UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ – CAMPUS CASCAVEL

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA

MÓDULOS COMPUTACIONAIS PARA SELEÇÃO DE VARIÁVEIS E ANÁLISE DE AGRUPAMENTO PARA DEFINIÇÃO DE ZONAS DE MANEJO

ALAN GAVIOLI

CASCAVEL – PR
FEVEREIRO DE 2017

ALAN GAVIOLI

MÓDULOS COMPUTACIONAIS PARA SELEÇÃO DE VARIÁVEIS E ANÁLISE DE AGRUPAMENTO PARA DEFINIÇÃO DE ZONAS DE MANEJO

Tese apresentada ao Programa de Pós-Graduação em Engenharia Agrícola, em cumprimento parcial aos requisitos para obtenção do título de Doutor em Engenharia Agrícola, área de concentração Sistemas Biológicos e Agroindustriais.

Orientador: Prof. Dr. Eduardo Godoy de Souza
Coorientador: Prof. Dr. Claudio Leones Bazzi

CASCAVEL – PR
FEVEREIRO DE 2017

Dados Internacionais de Catalogação-na-Publicação (CIP)

G243m

Gavioli, Alan

Módulos computacionais para seleção de variáveis e análise de agrupamento para definição de zonas de manejo. / Alan Gavioli. Cascavel, 2017.

128 f.

Orientador: Prof. Dr. Eduardo Godoy de Souza
Coorientador: Prof. Dr. Claudio Leones Bazzi
Revisão Português, Inglês e Normas: Dhandara Capitani

Tese (Doutorado) – Universidade Estadual do Oeste do Paraná, Campus de Cascavel, 2017

Programa de Pós-Graduação em Engenharia Agrícola

1. Agricultura de precisão. 2. Agrupamento de dados. 3. Análise de componentes principais. 4. Multispati-PCA. 5. Software para agricultura. I.Souza, Eduardo Godoy de. II. Bazzi, Claudio Leones. III. Capitani, Dhandara, rev. IV. Universidade Estadual do Oeste do Paraná. V. Título.

CDD 20.ed. 630
CIP-NBR 12899

Ficha catalográfica elaborada por Helena Soterio Bejio – CRB 9ª/965

ALAN GAVIOLI

Módulos Computacionais de Seleção de Variáveis e Análise de Agrupamento para Definição de Zonas de Manejo

Tese apresentada ao Programa de Pós-Graduação em Engenharia Agrícola em cumprimento parcial aos requisitos para obtenção do título de Doutor em Engenharia Agrícola, área de concentração Sistemas Biológicos e Agroindustriais, linha de pesquisa Geoprocessamento, Estatística Espacial e Agricultura de Precisão, APROVADO(A) pela seguinte banca examinadora:

Orientador(a) - Eduardo Godoy de Souza

Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Luciana Pagliosa Carvalho Guedes

Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Raimundo Pinheiro Neto

Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Antônio Carlos Andrade Gonçalves

Universidade Estadual de Maringá (UEM)

Marcio Furlan Maggi

Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Cascavel, 17 de fevereiro de 2017

BIOGRAFIA RESUMIDA

Alan Gavioli nasceu em 04 de maio de 1979, no município de Cianorte, estado do Paraná. Em 1999, iniciou o curso de Bacharelado em Ciência da Computação na Universidade Estadual de Londrina (UEL), concluindo-o em 2002. Em 2003, ingressou no curso de Mestrado em Ciência da Computação na Universidade Federal de São Carlos (UFSCar), sob orientação do professor Dr. Mauro Biajiz. Em 2005, concluiu o mestrado com a dissertação intitulada “Sistema para Recuperação de Imagens com base em Características Geométricas, Conjuntos Nebulosos e Indexação Métrica”. Em fevereiro de 2013, ingressou no curso de Doutorado em Engenharia Agrícola – área de concentração Sistemas Biológicos e Agroindustriais – na Universidade Estadual do Oeste do Paraná (UNIOESTE – campus Cascavel), sob orientação do professor Dr. Eduardo Godoy de Souza e coorientação do professor Dr. Claudio Leones Bazzi. Iniciou sua atuação profissional em docência no ensino superior em fevereiro de 2005, como professor e coordenador de cursos de graduação e pós-graduação lato sensu na área de Ciência da Computação na Faculdade de Ciências Aplicadas de Cascavel (UNIPAN). No período de 2005 a 2008 também exerceu a função de professor em outras duas instituições de ensino superior privadas: Faculdade UNICA e Faculdade Iguaçu. Desde setembro de 2008 é servidor público federal e atua na Universidade Tecnológica Federal do Paraná (UTFPR – campus Medianeira) como professor de graduação e de pós-graduação lato sensu, vinculado ao Departamento Acadêmico de Computação.

DEDICATÓRIA

Dedico este trabalho aos meus pais, Luiz e Nadir, que sempre fizeram tudo que era possível para me apoiar e incentivar.

AGRADECIMENTOS

A Deus, agradeço principalmente pela saúde e pelas oportunidades que coloca em minha vida. Graças a Ele, tenho conseguido superar os obstáculos que surgem e alcançar os objetivos que estabeleço para minha vida.

Ao meu orientador, Dr. Eduardo Godoy de Souza, por sua dedicação contínua ao meu projeto de doutorado nesses quatro anos de convivência. Também agradeço por sua confiança, pela amizade e pelas oportunidades que me ofereceu para que eu pudesse evoluir como pesquisador.

Ao meu coorientador, Dr. Claudio Leones Bazzi, por sua disposição para contribuir para a evolução do meu trabalho. Suas sugestões sempre foram relevantes, especialmente para a implementação dos módulos computacionais do projeto.

Ao Programa de Pós-Graduação em Engenharia Agrícola (PGEAGRI) da UNIOESTE, pela oportunidade oferecida de cursar o doutorado nesta renomada instituição.

À Universidade Tecnológica Federal do Paraná (UTFPR – campus Medianeira), pelas várias formas de apoio concedidas a mim durante a realização do doutorado. Esse suporte foi fundamental para que eu pudesse desenvolver este trabalho com dedicação e tranquilidade.

A todos os professores do PGEAGRI com os quais tive contato, por sempre terem mostrado disposição para compartilhar seu conhecimento. Em especial, à Drª. Luciana Pagliosa Carvalho Guedes, por suas relevantes sugestões relacionadas a métodos de análise multivariada e de agrupamento de dados.

À minha noiva, Maryana, pelo amor, a cumplicidade, o apoio e o incentivo contínuos. Não foram poucos os momentos em que modificou seus planos para que eu pudesse me dedicar a atividades exigidas pelo doutorado.

Aos meus pais, Luiz e Nadir, pelo amor, o incentivo e a educação irrepreensível que me ofereceram. Continuarei sendo grato a eles durante toda a minha vida.

A todos os colegas do Laboratório de Mecanização e Agricultura de Precisão (LAMAP) da UNIOESTE, pela convivência harmoniosa e pela parceria em atividades importantes para o meu projeto.

Aos proprietários das áreas agrícolas localizadas em Céu Azul (Aldo Tasca) e Serranópolis do Iguaçu (Wanderley Schenatto), por terem permitido que eu e os colegas do LAMAP realizássemos diversas atividades de amostragem nessas áreas.

Por fim, agradeço a todos que de alguma maneira me ajudaram ou torceram por mim ao longo dos quatro anos que dediquei à realização deste sonho.

MÓDULOS COMPUTACIONAIS PARA SELEÇÃO DE VARIÁVEIS E ANÁLISE DE AGRUPAMENTO PARA DEFINIÇÃO DE ZONAS DE MANEJO

RESUMO

A seleção de variáveis e a análise de agrupamento de dados são atividades fundamentais para a definição de zonas de manejo (ZMs) de qualidade. Para executar essas duas atividades, existem diversos métodos propostos, que devido à sua complexidade precisam ser executados por meio da utilização de sistemas computacionais. Neste trabalho, avaliaram-se 5 métodos de seleção de variáveis baseados em análise de correlação espacial, análise de componentes principais (ACP) e análise espacial multivariada baseada no índice de Moran e em ACP (MULTISPATI-PCA). Propôs-se um novo algoritmo de seleção de variáveis, denominado MPCA-SC, desenvolvido a partir da aplicação conjunta da análise de correlação espacial e de MULTISPATI-PCA. Avaliou-se a viabilidade de aplicação de 20 algoritmos de agrupamento de dados para a geração de ZMs: average linkage, bagged clustering, centroid linkage, clustering large applications, complete linkage, divisive analysis, fuzzy analysis clustering (fanny), fuzzy c-means, fuzzy c-shells, hard competitive learning, hybrid hierarchical clustering, k-means, median linkage, método de McQuitty (mcquitty), método de Ward, neural gas, partitioning around medoids, single linkage, spherical k-means e unsupervised fuzzy competitive learning. Apresentaram-se ainda dois módulos computacionais desenvolvidos para disponibilizar os métodos de seleção de variáveis e de agrupamento de dados para a definição de ZMs. As avaliações foram realizadas com dados obtidos entre os anos de 2010 e 2015 de três áreas agrícolas comerciais, localizadas no estado do Paraná, nas quais cultivaram-se milho e soja. Os experimentos efetuados para avaliar os 5 algoritmos de seleção de variáveis mostraram que o novo método MPCA-SC pode melhorar a qualidade de ZMs em diversos aspectos, mesmo obtendo-se resultados satisfatórios com os outros 4 algoritmos. Os experimentos de avaliação dos 20 métodos de agrupamento citados mostraram que 17 deles foram adequados para o delineamento de ZMs, com destaque para fanny e mcquitty. Por fim, concluiu-se que os dois módulos computacionais desenvolvidos possibilitaram a obtenção de ZMs de qualidade. Além disso, esses módulos constituem uma ferramenta computacional mais abrangente que outros softwares de uso gratuito, como FuzME, MZA e SDUM, em relação à diversidade de algoritmos disponibilizados para selecionar variáveis e agrupar dados.

Palavras-chave: agricultura de precisão; agrupamento de dados; análise de componentes principais; MULTISPATI-PCA; software para agricultura.

COMPUTATIONAL MODULES FOR VARIABLE SELECTION AND CLUSTER ANALYSIS FOR DEFINITION OF MANAGEMENT ZONES

ABSTRACT

Two basic activities for the definition of quality management zones (MZs) are the variable selection task and the cluster analysis task. There are several methods proposed to execute them, but due to their complexity, they need to be made available by computer systems. In this study, 5 methods based on spatial correlation analysis, principal component analysis (PCA) and multivariate spatial analysis based on Moran's index and PCA (MULTISPATI-PCA) were evaluated. A new variable selection algorithm, named MPCA-SC, based on the combined use of spatial correlation analysis and MULTISPATI-PCA, was proposed. The potential use of 20 clustering algorithms for the generation of MZs was evaluated: average linkage, bagged clustering, centroid linkage, clustering large applications, complete linkage, divisive analysis, fuzzy analysis clustering (fanny), fuzzy c-means, fuzzy c-shells, hard competitive learning, hybrid hierarchical clustering, k-means, McQuitty's method (mcquitty), median linkage, neural gas, partitioning around medoids, single linkage, spherical k-means, unsupervised fuzzy competitive learning, and Ward's method. Two computational modules developed to provide the variable selection and data clustering methods for definition of MZs were also presented. The evaluations were conducted with data obtained between 2010 and 2015 in three commercial agricultural areas, cultivated with soybean and corn, in the state of Paraná, Brazil. The experiments performed to evaluate the 5 variable selection algorithms showed that the new method MPCA-SC can improve the quality of MZs in several aspects, even obtaining satisfactory results with the other 4 algorithms. The evaluation experiments of the 20 clustering methods showed that 17 of them were suitable for the delineation of MZs, especially fanny and mcquitty. Finally, it was concluded that the two computational modules developed made it possible to obtain quality MZs. Furthermore, these modules constitute a more complete computer system than other free-to-use software such as FuzME, MZA, and SDUM, in terms of the diversity of variable selection and data clustering algorithms.

Keywords: data clustering; MULTISPATI-PCA; precision agriculture; principal component analysis; software for agriculture.

SUMÁRIO

LISTA DE FIGURAS

REVISÃO BIBLIOGRÁFICA

ARTIGO 1

ARTIGO 2

ARTIGO 3

LISTA DE TABELAS

REVISÃO BIBLIOGRÁFICA

ARTIGO 1

ARTIGO 2

ARTIGO 3

LISTA DE ABREVIATURAS E SIGLAS

ACP	Análise de componentes principais
ANOVA	Análise de variância
AP	Agricultura de precisão
ASC	Average silhouette coefficient
CP	Componente principal
CPE	Componente principal espacial
CVI	Cluster validation index
Embrapa	Empresa Brasileira de Pesquisa Agropecuária
FPI	Fuzziness performance index
GPS	Global positioning system
GSC	Group silhouette coefficient
ICVI	Improved cluster validation index
MPE	Modified partition entropy
MULTISPATI-PCA	Multivariate spatial analysis based on Moran's index and PCA
MZA	Management zone analyst
PAM	Partitioning around medoids
RSP	Resistência mecânica do solo à penetração
SDUM	Software para Definição de Unidades de Manejo
SC	Silhouette coefficient
SI	Smoothness index
UFCL	Unsupervised fuzzy competitive learning
UTM	Universal transversa de mercator
VR	Variance reduction
ZM	Zona de manejo

1 INTRODUÇÃO

A agricultura de precisão (AP) corresponde a um conjunto de técnicas e tecnologias desenvolvidas para melhorar o gerenciamento do solo e de plantas cultivadas em áreas agrícolas. Ela possibilita monitorar, em nível local, a variabilidade da produção e das variáveis (atributos) que exercem influência sobre o desenvolvimento das plantas.

Trata-se de uma abordagem para possibilitar a aplicação de insumos como água, fertilizantes e defensivos em quantidades adequadas, nos locais e momentos corretos, em subáreas delimitadas de acordo com critérios de homogeneidade. Em relação a resultados, essa forma de gerenciamento pode conduzir ao aumento da produtividade e à redução de impactos ambientais decorrentes de aplicações de certos insumos em excesso.

Para que resultados satisfatórios possam ser alcançados, a AP recomenda o uso de técnicas e tecnologias como sistemas de navegação global por satélite, grades amostrais densas, monitoramento instantâneo de variáveis do solo e de plantas cultivadas, mapeamento das variabilidades espacial e temporal dessas variáveis, aplicação de insumos a taxas variáveis e gerenciamento de dados por meio do uso de sistemas de informação geográfica. Além disso, profissionais especializados também são necessários, a fim de utilizar os recursos disponíveis para produzir resultados que sejam interpretados corretamente.

Todavia, apesar dos benefícios que a aplicação da AP pode proporcionar, ela ainda é uma abordagem pouco empregada por pequenos produtores. De forma geral, isso se deve principalmente aos altos custos para sua implantação e manutenção.

Neste contexto, a divisão de áreas produtivas em zonas de manejo (ZMs) é uma prática utilizada com resultados satisfatórios para viabilizar a AP para mais produtores. Cada ZM é uma subárea que possui um conjunto de características similares e que, por isso, pode ser tratada como uma área homogênea sob o ponto de vista de amostragem e gerenciamento.

A implantação de ZMs no campo possibilita a redução da quantidade de amostras que precisam ser coletadas e analisadas, bem como o aproveitamento de máquinas e equipamentos convencionais – já que as aplicações de insumos em cada zona podem ocorrer a taxa constante. Essas aplicações a taxa fixa dentro de cada subárea, mas a taxas diferentes para ZMs diferentes, geralmente possibilitam reduzir a variabilidade espacial dos fatores relacionados à fertilidade do solo. Como consequência disso, tem-se a possibilidade de aumentar a produtividade das áreas agrícolas.

Métodos ou algoritmos de agrupamento de dados constituem uma abordagem fundamental para a definição de ZMs. Eles têm o propósito de dividir os pontos georreferenciados de uma área, aos quais estejam associados valores de variáveis de interesse, em certo número de classes. Para isso, aplicam algum critério para determinar o

nível de similaridade entre os pontos. Na prática, essas classes, também chamadas de grupos, são empregadas para delimitar as ZMs.

Há muitas opções de algoritmos de agrupamento de dados que podem ser avaliados para a geração de ZMs. No entanto, poucos têm sido efetivamente testados e aplicados nesse contexto, com destaque para os métodos k-means (MACQUEEN, 1967) e fuzzy c-means (BEZDEK, 1981).

Os métodos de agrupamento podem empregar muitas variáveis para a criação de ZMs, que representem condições do solo, do relevo e/ou de plantas cultivadas. Contudo, a seleção das variáveis realmente necessárias para a definição das subáreas é uma das tarefas mais difíceis na análise de agrupamento.

Bazzi et al. (2013) descreveram a aplicação de um método de seleção de variáveis que pode ser usado com algoritmos de agrupamento, baseado na análise de correlação espacial entre variáveis. Cohen et al. (2013) e Moral, Terrón e Silva (2010) mostraram que o método denominado análise de componentes principais (ACP) (HOTELLING, 1933) pode ser útil para produzir novas variáveis a serem consideradas na definição de ZMs. Já Córdoba et al. (2013) e Peralta et al. (2015) mostraram que o método de análise multivariada baseada no índice de Moran e em ACP, denominado MULTISPATI-PCA (DRAY; SAID; DÉBIAS, 2008), também pode ser empregado com resultados satisfatórios para essa tarefa.

A seleção de variáveis e a geração de classes para o delineamento de ZMs por meio da aplicação das abordagens citadas são tarefas complexas, que dependem da utilização de ferramentas computacionais adequadas. Entretanto, as ferramentas desenvolvidas para a criação de ZMs disponibilizam poucos algoritmos para a execução dessas duas tarefas. É o caso de softwares relevantes, como FuzME (MINASNY; MCBRATNEY, 2002), Management Zone Analyst (MZA) (FRIDGEN et al., 2004) e Software para Definição de Unidades de Manejo (SDUM) (BAZZI et al., 2013).

Diante dessa constatação, considerou-se importante pesquisar e avaliar métodos de seleção de variáveis e de agrupamento de dados que não estavam disponíveis em softwares desenvolvidos para a geração de ZMs. Também se considerou relevante implementar computacionalmente os métodos avaliados como úteis e disponibilizar essas implementações à comunidade de AP.

Esta tese está organizada na forma de artigos científicos. Em razão disso, inicialmente apresenta-se no Capítulo 3 uma revisão bibliográfica geral, com o propósito de servir de fundamentação teórica abrangente para os três artigos que foram incluídos. No Capítulo 5, apresenta-se o primeiro artigo elaborado como parte do projeto de doutorado. Nesse artigo, propôs-se um novo método para seleção de variáveis para a geração de ZMs e avaliou-se a eficiência desse novo método e de mais quatro algoritmos destinados a essa finalidade. No Capítulo 6, apresenta-se o segundo artigo, no qual avaliou-se a possibilidade de utilização de 20 algoritmos de agrupamento para a definição de ZMs. O Capítulo 7

corresponde ao terceiro artigo, que teve o objetivo de apresentar dois módulos computacionais desenvolvidos para possibilitar a execução eficiente dos métodos de seleção de variáveis e de agrupamento de dados previamente selecionados. Por fim, no Capítulo 8 apresentam-se as conclusões gerais do trabalho desenvolvido e atividades a serem realizadas como trabalhos futuros.

2 OBJETIVOS

2.1 Objetivo geral

Avaliar métodos de seleção de variáveis e de agrupamento de dados para a definição de zonas de manejo e implementar os métodos considerados úteis em dois módulos computacionais.

2.2 Objetivos específicos

- Propor um método de seleção de variáveis para a definição de ZMs que seja fundamentado na aplicação conjunta de análise de correlação espacial e MULTISPATI-PCA;
- Avaliar de forma comparativa 5 métodos de seleção de variáveis, baseados em análise de correlação espacial e/ou análise multivariada (incluindo o método proposto), para a geração de ZMs;
- Avaliar a utilização de 20 algoritmos de agrupamento, selecionados por meio de mapeamento sistemático da literatura, para o delineamento de ZMs: average linkage, bagged clustering, centroid linkage, clustering large applications, complete linkage, divisive analysis, fuzzy analysis clustering, fuzzy c-means, fuzzy c-shells, hard competitive learning, hybrid hierarchical clustering, k-means, median linkage, método de McQuitty, método de Ward, neural gas, partitioning around medoids, single linkage, spherical k-means e unsupervised fuzzy competitive learning;
- Implementar e disponibilizar um módulo computacional para seleção de variáveis, que contenha os 5 métodos comparados;
- Implementar e disponibilizar um módulo computacional para geração de classes destinadas ao delineamento de ZMs que contenha os algoritmos de agrupamento considerados úteis dentre os 20 supracitados.

3 REVISÃO BIBLIOGRÁFICA

3.1 Variáveis do solo, produtividade e variabilidade

De acordo com Guedes et al. (2012) e Lima (2010), as variáveis (também chamadas de atributos) físicas e químicas do solo, dentre outros fatores, exercem influência sobre a produtividade de áreas agrícolas. Essas variáveis sofrem variabilidades espacial e temporal devido a interações complexas dos fatores e processos de formação do solo. Além disso, práticas de manejo são causadoras adicionais de variabilidade, podendo modificar características do solo, principalmente em suas camadas superficiais.

De acordo com Beutler et al. (2012), as variáveis físicas do solo estão relacionadas à capacidade de infiltração, retenção e disponibilização de água para as plantas, além da facilidade de circulação do ar e de penetração de raízes. Variáveis como macro e microporosidade, porosidade total e densidade do solo são úteis para analisar esses aspectos, além de atuarem como indicadores da existência de problemas de compactação do solo, que podem afetar o desenvolvimento de raízes.

Carmo et al. (2011) afirmam que duas fontes importantes de variabilidade espacial dos atributos físicos do solo são o tipo de manejo empregado e a taxa de utilização do solo, que normalmente resultam em impactos na produtividade. Devido a isso, estes dois fatores devem ser analisados e gerenciados de forma específica para cada área agrícola, conforme os valores de suas variáveis físicas.

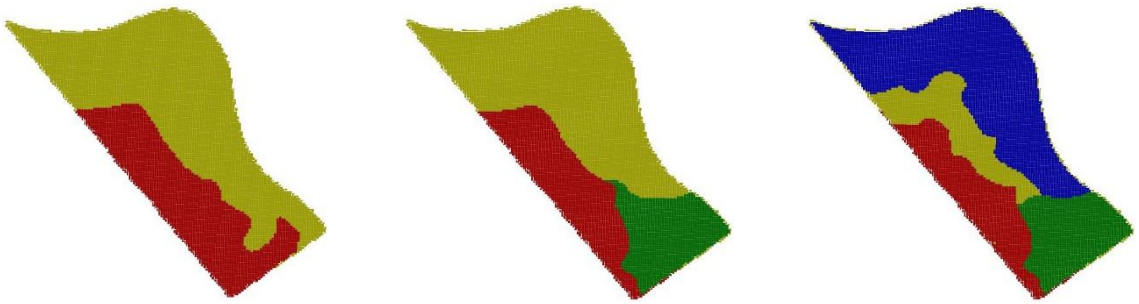
Segundo Klein et al. (2010), a estrutura do solo e sua textura são fatores relevantes no que diz respeito à retenção e à disponibilização de água para as plantas. A estrutura do solo geralmente é analisada com base em sua densidade e sua porosidade. Já a textura é avaliada com base na proporção de tamanho das partículas minerais constituintes do solo. Pode-se empregar análise granulométrica para classificar os componentes sólidos, conforme seus respectivos diâmetros, em argila, silte e areia.

Carmo et al. (2011) destacam que o estudo das variabilidades espacial e temporal das variáveis químicas e físicas do solo é relevante porque estão relacionadas com a produtividade das áreas, com a definição de grades de amostragem, com a interpretação de resultados e com a definição de aplicações racionais de insumos.

3.2 Geração de zonas de manejo

Segundo McBratney et al. (2005), as variáveis que em geral devem ser empregadas na geração de zonas de manejo (ZMs) são a produtividade de plantas cultivadas, dados químicos e físicos do solo, condutividade elétrica aparente do solo, dados topográficos da área, índices de vegetação ou combinações entre partes desses dados.

Para representar graficamente ZMs, costumam-se utilizar mapas temáticos. Além de serem representações de fácil compreensão, existem diversos softwares disponíveis para gerá-los. A Figura 1 exibe mapas temáticos que contêm duas, três e quatro zonas referentes a uma área agrícola, definidas a partir de uma combinação de variáveis avaliadas nessa área.



2 ZMs 3 ZMs 4 ZMs

Figura 1 Exemplos de mapas temáticos contendo duas, três e quatro zonas de manejo.

O processo de definição de ZMs é composto pela execução sucessiva das atividades representadas na Figura 2. Essas atividades são abordadas nas seções seguintes, com destaque para os métodos de seleção de variáveis e de agrupamento de dados.

Figura 2 Diagrama com as atividades geralmente executadas para a definição de zonas de manejo.

Mesmo sabendo que a utilização de ZMs não garante a obtenção dos melhores resultados, em muitos trabalhos relatam-se resultados satisfatórios com diferentes plantas cultivadas e tipos de solo, considerando diversas variáveis para a definição das subáreas. Arno et al. (2011), Bazzi et al. (2013), Cid-Garcia, Bravo-Lozano e Rios-Solis (2014), Córdoba et al. (2013), Ferraz et al. (2011), Peralta et al. (2015), Schenatto et al. (2016) e Tripathi et al. (2015) apresentaram alguns exemplos disso.

3.3 Métodos de seleção de variáveis

Conforme ilustrado na Figura 2, o processo de delineamento de ZMs pode ser realizado após a geração de classes, que ocorre por meio da aplicação de um método de agrupamento de dados. Por sua vez, a execução do algoritmo de agrupamento geralmente depende da execução prévia de um algoritmo de seleção das variáveis necessárias para esse processo. Nas próximas subseções, apresentam-se três abordagens de seleção de variáveis utilizadas na literatura em conjunto com algoritmos de agrupamento para a definição de ZMs.

3.3.1 Análise de correlação espacial

É um método fundamentado na estatística de autocorrelação espacial bivariada de Moran (CZAPLEWSKI; REICH, 1993), para analisar se amostras apresentam correlação espacial (autocorrelação) e se este tipo de correlação é significativo entre pares de variáveis. Para duas variáveis *X* e *Y* avaliadas nos mesmos pontos amostrais, pode-se calcular a correlação espacial bivariada de Moran (I_{XY}) por meio do uso da Equação 1 (REICH; CZAPLEWSKI; BECHTOLD, 1994):

$$I_{XY} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} * X_i * Y_j}{W\sqrt{m_X{}^2 * m_Y{}^2}} \qquad \text{Eq. (1)}$$

em que: W_{ij} é uma matriz de associação espacial, calculada por $W_{ij} = (1/(1+D_{ij}))$, sendo D_{ij} a distância entre os pontos *i* e *j*; X_i é o valor da variável X padronizada, no ponto *i*; Y_j é o valor da variável Y padronizada, no ponto *j*; W corresponde à soma dos graus de associação espacial, obtidos da matriz W_{ij}, para *i*≠*j*; $m_X{}^2$ corresponde à variância amostral de X; e $m_Y{}^2$ corresponde à variância amostral de Y. Há correlação espacial positiva se $I_{XY} > 0$, correlação negativa se $I_{XY} < 0$, ou não há correlação se $I_{XY} = 0$.

Conforme recomendação de Reich, Czaplewski e Bechtold (1994), deve-se interpretar a padronização de uma variável V como o procedimento executado sobre seus valores para que ela fique com média igual a 0. Para isso, aplica-se a equação $Z_i = (V_i - \bar{V})$,

em que V_i é o valor original da variável no ponto *i*, Z_i é o valor padronizado de V_i, e $\bar{V}$ representa a média de V.

Após a computação da autocorrelação para as amostras de cada variável e da correlação espacial para cada possível par de variáveis, pode-se então montar uma matriz de correlação espacial. Esta deve exibir todos os valores de I_{XY} calculados, bem como os valores de autocorrelação (I_{XX}). Assim, essa matriz poderá ser empregada para a identificação das variáveis necessárias para o delineamento de ZMs, de acordo com o seguinte procedimento definido por Bazzi et al. (2013):

1. Eliminar as variáveis com autocorrelação espacial não significativa a um dado nível de significância;
2. Remover as variáveis que não possuírem correlação espacial significativa com a produtividade;
3. Ordenar de modo decrescente as variáveis restantes, considerando o módulo do valor da correlação com a produtividade;
4. Eliminar as variáveis redundantes (que se correlacionem entre si), dando preferência para a exclusão das que possuírem menor correlação com a produtividade;
5. Finalmente, as variáveis restantes poderão ser utilizadas por um algoritmo de agrupamento para a definição de ZMs.

3.3.2 Análise de componentes principais

De acordo com Ferreira (1996) e Johnson e Wichern (2007), os métodos de análise multivariada foram desenvolvidos para resumir, representar e interpretar dados amostrados a partir de populações nas quais, em cada unidade experimental, estudam-se diversas variáveis. Este estudo simultâneo de várias variáveis é adequado quando se tem indícios de que nenhuma delas seja capaz de caracterizar, sozinha, cada unidade.

Segundo Jolliffe (2002), a análise de componentes principais (ACP) é um método de análise multivariada que tem como principal objetivo reduzir a dimensão de análise de conjuntos de dados associados a variáveis quantitativas correlacionadas entre si. Esse método permite identificar as variáveis que explicam a maior parte da variância total presente em conjuntos de dados. Além disso, pode explicitar relacionamentos existentes e eventualmente desconhecidos entre variáveis ou unidades experimentais.

Para utilizar a ACP, devem-se executar transformações a partir das variáveis originais, que resultem em um novo conjunto de variáveis sintéticas denominadas componentes principais (CPs). As CPs, que não apresentam correlação linear entre si, devem ser ordenadas de modo que as primeiras componentes retenham a maioria da variabilidade

presente no conjunto original. Isto geralmente permite notar que algumas das variáveis são menos importantes para um estudo (JOLLIFFE, 2002).

Para definir formalmente o conceito de CP, considere que se deseja estudar uma população por meio de *p* variáveis quantitativas $X_1, X_2, ..., X_p$, possivelmente com unidades de medida e escalas diferentes, avaliadas em *n* unidades amostrais independentes. Considere ainda que elas apresentam elevado grau de correlação entre si e que não dependam da suposição de distribuição normal dos dados. Logo, a partir dessas variáveis originais produzem-se *p* componentes principais $Y_1, Y_2, ..., Y_p$, denotadas conforme mostrado na Equação 2 (FERREIRA, 1996):

$$Y_k = e_{k1}X_1 + e_{k2}X_2 + \cdots + e_{kp}X_p \qquad k = 1, \ldots, p \qquad \text{Eq. (2)}$$

em que: e_{kj}, *j = 1, ..., p*, são números reais chamados de coeficientes da componente Y_k.

Observa-se na Equação 2 que cada componente é definida como uma combinação linear das *p* variáveis originais. Além disso, cada valor numérico e_{kj} indica o nível de contribuição de cada variável original no valor de Y_k. A origem dos coeficientes e_{kj} está relacionada ao vetor $\boldsymbol{X}$ das *p* variáveis aleatórias originais $X_1, X_2, ..., X_p$, representado na Equação 3 (FERREIRA, 1996).

$$\boldsymbol{X} = \begin{bmatrix} X_1 & X_2 & \ldots & X_p \end{bmatrix}^t \qquad \text{Eq. (3)}$$

O vetor $\boldsymbol{X}$ possui vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, de modo que $\boldsymbol{\Sigma}$ possui $\lambda_1, \lambda_2, ..., \lambda_p$ autovalores ordenados de forma decrescente e seus respectivos autovetores $\boldsymbol{e_1, e_2, ..., e_p}$. Por conseguinte, conforme representado na Equação 4, os coeficientes e_{kj}, *j = 1, ..., p*, da componente Y_k, são os elementos do autovetor $\boldsymbol{e_k}$.

$$Y_k = \boldsymbol{e_k^t X} = e_{k1}X_1 + e_{k2}X_2 + \cdots + e_{kp}X_p \qquad k = 1, \ldots, p \qquad \text{Eq. (4)}$$

Para cada CP, as variáveis que têm maior relevância em sua construção são as que possuem os coeficientes com maiores valores. As variáveis com coeficientes positivos exercem influência direta sobre a componente, enquanto que aquelas com coeficientes negativos exercem influência inversa. Segundo Johnson e Wichern (2007), as CPs possuem algumas propriedades importantes:

- O somatório das variâncias das *p* componentes sempre é igual ao somatório das variâncias das *p* variáveis originais;
- A componente Y_1 sintetiza o máximo possível da variância total das *p* variáveis originais; a componente Y_2 sintetiza o máximo possível da variância total residual dessas *p* variáveis, e assim sucessivamente até Y_p;

- A variância associada a qualquer componente Y_k é igual ao seu respectivo autovalor λ_k;
- A covariância $Cov(Y_i, Y_j) = 0$, para *i, j = 1, ..., p, i ≠ j*.

No caso de se ter uma amostra multivariada contendo *n* elementos e *p* variáveis, as definições anteriores continuam sendo válidas. Porém, deve-se utilizar a matriz de covariâncias ***S*** da amostra no lugar da matriz $\boldsymbol{\Sigma}$, que corresponde a uma população.

Contudo, antes da construção das CPs, deve-se considerar que geralmente as escalas de medidas das variáveis originais são diferentes. Neste caso, é necessário efetuar um pré-processamento dos dados, utilizando-se uma função de padronização dos valores de cada variável X como, por exemplo, a que é mostrada na Equação 5 (FERREIRA, 1996):

$$z_i = \frac{(X_i - \bar{X})}{\sqrt{s_{ii}}} \qquad \text{Eq. (5)}$$

em que: X_i é o valor da variável na localização *i*, Z_i é o valor padronizado de X_i, $\overline{X}$ é o valor da média de X, e s_{ii} é o valor da covariância da variável X com ela mesma (que se resume à variância de X).

A padronização de valores sobre as variáveis originais é equivalente a utilizar na ACP a matriz de coeficientes de correlação linear de Pearson ***R*** das variáveis originais no lugar da matriz $\boldsymbol{\Sigma}$ ou da matriz ***S***. Consequentemente, os autovalores e autovetores empregados para a construção das CPs serão, neste caso, obtidos a partir de ***R***. Assim, passa-se a definir as CPs conforme mostrado na Equação 6 (FERREIRA, 1996):

$$Y_k = e_{k1}Z_1 + e_{k2}Z_2 + \cdots + e_{kp}Z_p \qquad k = 1, \ldots, p \qquad \text{Eq. (6)}$$

Na prática, quando se aplica a ACP, é comum utilizar poucas componentes no lugar das *p* variáveis originais, sem perda significativa de informação. Alguns critérios utilizados para selecionar as CPs em uma análise são (FERREIRA, 1996; JOHNSON; WICHERN, 2007):

- Critério de Kaiser: devem ser consideradas as CPs que possuam respectivos autovalores maiores que 1,0;
- Critério apresentado em Ferreira (1996): devem ser consideradas as primeiras CPs que expliquem, juntas, ao menos 70% da variabilidade total das variáveis originais;
- Critério de Johnson e Wichern (2007): devem ser consideradas as primeiras CPs que expliquem, juntas, ao menos 80% da variabilidade total das variáveis originais;
- Procedimento de Horn (1965): baseado na comparação dos autovalores dos dados originais com os autovalores das CPs geradas por dados randômicos ou realocados;

- Procedimento de Cattell (1966): baseado na construção de um gráfico denominado scree plot, a partir do qual se define a quantidade de CPs necessárias para a análise.

Ao selecionar determinada quantidade de CPs por um ou mais desses critérios, pode-se calcular o valor de cada componente para cada observação amostral ou populacional, chamado de escore.

No processo de definição de ZMs, a ACP tem sido aplicada com sucesso para dois propósitos. Um deles é definir as CPs necessárias para serem empregadas diretamente na geração de ZMs, como fizeram, por exemplo, Li et al. (2013), Molin e Castro (2008), Moral, Terrón e Silva (2010) e Tripathi et al. (2015). O outro é identificar as variáveis que explicam a maioria da variância dos dados e usá-las na definição das subáreas, como optaram, por exemplo, Fraisse, Sudduth e Kitchen (2001) e Saleh e Belal (2014).

Li et al. (2013) utilizaram ACP sobre um conjunto de dados referente a sete variáveis químicas e físicas avaliadas em uma região da China. Eles constataram que as duas primeiras CPs eram suficientes para representar mais de 85% da variância original dos dados, o que os levou a aplicá-las na definição de ZMs com o uso conjunto de um método de agrupamento.

Moral, Terrón e Silva (2010) realizaram um experimento seguindo essa mesma metodologia em uma área agrícola de 33 ha localizada na Espanha. Os resultados obtidos e as conclusões foram similares às de Li et al. (2013).

Molin e Castro (2008) avaliaram o uso de ACP para seleção de variáveis e um algoritmo de agrupamento para a geração de ZMs, o que possibilitou substituir a análise da condutividade elétrica e de mais onze variáveis químicas e físicas do solo por duas CPs. A área agrícola experimental, localizada no estado do Paraná, foi dividida em três zonas, que representaram de modo satisfatório a variabilidade espacial da produtividade dessa área.

Tripathi et al. (2015) aplicaram ACP sobre um conjunto de dados correspondente a dez variáveis químicas e físicas medidas em 225 pontos amostrais de uma região da Índia. Eles consideraram suficiente usar as três primeiras CPs no lugar das variáveis originais para o delineamento de ZMs. As CPs selecionadas foram empregadas como entrada para um algoritmo de agrupamento, resultando no número ideal de três ZMs para a área.

Fraisse, Sudduth e Kitchen (2001) empregaram ACP para determinar quais de seis variáveis eram mais relevantes para o delineamento de ZMs por agrupamento. Foram consideradas duas áreas agrícolas, localizadas nos Estados Unidos da América. A aplicação de ACP evidenciou que duas das variáveis originais bastavam para representar quase a totalidade da variância dos dados. Diante disso, foram geradas subáreas e a análise de variância da produtividade de grãos nessas subáreas validou a abordagem seguida.

Saleh e Belal (2014) executaram um experimento em área agrícola no Egito, aplicando a mesma metodologia e as mesmas variáveis empregadas por Fraisse, Sudduth e Kitchen (2001). A ACP viabilizou a redução da quantidade de variáveis originais necessárias

para três e estas foram então utilizadas para a definição de ZMs com um algoritmo de agrupamento.

3.3.3 MULTISPATI-PCA

Segundo Dray, Said e Débias (2008), a ACP e os demais métodos tradicionais de análise multivariada não consideram possíveis relacionamentos espaciais entre os elementos do conjunto de dados, como, por exemplo, no caso de pontos georreferenciados. O motivo é que esses métodos não foram projetados para identificar tais relacionamentos. Diante disso, esses pesquisadores propuseram a abordagem denominada análise espacial multivariada baseada no índice de Moran (multivariate spatial analysis based on Moran's index - MULTISPATI), que introduziu uma restrição espacial aos métodos multivariados clássicos.

Com a abordagem MULTISPATI pode-se, por exemplo, executar a ACP considerando a existência de dependência espacial em um conjunto de dados que contenha os valores de *p* variáveis observadas em *n* pontos georreferenciados. O método resultante da aplicação de MULTISPATI combinada com ACP recebeu o nome de multivariate spatial analysis based on Moran's index and PCA (MULTISPATI-PCA) (CÓRDOBA et al., 2012).

Este método baseia-se na utilização de uma matriz de ponderação espacial $\boldsymbol{W}_{nxn}$, que é uma matriz de conectividade padronizada definida por Dray, Said e Débias (2008). Trata-se de uma matriz similar à matriz $\boldsymbol{W}$ citada para o cálculo da correlação espacial bivariada de Moran (Equação 1). Ela pode ser vista como uma representação matemática da distribuição geográfica dos pontos sob estudo, pois os pesos espaciais simbolizam a ausência ($w_{ij} = 0$) ou a intensidade ($w_{ij} > 0$) das relações espaciais entre os pontos dentro da área considerada (CÓRDOBA et al., 2012).

Assim, enquanto o método ACP tradicional é aplicado sobre uma matriz $\boldsymbol{X}_{nxp}$ chamada de tabela de dados, MULTISPATI-PCA é executado em dois passos: primeiro, constrói-se uma nova tabela de dados $\boldsymbol{Y} = \boldsymbol{WX}$ ($\boldsymbol{Y}$ também apresenta dimensão *nxp*), e então utiliza-se essa nova tabela como entrada para a ACP. Cada elemento desta nova tabela $\boldsymbol{Y}$ corresponde a um novo valor que substitui, como entrada para a ACP, o valor da mesma posição na matriz $\boldsymbol{X}$, correspondente ao valor de uma variável *p* em um ponto amostral *i*.

Diferente da ACP, a variância associada a cada componente gerada por meio da aplicação de MULTISPATI-PCA não é igual ao seu respectivo autovalor. Outra diferença importante é que a quantidade de CPs válidas geradas utilizando-se MULTISPATI-PCA pode ser menor que a quantidade de variáveis originais consideradas. De acordo com Arrouays et al. (2011), MULTISPATI-PCA tem uma vantagem importante em relação à ACP: seus escores maximizam a autocorrelação espacial entre pontos, enquanto os obtidos com ACP maximizam a variância total. Logo, os escores gerados com MULTISPATI-PCA mostram estruturas

espaciais fortes nas primeiras CPs, enquanto os escores da ACP podem mostrar estruturas espaciais em quaisquer componentes, até mesmo nas últimas, que na prática costumam ser desconsideradas.

Córdoba et al. (2013) demonstraram a aplicação de MULTISPATI-PCA em um comparativo com ACP, considerando dados georreferenciados multivariados de três áreas agrícolas localizadas na Argentina. Nesse comparativo, destacaram que MULTISPATI-PCA facilitou a seleção da quantidade de CPs necessárias para a geração de ZMs, além de ter identificado um relacionamento entre duas variáveis do solo que a ACP não identificou. Os autores também mostraram que MULTISPATI-PCA melhorou o desempenho do algoritmo de agrupamento utilizado na definição das ZMs.

Peralta et al. (2015) empregaram MULTISPATI-PCA com um método de agrupamento para definir ZMs para cinco áreas também localizadas na Argentina. Apesar de terem estabelecido objetivos diferentes daqueles de Córdoba et al. (2013), as conclusões quanto à obtenção de ZMs adequadas após a aplicação de MULTISPATI-PCA foram similares nesses dois trabalhos.

3.4 Interpolação espacial de dados

Os métodos de interpolação espacial de dados são empregados para estimar valores de variáveis para locais em que estas não foram medidas. Para isso, utilizam-se dados de pontos amostrados na mesma área. De acordo com Miranda (2010), o princípio básico da interpolação espacial é que valores de uma variável tendem, em média, a ser similares em locais mais próximos do que em locais mais distantes.

Deste modo, dados pontuais são transformados em campos contínuos com padrões espaciais que podem ser comparados a outros elementos espaciais contínuos. A interpolação espacial é necessária quando se tem poucos pontos amostrais em uma área e se deseja definir ZMs com delimitações precisas (MOLIN; FAULIN, 2013). Segundo Mazzini e Schettini (2009) e Schenatto et al. (2016), os métodos de interpolação espacial inverso da distância, inverso da distância ao quadrado e krigagem são frequentemente aplicados no processo de delineamento de ZMs.

3.5 Métodos de agrupamento de dados

Dados resultantes da interpolação espacial de variáveis frequentemente constituem a entrada de métodos de agrupamento utilizados para a definição de ZMs. Esses métodos incluem alguma função de distância (também chamada de função de similaridade) para determinar o nível de similaridade entre os valores de variáveis correspondentes a pontos resultantes da interpolação, a fim de dividi-los em uma determinada quantidade de classes.

Apesar dos algoritmos de agrupamento k-means e fuzzy c-means serem os mais empregados para a definição de ZMs, Delalibera, Weirich e Nagata (2012), Dobermann et al. (2003), Galambosová et al. (2014), Guastaferro et al. (2010), Ortega e Santibáñez (2007) e Tichý, Chytrý e Botta-Dukát (2014) mostraram que outros métodos de agrupamento podem conduzir ao delineamento de ZMs de qualidade, eventualmente até melhores que as geradas por k-means ou fuzzy c-means.

De acordo com Xu e Wunsch (2009), as duas principais desvantagens desses dois algoritmos são: o fato da definição inicial dos grupos ser aleatória, o que pode gerar elevado tempo de processamento até o algoritmo convergir para um agrupamento, além do risco desse agrupamento não ser satisfatório; e a robustez limitada, devido à sensibilidade em relação a valores discrepantes, o que também pode resultar em grupos distantes do ideal.

Diante disso, apresentam-se nas próximas subseções 20 métodos de agrupamento hierárquico e de particionamento que podem ser avaliados para a definição de ZMs (Tabela 1). Eles foram selecionados por meio de um mapeamento sistemático da literatura, com o intuito de avaliar métodos de diferentes paradigmas de agrupamento.

Tabela 1 Métodos de agrupamento que podem ser avaliados para a geração de zonas de manejo

Categoria	Algoritmos
Métodos hierárquicos	average linkage
	centroid linkage
	complete linkage
	median linkage
	método de McQuitty
	método de Ward
	single linkage
	divisive analysis
	hybrid hierarchical clustering
Métodos de particionamento	k-means
	fuzzy c-means
	bagged clustering
	fuzzy analysis clustering
	fuzzy c-shells
	hard competitive learning
	neural gas
	partitioning around medoids
	clustering large applications
	spherical k-means
	unsupervised fuzzy competitive learning

3.5.1 Métodos de agrupamento hierárquico

Os métodos de agrupamento hierárquico dividem os elementos de um conjunto de dados (como, por exemplo, pontos georreferenciados) em uma quantidade específica de grupos, em dois ou mais passos. Eles realizam o agrupamento por meio da definição de uma série de divisões aninhadas, podendo iniciar com um grupo contendo todos os *n* elementos para resultar na formação de *n* grupos com um elemento em cada, ou vice-versa. O primeiro

é chamado de agrupamento hierárquico divisivo, e se o processo for realizado no sentido inverso é chamado de aglomerativo. Mas ambos organizam os elementos em uma estrutura hierárquica, definida a partir de uma matriz de proximidade (EVERITT; LANDAU; LEESE, 2011; JAIN; DUBES, 1988).

Frequentemente, utiliza-se uma estrutura de árvore denominada dendrograma para representar graficamente os resultados de agrupamentos hierárquicos. Para exemplificar, exibe-se na Figura 3 um dendrograma correspondente ao agrupamento de sete elementos O_1, O_2, O_3, O_4, O_5, O_6 e O_7. Pode-se realizar um procedimento chamado de corte horizontal no dendrograma, a fim de obter uma quantidade desejada de grupos. Isto é ilustrado pela linha pontilhada na Figura 3, que neste exemplo evidencia que o corte resultou em dois grupos: um contendo os elementos O_1, O_2, O_3 e O_4, e o outro contendo O_5, O_6 e O_7.

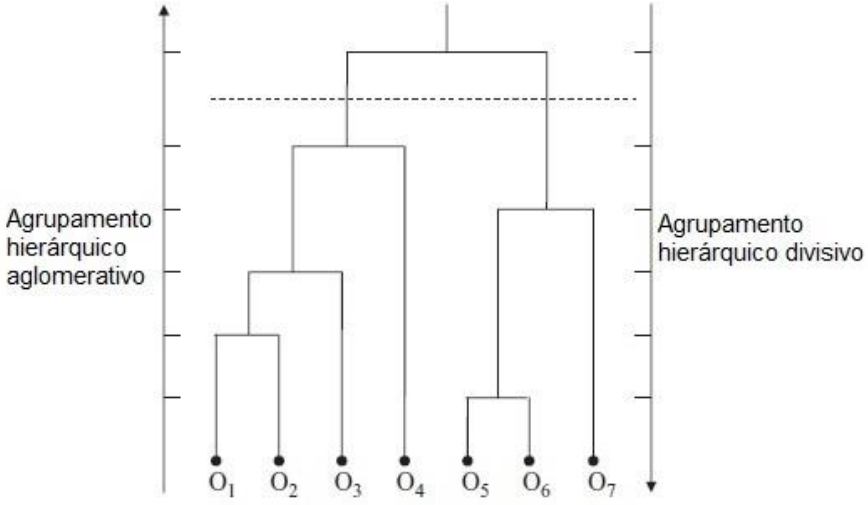


Figura 3 Exemplo de dendrograma de agrupamento hierárquico, em que a linha pontilhada de corte horizontal resulta em dois grupos.

Em um dendrograma, o nó raiz (posição no nível mais alto da árvore) representa o conjunto de dados inteiro, e cada nó folha (posição no nível mais baixo da árvore) representa um elemento do conjunto. Assim, os nós intermediários descrevem quais elementos estão próximos de cada um dos demais, e a distância entre cada par de grupos costuma ser representada pela altura do dendrograma.

Segundo Jain, Murty e Flynn (1999), os métodos hierárquicos aglomerativos são utilizados com mais frequência que os divisivos, pelo fato de serem executados com menor exigência de processamento computacional. Um agrupamento aglomerativo é iniciado com *n* grupos, em que cada um destes contém um elemento. Em seguida, realiza-se uma sequência de operações de combinação entre pares de grupos, que pode terminar com todos os elementos inseridos no mesmo grupo. Esse processo pode ser generalizado na seguinte sequência de passos:

1. Iniciar com *n* grupos unitários; utilizando uma função de distância, calcular a matriz de proximidade para os *n* grupos;
2. Na matriz de proximidade, buscar a distância mínima $D(G_i, G_j) = min(D(G_x, G_y))$, para $1 \leq (x, y) \leq n, x \neq y$, em que *D* representa a função de distância empregada para combinar os grupos G_i e G_j a fim de formar um novo grupo G_{ij};
3. Atualizar a matriz de proximidade por meio do cálculo das distâncias entre o grupo G_{ij} e os demais grupos;
4. Repetir os passos 2 e 3, até restar apenas um grupo que contenha todos os *n* elementos.

O passo 2 desse procedimento evidencia que a definição da função de distância exerce influência sobre a combinação de dois grupos quaisquer, para formar um novo grupo. A fórmula proposta por Lance e Williams (1967) (Equação 7) foi aplicada para generalizar uma grande quantidade de definições de distância entre um grupo G_x e um novo grupo G_{ij}, resultante da combinação de G_i e G_j:

$$D\left(G_x, \left(G_i, G_j\right)\right) = \alpha_i D(G_x, G_i) + \alpha_j D\left(G_x, G_j\right) + \beta D\left(G_i, G_j\right) + \gamma \left| D(G_x, G_i) - D\left(G_x, G_j\right) \right| \quad \text{Eq. (7)}$$

em que: *D* é a função de distância, e α_i, α_j, β e γ são coeficientes que têm seus valores determinados de acordo com o algoritmo aplicado. Os valores destes parâmetros para os métodos hierárquicos aglomerativos mais utilizados são mostrados na Tabela 2 (EVERITT; LANDAU; LEESE, 2011).

Tabela 2 Valores dos parâmetros propostos por Lance e Williams (1967), para métodos de agrupamento hierárquico aglomerativos comumente empregados

Algoritmos	α_i	α_j	β	γ
average linkage	$n_i / (n_i+n_j)$	$n_j / (n_i+n_j)$	0	0
centroid linkage	$n_i / (n_i+n_j)$	$n_j / (n_i+n_j)$	$-n_in_j / (n_i+n_j)^2$	0
complete linkage	1/2	1/2	0	1/2
median linkage	1/2	1/2	-1/4	0
método de McQuitty	1/2	1/2	0	0
método de Ward	$(n_i+n_x) / (n_i+n_j+n_x)$	$(n_j+n_x) / (n_i+n_j+n_x)$	$-n_x / (n_i+n_j+n_x)$	0
single linkage	1/2	1/2	0	-1/2

n_k: número de elementos no grupo *k*.

3.5.1.1 Average linkage

Para o algoritmo average linkage (JAIN; DUBES, 1988), a distância entre dois grupos é definida como uma média ponderada abrangendo as distâncias entre um grupo G_x e os grupos G_i e G_j, previamente combinados para formar o novo grupo G_{ij}. A ponderação é baseada na quantidade de elementos em cada grupo (n_i e n_j), tal que a Equação 7 é reescrita e resulta na Equação 8 (XU; WUNSCH, 2009).

$$D\left(G_x, \left(G_i, G_j\right)\right) = \frac{n_i}{n_i + n_j} D(G_x, G_i) + \frac{n_j}{n_i + n_j} D\left(G_x, G_j\right) \quad \text{Eq. (8)}$$

3.5.1.2 Centroid linkage

No algoritmo centroid linkage (JAIN; DUBES, 1988), a distância entre o grupo G_x e o novo grupo G_{ij} é determinada com base na distância entre seus centros geométricos (centroides). Assim, a Equação 7 passa a ser escrita da forma mostrada na Equação 9 (XU; WUNSCH, 2009).

$$D\left(G_x,\left(G_i,G_j\right)\right)=\frac{n_i}{n_i+n_j}D(G_x,G_i)+\frac{n_j}{n_i+n_j}D\left(G_x,G_j\right)-\frac{n_in_j}{\left(n_i+n_j\right)^2}D\left(G_i,G_j\right) \quad \text{Eq. (9)}$$

3.5.1.3 Complete linkage

No caso do algoritmo complete linkage (JAIN; DUBES, 1988), a distância entre dois grupos é determinada como a maior distância entre um elemento de um grupo e um elemento do outro grupo. Portanto, a Equação 7 torna-se, neste caso, a Equação 10 (XU; WUNSCH, 2009).

$$D\left(G_x,\left(G_i,G_j\right)\right)=max\left(D(G_x,G_i),D\left(G_x,G_j\right)\right) \quad \text{Eq. (10)}$$

3.5.1.4 Median linkage

O algoritmo median linkage (JAIN; DUBES, 1988) realiza o agrupamento de modo similar a centroid linkage. A única diferença é que, no caso de median linkage, a mesma ponderação é dada aos grupos G_i e G_j que foram combinados. Neste sentido, reescreve-se a Equação 7 para se obter a Equação 11 para este algoritmo (XU; WUNSCH, 2009).

$$D\left(G_x,\left(G_i,G_j\right)\right)=\frac{1}{2}D(G_x,G_i)+\frac{1}{2}D\left(G_x,G_j\right)-\frac{1}{4}D\left(G_i,G_j\right) \quad \text{Eq. (11)}$$

3.5.1.5 Método de McQuitty

Para o método de McQuitty (MCQUITTY, 1966), a distância entre dois grupos é estabelecida como a média dos valores da distância entre todos os pares de elementos desses grupos (tal que os elementos sejam de grupos diferentes). Desta forma, a distância entre o novo grupo G_{ij} e um grupo G_x é a média das distâncias $D(G_x, G_i)$ e $D(G_x, G_j)$ (Equação 12) (XU; WUNSCH, 2009).

$$D\left(G_x,\left(G_i,G_j\right)\right)=\frac{1}{2}\left(D(G_x,G_i)+D\left(G_x,G_j\right)\right) \quad \text{Eq. (12)}$$

3.5.1.6 Método de Ward

No método de Ward, também chamado de método da variância mínima (WARD, 1963), o objetivo é minimizar dentro do grupo G_{ij} resultante da combinação de G_i e G_j, a soma

dos quadrados das distâncias entre os elementos e o centroide (chamada de erro) (Equação 13) (XU; WUNSCH, 2009).

$$E_{ij} = \frac{n_i n_j}{n_i + n_j} \left\| \boldsymbol{m}_i - \boldsymbol{m}_j \right\|^2 \quad \text{Eq. (13)}$$

em que: $\left\| \boldsymbol{m}_i - \boldsymbol{m}_j \right\|$ é a distância euclidiana (BEZDEK, 1981) entre os centroides $\boldsymbol{m}_i$ e $\boldsymbol{m}_j$, dos grupos G_i e G_j, respectivamente.

Para o método de Ward, reescreve-se a Equação 7 para se obter a Equação 14 (XU; WUNSCH, 2009).

$$D\left(G_x, \left(G_i, G_j\right)\right) = \frac{n_i + n_x}{n_i + n_j + n_x} D(G_x, G_i) + \frac{n_j + n_x}{n_i + n_j + n_x} D\left(G_x, G_j\right) - \frac{n_x}{\left(n_i + n_j\right)^2} D\left(G_i, G_j\right) \quad \text{Eq. (14)}$$

3.5.1.7 Single linkage

Por fim, o algoritmo single linkage (JAIN; DUBES, 1988) funciona de forma oposta a complete linkage, pois a distância entre dois grupos é determinada como a menor distância entre um elemento de um grupo e um elemento do outro grupo. Portanto, a partir da Equação 7 obtém-se a Equação 15 (XU; WUNSCH, 2009).

$$D\left(G_x, \left(G_i, G_j\right)\right) = min\left(D(G_x, G_i), D\left(G_x, G_j\right)\right) \quad \text{Eq. (15)}$$

3.5.1.8 Divisive analysis

Proposto por Kaufman e Rousseeuw (1990), divisive analysis (diana) é um método hierárquico divisivo que tende a consumir mais tempo de processamento computacional que os métodos aglomerativos apresentados. Em cada etapa, esse algoritmo divide cada grupo existente em dois menores, até que finalmente todos os grupos contenham apenas um elemento cada. Portanto, executa-se a seguinte sequência de passos em cada etapa:

1. Para dividir os elementos de um grupo em dois, cria-se um novo grupo e coloca-se neste o elemento que apresentar a maior distância média em relação aos demais;
2. Calcula-se a distância média de cada elemento do grupo maior em relação a todos os demais deste grupo e também em relação aos elementos do novo grupo criado;
3. Considerando o menor valor obtido para a distância média, decide-se se cada elemento do grupo maior permanece neste ou se deve ser realocado ao novo grupo criado;
4. Após concluir essa divisão dos elementos entre os dois grupos, inicia-se uma nova etapa de divisões retornando ao passo 1.

Uma questão importante é como o algoritmo define qual grupo deve ser dividido na próxima etapa, quando há dois ou mais contendo mais de um elemento. O critério é que o

grupo que apresentar o maior valor para a distância entre dois de seus elementos é selecionado.

3.5.1.9 Hybrid hierarchical clustering

Foi proposto por Chipman e Tibshirani (2006) como um método de agrupamento híbrido, por combinar os pontos fortes dos algoritmos de agrupamento hierárquico aglomerativos e divisivos. Isto porque os algoritmos aglomerativos costumam ser adequados para a identificação de grupos com poucos elementos, mas não de grupos grandes. Por outro lado, os algoritmos divisivos geralmente são indicados para o oposto.

Assim, o método hybrid hierarchical clustering foi criado sobre a ideia de um agrupamento mútuo: um grupo de elementos mais próximos uns dos outros do que de quaisquer outros elementos do conjunto de dados. Em um grupo mútuo *G*, exige-se que a maior distância entre dois elementos seja menor do que a distância entre qualquer elemento de *G* e qualquer elemento que não pertença a *G*. Chipman e Tibshirani (2006) resumem o funcionamento do método hybrid hierarchical clustering em três passos:

1. Obter os grupos mútuos do conjunto de dados;
2. Executar um algoritmo hierárquico divisivo, mantendo cada grupo mútuo intacto;
3. Após concluir o agrupamento divisivo do conjunto inteiro, executar o agrupamento divisivo dentro de cada grupo mútuo, a fim de dividi-los.

No passo 3, pode-se substituir a utilização de um método divisivo por um método aglomerativo, sem que isso resulte em grandes mudanças no resultado final do agrupamento. Pode-se ainda utilizar um método aglomerativo nos passos 2 e 3. Ainda segundo os criadores desse algoritmo, hybrid hierarchical clustering evita a tendência dos algoritmos divisivos de realizarem divisões em grupos que deveriam ficar intactos. Consegue-se isso ao tratar cada grupo mútuo como um objeto indivisível.

3.5.1.10 Resultados com métodos hierárquicos

Dobermann et al. (2003) compararam o método de Ward e três métodos de particionamento para a definição de ZMs. Os experimentos ocorreram em duas áreas agrícolas com cultivo de soja e milho, com 62,7 e 60 ha e localizadas nos Estados Unidos da América. Os agrupamentos foram executados considerando-se a variável produtividade, medida no período de 1996 a 2001. Os autores mostraram que o método de Ward e um dos algoritmos de particionamento proporcionaram os melhores resultados, analisando-se diversas configurações dos dados de entrada.

Galambosová et al. (2014) também avaliaram o método de Ward e apresentaram resultados satisfatórios com a sua utilização para definir ZMs. Eles consideraram dados da

condutividade elétrica aparente do solo e da produtividade de cevada e trigo, coletados entre 2009 e 2011 em uma área experimental de 17 ha.

Ortega e Santibáñez (2007) relataram resultados satisfatórios com a utilização de análise hierárquica de agrupamento para geração de ZMs, considerando dados de seis variáveis químicas do solo avaliadas em 2002 e 2003. Os dados dessas variáveis e também da produtividade de milho foram coletados em 13 áreas agrícolas comerciais do Chile.

Delalibera, Weirich e Nagata (2012) também aplicaram a análise hierárquica de agrupamento para definição de ZMs e apresentaram resultados satisfatórios. Eles consideraram dados de variáveis químicas e físicas avaliadas em uma área de 22 ha localizada no estado do Paraná, com cultivo de soja, aveia e milho.

3.5.2 Métodos de agrupamento de particionamento

Os algoritmos de particionamento dividem um conjunto de elementos em *k* grupos sem a construção de uma estrutura hierárquica, seguindo o princípio de que elementos em um mesmo grupo devem ser mais similares que elementos pertencentes a grupos diferentes. Esses algoritmos realizam somente uma divisão dos dados, na tentativa de identificar grupos naturalmente presentes nesses dados (JAIN; DUBES, 1988). Nesse processo, geralmente tentam otimizar uma função de avaliação da partição, isto é, buscam organizar um conjunto de *n* elementos dentro de *k* grupos G_1, ..., G_k, enquanto maximizam ou minimizam uma função de avaliação pré-estabelecida. A quantidade de grupos normalmente é especificada, mas isto depende do algoritmo que se pretende usar (XU; WUNSCH, 2009).

O nível de homogeneidade dentro de cada grupo e o nível de separação (heterogeneidade) entre cada par de grupos são analisados com o uso das funções de avaliação. A mais utilizada para isso com algoritmos de particionamento é o critério da soma dos quadrados dos erros. Ao utilizá-lo, a partição que minimizar o valor desse critério será considerada ótima e chamada de partição de variância mínima (JAIN; MURTY; FLYNN, 1999; XU; WUNSCH, 2009).

Em relação aos algoritmos de particionamento que podem ser avaliados para a definição de ZMs, inicialmente apresentam-se k-means e fuzzy c-means. Estes dois métodos têm o propósito de dividir um conjunto de *n* elementos em determinada quantidade de grupos disjuntos, tendo como referência um centroide para cada grupo. Para isso, utilizam uma função de distância como, por exemplo, a distância euclidiana. Eles são capazes de realizar o agrupamento de maneira automática e buscam alcançar máxima similaridade entre os elementos de um mesmo grupo e o mínimo de similaridade entre grupos distintos. Para isso, seguem o critério da minimização da soma dos quadrados das distâncias entre os elementos e o centroide de cada grupo.

3.5.2.1 K-means

No algoritmo k-means, cada centroide é um elemento artificialmente gerado para representar o centro geométrico de um grupo. Sua execução segue um procedimento iterativo que visa alocar cada elemento p_i, $i = 1, ..., n$, do conjunto de dados ao grupo representado pelo centroide c_j, tal que c_j seja o centroide mais próximo de p_i. Este funcionamento é representado no fluxograma da Figura 4.

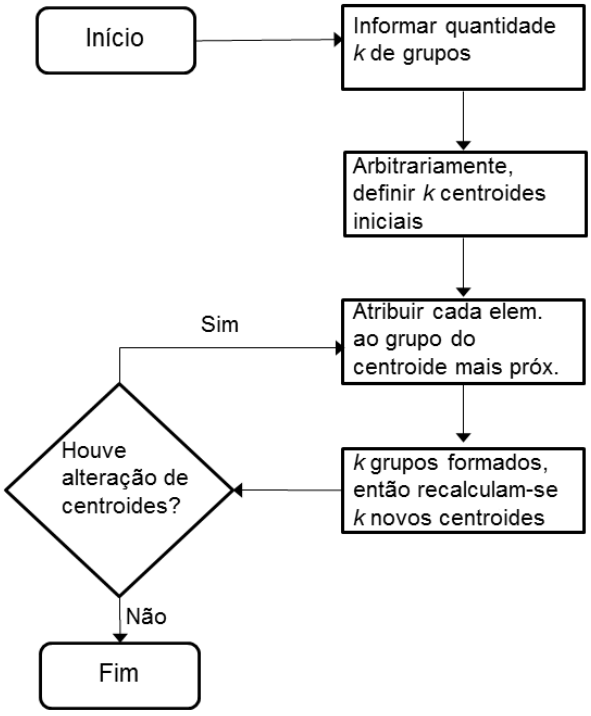


Figura 4 Fluxograma ilustrativo do funcionamento do método k-means.

3.5.2.2 Fuzzy c-means

No método fuzzy c-means, os elementos são inseridos de forma iterativa nos grupos, com base no critério da minimização da soma dos quadrados das distâncias entre os elementos e os centroides artificiais. Portanto, seguindo o mesmo princípio empregado por k-means. Entretanto, fuzzy c-means pode ser considerado mais robusto, pelo fato da flexibilidade própria da lógica difusa ou nebulosa (fuzzy logic) (ZADEH, 1965) ter sido incorporada a ele: cada elemento pode pertencer a todos os grupos com um certo grau de pertinência.

Isto pode ser útil quando as separações entre grupos não estão bem definidas. Além disso, os graus de pertinência eventualmente podem auxiliar usuários desse algoritmo a descobrirem relacionamentos inesperados entre determinados elementos alocados a grupos distintos (BEZDEK, 1981).

O seu processamento inicia a partir de um conjunto de *n* elementos $X = \{x_1, x_2, ..., x_n\}$, tal que para cada elemento tenham sido associados os valores de *p* variáveis de interesse. Busca-se encontrar uma partição que corresponda a *C* conjuntos difusos de X, que represente a estrutura dos dados da melhor forma possível e seja denotada por $P = \{A_1, A_2, ..., A_C\}$, tal que sejam respeitadas as seguintes restrições (MILNE et al., 2012): $\sum_{i=1}^{C} A_i(x_k) = 1$ e $0 < \sum_{k=1}^{n} A_i(x_k) < n$.

O algoritmo fuzzy c-means efetua seu processamento de acordo com parâmetros correspondentes à quantidade *C* de grupos que se deseja obter, uma função para determinar a distância entre os elementos e os centroides, e um valor chamado de erro ($\varepsilon > 0$), que é empregado como critério de parada. O funcionamento desse método é representado no fluxograma da Figura 5.

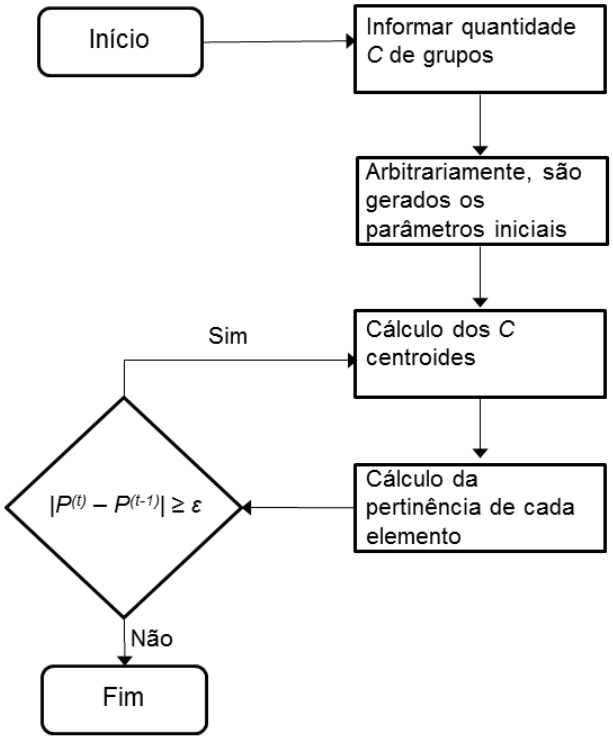


Figura 5 Fluxograma ilustrativo do funcionamento do método fuzzy c-means.

A execução iterativa é finalizada quando as pertinências calculadas nas iterações *t* e *t-1* apresentam uma diferença inferior ao valor do critério de parada, ou seja, quando se tem $|P^{(t)} - P^{(t-1)}| < \varepsilon$. Nesta situação, o algoritmo conclui seu processamento e os grupos são constituídos considerando as pertinências da última iteração.

3.5.2.3 Bagged clustering

O algoritmo bagged clustering foi definido por Leisch (1999) como uma combinação de método de particionamento com método hierárquico. A ideia de desenvolvê-lo surgiu do

problema de instabilidade de algoritmos como k-means e fuzzy c-means, devido à definição inicial dos centroides ser aleatória e à sensibilidade a valores discrepantes. O princípio de bagged clustering é executar repetidamente um método de particionamento e então combinar os resultados obtidos, a fim de estabilizá-los com a utilização de um método hierárquico. Segundo Leisch (1999), bagged clustering é executado seguindo-se estes cinco passos:

1. Realizar uma amostragem aleatória simples com reposição sobre os dados originais, a fim de obter diversos conjuntos de treinamento;
2. Executar um algoritmo de agrupamento de particionamento, neste caso chamado de método base, sobre cada um dos conjuntos de treinamento para obter agrupamentos com diferentes centroides;
3. Combinar todos os centroides obtidos por meio do uso de um método hierárquico; ou seja, os diversos centroides definidos pelo método base são combinados em um novo conjunto de dados, que passa a ser a entrada para um algoritmo hierárquico;
4. Determinar o centroide mais próximo de cada elemento do conjunto de dados original;
5. Realizar um corte no dendrograma em um certo nível, para gerar uma partição do conjunto original com a quantidade de grupos desejada.

3.5.2.4 Fuzzy analysis clustering

O algoritmo fuzzy analysis clustering (fanny) executa um agrupamento difuso para dividir um conjunto de dados em *k* grupos, similar ao realizado por fuzzy c-means. Porém, em comparação com este algoritmo, fanny possui duas vantagens: ele necessita apenas de uma matriz de distâncias entre os elementos para realizar o agrupamento, ou seja, não precisa dos valores medidos para as variáveis nos pontos; e apresenta maior robustez para lidar com valores discrepantes (KAUFMAN; ROUSSEEUW, 1990). Esse método de agrupamento não usa elementos representativos, como, por exemplo, centroides. Ao invés disso, tenta minimizar a função de avaliação mostrada na Equação 16 (KAUFMAN; ROUSSEEUW, 1990):

$$C = \sum_{v=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{iv}^{r} u_{jv}^{r} d(i,j)}{2 \sum_{j=1}^{n} u_{jv}^{r}} \qquad \text{Eq. (16)}$$

em que: u_{iv} significa a pertinência do elemento *i* em relação ao grupo *v*, *n* é o número de elementos que formam o conjunto de dados, *k* é a quantidade de grupos a serem formados, *r* é o expoente de pertinência e *d(i,j)* é a distância entre os elementos *i* e *j*.

O processamento de fanny ocorre de forma iterativa e é encerrado quando a função de avaliação converge, isto é, quando a diferença entre o valor de *C* nas iterações *t* e *t-1* se torna menor que um valor definido como critério de parada.

3.5.2.5 Fuzzy c-shells

O algoritmo fuzzy c-means original geralmente detecta grupos que correspondem espacialmente a formas esféricas. Porém, o uso de diferentes funções para a formação de matrizes de distâncias entre elementos produz variações dele (DAVE, 1992). Esse algoritmo foi generalizado por Bezdek (1981), ao permitir que, em termos de representação espacial, os centroides fossem objetos com muitas formas diferentes, de dimensões aleatórias e distintas.

Diante disso, Dave (1992) introduziu o método fuzzy c-shells como uma generalização de fuzzy c-means, por utilizar superfícies curvadas chamadas de hyper-spherical-shells como protótipos de grupos (equivalentes aos centroides). Esse conceito diferente para representação dos protótipos possibilitou que fuzzy c-shells obtivesse sucesso em algumas situações em que fuzzy c-means não obteve. Segundo Xu e Wunsch (2009), fuzzy c-shells pode ser melhor para a representação de grupos com formatos de contornos (círculos, elipses e linhas, por exemplo) com interiores vazios, para dados bidimensionais. Isto é exemplificado na Figura 6.

Figura 6 Exemplos de grupos com formatos curvos. Fonte: Xu e Wunsch (2009).

3.5.2.6 Hard competitive learning

A abordagem conhecida como hard competitive learning (online k-means) abrange métodos de agrupamento baseados na utilização de redes neurais artificiais. Resumidamente, segundo Xu e Wunsch (2009), a abordagem baseada em redes neurais consiste de três componentes básicos: iniciar com um conjunto de unidades de saída que seriam idênticas, exceto por algum parâmetro aleatoriamente distribuído que faz com que cada unidade responda de forma diferente das demais para um conjunto de dados de entrada *D*; limitar a força de cada unidade de saída em relação às demais; e permitir que as unidades disputem de alguma maneira o direito de responder a um determinado subconjunto de *D*.

Cada grupo está associado, na rede neural, a uma unidade de saída, que por sua vez está associada a um protótipo ou vetor $\boldsymbol{w}_j$, $j=1, ..., k$, em que k é o número de grupos. Basicamente, nos métodos de agrupamento do tipo hard competitive learning calcula-se a similaridade entre cada elemento de entrada e todos os vetores $\boldsymbol{w}_j$, mas apenas uma unidade de saída é ativada: aquela que apresenta maior valor de ativação da rede (XU; WUNSCH, 2009). Em termos de agrupamento, cada elemento deve ser inserido no grupo correspondente

à unidade de saída ativada, por ser a que apresentou maior nível de similaridade com o elemento.

O método online k-means é uma variação do k-means tradicional, com a diferença de que utiliza uma rede neural atualizada imediatamente após receber cada elemento de entrada. Assim, não é necessário ter o conjunto de dados completo antes de iniciar a execução de online k-means. Para cada elemento de entrada, emprega-se a distância euclidiana para determinar em qual grupo ele deve ser inserido e, então, adapta-se o protótipo do vencedor considerando a inserção do novo elemento no grupo.

3.5.2.7 Neural gas

O método neural gas (MARTINETZ; BERKOVICH; SCHULTEN, 1993) é do tipo soft competitive learning, ou seja, diferentemente do algoritmo online k-means, não apenas a unidade de saída vencedora na rede neural é adaptada depois da apresentação de um novo elemento de entrada. Além dela, todas as outras unidades de saída também são adaptadas.

Para cada elemento de entrada, o algoritmo neural gas ordena as unidades de saída da rede neural de acordo com a distância de cada protótipo de grupo em relação ao elemento. Após efetuar essa ordenação, todas as unidades de saída são adaptadas de acordo com sua posição na ordenação.

3.5.2.8 Partitioning around medoids

O algoritmo partitioning around medoids (PAM) (KAUFMAN; ROUSSEEUW, 1990) executa praticamente o mesmo procedimento que k-means para realizar o agrupamento de dados. A única diferença relevante é que PAM utiliza elementos reais, denominados medoids, como protótipos de grupos (abordagem denominada k-medoid). O medoid de um grupo é o elemento que apresenta a menor distância média em relação a todos os outros do mesmo grupo.

Para a definição dos *k* elementos que exercerão o papel de medoids de *k* grupos, testam-se muitas combinações diferentes de elementos candidatos. Os testes são finalizados quando se selecionam aqueles candidatos que proporcionem a menor distância média em relação aos elementos do conjunto de dados. De acordo com Kaufman e Rousseeuw (1990), essa estratégia do método PAM protege o procedimento de agrupamento da influência de valores discrepantes sobre os protótipos dos grupos, tornando-o mais robusto que k-means e fuzzy c-means.

3.5.2.9 Clustering large applications

O método PAM produz resultados satisfatórios em muitas situações. Todavia, conforme Kaufman e Rousseeuw (1986), ele poderia apresentar problemas relacionados a necessidades de tempo de processamento e quantidade de memória do computador, quando usado para agrupar grandes conjuntos de dados. Especialmente para particionar tais conjuntos, esses dois pesquisadores propuseram o algoritmo clustering large applications (clara), também baseado na abordagem k-medoid. Este algoritmo executa dois passos para particionar um conjunto de dados em *k* grupos:

1. Uma amostra aleatória simples é obtida do conjunto de dados e submetida ao algoritmo PAM, que seleciona os melhores medoids e particiona tal amostra em *k* grupos;
2. Cada elemento que não pertence à amostra é associado ao medoid mais próximo, obtendo-se assim uma partição do conjunto de dados completo.

Para medir a qualidade da partição, o valor médio da distância entre os elementos e seus correspondentes medoids é automaticamente calculado. Depois de gerar várias amostras distintas (a sugestão dos autores é para avaliar cinco) e executar o procedimento completo de agrupamento para cada uma, o resultado final do método é a partição que apresenta o menor valor para a distância média entre os elementos e seus medoids.

3.5.2.10 Spherical k-means

O algoritmo spherical k-means (DHILLON; MODHA, 2001) foi proposto principalmente para tratar de elementos de alta dimensionalidade e esparsos, ou seja, elementos representados por grande número de variáveis e sem valores conhecidos para muitas delas. Em sua definição, cada elemento representado por *d* variáveis quantitativas é tratado como um vetor no espaço $\boldsymbol{R}^d$. Para medir a distância entre os vetores, utiliza-se a função chamada de similaridade cosseno (DHILLON; MODHA, 2001). Com exceção disso, spherical k-means segue o mesmo procedimento de agrupamento de k-means, pois foi desenvolvido a partir deste.

Na avaliação de Xu e Wunsch (2009), as modificações que originaram spherical k-means o tornaram capaz de eventualmente gerar agrupamentos satisfatórios em menor tempo do que o exigido por k-means.

3.5.2.11 Unsupervised fuzzy competitive learning

Este método é a versão online de fuzzy c-means. Toda vez que um elemento do conjunto de dados é escolhido para ser uma entrada para unsupervised fuzzy competitive learning (UFCL) (PAL; BEZDEK; HATHAWAY, 1996), o grau de pertinência deste elemento

em relação aos *c* grupos é calculado e os centroides são atualizados. Portanto, diferentemente de fuzzy c-means, o algoritmo UFCL não necessita que todos os elementos do conjunto de dados estejam disponíveis para poder ser executado. Esta é a única diferença importante entre esses dois métodos de agrupamento.

3.5.2.12 Resultados com métodos de particionamento

Bazzi et al. (2013), Caires, Wuddivira e Bekele (2015), Fraisse, Sudduth e Kitchen (2001), Li et al. (2013), Molin e Castro (2008), Moral, Terrón e Silva (2010) e Schenatto et al. (2016) apresentaram resultados satisfatórios da aplicação de fuzzy c-means para definir ZMs, considerando diversos tipos de solo, variáveis e plantas cultivadas. Todos concluíram que as subáreas resultantes foram corretamente definidas, considerando as variabilidades espaciais e temporais da produtividade, além de serem úteis para a criação de planos de amostragem de solo. Já Arno et al. (2011), Ikenaga e Inamura (2008), Jipkate e Gohokar (2012), Perez-Quezada, Pettygrove e Plant (2003) e Rodrigues Junior et al. (2011) apresentaram resultados igualmente satisfatórios em relação à utilização do algoritmo k-means para gerar ZMs.

Tichý, Chytrý e Botta-Dukát (2014) mostraram que o método PAM pode proporcionar resultados melhores que k-means quando o agrupamento é realizado a partir de conjuntos de dados com valores discrepantes. Já Fu, Wang e Jiang (2010) apresentaram um algoritmo derivado de fuzzy c-means que pode, em algumas situações, melhorar o desempenho do algoritmo original na definição de ZMs.

3.6 Avaliação de zonas de manejo

Uma forma de avaliação de ZMs frequentemente realizada é constatar se as subáreas definidas apresentam diferenças estatisticamente significativas de potencial produtivo. Segundo Moral, Terrón e Silva (2010) e Saleh e Belal (2014), isso possibilita decidir se cada ZM pode ser tratada como subárea de gerenciamento diferenciado do restante da área. De acordo com Bazzi et al. (2013), embora qualquer variável possa ser utilizada nesse tipo de avaliação, a produtividade geralmente é a recomendada.

Também é relevante avaliar o resultado da geração de ZMs para inferir sobre a melhor quantidade de subáreas a serem implantadas. Sobre este item, Fraisse, Sudduth e Kitchen (2001) e Fridgen et al. (2004) destacam que quanto menor for essa quantidade, mais fácil será executar as operações de campo.

Alguns critérios relevantes para a avaliação do processo de definição e das próprias ZMs obtidas são análise de variância (ANOVA), teste de comparação de médias de Tukey (PIMENTEL-GOMES, 2000), índice de redução da variância (variance reduction - VR) (PING; DOBERMANN, 2003), coeficiente de silhueta médio (average silhouette coefficient - ASC) (ROUSSEEUW, 1987) e índice de suavidade (smoothness index - SI) (GAVIOLI et al., 2016).

A ANOVA e o teste de comparação de médias de Tukey podem ser usados para verificar se as ZMs geradas apresentam produtividades médias estatisticamente diferentes entre si, a determinado nível de significância. Contudo, segundo Bazzi et al. (2013), antes de usá-los deve-se ter certeza de que não ocorre dependência espacial entre as amostras em cada ZM.

O índice VR pode ser calculado a partir dos valores da variância da produtividade das ZMs e da área como um todo. A expectativa é que o somatório das variâncias das subáreas seja menor que a variância original da área (Equação 17). Portanto, quanto maior for o valor do índice VR, melhor terá sido a definição das ZMs em termos de redução da variância.

$$VR = \left(1 - \frac{\sum_{i=1}^{c} W_i * V_{zmi}}{V_{área}} \right) * 100 \qquad \text{Eq. (17)}$$

em que: *c* é a quantidade de ZMs; W_i é a proporção da área total referente à i-ésima ZM; V_{zmi} é a variância dos dados da i-ésima ZM; $V_{área}$ é a variância dos dados da área como um todo.

O coeficiente ASC é obtido a partir do coeficiente de silhueta (silhouette coefficient - SC), um critério de avaliação que mede a qualidade da formação interna e da separação externa de grupos. O valor do SC para um ponto *p*, denotado por sc_p, é calculado utilizando-se a Equação 18 (ROUSSEEUW, 1987):

$$sc_p = \frac{b_p - a_p}{Max(a_p, b_p)} \qquad \text{Eq. (18)}$$

em que: a_p é a média das distâncias entre o ponto *p* e todos os demais pontos pertencentes ao mesmo grupo, e b_p é a média das distâncias entre o ponto *p* e todos os pontos do grupo mais próximo ao que contém *p*.

O coeficiente de silhueta de grupo (group silhouette coefficient – GSC) é obtido calculando-se a média dos coeficientes de silhueta dos pontos desse grupo, e então o valor correspondente ao coeficiente ASC de *k* grupos é obtido calculando-se a média dos valores do GSC dos *k* grupos. Os valores do ASC variam entre -1 e 1, tal que -1 indica um agrupamento incorreto e 1 indica grupos com a melhor formação intra-grupo e a melhor separação inter-grupos possíveis.

O índice SI possibilita calcular a frequência de mudança de ZMs em um mapa temático nas direções horizontal e vertical, bem como nas diagonais, pixel a pixel (Equação 19). Na hipótese de um mapa possuir uma área totalmente homogênea, o resultado é SI = 100%, devido à ausência de mudanças de zonas. Por outro lado, se as ZMs forem muito fragmentadas, o índice SI terá um valor próximo de 0.

$$SI = 100 - \left(\left(\frac{\sum_{i=1}^{k} NM_{Hi}}{4P_H} + \frac{\sum_{j=1}^{k} NM_{Vj}}{4P_V} + \frac{\sum_{l=1}^{k} NM_{Ddl}}{4P_{Dd}} + \frac{\sum_{m=1}^{k} NM_{Dem}}{4P_{De}} \right) * 100 \right) \quad \text{Eq. (19)}$$

em que: *k* é o número de linhas, colunas ou diagonais; NM_{Hi} é o número de mudanças na linha *i* (horizontal); NM_{Vj} é o número de mudanças na coluna *j* (vertical); NM_{Ddl} é o número de mudanças na diagonal *l* (diagonal direita *Dd*); NM_{Dem} é o número de mudanças na diagonal *m* (diagonal esquerda *De*); P_H é a possibilidade de mudanças de pixels na horizontal; P_V é a possibilidade de mudanças de pixels na vertical; P_{Dd} é a possibilidade de mudanças na diagonal direita *Dd*; e P_{De} é a possibilidade de mudanças na diagonal esquerda *De*.

Especificamente no caso da aplicação de algoritmos de agrupamento baseados na lógica difusa, como fanny, fuzzy c-means, fuzzy c-shells e UFCL, pode-se considerar mais dois critérios de avaliação: o índice de desempenho fuzzy (fuzziness performance index - FPI) (FRIDGEN et al., 2004) e a entropia de partição modificada (modified partition entropy - MPE) (BOYDELL; MCBRATNEY, 2002).

O índice FPI permite determinar o grau de separação entre os grupos gerados pelo algoritmo. Seu valor varia entre 0 e 1, tal que quanto mais próximo for de 0, menor será o grau de compartilhamento de elementos entre os grupos gerados (Equação 20). Já o critério MPE é uma estimativa da quantidade de desorganização criada por um número específico de grupos c, tal que quanto mais próximo de 0 for seu valor, melhor (Equação 21).

$$FPI = 1 - \frac{c}{(c-1)} \left[1 - \sum_{j=1}^{n} \sum_{i=1}^{c} (m_{ij})^2 / n \right] \quad \text{Eq. (20)}$$

$$MPE = \frac{-\sum_{j=1}^{n} \sum_{i=1}^{c} m_{ij} \log(m_{ij}) / n}{\log c} \quad \text{Eq. (21)}$$

em que: *c* é a quantidade de grupos; *n* é a quantidade de elementos no conjunto de dados e m_{ij} é o valor correspondente ao grau de pertinência do j-ésimo elemento do conjunto em relação ao i-ésimo grupo.

3.7 Softwares para seleção de variáveis e agrupamento

A maioria dos softwares utilizados no processo de criação de ZMs não foi projetada especificamente para essa finalidade. Além disso, muitos foram desenvolvidos antes da difusão da abordagem de ZMs. Consequentemente, esses softwares costumam ser usados de forma adaptada nesta área de pesquisa.

Para executar o método de seleção de variáveis ACP, tem-se, por exemplo, os softwares R (R CORE TEAM, 2016), ArcGIS (Esri Software), Matlab (The MathWorks), SAS (SAS Institute), Scilab (Scilab Enterprises) e Statistica (Statsoft Inc.). Para utilizar o método MULTISPATI-PCA, tem-se apenas o software R disponível. Para efetuar análises espaciais sobre dados, destacam-se os softwares R, ArcGIS, GS+ (Gamma Design Software) e Surfer (Golden Software).

Os softwares FuzME e MZA são úteis para gerar agrupamentos por meio do algoritmo fuzzy c-means. Porém, não disponibilizam métodos para seleção das variáveis necessárias para a geração de ZMs e também não contêm outros métodos de agrupamento. Já o software SDUM é mais abrangente, pois disponibiliza os algoritmos de agrupamento k-means e fuzzy c-means, viabiliza a seleção de variáveis por meio do uso da análise de correlação espacial e possibilita a geração de mapas temáticos para representar as ZMs. Ele também possibilita avaliar as subáreas geradas, disponibilizando para isso a ANOVA, o teste de Tukey e um índice similar ao VR denominado eficiência relativa.

4 REFERÊNCIAS

ARNO, J.; MARTINEZ-CASASNOVAS, J. A.; RIBES-DASI, M.; ROSELL, J. R. Clustering of grape yield maps to delineate site-specific management zones. **Spanish Journal of Agricultural Research**, v. 9, n. 3, p. 721-729, 2011.

ARROUAYS, D.; SABY, N. P. A.; THIOULOUSE, J.; JOLIVET, C.; BOULONNE, L.; RATIÉ, C. Large trends in French topsoil characteristics are revealed by spatially constrained multivariate analysis. **Geoderma**, v. 161, p. 107-114, 2011.

BAZZI, C. L.; SOUZA, E. G.; URIBE-OPAZO, M. A.; NÓBREGA, L. H. P.; ROCHA, D. M. Management zones definition using soil chemical and physical attributes in a soybean area. **Engenharia Agrícola**, v. 33, n. 5, p. 952-964, 2013.

BEUTLER, A. N.; MUNARETO, J. D.; RAMÃO, C. J.; GALON, L.; DIAS, N. P.; POZZEBON, B. C. Propriedades físicas do solo e produtividade de arroz irrigado em diferentes sistemas de manejo. **Revista Brasileira de Ciência do Solo**, v. 36, p. 1601-1607, 2012.

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. New York: Plenum Press, 1981. 256 p.

BOYDELL, B.; MCBRATNEY, A. B. Identifying potential within-field management zones from cotton-yield estimates. **Precision Agriculture,** v. 3, n. 1, p. 9-23, 2002.

CAIRES, S. A.; WUDDIVIRA, M. N.; BEKELE, I. Spatial analysis for management zone delineation in a humid tropic cocoa plantation. **Precision Agriculture,** v. 16, p. 129-147, 2015.

CARMO, D. L.; NANNETTI, D. C.; DIAS JÚNIOR, M. S.; ESPÍRITO SANTO, D. J.; NANNETTI, A. N.; LACERDA, T. M. Propriedades físicas de um latossolo vermelho-amarelo cultivado com cafeeiro em três sistemas de manejo no sul de Minas Gerais. **Revista Brasileira de Ciência do Solo**, v. 35, n. 3, p. 991-998, 2011.

CATTELL, R. B. The scree test for the number of factors. **Multivariate Behavioral Research**, v. 1, p. 245-276, 1966.

CHIPMAN, H.; TIBSHIRANI, R. Hybrid Hierarchical Clustering with Applications to Microarray Data. **Biostatistics,** v. 7, p. 302-317, 2006.

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, p. 20-28, 2014.

COHEN, S.; COHEN, Y.; ALCHANATIS, V.; LEVI, O. Combining spectral and spatial information from aerial hyperspectral images for delineating homogenous management zones. **Biosystems Engineering,** v. 114, n. 4, p. 435-443, 2013.

CÓRDOBA, M.; BALZARINI, M.; BRUNO, C.; COSTA, J. L. Análisis de componentes principales con datos georreferenciados: Una aplicación en agricultura de precisión. **Revista de la Facultad de Ciencias Agrarias UNCUYO**, v. 44, n. 1, p. 27-39, 2012.

CÓRDOBA, M.; BRUNO, C.; COSTA, J. L.; 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.

CZAPLEWSKI, R. L.; REICH, R. M. **Expected value and variance of Moran's bivariate spatial autocorrelation statistic under permutation**. Research Paper RM-309. Fort Collins: USDA Forest Service, 1993. 13 p.

DAVE, R. N. Generalized fuzzy c-shells clustering and detection of circular and elliptical boundaries. **Pattern Recognition,** v. 25, n. 7, p. 713-721, 1992.

DELALIBERA, H. C.; WEIRICH, P. H.; NAGATA, N. Management zones in agriculture according to the soil and landscape variables. **Engenharia Agrícola,** v. 32, n. 6, p. 1197-1204, 2012.

DHILLON, I. S.; MODHA, D. S. Concept decompositions for large sparse text data using clustering. **Machine Learning,** v. 42, p. 143-175, 2001.

DOBERMANN, A.; PING, J. L.; ADAMCHUK, V. I.; SIMBAHAN. G. C.; FERGUSON, R. B. Classification of crop yield variability in irrigated production fields. **Agronomy Journal**, v. 95, n. 1, p. 1105-1120, 2003.

DRAY, S.; SAID, S.; DÉBIAS, F. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. **Journal of Vegetation Science,** v. 19, p. 45-56, 2008.

EVERITT, B.; LANDAU, S.; LEESE, M. **Cluster Analysis**. 5 ed. London: John Wiley & Sons, 2011. 330 p.

FERRAZ, G. A. E. S.; SILVA, F. M.; CARVALHO, F. M.; COSTA, P. A. N.; CARVALHO, L. C. C. Viabilidade econômica do sistema de adubação diferenciado comparado ao sistema de adubação convencional em lavoura cafeeira: um estudo de caso. **Engenharia Agrícola**, v. 31, n. 5, p. 906-915, 2011.

FERREIRA, D. F. **Análise multivariada**. Lavras: UFLA, 1996. 394 p.

FRAISSE, C. W.; SUDDUTH, K. A.; KITCHEN, N. R. Delineation of site–specific management zones by unsupervised classification of topographic attributes and soil electrical conductivity. **International Journal of the American Society of Agricultural and Biological Engineers,** v. 44, n. 1, p. 155–166, 2001.

FRIDGEN, J. J.; KITCHEN, N. R.; SUDDUTH, K. A.; DRUMMOND, S. T.; WIEBOLD, W. J.; FRAISSE, C. W. Management zone analyst (MZA): software for subfield management zone delineation. **Agronomy Journal**, v. 96, p. 100-108, 2004.

FU, Q.; WANG, Z.; JIANG, Q. Delineating soil nutrient management zones based on fuzzy clustering optimized by PSO. **Mathematical and Computer Modelling**, v. 51, n. 11-12, p. 1299-1305, 2010.

GALAMBOSOVÁ, J.; RATAJ, V.; PROKEINOVÁ, R.; PRESINSKÁ, J. Determining the management zones with hierarchic and non-hierarchic clustering methods. **Research in Agricultural Engineering**, v. 60, p. 44-51, 2014.

GAVIOLI, A.; SOUZA, E. G.; BAZZI, C. L.; GUEDES, L. P. C.; SCHENATTO, K. Optimization of management zone delineation by using spatial principal components. **Computers and Electronics in Agriculture,** v. 127, p. 302-310, 2016.

GUASTAFERRO, F.; CASTRIGNANO, A.; DE BENEDETTO, D.; SOLLITTO, D.; TROCCOLI, A.; CAFARELLI, B. A comparison of different algorithms for the delineation of management zones. **Precision Agriculture,** v. 11, p. 600-620, 2010.

GUEDES, E. M. S.; FERNANDES, A. R.; LIMA, H. V.; SERRA, A. P.; COSTA, J. R.; GUEDES, R. S. Impacts of different management systems on the physical quality of an Amazonian Oxisol. **Revista Brasileira de Ciência do Solo**, v. 36, n. 4, p. 1269-1278, 2012.

HORN, J. L. A rationale and test for the number of factors in factor analysis. **Psychometrika**, v. 30, p. 179-185, 1965.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of educational psychology,** v. 24, n. 6, p. 417-441, 1933.

IKENAGA, S.; INAMURA, T. Evaluation of site-specific management zones on a farm with 124 contiguous small paddy fields in a multiple-cropping system. **Precision Agriculture,** v. 9, p. 147-159, 2008.

JAIN, A. K.; DUBES, R. **Algorithms for clustering data**. Englewood Cliffs: Prentice-Hall, 1988. 320 p.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data Clustering: A Review. **ACM Computing Surveys,** v. 31, n. 3, p. 264-323, 1999.

JIPKATE, B. R.; GOHOKAR, V. V. A comparative analysis of Fuzzy C-Means clustering and K-Means clustering algorithms. **International Journal of Computational Engineering**, v. 2, n. 3, p. 737-739, 2012.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6 ed. New Jersey: Pearson, 2007. 773 p.

JOLLIFFE, I. T. **Principal Component Analysis**. 2 ed. New York: Springer, 2002. 487 p.

KAUFMAN, L.; ROUSSEEUW, P. J. Clustering Large Data Sets. In: GELSEMA, E. S. e KANAL, L. N. (Ed.). **Pattern Recognition in Practice II**. North-Holland: Elsevier, 1986. p. 425-435.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data**. Hoboken: John Wiley & Sons, 1990. 342 p.

KLEIN, V. A.; BASEGGIO, M.; MADALOSSO, T.; MARCOLIN, C. D. Textura do solo e a estimativa do teor de água no ponto de murcha permanente com psicrômetro. **Ciência Rural**, v. 40, n. 7, p. 1550-1556, 2010.

LANCE, G.; WILLIAMS, W. A general theory of classification sorting strategies: 1. Hierarchical systems. **Computer Journal,** v. 9, p. 373-380, 1967.

LEISCH, F. Bagged clustering. In: SFB ADAPTIVE INFORMATION SYSTEMS AND MODELLING IN ECONOMICS AND MANAGEMENT SCIENCE, 51, 1999, Vienna. **Anais...** Vienna: Vienna University of Economics and Business, 1999. p. 1-11.

LI, Y.; SHI, Z.; WU, H.; LI, F.; LI, H. Definition of management zones for enhancing cultivated land conservation using combined spatial data. **Environmental Management,** v. 52, n. 1, p. 792-806, 2013.

LIMA, A. F. **Desenvolvimento de métodos para o preparo de amostras de fertilizantes visando à determinação de cobre, cádmio e chumbo por espectrometria de absorção atômica com chama**. 2010. 66 f. Dissertação (Mestrado em Química) - Universidade Federal de Uberlândia, Uberlândia, 2010.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: PROCEEDINGS OF 5TH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 1967, Berkeley. **Anais...** Berkeley: University of California Press, 1967. p. 281–297.

MARTINETZ, T. M.; BERKOVICH, S. G.; SCHULTEN, K. J. "Neural-gas" network for vector quantization and its application to time-series prediction. **IEEE Transactions on Neural Networks,** v. 4, n. 4, p. 558-569, 1993.

MAZZINI, P. L. F.; SCHETTINI, C. A. F. Avaliação de metodologias de interpolação especial aplicadas a dados hidrográficos costeiros quase-sinópticos. **Brazilian Journal of Aquatic Science Technology**, v. 13, n. 1, p. 53-64, 2009.

MCBRATNEY, A.; WHELAN, B.; ANCEV, T.; BOUMA, J. Future Directions of Precision Agriculture. **Precision Agriculture,** v. 6, p. 7–23, 2005.

MCQUITTY, L. L. Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. **Educational and Psychological Measurement,** v. 26, p. 825-831, 1966.

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.

MINASNY, B.; MCBRATNEY, A. B. **FuzME 3.0**. Australian Centre for Precision Agriculture. The University of Sydney. Sydney. 2002.

MIRANDA, J. I. **Fundamentos de Sistemas de Informações Geográficas**. 2 ed. Brasília: Embrapa, 2010. 425 p.

MOLIN, J. P.; CASTRO, C. N. Establishing management zones using soil electrical conductivity and other soil properties by the fuzzy clustering technique. **Scientia Agricola**, v. 65, n. 6, p. 567-573, 2008.

MOLIN, J. P.; FAULIN, G. C. Spatial and temporal variability of soil electrical conductivity related to soil moisture. **Scientia Agricola**, v. 70, n. 1, p. 1-5, 2013.

MORAL, F. J.; TERRÓN, J. M.; SILVA, J. R. M. Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. **Soil and Tillage Research**, v. 106, n. 2, p. 335-343, 2010.

ORTEGA, R. A.; SANTIBÁÑEZ, O. A. 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, 2007.

PAL, N. R.; BEZDEK, J. C.; HATHAWAY, R. J. Sequential competitive learning and the fuzzy c-means clustering algorithm. **Neural Networks,** v. 9, n. 5, p. 787-796, 1996.

PERALTA, N. R.; COSTA, J. L.; BALZARINI, M.; FRANCO, M. C.; CÓRDOBA, M.; BULLOCK, D. Delineation of management zones to improve nitrogen management of wheat. **Computers and Electronics in Agriculture,** v. 110, p. 103-113, 2015.

PEREZ-QUEZADA, J. F.; PETTYGROVE, G. S.; PLANT, R. E. Spatial-Temporal Analysis of Yield and Soil Factors in Two Four-Crop–Rotation Fields in the Sacramento Valley, California. **Agronomy Journal**, v. 95, p. 676-687, 2003.

PIMENTEL-GOMES, F. **Curso de Estatística Experimental.** 14 ed. Piracicaba: Universidade de São Paulo, 2000. 477 p.

PING, J. L.; DOBERMANN, A. Creating spatially contiguous yield classes for site-specific management. **Agronomy Journal,** v. 95, n. 5, p. 1121-1131, 2003.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2016. 99 p.

REICH, R. M.; CZAPLEWSKI, R. L.; BECHTOLD, W. A. Spatial cross-correlation of undisturbed, natural shortleaf pine stands in northern Georgia. **Environmental and Ecological Statistics**, v. 1, p. 201-217, 1994.

RODRIGUES JUNIOR, F. A.; VIEIRA, L. B.; QUEIROZ, D. M.; SANTOS, N. T. Geração de zonas de manejo para cafeicultura empregando-se sensor SPAD e análise foliar. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 15, n. 8, p. 778-787, 2011.

ROUSSEEUW, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics,** v. 20, p. 53-65, 1987.

SALEH, A.; BELAL, A. A. Delineation of site-specific management zones by fuzzy clustering of soil and topographic attributes: a case study of East Nile Delta, Egypt. **IOP Conference Series: Earth and Environmental Science**, v. 18, p. 1-6, 2014.

SCHENATTO, K.; SOUZA, E. G.; BAZZI, C. L.; BIER, V. A.; BETZEK, N. M.; GAVIOLI, A. Data interpolation in the definition of management zones. **Acta Scientiarum,** v. 38, n. 1, p. 31-40, 2016.

TICHÝ, L.; CHYTRÝ, M.; BOTTA-DUKÁT, Z. Semi-supervised classification of vegetation: preserving the good old units and searching for new ones. **Journal of Vegetation Science**, v. 25, p. 1504-1512, 2014.

TRIPATHI, R.; NAYAK, A. K.; SHAHID, M.; LAL, B.; GAUTAM, P.; RAJA, R.; MOHANTY, S.; KUMAR, A.; PANDA, B. B.; SAHOO, R. N. Delineation of soil management zones for a rice cultivated area in eastern India using fuzzy clustering. **Catena,** v. 133, p. 128-136, 2015.

WARD, J. H. Hierarchical Grouping to Optimize an Objective Function. **Journal of the American Statistical Association,** v. 58, n. 301, p. 236-244, 1963.

XU, R.; WUNSCH, D. C. **Clustering**. Piscataway: IEEE Press, 2009. 358 p.

ZADEH, L. A. Fuzzy sets. **Information and Control**, v. 8, n. 1, p. 338-353, 1965.

5 ARTIGO 1 – OTIMIZAÇÃO DO DELINEAMENTO DE ZONAS DE MANEJO POR MEIO DO USO DE COMPONENTES PRINCIPAIS ESPACIAIS[1]

Resumo
A definição de zonas de manejo em áreas agrícolas consiste na delimitação de subáreas com características topográficas, de solo e/ou de plantas cultivadas similares. Nesse processo, é comum a aplicação de algoritmos de agrupamento como fuzzy c-means. Junto com esse tipo de algoritmo, também podem ser aplicadas abordagens para seleção de variáveis, como a análise de correlação espacial, a análise de componentes principais (ACP) e a análise espacial multivariada baseada no índice de Moran e em ACP (MULTISPATI-PCA). Neste artigo, propôs-se um novo método para seleção de variáveis, denominado MPCA-SC, baseado no uso conjunto da análise de correlação espacial e de MULTISPATI-PCA. Além disso, avaliou-se a eficiência de MPCA-SC e de mais quatro métodos baseados nas abordagens de seleção de variáveis citadas, quando utilizados conjuntamente com fuzzy c-means. As avaliações foram realizadas com dados coletados de 2010 a 2014 em três áreas agrícolas localizadas no estado do Paraná, com cultivo de milho e soja. A partir desses dados, foram geradas duas, três e quatro classes. As zonas de manejo delineadas a partir dessas classes ficaram visualmente diferentes, conforme o método usado. O método MPCA-SC proporcionou o melhor desempenho para o algoritmo fuzzy c-means e os melhores valores de redução da variância dos dados após a delimitação das subáreas. MPCA-SC também propiciou zonas de manejo com maior homogeneidade interna, tornando-as mais viáveis para implantação sob o ponto de vista de operações de campo.

Palavras-chave: agricultura de precisão; análise de componentes principais; fuzzy c-means; índice de Moran; MULTISPATI-PCA.

OPTIMIZATION OF MANAGEMENT ZONE DELINEATION BY USING SPATIAL PRINCIPAL COMPONENTS

Abstract
Definition of management zones is the delimitation of sub-areas with similar topographic, soil and/or crop characteristics within a field. Clustering algorithms such as fuzzy c-means are also frequently applied to define management zones. Three variable selection approaches that can be applied with clustering algorithms are spatial correlation analysis, principal component analysis (PCA), and multivariate spatial analysis based on Moran's index and PCA (MULTISPATI-PCA). In this study, the efficiency of each of these three approaches used in conjunction with the fuzzy c-means method was assessed. Furthermore, a new variable selection method, named MPCA-SC, based on the combined use of spatial correlation analysis and MULTISPATI-PCA, was proposed and assessed. The evaluation was performed by using data collected from 2010 to 2014 from three agricultural areas in Paraná State, Brazil, with corn and soybean crops, generating two, three, and four classes. The delineated management zones were different according to the method used, and MPCA-SC provided the best performance for the fuzzy c-means algorithm and the best variance reduction values of the data after the delimitation of the sub-areas. Furthermore, MPCA-SC provided management zones with greater internal homogeneity, making them more viable for implementation from the viewpoint of field operations.

Keywords: fuzzy c-means; Moran's index; MULTISPATI-PCA; precision agriculture; principal component analysis.

[1] Artigo publicado em 2016 no periódico internacional Computers and Electronics in Agriculture, com classificação A2 no QUALIS/CAPES na área de Ciências Agrárias.

5.1 Introdução

Segundo Schepers et al. (2004), definem-se zonas de manejo (ZMs) por meio da delimitação de subáreas com características do solo, do relevo e/ou de plantas cultivadas similares dentro de uma área agrícola. Isso permite gerenciar de maneira uniforme cada subárea, já que esta requer quantidades similares de insumos agrícolas em toda a sua extensão.

O delineamento de ZMs pode contribuir para viabilizar economicamente a agricultura de precisão (AP) para um maior número de produtores. Isto porque o tipo de manejo e as taxas fixas de aplicação de insumos dentro de cada subárea permitem o uso de máquinas e equipamentos da agricultura convencional, que normalmente custam menos que aqueles próprios da AP.

De acordo com Fraisse, Sudduth e Kitchen (2001), as ZMs também podem representar indicadores para amostragem do solo e das culturas plantadas, reduzindo o número de amostras a serem analisadas, sem comprometer a obtenção de resultados confiáveis. Dados de produtividade, dados químicos e físicos do solo, dados topográficos e de condutividade elétrica aparente do solo, índices de vegetação ou combinações entre estes podem ser utilizados para definir as subáreas.

Porém, recomenda-se que apenas variáveis (atributos) temporalmente estáveis que apresentem correlação espacial significativa com a produtividade sejam empregadas na geração das subáreas (DOERGE, 2000). Isto se deve à intenção de que as subáreas sejam utilizadas por vários anos e, em geral, conduz à eliminação das variáveis químicas do solo desse processo.

Para a delimitação de ZMs, também costumam-se empregar algoritmos de agrupamento, como fuzzy c-means (FRIDGEN et al., 2004; FU; WANG; JIANG, 2010; HORNUNG et al., 2006; LI et al., 2013; ZHANG et al., 2013).

De acordo com Gnanadesikan, Kettenring e Tsao (1995), a seleção de variáveis é uma tarefa difícil na análise de agrupamento. A capacidade de softwares de agrupamento de processarem um grande número de variáveis tende a incentivar a utilização de muitas nesse processo. Todavia, deve-se estar ciente de que a escolha das variáveis e dos pesos atribuídos a elas exercem influência sobre a formação dos grupos.

Três métodos de seleção de variáveis que podem ser aplicados em combinação com o algoritmo fuzzy c-means são: análise de correlação espacial (CZAPLEWSKI; REICH, 1993), aplicada conforme descrito por Bazzi et al. (2013); análise de componentes principais (ACP) (HOTELLING, 1933), usada por Cohen et al. (2013), Fraisse, Sudduth e Kitchen (2001), Li et al. (2007) e Moral, Terrón e Silva (2010); e análise espacial multivariada baseada no índice

de Moran e em ACP (MULTISPATI-PCA) (DRAY; SAID; DÉBIAS, 2008), empregada por Córdoba et al. (2013, 2016) e Peralta et al. (2015).

Para analisar a correlação espacial entre variáveis, utiliza-se a estatística de correlação espacial bivariada de Moran (CZAPLEWSKI; REICH, 1993). Essa estatística também é empregada para avaliar se amostras correspondentes a uma área apresentam autocorrelação espacial.

A ACP permite identificar as variáveis que explicam a maior parte da variância total existente em conjuntos de dados. Ao executá-la, realizam-se transformações a partir das variáveis originais que resultam em um novo conjunto de variáveis sintéticas, as componentes principais (CPs) (JOHNSON; WICHERN, 2007). O método MULTISPATI-PCA visa adicionar uma restrição espacial à ACP tradicional, melhorando-o para ser executado quando há dependência espacial em conjuntos de dados georreferenciados. Baseia-se na introdução de uma matriz de ponderação espacial ao método ACP, tal que esta matriz é construída utilizando-se a estatística de autocorrelação de Moran.

Uma vantagem de MULTISPATI-PCA em relação à ACP é que os escores obtidos com o primeiro método maximizam a autocorrelação espacial entre pontos, enquanto os escores obtidos com o segundo maximizam a variância total (CÓRDOBA et al., 2013). Assim, segundo Arrouays et al. (2011), os escores gerados com MULTISPATI-PCA mostram estruturas espaciais fortes nas primeiras CPs, ao passo que os escores da ACP podem mostrar estruturas espaciais em quaisquer componentes, até mesmo nas últimas, que na prática costumam ser desconsideradas.

O objetivo deste artigo é propor um método para seleção de variáveis a serem utilizadas na definição de ZMs, baseado no uso conjunto de MULTISPATI-PCA e análise de correlação espacial. Além disso, avalia-se a eficiência do novo método e de mais quatro algoritmos derivados da análise de correlação espacial, ACP e MULTISPATI-PCA, quando aplicados com o algoritmo fuzzy c-means.

5.2 Material e métodos

5.2.1 Conjuntos de dados

Utilizaram-se dados coletados entre 2010 e 2014 de três áreas agrícolas comerciais com cultivo de milho e soja, localizadas no estado do Paraná. Estas áreas estão representadas na Figura 1 (destaca-se que as imagens estão em escalas diferentes). Os solos foram classificados como LATOSSOLO VERMELHO Distroférrico típico (EMBRAPA, 2013) e cultivados em sistema de plantio direto. A área A possui 15,5 ha e está localizada no município de Céu Azul, com localização geográfica central de 25º06'32" S e 53º49'55" O e altitude média de 460 m. A área B se estende por 9,9 ha e está localizada no município de Serranópolis do Iguaçu, com localização geográfica central de 25º24'28" S e 54º00'17" O e altitude média de

355 m. Já a área C possui 19,8 ha e está localizada no município de Cascavel, com localização geográfica central de 24º57'08'' S e 53º33'59'' O e altitude média de 650 m.

Área A Área B Área C

Figura 1 As três áreas experimentais: área A, em Céu Azul - PR; área B, em Serranópolis do Iguaçu - PR; área C, em Cascavel - PR.

Para a definição das ZMs, utilizaram-se somente as variáveis consideradas temporalmente estáveis (Tabela 1), visando atender à recomendação de Doerge (2000). Com o uso de grades irregulares, realizou-se a amostragem em 40 (2,67 pontos ha^{-1}), 42 (4,24 pontos ha^{-1}) e 68 (3,43 pontos ha^{-1}) pontos georreferenciados nas áreas A, B e C, respectivamente. Os pontos amostrais ficaram localizados na linha imaginária central entre as curvas de nível de cada área.

Tabela 1 Variáveis avaliadas e anos de coleta de dados, para cada área agrícola

Variáveis (atributos)	Área A			Área B			Área C	
	2012	2013	2014	2012	2013	2014	2010	2011
RSP 0-0,1 m (MPa)	X	X	X	X	X	X	X	
RSP 0,1-0,2 m (MPa)	X	X	X	X	X	X	X	
RSP 0,2-0,3 m (MPa)	X	X	X	X	X	X	X	
pH	X			X			X	
Altitude (m)	X			X			X	
Declividade (˚)	X						X	
Densidade (g cm^{-3})	X						X	
Areia (%)	X			X			X	
Silte (%)	X			X			X	
Argila (%)	X			X			X	
Produtividade soja (t ha^{-1})	X	X	X	X	X	X	X	X
Produtividade milho (t ha^{-1})					X	X		

RSP: resistência mecânica do solo à penetração.

As amostras de solo foram coletadas à profundidade de 0-0,2 m por meio do uso de um perfurador de solo Stihl BT 45. A resistência mecânica do solo à penetração (RSP) foi determinada para as profundidades de 0-0,1 m, 0,1-0,2 m e 0,2-0,3 m, utilizando-se um medidor eletrônico de compactação do solo Falker PenetroLOG PLG1020. Os valores da altitude foram obtidos com o uso de uma estação total eletrônica Topcon GPT-7505. Posteriormente, foram calculados os valores de declividade em função da altitude dos pontos amostrais, com o uso do software Surfer.

Os valores da produtividade de soja para a área A foram determinados por um monitor de colheita CASE AFS PRO 600 acoplado a uma colhedora CASE IH 2388. Para as

áreas B e C, a produtividade foi determinada por meio da colheita manual de uma área de amostragem de 0,9 m^2 em cada um dos pontos amostrais. Em todos os casos, os valores da produtividade foram corrigidos para um teor de água de 13%.

Com o intuito de atender ao requisito de estabilidade dos dados da produtividade, que normalmente são bastante influenciados pelo clima e pela precipitação pluviométrica, realizou-se a padronização dos valores amostrais da produtividade de cada ano, empregando-se a técnica de escore padrão (Equação 1) (LARSCHEID; BLACKMORE, 1996). Em seguida, calculou-se a média aritmética dos valores padronizados dos anos disponíveis, gerando-se assim uma única variável correspondente à média das produtividades padronizadas (por simplicidade, chamada neste trabalho de produtividade média).

$$P_{iN} = \frac{(P_i - \overline{P})}{S} \qquad \text{Eq. (1)}$$

em que: P_{iN} é o valor padronizado da produtividade para o ponto amostral i; P_i é o valor original da produtividade no ponto amostral i; $\overline{P}$ corresponde à média aritmética dos valores originais da produtividade nos pontos amostrais; e *S* corresponde ao desvio padrão dos valores originais da produtividade.

5.2.2 Seleção de variáveis

Avaliaram-se cinco métodos de seleção de variáveis baseados na análise de correlação espacial, em ACP e em MULTISPATI-PCA, além de uma abordagem que consistiu simplesmente no uso de todas as variáveis estáveis:

1. All-Attrib: utilização de todas as variáveis estáveis disponíveis;
2. Spatial-Matrix: após calcular a estatística de autocorrelação espacial bivariada de Moran entre todas as variáveis usando o Software para Definição de Unidades de Manejo (SDUM) (BAZZI et al., 2013), selecionaram-se variáveis pelo procedimento proposto por Bazzi et al. (2013): a) eliminação das variáveis com autocorrelação espacial não significativa ao nível de 5% de significância; b) remoção das variáveis que não apresentaram correlação espacial significativa com a produtividade média; c) ordenação decrescente das variáveis restantes, considerando o módulo do valor da correlação com a produtividade média; e d) eliminação de variáveis que estavam correlacionadas com outras, dando preferência para a remoção das que apresentaram menor correlação com a produtividade média;
3. PCA-All (ACP tradicional): obtenção das CPs a partir de todas as variáveis estáveis, tal que para definir a quantidade de CPs a serem utilizadas, seguiu-se o critério da representação de ao menos 70% da variabilidade dos dados das variáveis originais (FERREIRA, 1996);

4. MPCA-All (MULTISPATI-PCA tradicional): obtenção das CPs, que no caso de MULTISPATI-PCA também são chamadas de componentes principais espaciais (CPEs), a partir de todas as variáveis estáveis; o número de CPEs selecionadas também foi baseado no critério da representação de ao menos 70% da variabilidade total dos dados originais;
5. PCA-SC: execução da ACP com os mesmos parâmetros do método PCA-All, porém aplicada somente sobre as variáveis estáveis que apresentaram correlação espacial significativa ao nível de 5% com a produtividade média;
6. MPCA-SC: novo método proposto, baseado na execução de MULTISPATI-PCA com os mesmos parâmetros de MPCA-All, mas aplicado apenas sobre as variáveis estáveis que apresentaram correlação espacial significativa ao nível de 5% com a produtividade média.

Os métodos PCA-All, MPCA-All, PCA-SC e MPCA-SC foram aplicados sobre os dados de cada área por meio do desenvolvimento de rotinas no software estatístico R (R CORE TEAM, 2016). Nessas rotinas, incluíram-se os pacotes geoR, gstat, ade4 (CHESSEL; DUFOUR; THIOULOUSE, 2004) e spdep (BIVAND, 2012).

Do pacote spdep, utilizou-se a função dnearneigh para identificar os vizinhos de cada ponto amostral, que foram necessários para a execução dos métodos MPCA-All e MPCA-SC. Essa função emprega a distância euclidiana para computar a distância de cada ponto em relação aos demais e gera como resultado uma lista de vizinhos de cada ponto. Ela se baseia no valor definido como raio de vizinhança, uma distância determinada para cada área agrícola por meio de um processo iterativo. No caso deste trabalho, foram definidos os seguintes raios: para a área A, 240 m; para a área B, 120 m; e para a área C, 200 m.

O sistema gerenciador de banco de dados PostgreSQL 9.0.5 (PostgreSQL Global Development Group) foi usado para o armazenamento de dados. O software PostGIS 1.5.5 (PostGIS Project Steering Committee), uma extensão para bancos de dados espaciais do PostgreSQL, também foi empregado. Além disso, o software pgAdmin III (pgAdmin Development Team) foi utilizado para a administração das bases de dados criadas.

5.2.3 Interpolação e agrupamento de dados

Para os dados das três áreas consideradas, notou-se que o método de interpolação espacial krigagem ordinária produziria os melhores resultados. Todavia, a vantagem de usar a krigagem ordinária ao invés do interpolador inverso da distância ao quadrado seria pequena. Além disso, o software SDUM, que era o único de uso gratuito capaz de interpolar, definir e avaliar ZMs, tem a limitação de ainda não disponibilizar a interpolação por krigagem. Diante disso, decidiu-se utilizar o SDUM para interpolar os dados das variáveis selecionadas pelo método inverso da distância ao quadrado, com pixels representando áreas de 5x5 m.

Como as escalas de medidas das variáveis consideradas eram diferentes, realizou-se a padronização de seus valores antes de passarem pela etapa de interpolação, empregando-se uma versão da técnica da amplitude (Equação 2) (MIELKE JR; BERRY, 2007). Com isso, manteve-se a mesma amplitude para os dados, independentemente da variável empregada.

$$P_{iN} = \frac{P_i - P_{\min}}{P_{\max} - P_{\min}} \qquad \text{Eq. (2)}$$

em que: P_{iN} é o valor padronizado da produtividade para o ponto amostral i; P_i é o valor original da produtividade para o ponto i; e $P_{\max}$ e $P_{\min}$ correspondem, respectivamente, aos valores amostrais máximo e mínimo da produtividade no conjunto de dados considerado.

Após a interpolação espacial, os dados resultantes foram usados como entrada para o algoritmo de agrupamento fuzzy c-means, com parâmetro de erro igual a 0,0001 e índice de ponderação igual a 1,3 (valores sugeridos em diversos trabalhos). Assim, foram geradas duas, três e quatro classes para cada área. O software SDUM também foi utilizado para executar o método fuzzy c-means e gerar os mapas temáticos das ZMs correspondentes às classes.

5.2.4 Avaliação dos métodos de seleção de variáveis

Avaliou-se o desempenho dos métodos por meio da aplicação de seis critérios:

1. Redução da variância (variance reduction – VR) (PING; DOBERMANN, 2003): é calculado para a produtividade média, com a expectativa de que o somatório das variâncias dos dados das ZMs seja menor que a variância da área como um todo (Equação 3):

$$VR = \left(1 - \frac{\sum_{i=1}^{c} W_i * V_{zmi}}{V_{área}} \right) * 100 \qquad \text{Eq. (3)}$$

   em que: *c* é a quantidade de ZMs; W_i é a proporção da área total referente à i-ésima ZM; V_{zmi} é a variância dos dados da i-ésima ZM; $V_{área}$ é a variância dos dados da área como um todo.

2. Índice de desempenho fuzzy (fuzziness performance index - FPI) (FRIDGEN et al., 2004): permite determinar o grau de separação entre os grupos difusos gerados por fuzzy c-means; seu valor varia entre 0 e 1, tal que quanto mais próximo for de 0, menor será o grau de compartilhamento de elementos entre os grupos gerados (Equação 4):

$$FPI = 1 - \frac{c}{(c-1)}\left[1 - \sum_{j=1}^{n}\sum_{i=1}^{c}(m_{ij})^2 / n\right] \quad \text{Eq. (4)}$$

em que: *c* é a quantidade de grupos; *n* é a quantidade de elementos no conjunto de dados e m_{ij} é o valor correspondente ao grau de pertinência do j-ésimo elemento do conjunto em relação ao i-ésimo grupo.

3. Entropia de partição modificada (modified partition entropy - MPE) (BOYDELL; MCBRATNEY, 2002): é uma estimativa do nível de dificuldade para a organização dos grupos gerados por fuzzy c-means, tal que quanto mais próximo de 0 for seu valor, menor terá sido essa dificuldade (Equação 5):

$$MPE = \frac{-\sum_{j=1}^{n}\sum_{i=1}^{c} m_{ij}\log(m_{ij})/n}{\log c} \quad \text{Eq. (5)}$$

em que: *c* é a quantidade de grupos; *n* é a quantidade de elementos no conjunto de dados e m_{ij} é o valor correspondente ao grau de pertinência do j-ésimo elemento do conjunto em relação ao i-ésimo grupo.

4. Índice de suavidade (smoothness index – SI): indica a frequência de mudanças de ZMs, pixel a pixel, nas direções horizontal e vertical em um mapa temático, assim como nas diagonais. Também representa a suavidade das curvas de contorno das zonas. Varia entre 0 e 100%, tal que caso um mapa corresponda a uma área completamente homogênea, o resultado será 100%, devido à ausência de mudanças de subáreas. Mas se as ZMs forem bastante fragmentadas, SI apresentará um valor próximo de 0% (Equação 6):

$$SI = 100 - \left(\left(\frac{\sum_{i=1}^{k} NM_{Hi}}{4P_H} + \frac{\sum_{j=1}^{k} NM_{Vj}}{4P_V} + \frac{\sum_{l=1}^{k} NM_{Ddl}}{4P_{Dd}} + \frac{\sum_{m=1}^{k} NM_{Dem}}{4P_{De}}\right) * 100\right) \quad \text{Eq. (6)}$$

em que: *k* é o número de linhas, colunas ou diagonais; NM_{Hi} é o número de mudanças na linha *i* (horizontal); NM_{Vj} é o número de mudanças na coluna *j* (vertical); NM_{Ddl} é o número de mudanças na diagonal *l* (diagonal direita *Dd*); NM_{Dem} é o número de mudanças na diagonal *m* (diagonal esquerda *De*); P_H é a possibilidade de mudanças de pixels na horizontal; P_V é a possibilidade de mudanças de pixels na vertical; P_{Dd} é a possibilidade de mudanças na diagonal direita *Dd*; e P_{De} é a possibilidade de mudanças na diagonal esquerda *De*.

5. Análise de variância: a ANOVA e o teste de comparação de médias de Tukey foram empregados para constatar se as ZMs geradas apresentaram diferenças de produtividade média estatisticamente significativas, ao nível de significância de 5%. Antes de aplicá-los, confirmou-se que não havia dependência espacial entre as amostras dentro de cada ZM.

6. Índice de validação de grupos aprimorado (Improved Cluster Validation Index – ICVI): baseado no índice CVI (SCHENATTO et al., 2016), o ICVI é proposto neste trabalho (Equação 7) para resolver um problema que pode ocorrer quando os valores dos índices VR, FPI e MPE não indicam a mesma abordagem de definição de ZMs como a melhor solução. O ICVI varia entre 0 e 1, tal que quanto maior for o valor de VR e menores forem os valores de FPI e MPE, mais próximo de 0 será o valor deste novo índice. Em uma comparação entre *n* abordagens de geração de ZMs que utilizem um mesmo algoritmo de agrupamento difuso (como fuzzy c-means, por exemplo), a melhor será aquela que apresentar o menor ICVI.

$$ICVI_i = \frac{1}{3} * \left(\frac{FPI_i}{Max\{FPI\}} + \frac{MPE_i}{Max\{MPE\}} + \left(1 - \frac{VR_i}{Max\{VR\}} \right) \right) \quad \text{Eq. (7)}$$

em que: FPI_i, MPE_i e VR_i são, respectivamente, os valores de FPI, MPE e VR da i-ésima abordagem de definição de ZMs, e $Max\{X\}$ representa o maior valor do índice *X* entre as *n* abordagens comparadas.

5.3 Resultados e discussão

5.3.1 Variáveis selecionadas

As variáveis selecionadas para a definição das classes e os valores da correlação espacial bivariada de Moran entre cada variável e a produtividade média são apresentados na Tabela 2. Como os valores da correlação espacial bivariada de Moran não estão padronizados, mesmo pequenos valores podem ser estatisticamente significativos (CZAPLEWSKI; REICH, 1993). Os valores foram considerados relevantes quando estatisticamente significativos ao nível de 5%.

Verificou-se que a altitude foi a variável com correlação espacial mais forte em relação à produtividade média, para as três áreas. Esta constatação está alinhada com as que foram apresentadas por Jaynes, Colvin e Kaspar (2005) e Peralta et al. (2013), que sugeriram que há associação espacial entre a altitude e a produtividade de soja e milho.

De acordo com o método Spatial-Matrix, as variáveis selecionadas para as áreas A e B foram altitude e RSP 0–0,1 m, enquanto que para a área C somente a altitude foi selecionada.

Tabela 2 Variáveis escolhidas por meio do uso de cada um dos seis métodos de seleção, e valores da estatística de correlação espacial bivariada de Moran com a produtividade média, para cada área

Área	Variáveis	CM com prod. média	AS	CPM	NR	Métodos de seleção de variáveis					
						All-Attrib	Spatial Matrix	PCA-All	MPCA-All	PCA-SC	MPCA-SC
A	RSP 0-0,1 m	-0,053*	S	S	S	S	S	S	S	S	S
	RSP 0,1-0,2 m	-0,017	N	N	N	S	N	S	S	N	N
	RSP 0,2-0,3 m	-0,022	N	N	N	S	N	S	S	N	N
	pH	-0,034*	N	S	N	S	N	S	S	S	S
	Altitude	0,100*	S	S	S	S	S	S	S	S	S
	Declividade	-0,016	N	N	N	S	N	S	S	N	N
	Densidade	0,023	N	N	N	S	N	S	S	N	N
	Areia	-0,075*	S	S	N	S	N	S	S	S	S
	Silte	0,028	N	N	N	S	N	S	S	N	N
	Argila	-0,040*	S	S	N	S	N	S	S	S	S
B	RSP 0-0,1 m	0,039*	S	S	S	S	S	S	S	S	S
	RSP 0,1-0,2 m	0,044*	N	S	N	S	N	S	S	S	S
	RSP 0,2-0,3 m	-0,014	N	N	N	S	N	S	S	N	N
	pH	-0,029*	N	S	N	S	N	S	S	S	S
	Altitude	0,051*	S	S	S	S	S	S	S	S	S
	Areia	0,007	N	N	N	S	N	S	S	N	N
	Silte	-0,013	S	N	N	S	N	S	S	N	N
	Argila	0,012	S	N	N	S	N	S	S	N	N
C	RSP 0-0,1 m	-0,002	N	N	N	S	N	S	S	N	N
	RSP 0,1-0,2 m	0,114*	S	S	N	S	N	S	S	S	S
	RSP 0,2-0,3 m	0,102*	S	S	N	S	N	S	S	S	S
	pH	0,024	N	N	N	S	N	S	S	N	N
	Altitude	0,137*	S	S	S	S	S	S	S	S	S
	Declividade	0,011	S	N	N	S	N	S	S	N	N
	Densidade	-0,029	N	N	N	S	N	S	S	N	N
	Areia	0,078*	S	S	N	S	N	S	S	S	S
	Silte	0,021	N	N	N	S	N	S	S	N	N
	Argila	-0,082*	S	S	N	S	N	S	S	S	S

*: valor significativo a 5%; RSP: resistência mecânica do solo à penetração; CM: correlação espacial bivariada de Moran; AS: autocorrelação espacial significativa; CPM: correlação significativa com a produtividade média; NR: variável não redundante; S: sim; N: não.

5.3.2 Componentes principais

Pode-se notar na Tabela 3 que, ao considerar todas as variáveis temporalmente estáveis para a obtenção das CPs, as quantidades necessárias dessas componentes foram iguais ou maiores do que aquelas correspondentes às situações em que foram consideradas somente as variáveis com correlação espacial significativa com a produtividade média (variáveis com valor igual a “S” na Tabela 2, para PCA-SC e MPCA-SC). Isto sugere que variáveis não correlacionadas espacialmente com a produtividade podem prejudicar a construção de CPs destinadas à geração de ZMs.

Comparando-se os quatro métodos baseados em ACP ou MULTISPATI-PCA, ou seja, PCA-All, PCA-SC, MPCA-All e MPCA-SC, constatou-se que MPCA-SC obteve o melhor desempenho na redução da dimensionalidade dos dados sem perda significativa de informação. Isto porque este método garantiu as maiores porcentagens acumuladas de representação da variância original com menor quantidade de CPs. Com MPCA-SC, apenas as duas primeiras componentes foram necessárias para cada uma das áreas, enquanto que os outros três métodos exigiram até cinco componentes (Tabela 3).

Tabela 3 Estatísticas das componentes principais necessárias para representar no mínimo 70% da variância total dos dados originais, geradas com PCA-All, MPCA-All, PCA-SC e MPCA-SC, para as três áreas

Área	Método / Variáveis	Variância	% da variância total dos dados	Soma das % da variância	AM
	PCA-All				
	CP1	2,98	27	27	0,23
	CP2	2,57	23	50	0,15
	CP3	1,50	14	64	-0,05
	CP4	1,15	10	74	-0,05
	MPCA-All				
A	CPE1	2,81	53	53	0,29
	CPE2	2,45	47	100	0,15
	PCA-SC				
	CP1	2,94	49	49	0,22
	CP2	1,27	21	70	0,09
	MPCA-SC				
	CPE1	2,77	71	71	0,25
	CPE2	1,11	29	100	0,13
	PCA-All				
	CP1	3,20	32	32	0,01
	CP2	1,93	19	51	0,01
	CP3	1,33	13	64	0,07
	CP4	1,18	12	76	0,03
	MPCA-All				
	CPE1	1,66	35	35	0,19
B	CPE2	1,50	32	67	0,11
	CPE3	0,68	15	82	0,08
	PCA-SC				
	CP1	2,56	43	43	0,03
	CP2	1,34	22	65	0,11
	CP3	0,92	15	80	-0,05
	MPCA-SC				
	CPE1	1,67	61	61	0,19
	CPE2	0,64	23	84	0,05
	PCA-All				
	CP1	3,44	31	31	0,34
	CP2	1,40	13	44	0,03
	CP3	1,27	12	56	0,22
	CP4	1,10	10	66	-0,02
	CP5	0,99	9	75	0,03
	MPCA-All				
	CPE1	3,07	48	48	0,44
C	CPE2	1,31	21	69	0,24
	CPE3	1,14	18	87	0,06
	PCA-SC				
	CP1	2,87	48	48	0,62
	CP2	1,12	19	67	0,36
	CP3	0,98	16	83	0,10
	MPCA-SC				
	CPE1	2,63	68	68	0,65
	CPE2	1,21	32	100	0,46

AM: autocorrelação espacial de Moran.

Comparando-se PCA-All a MPCA-All, ou PCA-SC a MPCA-SC, do ponto de vista da variância e da autocorrelação espacial (Tabela 3), a primeira componente principal espacial (CPE1) apresentou menor variância, porém maior autocorrelação espacial, do que a primeira componente principal (CP1), para as três áreas. Isto indica que os valores da autocorrelação espacial tendem a ser maiores quando se usa um método baseado em MULTISPATI-PCA.

Portanto, a abordagem MULTISPATI-PCA facilitou a seleção das CPs necessárias para a geração de ZMs para as áreas consideradas.

Córdoba et al. (2012, 2013) obtiveram resultados similares a esses. Embora um método equivalente a MPCA-SC não tenha sido empregado, eles executaram PCA-All e MPCA-All a partir das variáveis altitude, RSP e condutividade elétrica aparente do solo, avaliadas em áreas da Argentina. Nos dois trabalhos, destacou-se que MULTISPATI-PCA simplificou a seleção das CPs imprescindíveis para a definição de ZMs.

Na análise dos coeficientes das CPs, que atuam como ponderações sobre as variáveis originais que formam essas componentes (Tabelas 4, 5 e 6), notou-se que a primeira componente (CP1 ou CPE1) apresentou coeficientes de ponderação maiores (em valores absolutos) para as seguintes variáveis: altitude, argila e areia para a área A; altitude e RSP 0,1-0,2 m para a área B; e altitude, argila e RSP 0,1-0,2 m para a área C.

Tabela 4 Ponderações correspondentes às variáveis utilizadas na formação das CPs, para a área A

Variáveis	Altit.	RSP 0-0,1	pH	Argila	Areia	Silte	Decliv.	Dens.	RSP 0,1-0,2	RSP 0,2-0,3
PCA-All										
CP1	0,49	-0,26	-0,39	0,45	-0,49	-0,07	-0,12	0,07	-0,05	0,01
CP2	0,07	-0,41	0,12	-0,30	-0,04	0,46	0,04	-0,10	-0,49	-0,48
CP3	0,25	-0,03	0,23	-0,20	-0,07	0,35	0,49	0,43	0,38	0,25
CP4	-0,16	-0,38	-0,18	0,07	-0,02	-0,08	0,62	-0,58	0,09	0,19
MPCA-All										
CPE1	0,53	0,02	-0,26	0,44	-0,49	-0,18	-0,22	0,20	0,15	0,22
CPE2	0,45	-0,48	0,01	-0,14	-0,17	0,38	0,14	0,05	-0,29	-0,46
PCA-SC										
CP1	0,50	-0,29	-0,38	0,43	-0,50					
CP2	0,08	-0,55	0,24	-0,47	0,16					
MPCA-SC										
CPE1	0,56	-0,07	-0,26	0,52	-0,54					
CPE2	-0,39	0,72	-0,13	0,52	-0,01					

Altit.: altitude; RSP: resistência mecânica do solo à penetração; Decliv.: declividade; Dens.: densidade.

Tabela 5 Ponderações correspondentes às variáveis utilizadas na formação das CPs, para a área B

Variáveis	Altitude	RSP 0-0,1	RSP 0,1-0,2	pH	RSP 0,2-0,3	Areia	Argila	Silte
PCA-All								
CP1	-0,36	-0,29	-0,43	0,22	-0,29	-0,26	-0,31	0,34
CP2	0,10	0,19	0,19	-0,42	0,22	-0,30	-0,51	0,53
CP3	0,37	-0,21	-0,03	0,32	0,21	0,07	-0,22	0,18
CP4	0,37	0,61	0,04	0,30	-0,51	0,24	-0,18	0,15
MPCA-All								
CPE1	-0,76	0,07	-0,31	-0,02	0,07	-0,12	-0,01	0,07
CPE2	-0,09	0,04	0,44	-0,18	0,85	-0,13	-0,04	0,12
CPE3	0,18	-0,14	-0,16	-0,40	0,22	0,59	0,11	-0,27
PCA-SC								
CP1	-0,43	-0,43	-0,51	0,35				
CP2	0,42	-0,10	0,02	0,46				
CP3	0,27	-0,65	-0,08	-0,54				
MPCA-SC								
CPE1	-0,76	0,05	-0,35	-0,02				
CPE2	-0,43	-0,09	-0,05	0,60				

RSP: resistência mecânica do solo à penetração.

Tabela 6 Ponderações correspondentes às variáveis utilizadas na formação das CPs, para a área C

Variáveis	Altit.	RSP 0,1-0,2	RSP 0,2-0,3	Argila	Areia	RSP 0-0,1	Silte	Decliv.	Dens.	pH
PCA-All										
CP1	-0,40	-0,45	-0,42	0,27	-0,30	0,01	-0,24	0,04	0,39	-0,06
CP2	0,25	-0,19	-0,22	-0,01	-0,13	-0,63	0,07	0,42	0,05	-0,08
CP3	0,11	0,20	0,19	0,48	-0,51	-0,09	-0,03	-0,35	0,10	0,48
CP4	-0,06	0,29	0,34	-0,16	-0,14	-0,17	-0,49	-0,18	0,36	-0,55
CP5	-0,11	0,01	0,10	0,39	0,12	-0,41	0,54	-0,31	-0,05	-0,42
MPCA-All										
CPE1	-0,41	-0,32	-0,30	0,53	-0,36	0,07	-0,22	-0,15	0,27	-0,01
CPE2	-0,38	-0,37	-0,31	-0,66	0,27	0,01	0,03	0,15	0,11	-0,08
CPE3	0,03	-0,17	-0,12	0,09	-0,25	0,11	0,07	0,85	0,09	-0,04
PCA-SC										
CP1	-0,43	-0,52	-0,49	0,32	-0,30					
CP2	-0,16	-0,20	-0,20	-0,58	0,66					
CP3	0,41	-0,39	-0,49	-0,24	-0,05					
MPCA-SC										
CPE1	-0,44	-0,35	-0,32	0,58	-0,39					
CPE2	-0,39	-0,37	-0,31	-0,66	0,28					

Altit.: altitude; RSP: resistência mecânica do solo à penetração; Decliv.: declividade; Dens.: densidade.

Portanto, a variável altitude foi a única que apresentou coeficientes de ponderação em CP1 e CPE1 que ficaram entre os maiores valores, para as três áreas. Esse resultado da influência da altitude sobre CP1 é similar aos obtidos por Fraisse, Sudduth e Kitchen (2001), que aplicaram ACP para delinear ZMs para duas áreas com cultivo de milho e soja nos Estados Unidos da América. Saleh e Belal (2014) também aplicaram ACP sobre dados de uma área localizada no Egito e obtiveram resultados similares em relação à influência da altitude sobre CP1.

A influência da variável argila sobre CP1 também foi observada por Moral, Terrón e Silva (2010), que utilizaram ACP para definir ZMs em uma área na Espanha. Já a influência da altitude sobre CPE1 também foi detectada por Córdoba et al. (2016) e Peralta et al. (2015), na geração de subáreas para diversas áreas com plantio de trigo na Argentina.

5.3.3 Mapas temáticos de zonas de manejo

Para cada área, as ZMs delineadas apresentaram diferenças de formato e extensão, de acordo com o método de seleção de variáveis utilizado conjuntamente com o algoritmo fuzzy c-means (Figuras 2, 3 e 4). Considerando a aplicação do método All-Attrib no processo de geração de três ou quatro classes para a área A, as operações de campo seriam de difícil execução para ao menos uma das ZMs correspondentes às classes. Isto em razão do tamanho pequeno e do formato da ZM. O mesmo problema ocorreu no caso da aplicação de Spatial-Matrix na definição de quatro classes para a área C. Além disso, All-Attrib não poderia ser empregado para o delineamento de três ou quatro ZMs na área C, pois não foi possível obter essas quantidades de subáreas. Por outro lado, não ocorreram problemas similares a esses quando foram utilizados os métodos PCA-All, PCA-SC, MPCA-All e MPCA-SC.

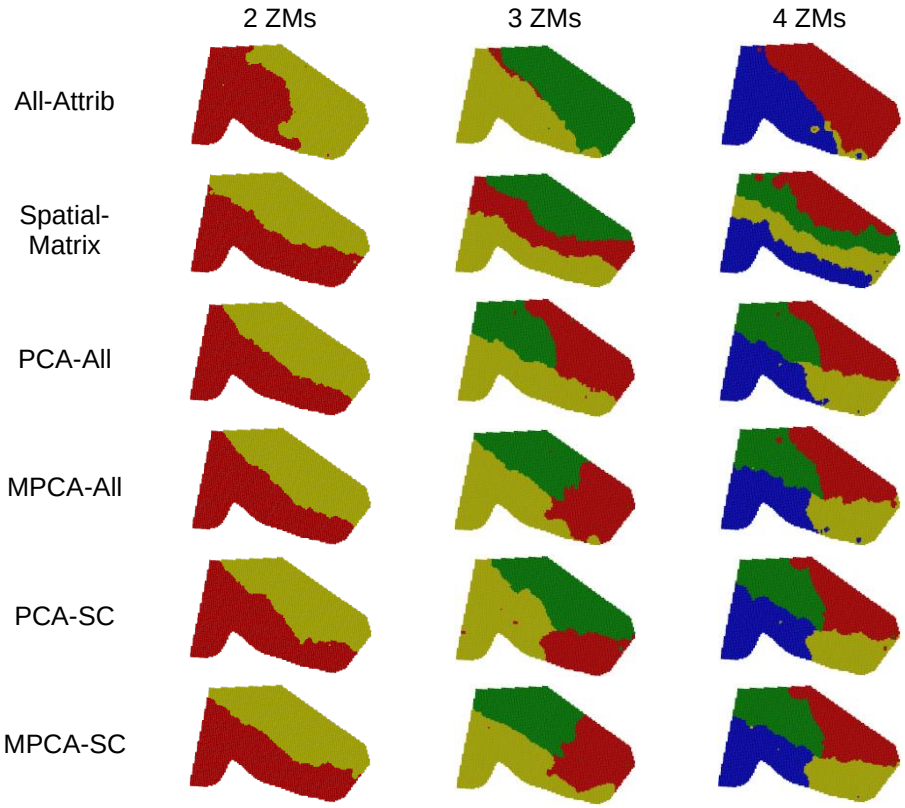


Figura 2 Mapas temáticos com 2, 3 e 4 zonas de manejo para a área A, gerados com a execução dos seis métodos de seleção de variáveis e do algoritmo de agrupamento fuzzy c-means.

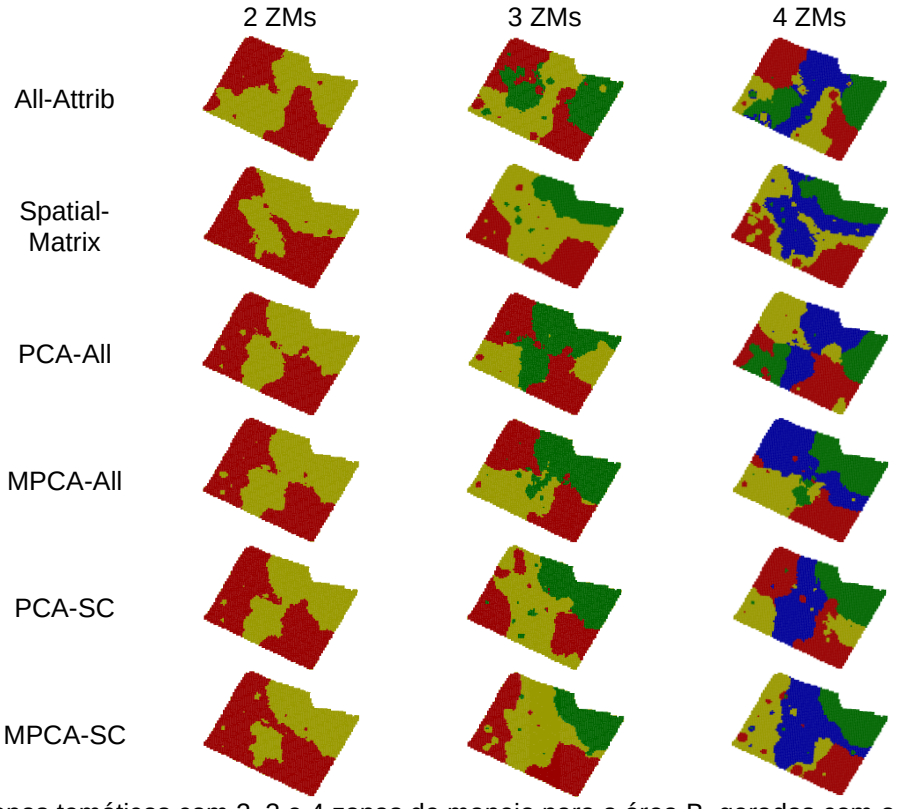


Figura 3 Mapas temáticos com 2, 3 e 4 zonas de manejo para a área B, gerados com a execução dos seis métodos de seleção de variáveis e do algoritmo de agrupamento fuzzy c-means.

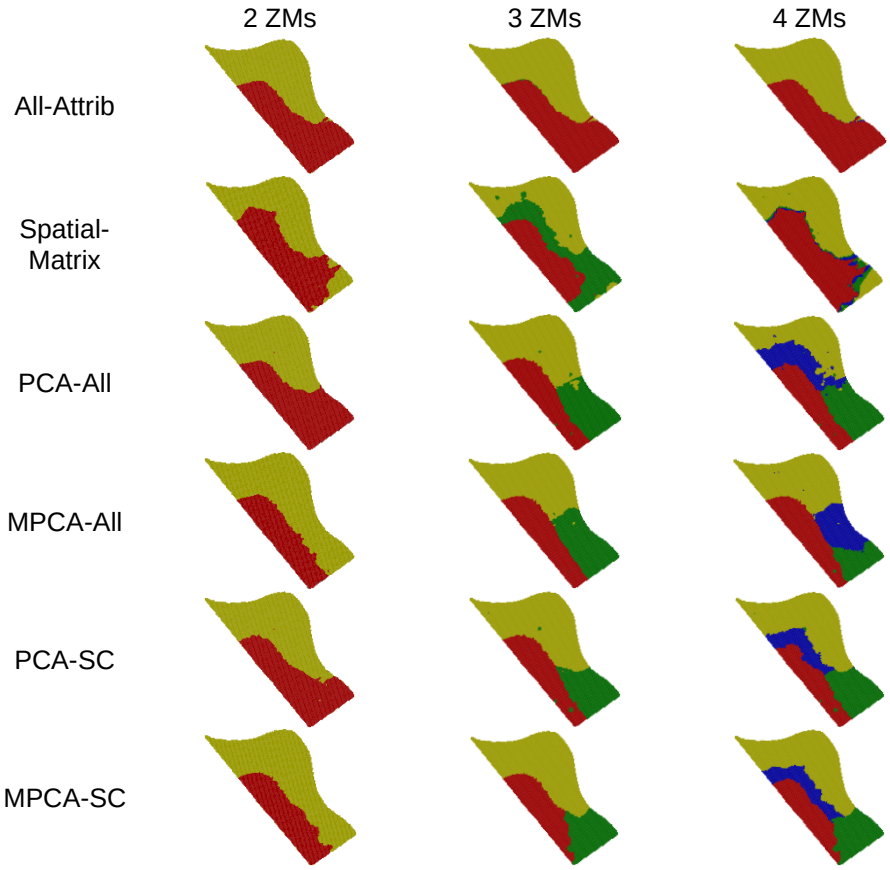


Figura 4 Mapas temáticos com 2, 3 e 4 zonas de manejo para a área C, gerados com a execução dos seis métodos de seleção de variáveis e do algoritmo de agrupamento fuzzy c-means.

Os resultados das avaliações das ZMs geradas, considerando-se a ANOVA (teste de Tukey) e os critérios VR, FPI, MPE, SI e ICVI (Tabela 7), mostraram que é possível dividir as três áreas em duas ZMs com potenciais de produtividade estatisticamente diferentes a 5% de significância. Para a área B, foi possível obter esse resultado somente com o uso dos métodos MPCA-All e MPCA-SC.

O método MPCA-SC obteve com maior frequência os melhores resultados para o índice VR. Em outras palavras, este método identificou ZMs com as maiores diferenças entre as respectivas produtividades médias e com os menores valores de desvio padrão. Diferenças entre ZMs em relação à produtividade média indicam que existem condições do solo influenciando a resposta das culturas plantadas. Como mencionado anteriormente, a variável altitude exerceu a maior influência sobre a componente principal CPE1. Portanto, trata-se de uma variável fundamental para os resultados obtidos com MPCA-SC. Córdoba et al. (2013) e Peralta et al. (2015) também constataram a influência da altitude sobre a primeira CPE, em seus experimentos com MPCA-All.

O índice de suavidade (SI) foi utilizado para avaliar a suavidade das linhas que delimitam as ZMs (Tabela 7). Nessa avaliação, confirmou-se que MPCA-SC foi o método que, com maior frequência, proporcionou os melhores resultados para as áreas – isto porque MPCA-SC geralmente conduziu a ZMs com delineamentos satisfatórios para operações de campo.

Tabela 7 Resultados para a ANOVA (teste de Tukey), VR, FPI, MPE, SI e ICVI, para as três áreas

Área	Classes	Métodos	ANOVA (Teste de Tukey)				VR(%)	FPI	MPE	SI(%)	ICVI
			C_1	C_2	C_3	C_4					
A	2	All-Attrib	a	a			0,0	0,500	0,079	98,4	1
		Spatial-Matrix	a	b			42,7	0,091	0,018	98,3	0,137
		PCA-All	a	b			42,5	0,185	0,035	98,5	0,273
		MPCA-All	a	b			25,5	0,161	0,030	98,6	0,368
		PCA-SC	a	b			24,4	0,177	0,032	98,4	0,396
		MPCA-SC	a	b			28,8	0,153	0,029	98,6	0,333
	3	All-Attrib	a	a	a		0,0	0,667	0,125	97,7	1
		Spatial-Matrix	a	b	b		22,6	0,156	0,032	96,8	0,307
		PCA-All	a	a	b		39,8	0,287	0,058	97,6	0,298
		MPCA-All	a	a	b		16,7	0,212	0,043	97,5	0,414
		PCA-SC	a	b	a		28,4	0,200	0,042	97,7	0,307
		MPCA-SC	a	b	b		33,6	0,210	0,043	97,7	0,272
	4	All-Attrib	a	a	a	a	0,0	0,750	0,158	97,1	1
		Spatial-Matrix	a	b	b	a	39,1	0,213	0,044	95,0	0,254
		PCA-All	a	b	b	a	28,1	0,314	0,069	96,9	0,427
		MPCA-All	a	ab	b	a	20,8	0,215	0,048	96,5	0,388
		PCA-SC	a	b	a	b	48,9	0,178	0,038	97,0	0,159
		MPCA-SC	a	a	b	b	33,7	0,182	0,041	97,2	0,271
B	2	All-Attrib	a	a			4,1	0,285	0,054	95,7	0,908
		Spatial-Matrix	a	a			5,2	0,146	0,029	95,5	0,573
		PCA-All	a	a			1,7	0,292	0,054	95,7	0,965
		MPCA-All	a	b			15,1	0,255	0,048	95,8	0,612
		PCA-SC	a	a			0,0	0,234	0,045	95,8	0,878
		MPCA-SC	a	b			16,3	0,161	0,032	95,8	0,381
	3	All-Attrib	a	a	a		8,5	0,667	0,132	91,7	0,919
		Spatial-Matrix	a	a	a		11,6	0,153	0,034	94,7	0,385
		PCA-All	a	a	a		2,2	0,357	0,076	94,3	0,683
		MPCA-All	a	ab	b		21,7	0,333	0,071	94,0	0,472
		PCA-SC	a	a	a		17,9	0,327	0,069	93,6	0,500
		MPCA-SC	a	b	a		34,9	0,176	0,038	94,7	0,184
	4	All-Attrib	a	b	ab	ab	22,8	0,536	0,119	91,2	0,774
		Spatial-Matrix	a	ab	b	ab	15,3	0,239	0,052	89,8	0,476
		PCA-All	a	a	a	a	7,7	0,415	0,095	93,1	0,781
		MPCA-All	a	ab	b	ab	-0,2	0,290	0,068	92,9	0,704
		PCA-SC	a	a	b	a	21,2	0,316	0,073	93,6	0,525
		MPCA-SC	a	a	b	a	33,7	0,205	0,046	93,8	0,256
C	2	All-Attrib	a	b			19,4	0,500	0,077	98,8	0,797
		Spatial-Matrix	a	b			31,9	0,495	0,076	97,9	0,659
		PCA-All	a	b			26,9	0,206	0,037	99,0	0,350
		MPCA-All	a	b			23,2	0,162	0,030	98,6	0,329
		PCA-SC	a	b			28,6	0,150	0,027	98,7	0,251
		MPCA-SC	a	b			23,7	0,117	0,021	98,6	0,255
	3	All-Attrib	a	a	a		0,0	0,667	0,122	98,3	1
		Spatial-Matrix	a	a	b		28,0	0,108	0,023	96,4	0,157
		PCA-All	a	b	a		25,4	0,189	0,040	97,8	0,271
		MPCA-All	a	b	b		20,1	0,147	0,031	98,2	0,281
		PCA-SC	a	b	a		28,9	0,127	0,027	97,9	0,168
		MPCA-SC	a	a	b		31,8	0,085	0,017	98,4	0,089
	4	All-Attrib	a	a	a	a	0,0	0,750	0,154	98,1	1
		Spatial-Matrix	a	b	bc	ac	35,9	0,535	0,111	94,7	0,512
		PCA-All	a	ab	b	c	26,6	0,286	0,061	96,3	0,371
		MPCA-All	a	b	ac	bc	26,4	0,166	0,037	97,6	0,267
		PCA-SC	a	b	a	a	38,5	0,175	0,039	97,1	0,175
		MPCA-SC	a	b	c	a	40,0	0,146	0,032	97,3	0,134

C_i: classe *i*.

Na Figura 5, apresentam-se gráficos com os valores dos índices FPI, MPE, ICVI e VR, obtidos por cada um dos seis métodos avaliados. Analisando os valores do FPI e da MPE, observou-se que o método MPCA-SC proporcionou o melhor desempenho ao algoritmo fuzzy

c-means durante a definição das ZMs. Isto em razão desse método ter sido o que mais conduziu FPI e MPE a seus melhores valores (valores mais próximos de 0). Por conseguinte, MPCA-SC também foi o que mais se destacou em relação aos valores do índice ICVI, que é definido a partir de VR, FPI e MPE.

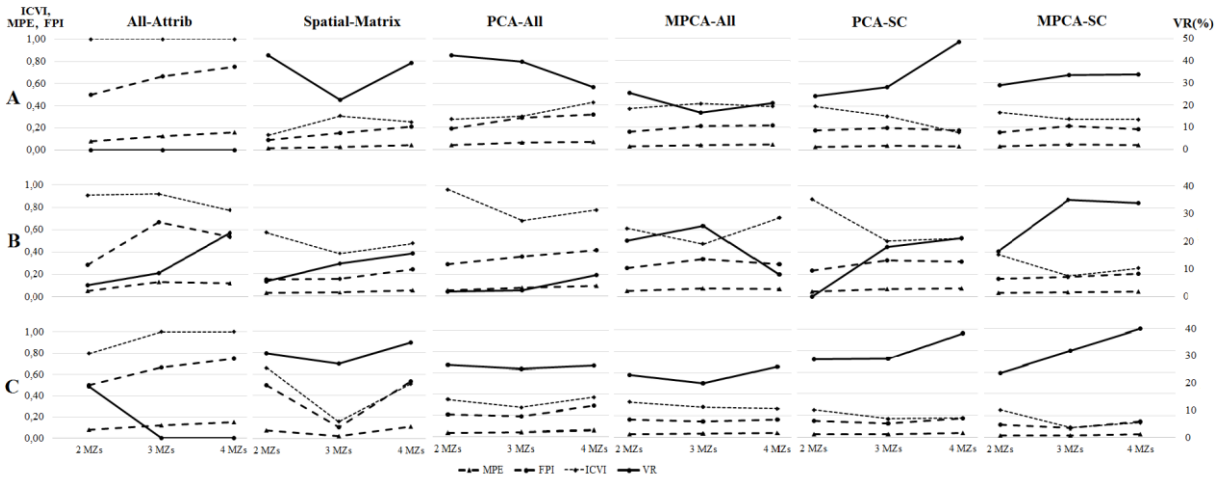


Figura 5 Gráficos para os índices FPI, MPE, ICVI e VR, para os seis métodos de seleção de variáveis avaliados, considerando duas, três e quatro ZMs para cada área.

A análise conjunta dos resultados do FPI, da MPE e da ANOVA confirmaram a recomendação de divisão de cada área em duas ZMs, aplicando-se MPCA-SC para definir as variáveis. Córdoba et al. (2016) e Peralta et al. (2015) também empregaram os menores valores obtidos para FPI e MPE, bem como a simplificação de operações de campo, como critérios para escolherem a opção de delineamento de duas ZMs.

5.4 Conclusões

O estudo de caso realizado com os dados das três áreas agrícolas mostrou que o novo método MPCA-SC, que combina análise de correlação espacial com a abordagem de análise multivariada espacial MULTISPATI-PCA, pode melhorar a qualidade de zonas de manejo. Com esse novo algoritmo, as subáreas definidas geralmente apresentaram tamanhos satisfatórios e contornos mais suaves, o que as tornou melhores para a execução de operações de campo. MPCA-SC foi o método que mais conduziu a ZMs com as maiores diferenças entre as respectivas produtividades médias e com os menores valores residuais internos. Ele também promoveu a melhor redução de dimensionalidade dos dados originais, sem perda significativa de informação, para as três áreas.

5.5 Referências

ARROUAYS, D.; SABY, N. P. A.; THIOULOUSE, J.; JOLIVET, C.; BOULONNE, L.; RATIÉ, C. Large trends in French topsoil characteristics are revealed by spatially constrained multivariate analysis. **Geoderma,** v. 161, p. 107-114, 2011.

BAZZI, C. L.; SOUZA, E. G.; URIBE-OPAZO, M. A.; NÓBREGA, L. H. P.; ROCHA, D. M. Management zones definition using soil chemical and physical attributes in a soybean area. **Engenharia Agrícola,** v. 33, n. 5, p. 952-964, 2013.

BIVAND, R. **spdep: Spatial Dependence: Weighting Schemes, Statistics and Models.** Vienna: R Foundation for Statistical Computing, 2012. 10 p.

BOYDELL, B.; MCBRATNEY, A. B. Identifying potential within-field management zones from cotton-yield estimates. **Precision Agriculture,** v. 3, n. 1, p. 9-23, 2002.

CHESSEL, D.; DUFOUR, A. B.; THIOULOUSE, J. The ade4 package-I-one-table methods. **R News**, v. 4, p. 5-10, 2004.

COHEN, S.; COHEN, Y.; ALCHANATIS, V.; LEVI, O. Combining spectral and spatial information from aerial hyperspectral images for delineating homogenous management zones. **Biosystems Engineering,** v. 114, n. 4, p. 435-443, 2013.

CÓRDOBA, M.; BALZARINI, M.; BRUNO, C.; COSTA, J. L. Análisis de componentes principales con datos georreferenciados: Una aplicación en agricultura de precisión. **Revista de la Facultad de Ciencias Agrarias UNCUYO,** v. 44, n. 1, p. 27-39, 2012.

CÓRDOBA, M.; BRUNO, C.; COSTA, J. L.; 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.

CÓRDOBA, M.; BRUNO, C.; COSTA, J. L.; PERALTA, N. R.; BALZARINI, M. Protocol for multivariate homogeneous zone delineation in precision agriculture. **Biosystems Engineering**, n. 143, p. 95-107, 2016.

CZAPLEWSKI, R. L.; REICH, R. M. **Expected value and variance of Moran's bivariate spatial autocorrelation statistic under permutation**. Research Paper RM-309. Fort Collins: USDA Forest Service, 1993. 13 p.

DOERGE, T. A. **Site-specific management guidelines**. Norcross: Potash & Phosphate Institute, 2000. 135 p.

DRAY, S.; SAID, S.; DÉBIAS, F. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. **Journal of Vegetation Science,** v. 19, n. 1, p. 45-56, 2008.

EMBRAPA. Centro Nacional de Pesquisa de Solos. **Sistema brasileiro de classificação de solo**. Rio de Janeiro: CNPSO, 2013. 412 p.

FERREIRA, D. F. **Análise multivariada**. Lavras: UFLA, 1996. 394 p.

FRAISSE, C. W.; SUDDUTH, K. A.; KITCHEN, N. R. Delineation of site–specific management zones by unsupervised classification of topographic attributes and soil electrical conductivity. **International Journal of the American Society of Agricultural and Biological Engineers,** v. 44, n. 1, p. 155-166, 2001.

FRIDGEN, J. J.; KITCHEN, N. R.; SUDDUTH, K. A.; DRUMMOND, S. T.; WIEBOLD, W. J.; FRAISSE, C. W. Management Zone Analyst (MZA): Software for subfield management zone delineation. **Agronomy Journal,** v. 96, p. 100-108, 2004.

FU, Q.; WANG, Z.; JIANG, Q. Delineating soil nutrient management zones based on fuzzy clustering optimized by PSO. **Mathematical and Computer Modelling,** v. 51, n. 11-12, p. 1299-1305, 2010.

GNANADESIKAN, R.; KETTENRING, J.; TSAO, S. Weighting and selection of variables for cluster analysis. **Journal of Classification,** v. 12, n. 1, p. 113-136, 1995.

HORNUNG, A.; KHOSLA, R.; REICH, R. M.; INMAN, D.; WESTFALL, D. G. Comparison of Site-Specific Management Zones: Soil-Color-Based and Yield-Based. **Agronomy Journal,** v. 98, n. 1, p. 407-415, 2006.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of educational psychology,** v. 24, n. 6, p. 417-441, 1933.

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, p. 309-327, 2005.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6 ed. New Jersey: Pearson, 2007. 800 p.

LARSCHEID, G.; BLACKMORE, B. S. Interactions between farm managers and information systems with respect to yield mapping. In: INTERNATIONAL CONFERENCE ON PRECISION AGRICULTURE, 3, 1996, Minneapolis. **Anais…** Minneapolis: American Society of Agronomy, 1996. p. 1153-1163.

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, p. 174–186, 2007.

LI, Y.; SHI, Z.; WU, H.; LI, F.; LI, H. Definition of management zones for enhancing cultivated land conservation using combined spatial data. **Environmental Management,** v. 52, n. 1, p. 792-806, 2013.

MIELKE JR, P. W.; BERRY, K. J. **Permutation methods: a distance function approach**. New York: Springer, 2007. 446 p.

MORAL, F. J.; TERRÓN, J. M.; SILVA, J. R. M. Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. **Soil and Tillage Research,** v. 106, n. 2, p. 335-343, 2010.

PERALTA, N. R.; COSTA, J. L.; FRANCO, M. C.; BALZARINI, M. Delimitación de zonas de manejo con modelos de elevación digital y profundidad de suelo. **Interciencia,** v. 38, n. 6, p. 418-424, 2013.

PERALTA, N. R.; COSTA, J. L.; BALZARINI, M.; FRANCO, M. C.; CÓRDOBA, M.; BULLOCK, D. Delineation of management zones to improve nitrogen management of wheat. **Computers and Electronics in Agriculture,** v. 110, p. 103-113, 2015.

PING, J. L.; DOBERMANN, A. Creating spatially contiguous yield classes for site-specific management. **Agronomy Journal,** v. 95, n. 5, p. 1121-1131, 2003.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2016. 99 p.

SALEH, A.; BELAL, A. A. Delineation of site-specific management zones by fuzzy clustering of soil and topographic attributes: A case study of East Nile Delta, Egypt. **IOP Conf. Ser.: Earth Environmental Science**, v. 18, p. 1–6, 2014.

SCHENATTO, K.; SOUZA, E. G.; BAZZI, C. L.; BIER, V. A.; BETZEK, N. M.; GAVIOLI, A. Data interpolation in the definition of management zones. **Acta Scientiarum,** v. 38, n. 1, p. 31-40, 2016.

SCHEPERS, A. R.; SHANAHAN, J. F.; LIEBIG, M. A.; SCHEPERS, J. S.; JOHNSON, S. H.; LUCHIARI, J. A. Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years. **Agronomy Journal,** v. 96, p. 195-203, 2004.

ZHANG, Z.; LU, X.; LV, N.; CHEN, J.; FENG, B.; LI, X. W.; MA, L. Defining agricultural management zones using GIS techniques: Case study of drip-irrigated cotton fields. **Information Technology Journal,** v. 12, n. 1, p. 6241-6246, 2013.

6 ARTIGO 2 – MÉTODOS DE AGRUPAMENTO DE DADOS PARA DEFINIÇÃO DE ZONAS DE MANEJO

Resumo

A definição de zonas de manejo (ZMs) em áreas agrícolas tem sido sugerida como uma abordagem economicamente viável da agricultura de precisão. Dentre os métodos mais aplicados para realizar essa definição, estão os algoritmos de agrupamento de dados denominados k-means e fuzzy c-means, que geram classes que são usadas para o delineamento de ZMs. No entanto, existem diversos outros algoritmos que podem ser avaliados para isso. Nesse contexto, o objetivo deste trabalho foi avaliar a utilização de 20 algoritmos de agrupamento para a geração de ZMs: average linkage, bagged clustering, centroid linkage, clustering large applications, complete linkage, divisive analysis, fuzzy analysis clustering, fuzzy c-means, fuzzy c-shells, hard competitive learning, hybrid hierarchical clustering, k-means, median linkage, método de McQuitty (mcquitty), método de Ward, neural gas, partitioning around medoids, single linkage, spherical k-means e unsupervised fuzzy competitive learning. A avaliação foi realizada com dados obtidos entre os anos de 2010 e 2015 em três áreas agrícolas comerciais, localizadas no estado do Paraná, nas quais ocorreu o cultivo de soja e milho. A partir das variáveis altitude, argila, areia, silte, resistência mecânica do solo à penetração, declividade e densidade, aplicou-se um método baseado em análise de componentes principais para gerar variáveis sintéticas, que então foram utilizadas pelos 20 métodos de agrupamento. Os resultados da análise de variância possibilitaram sugerir a divisão das três áreas em duas ZMs com potenciais produtivos estatisticamente distintos, além de uma das áreas em três ZMs. Essas divisões puderam ser realizadas satisfatoriamente por meio do uso de 17 dos algoritmos avaliados. Destes, mcquitty e fanny foram considerados os melhores, por proporcionarem as maiores reduções de variância da produtividade após as divisões, para as três áreas. Além disso, estes dois métodos geraram classes com elevada homogeneidade interna e delimitaram ZMs sem fragmentações – portanto, adequadas para as operações de campo. Já os algoritmos fuzzy c-means e k-means foram capazes de gerar classes adequadas somente para duas áreas. Nelas, obtiveram resultados inferiores aos de mcquitty e fanny em relação à redução de variância da produtividade e homogeneidade interna das classes definidas.

Palavras-chave: agricultura de precisão; análise de agrupamento; análise de componentes principais; manejo localizado; unidades de manejo.

DATA CLUSTERING METHODS FOR DEFINITION OF MANAGEMENT ZONES

Abstract

The definition of management zones (MZs) in agricultural fields has been suggested as an economically viable approach to precision agriculture. Common methods for this task include the cluster analysis algorithms fuzzy c-means and k-means, that generate classes which are used to define MZs. However, many other algorithms can generate these subareas. In this context, the objective of this study was to evaluate the use of 20 clustering algorithms to define MZs: average linkage, bagged clustering, centroid linkage, clustering large applications, complete linkage, divisive analysis, fuzzy analysis clustering (fanny), fuzzy c-means, fuzzy c-shells, hard competitive learning, hybrid hierarchical clustering, k-means, McQuitty's method, median linkage, neural gas, partitioning around medoids, single linkage, spherical k-means, unsupervised fuzzy competitive learning, and Ward's method. The evaluation was conducted with data obtained between 2010 and 2015 in three commercial agricultural fields cultivated with soybean and corn in the state of Paraná, Brazil. From variables elevation, clay, sand, silt, soil penetration resistance, slope and bulk density, a method based on principal component analysis (PCA) was applied to generate new variables that were employed as inputs for the

clustering algorithms. The results of the analysis of variance (ANOVA) suggested a division of the three fields into two classes with significantly different yields and a division of one of the fields into three classes. These divisions can be satisfactorily performed using 17 of the evaluated algorithms. Mcquitty and fanny were the best algorithms, because they produced the largest reductions in the variance of yield in the three fields. These methods generated classes with high internal homogeneity and delimited MZs without fragmentation – therefore suitable for field operations. Fuzzy c-means and k-means generated significantly adequate subareas in only two fields, in which the obtained results were lower than those obtained using mcquitty and fanny, for variance reduction and internal homogeneity of the defined classes.

Keywords: cluster analysis; management units; precision agriculture; principal component analysis; site-specific management.

6.1 Introdução

Segundo Fridgen et al. (2004), a identificação e o gerenciamento de subáreas com características distintas dentro de uma área agrícola são atividades incluídas no conceito de gerenciamento localizado de culturas. Cada subárea identificada, que na agricultura de precisão é frequentemente chamada de zona de manejo (ZM), é considerada homogênea com base em determinadas medidas quantitativas – as variáveis ou atributos. Dentre muitas variáveis que podem ser consideradas para a definição de ZMs, geralmente aquelas que representem fontes de informação espacial, que estejam correlacionadas com a produtividade e que sejam estáveis ao longo do tempo, são mais recomendadas.

De acordo com Cid-Garcia, Bravo-Lozano e Rios-Solis (2014), em comparação com os procedimentos da agricultura convencional, a utilização de ZMs possibilita reduzir a variabilidade espacial da produtividade das culturas e os danos ao meio ambiente associados à aplicação excessiva de determinados insumos. Essas subáreas também podem representar indicadores para amostragens do solo e das plantas cultivadas, reduzindo a quantidade de amostras que precisam ser coletadas e analisadas.

Os métodos aplicados para a geração de ZMs podem ser divididos em duas categorias (LI et al., 2007): os empíricos e os de análise de agrupamento. Os métodos empíricos são fundamentados no uso de conhecimento especializado e, normalmente, na distribuição da produtividade para dividir uma área em determinada quantidade de subáreas. São abordagens mais simples, que estão sujeitas a decisões subjetivas.

Já os métodos de análise de agrupamento têm o objetivo de colocar pontos de uma área agrícola que apresentem valores similares para certas variáveis na mesma classe (também chamada de grupo) e separar pontos que tenham valores pouco similares para essas variáveis em classes distintas. Na prática, essas classes são utilizadas para definir as ZMs. Segundo Boydell e McBratney (2002), embora os algoritmos de agrupamento sejam mais complexos que os empíricos, possibilitam maior diferenciação entre grupos por meio de critérios menos subjetivos – isto porque podem empregar diversas variáveis no processo de geração das subáreas.

Há muitas opções de algoritmos de agrupamento de dados que podem ser avaliados para a criação de ZMs, com destaque para k-means (MACQUEEN, 1967) e fuzzy c-means (BEZDEK, 1981). Arno et al. (2011), Ikenaga e Inamura (2008) e Taylor et al. (2003) relataram resultados satisfatórios com o uso de k-means. Já Boydell e McBratney (2002), Caires, Wuddivira e Bekele (2015), Milne et al. (2012), Moral, Terrón e Silva (2010) e Schenatto et al. (2016) apresentaram resultados adequados empregando fuzzy c-means.

No entanto, esses dois algoritmos de agrupamento eventualmente podem ser superados por outros na geração das melhores classes para ZMs. Dobermann et al. (2003) compararam os algoritmos k-means, fuzzy c-means, isodata (BALL; HALL, 1967) e método de Ward (WARD, 1963), e mostraram que os dois últimos podem proporcionar resultados melhores. Guastaferro et al. (2010) também mostraram que isodata pode ser superior a fuzzy c-means na geração de ZMs. Xu e Wunsch (2009) destacaram que o algoritmo partitioning around medoids (KAUFMAN; ROUSSEEUW, 1990) pode superar k-means. Já Delalibera, Weirich e Nagata (2012), Fleming et al. (2000), Ortega e Santibáñez (2007) e Russ e Kruse (2011) apresentaram resultados satisfatórios com a utilização de métodos de agrupamento classificados como hierárquicos.

Tendo em vista que esses e outros trabalhos abordaram poucos algoritmos alternativos a k-means e a fuzzy c-means, o objetivo deste trabalho foi avaliar a utilização de 20 métodos de agrupamento para a geração de ZMs, incluindo métodos de diferentes paradigmas de agrupamento e que possivelmente não foram avaliados em outros trabalhos para essa finalidade.

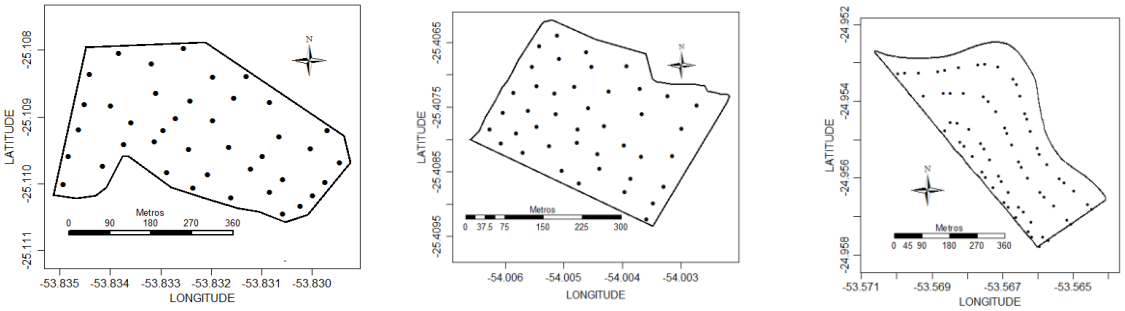
6.2 Material e métodos

6.2.1 Conjuntos de dados

Os experimentos foram realizados com dados coletados em três áreas agrícolas comerciais localizadas no Estado do Paraná, com cultivo de soja e milho, no período de 2010 a 2015. Os solos foram cultivados em sistema de plantio direto e classificados como LATOSSOLO VERMELHO Distroférrico típico (EMBRAPA, 2013). A área A está localizada no município de Céu Azul, possui 15,5 ha, apresenta localização geográfica central de 25º06'32'' S e 53º49'55'' O e altitude média de 460 m. A área B está localizada no município de Serranópolis do Iguaçu, apresenta 9,9 ha, possui localização geográfica central de 25º24'28'' S e 54º00'17'' O e altitude média de 355 m. Já a área C está no município de Cascavel, possui 19,8 ha, apresenta localização geográfica central de 24º57'08'' S e 53º33'59'' O e altitude média de 650 m.

Tendo em vista que as três áreas apresentam certo grau de declividade e comportam curvas de nível, optou-se por utilizar grades amostrais irregulares, definindo-se os pontos em locais que não coincidissem com as curvas (Figura 1, onde as imagens estão em escalas

diferentes). Isto devido à possibilidade de influência destas sobre a produtividade, em função das condições físicas do solo e do acúmulo de água nestes locais. Assim, os pontos amostrais ficaram localizados na linha central imaginária entre as curvas de nível.



Área A Área B Área C

Figura 1 Áreas experimentais e pontos amostrais: área A, situada em Céu Azul - PR; área B, situada em Serranópolis do Iguaçu - PR; área C, situada em Cascavel - PR.

Seguindo recomendação de Doerge (2000), utilizaram-se apenas variáveis consideradas temporalmente estáveis (Tabela 1) no processo de definição de ZMs. A intenção foi que as variáveis empregadas conduzissem ao delineamento de subáreas válidas por vários anos. Consequentemente, variáveis químicas não foram consideradas.

Tabela 1 Variáveis avaliadas entre 2010 e 2015, para cada área agrícola

Variáveis (atributos)	Área A				Área B				Área C	
	2012	2013	2014	2015	2012	2013	2014	2015	2010	2011
Argila (%)	X				X				X	
Densidade (g cm^{-3})	X				X				X	
Altitude (m)	X				X				X	
Areia (%)	X				X				X	
Silte (%)	X				X				X	
Declividade (˚)	X				X				X	
RSP 0-0,1 m (MPa)	X	X	X		X	X	X		X	
RSP 0,1-0,2 m (MPa)	X	X	X		X	X	X		X	
RSP 0,2-0,3 m (MPa)	X	X	X		X	X	X		X	
Produtividade milho (t ha^{-1})						X	X	X		
Produtividade soja (t ha^{-1})	X	X	X	X	X	X	X	X	X	X

RSP: resistência mecânica do solo à penetração.

As grades amostrais foram constituídas por 40 pontos para a área A (densidade de 2,67 pontos ha^{-1}), 42 pontos para a área B (4,24 pontos ha^{-1}) e 68 pontos para a área C (3,43 pontos ha^{-1}). As densidades de amostragem foram maiores que 2,5 pontos ha^{-1}, seguindo recomendação de Nanni et al. (2011) para viabilizar a detecção de dependência espacial entre amostras. Essas densidades foram diferentes pelo fato dessas áreas também serem utilizadas para a condução de experimentos de outros pesquisadores.

Além disso, ao menos 40 amostras compostas de solo foram coletadas em cada área durante cada atividade de coleta de solo, seguindo recomendação de Journel e Huijbregts (2004). Esta recomendação e a de Nanni et al. (2011) garantiram o mínimo de 30 pares de pontos para obtenção de cada semivariância (DIGGLE; RIBEIRO JR, 2007), durante a

definição do semivariograma experimental de Matheron, necessário para a interpolação por krigagem ordinária.

As amostras de solo destinadas à análise de características químicas e de textura (areia, argila e silte) foram coletadas na profundidade de 0-0,2 m, com auxílio de um perfurador de solo Stihl BT 45. Adaptando o método proposto por Wollenhaupt, Wolkowski e Clayton (1994), coletaram-se oito subamostras em cada ponto amostral, para formar uma amostra composta, dentro de um círculo imaginário com raio de 3 m e centro no ponto amostral registrado na grade. Posteriormente, as amostras compostas foram encaminhadas para um laboratório especializado em análise de solos.

Utilizando-se anel volumétrico, também se coletou uma amostra de solo não deformada em cada ponto amostral para determinar a densidade aparente do solo. Os procedimentos para obtenção dos valores dessa variável foram executados no Laboratório de Armazenamento de Amostras e no Laboratório de Solos da UNIOESTE, seguindo orientações do Manual de Métodos de Análise de Solo da Embrapa (EMBRAPA, 1997).

Para obter os valores da resistência mecânica do solo à penetração (RSP) nas profundidades de 0-0,1 m, 0,1-0,2 m e 0,2-0,3 m, empregou-se um medidor eletrônico de compactação do solo Falker PenetroLOG PLG1020. O valor da RSP para cada ponto amostral foi definido como a média dos valores correspondentes a quatro medições, realizadas em locais próximos àqueles em que se efetuou a coleta das oito subamostras de solo. Para determinar os valores da altitude dos pontos amostrais nas três áreas, utilizou-se uma estação total eletrônica Topcon GPT-7505. Posteriormente, foram calculados os valores de declividade a partir da altitude desses pontos, empregando-se o software Surfer.

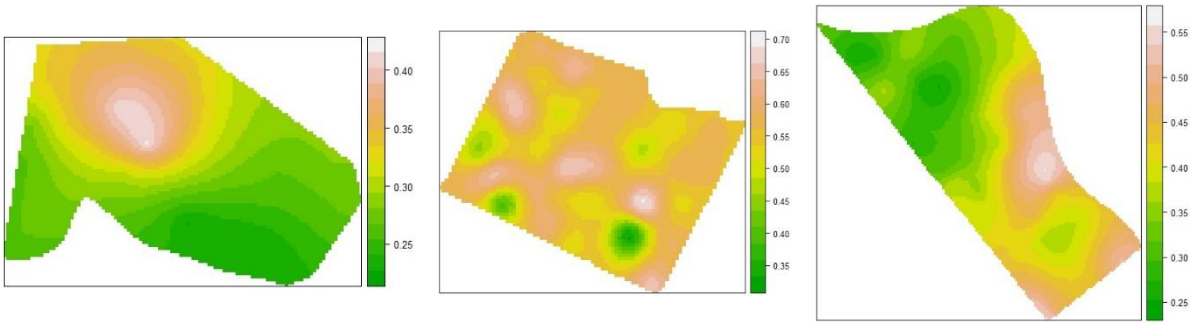
Para validação das ZMs geradas, foram usados dados da produtividade de milho e de soja dos anos citados na Tabela 1. Esses dados foram obtidos para os mesmos pontos amostrais onde ocorreram as avaliações das outras variáveis. Para a área A, os dados de produtividade de soja foram obtidos utilizando-se uma colhedora CASE IH 2388 com um monitor de colheita CASE AFS PRO 600. Para as áreas B (milho e soja) e C (soja), a produtividade foi determinada por meio da execução de colheita manual de uma área de amostragem de aproximadamente 0,9 m^2, em cada ponto amostral. Os valores da produtividade foram corrigidos para um teor de água de 13% em todas as situações.

Para cada safra de milho e de soja considerada, realizou-se a padronização dos valores da produtividade correspondentes aos pontos amostrais, utilizando-se uma versão da técnica da amplitude (Equação 1) (MIELKE JR; BERRY, 2007). Assim, a produtividade foi padronizada para valores entre 0 e 1 com o objetivo de atender ao requisito de estabilidade desses dados, que geralmente são bastante influenciados pelas variações do clima e da precipitação pluviométrica.

$$P_{iN} = \frac{P_i - P_{\min}}{P_{\max} - P_{\min}} \qquad \text{Eq. (1)}$$

em que: P_{iN} é o valor padronizado da produtividade para o ponto amostral i; P_i é o valor original da produtividade para o ponto i; e $P_{\max}$ e $P_{\min}$ correspondem, respectivamente, aos valores amostrais máximo e mínimo da produtividade no conjunto de dados considerado.

Em seguida, calculou-se, para cada área, a média aritmética dos valores padronizados dos anos disponíveis, gerando-se uma única variável correspondente à produtividade média padronizada. Com o intuito de representar graficamente a distribuição espacial da produtividade média padronizada, realizou-se a interpolação espacial dos valores correspondentes aos pontos amostrais, aplicando-se o método krigagem ordinária com uma grade de pixels representando 5x5 m. Os modelos teóricos esférico, exponencial e gaussiano foram ajustados ao semivariograma experimental. Em seguida, estatísticas de erro médio e desvio padrão do erro médio, obtidas por meio da técnica de validação cruzada (DIGGLE; RIBEIRO JR, 2007), foram empregadas para escolher o melhor modelo para interpolar os dados. Assim, para interpolar a variável produtividade média padronizada correspondente às áreas A e C, selecionou-se o modelo exponencial, enquanto que para a área B escolheu-se o modelo gaussiano. Os mapas dos valores interpolados são exibidos na Figura 2.



Área A Área B Área C

Figura 2 Mapas resultantes da interpolação espacial da produtividade média padronizada, utilizando-se krigagem ordinária.

Foram calculadas estatísticas descritivas para os dados da produtividade amostral de soja e milho, para cada safra considerada, bem como para os dados resultantes da interpolação da variável produtividade média padronizada (Tabela 2).

Para selecionar as variáveis necessárias para a geração de ZMs, Córdoba et al. (2013) e Gavioli et al. (2016) compararam diversos métodos com conjuntos de dados multivariados correspondentes a áreas em que havia dependência espacial. Eles mostraram que métodos fundamentados em uma abordagem de análise multivariada espacial denominada MULTISPATI-PCA (DRAY; SAID; DÉBIAS, 2008) proporcionaram os melhores resultados.

Tabela 2 Resumo de estatísticas descritivas para os dados da produtividade amostral de cada safra de soja e miho, bem como para os dados interpolados da variável produtividade média padronizada

Área	Ano / cultura	Média	Mediana	Máx.	Mín.	DP	CV (%)
Dados da produtividade amostral original (t ha^{-1})							
A	2012 / S	3,984	4,067	5,068	2,541	0,472	11,8
	2013 / S	3,941	4,012	6,166	2,057	0,518	13,1
	2014 / S	4,525	4,553	5,354	3,635	0,281	6,2
	2015 / S	3,656	3,726	4,686	2,199	0,556	15,2
B	2012 / S	5,255	5,255	6,980	2,563	0,749	14,2
	2013 / M	8,945	9,195	10,799	6,678	0,992	11,1
	2013 / S	5,079	5,107	6,808	3,366	0,586	11,5
	2014 / M	10,276	10,415	13,342	6,763	1,095	10,7
	2014 / S	3,888	3,819	4,759	2,934	0,496	12,8
	2015 / M	7,975	8,317	9,895	4,508	1,166	14,6
	2015 / S	4,676	4,708	5,742	3,354	0,518	11,1
C	2010 / S	2,638	2,565	4,340	1,550	0,606	23,0
	2011 / S	3,243	3,263	4,644	2,300	0,484	14,9
Dados da produtividade média padronizada (após interpolação)							
A	-	0,301	0,288	0,416	0,226	0,047	15,6
B	-	0,558	0,562	0,688	0,333	0,045	8,1
C	-	0,385	0,384	0,558	0,252	0,071	18,6

M: milho; S: soja; DP: desvio padrão; CV: coeficiente de variação.

Por isso, neste trabalho utilizou-se o método proposto por Gavioli et al. (2016), denominado MPCA-SC, que se baseia na aplicação da análise de correlação espacial entre variáveis (CZAPLEWSKI; REICH, 1993) em conjunto com MULTISPATI-PCA. A principal característica do método MPCA-SC é ter incluído uma restrição espacial à análise de componentes principais (ACP) clássica, possibilitando que esta seja executada considerando a existência de dependência espacial em conjuntos de dados georreferenciados. MPCA-SC tem como vantagem o fato de maximizar a autocorrelação espacial entre pontos amostrais, enquanto ACP maximiza a variância total. Aplicando-se MPCA-SC, foram geradas novas variáveis sintéticas adequadas para constituírem conjuntos de dados de entrada para os algoritmos de agrupamento. Cada variável sintética, chamada de componente principal espacial (CPE), é uma combinação linear das variáveis originais que apresentaram correlação espacial significativa com a produtividade média, ao nível de 5%.

Na formação das CPEs, utilizaram-se as seguintes variáveis originais: argila, altitude, areia e RSP 0-0,1 m, para a área A; altitude, RSP 0-0,1 m e RSP 0,1-0,2 m, para a área B; e argila, altitude, areia, RSP 0,1-0,2 m e RSP 0,2-0,3 m, para a área C. Para as três áreas, foram usados os valores obtidos para as duas primeiras CPEs (CPE1 e CPE2) em cada ponto amostral. Tomou-se a decisão de empregar apenas as duas primeiras CPEs com base no critério sugerido por Ferreira (1996), que recomenda utilizar a quantidade de componentes principais suficientes para representar ao menos 70% da variância total dos dados originais (Tabela 3).

Tabela 3 Estatísticas das CPEs utilizadas para as três áreas: variância associada à CPE, porcentagem da variância total dos dados representada pela CPE e somatório dessas porcentagens

Área	Variável	Variância	Porcentagem variância total	Somatório das porcentagens
A	CPE1	2,64	68,0	68,0
	CPE2	1,24	32,0	100
B	CPE1	1,68	61,1	61,1
	CPE2	0,61	22,2	83,3
C	CPE1	2,57	66,9	66,9
	CPE2	1,27	33,1	100

Para gerar ZMs com contornos suaves, decidiu-se interpolar espacialmente os dados das duas primeiras CPEs de cada área. Compararam-se os resultados da aplicação dos métodos de interpolação krigagem ordinária, inverso da distância e inverso da distância ao quadrado. Como a krigagem ordinária proporcionou os melhores resultados, optou-se por usar esse interpolador com uma grade de 5x5 m.

Nesse sentido, aplicou-se o estimador clássico de Matheron (MATHERON, 1963) para a obtenção do semivariograma experimental. Os modelos teóricos esférico, exponencial e gaussiano foram ajustados ao semivariograma e a técnica de validação cruzada foi empregada para determinar o melhor modelo teórico para cada caso. Identificou-se que o melhor modelo para interpolar as duas CPEs das áreas A e C era o exponencial, enquanto que para a área B era o gaussiano. Os dados resultantes da interpolação das CPEs foram usados como entrada para todos os algoritmos de agrupamento, a fim de dividir as áreas em duas, três e quatro ZMs.

6.2.2 Métodos de agrupamento de dados

Avaliaram-se 20 métodos de agrupamento de dados, sendo 9 hierárquicos e 11 de particionamento (Tabela 4). Optou-se por algoritmos que asseguram ao usuário o controle sobre a quantidade de grupos que devem ser formados. Trabalhos correlatos apresentaram avaliações de alguns desses algoritmos na definição de ZMs. Porém, além da pequena quantidade de métodos comparados, eles consideraram variáveis e procedimentos metodológicos diferentes dos que foram utilizados neste trabalho.

A função de avaliação de similaridade empregada com os métodos que necessitaram de sua definição explícita foi a distância euclidiana (BEZDEK, 1981). Esta foi selecionada por ser a função mais empregada com os algoritmos avaliados e por não haver risco de distorção de resultados devido a variáveis correlacionadas entre si (já que CPEs não apresentam correlação entre si).

Tabela 4 Métodos de agrupamento implementados e avaliados para a definição de zonas de manejo

Métodos	Siglas	Referências
average linkage[a]	AVG	Jain e Dubes (1988)
centroid linkage[a]	CEN	Jain e Dubes (1988)
complete linkage[a]	COM	Jain e Dubes (1988)
divisive analysis (diana)[a]	DIA	Kaufman e Rousseeuw (1990)
hybrid hierarchical clustering[a]	HHC	Chipman e Tibshirani (2006)
median linkage[a]	MED	Jain e Dubes (1988)
método de McQuitty (mcquitty)[a]	MCQ	McQuitty (1966)
método de Ward (ward)[a]	WAR	Ward (1963)
single linkage[a]	SIN	Jain e Dubes (1988)
bagged clustering[b]	BCL	Leisch (1999)
clustering large applications (clara)[b]	CLA	Kaufman e Rousseeuw (1990)
fuzzy analysis clustering (fanny)[b]	FNY	Kaufman e Rousseeuw (1990)
fuzzy c-means[b]	FCM	Bezdek (1981)
fuzzy c-shells[b]	FCS	Dave (1992)
hard competitive learning[b]	HCL	Xu e Wunsch (2009)
k-means[b]	KME	MacQueen (1967)
neural gas[b]	NGA	Martinetz, Berkovich e Schulten (1993)
partitioning around medoids[b]	PAM	Kaufman e Rousseeuw (1990)
spherical k-means[b]	SKM	Dhillon e Modha (2001)
unsupervised fuzzy competitive learning[b]	UFCL	Pal, Bezdek e Hathaway (1996)

[a]: método hierárquico; [b]: método de particionamento.

6.2.3 Validação dos agrupamentos

A utilização de algoritmos de agrupamento de categorias diferentes exigiu a adoção de critérios de avaliação adequados para todos eles. Assim, a validação dos agrupamentos foi efetuada aplicando-se a análise de variância (ANOVA), o teste de comparação de médias de Tukey (PIMENTEL-GOMES, 2000), o índice de redução da variância (variance reduction - VR) (LI et al., 2007), o coeficiente de silhueta médio (average silhouette coefficient - ASC) (ROUSSEEUW, 1987) e o coeficiente Kappa (Kp) (COHEN, 1960).

A ANOVA e o teste de Tukey foram aplicados para comparar as produtividades médias padronizadas das ZMs geradas, com o intuito de verificar se as diferenças entre essas produtividades foram estatisticamente significativas ao nível de 5% de significância. Antes dessa verificação, confirmou-se que não existia dependência espacial entre as amostras pertencentes a cada ZM.

A redução da variância foi medida para a produtividade média padronizada com a expectativa de que o somatório das variâncias dos dados das ZMs fosse inferior à variância total. O coeficiente ASC representou a qualidade da formação interna e da separação externa dos grupos gerados, e o coeficiente Kappa foi aplicado para avaliar o grau de concordância entre mapas de ZMs gerados com a execução dos algoritmos de agrupamento.

6.2.4 Softwares

Utilizou-se o software estatístico R 3.1.2 (R CORE TEAM, 2016) para o desenvolvimento de programas que contêm a implementação do método de seleção de variáveis MPCA-SC e dos métodos de agrupamento. Estes programas foram integrados ao

sistema de banco de dados PostgreSQL 9.0.5 (The PostgreSQL Global Development Group), empregado para o armazenamento dos dados, contando com a inclusão de sua extensão para bancos de dados espaciais PostGIS 1.5.5 (PostGIS Project Steering Committee).

Para a interpolação espacial dos dados, utilizou-se um programa desenvolvido no software R. Para gerar os mapas temáticos das ZMs e executar a ANOVA e o teste de Tukey, empregou-se o Software para Definição de Unidades de Manejo (SDUM) (BAZZI et al., 2013). Para obter os valores do índice VR e dos coeficientes ASC e Kappa, mais uma vez aplicou-se o software R.

Os pacotes do software R denominados geoR, gstat, ade4 e spdep foram necessários para executar o método MPCA-SC. Além de geoR e gstat, os pacotes classInt e MASS foram exigidos para a aplicação da krigagem ordinária e para as análises estatísticas e geoestatísticas. Os pacotes cclust, cluster, e1071, fastcluster, fclust, hybridHclust, optpart e skmeans foram usados para a execução dos algoritmos de agrupamento, e o pacote psych foi necessário para determinar os valores do coeficiente Kappa.

6.3 Resultados e discussão

Os métodos de agrupamento diana, fuzzy c-shells e single linkage geraram ZMs excessivamente fragmentadas e exigiram tempos de processamento muito longos em comparação com os outros algoritmos avaliados (exigiram no mínimo 18 vezes mais tempo de processamento que o tempo médio de execução de qualquer um dos outros 17 algoritmos). Por isso, os três foram considerados insatisfatórios para a definição de ZMs para as três áreas experimentais, e seus resultados não foram incluídos neste trabalho.

Os resultados da avaliação dos outros 17 algoritmos de agrupamento, considerados úteis para definir ZMs, utilizando a ANOVA, teste de Tukey, índice VR e coeficiente ASC, são apresentados nas Tabelas 5, 6 e 7. Os tempos necessários para a execução computacional desses métodos variaram, mas foram similares (todos inferiores a 40 segundos, quando executados em um computador pessoal com processador Intel Core i7 e 8 GB de memória RAM). Devido a isso, as diferenças de tempo não foram consideradas relevantes para efeito de comparação entre os métodos.

Os resultados da ANOVA e do teste de Tukey possibilitaram sugerir a divisão das três áreas experimentais em duas ZMs com produtividades médias estatisticamente distintas, empregando os métodos average, bagged clustering, mcquitty, spherical k-means e ward (métodos sublinhados nas Tabelas 5, 6 e 7). Os outros 12 algoritmos foram capazes de formar duas classes estatisticamente distintas para duas dessas áreas. No caso da área C, também foi possível efetuar a divisão em três classes distintas, porém apenas com o uso dos algoritmos complete e fanny.

Tabela 5 Resultados da avaliação dos métodos de agrupamento na geração de duas, três e quatro classes, considerando-se ANOVA (teste de Tukey), índice VR e coeficiente ASC, para a área A

Método	2 classes				3 classes					4 classes					
	C_1	C_2	VR(%)	ASC	C_1	C_2	C_3	VR(%)	ASC	C_1	C_2	C_3	C_4	VR(%)	ASC
AVG	a	b	15,9	0,55	a	b	b	18,4	0,45	a	ab	bc	c	20,6	0,46
BCL	a	b	16,7	0,58	a	b	b	36,3	0,45	a	b	ab	b	21,3	0,55
CEN	a	b	18,2	0,57	a	ab	b	20,4	0,45	a	a	a	a	0,0	0,41
CLA	a	b	21,0	0,59	a	b	b	25,3	0,47	a	ab	b	b	19,5	0,55
COM	a	a	9,5	0,55	a	ab	b	15,0	0,46	a	ab	b	b	22,2	0,38
FNY	a	b	21,2	0,59	a	b	b	30,2	0,46	a	ab	c	bc	29,6	0,39
FCM	a	b	34,1	0,59	a	b	b	35,5	0,46	a	a	b	b	35,6	0,54
HCL	a	b	21,6	0,59	a	a	b	26,2	0,46	a	b	ab	b	19,9	0,54
HHC	a	b	21,6	0,59	a	a	b	21,4	0,48	a	ab	b	b	21,5	0,38
KME	a	b	33,8	0,59	a	b	a	23,8	0,46	a	a	b	b	35,8	0,39
MCQ	a	b	39,2	0,59	a	b	b	38,3	0,43	a	ab	c	bc	37,4	0,35
MED	a	b	16,2	0,56	a	b	b	14,4	0,42	a	ab	bc	c	13,2	0,33
NGA	a	b	21,4	0,59	a	b	a	25,8	0,46	ac	b	c	ab	29,7	0,38
PAM	a	b	20,9	0,59	a	b	b	29,3	0,46	a	ab	b	b	23,5	0,54
SKM	a	b	22,4	0,59	a	b	a	41,6	0,47	a	b	b	a	46,9	0,49
UFCL	a	b	21,7	0,59	a	b	b	25,8	0,46	a	ab	bc	c	30,7	0,39
WAR	a	b	19,8	0,58	a	a	b	21,3	0,47	a	ab	c	bc	29,3	0,54

C_i: classe *i*; VR: índice de redução da variância; ASC: coeficiente de silhueta médio.
Métodos sublinhados proporcionaram classes com produtividades médias estatisticamente distintas.

Tabela 6 Resultados da avaliação dos métodos de agrupamento na geração de duas, três e quatro classes, considerando-se ANOVA (teste de Tukey), índice VR e coeficiente ASC, para a área B

Método	2 classes				3 classes					4 classes					
	C_1	C_2	VR(%)	ASC	C_1	C_2	C_3	VR(%)	ASC	C_1	C_2	C_3	C_4	VR(%)	ASC
AVG	a	b	9,5	0,70	a	ab	b	10,9	0,61	a	a	a	a	0,0	0,55
BCL	a	b	9,2	0,70	ab	a	b	8,1	0,60	a	b	ab	ab	5,9	0,47
CEN	a	b	9,8	0,70	ab	a	b	15,8	0,60	ab	ab	a	b	12,5	0,48
CLA	a	a	9,2	0,70	a	ab	b	11,8	0,61	a	a	ab	b	9,6	0,57
COM	a	b	9,5	0,70	a	ab	b	13,1	0,59	a	ab	ab	b	12,8	0,51
FNY	a	a	10,0	0,69	a	ab	b	10,8	0,61	a	a	ab	b	10,9	0,55
FCM	a	a	10,7	0,70	a	b	a	13,7	0,61	ab	a	bc	c	20,5	0,55
HCL	a	a	9,6	0,69	a	b	ab	12,0	0,61	a	b	ab	b	10,7	0,55
HHC	a	a	9,6	0,69	a	ab	b	8,5	0,59	a	ab	ab	b	10,9	0,55
KME	a	a	10,5	0,70	a	b	a	13,4	0,61	a	ab	b	b	13,8	0,55
MCQ	a	b	11,9	0,70	a	ab	b	12,4	0,56	a	ab	ab	b	11,0	0,46
MED	a	b	5,3	0,55	a	a	b	22,2	0,57	a	a	a	a	0,0	0,52
NGA	a	a	9,6	0,69	ab	a	b	11,1	0,61	a	b	ab	ab	11,9	0,55
PAM	a	a	9,3	0,70	a	ab	b	11,7	0,61	a	ab	ab	b	13,6	0,48
SKM	a	b	8,8	0,58	a	b	ab	10,6	0,46	a	a	a	b	22,0	0,45
UFCL	a	a	9,6	0,69	a	b	ab	10,1	0,61	a	ab	a	b	11,9	0,51
WAR	a	b	8,1	0,57	a	b	b	11,0	0,61	a	ab	ab	b	13,6	0,54

C_i: classe *i*; VR: índice de redução da variância; ASC: coeficiente de silhueta médio.
Métodos sublinhados proporcionaram classes com produtividades médias estatisticamente distintas.

Tabela 7 Resultados da avaliação dos métodos de agrupamento na geração de duas, três e quatro classes, considerando-se ANOVA (teste de Tukey), índice VR e coeficiente ASC, para a área C

Método	2 classes				3 classes					4 classes					
	C_1	C_2	VR(%)	ASC	C_1	C_2	C_3	VR(%)	ASC	C_1	C_2	C_3	C_4	VR(%)	ASC
AVG	a	b	19,5	0,70	a	b	b	19,5	0,66	a	a	a	a	0,0	0,65
BCL	a	b	21,3	0,61	a	a	b	27,7	0,68	a	b	ac	c	37,7	0,61
CEN	a	a	0,0	0,69	a	a	a	0,0	0,57	a	a	a	a	0,0	0,66
CLA	a	b	22,0	0,66	a	b	b	28,6	0,68	a	b	b	c	40,0	0,55
COM	a	b	24,0	0,70	a	b	c	32,8	0,64	a	b	a	b	36,0	0,66
FNY	a	b	31,6	0,67	a	b	c	34,1	0,68	a	b	b	c	41,4	0,56
FCM	a	b	30,1	0,65	a	b	a	29,6	0,68	a	b	c	ab	34,1	0,65
HCL	a	b	24,3	0,64	a	a	b	28,4	0,68	a	b	bc	c	39,0	0,63
HHC	a	b	24,3	0,64	a	b	b	29,0	0,66	a	b	b	c	35,1	0,55
KME	a	b	29,6	0,64	a	a	b	29,1	0,48	a	b	c	ab	34,0	0,65
MCQ	a	b	18,3	0,70	a	b	b	15,5	0,61	a	b	c	ab	33,2	0,60
MED	a	a	0,0	0,66	a	a	b	7,8	0,59	a	a	a	a	0,0	0,21
NGA	a	b	24,3	0,64	a	a	b	28,3	0,68	a	b	b	a	37,9	0,65
PAM	a	b	22,0	0,66	a	b	b	28,6	0,68	a	b	b	c	40,1	0,55
SKM	a	b	21,9	0,69	a	b	b	27,8	0,68	a	b	bc	ac	33,5	0,65

UFCL	a	b	24,3	0,65	a	b	b	28,6	0,68	a	ab	b	c	39,1	0,63
WAR	a	b	16,5	0,61	a	b	b	20,7	0,67	a	b	c	ab	37,1	0,65

C_i: classe *i*; VR: índice de redução da variância; ASC: coeficiente de silhueta médio.
Métodos sublinhados proporcionaram classes com produtividades médias estatisticamente distintas.

Esses resultados evidenciam que métodos de agrupamento amplamente utilizados para a definição de ZMs, como fuzzy c-means e k-means, eventualmente podem não ser adequados para essa tarefa em alguma nova área. Dobermann et al. (2003) e Guastaferro et al. (2010) também mostraram a inferioridade de fuzzy c-means em relação a outros algoritmos, em situações de geração de ZMs para áreas com cultivo de milho e trigo.

Considerando a divisão dos dados de cada área em duas classes, o maior valor do índice VR para as áreas A e B foi obtido empregando-se o método mcquitty (39,2 e 11,9%, respectivamente). Fuzzy c-means e k-means foram o segundo e o terceiro melhores, porém não geraram duas classes distintas com os dados da área B (Tabela 6). Para a área C, o maior valor do índice VR foi proporcionado pelo algoritmo fanny, tanto para a divisão em duas classes (31,6%) como para a divisão em três (34,1%). Para esta área, fuzzy c-means e k-means também proporcionaram o segundo e o terceiro melhores valores para VR, mas para dividir os dados em duas classes – isto porque ambos não foram capazes de definir três classes distintas (Tabela 7).

Embora tenha sido destacado o melhor método de agrupamento segundo o índice VR para cada área, todos os algoritmos que geraram classes estatisticamente distintas também estão habilitados para a definição de ZMs. Os valores de VR variaram de acordo com a área e com a quantidade de classes, mas destaca-se o desempenho superior do método hierárquico mcquitty. Um resultado similar em relação a algoritmos hierárquicos foi apresentado por Dobermann et al. (2003), que mostraram que ward também pode proporcionar os maiores valores para o índice VR na geração de ZMs.

Ao analisar os valores do coeficiente ASC dos métodos que obtiveram os melhores resultados para o índice VR, observou-se que, para os dados das áreas A e B, mcquitty obteve os maiores valores para ASC ao gerar duas classes. Para essas duas áreas, os valores alcançados por fuzzy c-means e k-means foram iguais aos de mcquitty. Para a divisão dos dados da área C em duas classes, o algoritmo fanny obteve o valor 0,67 para o coeficiente ASC, superando a maioria dos algoritmos (inclusive fuzzy c-means e k-means). Já para a divisão em três classes, fanny obteve o valor 0,68 para esse coeficiente, que foi o valor mais alto.

Portanto, os melhores algoritmos segundo o índice VR formaram grupos com elevada similaridade interna e separação inter-grupos adequada (alta dissimilaridade externa). Este resultado satisfatório em relação ao ASC também foi alcançado pelos demais algoritmos. Por outro lado, em algumas situações em que os algoritmos não formaram classes distintas, os valores desse coeficiente foram altos. Constatou-se isto nas três áreas, o que pode ser um

indicativo de que a avaliação de agrupamentos no contexto da definição de ZMs não deve ser efetuada exclusivamente por meio do uso do ASC.

Também se notou que alguns métodos hierárquicos que apresentaram valor 0 para o índice VR obtiveram valores altos para o ASC. Isto ocorreu com average, na definição de quatro classes com os dados das áreas B e C; com centroid, na definição de quatro classes com os dados da área A e qualquer número de classes com os dados da área C; e com median, na definição de quatro classes com os dados das áreas B e C.

Apesar desses resultados negativos, dois desses métodos hierárquicos, average e centroid, ficaram entre os melhores na definição de duas classes com os dados das áreas A e B (Tabelas 5 e 6). Além disso, conforme informado anteriormente, Delalibera, Weirich e Nagata (2012), Fleming et al. (2000), Ortega e Santibáñez (2007) e Russ e Kruse (2011) também relataram sucesso na aplicação de algoritmos hierárquicos na geração de ZMs.

Quanto ao delineamento das subáreas, sob o ponto de vista de aspectos visuais, ocorreram diferenças conforme o algoritmo utilizado (Figuras 3, 4 e 5). Essas diferenças de tamanho e de formato das ZMs ficaram mais evidentes ao aumentar a quantidade delas.

O foco dado neste trabalho para a valorização da minimização da variância interna de cada ZM e da maximização da variância entre essas subáreas, sem impor restrições quanto ao seu tamanho ou formato, foi similar ao foco dado por Dobermann et al. (2003) e Guastaferro et al. (2010). No entanto, embora tenham obtido valores de redução de variância satisfatórios, Dobermann et al. (2003) relataram o problema de muita fragmentação espacial nos mapas gerados.

Neste trabalho, obtiveram-se valores satisfatórios para o índice VR e, de modo geral, os algoritmos geraram subáreas com baixo grau de fragmentação, portanto satisfatórias para o manejo localizado. Haja vista diversos testes realizados e os relatos de Dobermann et al. (2003), acredita-se que a redução da fragmentação esteja associada à utilização do método de seleção de variáveis MPCA-SC em conjunto com a krigagem ordinária (com grade e modelo teórico adequados).

Em algumas situações, apesar de não ter ocorrido fragmentação, definiram-se subáreas com tamanho ou formato que dificultariam a execução de operações de campo: na área A, com o uso do método centroid para definir quatro ZMs; na área B, quando average e median foram aplicados para gerar quatro ZMs; e na área C, ao utilizar centroid e median para qualquer quantidade de subáreas, e average para quatro ZMs. Esses problemas apresentaram relação com o valor do índice VR, pois, excetuando-se uma situação, em todas as demais esse índice foi igual a 0. Este valor significa que a divisão em subáreas não resultou em redução da variância original. Dobermann et al. (2003) também apresentaram mapas com subáreas de tamanhos desproporcionais e destacaram que estas estavam relacionadas a valores de VR próximos de 0.

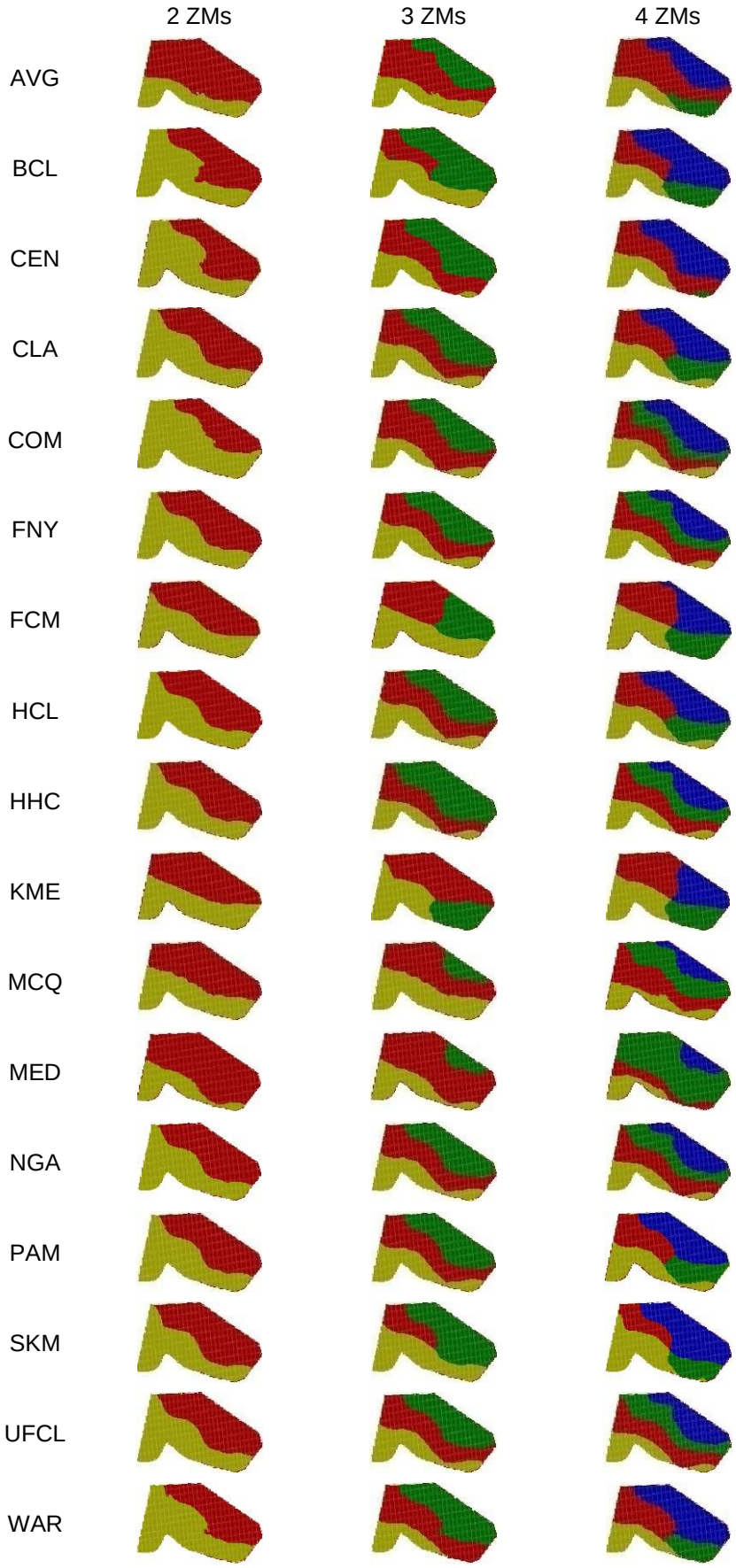


Figura 3 Mapas de zonas de manejo delineadas com o uso dos 17 algoritmos de agrupamento, para a área A.

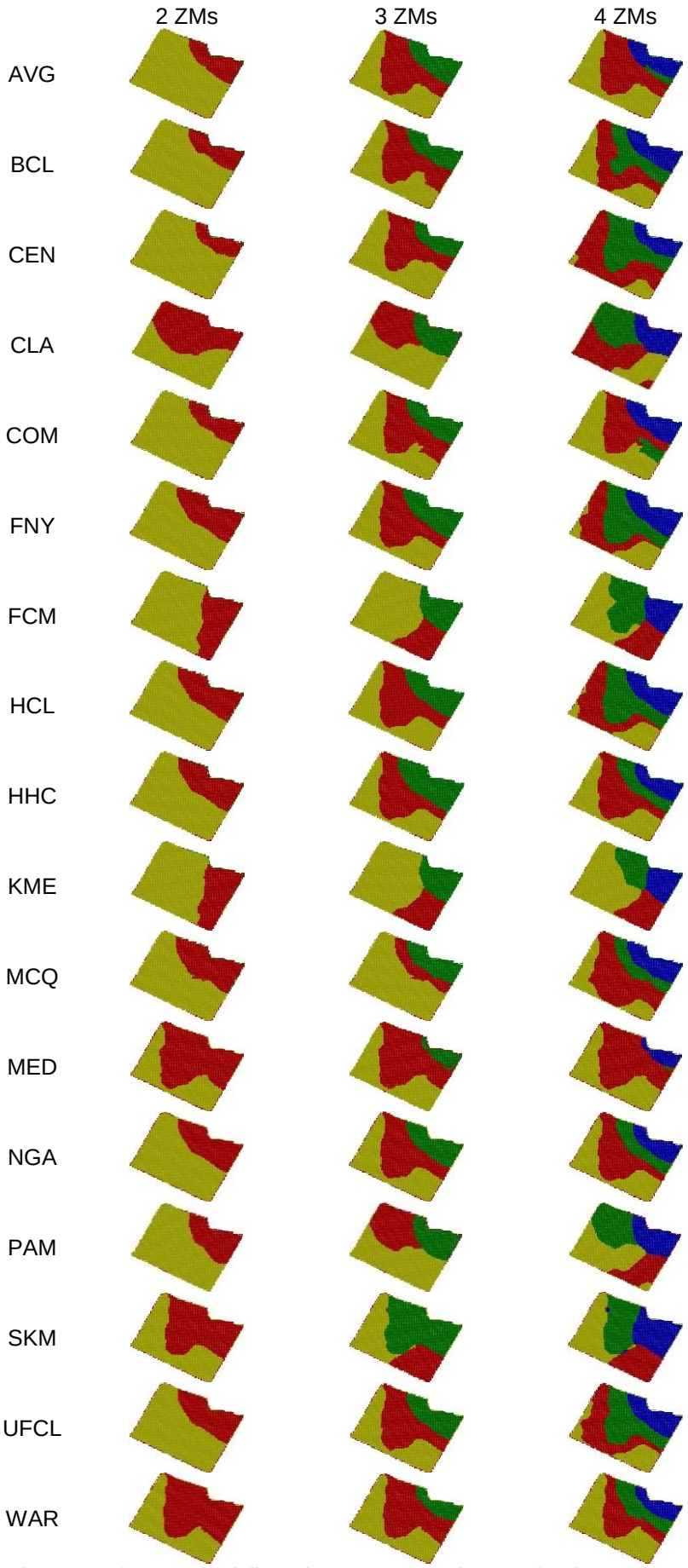


Figura 4 Mapas de zonas de manejo delineadas com o uso dos 17 algoritmos de agrupamento, para a área B.

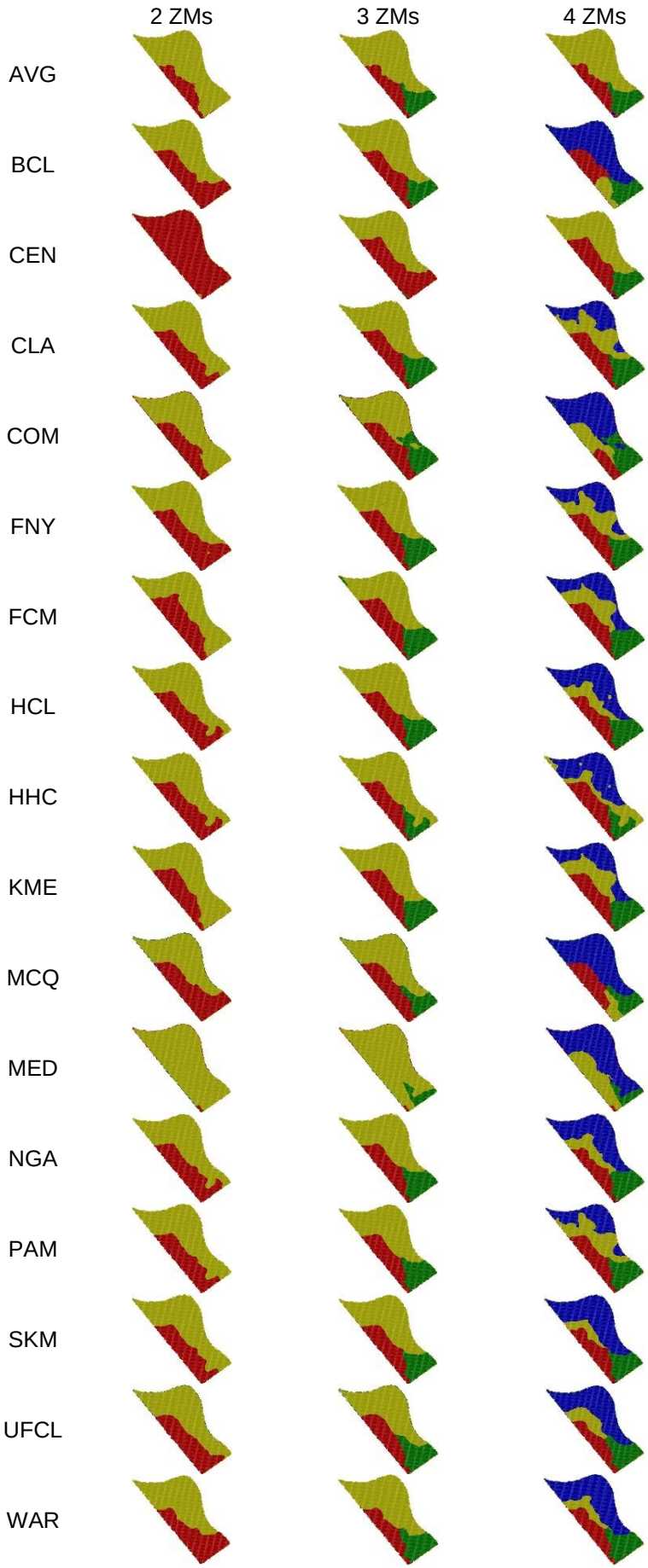


Figura 5 Mapas de zonas de manejo delineadas com o uso dos 17 algoritmos de agrupamento, para a área C.

A avaliação visual dos mapas de ZMs foi complementada com a avaliação quantitativa da concordância entre eles. Para isso, utilizou-se o coeficiente Kappa (Tabelas 8, 9 e 10), com os graus de concordância Kp sendo classificados conforme proposto por Landis e Koch (1977): $0 < Kp \leq 0,2$: não há concordância; $0,2 < Kp \leq 0,4$: fraca; $0,4 < Kp \leq 0,6$: moderada; $0,6 < Kp \leq 0,8$: forte; $0,8 < Kp \leq 1$: muito forte.

Para cada área, foram considerados apenas os mapas que apresentaram as subáreas com produtividades médias estatisticamente distintas. Assim, considerou-se a divisão das áreas em duas ZMs, além da divisão da área C em três. Os mapas de referência para análise das Tabelas 8, 9 e 10 foram aqueles correspondentes aos algoritmos com o maior valor de VR em cada caso (mcquitty para as áreas A e B e fanny para a área C).

Para a área A (Tabela 8), o grau de concordância entre o mapa de mcquitty e os quinze mapas dos demais algoritmos variou de 0,51 a 0,92, ou seja, de concordância moderada a muito forte. O mapa de k-means foi o mais similar ao de mcquitty, enquanto que o de centroid foi o menos similar. Considerando todas as comparações incluídas na Tabela 8, a concordância entre os mapas variou de 0,39 a 0,99. Para a área B (Tabela 9), a concordância entre mcquitty e os sete demais métodos variou de 0,46 a 0,63, isto é, de moderada a forte. Average foi o mais similar a mcquitty, e median foi o menos similar. A variação da concordância entre todos os mapas comparados foi de 0,17 a 0,95.

Para duas ZMs na área C (Tabela 10), a concordância entre fanny e os outros quatorze algoritmos variou de 0,47 a 0,96, ou seja, de moderada a muito forte. Mcquitty foi o mais similar a fanny, enquanto complete foi o menos similar. Já a concordância entre todos os mapas variou de 0,39 a 0,99. Por fim, para três subáreas em C, a concordância entre fanny e complete foi de 0,77, ou seja, concordância forte.

Os valores da concordância Kappa permitiram notar que para as áreas A e C houve maior similaridade entre os mapas de ZMs. Considerando apenas os mapas correspondentes aos três algoritmos com melhores valores de VR, a concordância foi muito forte na área A, fraca em B e forte em C. Dobermann et al. (2003) também mostraram que esse tipo de concordância varia de acordo com os algoritmos de agrupamento utilizados e as características de cada área. Porém, as concordâncias moderadas entre fuzzy c-means e k-means que esses autores obtiveram contrastam com os resultados obtidos neste trabalho, já que os mapas mais similares aos de fuzzy c-means foram os de k-means, e vice-versa.

Tabela 8 Graus de concordância Kappa entre os mapas com duas zonas de manejo estatisticamente distintas, para a área A

Kappa	AVG	BCL	CEN	CLA	FNY	FCM	HCL	HHC	KME	MCQ	MED	NGA	PAM	SKM	UFCL
BCL	0,51														
CEN	0,42	0,85													
CLA	0,53	0,95	0,79												
FNY	0,53	0,86	0,77	0,99											
FCM	0,56	0,67	0,59	0,80	0,81										
HCL	0,55	0,86	0,77	0,98	0,99	0,81									
HHC	0,56	0,86	0,77	0,98	0,99	0,81	0,99								
KME	0,64	0,54	0,44	0,64	0,65	0,92	0,65	0,65							
MCQ	0,68	0,62	0,51	0,71	0,72	0,87	0,72	0,72	0,92						
MED	0,71	0,47	0,39	0,48	0,49	0,57	0,50	0,50	0,59	0,63					
NGA	0,55	0,86	0,80	0,96	0,96	0,78	0,96	0,96	0,62	0,73	0,50				
PAM	0,53	0,87	0,79	0,99	0,99	0,80	0,98	0,98	0,64	0,71	0,48	0,96			
SKM	0,61	0,80	0,75	0,89	0,90	0,80	0,92	0,92	0,65	0,73	0,56	0,92	0,89		
UFCL	0,55	0,86	0,77	0,98	0,98	0,81	0,99	0,99	0,65	0,72	0,50	0,96	0,98	0,92	
WAR	0,42	0,88	0,85	0,86	0,85	0,72	0,83	0,84	0,56	0,62	0,39	0,83	0,86	0,76	0,83

Concordância: muito forte | forte | moderada | fraca | não há

Tabela 9 Graus de concordância Kappa entre os mapas com duas zonas de manejo estatisticamente distintas, para a área B

Kappa	AVG	BCL	CEN	COM	MCQ	MED	SKM
BCL	0,90						
CEN	0,79	0,85					
COM	0,95	0,94	0,80				
MCQ	0,63	0,58	0,47	0,62			
MED	0,23	0,22	0,17	0,23	0,46		
SKM	0,35	0,33	0,25	0,34	0,62	0,79	
WAR	0,27	0,26	0,20	0,27	0,53	0,85	0,87

Concordância: muito forte | forte | moderada | fraca | não há

Tabela 10 Graus de concordância Kappa entre os mapas com duas zonas de manejo estatisticamente distintas, para a área C

Kappa	AVG	BCL	CLA	COM	FNY	FCM	HCL	HHC	KME	MCQ	NGA	PAM	SKM	UFCL
BCL	0,52													
CLA	0,63	0,82												
COM	0,95	0,52	0,63											
FNY	0,48	0,92	0,80	0,47										
FCM	0,57	0,58	0,75	0,57	0,64									
HCL	0,58	0,86	0,94	0,58	0,86	0,71								
HHC	0,59	0,86	0,95	0,59	0,85	0,71	0,99							
KME	0,70	0,59	0,74	0,69	0,62	0,85	0,69	0,70						
MCQ	0,47	0,93	0,79	0,47	0,96	0,61	0,85	0,84	0,59					
NGA	0,59	0,85	0,94	0,58	0,86	0,72	0,99	0,99	0,70	0,84				
PAM	0,63	0,81	0,99	0,63	0,80	0,76	0,94	0,94	0,75	0,79	0,94			
SKM	0,67	0,80	0,96	0,67	0,76	0,75	0,89	0,91	0,75	0,76	0,90	0,96		
UFCL	0,39	0,74	0,74	0,39	0,77	0,59	0,77	0,77	0,53	0,76	0,77	0,74	0,70	
WAR	0,52	0,87	0,71	0,52	0,85	0,48	0,76	0,75	0,45	0,88	0,75	0,70	0,69	0,65

Concordância: muito forte | forte | moderada | fraca | não há

Em relação ao percentual da área total que correspondeu a cada ZM definida, os melhores métodos de agrupamento proporcionaram os valores apresentados na Tabela 11. Nesta tabela, também estão os valores do coeficiente de variação (CV) da produtividade média padronizada, antes e depois da geração das subáreas.

Tabela 11 Porcentagem da área total ocupada por cada zona de manejo e valores do coeficiente de variação da produtividade média padronizada, antes e depois da definição das subáreas

Área	Situação	Porcentagem da área	CV (%)
A	área sem ZMs	100	15,6
	mcquitty (2 ZMs)	ZM_1: 42,4	ZM_1: 4,2
		ZM_2: 57,6	ZM_2: 3,6
B	área sem ZMs	100	8,1
	mcquitty (2 ZMs)	ZM_1: 67,2	ZM_1: 3,8
		ZM_2: 32,8	ZM_2: 2,2
C	área sem ZMs	100	18,6
	fanny (2 ZMs)	ZM_1: 41,2	ZM_1: 5,1
		ZM_2: 58,8	ZM_2: 4,3
	fanny (3 ZMs)	ZM_1: 24,6	ZM_1: 3,8
		ZM_2: 19,2	ZM_2: 4,4
		ZM_3: 56,2	ZM_3: 3,9

CV: coeficiente de variação; ZM_i: zona de manejo *i*.

Nas três áreas, houve redução do CV com a definição das ZMs por meio da utilização de mcquitty e fanny. Na área A, reduziu-se de 15,6 para 4,2 e 3,6% dentro das subáreas; na área B, de 8,1 para 3,8 e 2,2%; e na área C, de 18,6% para valores entre 3,8 e 5,1%, para duas ou três zonas. Também se observou que apesar das ZMs corresponderem a percentuais bastante diferentes da área total, isto não interferiu na redução de todos os valores do CV. Dobermann et al. (2003) também obtiveram resultados similares a esses.

Utilizou-se novamente o teste de Tukey para verificar se as ZMs definidas pelos melhores algoritmos (mcquitty e fanny, além de fuzzy c-means e k-means como segundo e terceiro melhores) também apresentaram diferenças significativas em relação a variáveis químicas, físicas e topográficas das áreas (Tabelas 12, 13 e 14). Observou-se a frequência com que cada variável apresentou valores estatisticamente diferentes nas ZMs, para definir a ordenação destas variáveis em cada uma das tabelas.

Tabela 12 Teste de comparação de médias de Tukey para variáveis correspondentes à área A, considerando duas zonas de manejo definidas com os algoritmos mcquitty, fuzzy c-means e k-means

Variáveis*	2 ZMs		
	mcquitty	fuzzy c-means	k-means
Produtividade média	2	2	2
Altitude	2	2	2
Areia	2	2	2
Cu	2	2	2
RSP 0-0,1 m	2	2	2
Argila	2	2	
Declividade	2		
RSP 0,1-0,2 m	2		

*: Al, C, Ca, densidade, Fe, H+Al, K, macroporosidade, matéria orgânica, Mg, microporosidade, Mn, P, pH, RSP 0,2-0,3 m, silte e Zn: diferenças não significativas.
2: diferença significativa entre as médias da variável nas duas ZMs, a 5%.

Tabela 13 Teste de comparação de médias de Tukey para variáveis correspondentes à área B, considerando duas zonas de manejo definidas com os algoritmos mcquitty, fuzzy c-means e k-means

Variáveis*	2 ZMs		
	mcquitty	fuzzy c-means	k-means
Produtividade média	2		
Altitude	2		
Argila	2		
Ca	2		
K	2		
Matéria orgânica	2		
RSP 0-0,1 m	2		
RSP 0,1-0,2 m	2		

*: Al, areia, C, Cu, declividade, densidade, Fe, H+Al, macroporosidade, Mg, microporosidade, Mn, P, pH, RSP 0,2-0,3 m, silte e Zn: diferenças não significativas.
2: diferença significativa entre as médias da variável nas duas ZMs, a 5%.

Tabela 14 Teste de comparação de médias de Tukey para variáveis correspondentes à área C, considerando duas e três zonas de manejo definidas com os algoritmos fanny, fuzzy c-means e k-means

Variáveis*	2 ZMs			3 ZMs		
	fanny	fuzzy c-means	k-means	fanny	fuzzy c-means	k-means
Produtividade média	2	2	2	3	3	3
Altitude	2	2	2	3	3	3
Argila	2	2	2	3	3	3
Areia	2	2	2	3	2	2
RSP 0,1-0,2 m	2	2	2	3	2	2
RSP 0,2-0,3 m	2	2	2	2	2	2
Silte	2	2	2	2	2	2
C	2	2		2	2	2
Cu	2			3	2	2
Densidade	2			3	2	2
P				2	2	2
Declividade				2		

*: Al, Ca, Fe, H+Al, K, macroporosidade, matéria orgânica, Mg, microporosidade, Mn, pH, RSP 0-0,1 m e Zn: diferenças não significativas.
2 ou 3: diferença significativa entre as médias da variável em duas ou em três ZMs, a 5%.

As variáveis que mais apresentaram médias estatisticamente distintas após a divisão das áreas em ZMs foram altitude, argila, areia e RSP 0,1–0,2 m. Além disso, com exceção da areia, essas variáveis foram as únicas que mostraram diferenças significativas para as três áreas. Esses resultados parecem ser consequências do fato de que essas quatro variáveis foram as mais importantes na formação das variáveis CPE1 e CPE2, utilizadas diretamente na geração das subáreas. Vale destacar ainda os resultados obtidos por cobre, silte, RSP 0–0,1 m e RSP 0,2–0,3 m.

Analisando os dados das Tabelas 12, 13 e 14 sob a perspectiva de comparação entre os algoritmos de agrupamento, concluiu-se que mcquitty e fanny foram superiores a fuzzy c-means e k-means nas quatro situações de comparação. Isto corrobora os resultados discutidos em relação ao índice VR e ao coeficiente ASC. Dessas quatro situações, a correspondente a duas ZMs para a área B (Tabela 13) merece destaque, pois fuzzy c-means

e k-means não foram capazes de gerar subáreas estatisticamente diferentes em relação a qualquer uma das variáveis consideradas.

Na Figura 6, mostra-se o nível de concordância entre os mapas de ZMs gerados a partir das duas CPEs e os mapas da produtividade média interpolada.

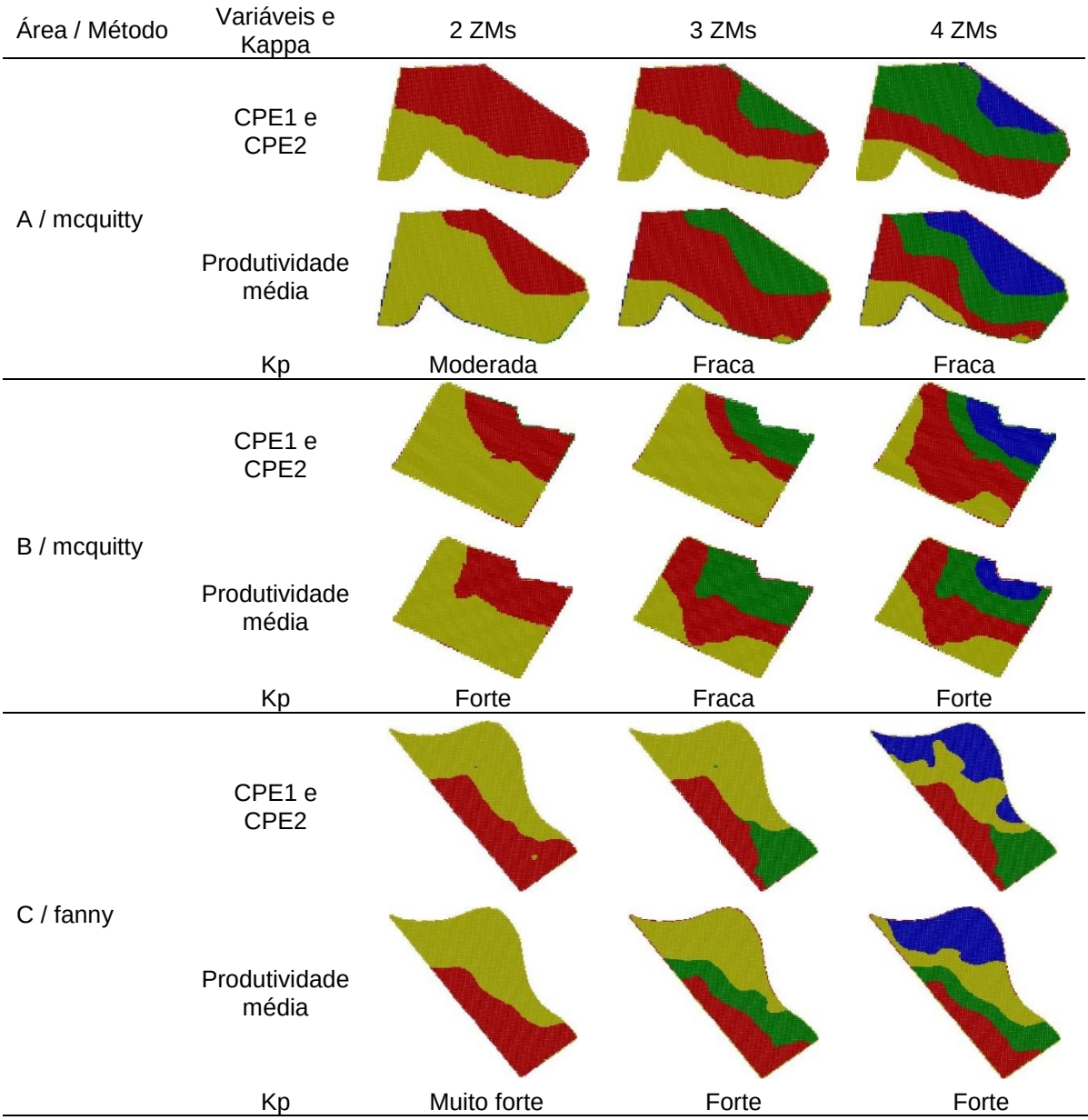


Figura 6 Mapas de zonas de manejo gerados a partir das duas primeiras CPEs e mapas para a produtividade média padronizada, com os respectivos níveis de concordância Kappa, para as três áreas.

Ao analisar a Figura 6, notou-se que houve maior concordância entre os mapas da área C, com níveis de Kappa entre forte e muito forte. Estes níveis foram considerados satisfatórios, principalmente pelo fato de que existem outros fatores limitantes da produtividade, além das variáveis do solo consideradas neste trabalho. Ao comparar os níveis de concordância dos mapas correspondentes à área C com aqueles da área A, observou-se que há uma grande diferença entre eles. Isto indica que características específicas do solo de

cada área devem ser consideradas no processo de definição de ZMs, conforme destacado por Kitchen et al. (2005). Também se considera importante obter o máximo possível de conhecimento sobre o histórico de manejo de cada área, como um auxílio para tentar compreender o porquê dessa diferença.

6.4 Conclusões

A aplicação dos algoritmos de agrupamento mcquitty e fanny resultou em zonas de manejo com vários tamanhos e formatos diferentes, pelo fato de serem influenciados pelas características de cada área e pela quantidade solicitada de subáreas. Todavia, isso não interferiu no desempenho superior desses métodos em comparação com os outros quinze que foram avaliados, principalmente na redução da variância original da produtividade. Com mcquitty e fanny foram geradas classes com elevada homogeneidade interna, que conduziram ao delineamento de subáreas praticamente sem fragmentações, isto é, subáreas adequadas para a execução de operações de campo. Portanto, esses resultados permitem concluir que esses dois métodos de agrupamento foram os melhores na definição de zonas de manejo para as áreas consideradas.

Os outros quinze algoritmos também alcançaram resultados que os habilitaram para serem usados na geração de duas zonas de manejo em pelo menos duas das áreas consideradas. Especificamente em relação aos métodos fuzzy c-means e k-means, sua utilização gerou classes que conduziram a subáreas com produtividades médias estatisticamente distintas para duas áreas; mas os resultados obtidos por esses dois algoritmos nessas áreas foram inferiores aos alcançados por mcquitty e fanny.

6.5 Referências

ARNO, J.; MARTINEZ-CASASNOVAS, J. A.; RIBES-DASI, M.; ROSELL, J. R. Clustering of grape yield maps to delineate site-specific management zones. **Spanish Journal of Agricultural Research**, v. 9, n. 3, p. 721-729, 2011.

BALL, G. H.; HALL, D. J. A clustering technique for summarizing multivariate data. **Systems Research and Behavioral Science,** v. 12, n. 2, p. 153-155, 1967.

BAZZI, C. L.; SOUZA, E. G.; URIBE-OPAZO, M. A.; NÓBREGA, L. H. P.; ROCHA, D. M. Management zones definition using soil chemical and physical attributes in a soybean area. **Engenharia Agrícola**, v. 33, n. 5, p. 952-964, 2013.

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. New York: Plenum Press, 1981. 256 p.

BOYDELL, B.; MCBRATNEY, A. B. Identifying potential within-field management zones from cotton-yield estimates. **Precision Agriculture,** v. 3, n. 1, p. 9-23, 2002.

CAIRES, S. A.; WUDDIVIRA, M. N.; BEKELE, I. Spatial analysis for management zone delineation in a humid tropic cocoa plantation. **Precision Agriculture,** v. 16, p. 129-147, 2015.

CHIPMAN, H.; TIBSHIRANI, R. Hybrid Hierarchical Clustering with Applications to Microarray Data. **Biostatistics,** v. 7, p. 302-317, 2006.

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, p. 20-28, 2014.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement,** v. 20, p. 37-46, 1960.

CÓRDOBA, M.; BRUNO, C.; COSTA, J. L.; 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.

CZAPLEWSKI, R. L.; REICH, R. M. **Expected value and variance of Moran's bivariate spatial autocorrelation statistic under permutation**. Research Paper RM-309. Fort Collins: USDA Forest Service, 1993. 13 p.

DAVE, R. N. Generalized fuzzy c-shells clustering and detection of circular and elliptical boundaries. **Pattern Recognition,** v. 25, n. 7, p. 713-721, 1992.

DELALIBERA, H. C.; WEIRICH, P. H.; NAGATA, N. Management zones in agriculture according to the soil and landscape variables. **Engenharia Agrícola,** v. 32, n. 6, p. 1197-1204, 2012.

DHILLON, I. S.; MODHA, D. S. Concept decompositions for large sparse text data using clustering. **Machine Learning,** v. 42, p. 143-175, 2001.

DIGGLE, P. J.; RIBEIRO JR, P. J. **Model-based Geostatistics**. New York: Springer-Verlag, 2007. 232 p.

DOBERMANN, A.; PING, J. L.; ADAMCHUK, V. I.; SIMBAHAN. G. C.; FERGUSON, R. B. Classification of crop yield variability in irrigated production fields. **Agronomy Journal**, v. 95, n. 1, p. 1105-1120, 2003.

DOERGE, T. A. **Site-specific management guidelines**. Norcross: Potash & Phosphate Institute, 2000. 135 p.

DRAY, S.; SAID, S.; DÉBIAS, F. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. **Journal of Vegetation Science,** v. 19, p. 45-56, 2008.

EMBRAPA. Centro Nacional de Pesquisa de Solos. **Manual de Métodos de Análise de Solo.** 2. ed. Rio de Janeiro: CNPSO, 1997. 212 p.

EMBRAPA. Centro Nacional de Pesquisa de Solos. **Sistema brasileiro de classificação de solo**. Rio de Janeiro: CNPSO, 2013. 412 p.

FERREIRA, D. F. **Análise multivariada**. Lavras: UFLA, 1996. 394 p.

FLEMING, K. L.; WESTFALL, D. G.; WIENS, D. W.; BRODAHL, M. C. Evaluating farmer developed management zone maps for variable rate fertilizer application. **Precision Agriculture,** v. 2, p. 201-215, 2000.

FRIDGEN, J. J.; KITCHEN, N. R.; SUDDUTH, K. A.; DRUMMOND, S. T.; WIEBOLD, W. J.; FRAISSE, C. W. Management zone analyst (MZA): software for subfield management zone delineation. **Agronomy Journal**, v. 96, p. 100-108, 2004.

GAVIOLI, A.; SOUZA, E. G.; BAZZI, C. L.; GUEDES, L. P. C.; SCHENATTO, K. Optimization of management zone delineation by using spatial principal components. **Computers and Electronics in Agriculture,** v. 127, p. 302-310, 2016.

GUASTAFERRO, F.; CASTRIGNANO, A.; DE BENEDETTO, D.; SOLLITTO, D.; TROCCOLI, A.; CAFARELLI, B. A comparison of different algorithms for the delineation of management zones. **Precision Agriculture,** v. 11, p. 600-620, 2010.

IKENAGA, S.; INAMURA, T. Evaluation of site-specific management zones on a farm with 124 contiguous small paddy fields in a multiple-cropping system. **Precision Agriculture,** v. 9, p. 147-159, 2008.

JAIN, A. K.; DUBES, R. **Algorithms for clustering data**. Englewood Cliffs: Prentice-Hall, 1988. 320 p.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining geostatistics**. London: The Blackburn Press, 2004. 600 p.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data**. Hoboken: John Wiley & Sons, 1990. 342 p.

KITCHEN, N. R.; SUDDUTH, K. A.; MYERS, D. B.; DRUMMOND, S. T.; HONG, S. Y. Delineating productivity zones on claypan soil fields using apparent soil electrical conductivity. **Computers and Electronics in Agriculture,** v. 46, p. 285-308, 2005.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics,** v. 33, n. 1, p. 159-174, 1977.

LEISCH, F. Bagged clustering. In: SFB ADAPTIVE INFORMATION SYSTEMS AND MODELLING IN ECONOMICS AND MANAGEMENT SCIENCE, 51, 1999, Vienna. **Anais...** Vienna: Vienna University of Economics and Business, 1999. p. 1-11.

LI, X.; PAN, Y.; GE, Z.; ZHAO, C. Delineation and scale effect of precision agriculture management zones using yield monitor data over four years. **Agricultural Sciences in China**, v. 6, n. 2, p. 180-188, 2007.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: PROCEEDINGS OF FIFTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 1967, Berkeley. **Anais...** Berkeley: University of California Press, 1967. p. 281–297.

MARTINETZ, T. M.; BERKOVICH, S. G.; SCHULTEN, K. J. "Neural-gas" network for vector quantization and its application to time-series prediction. **IEEE Transactions on Neural Networks,** v. 4, n. 4, p. 558-569, 1993.

MATHERON, G. Principles of Geostatistics. **Economic Geology,** v. 58, n. 8, p. 1246-1266, 1963.

MCQUITTY, L. L. Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. **Educational and Psychological Measurement,** v. 26, p. 825-831, 1966.

MIELKE JR, P. W.; BERRY, K. J. **Permutation methods: a distance function approach**. New York: Springer, 2007. 446 p.

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.

MORAL, F. J.; TERRÓN, J. M.; SILVA, J. R. M. Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. **Soil and Tillage Research**, v. 106, n. 2, p. 335-343, 2010.

NANNI, M. R.; POVH, F. P.; DEMATTÊ, J. A. M.; OLIVEIRA, R. B.; CHICATI, M. L.; CEZAR, E. Optimum size in grid soil sampling for variable rate application in site-specific management. **Scientia Agricola,** v. 68, n. 3, p. 386-392, 2011.

ORTEGA, R. A.; SANTIBÁÑEZ, O. A. 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, 2007.

PAL, N. R.; BEZDEK, J. C.; HATHAWAY, R. J. Sequential competitive learning and the fuzzy c-means clustering algorithm. **Neural Networks,** v. 9, n. 5, p. 787-796, 1996.

PIMENTEL-GOMES, F. **Curso de Estatística Experimental.** 14 ed. Piracicaba: Universidade de São Paulo, 2000. 477 p.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2016. 99 p.

ROUSSEEUW, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics,** v. 20, p. 53-65, 1987.

RUSS, G.; KRUSE, R. Machine Learning Methods for Spatial Clustering on Precision Agriculture Data. **SCAI,** v. 11, p. 40-49, 2011.

SCHENATTO, K.; SOUZA, E. G.; BAZZI, C. L.; BIER, V. A.; BETZEK, N. M.; GAVIOLI, A. Data interpolation in the definition of management zones. **Acta Scientiarum,** v. 38, n. 1, p. 31-40, 2016.

TAYLOR, J. C.; WOOD, G. A.; EARL, R.; GODWIN, R. J. Soil factors and their influence on within-field crop variability, part II: spatial analysis and determination of management zones. **Biosystems Engineering,** v. 84, p. 441-453, 2003.

WARD, J. H. Hierarchical Grouping to Optimize an Objective Function. **Journal of the American Statistical Association,** v. 58, n. 301, p. 236-244, 1963.

WOLLENHAUPT, N. C.; WOLKOWSKI, R. P.; CLAYTON, M. K. Mapping soil test phosphorus and potassium for variable-rate fertilizer application. **Journal of Production Agriculture**, v. 7, n. 4, p. 441-448, 1994.

XU, R.; WUNSCH, D. C. **Clustering**. Piscataway: IEEE Press, 2009. 358 p.

7 ARTIGO 3 – IMPLEMENTAÇÃO DE MÉTODOS DE SELEÇÃO DE VARIÁVEIS E AGRUPAMENTO DE DADOS PARA GERAÇÃO DE ZONAS DE MANEJO

Resumo
Para a definição de zonas de manejo (ZMs) de qualidade, duas atividades fundamentais são a seleção das variáveis a serem utilizadas e a análise de agrupamento de dados. Existem diversos métodos propostos para executá-las, mas que, devido à sua complexidade, precisam ser disponibilizados por ferramentas computacionais. Diante disso, o objetivo deste trabalho foi apresentar dois módulos computacionais desenvolvidos para possibilitar a execução eficiente dessas duas atividades. O módulo de seleção de variáveis disponibiliza 5 algoritmos baseados em análise da correlação espacial entre variáveis, análise de componentes principais (ACP) e no método derivado de ACP denominado MULTISPATI-PCA. Já o módulo de geração de classes disponibiliza 17 algoritmos de agrupamento de dados: average linkage, bagged clustering, centroid linkage, clustering large applications, complete linkage, fuzzy analysis clustering, fuzzy c-means, hard competitive learning, hybrid hierarchical clustering, k-means, median linkage, método de McQuitty, método de Ward, neural gas, partitioning around medoids, spherical k-means e unsupervised fuzzy competitive learning. Para exemplificar o funcionamento desses módulos, foram empregados dados obtidos entre os anos de 2012 e 2015 de uma área agrícola localizada no município de Céu Azul, estado do Paraná, na qual cultivou-se soja. Os módulos computacionais desenvolvidos mostraram-se eficientes para a definição de ZMs. Além disso, são mais abrangentes que outros softwares de uso gratuito, como FuzME, MZA e SDUM, no que tange à diversidade de algoritmos de seleção de variáveis e de agrupamento de dados disponibilizados.

Palavras-chave: agricultura de precisão; análise de componentes principais; clusterização de dados; MULTISPATI-PCA; software para agricultura.

IMPLEMENTATION OF VARIABLE SELECTION AND DATA CLUSTERING METHODS FOR GENERATION OF MANAGEMENT ZONES

Abstract
Two basic activities for the definition of quality management zones (MZs) are the variable selection task and the cluster analysis task. There are several methods proposed to execute them, but due to their complexity, they need to be made available by computer systems. In this context, the objective of this study was to present two computational modules developed to enable the efficient execution of these two activities. The variable selection module provides 5 algorithms based on spatial correlation analysis, principal component analysis (PCA) and the PCA-based method called MULTISPATI-PCA. The class generation module provides 17 data clustering algorithms, as follows: average linkage, bagged clustering, centroid linkage, clustering large applications, complete linkage, fuzzy analysis clustering, fuzzy c-means, hard competitive learning, hybrid hierarchical clustering, k-means, McQuitty's method, median linkage, neural gas, partitioning around medoids, spherical k-means, unsupervised fuzzy competitive learning, and Ward's method. To exemplify the execution of these modules, data were obtained between 2012 and 2015 in an agricultural area in the municipality of Céu Azul, state of Paraná, where soybean was cultivated. The computational modules developed proved to be efficient to define MZs. Furthermore, they were more complete than other free-to-use software such as FuzME, MZA, and SDUM, in terms of the diversity of variable selection and data clustering methods.

Keywords: data clusters; MULTISPATI-PCA; precision agriculture; principal component analysis; software for agriculture.

7.1 Introdução

A aplicação da tecnologia da informação (TI) e o desenvolvimento de dispositivos e máquinas têm sido fundamentais para o desenvolvimento da agricultura de precisão (AP) (EL-SHARKAWY et al., 2016). Diante do fato de que as práticas da AP requerem investimentos financeiros altos, uma alternativa para tornar essa abordagem mais atrativa para pequenos produtores é a implantação de zonas de manejo (ZMs) em áreas agrícolas.

De acordo com Chang et al. (2014), cada ZM é uma subárea que apresenta um conjunto de características similares, geralmente limitantes da produtividade, para a qual pode-se adotar a aplicação a taxa fixa de insumos agrícolas. Assim, cada ZM pode ser tratada como uma área homogênea sob o ponto de vista de amostragem e gerenciamento. Em comparação com a agricultura convencional, a utilização dessas subáreas possibilita reduzir a variabilidade espacial da produtividade das culturas e os danos ao meio ambiente provocados pela aplicação excessiva de determinados insumos.

Dentre os métodos mais utilizados para a geração de ZMs, estão os algoritmos de agrupamento de dados, que têm o propósito de dividir os pontos georreferenciados de uma área em certo número de classes, com o uso de algum critério para determinar o nível de similaridade entre esses pontos. Na prática, essas classes ou grupos são empregados para delimitar as ZMs no campo. Além dos algoritmos k-means (MACQUEEN, 1967) e fuzzy c-means (BEZDEK, 1981), outros métodos de agrupamento também podem ser aplicados com sucesso para definir essas subáreas (DOBERMANN et al., 2003; GUASTAFERRO et al., 2010; XU; WUNSCH, 2009).

Os métodos de agrupamento podem usar muitas variáveis na geração de ZMs, que denotem condições do solo, do relevo e/ou de plantas cultivadas. Entretanto, selecionar as variáveis necessárias é uma tarefa básica para assegurar que os algoritmos de agrupamento possam definir ZMs adequadas.

Cohen et al. (2013), Li et al. (2013) e Tripathi et al. (2015) utilizaram análise de componentes principais (ACP) (HOTELLING, 1933) para selecionar variáveis para a geração de ZMs. Para essa mesma tarefa, Córdoba et al. (2013) e Peralta et al. (2015) aplicaram a abordagem denominada análise multivariada espacial baseada no índice de Moran e em ACP (MULTISPATI-PCA) (DRAY; SAID; DÉBIAS, 2008). Bazzi et al. (2013) e Schenatto et al. (2016) empregaram uma abordagem fundamentada na análise de correlação espacial entre variáveis (REICH; CZAPLEWSKI; BECHTOLD, 1994). Já Gavioli et al. (2016) avaliaram 5 métodos de seleção de variáveis, incluindo um novo algoritmo baseado na aplicação conjunta de MULTISPATI-PCA e análise de correlação espacial.

A seleção de variáveis e a geração de classes por meio do uso de algoritmos de agrupamento são atividades de execução complexa, que dependem do uso de softwares apropriados. Porém, os principais sistemas computacionais desenvolvidos para tratar da

definição de ZMs disponibilizam poucos métodos para essas duas atividades. É o caso de FuzME (MINASNY; MCBRATNEY, 2002), Management Zone Analyst (MZA) (FRIDGEN et al., 2004) e do Software para Definição de Unidades de Manejo (SDUM) (BAZZI et al., 2013).

Neste contexto, o objetivo deste trabalho é apresentar dois módulos computacionais desenvolvidos para possibilitar a seleção de variáveis e a geração de classes para o delineamento de ZMs. Como diferencial, esses módulos disponibilizam 5 métodos de seleção de variáveis baseados em ACP, MULTISPATI-PCA e análise de correlação espacial, e 17 métodos de agrupamento de dados.

7.2 Material e métodos

7.2.1 Softwares utilizados

Durante o desenvolvimento dos dois módulos computacionais, empregaram-se softwares para a programação dos algoritmos de seleção de variáveis e de agrupamento de dados e para a construção da interface gráfica da futura versão desses módulos para a Internet. Todos os softwares selecionados estavam disponíveis para download na Internet e puderam ser utilizados de forma gratuita.

Os 5 algoritmos de seleção de variáveis e os 17 algoritmos de agrupamento foram programados em rotinas do software estatístico R (R CORE TEAM, 2016), com o principal objetivo de garantir rapidez na execução. Nas rotinas de implementação da seleção de variáveis, foram incluídos os pacotes geoR, gstat, ade4 e spdep. Já nas rotinas correspondentes aos métodos de agrupamento, foram utilizados os pacotes cclust, cluster, e1071, fastcluster, fclust, hybridHclust, optpart e skmeans.

Para o projeto da interface, inicialmente utilizou-se a ferramenta computacional Pencil 2.0.5 (Evolus Software). Com ela, realizou-se a prototipação das telas que futuramente constituirão a interface gráfica dos módulos computacionais implementados. Em seguida, empregou-se a linguagem de marcação de hipertexto HTML (HyperText Markup Language) para construir as páginas web correspondentes aos protótipos das telas. Mas como os dois módulos serão disponibilizados em uma aplicação web[2] em um momento futuro, as páginas web são apresentadas neste artigo apenas para exemplificar o uso dos métodos implementados nesses módulos.

7.2.2 Métodos implementados

Na Figura 1, representam-se atividades do processo de geração de ZMs, com destaque para os métodos incluídos nos módulos de seleção de variáveis e de agrupamento.

[2] De forma geral, é uma aplicação computacional projetada para ser utilizada empregando-se aplicativos de navegação na Internet (web browsers).

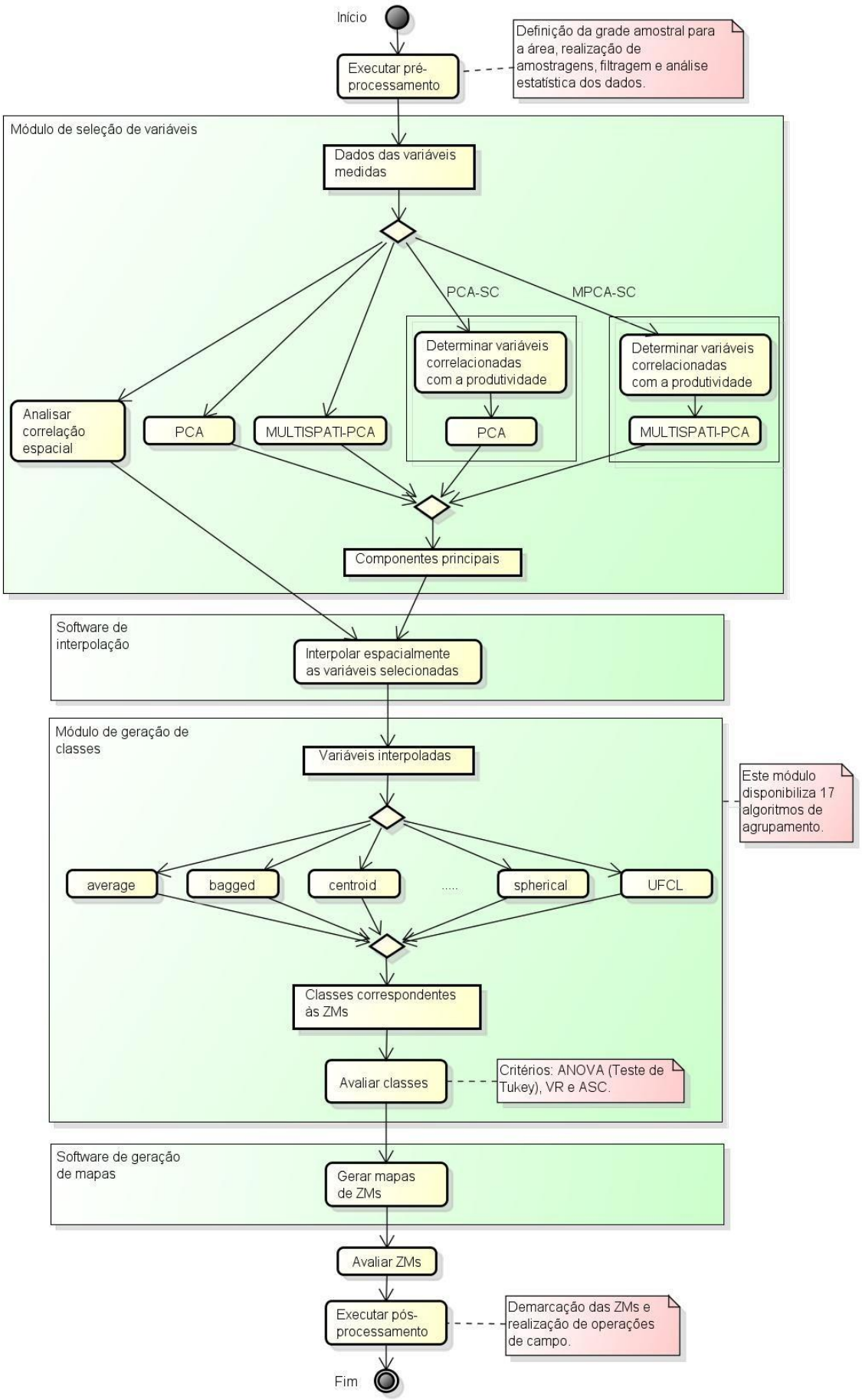


Figura 1 Diagrama representando atividades e algoritmos para a definição de zonas de manejo, com destaque para os módulos desenvolvidos para seleção de variáveis e geração de classes.

Para utilizar o módulo de seleção de variáveis, deve-se ter os dados georreferenciados correspondentes às variáveis do solo e/ou das plantas cultivadas de uma área. Esses dados devem estar armazenados em um arquivo de texto, em que as coordenadas geográficas (sistema UTM) dos pontos estejam nas duas primeiras colunas e, a partir da terceira, se tenha uma coluna para cada variável. Assim, podem-se executar as rotinas que constituem o módulo de seleção de variáveis para ler o arquivo de texto e realizar a seleção das variáveis necessárias para a definição de ZMs. As rotinas implementadas correspondem aos 5 métodos de seleção comparados por Gavioli et al. (2016), ou seja:

- PCA-All: executa a ACP sobre as variáveis originais disponíveis, para gerar as novas variáveis sintéticas, isto é, as componentes principais (CPs);
- MPCA-All: aplica a abordagem MULTISPATI-PCA sobre as variáveis originais disponíveis, a fim de definir as CPs correspondentes – que para MULTISPATI-PCA também são chamadas de CPs espaciais (CPEs);
- PCA-SC: verifica quais variáveis originais apresentam correlação espacial significativa com a produtividade da área, a um dado nível de significância; em seguida, sobre as variáveis que satisfizerem essa condição, aplica a ACP para gerar as CPs;
- MPCA-SC: executa-se o mesmo procedimento descrito para PCA-SC, mas substituindo a aplicação de ACP por MULTISPATI-PCA, para gerar as CPEs;
- Análise de correlação espacial: calcula a estatística de autocorrelação espacial bivariada de Moran (REICH; CZAPLEWSKI; BECHTOLD, 1994) entre todas as variáveis disponíveis; em seguida, seleciona as variáveis pelo procedimento proposto por Bazzi et al. (2013): 1) eliminação das variáveis com autocorrelação espacial não significativa a um dado nível de significância; 2) remoção das variáveis que não possuírem correlação significativa com a produtividade; 3) ordenação decrescente das variáveis restantes, considerando o módulo do valor da correlação com a produtividade; e 4) eliminação das variáveis redundantes (que apresentem correlação entre si), mantendo as que possuírem maior correlação com a produtividade.

No caso da aplicação de PCA-All, MPCA-All, PCA-SC ou MPCA-SC, devem-se selecionar as CPs que serão efetivamente utilizadas na geração das ZMs, seguindo um critério para definição da quantidade necessária dessas variáveis. Alguns desses critérios foram sugeridos por Ferreira (1996), Johnson e Wichern (2007) e Jolliffe (2002). Para os métodos MPCA-All e MPCA-SC, também deve-se definir um valor em metros para o parâmetro denominado raio de vizinhança. Esse valor é utilizado para a construção de uma matriz necessária para a execução desses métodos, chamada de matriz de vizinhos dos pontos georreferenciados (CÓRDOBA et al., 2013). Como o valor do raio de vizinhança é determinado de forma empírica para cada área, esses dois métodos devem ser testados com

diferentes valores a fim de descobrir um que conduza à definição de ZMs satisfatórias. Córdoba et al. (2012) sugeriram iniciar esses testes com o raio assumindo o valor de 50% da maior distância entre dois pontos amostrais da área e, em seguida, testar outros valores próximos a esse valor inicial para comparar os resultados gerados. Neste sentido, deve-se utilizar como raio de vizinhança o valor que conduzir à formação das melhores ZMs.

O resultado da execução das rotinas correspondentes aos métodos de seleção de variáveis é gravado em um arquivo de texto. O padrão de gravação dos dados nesse arquivo de resposta é o mesmo do arquivo de entrada. Portanto, as coordenadas geográficas dos pontos ocupam as duas primeiras colunas e, a partir da terceira, tem-se uma coluna para cada CP ou variável original selecionada. O código-fonte correspondente à implementação dos métodos de seleção de variáveis no software R é apresentado no Apêndice A.

Para executar a interpolação espacial das CPs ou variáveis originais selecionadas, deve-se utilizar um software que disponibilize algoritmos de interpolação recomendados para o contexto de criação de ZMs, como inverso da distância elevado a uma potência e, principalmente, krigagem. Durante o desenvolvimento deste trabalho, empregou-se o método krigagem ordinária, tendo em vista a constatação da existência de dependência espacial entre amostras. Executou-se a krigagem ordinária por meio do uso de uma rotina implementada no software R, que também possibilitou gerar gráficos para a análise visual da variabilidade espacial das variáveis consideradas.

Para utilizar o módulo de geração de classes para ZMs, são necessários dados georreferenciados resultantes da interpolação das variáveis ou CPs previamente selecionadas. Esses dados devem estar armazenados em um arquivo de texto com a mesma estrutura do arquivo de entrada de dados do módulo de seleção de variáveis. Assim, pode-se executar as rotinas que constituem o módulo de geração de classes para ler o arquivo de texto e dividir os dados interpolados em certa quantidade de classes. As rotinas implementadas correspondem aos 17 métodos de agrupamento (Tabela 1) que foram selecionados por meio da avaliação descrita no artigo 2 (Capítulo 6).

O resultado da execução da rotina correspondente a qualquer um dos 17 métodos de agrupamento também é gravado em um arquivo de texto. Nesse arquivo, as coordenadas dos pontos ocupam as duas primeiras colunas e na terceira são inseridos números inteiros que identificam as classes às quais os pontos foram associados. O código-fonte correspondente à implementação desses 17 algoritmos no software R é apresentado no Apêndice B.

Tabela 1 Métodos de agrupamento disponibilizados no módulo de geração de classes para zonas de manejo

Métodos de agrupamento	Referências
average linkage	Jain e Dubes (1988)
bagged clustering	Leisch (1999)
centroid linkage	Jain e Dubes (1988)
complete linkage	Jain e Dubes (1988)
clustering large applications	Kaufman e Rousseeuw (1990)
fuzzy analysis clustering	Kaufman e Rousseeuw (1990)
fuzzy c-means	Bezdek (1981)
hard competitive learning	Xu e Wunsch (2009)
hybrid hierarchical clustering	Chipman e Tibshirani (2006)
k-means	MacQueen (1967)
median linkage	Jain e Dubes (1988)
método de McQuitty	McQuitty (1966)
método de Ward	Ward (1963)
neural gas	Martinetz, Berkovich e Schulten (1993)
partitioning around medoids	Kaufman e Rousseeuw (1990)
spherical k-means	Dhillon e Modha (2001)
unsupervised fuzzy competitive learning	Pal, Bezdek e Hathaway (1996)

Após a execução do módulo de geração de classes, os dados referentes à classificação dos pontos georreferenciados podem ser utilizados para a geração do mapa temático das ZMs. Para isso, deve-se utilizar um software como, por exemplo, o SDUM, que foi empregado neste trabalho. Por fim, se as subáreas forem consideradas satisfatórias, encerra-se o processo de definição de ZMs.

Para avaliar o desempenho dos métodos de agrupamento em relação à qualidade das classes geradas, utilizaram-se a ANOVA (teste de comparação de médias de Tukey), o índice de redução da variância (variance reduction - VR) (LI et al., 2007) e o coeficiente de silhueta médio (average silhouette coefficient - ASC) (ROUSSEEUW, 1987).

7.2.3 Estudo de caso

Para exemplificar o funcionamento dos módulos de seleção de variáveis e de geração de classes para ZMs, utilizaram-se dados coletados entre 2012 e 2015 correspondentes a uma área agrícola comercial de 15,5 ha, localizada no município de Céu Azul, estado do Paraná. O solo dessa área foi classificado como LATOSSOLO VERMELHO Distroférrico típico (EMBRAPA, 2013) e tem sido cultivado em sistema de plantio direto com sequência de soja, trigo, milho e aveia há mais de 10 anos. Sua altitude média é de 460 m.

Para a demarcação dessa área, utilizou-se um dispositivo receptor GPS Trimble Geo Explorer XT 2005, em conjunto com o software Trimble GPS PathFinder Office para a geração de uma grade amostral com 40 pontos (2,67 pontos ha^{-1}) (Figura 2). Como essa área apresenta certo grau de declividade e comporta curvas de nível, decidiu-se pela definição de uma grade amostral irregular, de modo que os pontos amostrais ficassem posicionados na linha central imaginária entre as curvas de nível.

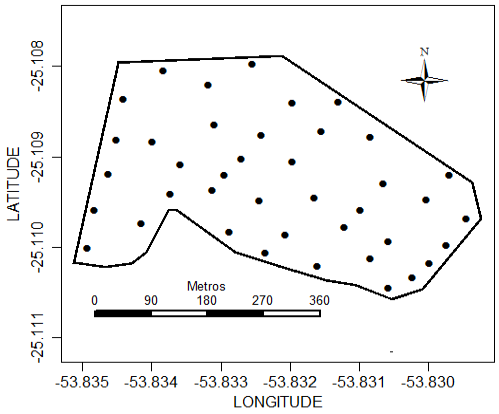


Figura 2 Área agrícola considerada no estudo de caso, com representação dos 40 pontos amostrais.

A densidade de amostragem foi estabelecida para seguir a recomendação de ser maior que 2,5 pontos ha^{-1} (NANNI et al., 2011), para que a dependência espacial entre amostras pudesse ser avaliada. Além disso, a quantidade total de pontos amostrais foi definida conforme a recomendação de Journel e Huijbregts (2004) de usar ao menos 40 pontos para coletas de amostras compostas de solo. Seguir essas recomendações foi importante para assegurar ao menos 30 pares de pontos para o cálculo de cada semivariância (DIGGLE; RIBEIRO JR, 2007), durante a geração do semivariograma experimental de Matheron (necessário para a interpolação por krigagem ordinária).

Seguindo recomendação de Doerge (2000), no estudo de caso empregaram-se somente variáveis consideradas temporalmente estáveis (Tabela 2) para a geração de ZMs. O objetivo dessa recomendação é que as variáveis usadas proporcionem a criação de subáreas válidas por vários anos.

Tabela 2 Variáveis da área experimental consideradas no estudo de caso

Variáveis	Anos			
	2012	2013	2014	2015
Argila	X			
Densidade	X			
Altitude	X			
Areia	X			
Silte	X			
Declividade	X			
RSP 0–0,1 m	X	X	X	
RSP 0,1–0,2 m	X	X	X	
RSP 0,2–0,3 m	X	X	X	
Produtividade de soja	X	X	X	X

A obtenção dos valores da altitude dos pontos amostrais foi realizada empregando-se uma estação total eletrônica Topcon GPT-7505. Posteriormente, foram calculados os valores de declividade a partir dos valores da altitude desses pontos, utilizando-se o software Surfer. Para cada ponto amostral, obteve-se uma amostra de solo composta a partir de oito subamostras coletadas dentro de um círculo imaginário com raio de 3 m e centro no ponto amostral (adaptado de WOLLENHAUPT; WOLKOWSKI; CLAYTON, 1994). Posteriormente,

as amostras compostas foram enviadas para um laboratório especializado em análise de solos. Essas amostras foram coletadas para a análise de características químicas e textura do solo (areia, argila e silte), na profundidade de 0-0,2 m, por meio do uso de um perfurador Stihl BT 45.

A fim de determinar a densidade aparente do solo, utilizaram-se anéis volumétricos para coletar mais uma amostra de solo em cada ponto da grade amostral. Posteriormente, executaram-se os procedimentos para obter os valores da densidade no Laboratório de Armazenamento de Amostras e no Laboratório de Solos da UNIOESTE. Tais procedimentos foram definidos de acordo com as orientações do Manual de Métodos de Análise de Solo da Embrapa (EMBRAPA, 1997).

O valor da resistência mecânica do solo à penetração (RSP) para cada ponto amostral foi definido como a média dos valores correspondentes a quatro medições, que foram realizadas praticamente nos mesmos locais em que se efetuou a coleta das oito subamostras de solo. A RSP foi determinada para as profundidades de 0-0,1 m, 0,1-0,2 m e 0,2-0,3 m, utilizando-se um medidor eletrônico de compactação do solo Falker PenetroLOG PLG1020.

Para validação das ZMs definidas, utilizaram-se dados da produtividade de soja do período de 2012 a 2015. Esses dados foram determinados para os mesmos pontos amostrais onde ocorreram as medições das variáveis listadas na Tabela 2. A coleta ocorreu por meio do uso de uma colhedora CASE IH 2388 com um monitor de colheita CASE AFS PRO 600. Com o intuito de prover estabilidade aos dados da produtividade de soja, que geralmente são influenciados pelas variações do clima e da precipitação pluviométrica, realizou-se a padronização dos valores obtidos para os pontos amostrais, para cada ano considerado. Para isso, aplicou-se uma das versões da técnica baseada na amplitude dos dados (Equação 1) (MIELKE JR; BERRY, 2007). Antes de serem padronizados, corrigiram-se os valores da produtividade de soja para um teor de água de 13%.

$$P_{iN} = \frac{P_i - P_{\min}}{P_{\max} - P_{\min}} \qquad \text{Eq. (1)}$$

em que: P_{iN} é o valor padronizado da produtividade de soja para o ponto amostral i; P_i é o valor original da produtividade de soja para o ponto i; e $P_{\max}$ e $P_{\min}$ correspondem, respectivamente, aos valores amostrais máximo e mínimo da produtividade de soja no conjunto de dados considerado. Por fim, calculou-se a média dos valores padronizados dos anos disponíveis, para gerar a variável chamada de produtividade média.

7.3 Resultados e discussão

7.3.1 Módulo de seleção de variáveis

A tela inicial projetada para esse módulo (Figura 3) apresentará, na parte superior, um campo para seleção da área agrícola de interesse. Após a escolha da área, o arquivo de texto com os dados das variáveis correspondentes a ela será automaticamente lido pelo software, para que os nomes dessas variáveis sejam exibidos na parte inferior da tela.

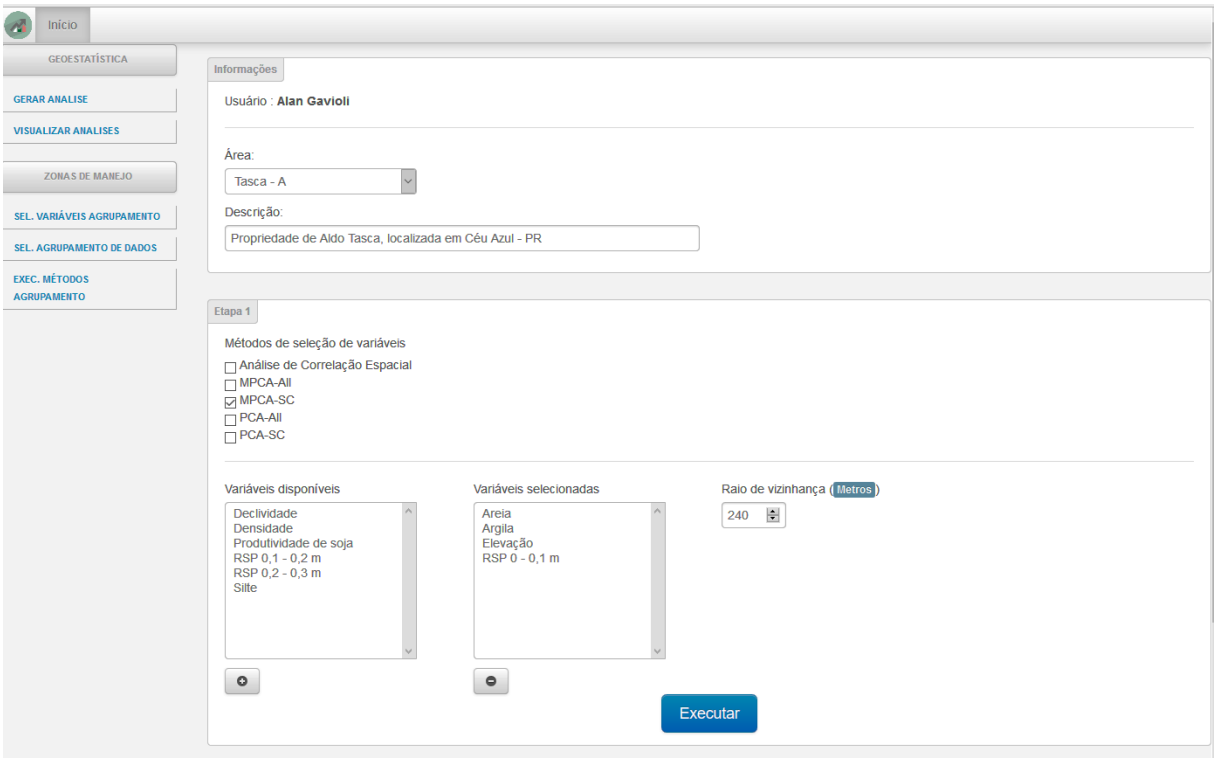


Figura 3 Tela inicial projetada para o módulo de seleção de variáveis, com os dados do estudo de caso considerado; neste exemplo, optou-se pelo método MPCA-SC.

Assim, será possível escolher nessa tela um dos 5 métodos de seleção de variáveis para ser utilizado. De acordo com essa escolha, o módulo automaticamente analisará as variáveis disponíveis e preencherá o campo correspondente às variáveis selecionadas, também na parte inferior da tela. Todavia, o módulo permitirá a livre modificação da lista de variáveis que tenham sido automaticamente selecionadas. Na versão atual desse módulo, pode-se configurar e aplicar os métodos por meio da execução das rotinas implementadas no software R (Apêndice A).

No exemplo mostrado na Figura 3, a opção pelo método MPCA-SC fez com que o módulo selecionasse quatro das dez variáveis disponíveis para serem empregadas na geração de CPs. Além disso, determinou-se de acordo com o procedimento explicado anteriormente que 240 m era um valor satisfatório para o raio de vizinhança. A rotina

implementada no software R possibilita gerar e exibir uma representação gráfica que ilustra quais são os vizinhos de cada ponto do conjunto de dados de entrada.

Após a execução dos métodos MPCA-All, MPCA-SC, PCA-All e PCA-SC, as CPs geradas serão apresentadas na parte inferior da tela de resultados (Figura 4), com os respectivos valores do percentual de representação da variância original dos dados e do percentual acumulado dessa variância.

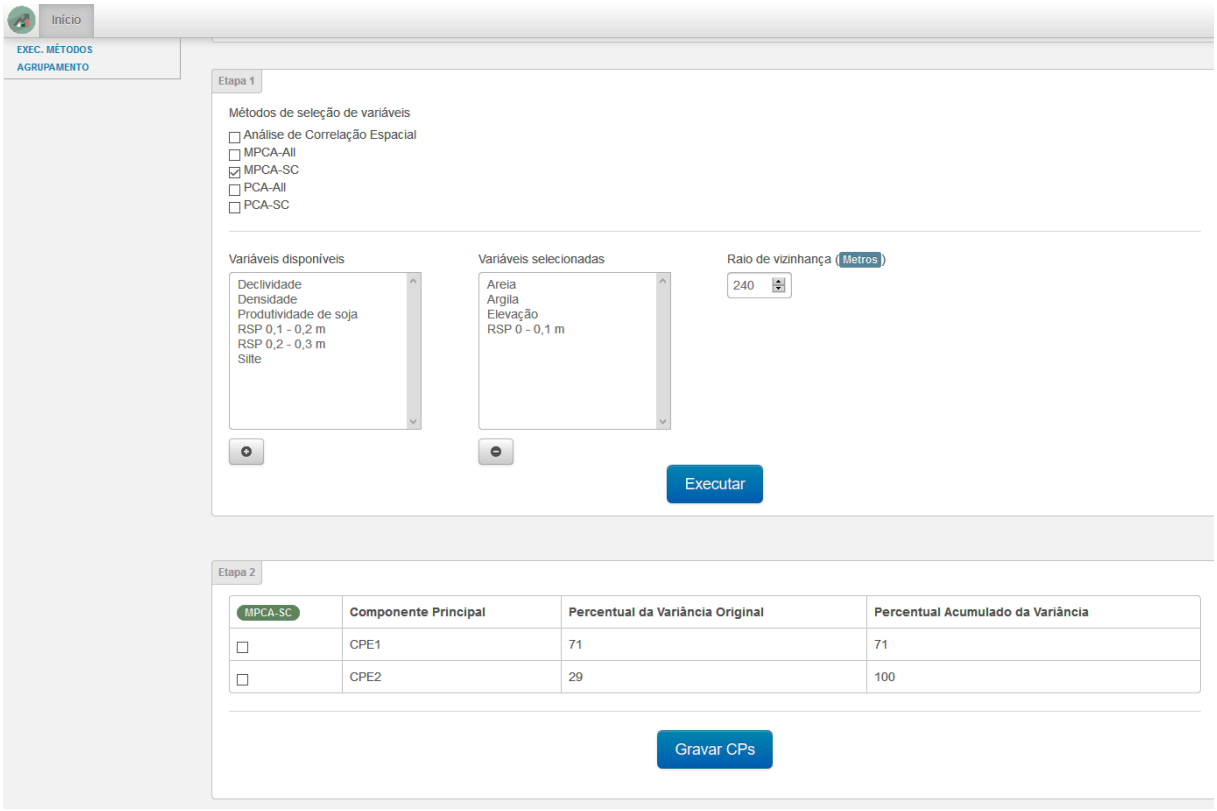


Figura 4 Tela projetada para exibir os resultados da execução dos métodos de seleção de variáveis baseados na criação de componentes principais.

Na versão atual desse módulo, esses resultados são exibidos na R Console, janela do software R na qual são apresentados resultados da execução de rotinas. Além disso, as rotinas implementadas no software R para esse módulo também permitem visualizar na R Console as seguintes informações correspondentes às CPs:

- Os autovalores e autovetores;
- Valores da estatística de autocorrelação espacial de Moran;
- Gráficos para análise das possíveis correlações existentes entre variáveis originais e entre essas variáveis e as CPs;
- Mapas de variabilidade espacial multivariada das CPs.

Em relação ao exemplo mostrado na Figura 4, ao encerrar a execução da rotina de implementação de MPCA-SC do módulo de seleção de variáveis, as coordenadas geográficas

dos pontos amostrais da área e os respectivos escores das componentes CPE1 e CPE2 foram gravados em um arquivo de texto de resposta.

O módulo de seleção de variáveis também permitirá selecionar simultaneamente dois ou mais métodos de geração de CPs, para serem executados com os dados da mesma área. Para isso, será necessário selecionar, na tela inicial desse módulo (Figura 3), os métodos que deverão ser comparados. Neste caso, para facilitar a interpretação dos resultados, os métodos executados serão exibidos em ordem decrescente de desempenho (Figura 5). O critério de comparação empregado será o maior percentual possível de representação da variância original dos dados nas primeiras CPs. Assim, no exemplo mostrado nesta figura, MPCA-SC obteve o melhor desempenho.

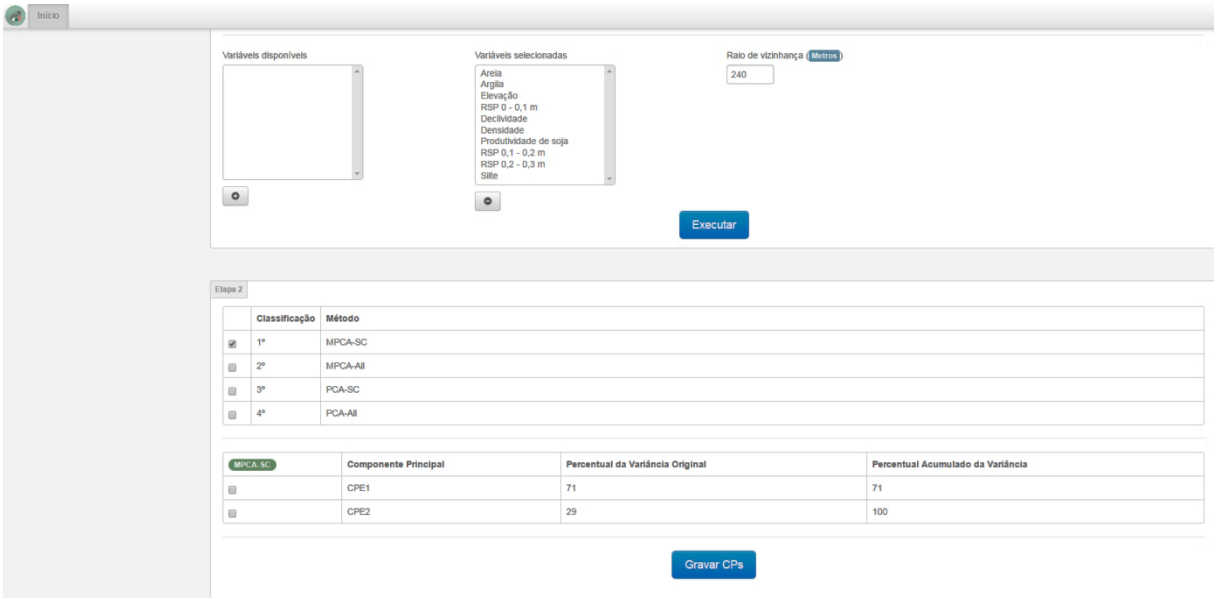


Figura 5 Tela projetada para exibir o resultado da comparação de métodos de seleção baseados em componentes principais, com destaque para o desempenho de MPCA-SC com os dados do estudo de caso.

7.3.2 Módulo de geração de classes

A tela inicial projetada para este módulo (Figura 6) exibirá, na parte superior, um campo para seleção da área agrícola de interesse. Após a escolha da área, os arquivos de texto com os dados das variáveis correspondentes a ela e que tenham sido interpoladas serão automaticamente lidos pelo software, para que os nomes dessas variáveis sejam exibidos na parte inferior da tela. Deste modo, será possível selecionar um dos 17 algoritmos de agrupamento e as variáveis de interesse para que sejam utilizados. Também será necessário informar as quantidades mínima e máxima de ZMs desejadas, além de possíveis parâmetros específicos do método de agrupamento selecionado. Para facilitar o uso desse módulo, os parâmetros específicos dos métodos serão automaticamente preenchidos com valores

geralmente recomendados. Na versão atual desse módulo, pode-se configurar e usar os 17 algoritmos por meio da execução das rotinas implementadas no software R (Apêndice B).

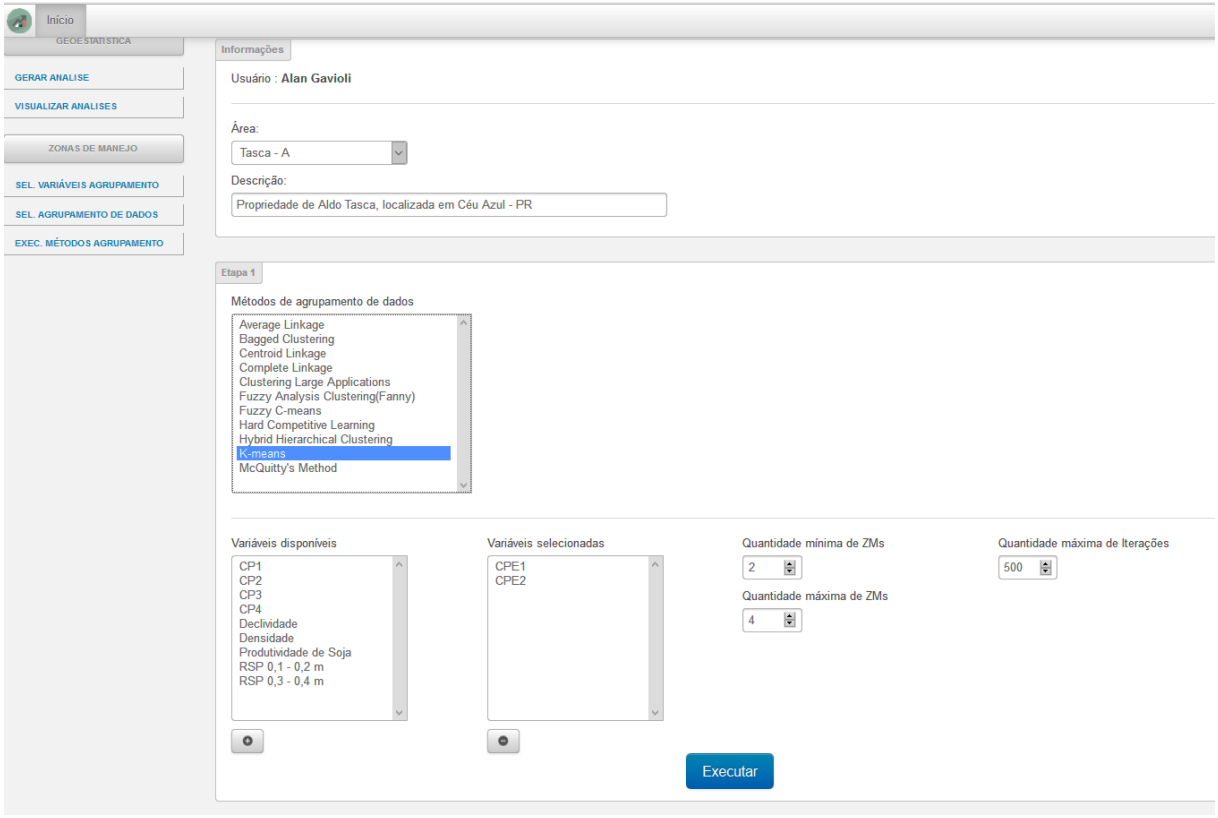


Figura 6 Tela inicial projetada para o módulo de geração de classes, mostrando a escolha do algoritmo K-means e das variáveis CPE1 e CPE2 para gerar duas, três e quatro zonas de manejo.

No exemplo mostrado na Figura 6, selecionou-se o algoritmo de agrupamento K-means para gerar duas, três e quatro classes, a partir dos valores resultantes da interpolação por krigagem ordinária das componentes principais CPE1 e CPE2. Para K-means, foi necessário informar a quantidade máxima de iterações que podem ocorrer durante a sua execução. A geração de diversas quantidades de classes, empregando o mesmo algoritmo de agrupamento com as mesmas variáveis de entrada, é importante quando não se sabe, a priori, em quantas ZMs uma área deve ser dividida para se obter os melhores resultados.

Na tela de resultados (Figura 7), será exibido o desempenho do algoritmo de agrupamento para cada quantidade requisitada de classes. Para facilitar a visualização dos resultados, estes serão exibidos em ordem crescente de quantidade de classes. Na versão atual desse módulo, os dados correspondentes à execução de qualquer método de agrupamento são apresentados na R Console. No exemplo ilustrado na Figura 7, o resultado do teste de Tukey mostrou que é possível dividir a área somente em duas ZMs com potenciais produtivos estatisticamente distintos, ao nível de 5% de significância. Essa divisão promoveu redução satisfatória da variância da produtividade (VR = 33,8%), com subáreas que apresentaram elevada homogeneidade interna (ASC = 0,59).

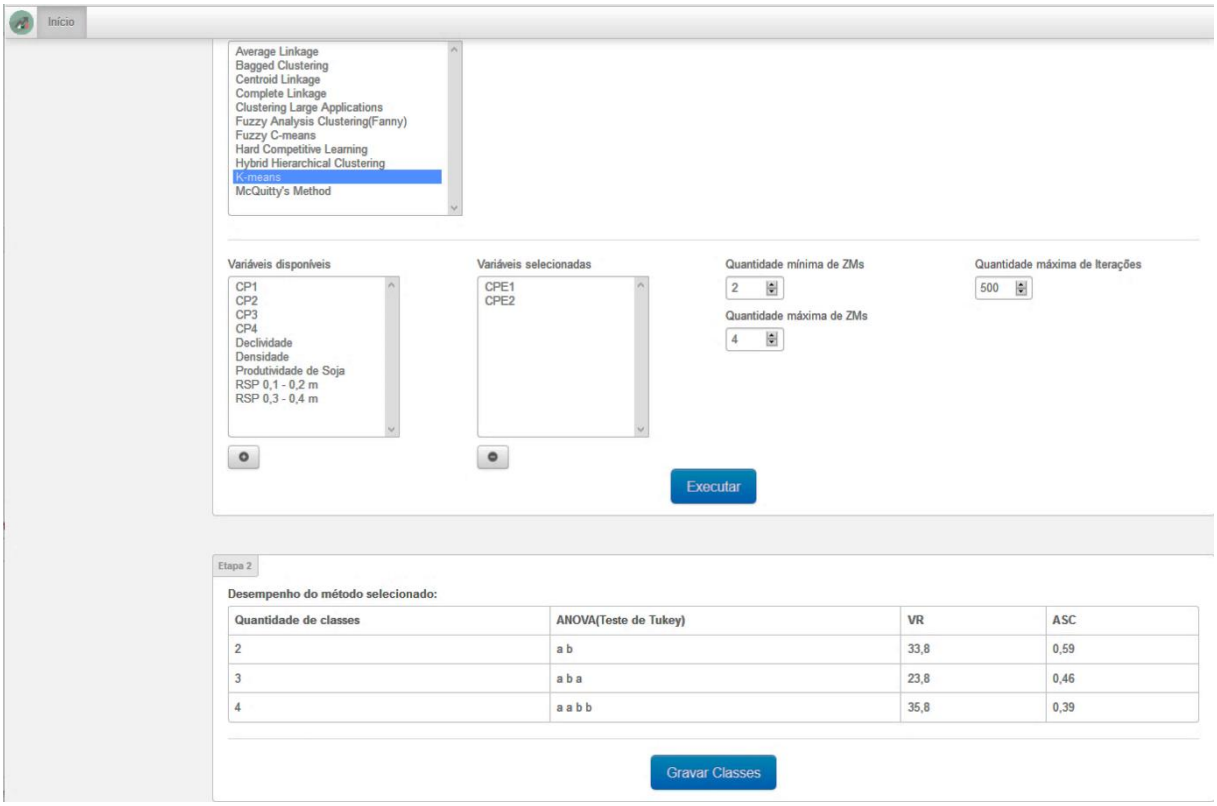
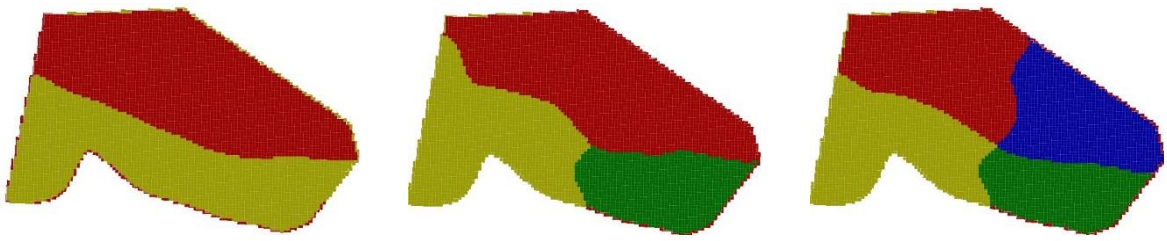


Figura 7 Tela projetada para exibir os resultados da avaliação do método de agrupamento selecionado: valores do teste de Tukey, do índice VR e do coeficiente ASC.

Ao concluir a execução da rotina correspondente a K-means para esse estudo de caso, as coordenadas geográficas dos pontos resultantes da interpolação e os respectivos números identificadores das classes associadas a esses pontos foram gravados em um arquivo de texto de resposta. Com essa rotina, foram criados três arquivos de resposta, correspondentes à geração de duas, três e quatro classes.

Utilizando o software SDUM, leram-se dos três arquivos de texto as coordenadas dos pontos e suas respectivas classes, para definirem-se os mapas temáticos com duas, três e quatro ZMs (Figura 8).



2 ZMs 3 ZMs 4 ZMs

Figura 8 Mapas temáticos com duas, três e quatro zonas de manejo, correspondentes ao estudo de caso executado nos dois módulos computacionais.

Existem outros softwares que possibilitam aplicar os métodos de seleção de variáveis baseados em ACP, como ArcGIS (Esri Software), Matlab (The MathWorks), SAS (SAS Institute), Scilab (Scilab Enterprises) e Statistica (Statsoft). No entanto, quase todos exigem o pagamento de algum valor para serem utilizados. Além disso, são de propósito mais geral que o módulo de seleção de variáveis desenvolvido, isto é, nenhum deles foi projetado especificamente para a definição de ZMs. Por isso, requerem mais tempo de aprendizado para poderem ser aplicados em uma tarefa que pode ser feita com simplicidade no módulo apresentado. Já para a execução dos métodos baseados em MULTISPATI-PCA, tem-se apenas o software R disponível para ser empregado.

Comparando os dois módulos desenvolvidos aos softwares FuzME e MZA, notou-se que estes dois não disponibilizam métodos para seleção das variáveis para a definição de ZMs, deixando para o usuário a tarefa de escolher manualmente essas variáveis. Em relação a algoritmos de agrupamento, FuzME e MZA contêm apenas a implementação do método fuzzy c-means. Já ao serem comparados os dois módulos e o software SDUM, este viabiliza a seleção de variáveis apenas por meio da análise de correlação espacial e disponibiliza dois métodos de agrupamento: fuzzy c-means e k-means.

7.4 Conclusões

Os módulos computacionais desenvolvidos proporcionam flexibilidade na seleção de variáveis para a geração de classes, por meio do uso dos 5 algoritmos de seleção disponibilizados, assim como no agrupamento de dados que efetivamente define as ZMs, por meio da aplicação dos 17 métodos implementados. Esses módulos constituem uma ferramenta computacional mais abrangente que outros softwares de uso gratuito, como FuzME, MZA e SDUM, em relação à diversidade de algoritmos para seleção de variáveis e agrupamento de dados disponibilizados.

7.5 Referências

BAZZI, C. L.; SOUZA, E. G.; URIBE-OPAZO, M. A.; NÓBREGA, L. H. P.; ROCHA, D. M. Management zones definition using soil chemical and physical attributes in a soybean area. **Engenharia Agrícola**, v. 33, n. 5, p. 952-964, 2013.

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. New York: Plenum Press, 1981. 256 p.

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.

CHIPMAN, H.; TIBSHIRANI, R. Hybrid Hierarchical Clustering with Applications to Microarray Data. **Biostatistics,** v. 7, p. 302-317, 2006.

COHEN, S.; COHEN, Y.; ALCHANATIS, V.; LEVI, O. Combining spectral and spatial information from aerial hyperspectral images for delineating homogenous management zones. **Biosystems Engineering,** v. 114, n. 4, p. 435-443, 2013.

CÓRDOBA, M.; BALZARINI, M.; BRUNO, C.; COSTA, J. L. Análisis de componentes principales con datos georreferenciados: Una aplicación en agricultura de precisión. **Revista de la Facultad de Ciencias Agrarias UNCUYO,** v. 44, n. 1, p. 27-39, 2012.

CÓRDOBA, M.; BRUNO, C.; COSTA, J. L.; 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.

DHILLON, I. S.; MODHA, D. S. Concept decompositions for large sparse text data using clustering. **Machine Learning,** v. 42, p. 143-175, 2001.

DIGGLE, P. J.; RIBEIRO JR, P. J. **Model-based Geostatistics**. New York: Springer-Verlag, 2007. 232 p.

DOBERMANN, A.; PING, J. L.; ADAMCHUK, V. I.; SIMBAHAN. G. C.; FERGUSON, R. B. Classification of crop yield variability in irrigated production fields. **Agronomy Journal**, v. 95, n. 1, p. 1105-1120, 2003.

DOERGE, T. A. **Site-specific management guidelines**. Norcross: Potash & Phosphate Institute, 2000. 135 p.

DRAY, S.; SAID, S.; DÉBIAS, F. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. **Journal of Vegetation Science,** v. 19, n. 1, p. 45-56, 2008.

EL-SHARKAWY, M. M.; SHETA, A. S.; EL-WAHED, M. S. A.; ARAFAT, S. M.; EL-BEHIERY, O. M. Precision Agriculture using Remote Sensing and GIS for Peanut Crop Production in Arid Land. **International Journal of Plant & Soil Science,** v. 58, n. 8, p. 1246-1266, 2016.

EMBRAPA. Centro Nacional de Pesquisa de Solos. **Manual de Métodos de Análise de Solo.** 2. ed. Rio de Janeiro: CNPSO, 1997. 212 p.

EMBRAPA. Centro Nacional de Pesquisa de Solos. **Sistema brasileiro de classificação de solo**. Rio de Janeiro: CNPSO, 2013. 412 p.

FERREIRA, D. F. **Análise multivariada**. Lavras: UFLA, 1996. 394 p.

FRIDGEN, J. J.; KITCHEN, N. R.; SUDDUTH, K. A.; DRUMMOND, S. T.; WIEBOLD, W. J.; FRAISSE, C. W. Management zone analyst (MZA): software for subfield management zone delineation. **Agronomy Journal**, v. 96, p. 100-108, 2004.

GAVIOLI, A.; SOUZA, E. G.; BAZZI, C. L.; GUEDES, L. P. C.; SCHENATTO, K. Optimization of management zone delineation by using spatial principal components. **Computers and Electronics in Agriculture,** v. 127, p. 302-310, 2016.

GUASTAFERRO, F.; CASTRIGNANO, A.; DE BENEDETTO, D.; SOLLITTO, D.; TROCCOLI, A.; CAFARELLI, B. A comparison of different algorithms for the delineation of management zones. **Precision Agriculture,** v. 11, p. 600-620, 2010.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of educational psychology,** v. 24, n. 6, p. 417-441, 1933.

JAIN, A. K.; DUBES, R. **Algorithms for clustering data**. Englewood Cliffs: Prentice-Hall, 1988. 320 p.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 6 ed. New Jersey: Pearson, 2007. 773 p.

JOLLIFFE, I. T. **Principal Component Analysis**. 2 ed. New York: Springer, 2002. 487 p.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining geostatistics**. London: The Blackburn Press, 2004. 600 p.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding groups in data**. Hoboken: John Wiley & Sons, 1990. 342 p.

LEISCH, F. Bagged clustering. In: SFB ADAPTIVE INFORMATION SYSTEMS AND MODELLING IN ECONOMICS AND MANAGEMENT SCIENCE, 51, 1999, Vienna. **Anais...** Vienna: Vienna University of Economics and Business, 1999. p. 1-11.

LI, X.; PAN, Y.; GE, Z.; ZHAO, C. Delineation and scale effect of precision agriculture management zones using yield monitor data over four years. **Agricultural Sciences in China**, v. 6, n. 2, p. 180-188, 2007.

LI, Y.; SHI, Z.; WU, H.; LI, F.; LI, H. Definition of management zones for enhancing cultivated land conservation using combined spatial data. **Environmental Management,** v. 52, n. 1, p. 792-806, 2013.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: PROCEEDINGS OF 5TH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY, 1967, Berkeley. **Anais...** Berkeley: University of California Press, 1967. p. 281–297.

MARTINETZ, T. M.; BERKOVICH, S. G.; SCHULTEN, K. J. "Neural-gas" network for vector quantization and its application to time-series prediction. **IEEE Transactions on Neural Networks,** v. 4, n. 4, p. 558-569, 1993.

MCQUITTY, L. L. Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. **Educational and Psychological Measurement,** v. 26, p. 825-831, 1966.

MIELKE JR, P. W.; BERRY, K. J. **Permutation methods: a distance function approach**. New York: Springer, 2007. 446 p.

MINASNY, B.; MCBRATNEY, A. B. **FuzME 3.0**. Australian Centre for Precision Agriculture. The University of Sydney. Sydney. 2002.

NANNI, M. R.; POVH, F. P.; DEMATTÊ, J. A. M.; OLIVEIRA, R. B.; CHICATI, M. L.; CEZAR, E. Optimum size in grid soil sampling for variable rate application in site-specific management. **Scientia Agricola,** v. 68, n. 3, p. 386-392, 2011.

PAL, N. R.; BEZDEK, J. C.; HATHAWAY, R. J. Sequential competitive learning and the fuzzy c-means clustering algorithm. **Neural Networks,** v. 9, n. 5, p. 787-796, 1996.

PERALTA, N. R.; COSTA, J. L.; BALZARINI, M.; FRANCO, M. C.; CÓRDOBA, M.; BULLOCK, D. Delineation of management zones to improve nitrogen management of wheat. **Computers and Electronics in Agriculture,** v. 110, p. 103-113, 2015.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna: R Foundation for Statistical Computing, 2016. 99 p.

REICH, R. M.; CZAPLEWSKI, R. L.; BECHTOLD, W. A. Spatial cross-correlation of undisturbed, natural shortleaf pine stands in northern Georgia. **Environmental and Ecological Statistics**, v. 1, p. 201-217, 1994.

ROUSSEEUW, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics,** v. 20, p. 53-65, 1987.

SCHENATTO, K.; SOUZA, E. G.; BAZZI, C. L.; BIER, V. A.; BETZEK, N. M.; GAVIOLI, A. Data interpolation in the definition of management zones. **Acta Scientiarum,** v. 38, n. 1, p. 31-40, 2016.

TRIPATHI, R.; NAYAK, A. K.; SHAHID, M.; LAL, B.; GAUTAM, P.; RAJA, R.; MOHANTY, S.; KUMAR, A.; PANDA, B. B.; SAHOO, R. N. Delineation of soil management zones for a rice cultivated area in eastern India using fuzzy clustering. **Catena,** v. 133, p. 128-136, 2015.

WARD, J. H. Hierarchical Grouping to Optimize an Objective Function. **Journal of the American Statistical Association,** v. 58, n. 301, p. 236-244, 1963.

WOLLENHAUPT, N. C.; WOLKOWSKI, R. P.; CLAYTON, M. K. Mapping soil test phosphorus and potassium for variable-rate fertilizer application. **Journal of Production Agriculture**, v. 7, n. 4, p. 441-448, 1994.

XU, R.; WUNSCH, D. C. **Clustering**. Piscataway: IEEE Press, 2009. 358 p.

8 CONSIDERAÇÕES FINAIS

8.1 Conclusões

Os experimentos realizados para avaliar de forma comparativa os 5 métodos de seleção de variáveis baseados em análise de correlação espacial, ACP e MULTISPATI-PCA possibilitaram concluir que o novo algoritmo proposto, denominado MPCA-SC, pode melhorar a qualidade de zonas de manejo (ZMs) para áreas agrícolas. Aplicando esse algoritmo, as subáreas tendem a apresentar tamanhos mais recomendados e contornos mais suaves do que as geradas por meio do uso dos outros quatro métodos (MPCA-All, PCA-SC, PCA-All e Spatial-Matrix). Na prática, isso pode torná-las melhores para a realização de operações de campo.

O método MPCA-SC também pode conduzir à formação de ZMs com as maiores diferenças entre as respectivas produtividades médias e com os menores valores para o desvio padrão. Além disso, ele tende a promover a melhor redução de dimensionalidade dos dados originais, sem perda significativa de informação. Todavia, apesar da tendência de superioridade apresentada por MPCA-SC, pôde-se notar que os outros quatro algoritmos de seleção de variáveis avaliados também podem proporcionar resultados satisfatórios, similares aos obtidos com a aplicação do novo método.

Os experimentos de avaliação de algoritmos de agrupamento de dados possibilitaram concluir que, dos 20 métodos inicialmente considerados, 17 geraram resultados que os tornaram adequados para a definição de ZMs em ao menos duas das três áreas agrícolas consideradas: average linkage, bagged clustering, centroid linkage, clustering large applications, complete linkage, fuzzy analysis clustering (fanny), fuzzy c-means, hard competitive learning, hybrid hierarchical clustering, k-means, median linkage, método de McQuitty (mcquitty), método de Ward, neural gas, partitioning around medoids, spherical k-means e unsupervised fuzzy competitive learning. Por isso, também podem vir a ser úteis para a geração de ZMs para outras áreas agrícolas.

Os algoritmos de agrupamento mcquitty e fanny apresentaram desempenho superior aos demais na redução da variância original da produtividade média das três áreas, após a criação das ZMs. Além disso, esses dois algoritmos produziram classes com alta homogeneidade interna, que resultaram em subáreas praticamente sem fragmentações, ou seja, apropriadas para a execução de operações de campo. Esses resultados permitiram concluir que mcquitty e fanny foram os melhores métodos para a definição de ZMs para as áreas consideradas.

Por fim, concluiu-se que com os dois módulos computacionais implementados para a seleção de variáveis e o agrupamento de dados obteve-se eficiência e flexibilidade na

execução dessas duas atividades, por meio do uso dos diversos métodos disponibilizados. Graças a essa variedade de métodos, obteve-se uma ferramenta computacional mais abrangente que outros softwares de uso gratuito, como FuzME, MZA e SDUM, no que tange à seleção de variáveis e ao agrupamento de dados.

8.2 Trabalhos futuros

Inicialmente, disponibilizar-se-ão os módulos computacionais de seleção de variáveis e de agrupamento de dados em uma aplicação web. Com isso, os métodos implementados nesses módulos poderão ser utilizados diretamente na Internet, por meio do uso de um software de navegação (web browser).

Outro trabalho a ser realizado é uma investigação sobre a possibilidade de aplicação de novos algoritmos para a seleção de variáveis e o agrupamento de dados, no âmbito da definição de ZMs. Se essa investigação conduzir a resultados satisfatórios, os novos algoritmos deverão ser implementados e disponibilizados para uso na aplicação web.

APÊNDICES

APÊNDICE A – IMPLEMENTAÇÃO DE MÉTODOS DE SELEÇÃO DE VARIÁVEIS

```
# Pacotes necessários.
library(ade4)
library(spdep)
library(geoR)
library(gstat)
library(data.table)

# Carregamento das coordenadas (UTM) e valores das variáveis, para os pontos.
dados <- read.table("dados_variaveis_area_ceuazul.txt", header = TRUE)

# Carregamento das coordenadas (UTM) do contorno da área.
contorno <- read.table("contorno_area_ceuazul.txt", header = TRUE)

num_atrib <- ncol(dados) - 2     # Quantidade de variáveis lidas do arquivo de entrada.

# Matriz de ponderação espacial para o método MULTISPATI-PCA.
coord <- coordinates(dados[,1:2])     # Colunas 1 e 2 contêm as coordenadas dos pontos

# O terceiro parâmetro, raio de vizinhança, deve ser definido empiricamente; pode-se testar
# inicialmente o valor da metade da maior distância entre 2 pontos da área.
gri <- dnearneigh(coord, 0, 240)

# Plotagem da grade amostral, ilustrando a rede de vizinhança dos pontos.
lw2 <- nb2listw(gri, style = "W")
plot(gri,coord,col = "red", pch = 20, cex = 1)

# Cálculo da estatística de autocorrelação espacial de Moran (índice de Moran).
# Colunas 3 a (num_atrib+2) da entrada contêm os valores das variáveis.
c.moran <- lapply(dados[,3:(num_atrib+2)], moran.mc, lw2, 999)
c.moran


################### EXECUÇÃO DO MÉTODO ACP TRADICIONAL #################

pca2 <- dudi.pca(dados[,3:(num_atrib+2)], center=T, scannf = FALSE, nf = num_atrib)

# Biplot, autovalores das CPs e gráfico representando correlações (CP1 e CP2).
par(mfrow=c(1,2))
scatter(pca2, xax = 1, yax = 2,clab.r=0.4, clab.c=0.7)
s.corcircle(pca2$co, clabel = 0.7)

# Biplot, autovalores das CPs e gráfico representando correlações (CP1 e CP3).
par(mfrow=c(1,2))
scatter(pca2, xax = 1, yax = 3,clab.r=0.4, clab.c=0.9)
s.corcircle(pca2$co,xax = 1, yax = 3, clabel = 0.9)

# Autovetores e autovalores associados a cada componente principal (CP).
# Autovetores
pca2$c1      # Exibe os coeficientes das CPs, correspondentes às variáveis originais.

# Autovalores
pca2$eig     # Úteis para avaliação sobre quantas CPs efetivamente utilizar.

# Proporção da variância total dos dados originais explicada por cada CP.
```

```
pca2$eig/sum(pca2$eig) * 100     # Úteis para decidir quantas CPs utilizar.

# Teste de significância do índice de Moran das CPs.
mc.pca <- lapply(pca2$l1, moran.mc, lw2, 999)
mc.pca

# Mapas de variabilidade espacial multivariada de CP1 e CP2.
par(mfrow = c(1,3))
plot(dados[,1:2], pch = 30, cex = 0.1, xlab="x", ylab="y", main="CP1")
s.value(coord, pca2$li[,1], add.plot = TRUE)
text(5576800, 5800100, paste("MC = ", round(mc.pca[[1]]$statistic, 3), " (",
mc.pca[[1]]$p.value, ")", sep = ""), cex = 1)
polygon(contorno)

plot(dados[,1:2],pch = 30, cex = 0.1,xlab="x",ylab="y",main="CP2")
s.value(coord, pca2$li[,2], add.plot = TRUE)
text(5576800, 5800100, paste("MC = ", round(mc.pca[[2]]$statistic, 3), " (",
mc.pca[[1]]$p.value, ")", sep = ""), cex = 1)
polygon(contorno)

plot(dados[,1:2],pch = 30, cex = 0.1,xlab="x",ylab="y",main="CP3")
s.value(coord, pca2$li[,3], add.plot = TRUE)
text(5576800, 5800100, paste("MC = ", round(mc.pca[[3]]$statistic, 3), " (",
mc.pca[[1]]$p.value, ")", sep = ""), cex = 1)
polygon(contorno)


#################### EXECUÇÃO DO MÉTODO MULTISPATI-PCA ###################

ms2 <- multispati(pca2, lw2, scannf = F, nfposi = num_atrib)
ms2

# Exibe dados referentes às CPs espaciais de MULTISPATI-PCA:
# eig: autovalores das CPs espaciais;
# var: variância representada por cada CP espacial;
# moran: índice de Moran de cada CP espacial.
sum.ms <- summary(ms2)

# Autovetores
ms2$c1     # Exibe os coeficientes das CPEs, correspondentes às variáveis originais.

# Biplot para CPE1 e CPE2 de MULTISPATI-PCA, incluindo autovalores das CPs.
s.arrow(ms2$c1,xax = 1, yax = 2, clabel = 1)
add.scatter.eig(ms2$eig, xax = 1, yax = 2, posi = "topright", ratio = 0.2)

# Teste de significância do índice de Moran das CPs espaciais.
mc.mpca <- lapply(ms2$li, moran.mc, lw2, 999)
mc.mpca

# Mapas ilustrando os escores das CPs espaciais, para os pontos.
par(mfrow = c(1, 2))
plot(dados[,1:2], pch = 30, cex = 0.1, xlab="x", ylab="y", main="CPE1")
s.value(coord, ms2$li[,1], add.plot = TRUE)
text(5576800, 5800100, paste("MC = ", round(mc.mpca[[1]]$statistic, 3), " (",
mc.mpca[[1]]$p.value, ")", sep = ""), cex = 1)
polygon(contorno)
```

```
plot(dados[,1:2], pch = 30, cex = 0.1, xlab="x", ylab="y", main="CPE2")
s.value(coord, ms2$li[,2], add.plot = TRUE)
text(5576800, 5800100, paste("MC = ", round(mc.mpca[[2]]$statistic, 3), " (",
mc.mpca[[2]]$p.value, ")", sep = ""), cex = 1)
polygon(contorno)


## GERAÇÃO DOS ARQUIVOS COM ESCORES DE ACP E MULTISPATI-PCA ##

CP_final <- pca2$li[,1:(pca2$nf)]     # CP_final recebe os escores das CPs para os pontos.

# result_pca recebe as coordenadas e escores das CPs, para os pontos.
result_pca <- data.table(cbind(dados[,1:2], CP_final))

CS_final <- ms2$li[,1:(ms2$nfposi)]      # CS_final recebe os escores das CPEs.

# result_mpca recebe as coordenadas e escores das CPEs, para os pontos.
result_mpca <- data.table(cbind(dados[,1:2], CS_final))

# Gravação das coordenadas e escores das CPs de ACP e MULTISPATI-PCA
# para os pontos, em arquivos de texto.
write.table(result_pca, "resultado_acp.txt")
write.table(result_mpca, "resultado_multispati-pca.txt")
```

APÊNDICE B – IMPLEMENTAÇÃO DE MÉTODOS DE AGRUPAMENTO DE DADOS

```
# Pacotes necessários.
library(data.table)
library(cluster)     # Necessário para Clara, Diana, Fanny e PAM, e coeficiente de silhueta.
library(e1071)     # Necessário para BCL, FCM, FCS e UFCL.
library(cclust)     # Necessário para Neural Gas e HCL.
library(hybridHclust)     # Necessário para hybrid hierarchical clustering.
library(skmeans)     # Necessário para Spherical k-Means Clustering.

# Pacote fastcluster é necessário para métodos hierárquicos aglomerativos;
# substitui hclust do pacote base "stats".
library(fastcluster)

library(vegan)     # Necessário para métodos hierárquicos.

# Carregamento dos dados interpolados das variáveis de interesse
# (CPs, CPEs ou variáveis), para serem usados pelos métodos de agrupamento.
frame_atrib <- read.table("dados_interpolados_area_ceuazul.txt", header = TRUE)

# x receberá os valores interpolados das variáveis,
# que serão efetivamente usados para o agrupamento.
x <- frame_atrib[,3:ncol(frame_atrib)]

num_tuplas <- nrow(frame_atrib)     # Qtde de pontos do conjunto de dados.
num_classes <- 2     # Qtde de classes a serem geradas.

# Construção da matriz de dados a partir do data frame "x".
matriz_x <- t(rbind(matrix(x[,1], ncol=num_tuplas), matrix(x[,2], ncol=num_tuplas)))

# Construção da matriz de distâncias entre pontos, usando distância euclidiana.
distmat <- dist(matriz_x)


################################################################################
#########  MÉTODOS PARA AGRUPAMENTO POR PARTICIONAMENTO  ##########
################################################################################

## AGRUPAMENTO DOS PONTOS USANDO MÉTODO BAGGED CLUSTERING ##

result_bclust <- bclust(x, centers=num_classes, iter.base=500, minsize=0,
dist.method="euclidian", hclust.method="ward.D2", base.method="kmeans",
base.centers=20, final.kmeans=FALSE)
vet_bclust <- t(result_bclust$cluster)
gri_medida <- data.table(result_bclust$cluster)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
final_bclust <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_bclust, "resultado_bclust.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_bclust, distmat)
```

```
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm


## AGRUPAMENTO USANDO MÉTODO CLUSTERING LARGE APPLICATIONS (CLARA)

clarax <- clara(x, num_classes, samples=(num_tuplas / 10))
vet_clara <- t(clarax$clustering)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- vet_clara[1,]
final_clara <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_clara, "resultado_clara.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_clara, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm


## AGRUPAMENTO USANDO MÉTODO FUZZY ANALYSIS CLUSTERING (FANNY) ##

fannyx <- fanny(x,num_classes,memb.exp = 1.3,metric = "euclidean",maxit = 500, tol = 1e-4)
vet_fanny <- t(fannyx$clustering)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- vet_fanny[1,]
final_fanny <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_fanny, "resultado_fanny.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_fanny, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm


## AGRUPAMENTO DOS PONTOS USANDO MÉTODO FUZZY C-MEANS (FCM) ##

result_cmeans <- cmeans(x, num_classes, iter.max=500, verbose=TRUE, dist="euclidean",
method="cmeans", m=1.3)
vet_cmeans <- t(result_cmeans$cluster)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- vet_cmeans[1,]
final_cmeans <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_cmeans, "resultado_fuzzy_cmeans.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_cmeans, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm
```

```
## AGRUPAMENTO DOS PONTOS USANDO MÉTODO FUZZY C-SHELL (FCS) ##

result_cshell <- cshell(x, num_classes, iter.max=500, verbose=FALSE, dist="euclidean",
method="cshell", m=1.3)
vet_cshell <- t(result_cshell$cluster)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- vet_cshell[1,]
final_cshell <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_cshell, "resultado_fuzzy_cshell.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_cshell, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm


## AGRUPAMENTO USANDO MÉTODO HARD COMPETITIVE LEARNING (HARDCL) ##

matriz_x <- t( rbind( matrix(x[,1], ncol=num_tuplas), matrix(x[,2], ncol=num_tuplas) ) )
res_hcl <- cclust(matriz_x, num_classes, iter.max=500, dist="euclidean", method="hardcl")
vet_hcl <- t(res_hcl$cluster)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- vet_hcl[1,]
final_hcl <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_hcl, "resultado_hardcl.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_hcl, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm


## AGRUPAMENTO DOS PONTOS USANDO MÉTODO "K-MEANS TRADICIONAL" ##

res_kmeans <- kmeans(x, num_classes)
vet_kmeans <- t(res_kmeans$cluster)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- vet_kmeans[1,]
final_kmeans <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_kmeans, "resultado_kmeans.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_kmeans, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm
```

```
## AGRUPAMENTO USANDO NEURAL GAS (SOFT COMPETITIVE LEARNING) ##

matriz_x <- t(rbind(matrix(x[,1], ncol=num_tuplas), matrix(x[,2], ncol=num_tuplas)))
res_ng <- cclust(matriz_x, num_classes, iter.max=500, dist="euclidean",
method="neuralgas")
vet_ng <- t(res_ng$cluster)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- vet_ng[1,]
final_ng <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_ng, "resultado_neuralgas.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_ng, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm


## AGRUPAMENTO USANDO MÉTODO PARTITIONING AROUND MEDOIDS (PAM) ##

pamx <- pam(x, num_classes)
vet_pam <- t(pamx$clustering)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- vet_pam[1,]
final_pam <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_pam, "resultado_pam.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_pam, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm


## AGRUPAMENTO USANDO SPHERICAL K-MEANS CLUSTERING (SKMEANS) ##

matriz_x <- t(rbind(matrix(x[,1], ncol=num_tuplas), matrix(x[,2], ncol=num_tuplas)))

# Particionamento soft/fuzzy em "num_classes" grupos.
sparty <- skmeans(matriz_x, num_classes, method = "pclust", m = 1.3, control = list(verbose
= TRUE, maxiter = 500))
vet_sk <- t(sparty$cluster)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- vet_sk[1,]
final_sk <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_sk, "resultado_spherical_kmeans.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_sk, distmat)
```

```
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm


## AGRUPAMENTO USANDO MÉTODO "UNSUPERVISED FUZZY COMPETITIVE
LEARNING (UFCL)" ##

result_ufcl <- cmeans(x, num_classes, iter.max=500, verbose=TRUE, dist="euclidean",
method="ufcl", m=1.3)
vet_ufcl <- t(result_ufcl$cluster)

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- vet_ufcl[1,]
final_ufcl <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_ufcl, "resultado_ufcl.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(vet_ufcl, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm


################################################################################
##############  MÉTODOS PARA AGRUPAMENTO HIERÁRQUICO  ##############
####################         PACOTE FASTCLUSTER         ####################
################################################################################

## AGRUPAMENTO DOS PONTOS USANDO MÉTODO AVERAGE ##

disteuc <- dist(x)
aa_avg <- hclust(disteuc, "average")     # Average com distância euclidiana.
hc <- t(cutree(aa_avg, num_classes))     # Divide os pontos na qtde de classes informada.

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- hc[1,]
final_avg <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_avg, "resultado_average.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(hc, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm

## AGRUPAMENTO DOS PONTOS USANDO MÉTODO CENTROID ##
disteuc <- dist(x)
aa_cen <- hclust(disteuc, "centroid")     # Centroid com distância euclidiana.
hc <- t(cutree(aa_cen, num_classes))     # Divide os pontos na qtde de classes informada.

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- hc[1,]
final_cen <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
```

```
write.table(final_cen, "resultado_centroid.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(hc, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm

## AGRUPAMENTO DOS PONTOS USANDO MÉTODO COMPLETE ##
disteuc <- dist(x)
aa_comp <- hclust(disteuc, "complete")     # Método Complete com distância euclidiana.
hc <- t(cutree(aa_comp, num_classes))     # Divide os pontos na qtde de classes informada.

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- hc[1,]
final_comp <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_comp, "resultado_complete.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(hc, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm


## MÉTODO HIERÁRQUICO HYBRID HIERARCHICAL CLUSTERING (HYBRIDHCLUST) #
hyb1 <- hybridHclust(x)     # Executa o algoritmo com a distância euclidiana.
res_hyb <- t(cutree(as.hclust(hyb1), num_classes))     # Divide os pontos em classes.

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- res_hyb[1,]
final_hyb <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_hyb, "resultado_hybridhclust.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(res_hyb, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm

## AGRUPAMENTO DOS PONTOS USANDO MÉTODO DE MCQUITTY ##
disteuc <- dist(x)
aa_mcq <- hclust(disteuc, "mcquitty")     # McQuitty com distância euclidiana.
hc <- t(cutree(aa_mcq, num_classes))     # Divide os pontos na qtde de classes informada.

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- hc[1,]
final_mcq <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_mcq, "resultado_mcquitty.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(hc, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm
```

```
## AGRUPAMENTO DOS PONTOS USANDO MÉTODO MEDIAN ##
disteuc <- dist(x)
aa_med <- hclust(disteuc, "median")     # Median com distância euclidiana.
hc <- t(cutree(aa_med, num_classes))     # Divide os pontos na qtde de classes informada.

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- hc[1,]
final_med <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_med, "resultado_median.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(hc, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm

## AGRUPAMENTO DOS PONTOS USANDO MÉTODO SINGLE ##
disteuc <- dist(x)
aa_sin <- hclust(disteuc, "single")     # Single com distância euclidiana.
hc <- t(cutree(aa_sin, num_classes))     # Divide os pontos na qtde de classes informada.

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- hc[1,]
final_sin <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_sin, "resultado_single.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(hc, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm

## AGRUPAMENTO DOS PONTOS USANDO MÉTODO DE WARD ##
disteuc <- dist(x)
aa_ward <- hclust(disteuc, "ward.D2")     # Ward com distância euclidiana.
hc <- t(cutree(aa_ward, num_classes))     # Divide os pontos na qtde de classes informada.

# Concatena coordenadas em UTM de cada ponto com a respectiva classe.
gri_medida <- hc[1,]
final_ward <- data.table(cbind(frame_atrib[,1:2], gri_medida))

# Gravação das coordenadas dos pontos e das respectivas classes, em arquivo de texto.
write.table(final_ward, "resultado_ward.txt")

# Obtenção do coef. de silhueta de grupo e do coef. de silhueta médio (ASC).
si2 <- silhouette(hc, distmat)
csm <- sum(si2[,3]) / num_tuplas     # csm recebe o Coef. de Silhueta Médio.
csm
```