11119-016-9446-9>.

BERNARDI, A. C. d. C. et al. *Agricultura de precisão: resultados de um novo olhar*. 1 ed. Brasília: Empresa Brasileira de Pesquisa Agropecuaria, 2014. 596 p. ISBN 978-85-7035-352-8.

BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981. ISBN 0306406713.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM : The Fuzzy c-Means Clustering Algorithm. *Computer & Geosciences*, v. 10, n. 2-3, p. 191–203, 1984. ISSN 0098-3004. Disponível em: <http://dx.doi.org/10.1016/0098-3004(84)90020-7>.

BOHLING, G. Kriging. *Kansas Geological Survey, Tech. Rep*, 2005.

BOOCH, G.; RUMBAUGH, J.; JACOBSON, I. *The Unified Modeling Language For Object-Oriented Development, Documentation Set Version 1.0*. Santa Clara, CA: Rational Software Corporation, 1997.

BORGES, K.; DAVIS, C.; LAENDER, A. OMT-G: An Object-Oriented Data Model for Geographic Applications. *GeoInformatica*, Kluwer Academic Publishers, v. 5, n. 3, p. 221–260, 2001. ISSN 1384-6175. Disponível em: <http://dx.doi.org/10.1023/A%3A1011482030093>.

BOYDELL, B.; MCBRATNEY, A. Identifying potential within-field management zones from cotton-yield estimates. *Precision Agriculture*, Kluwer Academic Publishers, v. 3, n. 1, p. 9–23, 2002. ISSN 1385-2256. Disponível em: <http://dx.doi.org/10.1023/A%3A1013318002609>.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. *Agricultura de precisão: boletim técnico*. 2a. ed. Brasília, DF, 2011. 36 p.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. *Agricultura de Precisão: boletim técnico*. 3a. ed. Brasília, DF, 2013. 36 p.

BRASIL. Ministério do Planejamento. Comissão Nacional de Cartografia. *Perfil de Metadados Geoespaciais do Brasil*. [S.l.], 2009.

BROCK, A. et al. Defining Yield-Based Management Zones for Corn-Soybean Rotations. *Agronomy Journal*, v. 97, n. 4, p. 1115–1128, jul. 2005. Disponível em: <http://dx.doi.org/10.2134/agronj2004.0220>.

BURROUGH, P. A. Principles of geographical information systems for land resources assessment. *Geocarto International*, v. 1, n. 3, p. 54–54, 1986. Disponível em: <http://dx.doi.org/10.1080/10106048609354060>.

CAI, W.; CHEN, S.; ZHANG, D. Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognition*, v. 40, n. 3, p. 825–838, mar. 2007. ISSN 0031-3203. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2006.07.011>.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Disponível em: <http://dx.doi.org/10.1080/03610927408827101>.

CÂMARA, G. et al. *Anatomia dos Sistemas de Informações Geográficas*. [S.l.]: Campinas-SP: Instituto de Computação, UNICAMP, 1996.

CARNIEL, A. C.; CIFERRI, R. R.; CIFERRI, C. D. A. The VagueGeometry abstract data type. *JIDM*, v. 7, n. 1, p. 18–34, 2016.

CASANOVA, M. A. et al. *Bancos de Dados Geográficos*. Curitiba: MundoGeo, 2005. 506 p.

CERNY, V. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, Kluwer Academic Publishers-Plenum Publishers, v. 45, n. 1, p. 41–51, 1985. ISSN 0022-3239. Disponível em: <http://dx.doi.org/10.1007/BF00940812>.

CHANG, D. et al. Delineation of management zones using an active canopy sensor for a tobacco field. *Computers and Electronics in Agriculture*, Elsevier B.V., v. 109, p. 172–178, 2014. ISSN 01681699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2014.09.019>.

CHEN, P. P. S. The Entity-relationship Model - Toward a Unified View of Data. *ACM Trans. Database Syst.*, ACM, New York, NY, USA, v. 1, n. 1, p. 9–36, mar. 1976. ISSN 0362-5915. Disponível em: <http://dx.doi.org/10.1145/320434.320440>.

CHOO, J. et al. MOSAIC: A Proximity Graph Approach for Agglomerative Clustering. In: SONG, I. Y.; EDER, J.; NGUYEN, T. M. (Ed.). *Data Warehousing and Knowledge Discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, (Lecture Notes in Computer Science, v. 4654). p. 231–240. ISBN 978-3-540-74552-5. Disponível em: <http://dx.doi.org/10.1007/978-3-540-74553-2_21>.

CID-GARCIA, N. M. et al. Rectangular shape management zone delineation using integer linear programming. *Computers and Electronics in Agriculture*, Elsevier B.V., v. 93, p. 1–9, 2013. ISSN 01681699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2013.01.009>.

CID-GARCIA, N. M.; BRAVO-LOZANO, A. G.; RIOS-SOLIS, Y. A. A crop planning and real-time irrigation method based on site-specific management zones and linear programming. *Computers and Electronics in Agriculture*, v. 107, n. Supplement C, p. 20 – 28, 2014. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2014.06.002>.

CIFERRI, R. R. *Um benchmark voltado à análise de desempenho de sistemas de informações geográficas*. Dissertação (Mestrado em Ciência da Computação) — Universidade Estadual de Campinas, 1995.

CIFERRI, R. R. *Análise da Influência do Fator Distribuição Espacial dos Dados no Desempenho de Métodos de Acesso Multidimensionais*. 279 p. Tese (Doutorado em Ciência da Computação) — Universidade Federal de Pernambuco, 2002.

CLEMENTINI, E.; DI FELICE, P. A Model for Representing Topological Relationships Between Complex Geometric Features in Spatial Databases. *Inf. Sci.*, Elsevier Science Inc., New York, NY, USA, v. 90, n. 1-4, p. 121–136, abr. 1996. ISSN 0020-0255. Disponível em: <http://dx.doi.org/10.1016/0020-0255(95)00289-8>.

CLEMENTINI, E.; SHARMA, J.; EGENHOFER, M. J. Modelling topological spatial relations: Strategies for query processing. *Computers & Graphics*, v. 18, n. 6, p. 815–822, 1994. ISSN 0097-8493. Disponível em: <http://dx.doi.org/10.1016/0097-8493(94)90007-8>.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, v. 20, n. 1, p. 37–46, 1960. Disponível em: <http://dx.doi.org/10.1177/001316446002000104>.

CÓRDOBA, M. et al. Subfield management class delineation using cluster analysis from spatial principal components of soil variables. *Computers and Electronics in Agriculture*, v. 97, p. 6–14, 2013. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2013.05.009>.

COX, S. et al. *OpenGIS Geography Markup Language (GML) Implementation Specification, version 3.0*. Open Geospatial Consortium Inc., 2002. Disponível em: <http://portal.opengeospatial.org/files/?artifact_id=7174>.

DAVIDSON, I.; RAVI, S. S. Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results. In: JORGE, A. et al. (Ed.). *Knowledge Discovery in Databases: PKDD 2005*. Springer Berlin Heidelberg, 2005, (Lecture Notes in Computer Science, v. 3721). p. 59–70. ISBN 978-3-540-29244-9. Disponível em: <http://dx.doi.org/10.1007/11564126_11>.

DAVIES, D. L.; BOULDIN, D. W. A Cluster Separation Measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1, n. 2, p. 224–227, 1979. ISSN 0162-8828. Disponível em: <http://dx.doi.org/10.1109/TPAMI.1979.4766909>.

DAVIS, B. E. *GIS: A visual approach*. [S.l.]: Cengage Learning, 2001.

DEEGREE: DeeGree. 2017. Disponível em: <http://www.deegree.org>. Acesso em: 18 mai. 2017.

DELAUNAY, B. Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, v. 7, n. 793-800, p. 1–2, 1934.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, Wiley for the Royal Statistical Society, v. 39, n. 1, p. 1–38, 1977. ISSN 00359246.

DERBY, N. E.; CASEY, F. X. M.; FRANZEN, D. W. Comparison of Nitrogen Management Zone Delineation Methods for Corn Grain Yield. *Agronomy Journal*, v. 99, n. 2, p. 405–414, mar. 2007. Disponível em: <http://dx.doi.org/10.2134/agronj2006.0027>.

DIAMOND, P. Fuzzy Kriging. *Fuzzy Sets and Systems*, v. 33, n. 3, p. 315–332, 1989. ISSN 0165-0114. Disponível em: <http://dx.doi.org/10.1016/0165-0114(89)90121-8>.

DOERGE, T. A. Management Zone Concepts. *Site-Specific Management Guidelines*, n. 2, p. 4, 1999.

DRAGIĆEVIĆ, S. The potential of Web-based GIS. *Journal of Geographical Systems*, Springer-Verlag, v. 6, n. 2, p. 79–81, 2004. ISSN 1435-5930. Disponível em: <http://dx.doi.org/10.1007/s10109-004-0133-4>.

DRAY, S.; CHESSEL, D.; THIOULOUSE, J. Co-inertia analysis and the linking of ecological data tables. *Ecology*, Ecological Society of America, v. 84, n. 11, p. 3078–3089, 2003. ISSN 0012-9658. Disponível em: <http://dx.doi.org/10.1890/03-0178>.

ELMASRI, R.; NAVATHE, S. B. *Fundamentals of Database Systems*. 6th ed. Boston: Addison-Wesley, 2011. 1172 p. ISSN 14337851. ISBN 9780136086208.

EMBRAPA. Súmula da 10a. Reunião Técnica de Levantamento de Solos. *Súmula. Rio de Janeiro: EMBRAPA-SNLCS, 1979b. 83p.(EMBRAPA. SNLCS. Série Miscelânea, 1)*, 1979.

ESTER, M. et al. Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support. *Data Min. Knowl. Discov.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 4, n. 2-3, p. 193–216, 2000. ISSN 1384-5810. Disponível em: <http://dx.doi.org/10.1023/A:1009843930701>.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J. Knowledge Discovery in Spatial Databases. In: FöRSTNER, W. et al. (Ed.). *Mustererkennung 1999 SE - 1*. Springer Berlin Heidelberg, 1999, (Informatik aktuell). p. 1–14. ISBN 978-3-540-66381-2. Disponível em: <http://dx.doi.org/10.1007/978-3-642-60243-6_1>.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. [S.l.]: AAAI Press, 1996. p. 226–231.

ESTIVILL-CASTRO, V.; LEE, I. Amoeba: Hierarchical clustering based on spatial proximity using delaunay diagram. In: *Proceedings of the 9th International Symposium on Spatial Data Handling*. Beijing, China: [s.n.], 2000. p. 7–26. Disponível em: <http://dx.doi.ogr/10.1.1.125.4505>.

ESTIVILL-CASTRO, V.; LEE, I. Multi-Level Clustering and its Visualization for Exploratory Spatial Analysis. *GeoInformatica*, Kluwer Academic Publishers, v. 6, n. 2, p. 123–152, 2002. ISSN 1384-6175. Disponível em: <http://dx.doi.org/10.1023/A%3A1015279009755>.

EVERITT, B. S. et al. *Cluster Analysis*. Chichester, UK: John Wiley & Sons, Ltd, 2011. (Wiley Series in Probability and Statistics). ISBN 9780470977811. Disponível em: <http://dx.doi.org/10.1002/9780470977811>.

FABBRI, S. et al. Managing Literature Reviews Information through Visualization. In: MACIASZEK, L. A.; CUZZOCREA, A.; CORDEIRO, J. (Ed.). *ICEIS 2012 - Proceedings of the 14th International Conference on Enterprise Information Systems*. Wroclaw, Poland: SciTePress, 2012. v. 2, p. 36–45. ISBN 978-989-8565-11-2.

FAN, B. A Hybrid Spatial Data Clustering Method for Site Selection: The Data Driven Approach of GIS Mining. *Expert Syst. Appl.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 36, n. 2, p. 3923–3936, 2009. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2008.02.056>.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM*, ACM, New York, NY, USA, v. 39, n. 11, p. 27–34, 1996. ISSN 0001-0782. Disponível em: <http://dx.doi.org/10.1145/240455.240464>.

FAYYAD, U. M. et al. (Ed.). *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. ISBN 0-262-56097-6.

FGDC. Federal Geographic Data Committee. *Content standard for digital geospatial metadata (FGDC-STD-001-1998)*. Washington, DC, jun. 1998.

FISHER, R. A. *Statistical methods for research workers*. [S.l.]: Genesis Publishing Pvt Ltd, 1925.

FLOREK, K. et al. Sur la liaison et la division des points d'un ensemble fini. In: *Colloquium Mathematicae*. [S.l.: s.n.], 1951. v. 2, p. 282–285.

FOUNTAS, S. et al. Farm management information systems: Current situation and future perspectives. *Computers and Electronics in Agriculture*, Elsevier, v. 115, p. 40–50, 2015. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2015.05.011>.

FRANKE, R. Scattered data interpolation: Tests of some methods. *Mathematics of computation*, American Mathematical Society, v. 38, n. 157, p. 181–200, 1982. ISSN 1088-6842. Disponível em: <http://dx.doi.org/10.1090/S0025-5718-1982-0637296-4>.

FRANZEN, D. W.; NANNA, T. Comparison of Nitrogen Management Zone Delineation Methods. In: *North Central Extension-Industry Soil Fertility Conference*. Des Moines, IA: [s.n.], 2003. v. 19, p. 114–118.

FRED, a. N. L.; JAIN, a. K. Combining multiple clusterungs using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 27, n. 6, p. 835–850, 2005. ISSN 0162-8828. Disponível em: <https://dx.doi.org/10.1109/TPAMI.2005.113>.

FRIDGEN, J. J. et al. Management Zone Analyst (MZA): Software for Subfield Management Zone Delineation. *Agronomy Journal*, v. 96, p. 100–108, 2004.

GATH, I.; GEVA, A. Unsupervised optimal fuzzy clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 11, n. 7, p. 773–780, Jul 1989. ISSN 0162-8828. Disponível em: <http://dx.doi.org/10.1109/34.192473>.

GDAL. *Geospatial Data Abstraction Library*. 2017. Disponível em: <http://www.gdal.org>. Acesso em: 05 jul. 2017.

GEOSERVER. *Open Source Server for Sharing Geospatial Data*. 2017. Disponível em: <http://geoserver.org>. Acesso em: 18 mai. 2017.

GHOSH, J.; ACHARYA, A. Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 1, n. 4, p. 305–315, 2011. ISSN 19424787. Disponível em: <http://dx.doi.org/10.1002/widm.32>.

GHOSH, S.; DUBEY, S. K. Comparative analysis of k-means and fuzzy c-means algorithms. *International Journal of Advanced Computer Science and Applications*, v. 4, n. 4, p. 35–39, 2013.

GOLDEN, R. R.; MEEHL, P. E. Detection of biological sex: An empirical test of cluster methods. *Multivariate Behavioral Research*, Taylor & Francis, v. 15, n. 4, p. 475–496, 1980. Disponível em: <http://dx.doi.org/10.1207/s15327906mbr1504_6>.

GREGO, C. R. et al. Estoque de carbono no solo e produtividade da cana-de-açúcar analisados quanto a variabilidade espacial. In: INAMASU, R. Y. et al. (Ed.). *Agricultura de Precisão: um novo olhar*. 1a. ed. São Carlos, SP: Embrapa Instrumentação Agropecuária, 2011. p. 240–244. ISBN 978-85-86463-31-0.

GREGO, C. R. et al. Geoestatística aplicada a condutividade elétrica do solo e altitude do solo cultivado com cana-de-açúcar. In: INAMASU, R. Y. et al. (Ed.). *Agricultura de Precisão: um novo olhar*. 1 ed. São Carlos, SP: Embrapa Instrumentação Agropecuária, 2011. p. 245–248. ISBN 978-85-86463-31-0.

GROSS, A. L. A Monte Carlo study of the accuracy of a hierarchical grouping procedure. *Multivariate behavioral research*, Taylor & Francis, v. 7, n. 3, p. 379–389, 1972.

GUO, D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, Taylor & Francis, v. 22, n. 7, p. 801–823, jul. 2008. ISSN 1365-8816. Disponível em: <http://dx.doi.org/10.1080/13658810701674970>.

GUO, D.; PEUQUET, D.; GAHEGAN, M. ICEAGE: Interactive Clustering and Exploration of Large and High-Dimensional Geodata. *GeoInformatica*, Kluwer Academic Publishers, v. 7, n. 3, p. 229–253, 2003. ISSN 1384-6175. Disponível em: <http://dx.doi.org/10.1023/A%3A1025101015202>.

GUSTAFSON, D.; KESSEL, W. Fuzzy clustering with a fuzzy covariance matrix. In: *Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on*. [s.n.], 1978. p. 761–766. Disponível em: <http://dx.doi.org/10.1109/CDC.1978.268028>.

GUTIN, G.; PUNNEN, A. P. *The traveling salesman problem and its variations*. [S.l.]: Springer Science & Business Media, 2006.

GÜTING, R. An introduction to spatial database systems. *The VLDB Journal*, Springer-Verlag, v. 3, n. 4, p. 357–399, 1994. ISSN 1066-8888. Disponível em: <http://dx.doi.org/10.1007/BF01231602>.

HADZILACOS, T.; TRYFONA, N. An Extended Entity-relationship Model for Geographic Applications. *SIGMOD Rec.*, ACM, New York, NY, USA, v. 26, n. 3, p. 24–29, set. 1997. ISSN 0163-5808. Disponível em: <http://dx.doi.org/10.1145/262762.262766>.

HALKIDI, M.; VAZIRGIANNIS, M.; BATISTAKIS, Y. Quality Scheme Assessment in the Clustering Process. In: ZIGHED, D.; KOMOROWSKI, J.; Å»YTKOW, J. (Ed.). *Principles of Data Mining and Knowledge Discovery*. Springer Berlin Heidelberg, 2000, (Lecture Notes in Computer Science, v. 1910). p. 265–276. ISBN 978-3-540-41066-9. Disponível em: <http://dx.doi.org/10.1007/3-540-45372-5_26>.

HAN, E.-H. et al. Clustering based on association rule hypergraphs. In: *Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*. [S.l.: s.n.], 1997. p. 9–13.

HAN, J. et al. Survey on NoSQL database. *Proceedings - 2011 6th International Conference on Pervasive Computing and Applications, ICPCA 2011*, p. 363–366, 2011. ISSN 978-1-4577-0207-5.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. New York, NY: Springer New York, 2009. (Springer Series in Statistics). ISBN 978-0-387-84857-0. Disponível em: <http://dx.doi.org/10.1007/978-0-387-84858-7>.

HATHAWAY, R. J.; BEZDEK, J. C. Local convergence of the fuzzy c-Means algorithms. *Pattern Recognition*, v. 19, n. 6, p. 477–480, 1986. ISSN 00313203. Disponível em: <http://dx.doi.org/10.1016/0031-3203(86)90047-6>.

HERNANDES, E. et al. Using GQM and TAM to evaluate StArt - a tool that supports Systematic Review. *CLEI Electronic Journal*, Scielo, v. 15, p. 3 – 3, 04 2012. ISSN 0717-5000.

HERRING, J. *The OpenGIS abstract specification, Topic 1: Feature geometry (ISO 19107 Spatial schema), version 5*. Open GeoSpatial Consortium Inc., 2001. Disponível em: <http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=26012>.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, Warwick & York, v. 24, n. 6, p. 417–441, 1933. Disponível em: <http://dx.doi.org/10.1037/h0071325>.

HU, C.; MENG, L.; SHI, W. Fuzzy clustering validity for spatial data. *Geo-spatial Information Science*, Wuhan University, v. 11, n. 3, p. 191–196, 2008. ISSN 1009-5020. Disponível em: <http://dx.doi.org/10.1007/s11806-008-0094-8>.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, n. 1, p. 193–218, 1985. ISSN 1432-1343. Disponível em: <http://dx.doi.org/10.1007/BF01908075>.

INAMASU, R. Y.; BERNARDI, A. C. D. C. Agricultura de Precisão. In: BERNARDI, A. C. d. C. et al. (Ed.). *Agricultura de Precisão: resultados de um novo olhar*. 1 ed. Brasília, DF: Empresa Brasileira de Pesquisa Agropecuaria, 2014. p. 21–33. ISBN 978-85-7035-352-8.

ISAAKS, E. H.; SRIVASTAVA, R. M. *An introduction to applied geostatistics*. 1st ed. [S.l.]: Oxford University Press, 1989. ISBN 978-0195050134.

ISO. International Organization for Standardization. *ISO 19115: 2003 Geographic information-Metadata*. [S.l.], 2003.

JACCARD, P. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull. Soc. Vaud. Sci. Nat.*, v. 37, p. 241–272, 1901. Disponível em: <http://dx.doi.org/10.5169/seals-266440>.

JAIN, A. K.; DUBES, R. C.; OTHERS. *Algorithms for clustering data*. [S.l.]: Prentice hall Englewood Cliffs, 1988. ISBN 978-0130222787.

JASKOWIAK, P. a. et al. On strategies for building effective ensembles of relative clustering validity criteria. *Knowledge and Information Systems*, Springer London, 2015. ISSN 0219-1377. Disponível em: <http://dx.doi.org/10.1007/s10115-015-0851-6>.

JAYAKUMARAN, C.; KARUPPANAN, K. Pattern identification using rough set clustering for spatio-temporal dataset. In: *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*. [s.n.], 2013. p. 1598–1603. Disponível em: <http://dx.doi.org/10.1109/ICACCI.2013.6637419>.

JAYNES, D. B.; COLVIN, T. S.; KASPAR, T. C. Identifying potential soybean management zones from multi-year yield data. *Computers and Electronics in Agriculture*, v. 46, n. 1-3, p. 309–327, 2005. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2004.11.011>.

JAYNES, D. B. et al. Cluster Analysis of Spatiotemporal Corn Yield Patterns in an Iowa Field. *Agronomy Journal*, v. 95, n. 3, p. 574–586, maio 2003. Disponível em: <https://www.agronomy.org/publications/aj/abstracts/95/3/574>.

JIN, H.; MIAO, B. The research progress of spatial data mining technique. In: *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*. [s.n.], 2010. v. 3, p. 81–84. Disponível em: <http://dx.doi.org/10.1109/ICCSIT.2010.5564659>.

JORGE, L. A. d. C.; INAMASU, R. Y.; CARMO, R. B. do. Desenvolvimento de um VANT totalmente configurado para aplicações em Agricultura de Precisão no Brasil. In: *Anais do XV Simpósio Brasileiro de Sensoriamento Remoto - SBSR*. Curitiba, PR: Instituto Nacional de Pesquisas Espaciais, 2011. p. 399–406.

JORGE, L. A. d. C. et al. GeoField-Net: Sistema para scouting no campo. In: INAMASU, R. Y. et al. (Ed.). *Agricultura de Precisão: um novo olhar*. 1 ed. São Carlos, SP: Embrapa Instrumentação Agropecuária, 2011. p. 51–54. ISBN 978-85-86463-31-0.

JOSHI, D. *Polygonal spatial clustering*. 215 p. Tese (Degree of Doctor of Philosophy - Major: Computer Science) — University of Nebraska, 2011.

JOSHI, D.; SOH, L.-K.; SAMAL, A. Redistricting Using Constrained Polygonal Clustering. *Knowledge and Data Engineering, IEEE Transactions on*, v. 24, n. 11, p. 2065–2079, Nov 2012. ISSN 1041-4347. Disponível em: <http://dx.doi.org/10.1109/TKDE.2011.140>.

JULIÃO, R. P. Geografia, Informação e Sociedade. *GEOINOVA - Revista do Departamento de Geografia e Planejamento Rural*, Lisboa, p. 95–108, 1999.

KANUNGO, T. et al. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 7, p. 881–892, 2002. ISSN 0162-8828. Disponível em: <http://dx.doi.org/10.1109/TPAMI.2002.1017616>.

KARYPIS, G.; KUMAR, V. Multilevelk-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed computing*, Elsevier, v. 48, n. 1, p. 96–129, 1998. ISSN 0743-7315. Disponível em: <http://dx.doi.org/10.1006/jpdc.1997.1404>.

KAUFMAN, L.; ROUSSEEUW, P. *Clustering by Means of Medoids*. [S.l.]: Faculty of Mathematics and Informatics, 1987. (Reports of the Faculty of Mathematics and Informatics). ISSN 0920-8577.

KHANI, F. et al. An Algorithm for Discovering Clusters of Different Densities or Shapes in Noisy Data Sets. In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2013. (SAC '13), p. 144–149. ISBN 978-1-4503-1656-9. Disponível em: <http://dx.doi.org/10.1145/2480362.2480392>.

KHOSLA, R. et al. A synthesis of multi-disciplinary research in precision agriculture: site-specific management zones in the semi-arid western Great Plains of the USA. *Precision Agriculture*, Springer US, v. 9, n. 1-2, p. 85–100, 2008. ISSN 1385-2256. Disponível em: <http://dx.doi.org/10.1007/s11119-008-9057-1>.

KHOSLA, R. et al. Spatial Variation and Site-Specific Management Zones. In: OLIVER, M. (Ed.). *Geostatistical Applications for Precision Agriculture*. Springer Netherlands, 2010. p. 195–219. ISBN 978-90-481-9132-1. Disponível em: <http://dx.doi.org/10.1007/978-90-481-9133-8_8>.

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by Simulated Annealing. *Science*, v. 220, n. 4598, p. 671–680, 1983. ISSN 1095-9203. Disponível em: <http://dx.doi.org/10.1126/science.220.4598.671>.

KITCHEN, N. et al. Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity. *Computers and Electronics in Agriculture*, v. 46, n. 1-3, p. 285–308, mar. 2005. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2004.11.012>.

KITCHEN, N. R.; GOULDING, K. W. T.; SHANAHAN, J. F. Proven practices and innovative technologies for on-farm crop nitrogen management. In: HATFIELD, J. L.; FOLLETT, R. F. (Ed.). *Nitrogen in the environment: sources, problems and management*. Amsterdam: Elsevier, 2008. cap. 15, p. 487–517.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering - A tertiary study. *Information and Software Technology*, Elsevier B.V., v. 52, n. 8, p. 792–805, 2010. ISSN 0950-5849. Disponível em: <http://dx.doi.org/10.1016/j.infsof.2010.03.006>.

KOPETZ, H. *Real-Time Systems: Design Principles for Distributed Embedded Applications*. 2nd ed. Springer Publishing Company, Incorporated, 2011. ISBN 978-1-4419-8237-7. Disponível em: <http://dx.doi.org/10.1007/978-1-4419-8237-7>.

KOSTERS, G.; PAGEL, B.-U.; SIX, H.-W. GeoOOA: object-oriented analysis for geographic information systems. In: *Requirements Engineering, 1996., Proceedings of the Second International Conference on*. [s.n.], 1996. p. 245–253. Disponível em: <http://dx.doi.org/10.1109/ICRE.1996.491453>.

KRIGE, D. G. A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, v. 52, p. 119–139, 1952.

KRUMMEL, J.; SU, H. Topographic Effect and Its Relation to Crop Production in an Individual Field. *Precision Agriculture*, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, p. 273–273, 1996. Disponível em: <http://dx.doi.org/10.2134/1996.precisionagproc3.c28>.

KUIPER, F. K.; FISHER, L. 391: A Monte Carlo comparison of six clustering procedures. *Biometrics*, JSTOR, p. 777–783, 1975.

LANCE, G. N.; WILLIAMS, W. T. A general theory of classificatory sorting strategies II. Clustering systems. *The Computer Journal*, Br Computer Soc, v. 10, n. 3, p. 271–277, 1967.

LE, T.; ALTMAN, T.; GARDINER, K. J. A probability based defuzzification method for fuzzy cluster partition. In: *Proc. Intl'Conf. on Artificial Intelligence (WORLDCOMP-ICAI'12)*. [S.l.: s.n.], 2012. v. 2, p. 1038–1043.

LI, M.; CHUNG, S.-O. Special issue on precision agriculture. *Computers and Electronics in Agriculture*, Elsevier B.V., v. 112, p. 1, 2015. ISSN 01681699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2015.03.014>.

LI, Y. et al. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. *Computers and Electronics in Agriculture*, v. 56, n. 2, p. 174–186, abr. 2007. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2007.01.013>.

LINSLEY, C. M.; BAUER, F. C. Test your soil for acidity. *University of Illinois, Agricultural Experiment Station*, n. 346, p. 2–4, 1929.

LISBOA-FILHO, J.; IOCHPE, C. Specifying Analysis Patterns for Geographic Databases on the Basis of a Conceptual Framework. In: *Proceedings of the 7th ACM International Symposium on Advances in Geographic Information Systems*. New York, NY, USA: ACM, 1999. (GIS '99), p. 7–13. ISBN 1-58113-235-2. Disponível em: <http://dx.doi.org/10.1145/320134.320139>.

LIU, X.; CHEN, F.; LU, C.-T. Robust Prediction and Outlier Detection for Spatial Datasets. In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. [s.n.], 2012. p. 469–478. ISSN 1550-4786. Disponível em: <http://dx.doi.org/10.1109/ICDM.2012.147>.

LLOYD, S. P. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, v. 28, n. 2, p. 129–137, 1982. ISSN 15579654.

LONGLEY, P. *Geographic information systems and science*. [S.l.]: John Wiley & Sons, 2005.

LOQUIN, K.; DUBOIS, D. Kriging and Epistemic Uncertainty: A Critical Discussion. In: JEANSOULIN, R. et al. (Ed.). *Methods for Handling Imperfect Spatial Information*. Springer Berlin Heidelberg, 2010, (Studies in Fuzziness and Soft Computing, v. 256). p. 269–305. ISBN 978-3-642-14754-8. Disponível em: <http://dx.doi.org/10.1007/978-3-642-14755-5_11>.

LOQUIN, K.; DUBOIS, D. A fuzzy interval analysis approach to kriging with ill-known variogram and data. *Soft Computing*, Springer-Verlag, v. 16, n. 5, p. 769–784, 2012. ISSN 1432-7643. Disponível em: <http://dx.doi.org/10.1007/s00500-011-0768-2>.

LOTT, R. *OGC abstract specification topic 2, spatial referencing by coordinates*. Open Geospatial Consortium, Inc., 2004. Disponível em: <http://portal.opengeospatial.org/files/?artifact_id=39049>.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA.. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, p. 281–297.

MAPSERVER: Open source platform for publishing spatial data and interactive mapping applications to the web. 2017. Disponível em: <http://mapserver.org>. Acesso em: 18 mai. 2017.

MARGULES, C. R.; FAITH, D. P.; BELBIN, L. An adjacency constraint in agglomerative hierarchical classifications of geographic data. *Environment and Planning A*, Pion Ltd, London, v. 17, n. 3, p. 397–412, 1985.

MASSRUHÁ, S. M. F. S. et al. *Tecnologias da informação e comunicação e suas relações com a agricultura.* 1 ed. Brasília, DF: Embrapa Informática Agropecuária, 2014. 411 p. ISBN 978-85-7035-414-3.

MATHERON, G. Principles of geostatistics. *Economic geology*, Society of Economic Geologists, v. 58, n. 8, p. 1246–1266, 1963.

MATHERON, G. Le krigeage universel. École nationale supérieure des mines de Paris, Paris, France, 1969.

MENEGATTI, L. A. A. Metodologia para identificação e caracterização de erros em mapas de produtividade. *Revista Brasileira de Engenharia Agrícola e Ambiental*, v. 7, n. 2, p. 367–374, 2003. ISSN 1807-1929.

METROPOLIS, N.; ULAM, S. The Monte Carlo method. *Journal of the American statistical association*, Taylor & Francis Group, v. 44, n. 247, p. 335–341, 1949.

MILLIGAN, G. W.; COOPER, M. C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, Springer, v. 50, n. 2, p. 159–179, 1985. Disponível em: <http://dx.doi.org/10.1007/BF02294245>.

MILLIGAN, G. W.; SCHILLING, D. A. Asymptotic and Finite Sample Characteristics of Four External Criterion Measures. *Multivariate Behavioral Research*, Taylor & Francis, v. 20, n. 1, p. 97–109, 1985. Disponível em: <http://dx.doi.org/10.1207/s15327906mbr2001_6>.

MILNE, A. E. et al. Spatial multivariate classification of an arable field into compact management zones based on past crop yields. *Computers and Electronics in Agriculture*, Elsevier B.V., v. 80, p. 17–30, 2012. ISSN 01681699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2011.10.007>.

MITCHELL, T. *Web mapping illustrated: using open source GIS toolkits*. [S.l.]: O'Reilly Media, Inc., 2005.

MITCHELL, T. M. *Machine Learning*. New York; London: McGraw-Hill, 1997. 414 p. ISBN 0070428077.

MOJENA, R. Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal*, Br Computer Soc, v. 20, n. 4, p. 359–363, 1977.

MOLIN, J. P. Agricultura de Precisão: Situação atual e perspectivas. In: FANCELLI, A. L.; NETO, D. D. (Ed.). *Milho: Estratégias de Manejo para Alta Produtividade*. Piracicaba: ESALQ/USP/LPV, 2003. p. 89–98.

MOLIN, J. P. Tendências da Agricultura de Precisão no Brasil. In: *Congresso Brasileiro de Agricultura de Precisão*. [S.l.: s.n.], 2004. p. 1–10.

MOLIN, J. P.; AMARAL, L. R. do; COLAÇO, A. *Agricultura de Precisão*. [S.l.]: Oficina de Textos, 2015. ISBN 9788579752148.

MORAL, F.; TERRÓN, J.; REBOLLO, F. Site-specific management zones based on the Rasch model and geostatistical techniques. *Computers and Electronics in Agriculture*, Elsevier B.V., v. 75, n. 2, p. 223–230, fev. 2011. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2010.10.014>.

MORALES, E. R. C.; MENDIZABAL, Y. Y. A New Contiguity-Constrained Agglomerative Hierarchical Clustering Algorithm for Image Segmentation. In: MESEGUER, P.; MANDOW, L.; GASCA, R. (Ed.). *Current Topics in Artificial Intelligence SE - 27*. Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 5988). p. 261–270. ISBN 978-3-642-14263-5. Disponível em: <http://dx.doi.org/10.1007/978-3-642-14264-2_27>.

MORARI, F.; CASTRIGNANÒ, a.; PAGLIARIN, C. Application of multivariate geostatistics in delineating management zones within a gravelly vineyard using geo-electrical sensors. *Computers and Electronics in Agriculture*, v. 68, n. 1, p. 97–107, ago. 2009. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2009.05.003>.

MURAKAMI, E. et al. An infrastructure for the development of distributed service-oriented information systems for precision agriculture. *Computers and Electronics in Agriculture*, v. 58, n. 1, p. 37–48, ago. 2007. ISSN 0168-1699. Disponível em: <dx.doi.org/10.1016/j.compag.2006.12.010>.

NALDI, M. C.; CARVALHO, a. C. P. L. F.; CAMPELLO, R. J. G. B. Cluster ensemble selection based on relative validity indexes. *Data Mining and Knowledge Discovery*, v. 27, n. 2, p. 259–289, 2013. ISSN 1384-5810. Disponível em: <http://dx.doi.org/10.1007/s10618-012-0290-x>.

NASCIMENTO, P. d. S. *Manejo da viticultura irrigada no semiárido com base em zonas homogêneas do solo e da planta*. 59 p. Tese (Doutorado em Agronomia) — Universidade Estadual Paulista, 2013.

NASH, E.; KORDUAN, P.; BILL, R. Applications of open geospatial web services in precision agriculture: a review. *Precision Agriculture*, Springer US, v. 10, n. 6, p. 546–560, 2009. ISSN 1385-2256. Disponível em: <http://dx.doi.org/10.1007/s11119-009-9134-0>.

NG, R. T.; HAN, J. Efficient and Effective Clustering Methods for Spatial Data Mining. In: *Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994. (VLDB '94), p. 144–155. ISBN 1-55860-153-8.

ODEH, I. O. A.; CHITTLEBOROUGH, D. J.; MCBRATNEY, A. B. Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. *Soil Science Society of America Journal*, Soil Science Society of America, v. 56, n. 2, p. 505–516, 1992.

OGC. *The Open Geospatial Consortium*. 2017. Disponível em: <http://www.opengeospatial.org>. Acesso em: 18 mai. 2017.

OLDONI, H.; BASSOI, L. H. Delineation of irrigation management zones in a Quartzipsamment of the Brazilian semiarid region. *Pesquisa Agropecuária Brasileira*, v. 51, n. 9, p. 1283–1294, sep 2016. ISSN 0100-204X. Disponível em: <http://dx.doi.org/10.1590/s0100-204x2016000900028>.

OLIVEIRA, R. P. de et al. Sistematização do índice de oportunidade na adoção da agricultura de precisão para diferentes sistemas produtivos. Empresa Brasileira de Pesquisa Agropecuaria, Brasília, DF, p. 173–179, 2014.

OPENGEO. *Opengeo Suite*. 2017. Disponível em: <http://suite.opengeo.org>. Acesso em: 18 mai. 2017.

OPENLAYERS. *Open Layers 3*: A high-performance, feature-packed library for all your mapping needs. 2017. Disponível em: <http://openlayers.org>. Acesso em: 18 mai. 2017.

ORACLE. *Oracle Documentation*. 2017. Disponível em: <http://docs.oracle.com>. Acesso em: 18 mai. 2017.

ORD, J. K.; GETIS, A. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, Wiley Online Library, v. 27, n. 4, p. 286–306, 1995. Disponível em: <http://dx.doi.org/10.1111/j.1538-4632.1995.tb00912.x>.

ORTEGA, R.; SANTIBÁÑEZ, O. Determination of management zones in corn (Zea mays L.) based on soil fertility. *Computers and Electronics in Agriculture*, v. 58, n. 1, p. 49–59, ago. 2007. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2006.12.011>.

PARENT, C. et al. Modeling Spatial Data in the MADS Conceptual Model. In: *Proceedings of the International Symposium on Spatial Data Handling, SDH*. Vancouver, Canada: [s.n.], 1998. p. 138–150.

PEARSON, K. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, JSTOR, v. 58, p. 240–242, 1895. Disponível em: <http://dx.doi.org/10.1098/rspl.1895.0041>.

PEDROSO, M. et al. A segmentation algorithm for the delineation of agricultural management zones. *Computers and Electronics in Agriculture*, v. 70, n. 1, p. 199–208, jan. 2010. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2009.10.007>.

PEETERS, A. et al. Getis-Ord's hot- and cold-spot statistics as a basis for multivariate spatial clustering of orchard tree data. *Computers and Electronics in Agriculture*, Elsevier B.V., v. 111, p. 140–150, 2015. ISSN 01681699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2014.12.011>.

PEI, T. et al. DECODE: A New Method for Discovering Clusters of Different Densities in Spatial Data. *Data Min. Knowl. Discov.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 18, n. 3, p. 337–369, 2009. ISSN 1384-5810. Disponível em: <http://dx.doi.org/10.1007/s10618-008-0120-3>.

PERALTA, N. R.; COSTA, J. L. Delineation of management zones with soil apparent electrical conductivity to improve nutrient management. *Computers and Electronics in Agriculture*, v. 99, p. 218–226, 2013. ISSN 01681699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2013.09.014>.

PERALTA, N. R. et al. Delineation of management zones to improve nitrogen management of wheat. *Computers and Electronics in Agriculture*, v. 110, p. 103–113, 2015. ISSN 01681699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2014.10.017>.

PERCIVALL, G. *The OpenGIS Abstract Specification-Topic 12: OpenGIS Service Architecture Version 4.3*. Open GeoSpatial Consortium Inc., 2002. Disponível em: <portal.opengeospatial.org/files/?artifact_id=1221>.

PERCIVALL, G. et al. *OGC Reference Model*. Open Geospatial Consortium Inc., 2003. 108 p. Disponível em: <http://www.opengeospatial.org/standards/orm>.

PERRUCHET, C. Constrained agglomerative hierarchical classification. *Pattern Recognition*, v. 16, n. 2, p. 213–217, 1983. ISSN 0031-3203. Disponível em: <http://dx.doi.org/10.1016/0031-3203(83)90024-9>.

PIERCE, F. J.; NOWAK, P. Aspects of Precision Agriculture. In: SPARKS, D. L. (Ed.). *Advances in Agronomy*. Academic Press, 1999, (Advances in Agronomy, v. 67). p. 1–85. Disponível em: <http://dx.doi.org/10.1016/S0065-2113(08)60513-1>.

POSTGRESQL: PostgreSQL Documentation. 2017. Disponível em: <http://www.postgresql.org/docs/9.4/static/docguide.html>. Acesso em: 18 mai. 2017.

POWERS, D. M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Bioinfo Publications, v. 2, p. 37–66, 2011.

PRINGLE, M. et al. A preliminary approach to assessing the opportunity for site-specific crop management in a field, using yield monitor data. *Agricultural Systems*, v. 76, n. 1, p. 273 – 292, 2003. ISSN 0308-521X. Disponível em: <http://dx.doi.org/10.1016/S0308-521X(02)00005-7>.

QGIS: A Free and Open Source Geographic Information System. 2017. Disponível em: <http://www.qgis.org/>. Acesso em: 04 mai. 2017.

QUEIRÓS, L. R. et al. Análise das possibilidades e tendências do uso das tecnologias da informação e comunicação em Agricultura de Precisão. In: BERNARDI, A. C. d. C. et al. (Ed.). *Agricultura de Precisão: resultados de um novo olhar*. Brasília, DF: Empresa Brasileira de Pesquisa Agropecuaria, 2014. p. 97–108. ISBN 978-85-7035-352-8.

QUEIRÓS, L. R. et al. *Gestão de recursos de informação em Agricultura de Precisão*. Campinas, SP: Embrapa Informática Agropecuária, 2011. 39 p. (Documentos).

QUEIROZ, G. R. D.; FERREIRA, K. R. SGBD com extensões espaciais. In: CASANOVA, M. et al. (Ed.). *Bancos de Dados Geográficos*. São José dos Campos, SP: MundoGeo, 2005.

R: The R Project for Satistical Computing. 2017. Disponível em: <http://www.r-project.org/>. Acesso em: 15 mai. 2017.

RABELLO, L. M. *Condutividade elétrica do solo, tópicos e equipamentos*. São Carlos, SP: Embrapa Instrumentação, 2009. 19 p.

RABELLO, L. M. et al. Mapeamento da condutividade elétrica do solo - sistema protótipo. In: INAMASU, R. Y. et al. (Ed.). *Agricultura de Precisão: um novo olhar*. 1 ed. São Carlos, SP: Embrapa Instrumentação Agropecuária, 2011. p. 41–45. ISBN 978-85-86463-31-0.

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 66, n. 336, p. 846–850, 1971. Disponível em: <http://dx.doi.org/10.1080/01621459.1971.10482356>.

REZENDE, S. O. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Editora Manole Ltda, 2003. ISBN 8520416837.

RIBEIRO-JÚNIOR, L. C. M. *Uma Arquitetura de Software para Sistemas Espaço-Temporais Baseados na Web para Agricultura de Precisão*. 190 p. Tese (Doutorado em Engenharia) — Universidade de São Paulo, 2007.

ROSA, H. J. A. et al. Sugarcane response to nitrogen rates, measured by a canopy reflectance sensor. *Pesquisa Agropecuária Brasileira*, v. 50, n. 9, p. 840–848, sep 2015. ISSN 0100-204X. Disponível em: <http://dx.doi.org/10.1590/S0100-204X2015000900013>.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, n. 0, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <http://dx.doi.org/10.1016/0377-0427(87)90125-7>.

RUMBAUGH, J. et al. *Object-Oriented Modeling and Design*. Englewood Cliffs, NJ: Prentice-Hall, 1991. ISBN 978-8120310469.

RUSS, G. *Spatial Data Mining in Precision Agriculture*. 251 p. Tese (Doktoringenieur) — Otto-von-Guericke-University of Magdeburg, 2012.

RUSS G.; KRUSE, R. Regression Models for Spatial Data: An Example from Precision Agriculture. In: *Proceedings of the 10th Industrial Conference on Advances in Data Mining: Applications and Theoretical Aspects*. Berlin, Heidelberg: Springer-Verlag, 2010. (ICDM'10), p. 450–463. ISBN 3-642-14399-7, 978-3-642-14399-1. Disponível em: <http://dx.doi.org/10.1007/978-3-642-14400-4_35>.

RUSS G.; KRUSE, R. Exploratory Hierarchical Clustering for Management Zone Delineation in Precision Agriculture. In: PERNER, P. (Ed.). *Advances in Data Mining. Applications and Theoretical Aspects*. Springer Berlin Heidelberg, 2011, (Lecture Notes in Computer Science, v. 6870). p. 161–173. ISBN 978-3-642-23183-4. Disponível em: <http://dx.doi.org/10.1007/978-3-642-23184-1_13>.

RUSS, G.; SCHNEIDER, M.; KRUSE, R. Hierarchical Spatial Clustering for Management Zone Delineation in Precision Agriculture. In: PERNER, P. (Ed.). *Industrial Conference on Data Mining - Workshops*. [S.l.]: IBaI Publishing, 2010. p. 95–104. ISBN 978-3-940501-15-8.

SALVADOR, A.; ANTUNIASSI, U. R. Imagens Aéreas Multiespectrais na Identificação de Zonas de Manejo em Áreas de Algodão para Aplicação Localizada de Insumos. *Revista Energia na Agricultura*, v. 26, n. 2, p. 1–19, 2011. ISSN 2359-6562. Disponível em: <http://dx.doi.org/10.17224/EnergAgric.2011v26n2p01-19>.

SANTOS, R. T.; SARAIVA, A. M. A Reference Process for Management Zones Delineation in Precision Agriculture. *IEEE Latin America Transactions*, v. 13, n. 3, p. 727–738, March 2015. ISSN 1548-0992. Disponível em: <http://dx.doi.org/10.1109/TLA.2015.7069098>.

SANTOS, R. T. dos; SARAIVA, A. M.; MOLIN, J. P. Avaliação do Uso das Coordenadas Geográficas como Parte do Conjunto de Atributos para Definição de Unidades de Gerenciamento Diferenciado. In: *Anais do Congresso Brasileiro de Agricultura de Precisao - COnBAP*. Ribeirão Preto, SP: [s.n.], 2012.

SAWYER, J. Concepts of variable rate technology with considerations for fertilizer application. *Journal of Production Agriculture*, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, v. 7, n. 2, p. 195–201, 1994. Disponível em: <http://dx.doi.org/10.2134/jpa1994.0195>.

SCARPARI, M. S.; GOMES, E.; BEAUCLAIR, F. D. Physiological model to estimate the maturity of sugarcane. *Scientia Agricola*, n. October, p. 622–628, 2009. ISSN 0103-9016. Disponível em: <http://dx.doi.org/10.1590/S0103-90162009000500006>.

SCHENATTO, K. et al. Use of the farmer's experience variable in the generation of management zones. *Semina: Ciências Agrárias*, v. 38, n. 4Supl1, p. 2305, aug 2017. ISSN 1679-0359.

SCHNEIDER, M.; BEHR, T. Topological Relationships Between Complex Spatial Objects. *ACM Trans. Database Syst.*, ACM, New York, NY, USA, v. 31, n. 1, p. 39–81, mar. 2006. ISSN 0362-5915. Disponível em: <http://dx.doi.org/10.1145/1132863.1132865>.

SCHUT, P.; WHITESIDE, A. *OpenGIS Web Processing Service*. Open GeoSpatial Consortium Inc., 2007. Disponível em: <http://portal.opengeospatial.org/files/?artifact_id=24151>.

SCHWALBERT, R. A. et al. Zonas de manejo: atributos de solo e planta visando a sua delimitação e aplicações na agricultura de precisão. *Revista Plantio Direto*, p. 21–32, 2014.

SCHWARTZBERG, J. E. Reapportionment, gerrymanders, and the notion of compactness. *Minn. L. Rev.*, HeinOnline, v. 50, p. 443, 1965.

SCUDIERO, E. et al. Delineation of site-specific management units in a saline region at the Venice Lagoon margin, Italy, using soil reflectance and apparent electrical conductivity. *Computers and Electronics in Agriculture*, Elsevier B.V., v. 99, p. 54–64, 2013. ISSN 01681699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2013.08.023>.

SHEKHAR, S. et al. Data Models in Geographic Information Systems. *Commun. ACM*, ACM, New York, NY, USA, v. 40, n. 4, p. 103–111, abr. 1997. ISSN 0001-0782. Disponível em: <http://dx.doi.org/10.1145/248448.248465>.

SHEPARD, D. A Two-dimensional Interpolation Function for Irregularly-spaced Data. In: *Proceedings of the 1968 23rd ACM National Conference*. New York, NY, USA: ACM, 1968. (ACM '68), p. 517–524. Disponível em: <http://dx.doi.org/10.1145/800186.810616>.

SILVA, T. et al. Large-scale study of city dynamics and urban social behavior using participatory sensing. *Wireless Communications, IEEE*, v. 21, n. 1, p. 42–51, February 2014. ISSN 1536-1284. Disponível em: <http://dx.doi.org/10.1109/MWC.2014.6757896>.

SNEATH, P. H. The application of computers to taxonomy. *Microbiology*, Microbiology Society, v. 17, n. 1, p. 201–226, 1957. Disponível em: <http://dx.doi.org/10.1099/00221287-17-1-201>.

SOKAL, R. R. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, v. 38, p. 1409–1438, 1958.

SONG, X. et al. The delineation of agricultural management zones with high resolution remotely sensed data. *Precision Agriculture*, Springer US, v. 10, n. 6, p. 471–487, 2009. ISSN 1385-2256. Disponível em: <http://dx.doi.org/10.1007/s11119-009-9108-2>.

SØRENSEN, C. et al. Functional requirements for a future farm management information system. *Computers and Electronics in Agriculture*, Elsevier, v. 76, n. 2, p. 266–276, 2011. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2011.02.005>.

SØRENSEN, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.*, v. 5, p. 1–34, 1948.

SOUSA, R. V. D. et al. Estudo dos elementos mínimos para projeto de sistemas embarcados compatíveis para máquinas e implementos agrícolas. In: INAMASU, R. Y. et al. (Ed.). *Agricultura de Precisão: um novo olhar*. 1a. ed. São Carlos, SP: Embrapa Instrumentação Agropecuária, 2011. p. 126–131. ISBN 978-85-86463-31-0.

SPERANZA, E. et al. A Cluster-Based Approach to Support the Delineation of Management Zones in Precision Agriculture. In: *e-Science (e-Science), 2014 IEEE 10th International Conference on*. [s.n.], 2014. v. 1, p. 119–126. Disponível em: <http://dx.doi.org/10.1109/eScience.2014.42>.

SPERANZA, E. A.; CIFERRI, R. R.; CIFERRI, C. D. A. Clustering Approaches and Ensembles Applied in the Delineation of Management Classes in Precision Agriculture. In: *Proceedings of XVII Brazilian Symposium on GeoInformatics*. Campos do Jordão: MCTIC/INPE, 2016. p. 152–165.

SPERANZA, E. A.; QUEIRÓS, L. R. Organização de Dados Georreferenciados: Estudo de Caso - Rede de Agricultura de Precisão. In: *Congresso Brasileiro de Agricultura de Precisão*. Ribeirão Preto, SP: SBEA, 2010.

SRINIVASAN, A. *Handbook of precision agriculture : principles and applications*. 1st ed. New York, NY: Food Products Press, 2006. ISBN 978-1560229551.

STREHL, A.; GHOSH, J. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, v. 3, p. 583–617, 2002. ISSN 1532-4435.

TANGERINO, G. T. *Método de amostragem de área agrícola com sensores embarcados: uma abordagem que leva em conta a variabilidade do campo*. 131 p. Tese (Doutorado) — Universidade de São Paulo, São Carlos, 2014.

TAYLOR, J. A.; MCBRATNEY, A. B.; WHELAN, B. M. Establishing Management Classes for Broadacre Agricultural Production. *Agronomy Journal*, v. 99, n. 5, p. 1366–1376, set. 2007. ISSN 1435-0645. Disponível em: <http://dx.doi.org/10.2134/agronj2007.0070>.

TING, K. et al. Information Technology and Agriculture Global Challenges and Opportunities. *Bridge*, National Academy of Engineering, 2101 Constitution Ave., NW Washington DC 20418 United States, v. 41, n. 3, p. 6–13, 2011.

TORRE-NETO, A. et al. Wireless Sensor Network for Variable Rate Irrigation in Citrus. In: *Fruit, Nut And Vegetable Production Engeneering Symposium, 7th - Information and Technology for Sustainable Fruit and Vegetable Production*. Montpellier: CEMAGREF, 2005. p. 563–570.

TRAUWAERT, E. On the Meaning of Dunn's Partition Coefficient for Fuzzy Clusters. *Fuzzy Sets Syst.*, Elsevier North-Holland, Inc., Amsterdam, v. 25, n. 2, p. 217–242, fev. 1988. ISSN 0165-0114. Disponível em: <http://dx.doi.org/10.1016/0165-0114(88)90189-3>.

VALENTE, D. S. M. *Desenvolvimento de um sistema de apoio à decisão para definir zonas de manejo em cafeicultura de precisão*. 120 p. Tese (Doutorado) — Universidade Federal de Viçosa, 2010.

VALENTE, D. S. M. et al. Definition of management zones in coffee production fields based on apparent soil electrical conductivity. *Scientia Agricola*, v. 69, n. 3, p. 173–179, jun 2012. ISSN 0103-9016. Disponível em: <http://dx.doi.org/10.1590/S0103-90162012000300001>.

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, Wiley Subscription Services, Inc., A Wiley Company, v. 3, n. 4, p. 209–235, 2010. ISSN 1932-1872. Disponível em: <http://dx.doi.org/10.1002/sam.10080>.

VIASIG: Medir o desempenho do PostGIS. 2017. Disponível em: <http://blog.viasig.com/2015/03/medir-o-desempenho-do-postgis/>. Acesso em: 05 jul. 2017.

VIEIRA, S. R. Geoestatística aplicada à Agricultura de Precisão. In: BORÉM, A. et al. (Ed.). *Agricultura de precisão*. Viçosa, MG: UFV, 2000. p. 93–108.

VIEIRA, S. R. Geoestatistica em Estudos de Variabilidade Espacial do Solo. In: NOVAIS, R. F.; ALVAREZ V H, S. G. R. (Ed.). *Tópicos em ciência do solo*. 1 ed. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, 2000. p. 1–54.

VRETANOS, P. A. *Web feature service implementation specification*. Open Geospatial Consortium Inc., 2005. 94 p. Disponível em: <http://portal.opengeospatial.org/files/?artifact_id=8339>.

WANG, B.; WANG, X. Spatial Entropy-Based Clustering for Mining Data with Spatial Correlation. In: HUANG, J. Z.; CAO, L.; SRIVASTAVA, J. (Ed.). *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2011. (Lecture Notes in Computer Science, v. 6634), p. 196–208. ISBN 978-3-642-20840-9. Disponível em: <http://dx.doi.org/10.1007/978-3-642-20841-6_17>.

WANG, W.; YANG, J.; MUNTZ, R. R. STING: A Statistical Information Grid Approach to Spatial Data Mining. In: *Proceedings of the 23rd International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. (VLDB '97), p. 186–195. ISBN 1-55860-470-7.

WANG, X.-Y.; BU, J. A Fast and Robust Image Segmentation Using FCM with Spatial Information. *Digit. Signal Process.*, Academic Press, Inc., v. 20, n. 4, p. 1173–1182, 2010. ISSN 1051-2004. Disponível em: <http://dx.doi.org/10.1016/j.dsp.2009.11.007>.

WARD JR, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, Taylor & Francis, v. 58, n. 301, p. 236–244, 1963. Disponível em: <http://dx.doi.org/10.1080/01621459.1963.10500845>.

WEISS, S. M.; INDURKHYA, N. *Predictive Data Mining: A Practical Guide*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. ISBN 978-1558604032.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. San Francisco, CA; London: Morgan Kaufmann, 2011. 664 p. ISBN 978-0-12-374856-0.

WU, X. et al. Top 10 Algorithms in Data Mining. *Knowl. Inf. Syst.*, Springer-Verlag New York, Inc., New York, NY, USA, v. 14, n. 1, p. 1–37, dez. 2007. ISSN 0219-1377. Disponível em: <http://dx.doi.org/10.1007/s10115-007-0114-2>.

XIN-ZHONG, W. et al. Determination of management zones for a tobacco field based on soil fertility. *Computers and Electronics in Agriculture*, v. 65, n. 2, p. 168–175, mar. 2009. ISSN 0168-1699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2008.08.008>.

XU, R.; WUNSCH-II, D. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, v. 16, n. 3, p. 645–678, May 2005. ISSN 1045-9227. Disponível em: <http://dx.doi.org/10.1109/TNN.2005.845141>.

XU, X. et al. A distribution-based clustering algorithm for mining in large spatial databases. In: *Data Engineering, 1998. Proceedings., 14th International Conference on*. [s.n.], 1998. p. 324–331. ISSN 1063-6382. Disponível em: <http://dx.doi.org/10.1109/ICDE.1998.655795>.

YANG, T.; BAI, P.; GONG, Y. Spatial data mining features. *Journal of Computational Information Systems*, Binary Information Press, v. 5, n. 3, p. 1503–1510, 2009. ISSN 1553-9105.

ZAIANE, O.; LEE, C.-H. Clustering spatial data in the presence of obstacles: a density-based approach. In: *Database Engineering and Applications Symposium, 2002. Proceedings. International*. [s.n.], 2002. p. 214–223. ISSN 1098-8068. Disponível em: <http://dx.doi.org/10.1109/IDEAS.2002.1029674>.

ZALIK, K. R.; ZALIK, B. A sweep-line algorithm for spatial clustering. *Advances in Engineering Software*, Elsevier, v. 40, n. 6, p. 445–451, jun. 2009. ISSN 0965-9978. Disponível em: <http://dx.doi.org/10.1016/j.advengsoft.2008.06.003>.

ZAMBONI, A. B. et al. StArt - Uma Ferramenta Computacional de Apoio à Revisão Sistemática. In: *Brazilian Conference on Software: Theory and Practice - Tools session*. Salvador: UFBA, 2010.

ZHANG, P.; AIKMAN, S. N.; SUN, H. Two Types of Attitudes in ICT Acceptance and Use. *International Journal of Human-Computer Interaction*, v. 24, n. 7, p. 628–648, 2008. Disponível em: <http://dx.doi.org/10.1080/10447310802335482>.

ZHANG, X. et al. An improved method of delineating rectangular management zones using a semivariogram-based technique. *Computers and Electronics in Agriculture*, Elsevier B.V., v. 121, p. 74–83, 2016. ISSN 01681699. Disponível em: <http://dx.doi.org/10.1016/j.compag.2015.11.016>.

ZHANG, X. et al. Zone mapping application for precision-farming: a decision support tool for variable rate application. *Precision Agriculture*, Springer US, v. 11, n. 2, p. 103–114, 2010. ISSN 1385-2256. Disponível em: <http://dx.doi.org/10.1007/s11119-009-9130-4>.

ZHONG, J.; LIU, L.; LI, Z. Advanced Data Mining and Applications. In: CAO, L.; FENG, Y.; ZHONG, J. (Ed.). *Proceedings of the 6th International Conference on Advanced Data Mining and Applications: Part I*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. (Lecture Notes in Computer Science, v. 6440), p. 302–309. ISBN 978-3-642-17315-8. Disponível em: <http://dx.doi.org/10.1007/978-3-642-17316-5>.

LISTA DE ABREVIATURAS E SIGLAS

AP – *Agricultura de Precisão*

CE – *Condutividade Elétrica do Solo*

CGD – *Classe de Gestão Diferenciada*

CONCAR – *Comissão Nacional de Cartografia*

CPSC – *Constrained Polygon Spatial Clustering*

CSPA – *Cluster-based Similarity Partitioning Algorithm*

DER – *Diagrama Entidade Relacionamento*

EM – *Expectation Maximization*

Embrapa – *Empresa Brasileira de Pesquisa Agropecuária*

FCM – *Fuzzy C-Means*

FMIS – *Farm Management Information System*

FPI – *Fuzzy Performance Index*

GML – *Geography Markup Language*

GPS – *Global Positioning System*

HACC-Spatial – *Hierarchical Agglomerative Clustering with Spatial Constraint*

HGPA – *HyperGraph Partitioning Algorithm*

IDW – *Inverse Distance Weighted*

INPE – *Instituto Nacional de Pesquisas Espaciais*

INPI – *Instituto Nacional de Propriedade Industrial*

ISO – *International Organization for Standardization*

IoT – *Internet of Things*

KDD – *Knowledge Discovery in Databases*

KDSD – *Knowledge Discovery in Spatial Databases*

MCLA – *Meta-Clustering Algorithm*

MER – *Modelo Entidade Relacionamento*

MZA – *Management Zones Analyst*

NCE – *Normalized Classification Entropy*

NDVI – *Normalized Difference Vegetation Index*

OGC – *Open Geospatial Consortium*

OMT-G – *Object Modeling Technique for Geographic Applications*

OMT – *Object Modeling Technique*

OO – *Orientação a Objetos*

OSAVI – *Optimized Soil-Adjusted Vegetation Index*

PCA – *Principal Component Analysis*

Perfil MGB – *Perfil de Metadados Geoespaciais do Brasil*

SGBDE – *Sistema Gerenciador de Bancos de Dados Espaciais*

SGBD – *Sistema Gerenciador de Banco de Dados*

SIG – *Sistema de Informações Geográficas*

SRE – *Sistema de Referência Espacial*

SWMU Clustering – *Spatial Ward's Management Units Clustering*

SWMU Polygon – *Spatial Ward's Management Units Polygon*

StArt – *State of the Art through Systematic Review*

TIC – *Tecnologia da Informação e Comunicação*

UGD – *Unidade de Gestão Diferenciada*

UML – *Unified Modeling Language*

UP – *Unidade Piloto*

UTM – *Universal Transverse Mercator*

VANT – *Veículo Aéreo Não Tripulado*

WFS – *Web Feature Service*

WMS – *Web Map Service*

WPS – *Web Processing Service*

XML – *eXtensible Markup Language*

Apendice A

ABORDAGENS PARA DELINEAMENTO DE UGDS EM AP BASEADAS EM TÉCNICAS EXISTENTES

Nesta apêndice, estão descritos os trabalhos relacionados à aplicação de técnicas de mineração de dados já existentes na literatura para auxílio ao delineamento de UGDs em AP. Conforme já descrito no Capítulo 2, existe uma variedade muito grande de atributos que podem ser capturados em campo agrícola para verificação do comportamento do solo e da cultura. Produtividade histórica, dados topográficos, propriedades físicas do solo, imagens aéreas e de sensoriamento remoto capazes de fornecer índices de vegetação e propriedades da cultura, são exemplos de atributos que podem ser utilizados para esse fim (AGGELOPOOULOU, 2013; MOLIN; AMARAL; COLAÇO, 2015). Por outro lado, atributos sensíveis a intervenções de rotina e ações humanas, como os relacionados à química do solo, devem ser evitados para que as UGDs definidas para uma área de cultivo possam se tornar consistentes ao longo do tempo e capazes de auxiliar efetivamente na gestão da lavoura (MOLIN; AMARAL; COLAÇO, 2015).

Considerando essa variedade de atributos, inicialmente são sumarizadas abordagens aplicadas presentes na literatura cujas UGDs são delineadas a partir de técnicas de geoestatística e algoritmos de agrupamento tradicionais. Essas abordagens foram divididas em seções que identificam o principal tipo de atributo ou técnica utilizada. Em uma última seção, são descritas abordagens que resultaram no desenvolvimento de sistemas de informação específicos para essa aplicação.

A.1 UGDs por Fertilidade do Solo

Abordagens utilizando dados de fertilidade do solo aplicados ao delineamento de UGDs em AP são facilmente encontradas na literatura. Entretanto, a utilização única e exclusiva desse tipo de atributo pode dificultar a obtenção de bons resultados. No trabalho de Ortega e Santibáñez (2007), foi realizada uma comparação dos resultados obtidos para o delineamento de UGDs fornecidos por três diferentes abordagens: baseada em agrupamento, utilizando o algoritmo *k-means*; SIPC, baseada na análise de componentes principais; e SICV, baseada em coeficiente de variação. Para tanto, foram utilizados dados de fertilidade do solo em uma área de cultivo de milho no Chile. Os dados foram inicialmente interpolados em uma grade espacial única, a partir do algoritmo da krigagem. Na aplicação das abordagens SIPC e SICV, os dados foram analisados e classificados em quatro CGDs. A abordagem baseada em agrupamento utilizou o algoritmo *k-means*, com $k = 4$. A quantidade preestabelecida de CGDs foi determinada, segundo os autores, levando em consideração a quantidade máxima que pode ser atribuída a equipamentos convencionais para aplicação de insumos. Os resultados fornecidos pelos três métodos foram considerados similares. Entretanto, análises estatísticas comparando os mapas de UGDs obtidos com relação a dados classificados de produtividade e da própria fertilidade do solo mostraram correlações baixas ou inexistentes. Com isso, os autores concluíram que as amostras de solo utilizadas, para esse caso, foram insuficientes para determinar as UGDs, pois não conseguiram expressar corretamente a variabilidade da produtividade nessa área. Como trabalho futuro, foi sugerida a utilização adicional de dados topográficos e de sensoriamento remoto, de maneira a verificar outros fatores que poderiam ajudar a expressar melhor essa variabilidade.

A abordagem de Xin-Zhong (2009), utilizou um algoritmo de agrupamento *fuzzy* e análise de componentes principais em dados referentes à fertilidade e propriedades químicas do solo para delineamento de UGDs. Com a quantidade ótima de CGDs determinada pelo índice FPI, as avaliações dessa abordagem foram realizadas em uma área de cultivo de tabaco, na China. Como resultado, a baixa resolução espacial das amostras gerou dúvidas quanto ao local correto das bordas das UGDs.

Mais recentemente, Cid-Garcia (2013) e Zhang (2016) utilizaram-se de atributos de fertilidade do solo para validar uma nova abordagem que permite o delineamento de UGDs exclusivamente retangulares, com o intuito de facilitar e reduzir o custo das intervenções realizadas em campo. Diferentemente das técnicas tradicionais de agrupamento, essa nova abordagem realiza o delineamento de UGDs com base em programação linear de número inteiro binário. Segundo os autores, os resultados obtidos mostram que a metodologia é eficiente, principalmente para grandes áreas. Entretanto, é possível utilizar apenas uma única propriedade relacionada à ferti-

lidade do solo para delinear um mapa de UGDs. Essa limitação é a principal desvantagem para a utilização dessa abordagem, a ser resolvida em trabalhos futuros.

A.2 UGDs por Processamento de Imagens e Sensoriamento Remoto

Pelo fato de possibilitarem a obtenção de diferentes parâmetros referentes ao solo e às plantas, a partir de diferentes bandas espectrais, as imagens de sensoriamento remoto, principalmente com alta resolução espacial, são frequentemente utilizadas em AP. O trabalho de Song (2009) utilizou imagens QuickBird com 60 centímetros de resolução espacial, em conjunto com dados de fertilidade do solo e produtividade, para avaliar uma abordagem baseada em agrupamento *fuzzy*. As imagens de sensoriamento remoto forneceram à essa pesquisa o índice OSAVI (*Optimized Soil-Adjusted Vegetation Index*), adaptado para monitoramento agrícola. Os experimentos foram realizados em uma estação experimental de AP na China, onde se concluiu que as informações fornecidas pelo OSAVI podem refletir na variabilidade espacial durante o crescimento da cultura, em propriedades do solo e na produtividade.

Inspirada em algoritmos de processamento de imagens para crescimento de regiões, a abordagem desenvolvida por Pedroso (2010) foi validada pelos autores utilizando dados de NDVI obtidos a partir de imagens aéreas capturadas em uma área de videiras, na Espanha. Essa validação foi realizada a partir de uma comparação de eficácia com relação ao algoritmo *k-means* para obtenção de mapas de UGDs. O algoritmo desenvolvido nessa abordagem transforma, inicialmente, cada ponto do conjunto de dados em uma UGD, por meio de uma tesselação de Voronoi. Na sequência, é iniciada a fase iterativa, onde são sempre fundidas as UGDs espacialmente vizinhas e mais similares considerando o espaço de atributos. Essa iteração perdura até que seja atingida uma restrição de parada preestabelecida, com um limiar de variabilidade intraunidade no espaço de atributos. Os resultados comparativos mostraram que o algoritmo desenvolvido gera UGDs mais contíguas espacialmente e com transições mais suaves entre elas, quando comparado aos resultados obtidos pelo *k-means*.

A.3 UGDs por Condutividade Elétrica (CE) do Solo

A CE do solo é considerada em AP como um dos mais importantes indicadores da variabilidade espacial de uma área agrícola, podendo, por exemplo, substituir ou atuar em conjunto com dados de fertilidade do solo. Devido ao crescimento da disponibilidade no mercado de sensores

capazes de capturar dados de CE em campo, trabalhos relevantes utilizando essa medida para auxiliar no delineamento de UGDs em AP começaram a surgir na literatura.

A abordagem proposta por Kitchen (2005), permite o delineamento de UGDs a partir de diferentes combinações de dados de CE e elevação. Essa abordagem propõe uma arquitetura computacional para essa tarefa, contendo: pré-processamento dos dados por meio de interpolação espacial utilizando a *krigagem*; geração de grupos utilizando o algoritmo FCM; e comparação final de mapas de UGDs com relação a mapas de zonas de produtividade, obtidos por meio de dados históricos. A avaliação dessa abordagem foi realizada a partir de dados obtidos em áreas de cultivo de soja e milho nos Estados Unidos. O resultado final mostrou um nível de concordância entre 60% e 70% das UGDs com relação às zonas de produtividade obtidas por processos similares, indicando que dados de elevação e combinações de CE em diferentes profundidades podem se tornar fortes indicativos ao delineamento das UGDs.

O trabalho de Li (2007) utilizou dados de CE em conjunto com dados físicos e químicos do solo, matéria orgânica e indicativos de vegetação das plantas (NDVI) de uma área de produção de algodão, na China. Devido à grande quantidade de variáveis, foi realizada uma análise que resultou na utilização de dois componentes principais capazes de explicar a maior parte da variabilidade do conjunto de dados. Utilizando esses dois componentes como entrada, o algoritmo FCM foi utilizado para delineamento de três CGDs, cujo número foi determinado pela convergência dos valores dos índices FPI e NCE. Como resultado, foram identificadas CGDs bastante claras, evidenciando uma delas como mais propícia ao crescimento das plantas do que as outras duas.

A abordagem de Morari, Castrignanò e Pagliarin (2009) utilizou uma combinação de análise geoestatística multivariada com o algoritmo FCM, para o delineamento de potenciais UGDs para uma área de produção de uvas na Itália. Para avaliação dessa abordagem, foram utilizados dados de CE obtidos nas direções horizontal e vertical, medidas físicas de solo e imagens de alta resolução obtidas a partir de tomografia de resistividade elétrica. Como resultado, ficou demonstrado que, abordagens que utilizam interpolação espacial geoestatística em conjunto com o agrupamento *fuzzy*, são eficazes para o delineamento de UGDs utilizando os tipos de dado disponíveis.

O trabalho de Moral, Terrón e Rebollo (2011) tem como principal contribuição a criação de um modelo, denominado *Rasch*, que sintetiza dados em diferentes unidades de medida em um ambiente analítico único. Esse modelo foi avaliado em uma área de produção na Espanha, a partir de dados de fertilidade e CE do solo medidos em duas profundidades distintas. Nessa avaliação, a sumarização obtida pelo modelo Rasch, em conjunto com técnicas de geoestatística,

permitiu o delineamento de UGDs com configuração similar a obtida pelo FCM utilizando todas as variáveis.

Em trabalhos mais recentes, Scudiero (2013) mostraram, utilizando o algoritmo FCM, que a combinação de atributos de solo obtidos a partir de amostras georreferenciadas com dados de CE podem contribuir para a identificação da variabilidade espacial em uma cultura por meio de mapas de UGDs. O trabalho de Peralta e Costa (2013) avaliou, por meio de técnicas estatísticas, que os atributos de CE do solo podem ser potenciais estimadores de propriedades e nutrientes do solo capazes de auxiliar no delineamento de UGDs. A partir dessas análises, um novo trabalho utilizando dados de CE do solo e altitude foi desenvolvido, com o intuito de delinear UGDs para melhorar o gerenciamento da aplicação de nitrogênio no solo utilizando a metodologia descrita na abordagem KM-sPC (PERALTA, 2015).

A.4 UGDs por Dados de Produtividade e Atributos da Cultura

Uma abordagem para determinação de UGDs a partir de dados históricos de produtividade foi proposta por Jaynes (2003) e aplicada em áreas de produção de milho (JAYNES, 2003) e soja (JAYNES; COLVIN; KASPAR, 2005) nos Estados Unidos. Essa abordagem utiliza um processo de agrupamento dividido em três partes: particionamento, que utiliza o algoritmo *k-means* aplicado aos dados históricos de produtividade para gerar uma espécie de pré-tesselação do espaço de atributos; interpretação, onde são realizadas análises de correção espacial entre os grupos gerados anteriormente, para que possam ser divididos em áreas espacialmente contíguas; e perfilamento, onde é realizada uma análise de descriminantes para verificação de relacionamentos dos grupos obtidos por produtividade com outras variáveis, como atributos de fertilidade e CE do solo. Os resultados obtidos na área de produção de soja, onde foram utilizados dados históricos de produtividade de cinco anos, mostraram que os atributos relacionados ao solo identificam cerca de 80% dos grupos obtidos por produtividade, permitindo que os mesmos também sejam utilizados para a determinação de potenciais UGDs. Entretanto, segundo os autores, novas pesquisas devem ser realizadas para verificar se os relacionamentos entre a produtividade fornecida por essas UGDs com relação à aplicação de insumos são únicos.

Recentemente, novos estudos foram realizados envolvendo não só dados de produtividade histórica, mas também outros atributos relacionados à cultura. O trabalho de Milne (2012) utilizou dados de produtividade histórica de uma cultura de trigo para avaliar três metodologias distintas para o delineamento de UGDs: utilizando suavização espacial por meio de geoestatís-

tica e agrupamento pelo algoritmo *k-means*; uma variação da primeira metodologia, utilizando suavização das similaridades a partir de uma função de covariância; e utilizando agrupamento *fuzzy* para obtenção de CGDs espacialmente coerentes e internamente homogêneas. Os mapas de UGDs foram utilizados posteriormente para prever a aplicação de nitrogênio de maneira espacialmente diferenciada.

Já o trabalho de Aggelopooulou (2013) utilizou, além de propriedades do solo, dados de produtividade histórica e qualidade do fruto para o delineamento de UGDs em um pomar de maçãs, a partir de uma análise geoestatística multivariada e de um algoritmo de agrupamento não-paramétrico baseado em densidade. Os autores constataram que, apesar de a metodologia adotada ser capaz de gerar mapas de UGDs, a variabilidade interna presente nas CGDs obtidas é bastante grande, mostrando a necessidade de novas pesquisas e estudos para que os produtores da área estudada possam efetivamente adotar a AP.

Finalizando, Chang (2014) utilizaram dados de NDVI obtidos a partir de um sensor de campo em conjunto com dados de produtividade histórica e atributos do solo para delinear UGDs para uma área de cultivo de tabaco. A variabilidade espacial dos atributos foi identificada com a utilização de métodos geoestatísticos e o agrupamento foi realizado com a utilização do FCM. Os resultados obtidos mostraram o alto potencial do sensor de campo utilizado em identificar a variabilidade espacial da cultura.

A.5 UGDs por Sobreposição de Camadas

A abordagem de Franzen e Nanna (2003) utilizou dados de CE do solo, produtividade, topografia, imagens aéreas e de satélite e uso do solo para delinear mapas de UGDs que permitam diferenciar espacialmente o nível de nitrato presente no solo. Para tanto, diferentes combinações dessas variáveis foram correlacionadas com valores obtidos de nitrato presente no solo. O subconjunto de variáveis contendo topografia, imagens de satélite e mapas de produtividade produziu a correlação mais consistente com os dados de nitrato. A partir dessas variáveis, as UGDs foram obtidas utilizando uma simples sobreposição de camadas com pesos associados. Os resultados obtidos por essa abordagem mostram que a sobreposição de camadas realizada é totalmente dependente da configuração dos pesos.

Já a abordagem de Derby, Casey e Franzen (2007) realizou duas abordagens similares para delineamento de UGDs. Primeiramente, com base em dados contendo níveis de nitrato no solo, condutividade elétrica do solo e produtividade, foram utilizados algoritmos de agrupamento para obtenção de quatro CGDs. Na sequência, um método de sobreposição de camadas uti-

lizando pesos foi desenvolvido, utilizando dados de elevação, condutividade elétrica do solo e média de cinco anos de produtividade. Apesar de a configuração de UGDs obtida pelos dois métodos ser similar, a comparação com mapas potenciais de produtividade não é clara o bastante para justificar a sua utilização.

A.6 Aplicativos para Delineamento de UGDs em AP

A grande variedade de abordagens presentes na literatura aplicadas à tarefa de delineamento de UGDs fez com que surgissem alguns aplicativos com o objetivo de auxiliar os usuários em sua execução a partir de uma interface única. Um dos aplicativos mais utilizados por essas abordagens é o MZA (*Management Zones Anayst*) (FRIDGEN, 2004). Esse aplicativo realiza o processo de agrupamento de conjuntos de dados espaciais a partir da utilização do algoritmo FCM no espaço de atributos. Algumas características importantes do MZA são: a possibilidade de escolha, por parte do usuário final, de qual métrica de distância será utilizada no processo; prover diversas configurações de UGDs, utilizando diferentes quantidades de grupos, para que o usuário final possa escolher a mais adequada; e fornecer ao usuário final estatísticas descritivas e resultados obtidos com a utilização dos índices FPI e NCE. Entretanto, algumas limitações fazem com que a adoção do MZA como ferramenta exclusiva para o delineamento de UGDs em AP se torne mais difícil atualmente. Dentre elas, destaca-se o fato de esse aplicativo ter sido desenvolvido exclusivamente para instalação e utilização em sistemas operacionais *Windows*. Além disso, adota um formato de arquivo exclusivo para entrada de dados, exigindo que as variáveis do espaço de atributos já estejam acomodadas em uma grade espacial única. Outra limitação importante é a indisponibilidade de acesso direto à bases de dados ou serviços *Web* para recuperação de dados espaciais, que é uma característica quase que indispensável atualmente para esse tipo de aplicação.

De modo a resolver boa parte dessas limitações, permitindo o acesso a ferramentas para delineamento de UGDs via *Web*, Zhang (2010) desenvolveram o aplicativo *ZoneMap*. O principal fator que motivou o desenvolvimento do *ZoneMAP* foi a baixa adoção da AP entre os seus potenciais usuários, devido a fatores como: alto custo; falta de percepção dos benefícios; conservadorismo; e dificuldade de acesso às tecnologias. Além disso, os pacotes disponíveis em SIGs para esse tipo de aplicação são difíceis de utilizar, e requerem tempo de aprendizado. Adicionalmente, ferramentas específicas como o MZA possuem diversas limitações, conforme já citado anteriormente. Por conta desses problemas, o principal foco para o desenvolvimento do *ZoneMAP* foi a criação de uma ferramenta simples, de fácil acesso a dados de sensoriamento remoto, e que permitisse aos usuários informar dados vetoriais coletados em sua região

de interesse.

A base de dados de sensoriamento remoto originalmente utilizada pelo *ZoneMap* possui 30 anos de imagens históricas limitadas a áreas específicas dos Estados Unidos. A partir dessas imagens, esse aplicativo gera índices úteis ao delineamento de UGDs, como o NDVI. Independentemente disso, a possibilidade de *upload* de dados vetoriais pelos usuários permite que sejam realizadas análises e delineadas UGDs para qualquer área do globo terrestre. O *ZoneMAP* utiliza em sua abordagem de agrupamento o algoritmo FCM aplicado ao espaço de atributos, onde a quantidade de CGDs obtidas é determinada por um método definido pelos autores. Com relação à medida de similaridade, ao contrário do MZA, o *ZoneMAP* utiliza sempre a distância diagonal, definida após experimentos realizados pelos autores em algumas bases de dados. Nesses experimentos, apesar de apresentar desempenho ligeiramente inferior à distância de Mahalanobis, o tempo extremamente menor de execução do FCM utilizando a distância diagonal foi determinante para a sua escolha, principalmente por se tratar do uso em uma aplicação *Web*. Os usuários do *ZoneMap* podem realizar o *download* dos mapas de UGDs obtidos em diferentes formatos de arquivos. Entretanto, após ser disponibilizado gratuitamente por 12 anos seguidos, o *ZoneMap* teve seus serviços interrompidos por conta de falhas de *hardware*.

O aplicativo SDMU (*Software* para Definição de Unidades de Manejo) foi desenvolvido por Bazzi (2011, 2012), com o intuito de permitir a execução de todas as etapas necessárias para definir e validar UGDs pelo usuário final, a partir de ferramentas livres e de distribuição gratuita. A partir dessa aplicação, desenvolvida com a utilização da linguagem Java e, portanto, independente de plataforma e sistema operacional, é possível, utilizando um conjunto de dados espaciais de entrada fornecido pelo usuário final, realizar análises de correlação espacial entre os atributos e interpolação espacial utilizando algoritmos conhecidos. Nesse aplicativo, os mapas de UGDs podem ser gerados de duas maneiras: a partir da utilização métodos empíricos que se utilizam de dados históricos de produtividade normalizados ou padronizados; ou a partir dos algoritmos de agrupamento *k-means* ou FCM. Um módulo de avaliação de UGDs permite validar os mapas obtidos a partir de análises de eficiência relativa e comparação de médias. Segundo Bazzi (2011), o SDUM permite gerenciar e armazenar dados espaciais de maneira hierárquica, facilitando a sua manipulação, seleção e apresentação. Além disso, o *software* possibilita ao usuário final analisar se um mapa de UGDs deve ser utilizado ou não para recomendação de aplicações agrícolas, além da escolha da melhor subdivisão a ser utilizada.

Diversos trabalhos relacionados ao delineamento de UGDs tem sido realizados com o apoio do aplicativo SDUM. Em Bazzi (2013), foram utilizados atributos relacionados à propriedades físicas e químicas do solo, além da produtividade histórica, para delineamento de UGDs em uma

área de produção de soja utilizando o algoritmo FCM. Mais recentemente, Schenatto (2017) utilizaram o aplicativo SDUM para avaliar a eficiência da utilização da experiência do produtor rural na definição das UGDs, em comparação com atributos estáveis de solo e relevo.

Finalizando, Valente (2010) desenvolveu um sistema de apoio à tomada de decisão para delineamento de UGDs em AP, com foco na cafeicultura. Esse sistema possui formulários específicos para as diversas etapas do KDSD necessárias para a execução dessa tarefa, permitindo ajustes de semivariogramas e interpolação espacial utilizando o algoritmo da krigagem; utilização do algoritmo FCM para agrupar amostras; índices como o FPI e a Entropia de Partição Modificada (similar ao NCE) para encontrar a quantidade ideal de CGDs; e o coeficiente Kappa para verificar a correlação entre os mapas de UGDs obtidos e propriedades do solo. Esse sistema foi utilizado em uma aplicação prática para validar a utilização de dados de CE do solo e altitude no delineamento de UGDs para uma lavoura de café (VALENTE, 2012).

A.7 Sumário das Abordagens para Delineamento de UGDs em AP Baseadas em Técnicas Existentes

As abordagens descritas neste apêndice mostram que diferentes tipos de atributos relacionados ao solo e à cultura podem ser utilizados para auxiliar na tarefa de delineamento de UGDs AP. Deve-se destacar o uso inicial de atributos medidos a partir de análises de solo, evoluindo para a utilização de dados oriundos de sensores de campo, até chegar a abordagens que utilizam índices provenientes de imagens de sensoriamento remoto ou aéreas. Além disso, o uso de abordagens tradicionais como as que utilizam o algoritmo FCM, mesmo que considerando apenas o espaço de atributos, aparenta ser um consenso entre os usuários especialistas. O fato de não considerar diretamente o espaço de coordenadas durante esse processo, negligenciando relacionamentos espaciais importantes que podem ocorrer entre as amostras, faz com que grande parte dessas abordagens possam se mostrar eficazes em determinados conjuntos de dados, e em outros não. Consequentemente, essas abordagens podem proporcionar a obtenção de mapas de UGDs estratificados e de difícil interpretação pelos usuários especialistas.

Apendice B

ANÁLISE DAS POSSIBILIDADES E TENDÊNCIAS DO USO DAS TECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO EM AGRICULTURA DE PRECISÃO

Este apêndice contém a transcrição do capítulo intitulado *Análise das Possibilidades e Tendências do Uso das Tecnologias da Informação e Comunicação em Agricultura de Precisão*, publicado no livro técnico-científico *Agricultura de Precisão: resultados de um novo olhar* (QUEIRÓS, 2014).

Análise das possibilidades e tendências do uso das tecnologias da informação e comunicação em Agricultura de Precisão

Leonardo Ribeiro Queirós*[1], Ariovaldo Luchiari Junior*[2], João Camargo Neto*[3], Sílvia Maria Fonseca Silveira Massruhá*[4], Ricardo Yassushi Inamasu*[5], Eduardo Antonio Speranza*[6], Silvio Roberto Medeiros Evangelista*[7]

[1]Analista Dr., Embrapa Informática Agropecuária
[2]Pesquisador Dr., Embrapa Informática Agropecuária
[3]Analista Dr., Embrapa Informática Agropecuária
[4]Pesquisadora Dra., Embrapa Informática Agropecuária
[5]Pesquisador Dr., Embrapa Instrumentação Agropecuária
[6]Analista doutorando em Ciência da Computação, Embrapa Informática Agropecuária
[7]Analista Dr., Embrapa Informática Agropecuária

E-mails: leonardo.queiros@embrapa.br, ariovaldo.luchiari@embrapa.br, joao.camargo@embrapa.br, silvia.masshura@embrapa.br, ricardo.inamasu@embrapa.br, eduardo.speranza@embrapa.br, silvio.evangelista@embrapa.br

Resumo: A Agricultura de Precisão (AP) tem em sua concepção a emergência de novas combinações agrotecnológicas baseadas no desenvolvimento e aplicação das tecnologias da informação e comunicação na agricultura. As Tecnologias da Informação e Comunicação (TIC's) são definidas pela Agência dos Estados Unidos para Cooperação Internacional-USAID como sendo: a combinação de hardware, software e os instrumentos de produção que permitem troca, processamento e manejo da informação e do conhecimento. Então, de acordo com a USAID, as TIC's incluem tecnologias e métodos para armazenar, manejar e processar informação (e.g. computadores, softwares, livros, dispositivos móveis, livrarias digitais e não digitais) e para comunicar a informação (e.g. correio, correio eletrônico, rádio, televisão, telefones, celulares, pagers, internet, etc). Devido à combinação de agrotecnologias com as tecnologias da informação e da comunicação, a Agricultura de Precisão é vista atualmente como uma das formas mais eficientes e eficazes de se garantir a produção de alimentos para atender as necessidades alimentares de nove bilhões de habitantes da terra em 2050, com a garantia da qualidade do produto e dos recursos naturais bióticos e abióticos. Este capítulo analisa o estado da arte e as tendências futuras das Tecnologias da Informação e Comunicação no contexto da Agricultura de Precisão. Serão abordados os seguintes temas: Padrões para Integração de Equipamentos Agrícolas, Sistemas de Informação e na Automação de Processos e Operações Agrícolas; Computação Ubíqua e em Nuvem; Aplicações Geoespaciais; Sistemas de Suporte a Decisão; Processos Produtivos Agrícolas - Protocolos e Normas de Produção.

Palavras-chave: Agricultura de Precisão, Tecnologia da Informação e Comunicação, Computação em Nuvem e Ubíqua, ISOBUS, Sistemas de Suporte a Decisão, Protocolo de Produção.

Analysis of the Possibilities and Future Trends in the Use of Information and Communication Technologies in Precision Agriculture

Abstract: *Precision Agriculture (PA) has embedded in its conception news agro-technological combinations based on the use of the Information and Communication Technologies. In this chapter the USAID's definition of information and communication technology will be used, i.e., "the combination of hardware, software, and the means of production that enable the exchange, processing, and management of information and knowledge". ICTs thus include technologies and methods for storing, managing, and processing information (e.g., computers, software, books, mobile devices, tablets, androids, digital and non-digital libraries) and for communicating information (e.g., mail and email,*

radio and television, telephones, cell phones, pagers, instant messaging, "the web," etc.) Due the combination of agricultural, information and communication technologies, Precision Agriculture has been seen as the most effective and efficient form of agricultural production able to feed 9 billion people in 2050, while maintaining the safety and quality of the product in harmony with the biotic and non-biotic natural resources. This Chapter analyses the state-of-the art and future trends of ICT´s within the context of Precision Agriculture. The following themes will be covered: Standards for the Integration Agricultural Machinery, Information Systems and Automation of Agricultural Processes and Operations; Ubiquitous and Cloud Computing; Geo-spatial Applications; Decision Support Systems; The Role of AP and TIC´s in Attending Agricultural Production Standards, Safety and Traceability.

Keywords: *Precision Agriculture, Information and Communications Technologies, Cloud and Ubiquitous Computing; ISOBUS, Decision Support Systems, Agricultural Production Protocols.*

1. Introdução

A agricultura convencional, principalmente com a produção em larga escala, fez com que a gestão da lavoura intuitiva, que tratava as diferenças do campo, fosse dissimulada. As novas tecnologias, como GNSS e Sistemas de Informação, trouxe a viabilidade operacional para tratar essas diferenças, inovando a nossa lavoura.

A AP, por ter inserido em sua concepção a emergência de novas combinações agrotecnológicas, baseadas no desenvolvimento e na aplicação das tecnologias da informação e comunicação (TIC's) na agricultura, com possibilidades de ganhos econômicos e benefícios ambientais, vem ganhando popularidade mundial (WOLF; WOOD 1997). Essa nova forma de produção agrícola tem atraído, desde o início de sua adoção, o interesse de formuladores de políticas públicas de pesquisa, de ensino e de desenvolvimento econômico e social; das indústrias de telecomunicações e informática; da mídia; das instituições de crédito e seguro rural; e também dos setores tradicionais do agronegócio - indústrias de insumos, máquinas e processamento –(WOLF; WOOD, 1997) (SCHEPERS; SHANAHAN; LUCHIARI JÚNIOR, 2000). Atualmente é vista como uma das formas mais eficientes e eficazes de garantir a produção de alimentos para atender as necessidades alimentares de nove bilhões de habitantes da terra em 2050.

Uma vez que a agricultura de precisão tem se beneficiado da utilização das tecnologias da informação e comunicação na agricultura, é importante lembrar que as TIC's são definidas, pela Agência dos Estados Unidos para Cooperação Internacional - USAID, como sendo: a combinação de hardware, software e os instrumentos de produção que permitam a troca, o processamento e o manejo da informação e do conhecimento. De acordo com a USAID, as TIC's incluem tecnologias e métodos para armazenar, manejar e processar informação (e.g. computadores, softwares, livros, PDAs, tablets, androides, livrarias digitais e não digitais) e para comunicar a informação (e.g. correio, correio eletrônico, rádio, televisão, telefones, celulares, pagers, internet, etc. Nesse contexto, as TIC's aqui são vistas como desempenhando as seguintes funções (RUSTEN; RAMIREZ, 2003) :1 - que o conhecimento tecnológico é um componente importante para o desenvolvimento do setor agrícola; 2– que as TIC's aceleram o desenvolvimento do setor por organizar e facilitar a organização e a transferência do conhecimento entre os atores que atuam no setor e 3 - que as organizações terão um papel fundamental na identificação de necessidades de métodos adequados de manejo e de tomada de decisões e na identificação de novas necessidades tecnológicas para que o uso das TIC´s em AP seja mais eficaz, eficiente e mais fácil de ser usado. Este capítulo analisa o estado da arte e tendências futuras das Tecnologias da Informação e Comunicação no contexto da Agricultura de Precisão. Serão abordados os seguintes temas: Padrões para Integração de Equipamentos Agrícolas, Sistemas de Informação e na Automação de Processos e Operações Agrícolas; Computação Ubíqua e em Nuvem; Aplicações Geoespaciais, Sistemas de Suporte a Decisão, Uso TIC em Processos Produtivos Agrícolas - Protocolos e Normas de Produção, Uso de Padrões para Integração de TIC em Equipamentos Agrícolas e Uso de TIC´s na Automação de Processos e Operações Agrícolas.

2. Uso de Padrões para Integração de TIC em Equipamentos Agrícolas

Nas últimas décadas, a Agricultura de Precisão tem se beneficiado com a automação de máquinas e implementos agrícolas por meio do uso de sistemas eletrônicos embarcados compostos por programas de computadores e dispositivos eletrônicos e de hardwares. No início, os fabricantes desses sistemas se preocupavam com a confiabilidade, facilidade de instalação e de uso. O problema é que pouca atenção foi dada para que esses sistemas fossem facilmente integrados com outros disponíveis produzidos por outros fabricantes (HASSALL, 2010). Dessa forma, inúmeros sistemas foram disponibilizados para o mercado com protocolos proprietários de comunicação, de forma que não havia compartilhamento de informação entre eles. Além disso, cada sistema demandava um terminal para interação do usuário operador da máquina com suas funcionalidades de controle, de forma que dentro da máquina agrícola, instrumentada com esses sistemas, existiam vários terminais e um grande emaranhado de cabos, que contribuíam para um ambiente não otimizado e complexo de interação homem-máquina.

Para suprir essa necessidade de integração dos diferentes sistemas eletrônicos embarcados, padrões de redes de comunicação têm sido desenvolvidos. Destaca-se o esforço entre diversos países, coordenado por Forças Tarefas da Europa e dos Estados Unidos, para a geração e aplicação de uma norma internacional denominada ISO 11783, também conhecida industrialmente por ISOBUS, para tratar essa falta de interoperabilidade. Ela é baseada no protocolo de comunicação digital serial "Controller Area Network" e segundo Saraiva e Cugnasca (2006) especifica uma rede serial para comunicação e controle de veículos agrícolas, como tratores e seus implementos, de modo a tornar disponível uma padronização para sistemas embarcados em máquinas e equipamentos agrícolas. Essa padronização permite o uso de apenas um terminal para reconhecer, monitorar e gerenciar automaticamente os implementos, compatíveis como o padrão ISOBUS, conectados ao trator (SOUSA et al., 2011).

Para a Agricultural Industry Electronics Foundation - AEF (AGRICULTURAL..., 2013) - uma organização internacional composta por mais de 150 empresas, associações e organizações - os fabricantes de equipamentos agrícolas em todo o mundo elegeram ISOBUS como o protocolo universal para comunicação eletrônica entre implementos, tratores e computadores. A AEF tem mantido um banco de dados acessível pela Web com os equipamentos compatíveis com o ISOBUS e funcionalidades que permitem selecionar uma combinação de equipamentos e verificar a compatibilidade entre as funções oferecidas. Segundo a Força Tarefa ISOBUS Brasil - FIT Brasil - o emprego de sistemas eletrônicos embarcados em máquinas agrícolas em consonância com essa norma tem sido restrito a produtos importados. Porém, o grupo do FIT Brasil tem buscado criar competência no País por meio de domínio das tecnologias envolvidas e divulgar o benefício do padrão (FORÇA..., 2013).

A tendência é que os sistemas eletrônicos embarcados em máquinas agrícolas estejam em consonância com essa norma e cada vez mais presentes na área agrícola.

3. Uso de Padrões em TIC para Armazenamento de Dados, Intercâmbio e Interoperabilidade entre Sistemas de Informação

A imensa quantidade de dados digitais produzidos pelo uso de tecnologias da AP está armazenada em diferentes formatos e padrões de arquivos, em diferentes sistemas de informação, sem muita atenção para a documentação mínima de informações que facilitam a recuperação e entendimento desses dados. O problema se torna mais visível quando é necessário reutilizar esses dados ou integrar diferentes sistemas de informação para uma análise mais apurada, a qual, muitas vezes, é inviabilizada pela falta de descrição do procedimento usado na coleta dos dados, falta de unidade de medida ou até mesmo impossibilidade de identificar a qual variável um conjunto de medidas está associado.

Assim como o problema de integração de equipamentos agrícolas tem sido resolvido com os esforços direcionados ao ISOBUS, é preciso caminhar na mesma direção para padronização de armazenamento de dados e arquitetura de sistemas de informação distribuídos que permitam

a integração desses dados, de forma simples e transparente. O projeto europeu FutureFarm (http://www.futurefarm.eu/) produziu uma especificação para um sistema de informação de gestão agrícola com atenção para essas questões. Nessa especificação, todos os dados devem ser documentados e armazenados na linguagem padronizada para troca de dados agroXML e a arquitetura distribuída deve ser a arquitetura SOA (*Service-Oriented Architecture*) (BLACKMORE; APOSTOLIDI, 2011). Nos Estados Unidos a AgGateway, uma organização sem fins lucrativos que tem por visão ser reconhecida internacionalmente por promover o uso das TIC's na Agricultura de Precisão, lançou o projeto 'padronizando o intercâmbio de dados da AP (SPADE) '. O projeto visa atender as demandas dos produtores no sentido de tornar mais amigável o uso de equipamentos e aplicativos em AP (AGGATEWAY,2013).

No Brasil, a Rede de Agricultura de Precisão da Embrapa - Rede AP - atenta à necessidade de adoção de padrões para armazenamento e intercâmbio de dados e informações, e de uma arquitetura orientada a serviços que permita a interoperabilidade entre sistemas, mantendo a memória, a preservação, a recuperação e o intercâmbio com qualidade dos dados produzidos pelas unidades pilotos, desenvolveu um repositório de recursos de informação (https://www.redeap.cnptia.embrapa.br) que usa o perfil de metadados 'Perfil de Metadados Geoespaciais do Brasil' - versão homologada em 2009 pelo Comitê de Planejamento da Infraestrutura Nacional de Dados Espaciais (CONCAR) - para catalogar os dados geoespaciais e com arquitetura que permite a integração e interoperabilidade de aplicações. Na Figura 1a, é mostrada a estrutura banco de dados e sua integração com a camada de aplicação. Os recursos de informação digitais suporta os formatos - shapefile, raster, txt, doc,xls, jpg e pdf - e estão associados a elementos de metadados. Já a camada de aplicação é composta pela integração de ferramentas de softwares livre - banco de dados PostgreSQL, WebGIS i3Geo e o aplicativo para catalogação de dados geoespaciais GeoNetwork - e de conversores de dados dos equipamentos de sensoriamento usados pela Rede AP, com a função de realizar a interface com os usuários. Na Figura 1b, são ilustrados os elementos de metadados selecionados, customizados, criados e em uso, para a catalogação de dados geográficos e não geográficos (somente tabulares). Nesse diagrama, os elementos de metadados foram agrupados numa generalização e especialização. A generalização do diagrama representa os elementos de metadados que devem ser preenchidos, independente do tipo de dado ser tabular ou geográfico (CDG). Já a especialização expressa os elementos de metadados que devam ser preenchidos somente para o tipo CDG. Vale ressaltar que dois novos elementos foram criados dentro da Seção Identificação do 'Perfil de Metadados Geoespaciais do Brasil': "Observação" para contemplar qualquer observação ou necessidade de documentação que por ventura não possa ser expressa pelos demais elementos de metadados selecionados; e "Responsável pela Catalogação" para identificar o autor de documentação dos metadados. Com relação ao elemento "Observação", a ideia é analisar a frequência de necessidade de uso desse descritor para, posteriormente, eleger ou criar novas seções ou elementos para atender as especificidades de documentação do projeto AP (QUEIROS et al., 2011). O grande avanço conseguido pelo projeto é permitir a obtenção de séries históricas espaciais e temporais de lavouras, sendo elemento chave não só para alimentar as novas necessidades de pesquisa, mas também para rastreabilidade e comparação entre sistemas que adotaram a AP. Considerando que o repositório da Rede AP e seus resultados permitiram o estabelecimento de padrões adequados para operacionalizar, armazenar, recuperar, intercambiar e interoperar os dados e informações obtidas nas unidades pilotos, de forma quantitativa e qualitativa, ele permitirá também que essa experiência seja extrapolada para o manejo de propriedades agrícolas. Esse repositório foi concebido para atender necessidades futuras de organização e tratamento de informação.

4. Computação Ubíqua

O avanço dos sistemas embarcados aliado ao custo decrescente de equipamentos digitais tem sido fecundo para realização de constantes investimentos em infraestrutura de telecomunicações em todo

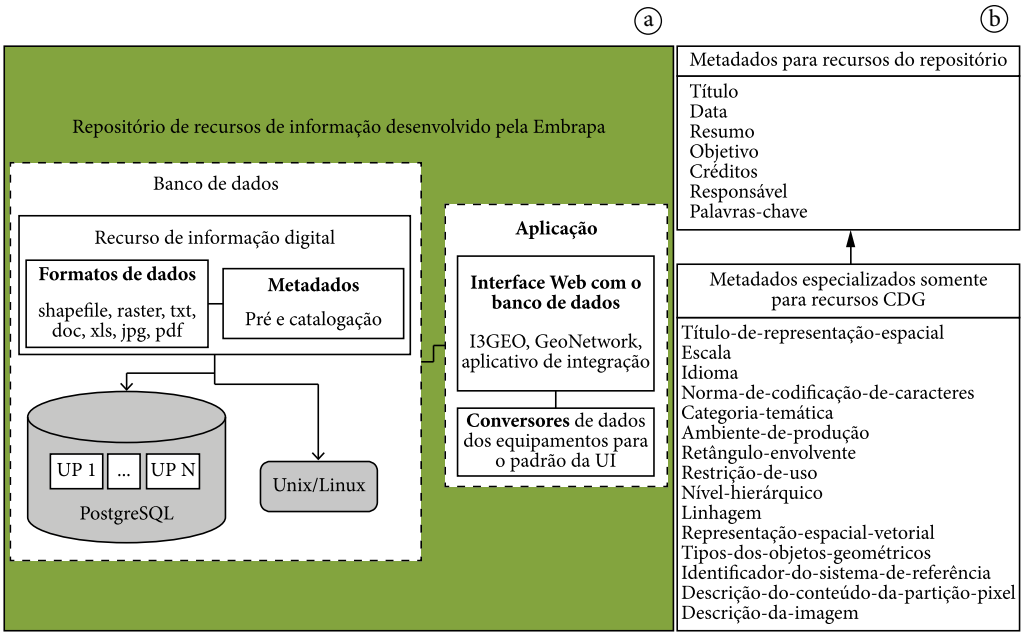


Figura 1. Repositório de recursos de informação desenvolvido pela Embrapa.

mundo (BALLANTYNE; MARU; PORCARI, 2010). Equipamentos como celulares, tablets, computadores pessoais - cada vez mais presentes no dia a dia das pessoas - conectados à Internet traz uma grande oportunidade de conectividade entre a ciência, produtores e demais atores relacionados ao contexto da Agricultura. Essa conectividade é facilitada quanto mais simples, autônomos e imperceptíveis forem os sistemas embarcados e equipamentos associados. A busca por não notoriedade da presença de computadores entre humanos, por meio da simplicidade de operação e maximização do funcionamento autônomo, tem sido conhecida por computação ubíqua. Torre Neto (2009) aponta como concepção da computação ubíqua a fusão dos computadores com o ambiente, a ponto de tornarem-se invisíveis para os usuários.

As tecnologias da AP tem se beneficiado dessa conectividade, em especial conectividades por meio de redes sem fio, e da computação ubíqua, nas quais sensores, redes de sensores, atuadores e sistemas de controle podem coletar dados, processá-los, realizar atuação e encaminhar informações para um computador servidor na sede da fazenda ou diretamente para algum serviço de nuvem disponível, conforme abordado na seção 'Computação em Nuvem', de forma autônoma e em tempo real. Como exemplo, a tecnologia de piloto automático, amplamente difundida na AP, permite que um veículo agrícola trafegue pela lavoura sem intervenção humana - sendo a coleta de informação dos sensores do motor, direção, localização espacial entre outros e a atuação na direção realizada de forma transparente e automática. Ainda nesse exemplo, o agricultor poderia monitorar em tempo real a rota realizada por meio de um aplicativo instalado em um tablet em qualquer lugar do mundo (HEST, 2013). Esses equipamentos, por atuarem de forma transparente, auxiliam o produtor a reduzir os erros e, portanto, reduz a variabilidade espacial antrópica e natural do campo.

No Brasil, a Rede de Agricultura de Precisão da Embrapa, tem abordado o monitoramento de controle de processos na agropecuária através do uso das inovadoras tecnologias de rede de sensores sem fio e da computação ubíqua, por meio das seguintes atividades de pesquisa: (i) a irrigação espacialmente diferenciada; (ii) a pulverização de precisão; (iii) o mapeamento da fertilidade do solo; (iv) a rastreabilidade animal e vegetal e (v) as mudanças climáticas e os problemas fitossanitários (TORRE NETO, 2009).

A conectividade tem se tornando cada vez mais pervasiva e móvel e mais dispositivos estão se tornando capazes de realizar múltiplas operações (BALLANTYNE; MARU; PORCARI, 2010)

em consonância com a computação ubíqua, de forma a ser uma tendência consolidada e cada vez mais refletida nas tecnologias da AP que serão produzidas no futuro.

5. Computação em nuvem

A geração de dados, em alta resolução, contínua e, muitas vezes, em tempo real, por meio do uso de tecnologias da AP ou por grandes bancos de dados públicos, disponíveis na Internet, com informações agrícolas em macro-escala, necessárias para a gestão das principais operações de cultivo - preparo do solo, semeadura, adubação, irrigação, pulverização e colheita - tem demandado uma crescente capacidade de armazenamento e processamento computacional que extrapola a capacidade de computadores pessoais alocados numa fazenda, trazendo à AP os desafios associados às pesquisas em Big Data. Soma-se, ainda, a necessidade do uso de procedimentos computacionais inerentes a um 'Data Center' para realizar o backup dos dados, instalação de programas de processamento, manutenção de rede cabeada ou sem fio para a transmissão, manutenção da rede elétrica, atualização de sistemas operacionais, entre outros que exigem uma dedicação integral de profissional com habilitação em TI e investimentos em hardware e software. Outra questão é que os produtores não têm recursos para manter seu próprio departamento de TI (WELTE et al. 2013).

Diante dessa realidade, muitas empresas ativas mundialmente no provimento de soluções para a AP têm oferecido serviços baseados na computação em nuvens que encapsulam toda a infraestrutura e gestão computacional de um Data Center e as oferecem como serviços disponíveis para acesso por meio da Internet (HEST, 2013) e (BALLANTYNE; MARU; PORCARI, 2010). A computação em nuvens provê serviços de acordo com três categorias distintas (KEPES, 2013): infraestrutura como serviço - servidores, rede, máquinas virtuais, armazenamento, balanceamento de carga, entre outros; plataforma como um serviço - banco de dados, ambiente de execução, servidor web, ferramentas de desenvolvimento, entre outros; e software como serviço disponibilizado para usuários finais e acessíveis pela Web - navegadores de Internet, aplicações para dispositivos móveis, sistemas embarcados em máquinas agrícolas, sistemas de suporte a decisão, e-mail, sistemas de informações geográficas, entre outros. Ballantyne, Maru e Porcari (2010) explica que já existem serviços que permitem ao usuário ter centenas ou milhares de computadores a sua disposição e ainda pagar por eles, por hora ou minuto, sem a necessidade de aquisição ou administração do hardware; a computação nas nuvens elimina a barreira de capacidade de processamento e os custos são bem menores, mesmo diante da queda de preço de hardware, uma vez que os custos de um 'Data Center' pode ser compartilhado entre vários usuários.

Como relatado por Hest (2013), algumas empresas já oferecem soluções nas nuvens, na qual equipamentos agrícolas estão conectados por rede sem fio e as informações são disponibilizadas em tempo real e acessíveis por navegadores de Internet ou por aplicativos instalados em dispositivos móveis. É possível, ainda, contar com informações processadas na nuvem por sistemas de suporte a decisão de maneira eficiente.

Trata-se de um grande atrativo a utilização de softwares sem que esses estejam instalados no computador e não ter que mantê-los ou se preocupar com infraestrutura e plataforma computacional, porém, algumas desvantagens são observadas: A falta de conectividade com a nuvem (Internet ou nuvens configuradas) pode comprometer a execução e a visualização das informações mantidas pelos sistemas associados; e o estabelecimento de uma clara política de propriedade e acesso aos dados armazenados na nuvem.

Um serviço brasileiro disponível para uso na AP não foi encontrado para avaliação de seu uso no país, porém, há softwares proprietários que se baseiam nas nuvens traduzidos para a língua portuguesa e que já estão disponíveis para uso. Com relação aos serviços de infraestrutura e plataforma existem empresas de propósitos gerais que tem oferecido serviços no país.

Uma vez que exista largura de banda suficiente para suportar a transferência de dados, e, as questões de política de propriedade e acesso aos dados armazenados estejam regulamentadas e bem asseguradas, o fornecimento e uso de serviços agrícolas nas nuvens para a AP tende a aumentar fortemente nos próximos anos.

6. Aplicações Geoespaciais

A principal hipótese para a adoção das tecnologias de Agricultura de Precisão é a existência da variabilidade espacial no campo. Dentre as várias tecnologias atualmente disponíveis para mapear esta variabilidade encontra-se o imageamento aéreo. Essa tecnologia traz embutida a característica espacial na qual cada pixel da imagem corresponde a uma amostragem espectral de uma região única no solo. Esse imageamento pode ser realizado por satélites orbitais, balões, aviões tripulados e, atualmente, pelos veículos autônomos não tripulados - VANT. Independente da tecnologia utilizada na obtenção das imagens, o objetivo final é correlacionar a características do objeto em estudo, no caso planta, solo e resíduo com reflectância espectral e temperaturas emitidas por esses objetos e registradas nas bandas espectrais das imagens.

Um dos primeiros estudos utilizando imageamento por satélite foi realizado com imagens do satélite Landsat 1, lançado em 23 de julho de 1972. Esse estudo teve como objetivo examinar as diferenças na vegetação no período da primavera e do verão na região das Grandes Planícies dos Estados Unidos correlacionando e quantificando as características biofísica da vegetação com as respostas espectrais. Como resultado desse estudo várias relações entre as bandas espectrais (índices) foram estudadas, sendo que o mais bem sucedido e utilizado até os dias de hoje é o NDVI (Índice de Vegetação por Diferença Normalizada). Esse índice tem sido usado em sensoriamento remoto para quantificar e monitorar o vigor das plantas, cobertura vegetal e produção de biomassa.

O estudo realizado por Moran, Inoue e Barnes (1997) descrevem as oportunidades na utilização de bandas espectrais e índices obtidos das imagens multispectrais em agricultura de precisão, tais como; utilização de imagens multispectrais adquiridas no período anterior a colheita para mapeamento de produtividade, do solo nu ou completamente coberta por vegetação para mapear variabilidade espectral e durante o crescimento da cultura, para monitorar variabilidade das condições da mesma.

Nessa linha de pesquisa outros estudos têm sido realizados utilizando estes índices obtidos das imagens multispectrais para detectar, georreferenciar e mapear regiões de variabilidade (LUCHIARI JUNIOR et al., 2011) causadas por doenças, deficiência nutricional, stress hídrico que refletem diretamente no vigor da planta, causando um declínio na produção de biomassa; e para mapear níveis de nitrogênio nas plantas que correlacionam resposta espectral com elevados níveis de clorofila e altas taxas de fotossínteses. Dentre esses índices podemos destacar o de vegetação de diferença normalizada na faixa do verde (GNDVI- Green Normalized Difference Vegetation Index) (SHANAHAN et al., 2001) e o de vegetação ajustado do solo - SAVI, (HUETE, 1988; RONDEAUX; STEVEN; BARET, 1996; BARET; GUYOT; MAJOR, 1998).

Recentemente vem crescendo a utilização de imagens multiespectrais adquiridas pelos Veículos Autonomos não Tripulados - VANT - pelas instituições de pesquisas e serviços de imageamento disponíveis no mercado por companias privadas. As vantagens da utilização dessa tecnologia são: aquisição de imagens multiespectrais com alta resolução espacial; custo de obtenção inferior a imagens de satélites ou fotos aéreas; aquisição de imagem a qualquer instante; permitir aquisição em tempo nublado por ser possível realizar voos abaixo da altura das núvens; capacidade de execução de trabalhos repetitivos e perigosos em locais de difícil acesso.

O uso de imagens, por estar diretamente relacionado à automação dos processos e de operações agrícolas; por poder ser usado em pequenas e grandes áreas, e em culturas de alto valor agregado como horticultura, e por seu custo estar em declínio é visto como uma tendência futura em expansão. Entretanto, algumas limitações terão que ser superadas, tais como: capacitação técnica de usuários; seguro para sobrevoos; legislação de uso e quebra de paradigmas tecnológicos.

7. Sistemas de Suporte a Decisão

Ao longo dos anos a Embrapa desenvolveu diversos sistemas especialistas que visam atender a demanda de certos nichos e entidades relacionadas ao negócio agrícola. Dentre eles é possível destacar os sistemas para: monitoramento

agrometeorológico (www.agritempo.gov.br); diagnose virtual de doenças de plantas (http://www.diagnose.cnptia.embrapa.br); previsão de safra de soja; recomendação para adubação; e o WebAgritec.

O WebAgritec é um sistema computacional de acesso e utilização via Web (http://www.agritec.cnptia.embrapa.br/) que agrega e torna disponíveis informações geradas pela pesquisa e permitem ao usuário planejar e conduzir a cultura plantada com as melhores práticas e material genético disponível. Sua finalidade é auxiliar os profissionais ligados ao setor agropecuário na tomada de decisões, para tanto, o sistema conta com 7 módulos (Zoneamento, Cultivar, Adubação, Previsão, Monitoramento, Diagnóstico, Videoteca), que orientam o Usuário desde o planejamento até a condução da cultura. Esses 7 módulos permitem uma visão geral do sistema produtivo.

Embora o resultado final alcançado, nesse primeiro protótipo do WebAgritec, tenha sido satisfatório, tanto do ponto de vista de arquitetura como da aplicação, novas tecnologias de computação móvel são avanços que devem ser contemplados em ações futuras no escopo da Agricultura de Precisão. Tendo em vista que as tecnologias da AP geram uma vasta gama de informações que estão dispersas e não estão sendo diretamente utilizadas no suporte à tomada de decisões do setor produtivo agrícola. Diante desse cenário, fica evidente a necessidade do desenvolvimento de infraestruturas que agreguem o conhecimento tecnológico e tácito gerado pela Agricultura de Precisão que suportem a tomada de decisão, em tempo real, e que facilitem a transferência e capacitação tecnológica, via web e dispositivos móveis, com o propósito de beneficiar os agricultores, os agentes da extensão e assistência técnica pública e privada, agências de fomento, de crédito, etc.

8. Uso de TIC's em Processos Produtivos Agrícolas - Protocolos e Normas de Produção

A Agricultura de Precisão utiliza GPS (Sistema de posicionamento global), GIS (Sistema de informações geográficas), instrumentos e sensores para medidas ou detecção de parâmetros ou de alvos de interesse no agroecossistema (solo, planta, insetos, doenças, etc.), de mapas de colheita, de métodos quantitativos e da mecatrônica. O uso desses conceitos e instrumentos permitem: i) utilizar mapas de colheita e variabilidades no solo e no clima, para diagnosticar as causas das variabilidades, espacial e temporal, quer sejam natural ou induzidas pelo homem, e analisar seus efeitos nas produtividades, ii) aplicar localizadamente os insumos em quantidades variáveis e em tempos específicos quer por taxa variada ou por zonas de manejo, e iii) controlar o manejo das culturas para que os níveis de produtividade pré-estabelecidos sejam atingidos iv) monitorar para que as práticas agrícolas estejam em harmonia com o meio ambiente e v) certificar-se de que os produtos obtidos sejam seguros.

Quando as tecnologias da Agricultura de Precisão são combinadas com as TIC´s, é possível de se obter, armazenar e processar informações que permitam ações de comando e controle da forma de produção. Assim, é possível atender, analisar, monitorar e rastrear a conformidade da produção com os requisitos de vários protocolos e normas, tais como: da Produção Integrada e da Produção Orgânica, do Ministério da Agricultura, da Pecuária e Abastecimento, do GLOBALG.A.P. da Europa, das Produções Étnicas, do Contrato de Produção de Alimentos Funcionais, entre outros.

Entretanto, o uso dessa forma de produção no Brasil não tem sido tão intenso. Furlaneto e Manzano (2010) citam o sucesso do uso de técnicas da agricultura na produção integrada e no processo de rastreabilidade do pêssego.

Em relação ao futuro, o uso de tecnologias e processos da agricultura de precisão para atender protocolos de certificação e rastreabilidade da produção é, ainda, uma incerteza crítica. Entretanto, o repositório concebido na Rede AP, pode transformar numa tendência consolidada com incremento do seu uso, por facilitar a organização e armazenamento de informações requeridas nas análises de conformidade constantes nos protocolos e normas de produção. Consequentemente permitirá que os produtores conquistem novos mercados com garantia de melhores preços, devido à certificação da qualidade, segurança e origem dos produtos.

9. Uso de TIC´s na Automação de Processos e Operações Agrícolas

A integração entre aquisição de dados obtidos por sensores ou por coletas georreferenciadas, TIC's, sistemas de suporte a decisão e de navegação são requisitos para o processo de automação agrícola. Para aplicação desse processo, é necessário que dados e informações obtidas por redes de sensores sem fio ou que dados espaciais e temporais dos agro-ecossistemas sejam tratados por padrões de representação e comunicação (agroXML, ISOBUS entre outros) entre sistemas numa arquitetura computacional distribuída como o SOA. Devido à vasta quantidade de dados e informações obtidas, o processamento e análise em infraestruturas de alto desempenho computacional como a computação em nuvens, grid, processamento paralelo entre outros, é necessário para o desenvolvimento de um sistema de informação de gestão agrícola automatizado que seja robusto e confiável.

A Figura 2 exemplifica o sistema de produção e os processos que nele ocorrem. Informações georreferenciadas dos atributos do solo (características físicas, químicas e biológicas) são coletadas, transmitidas e analisadas para que sejam estabelecidas as capacidades produtivas de áreas do terreno; em função dessa análise informações são transmitas às máquinas e aos equipamentos para a aplicação automatizada de corretivos e fertilizantes em taxas variáveis. Em seguida ocorre a operação de semeadura/ou plantio (mudas) automatizada com a utilização de plantas adequadas às diferentes capacidades produtivas do terreno, i.e., para explorar a máxima capacidade produtiva do solo. Posteriormente ocorrem as operações de manejo da cultura. Os estresses bióticos (patógenos, insetos e plantas daninhas) e abióticos (deficiências hídricas e nutricionais) podem ser determinados e georreferenciados pela utilização de sensores remotos (por satélite, avião, Vant). Todas essas informações são armazenadas e transmitidas numa linguagem padrão de intercâmbio, e.g. AgroXML, para uma central para serem processadas (computação em nuvens, grid, paralela entre outros) e analisadas por um sistema específico de decisão, que encaminhará as decisões, em conformidade com o padrão ISOBUS,

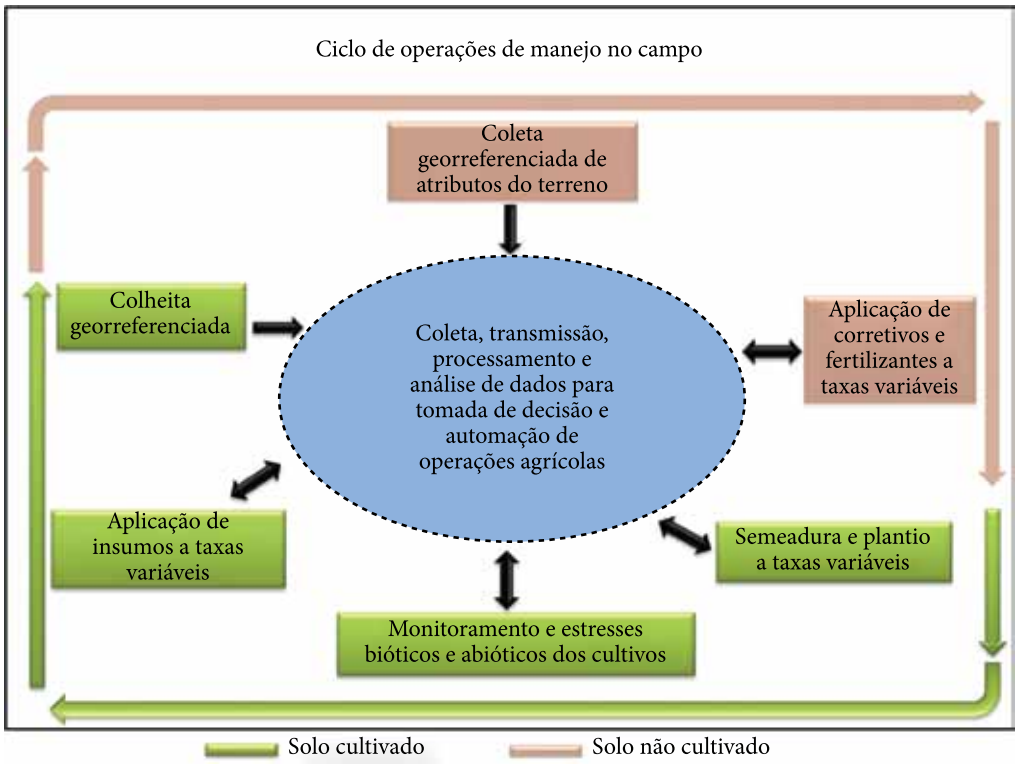


Figura 2. Ilustração das fases do sistema de produção e os processos que nele ocorrem.

para as máquinas equipadas com sistema de direção automática e equipamentos, que realizarão as operações de aplicação georreferenciadas em taxas variáveis de insumos (água, fertilizantes, defensivos, agentes de controle biológico,etc). O ciclo é iniciado novamente após a colheita, com a utilização de sensores de produtividade e/ou de qualidade (conteúdo de proteína, óleo ou outro parâmetro) cujos dados são enviados à central de processamento para a obtenção de mapas.

No mundo esses sistemas são utilizados para pequenas áreas de produção, como no Japão, Estados Unidos e na Europa. No Brasil são aplicados partes desse processo, como o piloto automático para operações de plantio.

Salienta-se que o repositório desenvolvido pela Rede AP terá um papel importante na organização do conhecimento científico e tecnológico que facilitará a coleta, transmissão, processamento e análise de dados para tomada de decisão e automação de operações agrícolas, visto que foi construído para atender requisitos de intercâmbio e interoperabilidade entre máquinas, equipamentos agrícolas e sistemas embarcados utilizados nos processos de automação.

Uma vez que a automação é uma tendência consolidada com evolução e expansão de seu uso, a Embrapa está criando um portfólio para definir necessidades de pesquisa e de inovação para consolidar, em bases científicas, os processos de automação agrícola.

10. Conectando Ciência e Tecnologia com a Extensão Rural, Agentes dos Setores Produtivos e Formuladores de Políticas Públicas

Bongiovanni e Lowenberg-Deboer (2001) definem AP como sendo: 'o monitoramento e controle eletrônico aplicado a coleta e ao processamento de uma base de dados e de informações para suporte a decisão na alocação espacial e temporal de insumos'. Portanto, baseados nessa definição e em tópicos citados em Ballantyne, Maru e Porcari (2010), faremos considerações sobre tendências futuras e quanto as formas de geração, transferência e uso das inovações tecnológicas. Será considerada a evolução das TIC's e das tecnologias de AP nos diversos setores envolvidos nas cadeias produtivas. Os avanços no desenvolvimento de hardware, software, formas de conectividade, volume de informações coletadas, processadas e disponibilizadas já estão transformando os métodos de promover inovações. O processo linear da transferência das tecnologias e conhecimentos originados da pesquisa para os produtores através da extensão, já vem sendo operacionalizado num processo de transferência em redes de conhecimentos e de informações.

As tendências futuras que indicam uma evolução e crescimento do uso das TIC's, nos temas de Computação Ubíqua e em Nuvens, Aplicações Geoespaciais, Sistemas de Suporte a Decisão, Processos e Equipamentos Agrícolas, Padronização de Dados e Automação já estão permitindo o acesso às informações e aos conhecimentos originados de fontes pluralísticas com diferentes formatos que estão sendo utilizados pelos usuários (GAKURU; WINTERS; STEPMAN, 2009; GANDHI et al., 2009). Isto significa que, não somente o conhecimento gerado pelas instituições de pesquisas vem sendo utilizado, mas também o conhecimento tácito obtido por produtores, provedores de serviço e extensionistas está sendo utilizado nas inovações. A grande maioria das inovações já está sendo transferida de modo ubíquo, ou seja, o usuário está acessando um volume enorme de informações e tendo que ter alguma forma de filtragem, para selecionar as tecnologias e conhecimentos mais relevantes para sua situação.

Considerando as mudanças ocorridas na sociedade, devido aos impactos das novas Tecnologias de Informação e de Comunicação, exigem da Embrapa novos procedimentos. A forma como foi concebido e desenvolvido o repositório da Rede AP, permitirá a organização das informações e dos conhecimentos existentes e será um instrumento efetivo e eficaz de transferência tecnológica, contribuindo para acelerar o desenvolvimento do processo de disseminação e adoção das tecnologias da Agricultura de Precisão.

Em função do quadro atual e das tendências futuras os formuladores de políticas públicas devem considerar o valor das TIC´s no desenvolvimento do setor agrícola, políticas que propiciem o acompanhamento da dinâmica da

evolução das TIC´s como a participação continua em programas de treinamento, capacitação dos pesquisadores, extensionistas, produtores e outros atores. Ações dessa natureza contribuirão paro uso pleno de suas capacidades de gerar, transferir, compartilhar dados, informações e conhecimentos, para o efetivo desenvolvimento do setor agrícola. Também é esperado que ocorram mudanças culturais para que assim culminem em transformações em direção a novos padrões tecnológicos de produção.

Agradecimentos

Agradecemos à Rede AP pelo apoio e oportunidade para a redação desse capítulo.

Referências

AGRICULTURAL INDUSTRY ELECTRONICS FOUNDATION - AEF. **O que é a AEF**. Disponível em: <http://www.aef-online.org/pt/o-que-e-a-aef.html>. Acesso em: 24 jun. 2013.

FORÇA TAREFA ISOBUS BRASIL. **Força Tarefa ISOBus.** Disponível em: <http://www.isobus.org.br/index.php?option=com_content&view=article&id=50&Itemid=37>. Acesso em: 24 jun. 2013.

AGGATEWAY. **SPADE PROJECT**. Disponível em: <http://www.aggateway.org/eConnectivity/Projects/CurrentOngoing/SPADE.aspx>. Acesso em: 11 jul. 2013.

BALLANTYNE, P.; MARU, A.; PORCARI, E. M. Information and communication technologies — opportunities to mobilize agricultural science for development. **Crop Science,** Madison, v. 50, p. S-63-S69, 2010. http://dx.doi.org/10.2135/cropsci2009.09.0527

BARET, F.; GUYOT, G.; MAJOR, D. J. TSAVI: a vegetation index which minimizes soil brightness effects on LAI and APAR estimation. In: CANADIAN SYMPOSIUM OF REMOTE SENSING, 10., 1990, Washington. **Proceedings**… New York: The Institute of Electrical and Electronics Engineers, 1990. p. 1355-1358.

BLACKMORE, S.; APOSTOLIDI, K. **Project – Integration of Farm Management Information Systems to support real-time management decisions and compliance of standards - final report**. 2011. Disponível em: <http://www.futurefarm.eu/system/files/FFD8.9_Final_Report_4.1_Final.pdf>. Acesso em: 22 jun. 2013.

BONGIOVANNI, R.; LOWENBERG-DEBOER, J. Precision agriculture: Economics of nitrogen management in corn using site-specific crop response estimates from a spatial regression model. In: AMERICAN AGRICULTURAL ECONOMISTS ASSOCIATION ANNUAL MEETING, 2001, Chicago. **Proceedings**… Chicago, 2001. Selected Paper.

KEPES, B. **CLOUDU UNDERSTANDING The Cloud Computing Stack SaaS, Paas, IaaS.** Disponível em: <http://broadcast.rackspace.com/hosting_knowledge/whitepapers/Understanding-the-Cloud-Computing-Stack.pdf>. Acesso em: 2 jul. 2013.

FURLANETO, F. B.; MANZANO, L. M. **Agricultura de precisão e a rastreabilidade de produtos agrícolas.** 2010. Disponível em: <http://www.infobibos.com/Artigos/2010_2/AgriculturaPrecisao/>. Acesso em: 2 jul. 2013.

GAKURU, M.; WINTERS, K.; STEPMAN, F. **Inventory of innovative farmer advisory services using ICTs.** Forum for Agricultural Research in Africa, 2009. p. 1-66.

GANDHI, R.; VEERARAGHAVAN, R.; TOYAMA, K.; RAMPRASAD, V. Digital green: Participatory video and mediated instruction for agricultural extension. **Information Technologies & International Development**, Los Angeles, v. 5, n. 1. p. 1-15, 2009.

HASSALL, J. **Future Trends in Precision Agriculture:** A look into the future of agricultural equipment. Nuffield Australia, 2010. p. 1-36.

HEST, D. **Capitalizing on the cloud:** Wireless connectivity in agriculture will make big gains in 2013. 2013. Disponível em: <https://www.onsiteag.com/news/capitalizing-on-the-cloud-wireless-connectivity-in-agriculture-will-make-big-gains-in-2013—21.html>. Acesso em 1 jul. 2013.

HUETE, A. R. A soil-adjusted vegetation index (SAVI). **Remote Sensing of Environment,** New York, v. 25, n. 3, p. 295-309, 1988. http://dx.doi.org/10.1016/0034-4257(88)90106-X

LUCHIARI JUNIOR, A.; BORGHI, E.; AVANZI, J. C.; FREITAS, A. A.; BORTOLON, L.; BORTOLON, E. S. O.; ;INAMASU, R., Y. Zonas de manejo: teoria e prática. In: INAMASU, R. Y.; NAIME, J. M.; RESENDE, A. V.; BASSOI, L. H.; BERNARDI, A. C. (Ed.). **Agricultura de precisão**: um novo olhar. São Carlos: Embrapa Instrumentação, 2011. p. 60-64.

MORAN, M. S.; INOUE, Y.; BARNES, E. M. Opportunities and limitations for image-based remote sensing in precision crop management. **Remote Sensing of Environment**, New York, v. 61, n. 3, p. 319-346, 1997. http://dx.doi.org/10.1016/S0034-4257(97)00045-X

QUEIROS, L. R.; SPERANZA, E. A.; BETTIOL, G. M.; FILIPPINI ALBA, J. M.; BERNARDI, A. C. C.; INAMASU, R. Y.; GREGO, C. R.; RABELLO, L. M. **Gestão de recursos de informação em Agricultura de Precisão.** Campinas: Embrapa Informática Agropecuária, 2011. 41 p. il. (Embrapa Informática Agropecuária. Documentos, n. 112). Disponível em: <http://www.infoteca.cnptia.embrapa.br/handle/doc/903343>. Acesso em: 10 jun. 2013.

RONDEAUX, G.; STEVEN, M.; BARET, F. Optimization of soil-adjusted vegetation indices. **Remote Sensing of Environment**, New York, v. 55, p. 95-107, 1996. http://dx.doi.org/10.1016/0034-4257(95)00186-7

RUSTEN, E.; RAMIREZ, S. **Future direction in agriculture and Information and Communication Technologies (ICTs) at USAID.** Academy for Educational Development, 2003. Disponível em: <http://www.winrock.org/agriculture/files/ag_ict.pdf>. Acesso em: 11 jul. 2013.

SARAIVA, A. M.; CUGNASCA, C. E. Redes de comunicação serial em máquinas agrícolas: uma revisão. **Revista Brasileira de Agroinformática**, Lavras, v. 8, p. 17-35, 2006.

SCHEPERS, J. S.; SHANAHAN, J. F.; LUCHIARI JÚNIOR, A. Precision Agriculture as a tool for sustainability. In: BALÁZS, E.; GALANTE, E.; LINCH, J. M.; SCHEPERS, J. S.; TOUTANT, J.-P.; WERNER, D; WERRY, P. A. T. J. (Ed.). **Biological resource management: connecting science and policy**. Berlin: Springer, 2000. p. 129-138. http://dx.doi.org/10.1007/978-3-662-04033-1_10

SHANAHAN, J. F.; SCHEPERS, J. S.; FRANCIS, D. D.; VARVEL, G. E.; WILHELM, W. W.; TRINGE, J. M.; SCHLEMMER, M. S.; MAJOR, D. J. Use of remote sensing imagery to estimate corn grain yield. **Agronomy Journal,** Madison, v. 93, p. 583-589, 2001. http://dx.doi.org/10.2134/agronj2001.933583x

SOUSA, R. V.; LOPES, W. C.; PEREIRA, R. R. D.; INAMASU, R. Y. Estudo dos elementos mínimos para projeto de sistemas embarcados compatíveis para máquinas e implementos agrícolas. In: INAMASU, R. Y.; NAIME, J. M.; RESENDE, A. V.; BASSOI, L. H.; BERNARDI, A. C. (Eds.). **Agricultura de precisão**: um novo olhar. São Carlos: Embrapa Instrumentação, 2011. p. 126-131.

TORRE NETO, A. **Rede de sensores sem fio e computação ubíqua na agropecuária.** São Carlos: Embrapa Instrumentação Agropecuária, 2009. 18 p. (Embrapa Instrumentação Agropecuária. Boletim de Pesquisa e Desenvolvimento, n. 31). Disponível em: <http://www.infoteca.cnptia.embrapa.br/handle/doc/658283>. Acesso em: 22 jun. 2013.

WELTE, J.; AULT, A.; BOWMAN, C.; LAYTON, A.; NOEL, S.; KROGMEIER, J.; BUCKMASTER, D. **Autogenic mobile computing technologies in agriculture: applications and sensor networking for smart phones and tablets**. St. Joseph: American Society of Agricultural and Biological Engineers, 2013. http://dx.doi.org/10.13031/aim.20131579954

WOLF, S. A.; WOOD, S. D. Precision Farming: environmental legitimation, commodification of information, and industrial coordination. **Rural Sociology**, Knoxville, v. 62, n. 2, p. 180-206, 1997. http://dx.doi.org/10.1111/j.1549-0831.1997.tb00650.x

Apendice C

A Cluster-Based Approach to Support the Delineation of Management Zones in Precision Agriculture

Este apêndice contém a transcrição do artigo científico intitulado *A Cluster-Based Approach to Support the Delineation of Management Zones in Precision Agriculture*, apresentado oralmente e publicado nos anais da *10th IEEE International Conference on eScience* (SPERANZA, 2014).

2014 IEEE 10th International Conference on eScience

A Cluster-Based Approach to Support the Delineation of Management Zones in Precision Agriculture

Eduardo Antonio Speranza
National Research Center for Computer Science in Agriculture
Brazilian Agricultural Research Corporation
Campinas, SP, Brazil
eduardo.speranza@embrapa.br

Ricardo Rodrigues Ciferri
Department of Computer Science
Federal University of São Carlos
São Carlos, SP, Brazil
ricardo@dc.ufscar.br

Célia Regina Grego, Luiz Eduardo Vicente
National Research Center for Satellite Monitoring
Brazilian Agricultural Research Corporation
Campinas, SP, Brazil
{celia.grego,luiz.vicente}@embrapa.br

***Abstract*—In this paper we propose a cluster-based approach for the delineation of management zones in precision agriculture. The proposed approach was built following the steps of data mining for the clustering task, resulting in a computer application that generates maps of management zones and yield areas, allowing to compare them using known statistical indexes. The basis for this implementation was a model previously published in the literature that uses only historical productivity, soil electrical conductivity and relief data to generate the maps. The main difference of our work with respect to the previous model is the clustering algorithms used in the step of extracting patterns. While the original model uses only the fuzzy c-means algorithm, the model developed in this study uses the GKCluster extension to this algorithm, able to detect clusters with different geometrical shapes. From the tests performed with the new proposed model, we achieved about 76% of correlation between maps of yield and management zones from kappa index, and about 85% of correlation from overall accuracy. The original model reached, according to the authors, a maximum correlation of 49% from kappa index, and 70% from overall accuracy.**

Keywords—precision agriculture; spatial data mining; clustering; management zones.

I. INTRODUCTION

Precision agriculture is usually defined as a farm management system, composed by technologies and process used for the optimization of crops and production systems based on the space variability management [1]. In traditional agriculture, the application of agricultural supplies, like soil fertilizers and correctives, is performed in an equivalent manner for all the crops, namely, an average value is applied. Thus, areas with very different yield levels are not treated differently, providing a waste in the application of these materials and preventing yield growth in less productive areas, not worrying about the environment. With the adoption of precision agriculture, the farmer has the ability to manage rationally the supplies to be used in tillage, applying them differently in georeferenced plots of the production area, known as management zones.

Management zones are defined as spatial regions that express a homogeneous combination of factors limiting the yield in which the application of a uniform dose of a particular input is appropriate [2]. They can be treated as subfields where the intervention takes into account the homogeneity in management terms. This approach can facilitate the adoption of an action strategy in precision agriculture without the need of using very sophisticated machinery and implements [3]. Thus, it can be adopted by both small and large producers. The main benefits of adopting this technique are the use of a minimum amount of agricultural supplies, water and energy, as well as environment preservation. As a result, is expected to increase the crop yield in areas known as less productive.

With the goal of allowing a quick and easy generation of management zones without worrying about the magnitude of agricultural production, this paper proposes the creation of an intelligent system using a limited number of variables, such as soil electrical conductivity (EC), relief and productivity, and clustering algorithms used for pattern extraction in data mining. Our experiments were performed using data obtained from the culture of sugar cane and preprocessed by using geostatistical techniques.

This paper is organized as follows. Section II surveys the background and related work, Section III describes the proposed cluster-based approach, Section IV summarizes a case study in sugar cane, Section V presents a discussion about the obtained results and Section VI concludes the paper and proposes the future work.

978-1-4799-4287-9/14 $31.00 © 2014 IEEE
DOI 10.1109/eScience.2014.42

II. Background and Related Work

A. Geostatistics and Spatial Variability

The starting point to enable the use of precision agriculture must be checking the spatial variability of the crop, which can be performed collecting and analyzing georeferenced information capable of providing data to the delineation of management zones. Because this information is usually collected from different sensors at different times, the collected data will not be georeferenced always in the same spots. Thus, first, it is necessary that this information be placed in a single grid of cells, so that clustering algorithms can be valid. The most traditional way for checking spatial variability and placing different variable data in the grid is by geostatistical techniques. According to [4], geostatistics is applied statistics which deals with problems involving regionalized variables. These variables display spatial behavior and show intermediate characteristics between truly random and completely deterministic variables, presenting an unpredictable variation from one point to another and relationships between the points in space. Geostatistical methods have two basic tools: the semivariogram and the kriging algorithm. The semivariogram shows a measure of the degree of spatial dependence between samples along a georreferenced area and, from this information, kriging finds optimal weights to be associated with samples that will estimate a value at a point. On the other hand, kriging is a process of estimating variable values from adjacent values distributed in space and considered as interdependent by the semivariogram [5]. The work of [6] illustrates spatial dependency and variability checking from a crop using geostatistics. In this work, the authors perform a geostatistical analysis using EC at 30 and 90 cm depth and topography from altimetry obtained from a culture of sugar cane. The results of the study showed spatial dependence of all variables, and consequently, their possible use for delineating management zones. The data and analysis obtained in [6] were used for the experiments performed in our work.

When we estimate a set of spatially correlated variables, multivariate geostatistics must be used [7]. In this case, unknown values should be estimated using the cross semivariogram and the cokriging method. The first describes the simultaneous spatial variation of two or more random variables, and the last is a process where several regionalized variables can be estimated together based on the set spatial relationship to each other, creating a multivariate extension of the kriging method [8]. In the literature, the work of [9] is an example of using multivariate geostatistics. In this case, through the cross semivariogram and cokriging, a high correlation between rainfall and elevation data was verified, which allows for unsampled points for the first variable to be obtained with assistance of the last.

Regardless of the large number of published papers related to precision agriculture using geostatistical analysis, this process is still performed manually, i.e., users need to perform filtering of the data, then analyze the semivariogram to verify if indeed the spatial variability exists at the study site, to finally obtain a model that can perform data interpolation and make them useful in the design of management zones. The use of spatial databases and other techniques, like spatial regression, can be an alternative to solve this problem and render the processes of spatial data interpolation and spatial variability checking more automated and out of sight of the user.

B. Clustering Algorithms for Extracting Patterns

When there is no prior knowledge about a particular problem, either for the absence of domain experts, or due to the high cost of performing a classification of samples, unsupervised machine learning algorithms are used. Some of these, named clustering algorithms, have the skill of cluster samples using the inherent knowledge in the data, through data mining tasks.

In data mining, the classical method used for clustering is the k-means [10]. This method aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The result is a partitioning of the data space into Voronoi cells [11]. An important characteristic from k-means is that, as a result of its execution, each sample belongs just to a unique cluster, characterizing it as a hard cluster algorithm.

In precision agriculture, clustering algorithms have been extensively used to aid in the delineation of management zones. The geographic coordinates associated with each sample obtained in the field leads us to argue that the data can be grouped not only by similarity of the values obtained at each point, but also by geographical distance determined between samples. However, because data are associated with location this causes an inherent degree of vagueness, because of several factors such as the accuracy of the GPS capture device. Moreover, this vagueness is aggravated because of geostatistical transformations that enable your use by clustering algorithms, making these groupings not wholly reliable if obtained as hard clusters, as by the k-means algorithm. For this reason, clustering algorithms that take into account the data imprecision, such as those obtained by georeferenced gathering or geostatistical processing, should be used for applications in precision agriculture. These algorithms are classified as soft or fuzzy clustering, and allow a sample to belong to more than one cluster with a certain degree of membership.

One of the most widely used fuzzy clustering algorithms is the fuzzy c-means (FCM) algorithm [12]. While the k-means calculates Euclidean distance of a sample to the center of all clusters linking the sample to the nearest cluster, FCM performs this step by calculating the relevance degree of each sample with respect to each cluster center, as described in (1).

$$\omega_k(x) = 1 \,/\, \sum (d(\text{center } k,x)/d(\text{center } j,x)^{2/(m-1)}) \qquad (1)$$

In equation (1), $\omega_k(x)$ represents the membership degree of a sample x with respect to a cluster k; j is the desired number of clusters; d is a distance function (usually the Euclidean distance); and m is the fuzzifier parameter that determines the level of cluster fuzziness, whose default value is 2. Large values of m results in smaller membership degrees, and consequently, fuzzier clusters. In the limit m=1, the membership degree converges to 0 or 1, implying a crisp partitioning. Thereby, samples on the edge of a cluster may be considered *in the cluster* to a lesser degree membership than

samples in the center of cluster. This calculation must be performed iteratively until the values of membership degree are not changed up than a certain threshold value between two iterations. At each iteration, new centroids are computed for each cluster, according to (2).

$$c_k = \sum_x \omega_k(x)^m x \; / \; \sum_x \omega_k(x)^m \quad (2)$$

In equation (2), c_k is the centroid of cluster k. The remaining variables are described in (1). In this case, the centroid is calculated taking into account all dataset samples and their membership degree with respect to the cluster.

Regarding the use of fuzzy clustering algorithms to assist delineation of management zones, the work of [13] can be highlighted. In this study, the authors found that data from EC and elevation can be used for obtaining clusters to generate management zone maps, and accuracy is comparable with maps obtained from clustering of historical yield data. The model developed creates two types of clusters: first, a set of clusters combining EC, EC ratios and elevation variables; last, a cluster that combines annual, mean and standard deviation values of yield. Both types of clusters are obtained running the FCM algorithm, and later compared using statistical analysis like the kappa index to obtain the management zone map best suited to the yield of the crop. To validate this model, the authors used data from EC, elevation and previous harvests yield of corn and soybean obtained from two areas with clay soil near the town of Centralia, Missouri, USA. All these data were preprocessed and sampled in the same grid with 10 meters of spatial resolution using geostatistical techniques. The correlation between the management zones was found to be 70%, considered good by experts.

Although it is very useful for performing soft clustering, the FCM algorithm fails to observe some characteristics that may be important in cluster formation. Among these characteristics is the ability to detect clusters with different geometrical shapes in a data set. To address this problem, an extension to fuzzy clustering was developed by [14], known as GKCluster. The GKCluster employs an adaptive distance norm, used as optimization variables in fuzzy clustering algorithms, thus allowing each cluster to adapt the distance norm to the local topological structure of data. Taking into account the application of precision agriculture, it is clear that due to the large number of variables that can influence the spatial variability of a culture, as EC and relief, the management zones to be obtained for an area must have different geometric shapes. Thus, the GKCluster extension tends to provide better results compared to what can be achieved with pure fuzzy clustering algorithms like FCM.

C. Statistical Indexes for Maps Comparison

According to [13], indexes like overall accuracy and kappa provide higher values when the spatial agreement between the two compared datasets is maximal. The overall accuracy used in this study is simply obtained by dividing the number of samples with coinciding clusters by the total number of samples. The kappa coefficient [15] is a statistical index that verifies the accuracy between two datasets from a contingency table cross-classification. The kappa index calculation is performed according to equation (3).

$$k = (Pr(a) - Pr(e)) \; / \; (1 - Pr(e)) \quad (3)$$

In this equation, $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying for each category. The kappa index value (k) varies from 1 (complete agreement) to 0 (no agreement).

Typically, the overall accuracy values obtained by comparing two sets of data are greater than the values obtained by the kappa index, because the second makes use of probability based on the original data.

D. Closing Remarks

In our research, the model described by [13] was used as the basis for developing our cluster-based approach. However, some changes - described in section III - were made to improve accuracy of the obtained results. First, we established steps that must be observed in data mining and related them with the procedures implemented in the model. Then, more specifically in the pattern extraction step, we used GKCluster extension for fuzzy clustering algorithms, that aims to improve the quality of the clusters obtained. For this, we used the FCM and GKCluster implementations available in the Fuzzy Clustering and Data Analysis Toolbox for Matlab[1] [16]. Finally, in the step of comparing maps, we implemented an algorithm for calculating the overall accuracy, and used the algorithm implemented by [17] for calculating the kappa index, both in Matlab.

III. CLUSTER-BASED APPROACH DEVELOPMENT

According to [18], a cluster-based approach should be used to solve a problem when there is no predicted class, but the samples tend to split into natural groupings. Thus, clustering is included in the group of unsupervised machine learning algorithms. There are different ways in which the result of clustering can be expressed: any sample belongs in only one group; any instance may fall into several groups; instances can be roughly divided into groups at an hierarchical top level and each group can be refined further; or any instance belongs to each group with a certain probability. Due to the fact that samples in precision agriculture possess a degree of uncertainty, as described in Section I, clustering algorithms with probabilistic characteristics tend to be more suited to assist the delineation of management zones.

When we need to develop an intelligent system with several steps for processing and use of data, the context of data mining should be considered. According to [19], data mining consists of the following steps: problem identification; preprocessing; patterns extraction; and knowledge use. The definition of our approach takes into account these steps, and is described below.

[1] Matlab: high-level language and interactive environment for numerical computation, visualization and programming.

A. Problem Identification and Design

The delineation of management zones of a crop for use in precision agriculture can take into account a series of variables related to yield, soil, plant and environment. This type of data may be obtained from field sensors or aerial images that can identify, for example, water shortage in some locations of the cultivation. However, all these data must have an essential factor to design these zones: an associated geographic coordinate. One way to verify the accuracy of management zones maps is comparing them with yield maps of the same area. This comparison is considered valid because the increased yield in less productive areas is one of the key factors for the use of precision agriculture.

The general implementation model used by [13] was considered as reference for this work due to the availability of a georeferenced database of soil electrical conductivity (EC), elevation and productivity for tests. However, although it seems obvious that it is an implementation of an intelligent system based on acquired knowledge from the data, the authors did not formally describe the steps that must be performed to the application of data mining techniques based on clustering algorithms. In adition, they used a well-known algorithm for this purpose (FCM) only in its original version, which does not care about some features that may be important in the clusters formation, such as its shape. Thus, the identification of data mining steps in the design of the proposed model to improve its understanding, and the use of an extension for fuzzy clustering algorithms to improve the accuracy of the obtained management zones were both added in our work. The new model is shown in Figure 1.

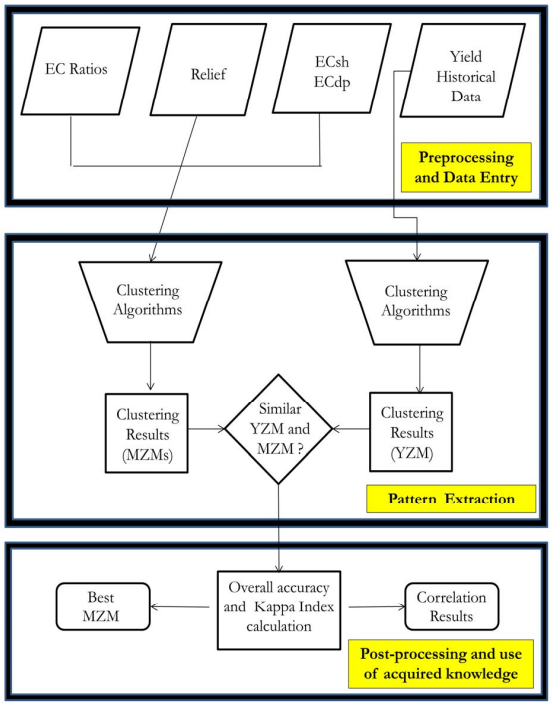


Fig. 1. Model for development of intelligent system

The following three sections describe the implementation of the three steps of the model, highlighted in yellow in Fig. 1.

B. Preprocessing and Data Entry

The preprocessing step provides the suitability of the input data so that they can be used correctly by the intelligent system. According to [19], several preprocessing methods, such as treatment, cleaning and data volume reduction can be performed. In this work, because data to be used were collected in points with known locations, all variables must be available in a same regular grid before to be used by clustering algorithms. To solve this problem, geostatistical techniques such as kriging are used to suit the input data. As kriging generates a rectangular grid, after this step the clipping of the resulting grid is performed, so only the cells that intersect the polygon that represents the study area are used. Here, this intersection topological relationship is established as described by [20]. Both preprocessing data using geostatistics, as the clipping for use only the cells of interest, should be performed by a geographic information system with these functions.

After preprocessing, the user needs to submit two input files for the system: one containing the EC and elevation data used to generate management zone maps (MZMs) and another containing yield data for past vintages to generate the yield zone map (YZM). For the creation of MZMs, data from EC have measures at 30 cm depth, which we will call *ECsh*, and values at 90 cm depth, which we will call *ECdp*. Furthermore, the system allows that *EC ratios* also be used as input variables, allowing different combinations of them. On the other hand, for the creation of YZM, the user can choose to use one of these three combinations: variables with data from all available vintages; only data from one vintage; mean and standard deviation of all vintages.

Finally, how the algorithms used in the pattern extraction step are based on the distance between the samples, it is possible that a variable with a value range much larger than another can excessively influence the cluster formation. To prevent this, when data are reported to the system, it automatically performs a normalization of each variable for values between 0 and 1.

C. Pattern Extraction

The step of pattern extraction should be developed to achieve the objectives defined in problem identification [19]. Since our system searches the intrinsic knowledge in the data to solve a problem, because samples do not have a specified class, it is understood that the task of clustering is more suitable for solving it. Moreover, the fact that data is associated with location indicates some evidenced degree of uncertainty, as described above. Thus, it is feasible to use clustering algorithms capable of handling such uncertainties, like FCM and extensions described in Section II. At this step, as many MZMs are generated as necessary and also a single YZM, aiming to meet all combinations of variables determined by the user. The next step, described below, performs through statistical indices the comparison between all MZMs generated regarding the YZM generated, returning to the user the best possible recommendation for MZM.

D. *Post-processing and use of acquired knowledge*

At this step, a comparison between all MZMs regarding YZM obtained in the previous step through the global accuracy and kappa statistical indices (described in Section II) is performed. The MZM that obtains the best value of the kappa index when compared to YZM will be counted as the most useful map of management zones obtained from the reported data.

IV. Case Study in Sugar Cane

In this section, we describe the results obtained from tests carried out in a study area used by the Brazilian Agricultural Research Corporation (Embrapa) for research in precision agriculture. This area corresponds to a field of about 17 hectares at Fazenda Aparecida, located in Mogi-Mirim, São Paulo state, Brazil, with coordinates 7505136N, 299621E in UTM Zone 23. This area boasts cultivation of sugar cane (*Saccharum officinarum* L.) by tillage for 12 years. Fig. 2 shows the georeferenced polygon edge corresponding to this field.

Tests were performed at this area using combinations of EC and elevation from real data collected in the field, and all variations of possible clusters, with settings ranging from 2 to 5 clusters. The following data were used for these tests: *ECsh* and *ECdp* from direct contact VERIS sensor [21] measured in 2010 and 2012; elevation data obtained by GPS and altimetry in 2010; and yield data in 2010, 2012 and 2013. All possible combinations of these variables allow the generation of 80 distinct test cases, allowing detailed analysis on them. At the preprocessing step, agronomy specialists reported that cellular spaces with 30 meters side length would be sufficient to identify the spatial variability of this field. Thus, from geostatistical techniques and a cropping only considering the area of interest, as described in section III, our study area was divided into 196 representative cells. Table 1 shows the model configuration for the better results obtained varying the number of clusters.

Fig. 2. Study Area

TABLE I. Better Results by Number of Clusters

# Clusters	Better Results Test Configurations			
	MZM variables	*Algorithm*	*OA*[a]	*kappa*
2	ECdp/ECsh; Elevation	GKCluster extension	0.88	0.76
3	ECsh/ECdp; Elevation	FCM	0.61	0.41
4	ECdp/ECsh; Elevation	GKCluster extension	0.41	0.23
5	ECdp/ECsh; Elevation	GKCluster extension	0.42	0.27

[a] Overall Accuracy

The best result was obtained using the EC ratio *ECdp/ECsh* and altimetry measured in 2010 to generate clusters that resulted in the best MZM, and all the historical yield data to generate clusters that resulted in YZM. The algorithm used for this result was the GKCluster extension, with the parameter number of clusters equal to 2. The value of overall accuracy obtained in this result was 0.88, and the kappa index was 0.76, indicating substantial agreement between obtained MZM and YZM. This result can be seen in Fig. 3.

Through this figure, we can visually observe a substantial agreement between the management area 1 (MZ-1) and the yield zone 1 (YZ-1); and between management area 2 (MZ-2) and yield zone 2 (YZ-2).

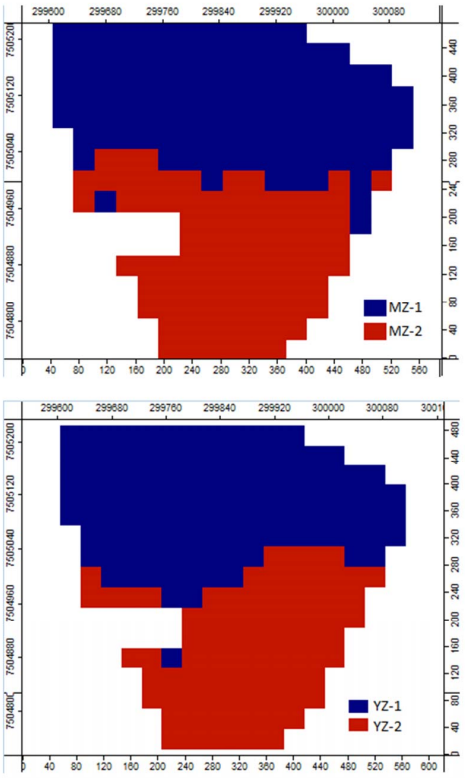


Fig. 3. Best result obtained from the tests

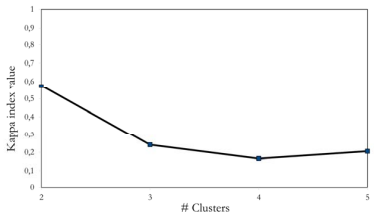
This result indicates that this is the best configuration for management zones found in this area, taking into account the data available for use.

In the next section, some analyses involving all the obtained results will be described, allowing a discussion of our work.

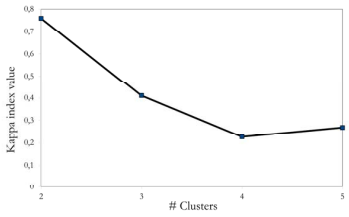
V. Discussion

After the tests, some analysis about the used clustering algorithms (pure FCM and GKCluster extension) was performed. Fig. 4 graphically shows the average (a) and maximum value (b) of the kappa index obtained considering all algorithms using different amounts of clusters.

Fig. 4 shows that both the mean and the maximum value decreases considerably when the number of clusters increases, returning to slightly improve only when we move from 4 to 5 clusters. Since this area is considered small by the experts, the observed spatial variability should provide the creation of a small number of clusters. This statement can be proven by the best result, shown in Fig. 3. If we analyze the EC and elevation data in a separate way, we can verify that the result shown in Fig. 3 was strongly influenced by the elevation variable, since the study area has a steep slope from north to south direction. The result obtained for the YZM proves this influence, identifying two distinct areas of yield quite compatible with two distinct areas obtained for the best MZM .



(a)



(b)

Fig. 4. Average (a) and maximum (b) kappa index from all algorithms varying the number of clusters.

The EC data, in this case, are useful to improve the accuracy of the obtained MZM, especially in places where the degree of membership of an instance with respect to all clusters is very similar if we consider only the elevation variable.

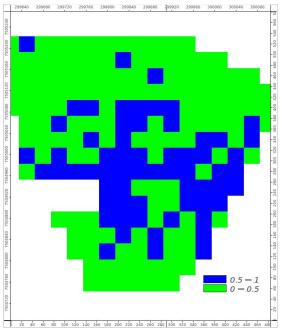
Fig. 5 shows the maximum value of Kappa index achieved by an algorithm when it was considered the best choice. From the graph displayed it can be verified, in general, that the FCM algorithm with the GKCluster extension obtained a much better performance than the traditional FCM algorithm. This result can be explained by the fact that GKCluster extension is adapted to identify clusters with different geometric shapes, allowing a better separation of clusters with georeferenced features, such as the management zones.

Considering all executed test cases, the settings with the best rates of correlation between MZM and YZM indicate more accurate management zones with respect to the model when: 2 clusters (75% of cases) are used; GKCluster FCM extension is used (60% of cases); and EC ratios in data entry are used (93% of cases).

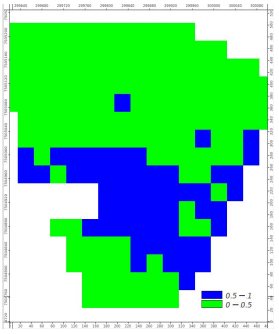
The first two cases, concerning the amount of clusters and the used algorithm, can be explained as previously described. Regarding the best of accuracy using EC ratios, [13] used this approach because it has already been successfully used in the work of [22], to assist in the explanation of the variability level of the culture. According to this work, the EC ratios allow an indication of the bulk electrical conductivity profile. Considering our best result, ratios > 1 indicate a growth profile of EC with increasing depth; ratios = 1 indicate a stable profile; and ratios < 1 indicate a decreased EC profile with increasing depth.

Fig. 6 shows maps for *ECdp* (a), *ECsh* (b) and *ECdp/ECsh* (c) used in our best result, all using two clusters or zones. The maps from Fig. 6 (a) and 6 (b) have values normalized between 0 and 1 and two EC zones equally classified: zone 1, from 0 to 0.5; and zone 2, from 0.5 to 1. If we visually analize these maps we can observe, for both cases, the appearance of a spot with higer EC values in the central region, and two spots with lower EC values in the ends of the map. Already the map from Fig. 6 (c) shows EC ratios also with two zones equally classified: zone 1, for ratios <= 1 and zone 2, for ratios > 1, following, in a simplified way, the EC profiles as explained above. In this case, we can no longer identify the spots displayed in previous maps. This shows that when we visualize de EC profiles and not just measures, the spatial variability regarding this variable changes, contributing to better accuracy in clustering for management zones with respect to the areas of yield, as noted in Fig. 3.

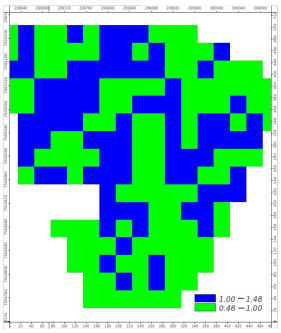
Fig. 5. Maximum value of kappa index for each algorithm as this is considered the best choice.



(a)



(b)



(c)

Fig. 6. ECdp (a), ECsh (b) and EC ratio (ECdp/ECsh)

VI. Conclusions and Future Work

According to the results obtained in the previous section, we can conclude that the use of EC ratios improves the accuracy in obtaining the MZMs compared to YZM. These results showed too that the use of optimized clustering algorithms to detect clusters with different geometrical shapes can be more effective in delineating management zones for precision agriculture, compared to algorithms in their traditional form, such as fuzzy c-means.

In our study, only EC and elevation variables were used. However, we know that many other variables can be collected, whether from soil samples, which can provide the values of chemical and physical soil properties such as the amount of clay, silt, sand, and indexes of pH, phosphorus, potassium, aluminum, etc., or from other field sensors or remote sensing images, capable of providing indexes such as the Normalized Difference Vegetation Index (NDVI) in precisions of centimeters. Anyway, different variables related to soil, plant and climate may be used, to make the delineation of management zones as close to reality as possible.

Although the area used for this study is somewhat biased because of their steep topography, it was useful to verify that FCM extensions able to detect clusters with different geometrical shapes in a data set, such as GKCluster, can be used to improve the accuracy in the delineation of management zones. Anyway, in future work, other areas producing even other crops and data collected in a higher temporal resolution should be used to improve the verification of results.

Due to the large amount and variety of data that can be generated in precision agriculture, it is increasingly necessary to create solutions for database and data mining to storage and subsequent use of these data. Regarding the storage of data, solutions involving data warehouses that deal with spatial and temporal dimensions should be generated in order to allow greater ease of the users in retrieving the necessary information for data mining tasks in a quick and organized way. The geostatistical techniques are very useful and widely used in most studies related to precision agriculture, but sometimes require a more accurate knowledge from the user to use them, as well as specific software for this. One of our future efforts will focus on developing techniques that allow preprocessing of this large amount of data directly in the database, using techniques such as spatial regression to creating spatio-temporal data warehouses able to properly prepare the information for use.

The arrangement of data from different sources and locations into a single spatial grid as accurately as possible, in order to provide better results in the pattern extraction process, is the great challenge of our research. For this purpose, uncertainties regarding the position and value of the predicted data from the original data should be handled. According to [23], there is a need to elaborate a spatial prediction method simple enough to remain tractable under uncertainty about unsampled data and realistic enough to provide reliable information about a region where some information is available.

With respect to the use of data, data mining also has an important role, and has been used in several studies through its grouping task. In our future efforts, we also intend to deal with algorithms related to clustering and its extensions, in order to find or propose a variation that better suits assist in the delineation of management zones for precision farming.

ACKNOWLEDGMENT

Thanks to Fazenda Aparecida for the availability of the area for data collection that allowed the achievement of this study. We also thank the following Brazilian research agencies: CNPq, CAPES, FAPESP and FINEP.

REFERENCES

[1] J.P. Molin, "Agricultura de Precisão: situação atual e perspectivas", in Milho: Estratégias de Manejo para Alta Produtividade, vol. I, A.L. Fancelli and D. Dourado-Neto, Eds. Piracicaba:ESALQ/USP, 2003, pp. 89-98.

[2] T.A. Doerge, "Management zone concepts", Site-specific management guidelines. Atlanta, USA: Potash and Phosphate Institute, 1999.

[3] L.S. Schiratsuchi, L.R. Queiros and G.C. Faccioni, "Classificação não-supervisionada no delineamento de zonas de manejo", Report of Research and Development. Planaltina, DF: Embrapa Cerrados, 2005, 16p.

[4] P.M.B Landim, "Sobre geoestatistica e mapas", Terra e Didática, vol. II, issue 1. Campinas, 2006, pp. 19-33.

[5] D.G. Krige, "A statistical approach to some basic mine valuation problems on the Witwatersrand", Jnl. Chem. Matal. Min. Soc. South Africa, vol. 52, 1951, pp. 119-139.

[6] C.R. Grego, L.M. Rabello, S.R. Brancalião, S.R. Vieira and A. Oliveira, "Geoestatística aplicada a condutividade elétrica do solo e altitude do solo", in Agricultura de precisão: um novo olhar, 1st ed., R.Y. Inamasu, J.M. Naime, A.V. Resende, L.H. Bassoi and A.C.C. Bernardi, Eds. São Carlos, SP: Embrapa, 2011, pp. 245-248.

[7] H. Wackernagel. Multivariate geostatistics, 3rd. ed. Berlin: Springer, 2003.

[8] S.R. Vieira, "Geoestatística aplicada à agricultura de precisão", in Agricultura de Precisão, A. Borém, M.P. Giúdice, D.M. Queiroz, E.C. Mantovani, L.R. Ferreira, F.X.R Valle and R.L. Gomide, Eds. Viçosa: UFV, 2000, pp. 93-108.

[9] J.R.P. Carvalho and E.D. Assad, "Comparação de interpoladores espaciais multivariados para precipitação pluvial anual no estado de São Paulo", Technical Report. Campinas: Embrapa, 2003, 5p.

[10] J. MacQueen, "Some methods for classification and analysis of multivariate observations", Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Berkley, CA: University of California Press, vol. 1, pp. 281-297, 1967.

[11] F. Aurenhammer, "Voroni diagrams – a survey of a fundamental geometric data structure", ACM Computer Surveys (CSUR). New York: ACM, vol. 23, issue 3, pp. 345-405, September 1991.

[12] J.C. Bezdek, R. Ehrlich and W. Full, "FCM: The fuzzy c-means clustering algorithm", Computers & Geosciences, vol. 10, issue 2, p. 191-203, 1984.

[13] N.R. Kitchen, K.A. Sudduth, D.B. Myers, S.T. Drummond and S.Y. Hong, "Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity", Computer and Electronics in Agriculture, vol. 46, issue 1-3, pp. 285-308, 2005.

[14] D. Gustafson and W. Kessel, "Fuzzy clustering with fuzzy covariance matrix", Decision and Control including the 17th Symposium on Adaptative Processes IEEE Conference. IEEE: San Diego, p. 761-766, 1978.

[15] J. Carletta, "Assessing agreement on classification tasks: the Kappa statistic", Computacional Linguistics, vol. 22, issue 2, pp. 249-254, 1996.

[16] B. Balasko, J. Abonyi, and B. Feil, "Fuzzy Clustering and Data Analysis Toolbox". Veszprem: Department of Process Engineering, University of Veszprem, 2005.

[17] G. Cardillo, "Cohen's Kappa". [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/15365-cohens-kappa.

[18] H. I. Witten, E. Frank and M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques", 3rd ed. Burlington: Morgan Kaufmann, 2011.

[19] S. Rezende, "Sistemas inteligentes: fundamentos e aplicações". Barueri: Manole, 2003.

[20] M.J. Egenhofer and J.R. Herring, "Categorizing Binary Topological Relations Between Regions, Lines, and Point in Geographic Databases", Technical Report. Department of Surveying Engineering, University of Maine, 28p, 1990.

[21] Veris Technologies. [Online]. Available: http://www.veristech.com.

[22] L.D. Corwin, M.L.K. Carrillo, P.J. Vaughan, J.D. Rhoades and D.G. Cone, "Evaluation of a GIS-linked mode of salt loading to groundwater", Journal of Environmental Quality, vol. 28, issue 2, pp. 471-480.

[23] K. Loquin and D. Dubois. "Kriging and epstemic uncertainty: a critical discussion", in Methods for Handling Imperfect Spatial Information. Springer Berlin Heidelberg, 2010, pp. 269-305.

Apendice D

UTILIZAÇÃO DE DADOS ESPACIAIS NA TAREFA DE AGRUPAMENTO DE DADOS APLICADA NA GESTÃO AGRÍCOLA

Este apêndice contém a transcrição do artigo científico intitulado *Utilização de Dados Espaciais na Tarefa de Agrupamento de Dados Aplicada na Gestão Agrícola*, aceito para publicação e em fase de edição para o periódico *Revista T.I.S. - Tecnologias, Infraestrutura e Software*.

Utilização de Dados Espaciais na Tarefa de Agrupamento de Dados Aplicada na Gestão Agrícola

Exploring Spatial Data in the Clustering Task Applied to Farm Management

Lucas Antoniale Callegari, Ricardo Rodrigues Ciferri, Eduardo Antonio Speranza

Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905 – São Carlos – SP – Brasil

lucascallegari93@gmail.com, ricardo@dc.ufscar.br, eduardo.speranza@dc.ufscar.br

Resumo. *A Agricultura de Precisão (AP) é uma prática agrícola baseada em tecnologia da informação e características de clima, do solo e das plantas para a gestão da produção de maneira sítio-específica. Os dados obtidos por meio dessa prática permitem ao usuário gerar sub-regiões homogêneas internas a uma área de produção conhecidas como unidades de gestão diferenciada (UGDs). Devido à grande quantidade e variedade de dados que pode ser coletada em uma área de produção agrícola, o delineamento de UGDs deve ser inserido em um processo computacional conhecido como Descoberta de Conhecimento em Bancos de Dados (DCB). A principal etapa desse processo é a mineração de dados, onde devem ser especificadas tarefas de agrupamento de dados sem rotulagem prévia, capazes de auxiliar o usuário na identificação das UGDs. O objetivo deste trabalho foi analisar as diferenças na obtenção de UGDs utilizando diferentes algoritmos de agrupamento a partir de configurações contendo apenas dados convencionais, e configurações incluindo dados de localização espacial (latitude e longitude). Essas análises foram realizadas de maneira visual, ou seja, em forma de mapas; e a partir de critérios de validação interna que avaliam o grau de coesão intragrupo e de separação intergrupo. Os resultados mostraram que dados de localização espacial influenciam significativamente no resultado final quando utilizados em associação com os dados convencionais, indicando a necessidade de tratamentos diferenciados por parte dos algoritmos de agrupamento.*

Palavras-chave: *agrupamento de dados espaciais, agricultura de precisão, unidades de gestão diferenciada.*

Abstract. *Precision Agriculture (PA) is a farm practice based on information technology and features of weather, soil and plants for site-specific production management. Data obtained from this practice allows its users to generate internal and homogeneous sub-regions to a production area known as management units. Due to the large amount and variety of collected data from an agricultural production area, the delineation of management units must be inserted in a computational process known as Knowledge Discovery in Databases (KDD). The main step in this process is data mining, which specifies clustering tasks for data without labeling, able to assist the user in identifying management units. The aim*

of this study was to analyze the differences in the delineation of management units using different clustering algorithms from settings containing only conventional data, and settings including spatial data (latitude and longitude). These analysis were performed visually, i.e., in the form of maps; and from internal validation criteria able to evaluate the achieved degree of intra-group cohesion and inter-group separation. Results showed that spatial data strong influence in the final result when used in association with conventional data, indicating the necessity for using spatial data to provide different treatments by clustering algorithms.

Keywords: *spatial data clustering, precision agriculture, management zones.*

1. Introdução

Em uma área de produção agrícola existem discrepâncias sobre a qualidade e a quantidade do produto final obtido, que podem estar associadas a diferenças relacionadas à gênese do solo e ao potencial produtivo da área. Na agricultura tradicional, insumos e corretivos agrícolas são distribuídos de maneira uniforme dentro de uma lavoura, e essas discrepâncias não são consideradas. Ações desse tipo podem prejudicar o meio ambiente e diminuir o lucro obtido, uma vez que nem todas as áreas necessitam da mesma quantidade de insumos, e outras necessitam de mais insumos do que realmente foi aplicado. Nesse contexto, técnicas de Agricultura de Precisão (AP) devem ser utilizadas, de maneira a impulsionar o gerenciamento sítio-específico da produção (MOLIN et al., 2015). Dentre essas técnicas, destaca-se o delineamento de unidades de gestão diferenciada (UGDs), o qual utiliza dados relacionados ao clima, solo e plantas obtidos a partir de ferramentas de tecnologia da informação, para subdividir as áreas de cultivo em sub-regiões suficientemente homogêneas com relação a esses dados. Doerge et al. (1999), definem UGDs como regiões espaciais internas a uma área de produção agrícola e que expressam combinações de fatores limitantes à produtividade. Já Taylor et al. (2007) definem UGDs como áreas espacialmente contíguas para as quais um tratamento particular deve ser aplicado. Sob a perspectiva de mineração de dados, Ruß (2012) define UGDs como sendo a transformação de coleções de dados espaciais pontuais obtidos em uma área agrícola em agrupamentos para uma finalidade específica. Em geral, as UGDs devem ser consideradas como regiões com mínima variabilidade, imutáveis, insensíveis a tratamentos agronômicos de rotina e independentes de fatores antrópicos (MOLIN et al., 2015).

Em termos práticos e levando-se em consideração a definição de Ruß (2012), o delineamento de UGDs deve ser considerado, no contexto da tecnologia da informação, como um processo de Descoberta de Conhecimento em Banco de Dados (DCB). O processo de DCB compreende diversas etapas, onde a principal delas é a mineração de dados (FAYYAD et al., 1996). Nessa etapa, quando existe a necessidade de encontrar dados similares e que não estão previamente rotulados, algoritmos de agrupamento devem ser utilizados. Os resultados obtidos com a utilização desses algoritmos podem ser avaliados, com relação à sua estrutura, por critérios de validação interna que permitem avaliar o grau de coesão intergrupo e de separação intragrupo obtidos (VENDRAMIN et al., 2010).

Este trabalho teve como objetivo realizar um estudo sobre a etapa de mineração do DCB com uso de dados espaciais aplicada no auxílio ao delineamento de UGDs em AP. Para tanto, foram utilizados dados espaciais referentes à cultura de cana-de-açúcar com diferentes atributos, entre eles condutividade elétrica do solo em diferentes

profundidades, altitude, e dados históricos de produtividade. Diferentes algoritmos de agrupamento e configurações utilizando ou não os atributos espaciais de latitude e longitude foram avaliados visualmente por meio de um Sistema de Informações Geográficas (SIG) e estruturalmente a partir de critérios de validação interna conhecidos na literatura. As próximas seções estão distribuídas como se segue: a seção 2 descreve resumidamente os algoritmos de agrupamento e os critérios de validação internos; a seção 3 descreve resumidamente a metodologia; a seção 4 descreve os principais resultados obtidos; e a seção 5 contém as conclusões.

2. Algoritmos de agrupamento e critérios de validação

Nesta seção estão descritos, de maneira resumida, os algoritmos de agrupamento e critérios de validação interna que foram utilizados neste trabalho.

2.1 Algoritmos de agrupamento

O principal objetivo dos algoritmos de agrupamento é agrupar amostras de maneira natural, fazendo com que aquelas mais similares com relação aos atributos do domínio sejam atribuídas a um mesmo grupo e aquelas menos similares sejam atribuídas a grupos distintos. Essa similaridade pode ser determinada por medidas algébricas conhecidas, como a distância Euclidiana para dados numéricos, e a correspondência simples para dados nominais (WITTEN et al., 2011). Diferentes classes de algoritmos podem ser utilizadas para o agrupamento de dados, e dentre elas estão os algoritmos particionais e hierárquicos, descritos a seguir.

2.1.1. Algoritmos de Agrupamento Particionais

Os algoritmos particionais fornecem ao usuário uma única solução de agrupamento, com a quantidade de grupos *k* previamente fornecida pelo usuário. Essa classe de algoritmos trabalha com o intuito de minimizar uma função objetivo, normalmente o erro quadrático médio entre as amostras de cada grupo e o seu centro considerando todas as dimensões dos dados de entrada (centroide).

O algoritmo particional mais conhecido e utilizado na literatura por conta de sua simplicidade é o *k-means* (MACQUEEN et al., 1967). Esse algoritmo gera partições rígidas da base de dados (conhecidas como *hard* ou *crisp*), fazendo com que cada objeto pertença a um único grupo. Os principais passos para execução do *k-means* são:

1) Escolher aleatoriamente os *k* protótipos (centroides) para os grupos, onde *k* é definido pelo usuário.
2) Atribuir cada amostra ao grupo com centroide mais próximo.
3) Recalcular o centroide de cada grupo como sendo a média entre todas as amostras do grupo.
4) Repetir os passos 2 e 3 até que algum critério de convergência seja obtido, como número máximo de iterações ou limiar mínimo de mudanças nos centroides.

O *k-means* é mais susceptível a erros quando os grupos naturais do conjunto de dados são de diferentes tamanhos e densidades e possuem formas não globulares. Contudo, possui vantagens como o fato de ser simples, intuitivo, eficaz em muitos

cenários de aplicação, produzindo resultados de interpretação relativamente simples, sendo assim considerado um dos dez mais influentes algoritmos em mineração de dados (WU et al., 2008). As principais desvantagens do *k-means* são a necessidade de informar um valor inicial para *k*, e ser sensível à inicialização dos protótipos, podendo o resultado final fornecido tratar-se de um mínimo local e não global.

Similar ao *k-means*, o algoritmo *Fuzzy c-means* (FCM) (BEZDEK et al., 1984) possui como principal característica permitir a sobreposição de amostras em grupos distintos. Desse modo, cada amostra possui um grau de pertinência para cada grupo, gerando partições não rígidas da base de dados, conhecidas como *soft*. O FCM pode ser utilizado, ao invés do *k-means*, em aplicações onde os dados utilizados e os resultados obtidos podem possuir natureza *fuzzy*, ou seja, onde não existem mudanças bruscas entre os grupos formados, como é o caso dos dados espaciais utilizados em AP.

2.1.2. Algoritmos de Agrupamento Hierárquicos

Ao contrário dos algoritmos particionais, onde a quantidade de grupos deve ser previamente definida, os algoritmos hierárquicos constroem hierarquias de partições, que podem ser constituídas por todas as quantidades de grupos possíveis. Os métodos clássicos para agrupamento hierárquico são normalmente baseados na construção de hierarquias *bottom-up* (aglomerativo) ou *top-down* (divisivo). O método aglomerativo inicia associando cada amostra a um grupo, e em cada passo encontra o par de grupos mais similar para unir, repetindo-se esse processo até que todas as amostras estejam reunidas em um único grupo. O método divisivo, por sua vez, inicia com todas as amostras em um único grupo, onde o grupo com amostras menos similares é subdividido em dois novos grupos em cada passo, até que cada amostra forme um grupo por si só (JAIN; DUBES, 1988). Por conta das dificuldades em se encontrar pontos de divisão em conjuntos de dados, os métodos divisivos praticamente não são utilizados, fazendo com que este trabalho se concentrasse apenas em métodos aglomerativos.

Na literatura, os algoritmos hierárquicos aglomerativos *single-linkage* (SNEATH, 1957) *complete-linkage* (SØRENSEN, 1948) e *average-linkage* (SOKAL, 1958) são os mais utilizados. O algoritmo *average-linkage*, utilizado neste trabalho, representa um compromisso entre o *single-linkage* e o *complete-linkage*. As similaridades entre os grupos são calculadas considerando a distância média entre pares de amostras de grupos distintos, favorecendo a presença de grupos em formas globulares e menos sensíveis a ruídos. Outro método hierárquico aglomerativo bastante utilizado é o algoritmo de *Ward* (WARD, 1963), considerado como análogo hierárquico ao *k-means*, pois em cada passo da hierarquia procura fundir os grupos que minimizam o aumento do erro quadrático. Assim como o *average-linkage*, esse método tende a favorecer grupos em formas globulares e menos sensíveis a ruídos.

2.2. Critérios de Validação Interna

Validação é um termo que se refere de forma ampla aos diferentes procedimentos para avaliar de maneira objetiva e quantitativa os resultados de análise de agrupamento. Os critérios de validação interna avaliam o grau de compatibilidade entre a estrutura de grupos sob avaliação e os dados, utilizando apenas os próprios dados (JAIN; DUBES, 1988). Para este trabalho, foram utilizados dois critérios de validação interna presentes na literatura, descritos a seguir.

2.2.1 Critério SD

O critério *SD* (HALKIDI et al., 2000) se baseia na média de espalhamento dos dados, sendo definido a partir de duas quantidades: *SD_Scat*, que é a dispersão média dos grupos; e *SD_Dist*, que é a separação total entre os grupos.

O valor de *SD_Scat* é definido considerando um vetor de variâncias *V* contendo as variâncias de cada um dos *p* atributos convencionais do conjunto de dados: $V = (Var(V_1), \ldots, Var(V_p))$. Do mesmo modo, definem-se variâncias vetoriais $V^{(K)}$ para cada grupo C_k: $V^{(k)} = (Var(V_1^{(k)}), \ldots, Var(V_p^{(k)}))$. O valor de *SD_Scat* é calculado então como sendo a média das normas dos vetores $V^{(K)}$ dividida pela norma do vetor *V*, como segue:

$$SD_{Scat} = \frac{\frac{1}{K}\sum_{k=1}^{K}\|V^{\{k\}}\|}{\|V\|}, \text{ onde K é o número máximo de grupos obtidos}$$

Já o valor de *SD_Dis* considera inicialmente a maior (D_{max}) e menor (D_{min}) distância entre os centroides dos *k* grupos obtidos:

$$D_{max} = max_{k \neq k'} \left\| G^{(k)} - G^{(k')} \right\|$$

$$D_{min} = min_{k \neq k'} \left\| G^{(k)} - G^{(k')} \right\|$$

A partir desses valores, *SD_Dis* é calculado como se segue:

$$SD_{Dis} = \frac{D_{max}}{D_{min}} \sum_{k=1}^{K} \frac{1}{\sum_{\substack{k'=1 \\ k' \neq k}}^{K} \left\| G^{\{k\}} - G^{\{k'\}} \right\|}, \text{ onde K é o número máximo de grupos obtidos}$$

Assim, temos que o critério *SD* é finalmente definido como:

$$SD = \alpha * SD\_Scat + SD\_Dis$$

onde α é um peso igual ao valor de *SD_Dis* obtido pela partição com o maior número de grupos, a fim de comparar as várias partições de dados. Apesar de serem considerados componentes de coesão e separação do critério *SD*, os valores de *SD_Scat* e *SD_Dis* foram utilizados individualmente nas análises realizadas neste trabalho.

2.2.2 Critério da Largura de Silhueta

O critério da *largura de silhueta* (Rousseaw et al., 1987), assim como o *SD*, tem o objetivo de avaliar agrupamentos a partir da coesão intragrupos e da separação intergrupos. No entanto, sua metodologia de cálculo é executada inicialmente em nível de amostra, até chegar a um valor total para o agrupamento gerado.

Primeiramente, para cada amostra M_i, a sua distância média para cada grupo deve ser considerada. O valor de *a(i)* é a distância média da amostra M_i com relação às outras amostras do grupo a qual ela pertence ($M_{i'}$). Considerando que $M_i \in C_k$, então temos:

$$a(i) = \frac{1}{n_k - 1} \sum_{\substack{i' \in I_k \\ i' \neq i}} d(M_i, M_{i'}),$$

onde I_k é o conjunto de amostras de C_k e n_k é a quantidade de amostras de C_k.

Similarmente, deve ser calculada a distância média $\delta(M_i, C_K')$ de M_i para as amostras pertencentes a cada um dos outros grupos C_k':

$$\delta(M_i, M_{i'}) = \frac{1}{n_{k'}} \sum_{i' \in I_{k'}} d(M_i, M_{i'})$$

Assim, é obtido o valor de *b(i)*, que é a menor distância entre as distâncias médias:

$$b(i) = min_{k' \neq k} \delta(M_i, M_{i'})$$

O valor *k'* para qual é obtido o valor mínimo indica a melhor escolha para reagrupar, se necessário, a amostra M_i em outro grupo diferente do que esta atualmente pertence.

Para cada amostra M_i, temos então o cálculo de *s(i)*:

$$s(i) = \frac{b(i) - a(i)}{max\left(a(i), b(i)\right)}$$

Esse valor é chamado de *largura de silhueta* da amostra, representado por valores entre -1 e 1. Valores próximos a 1 indicam que a amostra M_i tende a estar corretamente agrupada, enquanto que valores próximos de -1 indicam que a amostra poderia ter sido melhor associada a outro grupo.

A partir desses valores, obtém-se a *largura de silhueta* para um determinado grupo C_k, denotada por S_k:

$$S_k = \frac{1}{n_k} \sum_{i \in I_k} s(i)$$

Finalmente, a *largura de silhueta* global do agrupamento é a média das larguras de silhueta de todos os grupos, denotada por:

$C = \frac{1}{K} \sum_{k=1}^{K} S_k$, onde K é a quantidade máxima de grupos

3. Metodologia

A metodologia para o desenvolvimento deste trabalho considerou três etapas. Inicialmente, foram realizados estudos sobre o funcionamento dos algoritmos particionais *k-means* e FCM, e dos algoritmos hierárquicos *average-linkage* e *Ward's*, bem como dos critérios de validação interna *SD* e *largura de silhueta.* A segunda etapa constituiu-se do planejamento e execução dos algoritmos utilizando dados reais, com diferentes configurações utilizando ou não as coordenadas geográficas durante a tarefa de agrupamento. Finalmente, a terceira etapa constituiu-se de análises dos resultados obtidos, a partir da utilização dos critérios de validação interna para verificação da coesão e separação obtidas nos agrupamentos gerados e análises visuais dos agrupamentos em forma de mapas, já sendo considerados como UGDs.

4. Resultados e Discussão

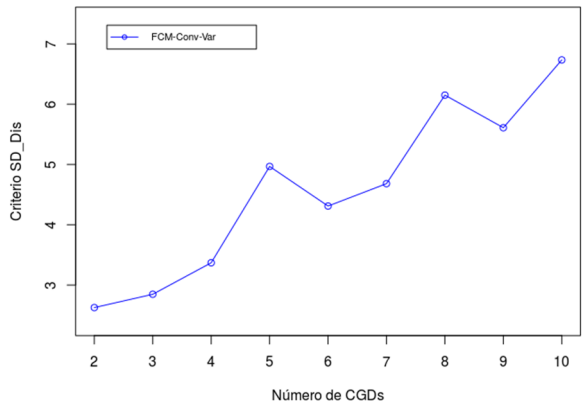
Nesta seção serão apresentados os dados reais utilizados e os resultados obtidos durante a execução deste trabalho. A Figura 1, abaixo, mostra o contorno da área piloto onde foram coletados os dados. Essa área corresponde a um talhão de aproximadamente 17 hectares de área, e está localizada na Fazenda Aparecida, em Mogi-Mirim, SP. Os dados convencionais correspondem a medidas de condutividade elétrica do solo nas profundidades de 30 e 90 cm, obtidas no ano de 2010; cota altimétrica, obtidas no ano de 2010; e produtividade, obtidas nos anos de 2010, 2012 e 2013. Todos os dados foram previamente interpolados em uma grade espacial regular de 20 m, constituindo 415 amostras georreferenciadas compostas pelos atributos convencionais supracitados e coordenadas geográficas de latitude e longitude, referenciadas nos resultados como atributos espaciais.

Figura 1 - Área piloto de coleta dos dados utilizados no trabalho

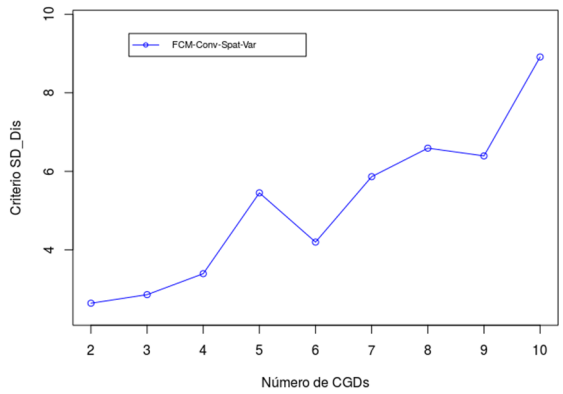
Os resultados de execução dos critérios de validação interna são compostos por gráficos de linha, que apresentam o valor de cada critério obtido em níveis progressivos, variando de 2 até 10 grupos. Essa variação serve para análise dos critérios de validação e verificação da tendência das curvas. Porém, na aplicação prática, são utilizados normalmente de 2 a 5 grupos, com cada um deles representando uma classe de gestão diferenciada (CGD), que pode ser constituída por uma ou mais UGDs não adjacentes espacialmente. Essa análise considerou os critérios que obtiveram maior variação, por isso não serão apresentados todos os critérios para o mesmo tipo de agrupamento e nem o mesmo tipo de critério para todos os agrupamentos, evitando-se assim repetições de conclusões. Para verificação dos efeitos da inclusão dos atributos espaciais nos algoritmos de agrupamento, em todos os experimentos foram utilizados apenas os atributos convencionais nos vetores utilizados pelas fórmulas que calculam os critérios de validação interna.

De acordo com o que foi descrito na seção 2, um bom nível de separação de grupos é desejável, pois na teoria de agrupamento de dados, quanto mais separados os grupos estão, maiores são as chances de se obter agrupamentos válidos capazes de serem diferenciados pelas características analisadas.

Considerando o algoritmo de particionamento FCM e o cálculo de *SD_Dist* para o critério de validação *SD*, a Figura 2 mostra a diferença entre os valores de separação obtidos para agrupamentos gerados sem e com a utilização dos atributos espaciais.



(a) Atributos convencionais

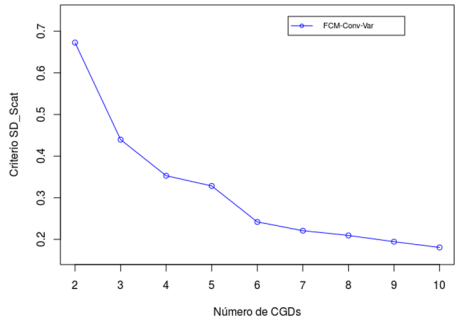


(b) Atributos convencionais e espaciais

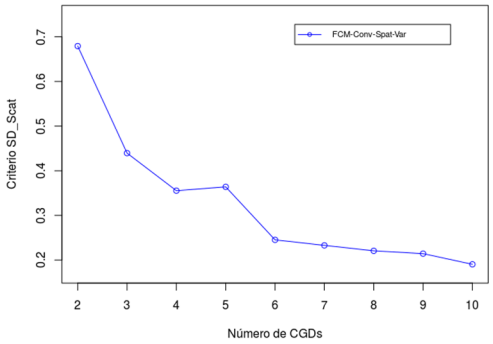
Figura 2 – Cálculo de SD_Dist para o algoritmo *k-means*: (a) Agrupamento utilizando apenas atributos convencionais e (b) Agrupamento utilizando atributos convencionais e espaciais.

Ainda de acordo com a equação do cálculo de *SD_Dist*, valores menores garantem que os grupos estão mais bem separados, ou seja, no gráfico estão próximos de 0. Analisando os dois gráficos, podemos notar que a utilização de atributos espaciais proporcionou uma diminuição na separação de grupos (Figura 2(b)) do que quando estes não foram utilizados (Figura 2(a)).

Ainda considerando o algoritmo de particionamento FCM e o cálculo de *SD_Scat*, para o critério *SD*, a Figura 3 mostra a diferença entre os valores de coesão obtidos para agrupamentos gerados sem e com a utilização dos atributos espaciais.



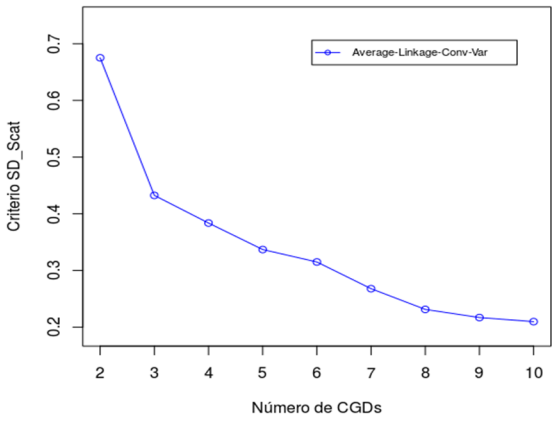
(a) Atributos convencionais



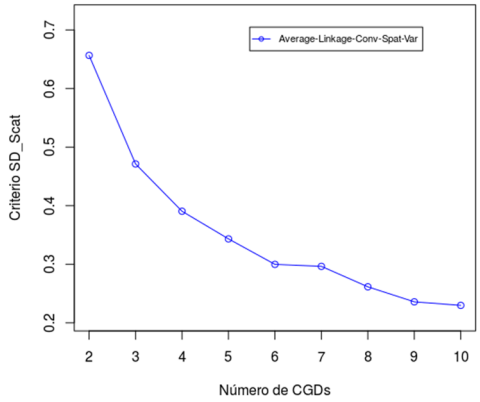
(b) Atributos convencionais e espaciais

Figura 3 – Cálculo de SD_Scat para o algoritmo FCM: (a) Agrupamento utilizando apenas atributos convencionais e (b) Agrupamento utilizando atributos espaciais e convencionais.

Ainda de acordo com os cálculos da seção 2, valores menores de *SD_Scat* garantem que os grupos estão mais coesos, ou seja, no gráfico estão próximos de 0. Assim como ocorreu para os valores de *SD_Dis*, os valores de *SD_Scat* obtidos incluindo os atributos espaciais contribuíram para diminuir, mesmo que muito sutilmente, a coesão entre os grupos. Um resultado similar para valores de *SD_Scat* pode ser observado considerando o agrupamento gerado pelo algoritmo hierárquico *average-linkage*, conforme exibido na Figura 4:



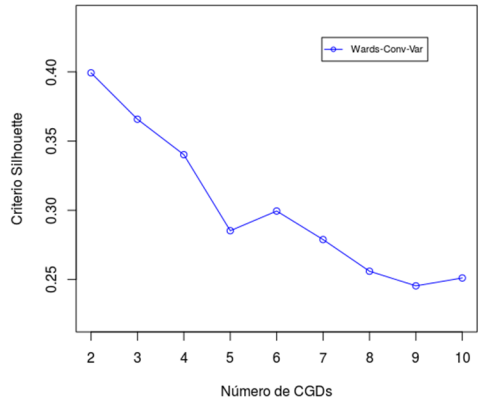
(a) Atributos Convencionais



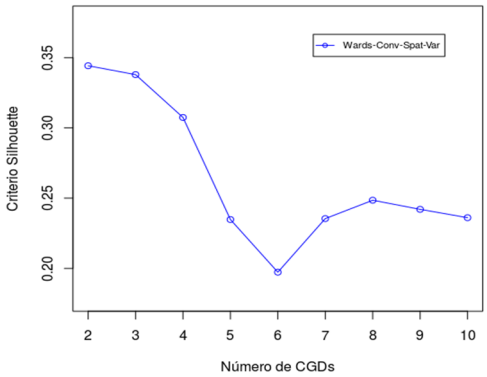
(b) Atributos Convencionais e Espaciais

Figura 4 – Cálculo de SD_Scat para o algoritmo *average-linkage*: (a) Agrupamento utilizando apenas atributos convencionais e (b) Agrupamento utilizando atributos espaciais e convencionais.

Em um novo experimento, considerando o algoritmo *Ward's* e o critério de validação *largura de silhueta*, a Figura 5 mostra a diferença de valores do critério entre agrupamentos gerados sem e com a utilização dos atributos espaciais.



(a) Atributos convencionais



(b) Atritubos convencionais e espaciais

Figura 5 – Cálculo da largura de silhueta utilizando o algoritmo de *Ward*: (a) Agrupamento utilizando apenas atributos convencionais. (b) Agrupamento utilizando atributos espaciais e convencionais.

De acordo com o que foi descrito nos cálculos da seção 2, o critério *largura de silhueta* garante que os dados estão mais bem agrupados com relação à coesão e separação quando os seus valores estão mais próximos de 1. O que podemos notar, a partir da Figura 5, é que a inclusão dos atributos espaciais no agrupamento contribuiu, em geral, para a diminuição dos índices do critério, indicando uma dúvida maior sobre a correta alocação das amostras em seus respectivos grupos, ou seja, são obtidos grupos menos coesos e com menor separação entre si.

Outras análises puderam ser realizadas a partir da visualização das UGDs em forma de mapas. As Figuras 6, 7, 8 mostram, respectivamente, como seria a delineamento das UGDs considerando os algoritmos FCM, *average-linkage* e *Ward's* para diferentes quantidades de grupos, sem e com a utilização de atributos espaciais.

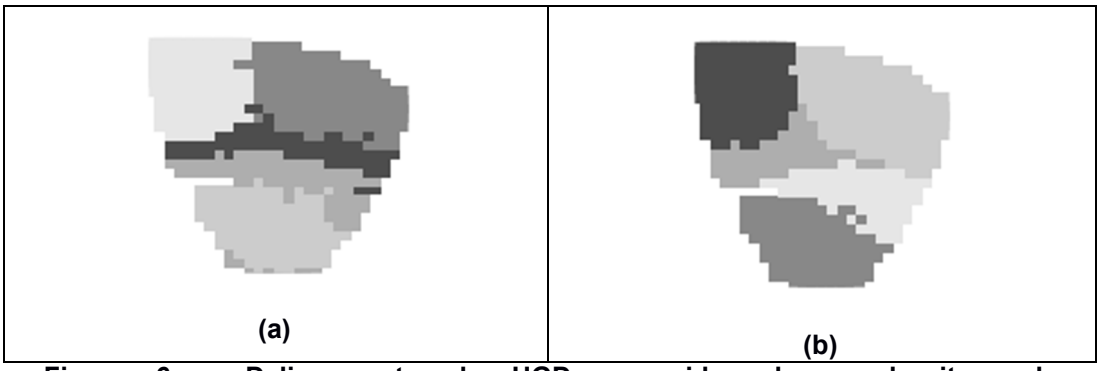


Figura 6 – Delineamento de UGDs, considerando o algoritmo de particionamento FCM com *k*=5: (a) Sem a utilização de atributos espaciais e (b) Com a utilização de atributos espaciais.

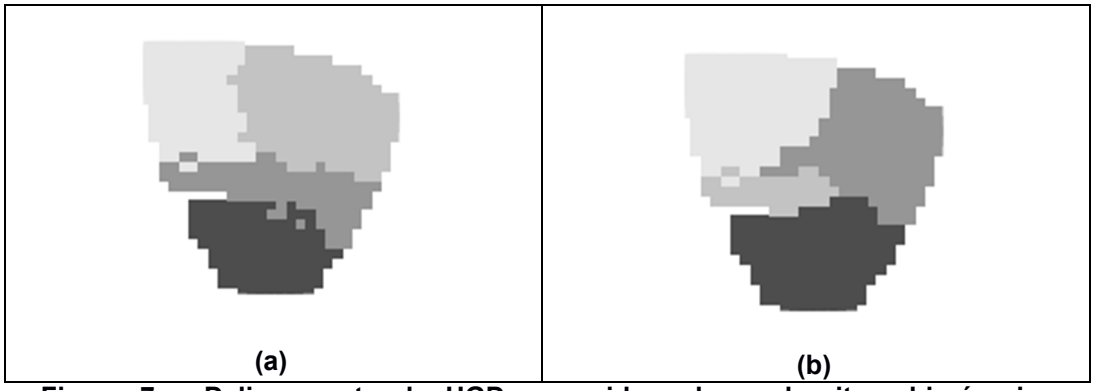


Figura 7 – Delineamento da UGDs, considerando o algoritmo hierárquico *average-linkage* para cortes na hierarquia em *k*=4: (a) Sem a utilização de dados espaciais e (b) Com a utilização de atributos espaciais.

Figura 8 – Delineamento de UGDs, considerando o algoritmo hierárquico *Ward's* para cortes na hierarquia em *k*=3: (a) Sem a utilização de atributos espaciais e (b) Com a utilização de dados espaciais.

Por meio das figuras acima, podemos notar que, a utilização de coordenadas geográficas pode contribuir para reduzir o nível de estratificação dos mapas. A estratificação está relacionada à perda de contiguidade espacial das CGDs, quando estas são visualizadas em forma de mapa, fazendo com que as mesmas sejam constituídas por diversas UGDs fragmentadas. Por outro lado, CGDs geradas dessa maneira podem não ser ideais para a aplicação, pois a utilização das coordenadas geográficas pode enviesar a obtenção de formas geométricas convexas. No caso da Figura 8, por exemplo, o resultado final foi bastante alterado, o que pode "esconder" características relacionadas ao solo e as plantas que podem ser importantes para o manejo da lavoura.

5. Conclusões

A inclusão dos atributos espaciais proporcionou um prejuízo nos resultados obtidos pelos critérios de validação interna, sendo estes mais sutis com relação aos critérios *SD_Scat* e *SD_Dis*, que tratam a coesão e a separação de maneira independente; e mais fáceis de visualizar com relação ao critério da *largura de silhueta*, que trata a coesão e a separação ao mesmo tempo. Com relação aos mapas de UGDs obtidos, notou-se que a inclusão de dados espaciais proporcionou a obtenção de áreas menos estratificadas. No entanto, ao mesmo tempo em que essas formas melhoram a acuidade visual dos resultados, características importantes de solo e da cultura em questão podem ser inibidas. Em um estudo similar ao apresentado neste trabalho, Santos et al. (2012) chegaram a uma conclusão semelhante, porém utilizaram apenas um algoritmo de agrupamento e medidas simples de desvio padrão ao invés de critérios para validação estrutural agrupamentos.

Complementarmente, conclui-se que a utilização de coordenadas geográficas como atributos convencionais deve ser utilizada com cautela em tarefas de agrupamento, devendo estas ser tratadas de maneira diferenciada por conta de sua natureza complexa, ou como forma de restringir ou permitir a fusão ou a separação de determinados grupos.

6. Agradecimentos

Agradecemos à Fazenda Aparecida e aos pesquisadores Célia Grego e Luiz Vicente, da Embrapa Monitoramento por Satélite, pela coleta dos dados utilizados neste trabalho. O segundo autor foi apoiado pela bolsa de produtividade em pesquisa 311868/2015-0, do CNPq.

7. Referências Bibliográficas

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM : The Fuzzy c-Means Clustering Algorithm. Computer & Geosciences, v. 10, n. 2-3, p. 191–203, 1984.

DOERGE, T. A. Management Zone Concepts. Site-Specific Management Guidelines, n. 2, p. 4, 1999.

FAYYAD, U.; SMYTH, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. v. 39, n. 11, p. 27–34, 1996.

HALKIDI, M.; VAZIRGIANNIS, M.; BATISTAKIS, Y. Quality Scheme Assessment in the Clustering Process. In: ELOMAA, T.; MANNILA, H.; TOIVONEN, H. (Eds.). . Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. v. 2431p. 265–276.

JAIN, A. K.; DUBES, R. C.; OTHERS. Algorithms for clustering data. [s.l.] Prentice hall Englewood Cliffs, 1988. v. 6

MACQUEEN, J. Some methods for classification and analysis5th Berkeley Symposium on Mathematical Statistics and Probability. Anais...University of California, 1967

MOLIN, José Paulo; DO AMARAL, Lucas Rios; COLAÇO, André. Agricultura de precisão. Oficina de Textos, 2015.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. v. 20, p. 53–65, 1987.

RUß, G. Spatial Data Mining in Precision Agriculture. [s.l.] Otto-von-Guericke-University of Magdeburg, 2012.

SANTOS, R. T.; SARAIVA, A. M.; MOLIN, J. P. Avalição do uso das coordenadas geográficas como parte do conjunto de atributos para definição de unidades de gerenciamento diferenciado. Congresso Brasileiro de Agricultura de Precisão - COnBAP. Anais...Ribeirão Preto, SP: 2012.

SNEATH, Peter HA. The application of computers to taxonomy. Microbiology, v. 17, n. 1, p. 201-226, 1957.

SOKAL, Robert R. A statistical method for evaluating systematic relationships. Univ Kans Sci Bull, v. 38, p. 1409-1438, 1958.

SØRENSEN, Thorvald. {A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}. Biol. Skr., v. 5, p. 1-34, 1948.

TAYLOR, J. A.; MCBRATNEY, A. B.; WHELAN, B. M. Establishing Management Classes for Broadacre Agricultural Production. Agronomy Journal, v. 99, n. 5, p. 1366, set. 2007.

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining, v. 3, n. 4, p. 209–235, 2010.

WARD JR, Joe H. Hierarchical grouping to optimize an objective function. Journal of the American statistical association, v. 58, n. 301, p. 236-244, 1963.

WITTEN, I. H.; FRANK, E.; HALL, M. A. Data Mining: Practical Machine Learning Tools and Techniques. 3rd. ed. San Francisco, CA; London: Morgan Kaufmann, 2011.

WU, Xindong et al. Top 10 algorithms in data mining. Knowledge and information systems, v. 14, n. 1, p. 1-37, 2008. MINERAÇÃO DE DADOS ESPACIAIS. Disponível em: < http://www.dsc.ufcg.edu.br/~sampaio/cursos/2008.2/PosGraduacao/MineracaoDeDados/Apresentacoes/MineracaoDeDadosEspaciais.pdf> Acesso em: 28 abr. 2016.

WITTEN, I. H.; FRANK, E.; HALL, M. A. Data Mining: Practical Machine Learning Tools and Techniques. 3rd ed. San Francisco, CA; London: Morgan Kaufmann, 2011. 664 p. ISBN 978-0-12-374856-0. Acesso em: 28 abr. 2016.

Apendice E

CLUSTERING APPROACHES AND ENSEMBLES APPLIED IN THE DELINEATION OF MANAGEMENT CLASSES IN PRECISION AGRICULTURE

Este apêndice contém a transcrição do artigo científico intitulado *Clustering Approaches and Ensembles Applied in the Delineation of Management Classes in Precision Agriculture*, apresentado oralmente e publicado nos anais do *XVII Brazilian Symposium on Geoinformatics (GEOINFO)* (SPERANZA; CIFERRI; CIFERRI, 2016).

Clustering Approaches and Ensembles Applied in the Delineation of Management Classes in Precision Agriculture

Eduardo A. Speranza[1], **Ricardo R. Ciferri**[2], **Cristina D. A. Ciferri**[3]

[1]Embrapa Agricultural Informatics – Brazilian Research Agricultural Corporation
13083-886 – Campinas – SP – Brazil

eduardo.speranza@embrapa.br

[2]Departament of Computer Science – Federal University of São Carlos
13.565-905 – São Carlos – SP – Brazil

ricardo@dc.ufscar.br

[3]Department of Computer Science – University of São Paulo at São Carlos
13.560-970 – São Carlos – SP – Brazil

cdac@icmc.usp.br

Abstract. *This paper describes an experiment performed using different approaches for spatial data clustering, aiming to assist the delineation of management classes in Precision Agriculture (PA). These approaches were established from the partitional clustering algorithm Fuzzy c-Means (FCM), traditionally used in this context, and from the hierarchical clustering algorithm HACC-Spatial, especially designed for this PA task. We also performed experiments using traditional ensembles approaches from the literature, evaluating their behavior to achieve consensus solutions from individual clusterings obtained from features splitting or running one of the abovementioned algorithms. Results showed some differences between FCM and HACC-Spatial, mainly for the visualization of management classes in the form of maps. Considering the consensus clusterings provided by ensembles, it became clear the attempt to achieve an agreement result that most closely matches the original clusterings, showing us some details that may go undetected when we analyse only the individual clusterings.*

1. Introduction

Precision Agriculture (PA) is an agricultural management system driven by spatio-temporal variability of soil and culture features of a crop. These parameters may be obtained from particular procedures and techniques based on information technology, remote sensing and Global Positioning System (GPS) [Molin 2003, Vendrusculo and Kaleita 2011]. Unlike conventional agriculture, where agricultural inputs and correctives are evenly applied across the cultivation area, PA enables its users to manage them in a site-specific way, aiming the maximization of profit cutting of yield limiting factors. Moreover, this system allows farmers to fit crop needs and supply of inputs, helping to reduce the environmental damage [Schwalbert et al. 2014]. Because of its highly dependency of the spatio-temporal variability built-in data collected on the field, the adoption of decision-making processes based on PA suggests data collection at high spatial resolutions. However, this usually is not possible for most farmers, because several factors such

as the high cost of acquiring satellite images and gathering data on the field, beyond the need to acquire services and automated machinery able to perform variable rate interventions. In these cases, the delineation of subfields spatially internal to the crop area, which the internal spatial variability is so negligible as to allow for evenly distributed internal interventions, is a way to disseminate the adoption of PA even using accurate spatial resolutions (e.g. between 10 and 30 meters). These subfields, known as management classes, may be composed by one or many spatially contiguous areas in the coordinates space, known as management zones [Taylor et al. 2007]. Taking into account these concepts, it is really intuitive to relate the delineation of management classes with traditional clustering algorithms, such as Fuzzy c-Means (FCM) [Bezdek et al. 1984]. However, PA tasks produce complex and non-conventional data, composed by two distinct spaces: features, regarding the events occurring in the crop; and coordinates, regarding the spatial location where these events took place. Thereby, because of its complexity, the coordinates space must to be handled in different ways by clustering algorithms. With the purpose of solving this challenge, Ruß and Kruse 2011 developed an agglomerative hierarchical clustering algorithm, known as HACC-Spatial. The HACC-Spatial enables the delineation of management classes preserving the spatial contiguity as much as possible, in order to facilitate easy visual interpretation of the user while maintain the coherence of the clustering obtained by events related to soil and plants.

Using algorithms composed by different features and parameters, such as FCM and HACC-Spatial, to solve clustering problems present in any domain, can generate different results and hence questions regarding which of them is the best solution. In order to clarify such questions, several approaches enabling consensual and more robust clusterings have been emerged in the literature. These clusterings, known as ensembles, must be obtained from different ways, such as individual clusterings using different kinds of algorithms, parameters configurations or subsets of features at the same data set [Ghosh and Acharya 2011]. Our work described in this paper were aimed to evaluate, from internal clustering validation measures, the accuracy of clusterings representing management classes that were obtained individually using the FCM and HACC-Spatial algorithms, as well as using more robust and consensual clustering ensembles to consolidate individual results and feature space partitioning.

The remainder of the paper is structured as follows. In section 2, we briefly describe the FCM and HACC-Spatial algorithms and approaches commonly used to delineate management classes in PA, beyond the ensemble approach used in our work. In section 3, we present the methodology used for the experiments. In section 4, we present results for experiments using reald data. Finally, in section 5, we present our conclusions and provide suggestions for future work proposals.

2. Background and Related Work

Some clustering algorithms have been used to assist the delineation of management zones in PA. Nevertheless, most of the approaches available in the literature use the Fuzzy c-Means algorithm (FCM) as a basis for this task. Based on the standard clustering algorithm k-means [MacQueen et al. 1967], the Fuzzy c-Means algorithm (FCM) [Bezdek et al. 1984] calculates, at each iteration, the membership (ω_k) of each data sample with respect to each one of the desired clusters. This calculation takes into account the distance (d) from any particular data sample to each cluster centroid and a fuzzification parameter

(m), defined by the user with default value of 2. At the end of each iteration, clusters centroids are recalculated taking into account all dataset samples and their membership values to each cluster. Instead of k-means, FCM convergence results not only to assign each sample to a unique cluster (hard clustering), but in a membership matrix with 0 to 1 values for each sample with respect to each cluster, known as fuzzy partition matrix (soft clustering). This matrix is one of the FCM advantages regarding hard clustering algorithms, providing better results for situations that have a difficult separation and overlapping datasets. However, like k-means, FCM centroids are randomly initialized, making the results susceptible to a local minima.

The main reason for using FCM in the context of this application is linked with the fact that abrupt changes do not occurs in soil and plant attributes in small enough parcels of the crop, causing input data and the obtained clusters to consider a membership degree. Over the years, several approaches in the literature using FCM and considering different types of these attributes have been developed. Brock et al. 2005 used FCM to delineate management zones considering historical yield data from corn-soybean rotation crops, indentifying the spatial association of the obtained maps with soil maps. Already Kitchen et al. 2005 used FCM to delineate management zones considering ratios of soil electrical conductivity (EC) in different depths (bulk of EC) and relief data, comparing them with yield zones obtained from historical yield data. As a result, it was found that the bulk of EC combined with relevant data are strong indications for management zones. Similar conclusions were obtained by Morari et al. 2009, including measures of soil and electrical resistivity data. The work of Li et al. 2007 used, in addition with abovementioned attributes, features indicating rates of organic matter and biomass. In this case, due to the large number of attributes, an intermediate phase of principal component analysis before getting the management zones by FCM was performed. High-resolution satellite images also appears as inputs to obtain management zones using the FCM, as in works of Song et al. 2009 and Zhang et al. 2010. More recently, Milne et al. 2012 used FCM to find management zones from smoothed spatial data obtained from three different methods. The results were compared with crop responses regarding the application of different nitrogen rates. The work of Scudiero et al. 2013 shows, using FCM to obtain management zones, that combined bare-soil and EC data can contribute to find spatial variability of a crop. The KM-sPC approach [Córdoba et al. 2013] allowed to show the importance of a principal component analysis considering the coordinate space to reduce the stratification provided by FCM when management zones are displayed in form of maps. This approach were used again in a pratical nitrogen management of wheat [Peralta et al. 2015]. The study of Chang et al. 2014 compared management zones generated by FCM using reflectance data regarding the soil properties and productivity, showing that it is feasible the use of an active canopy sensor for this PA application.

Despite the widespread use of FCM for this task, the coordinates space of PA datasets, composed by spatial coordinates variables (e.g., latitude and longitude), have been used only in preprocessing steps or to show the management classes provided by clustering in the form of maps. This fact does not prevent the use of these maps by automated machinery for variable rate interventions, but the reduction of spatial contiguity, causing stratification of management classes in too many areas, can confuse visual analysis by experts. In order to solve this problem, the HACC-Spatial hierarchical clustering algorithm were developed by Ruß and Kruse 2011. This approach takes into account spa-

tial restrictions for clustering samples, and considers a preprocessing step to perform an initial tessellation of them in small spatial clusters, using the k-means algorithm at the coordinates space. Such subdivision aims to reduce computational costs by decreasing the number of steps of the construction of the hierarchical tree (or dendrogram) produced by the algorithm, regarding the geostatistics principle claiming that spatially very close samples tends to have close enough values in the features space [Matheron 1963]. As a result, a structure similar to a Voronoi diagram should be obtained by the preprocessing step. From this moment, each dendrogram step merges the most similar clusters, according to the feature space. First, only spatially adjacent clusters can be merged, providing the maintenance of spatial contiguity. However, when a user-defined contiguity threshold *cp* is reached, this restriction is switched off. This threshold is associated to the ratio of the average distances between the samples belonging to adjacent clusters and the average distances between samples belonging to non-adjacent clusters.

Because of the differing nature of FCM and HACC-Spatial (partitional and hierarchical, respectively) and the spatial restrictions used for one of them, are expected distinct clustering results for the same dataset, making it difficult for the user to choose the best approach. A feasible solution to solve this question can be achieved using ensembles. Ensembles are able to combine multiple sample clusterings in a unique and consolidated one, known as consensus solution. These kind of approach can be used to meet several requirements, such as: increase the quality of the solution, providing more robust clusterings; select models; reuse knowledge; find consensus between clusterings obtained from subsets of features or subsamples, among others [Ghosh and Acharya 2011].

The main aim of a clustering ensemble is to find a consensus solution composed by an unique clustering to share as much information as possible derived from original clusterings. This sharing can be measure by the average of normalized mutual information (ANMI), where the desired optimal value is ANMI equal to 1 [Strehl and Ghosh 2002]. The main goal of the three ensembles algorithms developed by Strehl and Ghosh 2002 is to build general approaches to obtain consensus from individual clusterings aiming at maximizing the ANMI value. These algorithms were evaluated by the authors in scenarios where individual clusterings were composed by distinct features, distinct subsamples or distinct clustering algorithms. The Cluster-based Similarity Partitioning Algorithm (CSPA) is the simplest and most obvious heuristic. It is based on the fact that two objects have a similarity of 1 if they are in the same cluster and 0 otherwise. Thus, a n x n binary matrix, where n is the number of samples, is created for each original clustering. To recluster these samples, a similarity-based clustering algorithm based on graph partitioning is used [Karypis and Kumar 1998]. The computational and storage complexity of this algorithm are both quadratic in n. The HyperGraph Partitioning Algorithm (HPGA) addresses the clustering ensemble as a hypergraph partitioning problem, where hyperedges represent the original given clusters as indications of strong bonds. To recluster the samples, a partitional hypergraph algorithm, cutting a minimal number of hyperedges is used [Han et al. 1997]. In this case, while CPSA only considers pairwise relationships, HPGA includes original clustering relationships. Finally, the Meta-Clustering Algorithm (MCLA) represent each cluster by a hyperedge, and then group and collapse related hyperedges (or clusters), attaching each sample to the collapsed hyperedge in which it belongs more actively. At the end, a graph-based clustering of hyperedges is performed, indentifying consolidated ”clusters of clusters”. In contrast to CPSA, HPGA e MCLA

have linear computational and storage complexity. Still according to [Strehl and Ghosh 2002], the MCLA tends to provide better ANMI values when the consensus solution were obtained from individual clusterings with low noise rates and diversity; and HPGA and CSPA are usually better were obtained from individual clusterings with high noise rates and diversity.

From the abovementioned algorithms, it were possible for us to prepare some experiments, described in section 3, combining distinct approaches that can be applied in the delineation of management classes in PA. Results of these experiments are presented in section 4.

3. Methodology

The methodology used in ours experiments follows the concepts of Knowledge Discovery in Databases (KDD). According to Fayyad et al. 1996 and Weiss and Indurkhya 1998, at least three main steps of KDD process should be taken into account when it will be used: preprocessing, data mining (or pattern extraction) and post processing. The planned activities for each one of these steps, in the context of management classes in PA, are described below.

3.1 Preprocessing

The preprocessing step comprises the changes that should be made in a raw dataset when it will be used by a KDD process, preparing it to the next steps. Regarding to spatial data, in addition to very common preprocessing activities, such as standardization, cleaning and feature selection, the spatial interpolation must be performed in order to accommodate data samples in a single and regular spatial grid [Vieira 2000]. This activity is required, because PA datasets are caught using different kinds of sensors and samples densities, usually at distinct spatial spots in the same area. Another important activities in this step are: verifying data distribution using probabilistic density functions, as a preassessment of possible distortions that can occur in clustering algorithms when using non-Gaussians distributed features; verifying features correlations, using methods such as Pearson's Coefficient Correlation [Benesty et al. 2009]; and data standardization, reducing the bias caused by features with highly predominant scales relative to the others.

3.2 Data Mining

The data mining step can be viewed as an iterative process, where should be used different solutions to improve the accuracy of the results. In the context of our work, due to the fact that datasets had no previous classification, clusterings tasks need to be considered. Therefore, the approaches to be used are classified as non-supervised machine learning algorithms [Mitchell 1997]. In this step, we used the HACC-Spatial and FCM algorithms in the traditional way and also combining results by ensembles. HACC-Spatial was run using non-spatial features of the whole dataset to calculate dissimilarity values at each step of dendrogram, and spatial features to build the initial tessellation and to support adjacency treatments at each step of dendrogram (Approach I). In the other hand, FCM was run in its traditional way, i.e., using only non-spatial features (Approach II). Regarding ensembles, it was created an approach to found consensus clusterings from individual results provided by Approach I and Approach II (Approach III); and another two approaches to found consensus clustering from individual results provided by non-spatial

features subsets of soil, altimetry and yield using HACC-Spatial (Approach IV) and FCM (Approach V). The ensembles approaches was run using CSPA, HPGA and MCLA algorithms described in section 2, and the results with best values of ANMI were chosen as the best solution for each approach.

According to domain expert users, at least 2 and at most 5 management classes should be considered for a crop [Molin et al. 2015]. Thereby, the five abovementioned approaches were run using k=2 to 5 clusters for the experiments, when using FCM (partitional), and the same values for dendrogram cuts, when using HACC-Spatial (hierarchical). Regarding to dissimilarity measures, the Euclidean distance were used for all approaches. In relation to other parameters and customizations, for approaches using FCM, the standard fuzzification value m=2 was fixed, and samples were associated with the cluster where were achieved a higher membership degree. For approaches using HACC-Spatial, were used a binding criteria similar to average-linkage algorithm [Sokal 1958], because of its ability to handle data sets with presence of outliers. Other HACC-Spatial parameters, like initial tessellation number of clusters (k) and cp, were defined during the experiments.

3.3 Post Processing

Finally, in the post processing step, we used two internal validation criteria: the SD criteria and the silhouette width criteria. These criteria allow comparing and evaluating the effectiveness of the five approaches when they are run at the same number of clusters. The SD criteria [Halkidi et al. 2000, Halkidi and Vazirgiannis 2001] allows to verify, for each obtained clustering, how cohesive and well separated are the clusters, from average values of intra-cluster variance and distances between clusters centroids. In this case, optimal values should be closer to 0. The silhouette width criteria [Rousseeuw 1987] follows the same principles of SD, but using dissimilarity values of a sample regarding its associated cluster and the nearest neighbor cluster. In this case, values closer to 1 indicates that the sample has been allocated to the correct cluster; and values closer to -1 indicates that the sample could have been better allocated to the nearest neighbor cluster. According to Vendramin et al. 2010, the silhouette width criteria, in comparison to other internal criteria in the literature, can provide, in general, more effective assessments about the internal structure of the clusters.

4. Experiments

In this section, we present the results obtained from experiments using real data, following the methodology described in section 3. These data are composed by samples collected on an experimental crop field of sugarcane culture. This field has an area around 17 hectares belonging to Fazenda Aparecida, located in Mogi-Mirim, São Paulo state, Brazil, with central coordinates 7505136N (latitude) and 299621E (longitude), given the spatial reference system UTM Zone 23S. Figure 1 shows the contour shape and a cropped image of the experimental field.

The raw datasets used in our work comprises measures of soil electrical conductivity (EC), in milisiemens per meter; altimetry quota, in meters; and historical yield, in tons per hectare or culms per square meter. The samples were collected at different times and by different sensors or processes, providing us six conventional features associated with spatial coordinates: soil electrical conductivity at 30 e 90 cm deep in 2010

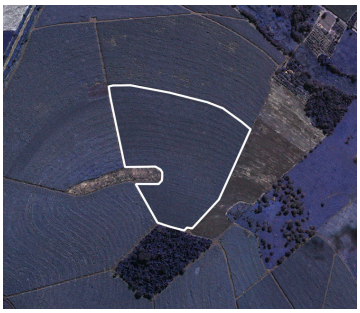


Figure 1. Experimental crop field of sugarcane (white contour) with a cropped image in the background provided by the World View 2 satellite (April 30, 2011).

(EC30 and EC90); altimetry quota (Quota); and historical yield in 2010 (Yield2010), 2012 (Yield2012) and 2013 (Yield2013). It is worth mentioning the need for historical yield data, because they could be considered susceptible to anthropic and climatic factors over the years. In addition, the rainfall data of the whole farm in the agricultural years should be considered to support some analysis: 1601 mm in 2010 (July 2009 to June 2010), 1538 mm in 2012 (July 2011 to June 2012) and 1599 mm in 2013 (July 2012 to June 2013). The probabilistic density distribution of EC30, EC90 and Yield2010 features could be described by Gaussians, with most values around the mean. On the other hand, the distributions of Yield2012 and Yield2013 indicates, respectively, predominance of higher and lower yield values, probably affected by the abovementioned factors. A special case occurs with the Quota feature, where average values are the minority because the experimental area has a slight slope and narrow in the central region. These distributions are shown in Figure 2.

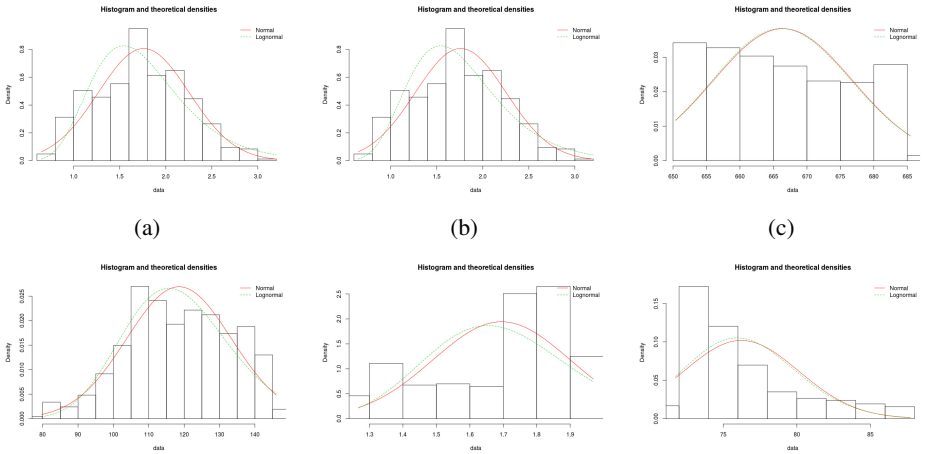


Figure 2. Probabilistic density distributions of dataset features: (a) EC30; (b) EC90; (c) Quote; (d) Yield2010; (e) Yield2012; e (f) Yield2013.

Applying the Pearson's Coefficient Correlation between pairs of features, were verified that EC30 and EC90 hold the most positive correlation of the dataset. In general,

the Quote feature was well correlated with all other features, and negatively (oppositely) correlated with Yield2010. Regarding to yield data, Yield2012 and Yield2013 features are highly correlated, and negatively correlated with Yeld2010 feature. The negative correlation of Yield2010 with other yield years could be influenced again by the anthropic and climatological factors.

Using the concepts of preprocessing described above, the dataset features were interpolated in a single regular spatial grid with spatial resolution of 20 meters. This value was calculated using the average coordinates spacing between samples for each one of the six features of the original data set. Simple algorithms, like the average of k nearest neighbors [Altman 1992], were used to interpolate features with higher sample densities. On the other hand, more sophisticated algorithms, like kriging [Matheron 1969], were used to interpolate features with smaller sample densities. After applying this process, each dataset feature were distributed in 415 samples spatially represented by points with latitude and longitude coordinates. Figure 3 shows raw samples of soil electrical conductivity (high density) and yield (medium density) and their respective interpolated samples in the same regular spatial grid. Lower values are represented by lighter colors, while higher values are represented by darker colors.

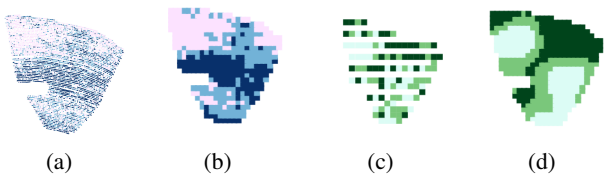


Figure 3. Example of raw and interpolated data in 3 classified intervals: (a) EC30 raw data (9046 samples); (b) EC30 interpolated data (415 samples); (c) Yield2010 raw data (111 samples); (d) Yield2010 interpolated data (415 samples).

Especially for the HACC-Spatial algorithm, when it was run in the context of approaches I, III and IV, the *cp* parameter was set to 0.5, according to the best results obtained by Ruß and Kruse 2011. Initial tessellation (k) was set to 200, after checking a significant increase in internal variance of the clusters for the following levels of the dendrogram.

Figures 4 and 5 show, respectively, linear charts containing values achieved by both SD and silhouette width criteria for the five proposed approaches, regarding k values between 2 and 5. Through these charts, we can observe better results for k=3, where we can found, in general, smaller values of SD and larger values of silhouette width.

By analyzing the results using ensembles, the charts of figures 4 and 5 show us that the approach IV, in the most of cases, achieved poor results regarding both the internal criteria. Therefore, we can conclude that the heuristic of HACC-Spatial algorithm, considering spatial relationships during the construction of the hierarchy, tends to be more consistent when using all features (approach I) than when using individual clusterings by features split to obtain a subsequent consensus by ensembles (approach IV). On the other hand, approach V achieved better results than approach IV, showing that in some cases consensus solutions from individual FCM clusterings by features division can be used to replace solutions provided by approach II. Finally, the approach III results shown, for all k

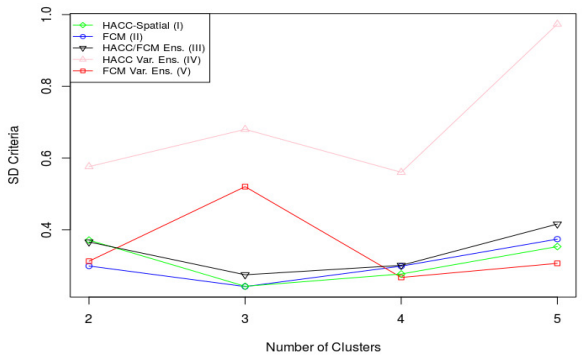


Figure 4. SD criteria values. Each line corresponds to values of SD achieved by the respective approach, considering *k*=2 to 5 clusters.

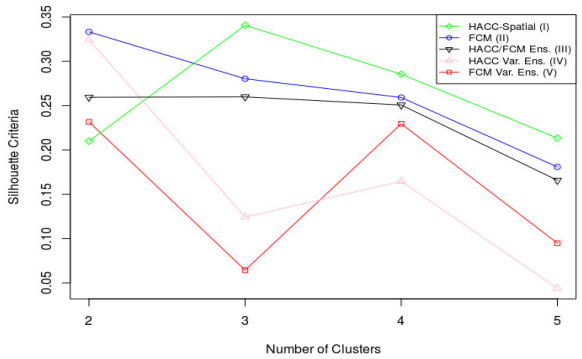
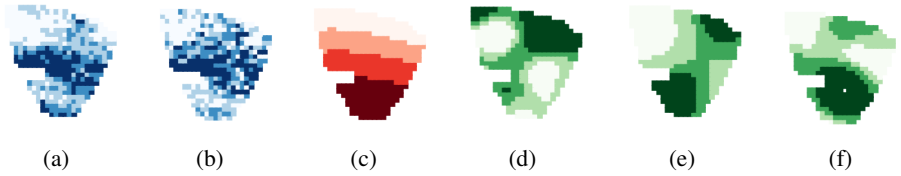


Figure 5. Silhouette width criteria values. Each line corresponds to values of SD achieved by the respective approach, considering *k*=2 to 5 clusters.

values, attempts to find consensus from clusterings obtained by approaches I and II, with slight variations in both the internal criteria values.

Beyond the analysis using internal criteria, we used the visualization of management classes in the form of maps to perform some observations. These analysis were performed from k=5 to 2 clusters, in order to observe some effects of agglomerative hierarchy provided by HACC-Spatial approaches. For k=5, management classes exhibited generally pronounced stratification, hindering the analysis and an accurate understanding by expert users. For k=4, were used, for comparsion with the clustering results, the interpolated dataset from each feature, classified in 4 classes of equal intervals (Figure 6). For each feature, lighter colors represent samples with higher values, while darker colors represent samples with lower values.

Figure 7 shows the results obtained by approaches I, II and III for k=4. As can be seen, the results are quite similar for management classes identified with the same color. We can observe the shaping of an isolated area on the top left of the map (blue),



(a) (b) (c) (d) (e) (f)

Figure 6. Interpolated data classified in 4 equal intervals: (a) EC30; (b) EC90; (c) Quote; (d) Yield2010; (e) Yield2012; e (f) Yield2013.

(a) (b) (c)

Figure 7. Results for *k*=4: (a) HACC-Spatial (I); (b) FCM (II); and (c) HACC/FCM Ensemble(III).

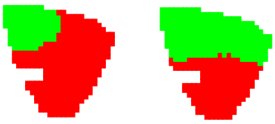
representing a low elevation region with lower rates of soil EC and historical yield. At the green area, also located in a low elevation region, can be observed medium values of yield and soil EC. Already at the red area, corresponding to a middle elevation region, can be observed a strong influence of extreme values of soil EC for its formation. Finally, the yellow area, located at a high elevation region, shows higher rates of yield.

Figure 8 shows the results obtained for k=3, regarding the approaches I, II and III, where were achieved the lowest value of SD criteria (approach II) and the highest value of silhouette width criteria (approach I) of the whole experiment. From this figure, can be observed that approaches I and II achieved very similar results. In both cases, the green and blue areas obtained for k=4 were practically kept. The main difference between these results is focused at the subdivision between green and red areas. While approach I is forced to merge two clusters because of the hierarchical caracteristics of HACC-Spatial, making the red area be composed by the most similar areas in k=4 (red and yellow), the approach II recalculates again which are the clusters where all samples should be assigned, promoting a greater amount of change. Neverthless, the differences observed between both approaches are quite small, which may still be noticed a strong influence of the low frequency of medium values of Quota in approach II, contributing for the user to clearly note the red region with higher values and blue and green regions with lower values of this feature. Regarding to ensemble approaches, Figure 8 (c) further reinforces that approach III, in turn, tried to find a consensus for these subdivision differences, turning the final map quite stratified.

Finally, for k=2 (Figure 9), we can verify many differences between approach I, that achieved the worst value of silhouette width criteria, and approach II, that achieved the best values for both internal criteria. While approach I strongly took into account low levels of historical yield in order to identify an isolated area at the top left region (green), merging clusters representing green and red classes for k=3, the approach II was affected again by the low frequency of average altitude values, clearly separating a low (green) from a high elevation region (red).

(a) (b) (c)

Figure 8. Results for *k*=3: (a) HACC-Spatial (I); (b) FCM (II); and (c) HACC/FCM Ens.(III).



(a) (b)

Figure 9. Results for *k*=2: (a) HACC-Spatial (I); e (b) FCM (II).

5. Conclusions and Future Work

If we take into account visual analysis and measures of cohesion and separation provided by SD criteria, approaches purely based on FCM (II e V) achieved, in general, better results in comparison to the approaches using the HACC-Spatial (I, III e IV) . Due to the fact that FCM is based in *k-means* algorithm, its bias is always performed to achieve the minimization of intracluster variance and maximization of intercluster dissimilarity. Because of this, internal criteria based on these measures, like SD and silhouette width, tends to provide suitable results for clusterings obtained by this algorithm. On the other hand, in some cases the visual perception of the expert user, of major importance in PA tasks, may be harmed. However, for the silhouette width criteria, these approaches achieved, in general, worst results in relation to those obtained by approach I, except for k=2. These results were likely influenced by intrinsic FCM fuzzy features, which can generate doubts if a sample was properly associated with a particular cluster or whether it will be better allocated to the nearest neighbor cluster.

Regarding to the use of ensembles, splitting of features (approaches IV and V) was important for clarifying some details that can get unnoticed in clusterings obtained using all features. However, the high stratification rates generated in the final maps can be very harmful to the users analysis. In the consensus approach between different kinds of algorithms (III), we can observe an increased stratification, causing damage to the visual user analysis. On the other hand, were observed slight variations in SD and silhouette width criteria for different values of k, indicating that this approach can be used as solution in some specific cases.

The ensembles approach used in this work is rather general and try to find consensus using only final clusterings obtained from splitting of features or from different algorithms. In an future work, could be used ensembles approaches that allow extracting the main features of each algorithm, making useful data like the membership values provided by FCM, might be used to obtain a better consensus solution.

6. Acknowledgement

Thanks to Fazenda Aparecida for the availability of the area for data collection that allowed the achievement of this study, and to researchers from Embrapa, Celia Grego and Luiz Vicente, responsible for collecting such data. We also thank the following Brazilian research agencies: CNPq, CAPES, FAPESP and FINEP. The second author have been supported by the grant 311868/2015-0 from CNPq.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing SE - 5*, volume 2 of *Springer Topics in Signal Processing*, pages 1–4. Springer Berlin Heidelberg.

Bezdek, J. C., Ehrlich, R., and Full, W. (1984). FCM : The Fuzzy c-Means Clustering Algorithm. *Computer & Geosciences*, 10(2-3):191–203.

Brock, A., Brouder, S. M., Blumhoff, G., and Hofmann, B. S. (2005). Defining Yield-Based Management Zones for Corn-Soybean Rotations. *Agronomy Journal*, 97(4):1115–1128.

Chang, D., Zhang, J., Zhu, L., Ge, S. H., Li, P. Y., and Liu, G. S. (2014). Delineation of management zones using an active canopy sensor for a tobacco field. *Computers and Electronics in Agriculture*, 109:172–178.

Córdoba, M., Bruno, C., Costa, J., and Balzarini, M. (2013). Subfield management class delineation using cluster analysis from spatial principal components of soil variables. *Computers and Electronics in Agriculture*, 97:6–14.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM*, 39(11):27–34.

Ghosh, J. and Acharya, A. (2011). Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):305–315.

Halkidi, M. and Vazirgiannis, M. (2001). Clustering validity assessment: finding the optimal partitioning of a data set. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 187–194.

Halkidi, M., Vazirgiannis, M., and Batistakis, Y. (2000). Quality scheme assessment in the clustering process. In Zighed, D., Komorowski, J., and Zytkow, J., editors, *Principles of Data Mining and Knowledge Discovery*, volume 1910 of *Lecture Notes in Computer Science*, pages 265–276. Springer Berlin Heidelberg.

Han, E.-H., Karypis, G., Kumar, V., and Mobasher, B. (1997). Clustering based on association rule hypergraphs. In *DMKD*, page 0.

Karypis, G. and Kumar, V. (1998). Multilevelk-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed computing*, 48(1):96–129.

Kitchen, N., Sudduth, K., Myers, D., Drummond, S., and Hong, S. (2005). Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity. *Computers and Electronics in Agriculture*, 46(1-3):285–308.

Li, Y., Shi, Z., Li, F., and Li, H.-Y. (2007). Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. *Computers and Electronics in Agriculture*, 56(2):174–186.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, 58(8):1246–1266.

Matheron, G. (1969). Le krigeage universel.

Milne, A. E., Webster, R., Ginsburg, D., and Kindred, D. (2012). Spatial multivariate classification of an arable field into compact management zones based on past crop yields. *Computers and Electronics in Agriculture*, 80:17–30.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York; London.

Molin, J. P. (2003). Agricultura de Precisão: Situação atual e perspectivas. In Fancelli, A. L. and Neto, D. D., editors, *Milho: Estratégias de Manejo para Alta Produtividade*, pages 89–98. ESALQ/USP/LPV, Piracicaba.

Molin, J. P., do Amaral, L. R., and Colaço, A. (2015). *Agricultura de precisão*. Oficina de Textos.

Morari, F., Castrignanò, a., and Pagliarin, C. (2009). Application of multivariate geostatistics in delineating management zones within a gravelly vineyard using geo-electrical sensors. *Computers and Electronics in Agriculture*, 68(1):97–107.

Peralta, N. R., Costa, J. L., Balzarini, M., Castro Franco, M., C??rdoba, M., and Bullock, D. (2015). Delineation of management zones to improve nitrogen management of wheat. *Computers and Electronics in Agriculture*, 110:103–113.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0):53–65.

Ruß G. and Kruse, R. (2011). Exploratory hierarchical clustering for management zone delineation in precision agriculture. In Perner, P., editor, *Advances in Data Mining. Applications and Theoretical Aspects*, volume 6870 of *Lecture Notes in Computer Science*, pages 161–173. Springer Berlin Heidelberg.

Schwalbert, R. A., Amado, T. J. C., Gebert, F. H., Santi, A. L., and Tabaldi, F. (2014). Zonas de manejo: atributos de solo e planta visando a sua delimitação e aplicações na agricultura de precisão. *Revista Plantio Direto*, pages 21–32.

Scudiero, E., Teatini, P., Corwin, D. L., Deiana, R., Berti, A., and Morari, F. (2013). Delineation of site-specific management units in a saline region at the Venice Lagoon margin, Italy, using soil reflectance and apparent electrical conductivity. *Computers and Electronics in Agriculture*, 99:54–64.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, 38:1409–1438.

Song, X., Wang, J., Huang, W., Liu, L., Yan, G., and Pu, R. (2009). The delineation of agricultural management zones with high resolution remotely sensed data. *Precision Agriculture*, 10(6):471–487.

Strehl, A. and Ghosh, J. (2002). Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3:583–617.

Taylor, J. A., McBratney, A. B., and Whelan, B. M. (2007). Establishing Management Classes for Broadacre Agricultural Production. *Agronomy Journal*, 99(5):1366–1376.

Vendramin, L., Campello, R. J. G. B., and Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235.

Vendrusculo, L. G. and Kaleita, A. L. (2011). Modeling zone management in precision agriculture through Fuzzy C-Means technique at spatial database. In *2011 Louisville, Kentucky, August 7 - August 10, 2011*, St. Joseph, MI. American Society of Agricultural and Biological Engineers.

Vieira, S. R. (2000). Geoestatistica em Estudos de Variabilidade Espacial do Solo. In Novais, R. F. and Alvarez, V H, S. G. R., editors, *Tópicos em ciência do solo*, pages 1–54. Sociedade Brasileira de Ciência do Solo, Viçosa, MG, 1 edition.

Weiss, S. M. and Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Zhang, X., Shi, L., Jia, X., Seielstad, G., and Helgason, C. (2010). Zone mapping application for precision-farming: a decision support tool for variable rate application. *Precision Agriculture*, 11(2):103–114.

Apendice F

UTILIZANDO ENSEMBLES COM ABORDAGENS DE AGRUPAMENTO ESPACIAL PARA O DELINEAMENTO DE CLASSES DE MANEJO EM AGRICULTURA DE PRECISÃO

Este apêndice contém a transcrição do artigo científico intitulado *Utilizando Ensembles com Abordagens de Agrupamento Espacial para o Delineamento de Classes de Manejo em Agricultura de Precisão*, aceito para publicação e em fase de edição para o periódico *Revista Brasileira de Cartografia*.

UTILIZANDO *ENSEMBLES* COM ABORDAGENS DE AGRUPAMENTO ESPACIAL PARA O DELINEAMENTO DE CLASSES DE MANEJO EM AGRICULTURA DE PRECISÃO

Using Ensembles with Spatial Clustering Approaches Applied in the Delineation of Management Classes in Precision Agriculture

Eduardo Antonio Speranza[1]
Ricardo Rodrigues Ciferri[2]

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informatica Agropecuária
Av. Andre Tosello, 209 – Barão Geraldo – Campinas – SP – CEP 13083-886 – Brasil
eduardo.speranza@embrapa.br

Universidade Federal de São Carlos
Departamento de Computação
Rod. Washington Luis, km 235 – SP-310 – São Carlos – SP – CEP 13565-905 – Brasil
ricardo@dc.ufscar.br

RESUMO

Este artigo descreve experimentos realizados utilizando diferentes abordagens para agrupamento de dados espaciais, com o objetivo de auxiliar no delineamento de classes de manejo em Agricultura de Precisão (AP). Essas abordagens foram estabelecidas a partir do algoritmo de agrupamento particional Fuzzy c-Means (FCM), tradicionalmente utilizado em AP, e do algoritmo de agrupamento hierárquico HACC-Spatial, especialmente desenvolvido para AP. Também foram realizados experimentos utilizando diferentes abordagens de *ensembles* para agrupamentos disponíveis na literatura, avaliando o seu funcionamento para obter soluções de consenso para agrupamentos individuais obtidos a partir do particionamento do conjunto de atributos ou da utilização exclusiva do FCM ou do HACC-Spatial. Os resultados obtidos mostraram algumas diferenças entre o FCM e o HACC-Spatial, principalmente com relação a visualização das classes de manejo em forma de mapas. O algoritmo HACC-Spatial, alcançou, de uma maneira geral, melhores resultados quando comparado ao FCM e as abordagens de *ensembles*. Levando-se em consideração os agrupamentos consensuais obtidos pelas abordagens de *ensembles*, ficou evidente a tentativa de se obter resultados concordantes que se aproximam das soluções fornecidas pelos agrupamentos originais, proporcionando o aumento ou a diminuição da estratificação dos mapas de classes de manejo.

Palavras chaves: Agricultura de precisão, classes de manejo, agrupamento de dados espaciais, ensembles.

ABSTRACT

This paper describes experiments performed using different approaches for spatial data clustering, aiming to assist the delineation of management classes in Precision Agriculture (PA). These approaches were established from the partitional clustering algorithm Fuzzy c-Means (FCM), traditionally used in PA, and from the hierarchical clustering algorithm HACC-Spatial, especially designed for PA. We also performed experiments using different clustering ensembles approaches, evaluating their behavior to achieve consensus solutions from individual clusterings obtained from attribute splitting or using exclusively FCM or HACC-Spatial. The achieved results exhibited some differences between FCM and HACC-Spatial, mainly for the visualization of management classes in the form of maps. The HACC-Spatial algorithm achieved, in general, better results when compared to FCM and ensembles approaches.

Regarding the consensus clusterings provided by ensembles, we can point out the attempt to achieve agreement results which most closely matches the original clusterings, decreasing or increasing the stratification of the management classes maps.

Keywords: precision agriculture, management classes, spatial data clustering, ensembles.

1. INTRODUCTION

Precision Agriculture (PA) is an agricultural management system driven by spatio-temporal variability of soil and culture attributes of a crop. These parameters may be obtained from particular procedures and techniques based on information technology, remote sensing and Global Positioning System (GPS) (MOLIN, 2003; VENDRUSCULO & KALEITA, 2011). Unlike conventional agriculture, where agricultural inputs and correctives are evenly applied across the cultivation area, PA enables its users to manage them in a site-specific way, allowing farmers to fit crop needs and supply of inputs (SCHWALBERT et al., 2014). Therefore, the main aim of PA is to increase yield in a sustainable way, reducing the environmental inpacts with the site-specific use of agricultural inputs and, consequently, increasing the profit (BERNARDI et al., 2014). Because of its highly dependency of the spatio-temporal variability built-in data collected on the field, the adoption of decision-making processes based on PA suggests data collection at high spatial resolutions. However, this usually is not possible for most farmers, because several factors such as the high cost of acquiring satellite images and gathering data on the field, beyond the need to acquire services and automated machinery able to perform variable rate interventions. In these cases, the delineation of subfields spatially internal to the crop area, which the internal spatial variability is so negligible as to allow for evenly distributed internal interventions, is a way to disseminate the adoption of PA even using accurate spatial resolutions (e.g. between 10 and 30 meters). These subfields, known as management classes, may be composed by one or many spatially contiguous areas in the coordinate space, known as management zones (TAYLOR et al., 2007). Taking into account these concepts, it is really intuitive to relate the delineation of management classes with traditional clustering algorithms, such as Fuzzy c-Means (FCM) (BEZDEK et al., 1984). However, PA tasks produce complex and non-conventional data, composed by two distinct spaces: attribute space, regarding the events occurring in the crop; and coordinate space, regarding the spatial location where these events took place. Thereby, because of its complexity, the coordinate space must to be handled in different ways by clustering algorithms. With the purpose of solving this challenge, Ruß & Kruse (2011) developed an agglomerative hierarchical clustering algorithm, known as HACC-Spatial. The HACC-Spatial enables the delineation of management classes preserving the spatial contiguity as much as possible, in order to facilitate easy visual interpretation of the user while maintain the coherence of the clustering obtained by events related to soil and plants.

Using algorithms composed by different attributes and parameters, such as FCM and HACC-Spatial, to solve the delineation of management classes in PA, may generate different results and hence questions regarding which of them is the best solution. In order to clarify such questions, several approaches enabling consensual and more robust clusterings have been emerged in the literature. These clusterings, known as ensembles, must be obtained from different ways, such as individual clusterings using different kinds of algorithms, parameters configurations or subsets of attributes at the same data set (GHOSH & ACHARYA, 2011).

In a preliminary version of this work, clustering ensembles approaches based on graph and hypergraph partitioning (STREHL & GHOSH, 2002) were evaluated in their ability to provide more robust management classes maps regarding both the individual clusterings from FCM and HACC-Spatial algorithms using all available attribute space, and the same clustering algorithm spliting the attribute space (SPERANZA et al., 2016). Here, we extend this evaluation performing new experiments, now using a more recent and simple clustering ensemble approach, based on evidence accumulation obtained accross the individual clusterings (FRED & JAIN, 2005). Therefore, it was possible to achieve a more complete evaluation on which situations each approach should be used, either individually or using clustering ensembles.

The remainder of this paper is structured as follows. In section 2, we describe the FCM and HACC-Spatial algorithms and approaches commonly used to delineate management classes in PA, beyond the ensemble approaches used in this work and in its preliminary version. In section 3, we present the methodology used for the experiments. In section 4, we present results for experiments using real data. Finally, in section 5, we present our conclusions and provide suggestions for future work proposals.

2. BACKGROUND AND RELATED WORK

Some clustering approaches have been used to assist the delineation of management classes in PA. Nevertheless, most of the approaches available in the literature use the Fuzzy c-Means algorithm (FCM) as a basis for this task. Based on the standard clustering algorithm k-means (MACQUEEN et al., 1967), the Fuzzy c-Means algorithm (FCM) (BEZDEK et al., 1984) calculates, at each iteration, the membership degree (ω)

of each data sample i with respect to each cluster j, for j varying from 1 to K, when K should be defined by the end user (Equation (1)).

$$\omega_{(i,j)} = \frac{1}{\sum_{k=1}^{K}\left[\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}\right]} \quad (1)$$

In Equation (1), m is a fuzzification parameter, defined by the user with default value of 2, K is the number of desired clusters, c_j is the centroid of the cluster j and c_k is the centroid of the cluster k, also for k varying from 1 to K. At the end of each iteration, the centroids of each cluster j are recalculated, taking into account all N dataset samples and their membership values for the respective cluster (Equation (2)).

$$c_j = \frac{\sum_{i=1}^{N}\omega(i,j)^m * x_i}{\sum_{i=1}^{N}\omega(i,j)^m} \quad (2)$$

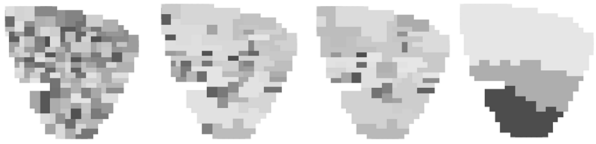
Instead of k-means, FCM convergence results not only to assign each sample to a unique cluster (non-overlapping clustering), but in a membership matrix with 0 to 1 values for each sample with respect to each cluster, known as fuzzy partition matrix (overlapping clustering). This matrix is one of the FCM advantages regarding non-overlapping clustering algorithms, providing better results for situations having difficult separation and overlapping datasets. However, like k-means, in the original version of FCM the centroids are randomly initialized. While this feature helps to reduce the computational cost in running FCM, it can also make the results susceptible to a local minima. Consequently, FCM may provide different results to the end user for different runs using the same parameters, which allows us to classify it as a non-deterministic algorithm.

The main reason for using FCM in the context of the delineation of management classes in PA is linked with the fact that abrupt changes do not occurs in soil and plant attributes in small enough parcels of the crop, causing input data and the obtained clusters to consider a membership degree. Over the years, several approaches in the literature using FCM and considering different types of these attributes have been developed. Brock et al. (2005) used FCM to delineate management classes considering historical yield data from corn-soybean rotation crops, indentifying the spatial association of the obtained maps with soil maps. Already Kitchen et al. (2005) used FCM to delineate management classes considering ratios of soil electrical conductivity (EC) in different depths (bulk of EC) and relief data, comparing them with yield classes obtained from historical yield data. As a result, it was found that the bulk of EC combined with relevant data are strong indications for management classes. Similar conclusions were obtained by Morari et al. (2009), including measures of soil and electrical resistivity data. The work of Li et al. (2007) used, in addition with abovementioned attributes, features indicating rates of organic matter and biomass. In this case, due to the large number of attributes, an intermediate phase of principal component analysis before getting the management classes by FCM was performed. High-resolution satellite images also appears as inputs to obtain management classes using the FCM, as in works of Song et al. (2009) and Zhang et al. (2009). Milne et al. (2012) used FCM to find management classes from smoothed spatial data obtained from three different methods. The results were compared with crop responses regarding the application of different nitrogen rates. The work of Scudiero et al. (2013) shows, using FCM to obtain management classes, that combined bare-soil and EC data can contribute to find spatial variability of a crop. The KM-sPC approach (CORDOBA et al., 2013) allowed to show the importance of a principal component analysis considering the coordinate space to reduce the stratification provided by FCM when management classes are displayed in form of maps. This approach were used again in a pratical nitrogen management of wheat (PERALTA et al., 2015). The study of Chang et al. (2014) compared management classes generated by FCM using reflectance data regarding the soil properties and productivity, showing that it is feasible the use of an active canopy sensor for this PA application.

Despite the widespread use of FCM for this task, the coordinate space of PA datasets, composed by spatial coordinates variables (e.g., latitude and longitude) have been used only to show the management classes provided by clustering in the form of maps. This fact does not block the use of these maps by automated machinery for variable rate interventions, but can reduce the spatial contiguity, causing stratification of management classes in too many management zones which can confuse visual analysis by experts. The approach proposed by Cordoba et al. (2013) attempts to reduce the effect of the contiguity loss treating the coordinate space during the preprocessing of the data. However although it is possible to achieve better results using this approach rather than traditional FCM, there is still very difficult for the end user to differentiate management classes in a visual way.

In order to solve this kind of problem, Ruß & Kruse (2011) proposed the HACC-Spatial hierarchical clustering algorithm. This approach takes into account spatial restrictions for clustering samples, and considers a preprocessing step to perform an initial tessellation of them in small spatial clusters, using the k-means algorithm at the coordinate space. Such subdivision aims to reduce computational costs by decreasing the number of steps of the construction of the hierarchical tree (or dendrogram) produced by the algorithm, considering the geostatistics principle that spatially very close samples tends to have close enough values in the attribute space

(MATHERON, 1963). As a result, a structure similar to a Voronoi diagram should be obtained by the preprocessing step (Fig. 1 (a)). From this moment, each dendrogram step merges the most similar clusters, according to the feature space. First, only spatially adjacent clusters can be merged, providing the maintenance of spatial contiguity (Fig. 1 (b)). However, when a user-defined contiguity threshold *cp* is reached, this restriction is switched off and from this point non-adjacent clusters can also be merged (Fig. 1 (c)). This threshold is associated to the ratio of the average distances between the samples belonging to adjacent clusters and the average distances between samples belonging to non-adjacent clusters. At the end of its run, it is expected that HACC-Spatial will provide maps of contiguous management classes as much as possible, regarding the parameters values provided by the end user (Fig.1 (d)).



(a) (b) (c) (d)

Fig. 1 - Clusterings obtained running HACC-Spatial, represented in the form of management classes maps, in sequencial dendrogram steps: (a) initial tesselation; (b) 60 clusters before reaching the cp threshold; (c) 30 clusters after reaching the cp threshold; (d) 3 clusters, representing management classes useful in practice. Source: Elaborated by the authors.

Because of the distinct nature of FCM and HACC-Spatial (partitional and hierarchical, respectively) and the spatial restrictions used by the second algorithm, it is expected distinct clustering results for the same dataset, making it difficult for the user to choose the best approach. A feasible solution to solve this question can be achieved using ensembles. Ensembles are able to combine multiple sample clusterings in a unique and consolidated one, known as consensus solution. These kind of approach can be used to meet several requirements, such as: increase the quality of the solution, providing more robust clusterings; select models; reuse knowledge; find consensus between clusterings obtained from subsets of features or subsamples, among others (GHOSH & ACHARYA, 2011).

The main aim of a clustering ensemble is to find a consensus solution composed by an unique clustering to share as much information as possible derived from original clusterings. This sharing can be measure by the average of normalized mutual information (ANMI), where the desired optimal value is ANMI equal to 1 (STREHL & GHOSH, 2002). The main goal of the three ensembles algorithms developed by Strehl & Ghosh (2012) is to build general approaches to obtain consensus from individual clusterings aiming at maximizing the ANMI value. These algorithms were evaluated by the authors in scenarios where individual clusterings were composed by distinct features, distinct subsamples or distinct clustering algorithms. The Cluster-based Similarity Partitioning Algorithm (CSPA) is the simplest and most obvious heuristic. It is based on the fact that two samples have a similarity of 1 if they are in the same cluster and 0 otherwise. Thus, a n x n binary matrix, where n is the number of samples, is created for each original clustering. To recluster these samples, a similarity-based clustering algorithm based on graph partitioning is used (KARYPIS & KUMAR, 1998). The HyperGraph Partitioning Algorithm (HGPA) addresses the clustering ensemble as a hypergraph partitioning problem, where hyperedges represent the original given clusters as indications of strong bonds. To recluster the samples, a partitional hypergraph algorithm, cutting a minimal number of hyperedges is used (HAN et al., 1997). In this case, while CPSA only considers pairwise relationships, HGPA includes original clustering relationships. Finally, the Meta-Clustering Algorithm (MCLA) represents each cluster by a hyperedge, and then group and collapse related hyperedges (or clusters), attaching each sample to the collapsed hyperedge in which it belongs more actively. At the end, a graph-based clustering of hyperedges is performed, indentifying consolidated *clusters of clusters*. According to Strehl & Ghosh (2002), the MCLA tends to provide better ANMI values when the consensus solution were obtained from individual clusterings with low noise rates and diversity; and HGPA and CSPA are usually better were obtained from individual clusterings with high noise rates and diversity.

Despite their effectiveness of obtain robust clusterings, the CPSA, HGPA and MCLA approaches are dependent of graph and hypergraph partitioning complex algorithms run by external software. In this way, and with the aim of extending the experiments and analyzes performed in our previous work (SPERANZA et al., 2016), here a simplified concept of clustering ensembles was used, based on evidence accumulation. This concept treats each original clustering as an independent evidence of data organization. Fred & Jain (2005) developed an approach based on this concept, where the original clusterings are combined by a voting mechanism, building a new similarity matrix known as co-association matrix. Equation (3) defines the co-association matrix calculation for n samples grouped by N different original clusterings.

$$M(i, j) = \frac{s_{ij}}{N} \tag{3}$$

In Equation (3), s_{ij} represents the number of times which the pair of samples (i, *j*), for *i* different from *j* and *i* and *j* less than or equal to *n*, are associated with the same cluster, considering the N original clusterings. Next, this matrix is used as input for an hierarchical clustering algorithm which regroups the samples in order to obtain a more robust result.

According to experiments performed by Fred & Jain (2005), the evidence accumulation approach presents, in general, better performance than the approaches based on partitioning of graphs and hypergraphs, mainly regarding the ability of this approach to recognize clusters with arbitrary forms in the

attribute space, making it possible to use it in data sets with well-correlated attributes.

From the concepts described in this section, we extend the experiments developed in our previous work (SPERANZA et al., 2016), now comparing the use of the clustering ensemble approaches based on graph and hypergraph partitioning with the evidence accumulation approach in generating more robust clusterings for the delineation of management classes in PA.

3. METHODOLOGY

The methodology used in ours experiments follows the concepts of Knowledge Discovery in Databases (KDD). According to Fayyad et al. (1996) and Weiss & Indurkhya (1998), at least three main steps of KDD process should be taken into account when it will be used: preprocessing, data mining (or pattern extraction) and post processing. The planned activities for each one of these steps, in the context of management classes in PA, are described below.

3.1 Preprocessing

The preprocessing step comprises the changes that should be made in a raw dataset when it will be used by a KDD process, preparing it to the next steps. Regarding to spatial data, in addition to very common preprocessing activities, such as standardization, cleaning and feature selection, the spatial interpolation must be performed in order to accommodate data samples in a single and regular spatial grid (VIEIRA, 2000). This activity is required, because PA datasets are caught using different kinds of sensors and samples densities, usually at distinct spatial spots in the same area. Another important activities in this step are: verifying data distribution using probabilistic density functions, as a preassessment of possible distortions that can occur in clustering algorithms when using non-Gaussians distributed features; verifying features correlations, using methods such as Pearson's Coefficient Correlation (BENESTY et al., 2009); and data standardization, reducing the bias caused by features with highly predominant scales relative to the others.

3.2 Data Mining

The data mining step can be viewed as an iterative process, where should be used different solutions to improve the accuracy of the results. In the context of our work, due to the fact that datasets had no previous classification, clusterings tasks need to be considered. Therefore, the approaches to be used are classified as non-supervised machine learning algorithms (MITCHELL, 1997). In this step, we used the HACC-Spatial and FCM algorithms in the traditional way and also combining results by ensembles. HACC-Spatial was run using non-spatial attributes of the whole dataset to calculate dissimilarity values at each step of dendrogram, and spatial attributes to build the initial tessellation and to support adjacency treatments at each step of dendrogram (Approach I). In the other hand, FCM was run in its traditional way, i.e., using only non-spatial attributes (Approach II). Regarding ensembles, it was created new approaches to found consensus clusterings from individual results provided by Approach I and Approach II, using both the graph and hypergraph partitioning algorithms (Approach III-A) and the evidence accumulation algorithm (Approach III-B). In the same way, it was created another four approaches to found consensus clustering from individual results provided by attributes subsets of soil, altimetry and yield using HACC-Spatial (Approaches IV-A and IV-B) and FCM (Approaches V-A and V-B). Regarding the ensembles approaches run using the graph and hypergraph partitioning algorithms (Approaches III-A, IV-A and V-A), where chosen the result which achieved the best values of ANMI considering different runs using CSPA, HGPA and MCLA.

According to domain expert users, at least 2 and at most 5 management classes should be considered for a crop (MOLIN et al., 2015). Thereby, the eight abovementioned approaches were run using k=2 to 5 clusters for the experiments, when using FCM (partitional), and the same values for dendrogram cuts, when using HACC-Spatial (hierarchical). Regarding to dissimilarity measures, the Euclidean distance were used for all approaches. In relation to other parameters and customizations, for approaches using FCM, the standard fuzzification value m=2 was fixed, and samples were associated with the cluster where were achieved a higher membership degree. HACC-Spatial parameters, like initial tessellation number of clusters (k) and cp, were defined during the experiments.

3.3 Post Processing

Finally, in the post processing step, we used two internal validation criteria: the SD criteria and the silhouette width criteria. These criteria allow comparing and evaluating the effectiveness of the eight approaches when they are run at the same number of clusters. The SD criteria (HALKIDI et al. 2000; HALKIDI & VAZIRGIANNIS, 2001) allows to verify, for each obtained clustering, how cohesive and well separated are the clusters, from average values of intra-cluster variance and distances between clusters centroids. In this case, optimal values should be closer to 0. The silhouette width criteria (ROUSSEEUW, 1987) follows the same principles of SD, but using dissimilarity values of a sample regarding its associated cluster and the nearest neighbor cluster. In this case, values closer to 1 indicates that the sample has been allocated to the correct cluster; and values closer to -1 indicates that the sample could have been better allocated to the nearest neighbor cluster. According to Vendramim et. al (2010), the silhouette width criteria, in comparison to other internal criteria in the literature, can provide, in general, more effective assessments about the internal structure of the clusters.

4. EXPERIMENTS

In this section, we present the results obtained

from experiments using real data, following the methodology described in section 3 and extending the results achieved in Speranza et al. (2016). These data are composed by samples collected on an experimental crop field of sugarcane culture. This field has an area around 17 hectares belonging to Fazenda Aparecida, located in Mogi-Mirim, São Paulo state, Brazil, with central coordinates 7505136N (latitude) and 299621E (longitude), given the spatial reference system UTM Zone 23S.

The raw datasets used in our work comprises measures of soil electrical conductivity (EC), in milisiemens per meter; altimetry quota, in meters; and historical yield, in tons per hectare or culms per square meter. The samples were collected at different times and by different sensors or processes, providing us six conventional attributes associated with spatial coordinates: soil electrical conductivity at 30 e 90 cm deep in 2010 (EC30 and EC90); altimetry quota (Quota); and historical yield in 2010 (Yield2010), 2012 (Yield2012) and 2013 (Yield2013). It is worth mentioning the need for historical yield data, because they could be considered susceptible to climatic factors over the years. In addition, the rainfall data of the whole farm in the agricultural years should be considered to support some analysis: 1601 mm in 2010 (July 2009 to June 2010), 1538 mm in 2012 (July 2011 to June 2012) and 1599 mm in 2013 (July 2012 to June 2013). The probabilistic density distribution of EC30, EC90 and Yield2010 attributes could be described by Gaussians, with most values around the mean. On the other hand, the distributions of Yield2012 and Yield2013 indicates, respectively, predominance of higher and lower yield values, probably affected by the climatic factors. A special case occurs with the Quota attribute, where average values are the minority because the experimental area has a slight slope and narrow in the central region. These distributions are shown in Fig. 2.

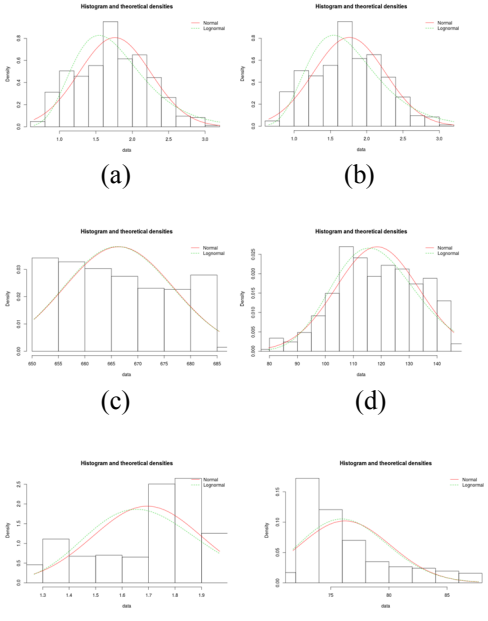


Fig. 2 - Probabilistic density distributions of dataset attributes: (a) EC30; (b) EC90; (c) Quote; (d) Yield2010; (e) Yield2012; e (f) Yield2013.

Applying the Pearson's Coefficient Correlation between pairs of attributes, were verified that EC30 and EC90 hold the most positive correlation of the dataset. In general, the Quote attribute was well correlated with all other attributes, and negatively (oppositely) correlated with Yield2010. Regarding to yield data, Yield2012 and Yield2013 attributes are highly correlated, and negatively correlated with Yeld2010 attribute. The negative correlation of Yield2010 with other yield years could be influenced again by the climatic factors.

Using the concepts of preprocessing described above, the dataset features were interpolated in a single regular spatial grid with spatial resolution of 20 meters. This value was calculated using the average coordinates spacing between samples for each one of the six features of the original data set. Simple algorithms, like the average of *k* nearest neighbors (ALTMAN, 1992), were used to interpolate attributes with higher sample densities. On the other hand, more sophisticated algorithms, like kriging (MATHERON, 1969), were used to interpolate attributes with smaller sample densities. After applying this process, each dataset feature were distributed in 415 samples spatially represented by points with latitude and longitude coordinates. Fig. 3 shows raw samples of soil electrical conductivity (high density) and yield (medium density) and their respective interpolated samples in the same regular spatial grid. Lower values are represented by lighter shades of gray, while higher values are represented by darker shades of gray.

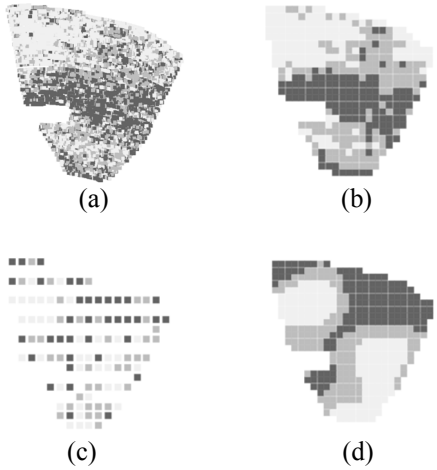


Fig. 3 - Example of raw and interpolated data in 3 classified intervals: (a) EC30 raw data (9046 samples); (b) EC30 interpolated data (415 samples); (c) Yield2010 raw data (111 samples); (d) Yield2010 interpolated data (415 samples).

Especially for the HACC-Spatial algorithm, when it was run in the context of approaches I, III-A and IV-A, the *cp* parameter was set to 0.5, according to the best results obtained by Ruß & Kruse (2011). Initial tessellation (*k*) was set to 200, after checking a significant increase in internal variance of the clusters for the following levels of the dendrogram.

First, the results achivied with the two approaches of clusterings ensembles used in this paper were qualitatively compared to each other and in relation

to the results achieved by the original clusterings using the internal validation criteria SD and silhouette width. Fig. 4 shows charts containing the indices of SD and silhouette width criteria achieved by the approaches I, II, III-A and III-B.

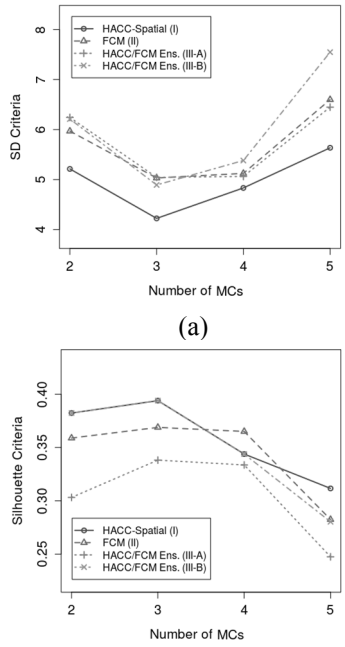


Fig. 4 - Indices of SD (a) and silhouette width (b) internal validation criteria achivied by clusterings generated using approaches I, II, III-A and III-B.

As already verified in previous experiments, charts from Fig. 4 show that ensemble approaches (III-A and III-B) attempt to find consensus solutions for the results obtained by the original clustering approaches. Therefore, it can be observed the trend maintenance in the value of the criteria for clusterings regarding 2 to 5 management classes, with slight variations in values favoring one or another approach. In addition, these results also show that HACC-Spatial (Approach I) achieved, in general, better performance when compared with FCM (Approach II) and ensembles approaches (III-A and III-B).

Completing this first analysis, Fig. 5 shows maps with four management classes generated by these approaches. In this figure, it can be observed very similar maps obtained by approaches I (Fig. 5 (a)) and II (Fig. 5(b)), suggesting the use of an ensemble approach. Regarding the esembles approaches, its easy to observe the influence of FCM on the final result of Approach III-A and the influence of HACC-Spatial on the final result of Approach III-B. Therefore, the Approach III-A expanded the stratification of the map obtained by Approach II, and the Approach III-B took advantage from the best spatial structure of the map from Approach I to generate a spatially better-distributed map. Due to the fact that the validation criteria used in our experiments do not have the ability to evalute the results regarding both attribute and coordinate space, spatially-stratified maps may obtain better indices than spatially well-distributed maps, as occurred in this case.

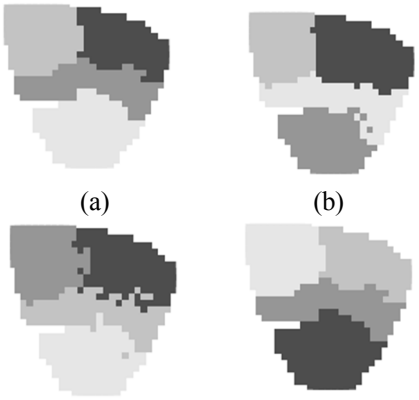


Fig. 5 - 4 Management classes maps generated using approaches I (a), II (b), III-A (c) and III-B (d).

Another experiment was performed using approaches IV-A and IV-B, where original clusterings were obtained using splits of the attribute space and the HACC-Spatial algorithm. Fig. 6 shows charts containg the SD and silhouette width indices achieved by these approaches, in comparison with the indices achived by Approach I.

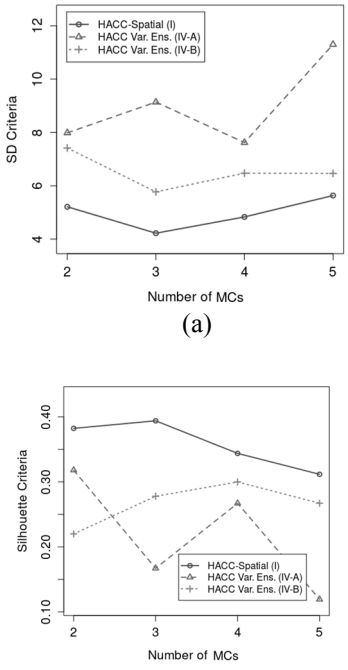
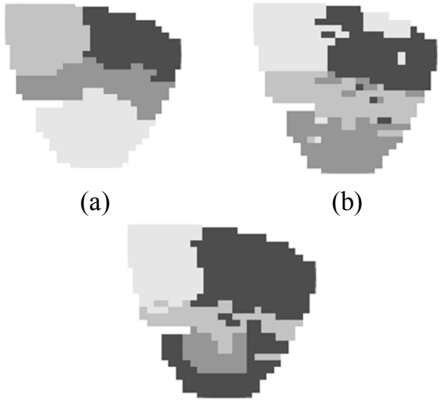


Fig. 6 - Indices of (a) and silhouette width (b) internal validation criteria achivied by clusterings generated using approaches I, IV-A and IV-B.

According to Fig. 6, the ensemble approach using evidence accumulation (IV-B) achieved better performance in this experiment than the approach using graph and subgraph partitioning (IV-A), confirming its better ability to take advantage of the better spatial distribution of the results provided by the HACC-Spatial algorithm. Nevertheless, the results achieved by the Approach IV-A can not be considered more robust than those achieved by Approach I, since in the same way as in our previous work, they may cause spatial contiguity

loss, harming the analysis of management classes maps by the end users (Fig. 7).

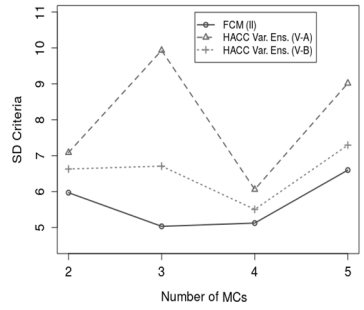


(a) (b)

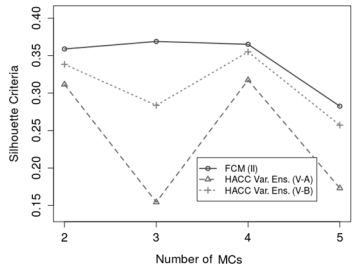
(c)

Fig. 7 - 4 Management classes maps generated using approaches I (a), IV-A (b) and IV-B (c).

Finally, an experiment similar to the previous one was performed, now regarding the approaches V-A and V-B, where the original clusterings were obtained using splits of the attribute space and the FCM algorithm. Fig. 8 shows charts containg the SD and silhouette width indices achieved by these approaches, in comparison with the indices achived by Approach II.



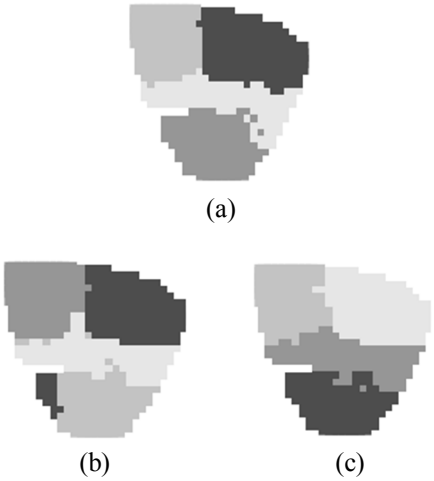
(a)



(b)

Fig. 8 - Indices of SD (a) and silhouette width (b) internal validation criteria achivied by clusterings generated using approaches II, V-A and V-B

In the same way, the ensemble approach using evidence accumulation (V-B) achieved better performance on this experiment than the approach using graph and subgraph partitioning (V-A). From the resulting management class maps shown in Fig. 9, it can be noted the better ability of the Approach V-B to dealing with the stratification issue, in comparison with the Approach V-A. In addition, for this case, the Approach V-B was able to obtain management classes maps quite compatible with the Approach II, without generating aditional stratification.



(a)

(b) (c)

Fig. 9 - 4 Management classes maps generated using approaches II (a), V-A (b) and V-B (c).

5. CONCLUSIONS AND FUTURE WORK

In this paper, new experiments and analysis were performed using clustering ensembles approaches applied in the delineation of management classes in PA. Here, we used an approach based on evidence accumulation and compare the results of the new experiments with the results performed by an approach based on partitioning of graphs and hypergraphs algorithms, already evaluated in a preliminary work.

If we consider maps delineated by traditional clustering approaches using all attributes (approaches I and II), according to both the internal validation criteria and visual analysis, the HACC-Spatial algorithm achieved in general better results compared to FCM algorithm. This finding is different from the preliminary work, and it can be explained by the non-determinism also verified in running HACC-Spatial, which just like FCM, uses random intiation of centroids in the initial tessellation step that may provide different results to the end user using the same values for parameters and attributes.

Regarding the use of ensembles, the results obtained by approaches III-A and III-B show an attempt to obtain consensus clusterings extracting the main features from both algorithms used in obtaining individual clusterings (HACC-Spatial and FCM). In this case, the ensemble approach based on graphs and hypergraphs partitioning (III-A) provided solutions closer to those obtained by the FCM algorithm, causing an increase in stratification. In the other hand, the ensemble approach based on accumulation of evidence (III-B) favored the best spatial arrangement generated by the HACC-Spatial algorithm, providing results with no stratification and easier interpretation by the end user. However, the indices for internal validation criteria achieved by the Approach III-B have not always been better in comparison to the indicies achived by other approaches, evidencing the fact that the internal

validation criteria used in this experinents only address issues related to attribute space.

Finally, although the ensembles approaches using subdivision of the attribute space have obtained worse results in relation to those achieved by the traditional algorithms using all attributes, the performed experiments were useful to confirm the skill of the Approach V-B in dealing better with stratification issues than the Approach V-A.

The clustering ensembles approaches used in this work is rather general and try to find consensus clusterings using only final clusterings obtained from splitting of features or from different algorithms. In a future work, could be proposed new clustering ensembles approaches which allow extracting the main features of each algorithm, as the membership values provided by FCM, or the contiguity threshold provided by HACC-Spatial. In addition, new internal validation criteria can be proposed to evaluate clusterings regarding both the attribute and coordinate space, in order to improve the analysis performed by the end user.

ACKNOWLEDGEMENT

Thanks to Fazenda Aparecida for the availability of the area for data collection that allowed the achievement of this study, and to researchers from Embrapa, Celia Grego and Luiz Vicente, responsible for collecting such data. We also thank the following Brazilian research agencies: CNPq, CAPES, FAPESP and FINEP. The second author has been supported by the grant 311868/2015-0 from CNPq.

REFERENCES

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **The American Statistician**, v. 46, n. 3, p. 175–185, 1992.

BENESTY, J.; CHEN, J.; HUANG, Y.; COHEN, I. Pearson Correlation Coefficient. In: **Noise Reduction in Speech Processing SE - 5**. Springer Berlin Heidelberg, 2009. v.2 of Springer Topics in Signal Processing, p. 1–4.

BERNARDI, A. C. d. C. et al. **Agricultura de Precisão - Resultados de um Novo Olhar**. 1. ed.Brasília: Empresa Brasileira de Pesquisa Agropecuaria, 2014. 596 p.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM: The Fuzzy c-Means Clustering Algorithm. **Computer & Geosciences**, v. 10, n. 2-3, p. 191–203, 1984.

BROCK, A.; BROUDER, S. M.; BLUMHOFF, G.; HOFMANN, B. S. Defining Yield-Based Management Zones for Corn-Soybean Rotations. **Agronomy Journal**, v. 97, n. 4, p. 1115–1128, July 2005.

CHANG, D.; ZHANG, J.; ZHU, L.; GE, S. H.; LI, P. Y.; LIU, G. S. Delineation of management zones using an active canopy sensor for a tobacco field. **Computers and Electronics in Agriculture**, v. 109, p. 172–178, 2014.

CÓRDOBA, M.; BRUNO, C.; COSTA, J.; BALZARINI, M. Subfield management class delineation using cluster analysis from spatial principal components of soil variables. **Computers and Electronics in Agriculture**, v. 97, p. 6–14, 2013.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996.

FRED, A. N. L.; JAIN, A. K. Combining multiple clusterungs using evidence accumulation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, n. 6, p. 835–850, 2005.

GHOSH, J.; ACHARYA, A. Cluster ensembles. Wiley Interdisciplinary Reviews: **Data Mining and Knowledge Discovery**, v. 1, n. 4, p. 305–315, 2011.

HALKIDI, M.; VAZIRGIANNIS, M.; BATISTAKIS, Y. Quality scheme assessment in the clustering process. In: ZIGHED, D.; KOMOROWSKI, J.; ZYTKOW, J. (Eds.) . **Principles of Data Mining and Knowledge Discovery**. Springer Berlin Heidelberg, 2000. v. 1910 of Lecture Notes in Computer Science, p. 265–276.

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment: finding the optimal partitioning of a data set. In: **Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on**. IEEE, 2001. p. 187-194.

HAN, E.-H.; KARYPIS, G.; KUMAR, V.; MOBASHER, B. Clustering based on association rule hypergraphs. In: **DKMD**, 1997.

KARYPIS, G.; KUMAR, V. Multilevelk-way partitioning scheme for irregular graphs. **Journal of Parallel and Distributed Computing**, v. 48, n. 1, p. 96–129, 1998.

KITCHEN, N.; SUDDUTH, K.; MYERS, D.; DRUMMOND, S.; HONG, S. Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity. **Computers and Electronics in Agriculture**, v. 46, n. 1-3, p. 285–308, Mar. 2005.

LI, Y.; SHI, Z.; LI, F.; LI, H.-Y. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. **Computers and Electronics in Agriculture**, v. 56, n. 2, p. 174–186, Apr. 2007.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate ob- servations. In: **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**, v. 1, p. 281–297.

MATHERON, G. Principles of geostatistics. **Economic**

geology, v. 58, n. 8, p. 1246–1266,1963.

MATHERON, G. **Le krigeage universel**. Paris, France, 1969.

MILNE, A. E.; WEBSTER, R.; GINSBURG, D.; KINDRED, D. Spatial multivariate classification of an arable field into compact management zones based on past crop yields. **Computers and Electronics in Agriculture**, v. 80, p. 17–30, 2012.

MITCHELL, T. M. **Machine Learning**. New York; London: McGraw-Hill, 1997

MOLIN, J. P. Agricultura de Precisão: Situação atual e perspectivas. In: FANCELLI, A. L.; NETO, D. D. (Eds.) **Milho: Estratégias de Manejo para Alta Produtividade**. Piracicaba: ESALQ/USP/LPV, 2003. p. 89–98.

MOLIN, J. P.; DO AMARAL, L. R.; COLAÇO, A. **Agricultura de precisão**. Oficina de Textos, 2015.

MORARI, F.; CASTRIGNANÒ, A.; PAGLIARIN, C. Application of multivariate geostatistics in delineating management zones within a gravelly vineyard using geo-electrical sensors. **Computers and Electronics in Agriculture**, v. 68, n. 1, p. 97–107, Aug. 2009.

PERALTA, N. R.; COSTA, J. L.; BALZARINI, M.; Castro Franco, M.; CORDOBA, M.; BULLOCK, D. Delineation of management zones to improve nitrogen management of wheat. **Computers and Electronics in Agriculture**, v. 110, p. 103–113, 2015.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, n. 0, p. 53–65, 1987.

RUSS G.; KRUSE, R. Exploratory hierarchical clustering for management zone delineation in precision agriculture. In: PERNER, P. (Ed.) **Advances in Data Mining. Applications and Theoretical Aspects**. Springer Berlin Heidelberg, 2011. v. 6870 of Lecture Notes in Computer Science, p. 161–173.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis**. Journal of Computational and Applied Mathematics**, v. 20, n. 0, p. 53–65, 1987.

SCHWALBERT, R. A.; AMADO, T. J. C.; GEBERT, F. H.; SANTI, A. L.; TABALDI, F. Zonas de manejo: atributos de solo e planta visando a sua delimitação e aplicações na agricultura de precisão. **Revista Plantio Direto**, p. 21–32, 2014.

SCUDIERO, E.; TEATINI, P.; CORWIN, D. L.; DEIANA, R.; BERTI, A.; MORARI, F. Delineation of site-specific management units in a saline region at the Venice Lagoon margin, Italy, using soil reflectance and apparent electrical conductivity. **Computers and Electronics in Agriculture**, v. 99, p. 54–64, 2013.

SONG, X.; WANG, J.; HUANG, W.; LIU, L.; YAN, G.; PU, R. The delineation of agricultural management zones with high resolution remotely sensed data. **Precision Agriculture**, v. 10, n. 6, p. 471–487, 2009.

SPERANZA, E. A.; CIFERRI, R. R.; CIFERRI, C. D. A. Clustering Approaches and Ensembles Applied in the Delineation of Management Classes in Precision Agriculture. In: **XVII Brazilian Symposium on GeoInformatics**. Campos do Jordão: MCTIC/INPE, 2016. p. 152–165.

STREHL, A.; GHOSH, J. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. **Journal of Machine Learning Research**, v. 3, p. 583–617, 2002.

TAYLOR, J. A.; MCBRATNEY, A. B.;WHELAN, B. M. Establishing Management
Classes for Broadacre Agricultural Production. **Agronomy Journal**, v. 99, n. 5, p. 1366–1376, Sept. 2007.

VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. **Statistical Analysis and Data Mining**, v.3, n. 4, p. 209–235, 2010.

VENDRUSCULO, L. G.; KALEITA, A. L. Modeling zone management in precision agriculture through Fuzzy C-Means technique at spatial database. In: . **St. Joseph, MI: American Society of Agricultural and Biological Engineers**, 2011.

VIEIRA, S. R. Geoestatistica em Estudos de Variabilidade Espacial do Solo. In: NOVAIS, R. F.; ALVAREZ, V H, S. G. R. (Eds.) **Tópicos em ciência do solo**. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, 1. ed., 2000. p. 1–54.

WEISS, S. M.; INDURKHYA, N. **Predictive Data Mining: A Practical Guide**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998.

ZHANG, X.; SHI, L.; JIA, X.; SEIELSTAD, G.; HELGASON, C. Zone mapping application for precision-farming: a decision support tool for variable rate application. **Precision Agriculture**, v. 11, n. 2, p. 103–114, 2010.

Apendice G

INTEGRAÇÃO DE FERRAMENTAS DE SIG E MINERAÇÃO DE DADOS PARA UTILIZAÇÃO EM ATIVIDADES DE GESTÃO ESPACIALMENTE DIFERENCIADA APLICADA NA AGRICULTURA DE PRECISÃO

Este apêndice contém a transcrição do artigo intitulado *Integração de Ferramentas de SIG e Mineração de Dados para Utilização em Atividades de Gestão Espacialmente Diferenciada Aplicada na Agricultura de Precisão*, aceito para apresentação oral e publicação nos anais do evento *XI Congresso Brasileiro de Agroinformática*, a ser realizado no período de 2 à 6 de outubro de 2017.

Integração de Ferramentas de SIG e Mineração de Dados para Utilização em Atividades de Gestão Espacialmente Diferenciada Aplicada na Agricultura de Precisão

Eduardo Antonio Speranza[1], *Ricardo Rodrigues Ciferri*[2]

[1]Embrapa Informática Agropecuária
Campinas, São Paulo, Brasil
eduardo.speranza@embrapa.br

[2]Departamento de Computação, Universidade Federal de São Carlos
São Carlos, São Paulo, Brasil
ricardo@dc.ufscar.br

RESUMO

A agricultura de precisão é uma abordagem agrícola que se utiliza de tecnologias da informação e comunicação para possibilitar uma gestão diferenciada da lavoura voltada para o aumento da produtividade de maneira sustentável, reduzindo os impactos ao meio ambiente a partir da aplicação espacialmente diferenciada de insumos agrícolas e consequentemente proporcionando o aumento do retorno econômico. Um dos conceitos utilizados por essa abordagem é o delineamento de unidades de gestão diferenciada, permitindo tratamentos localizados de acordo com as características do solo e das plantas. Devido ao crescimento da disponibilidade de ferramentas computacionais que podem auxiliar usuários finais, a criação de modelos e arquiteturas acessíveis e capazes de agregar essas ferramentas, de forma a possibilitar o seu uso de maneira integrada, tem se tornado intenso objeto de estudo. Este artigo descreve a proposta de uma arquitetura voltada para apoiar o processo de delineamento de unidades de gestão diferenciada em agricultura de precisão utilizando *software* livre e de distribuição gratuita. A arquitetura proposta é comparada com outras abordagens disponíveis na literatura, onde são identificadas as suas vantagens e desvantagens e propostas novas alternativas para trabalhos futuros. O trabalho realizado nesse artigo permitiu verificar que a solução proposta é viável para essa aplicação de AP, devido à disponibilidade das ferramentas e a utilização dos padrões OGC, amplamente disseminados nas áreas de aplicação que utilizam dados geoespaciais.

PALAVRAS-CHAVE: Agricultura de Precisão, Unidades de Gestão Diferenciada, Zonas de Manejo, Sistemas de Informações Geográficas, Mineração de Dados, Dados Espaciais.

ABSTRACT

Precision agriculture is an agricultural approach which uses information and communication technologies to enable differentiated crop management to increase yield in a sustainable way, reducing impacts to the environment from spatially differentiated application of agricultural inputs and consequently providing the increase of profits. One of the concepts used by this approach is the delineation of differentiated management units, allowing site-specific treatments according to soil and plant features. Due to the increasing availability of computational tools which can help end users to apply this concept, the proposal of models and achitetures accessible and able to aggregate these tools, so they can be used in an integrated way. This paper describes the proposal of an architeture aimed to support the delineation of differentiated management units in precision agriculture, using free available software. The proposed architeture is compared with other approaches available in the literature, where its advantages and drawbacks are identified and new alternatives are proposed for future work. The work performed in this article allowed to verify that the proposed solution is feasible for this AP application, due to the availability of tools and the use of OGC standards, widely disseminated in applied areas using geospatial data.

KEYWORDS: Precision Agriculture, Differentited Management Units, Management Zones, Geographic Information Systems, Data Mining, Spatial Data.

INTRODUÇÃO

A Agricultura de Precisão (AP) é definida como uma abordagem de gerenciamento agrícola baseada na variabilidade espacial e temporal da lavoura, com o intuito de aumentar a produtividade de maneira sustentável, reduzindo os impactos ao meio ambiente com a aplicação espacialmente diferenciada de insumos agrícolas e, consequentemente, aumentando o retorno econômico (BERNARDI et al., 2014; MOLIN; AMARAL; COLAÇO, 2015). Para atingir esse objetivos, essa abordagem comumente utiliza tecnologias de informação e comunicação (TICs) associadas a processos que visam a otimização das culturas (MOLIN, 2003; MASSRUHÁ et al., 2014). Assim, os processos de preparação do solo, plantio, acompanhamento da lavoura e colheita passam a ser auxiliados por atividades de coleta e interpretação de dados, visando a realização de intervenções sítio específicas (COMPARETTI, 2011). Um dos principais fatores contribuintes para impulsionar a adoção da AP no mundo foi a operacionalização do sinal de GPS[1], em meados da década de 90. Desse modo, os Sistemas de Informações Geográficas (SIGs) (BURROUGH, 1986; CIFERRI, 1995), em conjunto com bancos de dados espaciais (CASANOVA et al., 2005; CIFERRI, 1995), se tornaram ferramentas computacionais essenciais para dar suporte às atividades de armazenamento, manipulação, análise e recuperação de dados georreferenciados que podem ser executadas em um ciclo produtivo utilizando AP.

Nos últimos anos, com o crescimento das TICs e, consequentemente, das formas de se

[1]GPS: do inglês *Global Positioning System* - Sistema de Posicionamento Global.

obter dados georreferenciados em campo, seja por meio de sensores específicos, imagens aéreas e de satélite com altíssima resolução espacial, ou a partir de ferramentas de inspeção instaladas em *tablets* ou *smartphones*, novos conceitos relacionados ao armazenamento, manipulação e recuperação de dados, como *big data* e computação em nuvem, passaram a fazer parte do contexto das pesquisas em AP (LI; CHUNG, 2015). Desse modo, as atividades de análise, que estão fortemente relacionadas ao suporte à decisão para o usuário final, devem ser realizadas com o auxílio de processos de descoberta de conhecimento em bancos de dados (KDD[2]) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; WEISS; INDURKHYA, 1998). Uma das atividades de AP que pode se beneficiar desse processo é o delineamento de unidades de gestão diferenciada (UGDs), também conhecidas como zonas de manejo. As UGDs são regiões geográficas delimitadas na lavoura com variabilidade interna desprezível, a ponto de poderem ser consideradas intervenções homogêneas dentro dos seus limites (RUSS, 2012; MOLIN; AMARAL; COLAÇO, 2015). As UGDs obtidas devem ser consistentes ao longo do tempo e caracterizarem o potencial de resposta de uma área de produção agrícola, podendo ser obtidas a partir de tarefas de mineração de dados. Desse modo, as UGDs podem ser utilizadas para diversos fins gerenciais, que vão desde a determinação da densidade apropriada para coleta de amostras em solo, até o direcionamento de investimentos em insumos agrícolas.

Este artigo descreve a proposta de uma arquitetura computacional para suporte à tarefa de delineamento de UGDs em AP, utilizando a integração de SIGs e bancos de dados espaciais com algoritmos capazes de realizar as etapas de KDD necessárias para a execução dessa tarefa, tais como interpolação espacial, agrupamento de dados espaciais e validação. O restante do artigo é organizado da seguinte forma: a seção Material e Métodos descreve as TICs e conceitos utilizados para a construção da arquitetura; a seção Resultados e Discussão descreve a arquitetura em si, bem como resultados que podem ser obtidos com a sua utilização, discussões e comparações com outras soluções disponíveis na literatura; e a seção Conclusões resume os resultados obtidos e indica trabalhos futuros que podem ser realizados com o decorrer desta pesquisa.

MATERIAL E MÉTODOS

Nesta seção são apresentados os conceitos sobre TICs que dão suporte ao desenvolvimento da arquitetura proposta neste artigo, partindo desde o armazenamento dos dados até a interface com o usuário final. Também são descritos os dados reais utilizados em experimentos realizados com uma implementação da arquitetura proposta.

Sistemas Gerenciadores de Bancos de Dados Espaciais (SGBDEs)

Os Sistemas Gerenciadores de Bancos de Dados (SGBDs) são programas de computador que auxiliam nos processos de definição, construção, carga, manipulação e compartilhamento de

[2]KDD: do inglês *Knowledge Discovery in Databases*.

bancos de dados entre diversos usuários e aplicações (ELMASRI; NAVATHE, 2011). Dentre os SGBDs mais conhecidos e utilizados atualmente, destacam-se o PostgreSQL (POSTGRESQL, 2017) e o Oracle (ORACLE, 2017), que, dentre outras características, possuem mecanismos de extensibilidade que possibilitam o desenvolvimento de extensões geográficas capazes de tratar dados espaciais, que também são conhecidos como dados georreferenciados, dados geoespaciais ou dados geográficos (CASANOVA et al., 2005; CIFERRI, 1995), tornando-os SGBDs Espaciais (SGBDEs). Para as atividades de AP, pode ser utilizado qualquer SGBDE que possua a característica de armazenar dados espaciais, utilizando objetos bidimensionais para representar a disposição espacial; e dados convencionais associados a esse objetos representando fenômenos naturais que ocorrem na lavoura.

Padrões Open Geospatial Consortium

A OGC (*Open Geospatial Consortium*) é uma organização internacional sem fins lucrativos criada para o desenvolvimento de padrões de consenso para dados geoespaciais. Os padrões OGC são baseados na gramática de modelagem GML (*Geography Markup Language*), que atua no transporte, armazenamento e intercâmbio de dados geoespaciais na *Web* (PERCIVALL et al., 2003). Dentre esses padrões, destacam-se três importantes serviços: WMS (*Web Map Service*), utilizado para a recuperação e visualização de mapas como imagens; WFS (*Web Feature Service*), utilizado para acesso e edição de dados geoespaciais; e o WPS (*Web Processing Service*), utilizado para a execução de algoritmos de geoprocessamento. Diversos módulos da arquitetura proposta neste artigo são integrados utilizando os padrões OGC, que vêm sendo amplamente disseminados na comunidade produtora de dados geoespaciais.

Servidores de Mapas

A disponibilização dos SIGs na *Web* impulsionou o desenvolvimento de soluções para disseminar informações espaciais em forma de mapa, permitindo a criação de aplicações interativas com diversas funcionalidades disponíveis para o usuário final (MITCHELL, 2005). Desse modo, surgiram os servidores de mapas, como o GeoServer (GEOSERVER, 2017), capazes de acessar, manipular e entregar dados geoespaciais seguindo os padrões OGC. *Software* livre e de distribuição gratuita, o GeoServer possui suporte a diversos serviços OGC, e compatibilidade para manipular dados especiais armazenados por diversos SGBDEs, podendo ser facilmente integrado aos SIGs.

Sistemas de Informações Geográficas

Os sistemas de informações geográficas (SIGs) são utilizados para manipulação de dados que representam objetos e fenômenos onde a localização geográfica é uma característica intrínseca indispensável à análise da informação (ARONOFF, 1989). Os SIGs são compostos por vários subsistemas integrados, e devem prover suporte a diferentes tipos de dados e aplicações, com

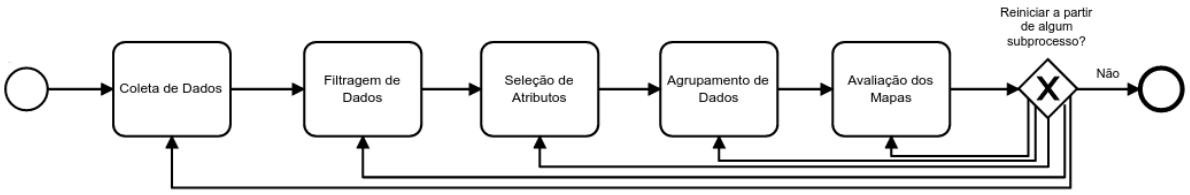
o intuito de centralizar os dados coletados de fontes heterogêneas de maneira transparente ao usuário final, por meio de funcionalidades de entrada, integração, processamento, visualização, plotagem, armazenamento e recuperação de dados (CIFERRI, 1995; CÂMARA et al., 1996). Atualmente, os SIGs trabalham de forma integrada com os SGBDEs para o armazenamento e recuperação de dados espaciais.

Dentre as ferramentas de SIG disponíveis no mercado, deve-se destacar a plataforma proprietária ArcGIS (ARCGIS, 2017), criada em 2001 pela empresa ESRI e amplamente disseminada, não só na comunidade de geoprocessamento, mas também na maioria das áreas de aplicação. Quando consideramos apenas o mercado de *software* livre e de distribuição gratuita, o QuantumGIS (ou QGIS) (QGIS, 2017) é o SIG mais utilizado. Além dos diversos complementos disponíveis para atividades de geoprocessamento, a possibilidade de implementação de algoritmos pelos próprios usuários e integração com outros SIGs e plataformas que possuem diversas bibliotecas relacionadas à geoestatística, geoprocessamento e mineração de dados, como o ambiente R (R, 2017), fazem com que o QGIS seja bastante disseminado entre os usuários e produtores de dados geoespaciais, principalmente quando o custo relacionado à aquisição de *software* é um fator relevante para o usuário.

Mineração de Dados

A transformação dos dados em conhecimento, com o intuito de fornecer suporte ao usuário final em atividades que exigem a tomada de decisão, é um dos principais motivadores para a utilização de conceitos de mineração de dados em AP, onde os valores dos atributos, que estão normalmente associados a coordenadas geográficas, podem ser influenciados por seus vizinhos espaciais (ESTER et al., 2000). A etapa de mineração de dados é considerada, em conjunto com o pré-processamento e o pós-processamento, como essencial em um processo de KDD (REZENDE, 2003). No contexto da AP, Santos, Molin e Saraiva (2013), Santos (2014) propuseram um modelo de referência para o processo de delineamento de UGDs (Figura 1). Esse modelo contempla todas as etapas do processo de KDD no âmbito da aplicação e foi utilizado como guia para a construção da arquitetura proposta neste artigo.

Figura 1: Modelo de referência para o processo de delineamento de UGDs.



Fonte: Extraído de Santos (2014)

Para a execução dos algoritmos que envolvem processos de KDD, o ambiente de desenvolvimento R (R, 2017) tem sido amplamente utilizado. O ambiente R permite o acesso aos seus

algoritmos que envolvem geoprocessamento por meio da interface do QGIS, com a utilização do complemento *Processing*. Esse complemento permite a criação de interfaces para usuários finais, e pode retornar como resultado tanto dados geoespaciais vetoriais como gráficos gerados pelos próprios algoritmos do ambiente R. Além disso, permite a criação de modelos para execução sequencial das interfaces de acesso criadas que definem *workflows* para um conjunto de aplicações.

Publicação de Dados Espaciais

Uma vez que o processo de delineamento de UGDs é finalizado, os mapas escolhidos pelo usuário final para serem utilizados em determinadas atividades devem ser armazenados para futuras consultas e até mesmo para disseminação via rede. Além do armazenamento direto dos dados espaciais em sua forma bruta no SGBDE, os mapas de UGDs também podem ser publicados no GeoServer diretamente pela interface do QGIS. Uma vez publicados, esses mapas podem ser acessados via *Web* por meio de sítios com funcionalidades de SIG compatíveis com os padrões OGC, conhecidos como *WebGIS* (MITCHELL, 2005). Um exemplo de sítios desse tipo é o GeoNode (GEONODE, 2017), uma plataforma de código aberto para compartilhamento de dados geoespaciais que permite ao usuário final explorar mapas disponibilizados por servidores como o GeoServer. Os mapas podem ser publicados diretamente no GeoNode por meio da interface gráfica do QGIS, utilizando o complemento OpenGeo.

Dados Geoespaciais

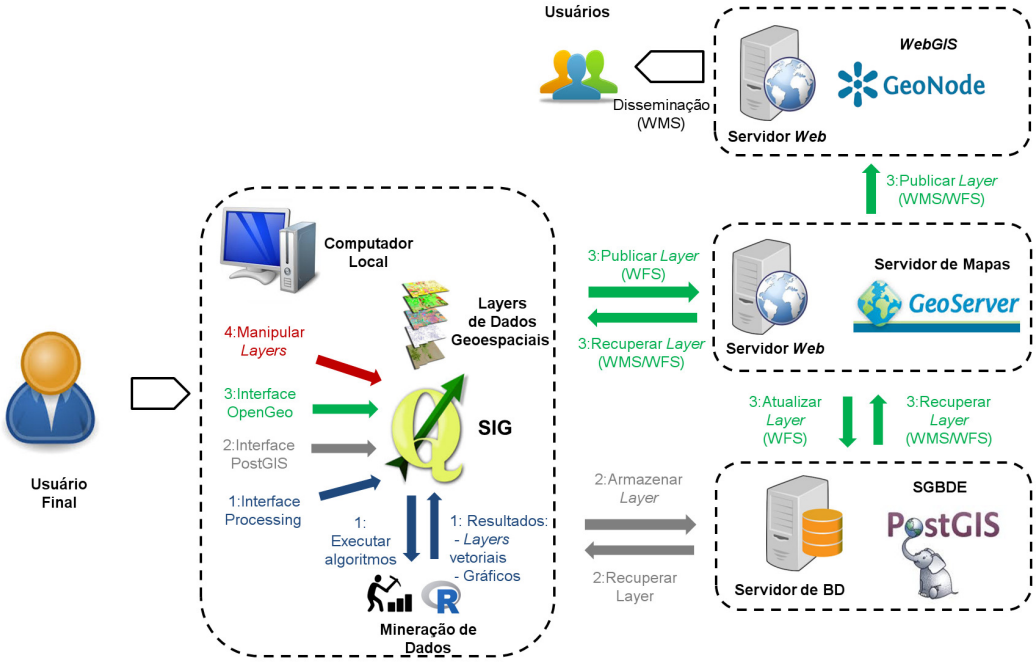
O conjunto de dados geoespaciais reais usado nos experimentos deste artigo foi obtido em um talhão de cultura de cana-de-açúcar de aproximadamente 17 hectares da Fazenda Aparecida, localizada em Mogi-Mirim, SP, com 20 atributos relacionados ao solo e a cultura, tais como pedologia, textura e condutividade elétrica do solo, índices de biomassa e produtividade. Para o delineamento de UGDs, a arquitetura proposta neste artigo permite ao usuário final: escolher os atributos que serão utilizados; submeter os dados a processos de limpeza para exclusão de valores extremos; realizar processos de interpolação espacial para acomodar os atributos em uma grade espacial única; delinear as UGDs utilizando algoritmos de agrupamento espacial; analisar os resultados obtidos por meio de critérios de validação; e publicar mapas finais de UGDs.

Na próxima seção, a arquitetura proposta é apresentada e comparada com outras soluções disponíveis na literatura, com o intuito de analisar as suas vantagens e desvantagens e também indicar melhorias que podem ser propostas para uma adequação maior ao processo à qual está sendo submetida.

RESULTADOS E DISCUSSÃO

Considerando as TICs descritas na seção Material e Métodos, foi proposto um modelo de arquitetura para permitir a realização da tarefa de delineamento de UGDs em AP, com a utilização de *software* livre e de distribuição gratuita (Figura 2).

Figura 2: Modelo de arquitetura para a tarefa de delineamento de UGDs em AP utilizando *software* livre e padrões OGC.



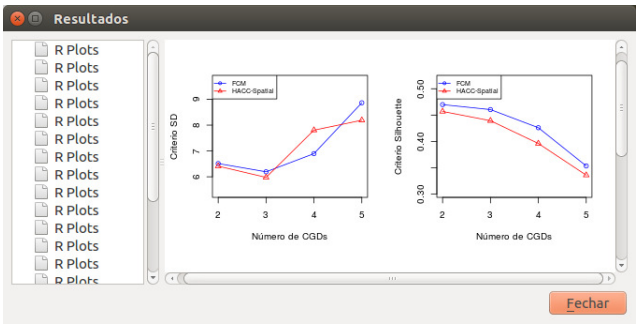
Nessa arquitetura, são observados quatro macroprocessos distintos que podem ser executados pelo usuário final. O primeiro, identificado pelo número 1, permite a integração da interface principal do QGIS com os algoritmos de mineração de dados disponibilizados pelo ambiente R, por meio da criação de *scripts* anotados utilizando o complemento *Processing*. Esses *scripts* geram janelas gráficas para que o usuário final informe os parâmetros e dados espaciais de entrada que serão utilizados pelos algoritmos de mineração de dados. Para os dados de saída, são permitidos tanto camadas vetoriais, que podem ser posteriormente manipuladas utilizando as funcionalidades específicas do QGIS, quanto figuras representando gráficos ou mapas gerados pelos próprios algoritmos de mineração de dados do ambiente R, visualizadas por meio do complemento *Processing* (Figura 3).

Além das diversas alternativas para entrada e saída de dados que a arquitetura proporciona, também é possível executar diversos algoritmos de mineração de dados de maneira sequencial, por meio da ferramenta de criação de modelos do complemento *Processing*. A Figura 4 mostra um exemplo de criação de um modelo para delineamento de UGDs, onde são executados sequencialmente algoritmos para limpeza de dados, criação de grade espacial única, interpolação espacial e agrupamento, até que o mapa vetorial de UGDs seja obtido.

O segundo macroprocesso definido na arquitetura da Figura 2, identificado pelo número 2,

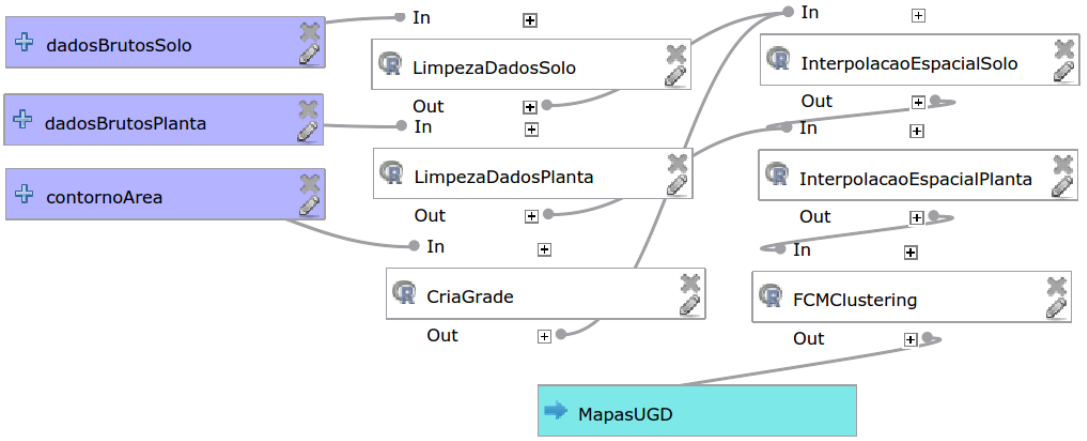
Figura 3: Exemplos de resultados retornados por algoritmos de mineração de dados executados pelo ambiente R e integrados a interface do QGIS: (a) mapa de UGDs no formato vetorial com imagem de sensoriamento remoto ao fundo; (b) gráfico com análises considerando critérios de validação.

(a)



(b)

Figura 4: Exemplo de modelo para delineamento de UGDs com execução sequencial de diversos algoritmos implementados no ambiente R.



permite ao usuário final, a partir de funcionalidades específicas do QGIS, armazenar e recuperar os mapas de UGDs obtidos a partir do macroprocesso 1 diretamente no SGBDE PostGIS. Com os dados armazenados, o usuário final pode realizar o terceiro macroprocesso, identificado na arquitetura da Figura 2 pelo número 3, para disseminar pela rede os mapas de UGDs obtidos. As funcionalidades descritas para esse macroprocesso devem ser acessadas a partir da interface do complemento OpenGeo, disponível no QGIS. Por meio dessa interface, é possível publicar e recuperar mapas de UGDs já armazenados no SGBDE a partir do GeoServer, utilizando para isso os serviços OGC WFS e WMS. Utilizando a mesma interface do complemento OpenGeo, os mapas de UGDs disponibilizados no servidor de mapas GeoServer podem ser publicados no GeoNode, permitindo a sua disseminação pela rede por meio de uma interface *WebGIS* disponível para qualquer usuário que desejar utilizá-los.

Finalmente, um quarto macroprocesso, relacionado à manipulação de *layers*[3] localmente

[3] *Layer:* Termo comumente utilizado pela comunidade de SIG para se referir a conjuntos de dados geoespaciais temáticos que podem ser visualizados ou manipulados.

pelo usuário final utilizando o SIG QGIS também foi definido na Figura 2, identificado pelo número 4. A manipulação local de *layers*, sejam elas representando dados espaciais brutos disponibilizados pelo GeoServer, ou mapas de UGDs delineadas a partir do macroprocesso 1, permite ao usuário final editá-las de maneira *off-line* utilizando as funcionalidades de geoprocessamento disponibilizadas pelo QGIS. Isso permite que essas *layers* sejam publicadas em um servidor de mapas apenas quando estiverem em sua versão final.

As ferramentas e complementos disponíveis no QGIS para a sua integração com o ambiente R e posterior publicação de mapas em *WebGIS* podem ajudar muito no processo de delineamento de UGDs em AP. Entretanto, algumas dificuldades foram encontradas durante a implementação e devem ser registradas para que melhorias possam ser propostas. Com relação à anotação de *scripts* para geração de interfaces, a limitação de tipos de campos disponíveis restringe a utilização de soluções alternativas em alguns casos, por exemplo quando existe a necessidade de informar mais de um atributo da camada de dados vetorial para ser utilizado pelo algoritmo a ser executado. Com relação aos gráficos gerados pelo ambiente R e exibidos na interface do QGIS, a janela para exibição de resultados (Figura 3 (b)) utiliza imagens de tamanho fixo que foram obtidas como retorno do algoritmo executado, limitando a sua visualização e manipulação pelo usuário final. Outras dificuldades também são encontradas com relação às mensagens de retorno de execução dos algoritmos, realizada por meio de uma janela textual que dificulta a identificação de possíveis erros que possam vir a ocorrer.

Com relação às ferramentas disponíveis na literatura, algumas questões podem ser discutidas. A primeira ferramenta computacional que surgiu para auxiliar o usuário final no delineamento de UGDs em AP foi o *software Management Zones Analyst* (MZA) (FRIDGEN et al., 2004), o qual é ainda muito utilizado em diversas abordagens da literatura. O MZA possui funcionalidades importantes que fornecem estatísticas descritivas para análise e escolha do melhor mapa de UGDs a ser utilizado. Entretanto, sua disponibilidade para instalação apenas sob o sistema operacional Windows, exigência de dados de entrada já em uma grade espacial única, e indisponibilidade de acesso a bases de dados e serviços *Web*, limitam o seu uso se considerarmos os conceitos atuais de TICs e o processo de KDD que envolvem o delineamento de UGDs em AP. A arquitetura proposta neste artigo visa resolver algumas dessas questões, possibilitando ao usuário final obter dados de diversas fontes, e que podem passar por todas as etapas do KDD até a extração final do conhecimento em mapas de UGDs.

Com o objetivo de prover uma solução acessível pela *Web* aos usuários finais, disponibilizando imagens de sensoriamento remoto e uma interface amigável e simples de ser utilizada, Zhang et al. (2010) desenvolveram o aplicativo ZoneMap. Segundo experimentos realizados pelos autores, o ZoneMap permite o delineamento de mapas de UGDs bastante compatíveis com mapas delineados manualmente pelos usuários finais. Entretanto, uma dificuldade que pode ser observada é a necessidade dos usuários finais informarem os seus dados vetoriais a um sistema externo, mesmo que este esteja instalado em servidores de universidades ou órgãos governamentais. Com relação à essa questão, a arquitetura proposta neste artigo permite que

o usuário final trabalhe os seus dados em seu ambiente local, utilizando um SIG, e publique, por exemplo, apenas mapas de UGDs finais em um servidor que pode ser externo à sua rede de trabalho.

A possibilidade de utilização dos padrões OGC para aplicações em AP já vem sendo estudada a algum tempo. Murakami et al. (2007) propuseram uma estrutura completa para desenvolvimento de sistemas em AP utilizando esses padrões, que posteriormente foi atualizada e incluída em um portal *Web* para serviços agrícolas (RIBEIRO-JÚNIOR, 2007). O trabalho de Nash, Korduan e Bill (2009) também mostrou a eficiência do uso de padrões OGC em atividades de AP que exigem processamento paralelo, possibilitando a execução dessas atividades utilizando diversos servidores distintos. Nesse sentido, a arquitetura proposta neste artigo também teve como objetivo estabelecer um modelo baseado nos padrões da OGC, porém já estabelecendo opções de ferramentas livres e de distribuição gratuita que podem ser utilizadas sem custos pelo usuário final. A Tabela 1 resume as principais características das soluções encontradas na literatura no que diz respeito ao acesso, entrada e saída de dados, em comparação com a arquitetura proposta.

Tabela 1: Resumo das características das soluções da literatura com relação à proposta de arquitetura apresentada.

Ferramenta	**Tipo Aplicação**	**Fontes de Dados**	**Saídas**
Fridgen et al. (2004)	Desktop	Arquivo ASCII	Arquivo ASCII Arquivo Estatísticas Arquivo Indicadores Gráficos
Murakami et al. (2007)	Web	OGC: WMS,WFS,WCS	OGC: WCS, WMS,WFS
Nash, Korduan e Bill (2009)	Web	OGC: WMS,WFS,WCS	OGC: WMS,WFS,WCS
Zhang et al. (2010)	Web	Arquivo ASCII Arquivo Raster BD Imagens	Arquivo ASCII Arquivo Raster Arquivo Shape
Arquitetura Proposta	Desktop / Web	OGC: WMS,WFS	OGC: WMS,WFS Gráficos

A partir da Tabela 1, verifica-se que, diferentemente das soluções disponíveis na literatura, a arquitetura proposta possibilita ao usuário final manipular os dados espaciais por meio de um SIG *Desktop*, contendo todas as funcionalidades de geoprocessamento úteis para a aplicação e que, por muitas vezes, possuem custo computacional elevado de execução. Entretanto, as fontes de dados devem ser remotas, e serem acessadas a partir de requisições a provedores de serviços da OGC. O acesso às fontes de dados é realizado de maneira semelhante pelas outras soluções, exceto para a solução desenvolvida por Fridgen et al. (2004), onde é permitida apenas a utilização de um formato específico de arquivo, a ser informado pelo usuário final. Com relação à saída de resultados, a arquitetura proposta proporciona, além das camadas de dados

geoespaciais nos formatos OGC, gráficos de avaliações qualitativas a respeito dos resultados obtidos. Esse tipo de indicador também é fornecido pela abordagem desenvolvida por Fridgen et al. (2004).

CONCLUSÕES

A utilização de TICs livres e de distribuição gratuita integradas em uma arquitetura única possibilitou a análise de que é possível executar processos importantes como o delineamento de UGDs em AP utilizando os conceitos de descoberta de conhecimento em bancos de dados. Outras soluções disponíveis na literatura mostram que, ao longo do tempo, está se estabelecendo uma tendência de utilização de novas tecnologias para processos agrícolas que envolvem TICs, capazes de aumentar a interoperabilidade entre sistemas e auxiliar a tomada de decisão do usuário final.

Com relação à arquitetura proposta neste artigo, deve-se destacar o uso de ferramentas livres e de padrões OGC, realizado de maneira similar em outras soluções disponíveis na literatura. Por outro lado, a arquitetura proposta possibilita ao usuário final a manipulação dos dados espaciais de maneira local, permitindo a utilização de funcionalidades de geoprocessamento computacionalmente custosas. Essa característica pode ser interessante ao usuário final, à medida que é desejável que apenas mapas finalizados devem ser publicados e visualizados por outros usuários.

A arquitetura proposta ainda deve levar em consideração o estabelecimento de melhorias que podem tornar os seus módulos mais independentes. Uma dessas melhorias está relacionada ao acesso aos algoritmos de geoprocessamento e mineração de dados executados utilizando o ambiente de desenvolvimento R, que na versão atual da arquitetura devem estar instalados localmente em conjunto com o SIG no computador local do usuário final. Para tornar esse módulo mais independente, uma nova proposta da arquitetura, considerando a utilização de uma versão servidor do ambiente R, o RServe (RSERVE, 2017), está sendo elaborada. Nessa nova proposta, os algoritmos serão executados diretamente no servidor, por meio de requisições do serviço WPS diretamente do SIG para o RServe, utilizando a ferramenta WPS4R (HINZ et al., 2013). Assim, algoritmos específicos poderão ser disponibilizados apenas no servidor, evitando a necessidade de instalação e atualização em computadores locais.

AGRADECIMENTOS

Agradecemos à Fazenda Aparecida e as pesquisadoras da Embrapa Célia Regina Grego e Cristina Aparecida Rodrigues, pela disponibilidade dos dados coletados para uso experimental neste artigo. Agradecemos também as agências de fomento em pesquisa CNPq, CAPES, FAPESP e FINEP. O primeiro autor é apoiado pelo programa de pós-graduação da Embrapa, e o segundo autor, pela bolsa de produtividade em pesquisa do CNPq número 311868/2015-0.

REFERÊNCIAS

ARCGIS: Esri arcgis. 2017. Disponível em: <http://www.arcgis.com/>. Acesso em: 02 mai. 2017.

ARONOFF, S. Geographic information systems: A management perspective. *Geocarto International*, v. 4, n. 4, p. 58, 1989. Disponível em: <http://dx.doi.org/10.1080/10106048909354237>.

BERNARDI, A. C. d. C. et al. *Agricultura de Precisão - Resultados de um Novo Olhar*. 1. ed. Brasília: Empresa Brasileira de Pesquisa Agropecuaria, 2014. 596 p.

BURROUGH, P. A. Principles of GIS for land resources assessment. *Monographs on soil and resources survey. Clarendon, Oxford*, 1986.

CÂMARA, G. et al. *Anatomia dos Sistemas de Informações Geográficas*. [S.l.]: Campinas-SP: Instituto de Computação, UNICAMP, 1996.

CASANOVA, M. A. et al. *Bancos de Dados Geográficos*. Curitiba: MundoGeo, 2005. 506 p.

CIFERRI, R. R. *Um benchmark voltado a análise de desempenho de sistemas de informações geográficas*. Dissertação (Mestrado em Ciência da Computação) — Universidade Estadual de Campinas, 1995.

COMPARETTI, A. Precision Agriculture: Past, Present and Future. In: *Agricultural Engineering and Environment*. Akademija, Lithuania: [s.n.], 2011. p. 17.

ELMASRI, R.; NAVATHE, S. B. *Fundamentals of Database Systems*. 6th. ed. Boston: Addison-Wesley, 2011. 1172 p. ISSN 14337851. ISBN 9780136086208.

ESTER, M. et al. Spatial Data Mining: Database Primitives, Algorithms and Efficient DBMS Support. *Data Min. Knowl. Discov.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 4, n. 2-3, p. 193–216, 2000. ISSN 1384-5810. Disponível em: <http://dx.doi.org/10.1023/A:1009843930701>.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Commun. ACM*, ACM, New York, NY, USA, v. 39, n. 11, p. 27–34, 1996. ISSN 0001-0782. Disponível em: <http://dx.doi.org/10.1145/240455.240464>.

FRIDGEN, J. J. et al. Management Zone Analyst (MZA): Software for Subfield Management Zone Delineation. *Agronomy Journal*, v. 96, p. 100–108, 2004.

GEONODE: Open source geospatial content management system. 2017. Disponível em: <http://geonode.org/>. Acesso em: 04 mai. 2017.

GEOSERVER. *Open Source Server for Sharing Geospatial Data*. 2017. Disponível em: <http://geoserver.org>. Acesso em: 08 mai. 2017.

HINZ, M. et al. Spatial Statistics on the Geospatial Web. In: *The 16th AGILE International Conference on Geographic Information Science, Short Papers*. [S.l.: s.n.], 2013.

LI, M.; CHUNG, S.-O. Special issue on precision agriculture. *Computers and Electronics in Agriculture*, Elsevier B.V., v. 112, p. 1, 2015. ISSN 01681699. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0168169915000897>.

MASSRUHÁ, S. M. F. S. et al. *Tecnologias da informação e comunicação e suas relações com a agricultura.* 1. ed. Brasília, DF: Embrapa Informática Agropecuária, 2014. 411 p. ISBN 978-85-7035-414-3.

MITCHELL, T. *Web mapping illustrated: using open source GIS toolkits*. [S.l.]: O'Reilly Media, Inc., 2005.

MOLIN, J. P. Agricultura de Precisão: Situação atual e perspectivas. In: FANCELLI, A. L.; NETO, D. D. (Ed.). *Milho: Estratégias de Manejo para Alta Produtividade*. Piracicaba: ESALQ/USP/LPV, 2003. p. 89–98.

MOLIN, J. P.; AMARAL, L. R. do; COLAÇO, A. *Agricultura de precis{ã}o:*. Oficina de Textos, 2015. ISBN 9788579752148. Disponível em: <https://books.google.com.br/books?id=MX7jCgAAQBAJ>.

MURAKAMI, E. et al. An infrastructure for the development of distributed service-oriented information systems for precision agriculture. *Computers and Electronics in Agriculture*, v. 58, n. 1, p. 37–48, ago. 2007. ISSN 0168-1699. Disponível em: <dx.doi.org/10.1016/j.compag.2006.12.010>.

NASH, E.; KORDUAN, P.; BILL, R. Applications of open geospatial web services in precision agriculture: a review. *Precision Agriculture*, Springer US, v. 10, n. 6, p. 546–560, 2009. ISSN 1385-2256. Disponível em: <http://dx.doi.org/10.1007/s11119-009-9134-0>.

ORACLE. *Oracle Documentation*. 2017. Disponível em: <http://docs.oracle.com>. Acesso em: 02 mai. 2017.

PERCIVALL, G. et al. *OGC Reference Model*. Open Geospatial Consortium Inc., 2003. 108 p. Disponível em: <http://www.opengeospatial.org/standards/orm>.

POSTGRESQL. *PostgreSQL Documentation*. 2017. Disponível em: <http://www.postgresql.org/docs/9.5/static/docguide.html>. Acesso em: 02 mai. 2015.

QGIS: A free and open souerce geographic information system. 2017. Disponível em: <http://www.arcgis.com/>. Acesso em: 04 mai. 2017.

R. *The R Project for Satistical Computing*. 2017. Disponível em: <http://www.r-project.org/>. Acesso em: 15 mai. 2017.

REZENDE, S. O. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Editora Manole Ltda, 2003. ISBN 8520416837.

RIBEIRO-JÚNIOR, L. C. M. *Uma Arquitetura de Software para Sistemas Espaço-Temporais Baseados na Web para Agricultura de Precisão*. 190 p. Tese (Doutorado em Engenharia) — Universidade de São Paulo, 2007. Disponível em: <http://www.teses.usp.br/teses/disponiveis/3/3141/tde-01082007-182328>.

RSERVE: Binary r server. 2017. Disponível em: <http://rforge.net/Rserve>. Acesso em: 08 mai. 2017.

RUSS, G. *Spatial Data Mining in Precision Agriculture*. 251 p. Tese (Doktoringenieur) — Otto-von-Guericke-University of Magdeburg, 2012.

SANTOS, R. T. dos. *Um modelo de referência para o processo de definição de zonas de manejo em agricultura de precisão*. 115 p. Tese (Doutorado em Engenharia) — Universidade de São Paulo, 2014.

SANTOS, R. T. dos; MOLIN, J. P.; SARAIVA, A. M. A Reference Process for Management Zone Delineation. In: *EFITA Conference - Sustainable Agriculture through ICT Innovation*. [S.l.: s.n.], 2013. p. 8.

WEISS, S. M.; INDURKHYA, N. *Predictive Data Mining: A Practical Guide*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. ISBN 978-1558604032.

ZHANG, X. et al. Zone mapping application for precision-farming: a decision support tool for variable rate application. *Precision Agriculture*, Springer US, v. 11, n. 2, p. 103–114, 2010. ISSN 1385-2256. Disponível em: <http://dx.doi.org/10.1007/s11119-009-9130-4>.

Apendice H

REVISÃO SISTEMÁTICA - AGRUPAMENTO DE DADOS ESPACIAIS

Este apêndice descreve as principais etapas do processo de revisão de trabalhos correlatos à esta tese, no que diz respeito a agrupamento de dados espaciais. Esse processo foi realizado seguindo o conceito de revisão sistemática, aplicado em bases de dados relevantes com o auxílio da ferramenta StArt (ZAMBONI, 2010; HERNANDES, 2012). Os procedimentos e análises apresentados neste apêndice resultaram na seleção dos trabalhos sumarizados nas Subseções 5.2.2 e 5.2.3 do Capítulo 5.

H.1 Revisão sistemática e a ferramenta StArt

A revisão sistemática é uma metodologia de pesquisa que tem como objetivo buscar, de uma maneira formal, evidências na literatura científica, a partir de etapas bem definidas que levam em conta um protocolo previamente elaborado (FABBRI, 2012). Como isso, possibilita o esclarecimento do estado da arte de uma determinada área de estudo, facilitando a elaboração de novas pesquisas (KITCHENHAM, 2010). A ferramenta StArt (*State of the Art through Systematic Review*) (ZAMBONI, 2010; HERNANDES, 2012) foi utilizada para auxiliar no processo de revisão sistemática realizado para esta tese. Essa ferramenta fornece suporte para aplicação de protocolos, por meio de etapas bem definidas para identificação, seleção, extração e sumarização de pesquisas presentes na literatura. A seguir, são descritos os principais passos realizados durante a revisão sistemática, no que diz respeito ao planejamento, execução e seleção de estudos.

H.1.1 Planejamento

Durante essa etapa, devem ser formuladas questões que estabelecem a área de interesse da pesquisa a ser realizada, com o intuito de auxiliar na formulação de *strings* de busca a serem aplicadas nas bases de pesquisa selecionadas. No caso desta tese, o principal questionamento a ser considerado é se existem soluções para agrupamento de dados espaciais que tratam o espaço de coordenadas e o espaço de atributos de maneiras distintas.

Com os questionamentos formulados, as fontes de busca e seleção de estudos devem ser escolhidas, levando-se em conta principalmente a disponibilidade de trabalhos na área de ciência da computação e do texto completo dos artigos. Desse modo, as fontes escolhidas foram: ACM Digital Library; IEEEXplore Digital Library; Scopus; e Digital Bibliography & Library Project (DBLP). Com relação aos idiomas, foram escolhidos o inglês, por ser a língua internacional mais aceita para a redação de trabalhos científicos; e o português, para possibilitar a busca por trabalhos realizados por pesquisadores brasileiros publicados em eventos ou periódicos científicos nacionais. Além disso, também foram determinadas as palavras-chave utilizadas durante o processo, agrupadas em termos mais abrangentes:

- Agrupamento espacial: *clustering*, *spatial clustering*, *spatial data clustering*, *polygonal clustering*, *point clustering*;
- Relacionamentos espaciais e dimensionalidade: *spatial contiguity*, *spatial correlation*, *spatial autocorrelation*, *high dimensional*, *high spatial*, *large spatial*, *large dimensional*, *spatial obstacle*, *spatial distribution*;
- Tipos de agrupamento: *agglomerative*, *hierarchical*, *membership*, *hybdrid*, *partitioning*, *density*.

As palavras-chave supracitadas na língua inglesa, foram substituídas por seus sinônimos na língua portuguesa durante as buscas realizadas para esse idioma nas fontes selecionadas.

Nessa etapa, também foram determinados os critérios de inclusão e exclusão a serem utilizados na seleção dos estudos. Esses critérios permitem aumentar a credibilidade da revisão, de forma a garantir a manutenção do seu foco até o final. Os critérios de inclusão selecionados foram:

- Estudos voltados ao agrupamento espacial de pontos;
- Estudos voltados ao agrupamento espacial de polígonos;

- Estudos que consideram obstáculos espaciais durante o agrupamento;
- Estudos que consideram restrições espaciais;
- Estudos aplicados ao delineamento de UGDs em AP.

Por sua vez, os critérios de exclusão selecionados foram:

- Estudos que consideram apenas o espaço de coordenadas para o agrupamento;
- Estudos que consideram o espaço de coordenadas apenas para encontrar relacionamentos entre os atributos ou reduzir sua dimensionalidade;
- Estudos que consideram apenas o espaço de atributos para o agrupamento;
- Abordagens de agrupamento baseadas na densidade de distribuição espacial das amostras;
- Estudos com quantidade não significativa de palavras-chave presentes no título e no *abstract* da publicação.

Finalizando a etapa de planejamento, foram definidos os procedimentos a serem realizados para a seleção dos estudos. Inicialmente, foi definido um processo de seleção preliminar, que deve construir *strings* de busca utilizando as palavras-chave definidas e os operados lógicos 'OR' e 'AND' para aplicação nas fontes. Os trabalhos duplicados obtidos são identificados, e uma primeira filtragem deve ser realizada, considerando os critérios de inclusão e exclusão verificados a partir da leitura dos *abstracts*. Na sequência, foi definido um processo de seleção final, onde é realizada a leitura completa dos trabalhos selecionados na etapa anterior. Finalizando, foi definido um processo de extração dos resultados, onde são sintetizadas as características importantes e pertinentes de cada trabalho selecionado para leitura, e que possuem correlação e aplicação nesta tese.

H.1.2 Execução

O processo de execução foi realizado com a construção de *strings* de busca para as fontes selecionadas, conforme descrito na etapa de planejamento. Algumas dessas fontes possuem limitações com relação à quantidade de termos utilizados, impossibilitando o uso de todas as palavras-chaves previstas. As *strings* de busca definidas em inglês foram:

- **ACM Digital Library**: ((acmdlTitle:"clustering"OR acmdlTitle:"spatial clustering"OR acmdlTitle:"spatial data clustering"OR acmdlTitle:"polygonal clustering"OR acmdlTitle:"point clustering") AND (acmdlTitle:"spatial contiguity"OR acmdlTitle:"spatial correlation"OR acmdlTitle:"spatial autocorrelation"OR acmdlTitle:"high dimensional"OR acmdlTitle:"high spatial"OR acmdlTitle:"large spatial"OR acmdlTitle:"large dimensional"OR acmdlTitle:"spatial obstacle"OR acmdlTitle:"spatial distribution") AND ((acmdlTitle:"agglomerative"OR acmdlTitle:"hierarchical"OR acmdlTitle:"membership"OR acmdlTitle:"hybrid"OR acmdlTitle:"partitioning") AND NOT acmdlTitle:"density")) OR ((recordAbstract:"clustering"OR recordAbstract:"spatial clustering"OR recordAbstract:"spatial data clustering"OR recordAbstract:"polygonal clustering"OR recordAbstract:"point clustering") AND (recordAbstract:"spatial contiguity"OR recordAbstract:"spatial correlation"OR recordAbstract:"spatial autocorrelation"OR recordAbstract:"high dimensional"OR recordAbstract:"high spatial"OR recordAbstract:"large spatial"OR recordAbstract:"large dimensional"OR recordAbstract:"spatial obstacle"OR recordAbstract:"spatial distribution") AND ((recordAbstract:"agglomerative"OR recordAbstract:"hierarchical"OR recordAbstract:"membership"OR recordAbstract:"hybrid"OR recordAbstract:"partitioning") AND NOT recordAbstract:"density"))

- **IEEEXplore Digital Library**: ((clustering OR "spatial clustering"OR "polygonal clustering "OR "point clustering") AND ("spatial contiguity"OR "spatial correlation"OR "high dimensional"OR "spatial obstacle") AND (agglomerative OR hierarchical OR membership OR hybrid OR partitioning))

- **Scopus**: TITLE-ABS-KEY ((clustering OR spatial clustering OR spatial data clustering OR polygonal clustering OR point clustering) AND (spatial contiguity OR spatial correlation OR spatial autocorrelation OR high dimensional OR high spatial OR large spatial OR large dimensional OR spatial obstacle OR spatial distribution) AND ((agglomerative OR hierarchical OR membership OR hybrid OR partitioning) AND NOT density))

- **DBLP**: spatial.clustering agglomerative|hierarchical|membership|hybrid|partitioning

Foram realizadas buscas similares, com os temos em português, mas que não retornaram resultados em nenhuma das fontes selecionadas.

H.2 Seleção de estudos

Os resultados obtidos na etapa de execução foram salvos em formato *BibTex* e importados para a ferramenta StArt. Após a execução desse processo e eliminando-se os trabalhos duplicados, restaram 1427 trabalhos selecionados. A partir da aplicação dos critérios de inclusão e exclusão e da leitura dos *abstracts*, foram selecionados 27 trabalhos para leitura completa. Após a leitura e análise crítica dos trabalhos, foi realizada uma sumarização com as características mais importantes de cada um deles relacionadas à esta tese, conforme descrito nas Subseções 5.2.2 e 5.2.3 do Capítulo 5. A Figura H.1 mostra a distribuição por fonte de busca dos trabalhos obtidos na seleção preliminar dos estudos. Essa proporção foi mantida durante as seleções seguintes, destacando a base Scopus como a principal fonte de artigos relacionados à linha de pesquisa desta tese.

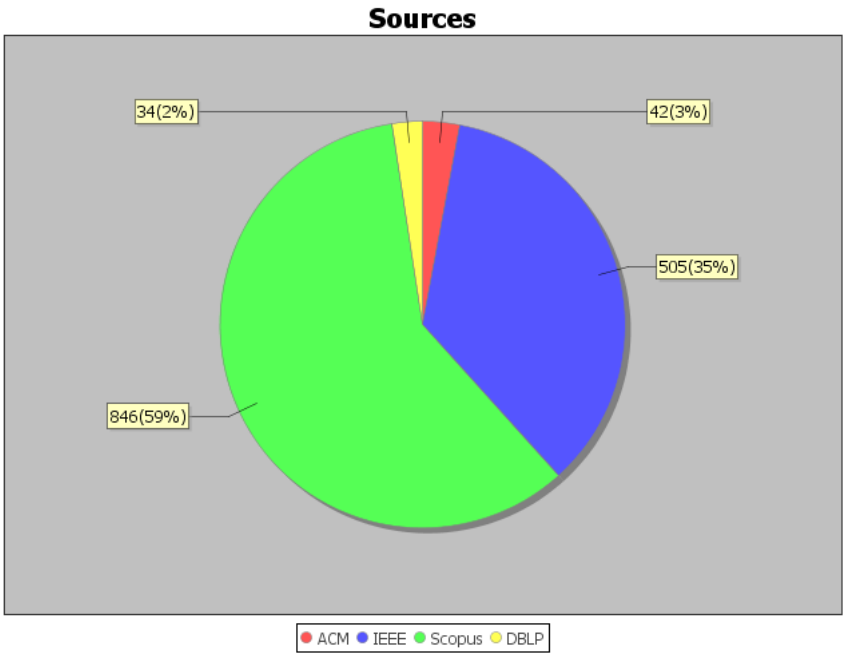


Figura H.1: Distribuição dos trabalhos obtidos na seleção preliminar dos estudos, por fonte de busca.

Anexo A

ESTATÍSTICA DESCRITIVA DOS ATRIBUTOS

Este anexo contém a estatística descritiva referentes aos dados oriundos das UPs descritas na Seção 6.5 do Capítulo 6 e utilizadas nos experimentos desta tese, bem como o coeficiente de correlação de Pearson entre pares de atributos.

Tabela A.1: Média, Variância e Desvio Padrão referente aos atributos da UP-CA, após a interpolação espacial.

Atributo	Média	Variância	Desvio Padrão
Areia - 15 cm (g/g)	36,54	1,24	1,12
Areia - 30 cm (g/g)	36,82	9,16	3,03
Argila - 15 cm (g/g)	49,89	6,12	2,47
Argila - 30 cm (g/g)	50,52	12,59	3,55
CE - 15 cm (dS/m)	0,51	0,002	0,04
CE - 30 cm (dS/m)	0,50	0,002	0,04
CE - 30 cm (mS/m)	1,76	0,24	0,49
CE - 90 cm (mS/m)	1,63	0,36	0,60
Cota Altimétrica (m)	666,42	108,62	10,42
Densidade - 15 cm (g/cm3)	1,24	0,002	0,05
Densidade - 30 cm (g/cm3)	1,30	0,001	0,03
NDVI 2011	0,75	0,002	0,04
NDVI 2012	0,31	0,001	0,03
NDVI 2013	0,77	0,002	0,04
NDVI 2014	0,65	0,001	0,03
Produtividade 2010 (ton/ha)	118,55	220,03	14,83
Produtividade 2012 (colmos/m2)	1,69	0,043	0,21
Produtividade 2013 (ton/ha)	76,23	15,35	3,92
Umidade - 15 cm (g/g)	0,27	0,001	0,01
Umidade - 30 cm (g/g)	0,27	0,001	0,004

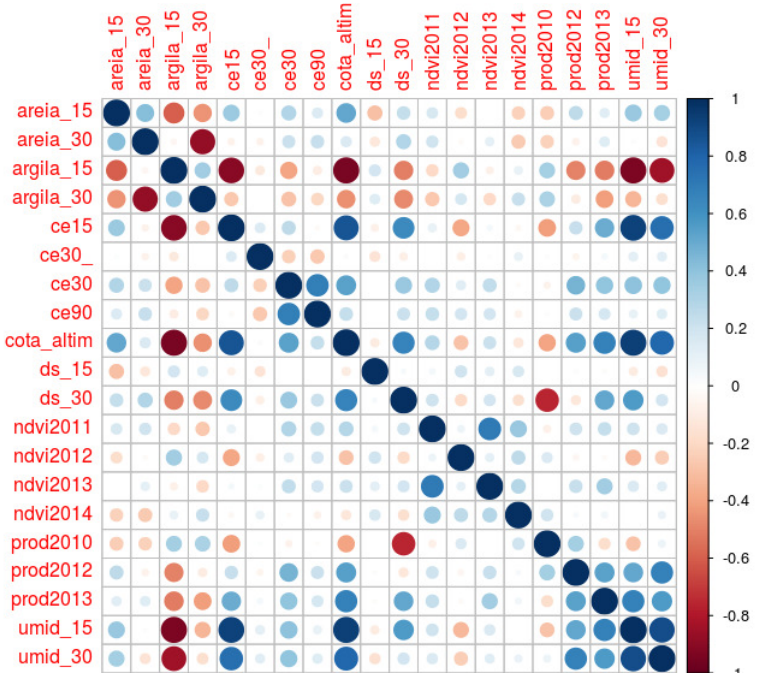


Figura A.1: Coeficiente de correlação de Pearson entre pares de atributos da UP-CA. Os atributos estão distribuídos na mesma ordem da Tabela A.1; círculos maiores representam correlações maiores.

Tabela A.2: Média, Variância e Desvio Padrão referente aos atributos da UP-UV, após a interpolação espacial.

Atributo	Média	Variância	Desvio Padrão
CE - 20 cm (mS/m)	6,73	1,55	1,24
CE - 40 cm (mS/m)	6,96	1,23	1,11
Teor de Clorofila A (SPAD)	41,92	3,80	1,95
Teor de Clorofila B (SPAD)	11,99	1,17	1,08
Teor de Clorofila Total (SPAD)	53,91	8,38	2,89

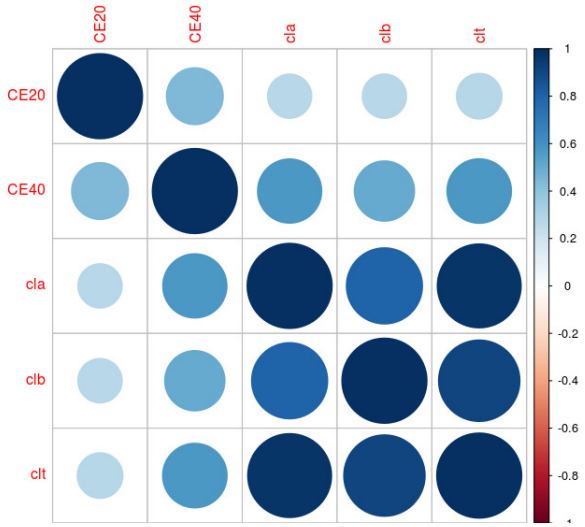


Figura A.2: Coeficiente de correlação de Pearson entre pares de atributos da UP-UV. Os atributos estão distribuídos na mesma ordem da Tabela A.2; círculos maiores representam correlações maiores.

Tabela A.3: Média, Variância e Desvio Padrão referente aos atributos da UP-G, após a interpolação espacial.

Atributo	Média	Variância	Desvio Padrão
Altitude (m)	1020,55	138,29	11,76
CE - 30 cm (mS/m)	8,47	1,31	1,14
CE - 90 cm (mS/m)	4,03	0,20	0,45
Produtividade Milho (ton/ha)	7,53	3,78	1,94
Produtividade Soja (ton/ha)	3507,37	51152,27	226,17

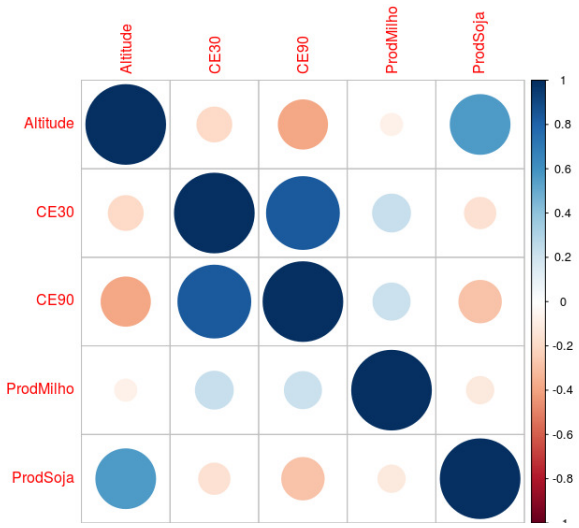


Figura A.3: Coeficiente de correlação de Pearson entre pares de atributos da UP-G. Os atributos estão distribuídos na mesma ordem da Tabela A.3; círculos maiores representam correlações maiores.