UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA

DEFINIÇÃO DE ZONAS DE MANEJO UTILIZANDO ALGORITMO DE AGRUPAMENTO FUZZY C-MEANS COM VARIADAS MÉTRICAS DE DISTÂNCIAS

FABIANE SORBAR FONTANA

CASCAVEL – PARANÁ

2017

FABIANE SORBAR FONTANA

DEFINIÇÃO DE ZONAS DE MANEJO UTILIZANDO ALGORITMO DE AGRUPAMENTO FUZZY C-MEANS COM VARIADAS MÉTRICAS DE DISTÂNCIAS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Agrícola em cumprimento aos requisitos para obtenção do título de **Mestre 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

DEZEMBRO DE 2017

Dados Internacionais de Catalogação-na-Publicação (CIP)

F756d

Fontana, Fabiane Sorbar

Definição de zonas de manejo utilizando algoritmo de agrupamento fuzzy c-means com variadas métricas de distâncias. / Fabiane Sorbar Fontana. -- Cascavel, 2017.

69 f.

Orientador: Prof. Dr. Eduardo Godoy de Souza

Coorientador: Prof. Dr. Claudio Leones Bazzi

Dissertação (Mestrado) – Universidade Estadual do Oeste do Paraná, Campus de Cascavel, 2017

Programa de Pós-Graduação em Engenharia Agrícola

1. Agricultura de precisão. I. Souza, Eduardo Godoy de. II. Bazzi, Claudio Leones. III. Universidade Estadual do Oeste do Paraná. IV. Título.

CDD 20.ed. 630.285

CIP-NBR 12899

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

Revisores:

Português: Dhandara Capitani em 12 de dezembro de 2017.

Ingles: Dhandara Capitani em 12 de dezembro de 2017.

Normas: Dhandara Capitani em 12 de dezembro de 2017.

FABIANE SORBAR FONTANA

Definição de Zonas de Manejo Utilizando Algoritmo Fuzzy C-means com Diferentes Medidas de Similaridade

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Agrícola em cumprimento parcial aos requisitos para obtenção do título de Mestra 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)

Kelyn Schenatto

Universidade Tecnológica Federal do Paraná (UTFPR)

Marcio Furlan Maggi

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

Jerry Adriani Johann

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

Cascavel, 19 de julho de 2017

BIOGRAFIA RESUMIDA

Fabiane Sorbar Fontana nasceu em 29 de outubro de 1982, no município de Cascavel, Estado do Paraná. Em 2003, iniciou o curso de Bacharelado em Ciência da Computação na Faculdade de Ciências Aplicadas de Cascavel (FACIAP), concluindo-o em 2006. Em 2007, ingressou no curso de Especialização em Tecnologia Java na Faculdade de Ciências Aplicadas de Cascavel (FACIAP), sob orientação do professor Ms. Alessandro Kraemer. Em 2008, concluiu a especialização com o artigo intitulado “Comparação entre as Plataformas Java JME e Java Cards para Dispositivos Móveis considerando a Convergência Tecnológica”. Em 2013, ingressou no curso de Especialização em Computação Aplicada à Agroindústria na Universidade Tecnológica Federal do Paraná (UTFPR) – Medianeira, sob orientação do professor Dr. Jean Metz. Em 2014, concluiu a especialização com o artigo intitulado “Utilização de Algoritmos Inteligentes para a Descoberta de Conhecimentos: um Estudo de Caso da Agricultura de Precisão”. Em fevereiro de 2015, ingressou no curso de Mestrado em Engenharia Agrícola – área de concentração Sistemas Biológicos e Agroindustriais – na Universidade Estadual do Oeste do Paraná (UNIOESTE), sob orientação do professor Dr. Eduardo Godoy de Souza. Suas experiências profissionais na área de docência no ensino superior iniciaram em 2014, como professora nos cursos de Bacharelado em Administração e Engenharia de Produção na Faculdade de Ciências Aplicadas de Cascavel (FACIAP/UNIPAN), onde atuou até 2015. Desde 2015 atua como docente nos cursos de Ciência da Computação, Engenharia Agrícola e Engenharia Civil na Universidade Estadual do Oeste do Paraná (UNIOESTE) e nos cursos de Engenharia Mecânica, Engenharia Elétrica e Engenharia de Controle e Automação no Centro Universitário Assis Gurgacz – FAG.

DEFINIÇÃO DE ZONAS DE MANEJO UTILIZANDO ALGORITMO DE AGRUPAMENTO FUZZY C-MEANS COM VARIADAS MÉTRICAS DE DISTÂNCIAS

RESUMO

A Agricultura de Precisão (AP) utiliza tecnologias objetivando o aumento da produtividade e redução do impacto ambiental por meio de aplicação localizada de insumos agrícolas. Para viabilizar economicamente a AP, é essencial aprimorar as metodologias atuais, bem como propor novas, como, por exemplo, o delineamento de zonas de manejo (ZMs) a partir de dados de produtividade, atributos topográficos e do solo, entre outros, utilizados a fim de determinar subáreas heterogêneas entre si em uma mesma área. Neste contexto, este trabalho teve como principal objetivo avaliar três métricas de distâncias (Diagonal, Euclidiana e Mahalanobis) junto aos Softwares FUZME e SDUM (Software para a definição de unidades de manejo), que utilizam o algoritmo fuzzy c-means, e, em um segundo momento, avaliar também as culturas de soja e milho, assim como a associação entre elas. No primeiro artigo, utilizando dados correspondentes a quatro áreas distintas, avaliaram-se as três métricas com dados originais e normalizados associados à produtividade de soja. Para a área A, as distâncias Diagonal e Mahalanobis dispensaram a necessidade de normalização das variáveis, apresentando áreas idênticas para as duas versões. Após a normalização dos dados, a distância Euclidiana apresentou um melhor delineamento em suas ZMs para a área A. Para as áreas B, C e D não foi possível obter conclusões quanto ao melhor desempenho, visto que o fato de ser utilizado apenas uma variável para o processo de definição de ZMs influenciou diretamente nos resultados obtidos. No segundo artigo, dados correspondentes a três áreas distintas foram utilizados para analisar o uso de produtividades de soja e milho, assim como a associação entre elas, na seleção de variáveis para definição de ZMs. A partir das variáveis disponíveis para cada uma das áreas foi realizada a seleção destas através do método da correlação espacial, levando em consideração, para cada uma das áreas, as três produtividades-alvo (soja, milho e soja+milho). O tipo de produtividade utilizada repercutiu de duas formas diferentes: primeiro no processo de seleção de variáveis, onde a sua alternância resultou em seleções diferenciadas para uma mesma área; e em um segundo momento, na avaliação das ZMs definidas, onde mesmo quando as mesmas variáveis foram selecionadas na definição das ZMs, os desempenhos das ZMs foram diferentes. Após os métodos de validação aplicados, verificou-se que a melhor produtividade-alvo foi soja+milho, reforçando a ideia de ser útil a utilização destas duas culturas, em conjunto, na definição das ZMs de uma área com alternância de produção de soja e milho.

Palavras-chave: Agricultura de Precisão, Unidades de Manejo, Métodos de Agrupamento de Dados, Clusterização.

MANAGEMENT ZONES DEFINITION USING THE CLUSTERING ALGORITHM FUZZY C-MEANS WITH ASSOCIATED VARIED DISTANCE METRICS

ABSTRACT

Precision Agriculture (AP) uses technologies aimed at increasing productivity and reducing environmental impact through localized application of agricultural inputs. In order to make AP economically feasible, it is essential to improve current methodologies, as well as to propose new ones, such as the design of management areas (MZs) from productivity data, topographic, and soil attributes, among others, to determine which are heterogeneous subareas among themselves in the same area. In this context, the main objective of this research was to evaluate three distance metrics (Diagonal, Euclidian, and Mahalanobis) through FUZME and SDUM software (for the definition of management units) using the fuzzy c-means algorithm, and, at a further moment, to evaluate the cultures of soybeans and corn, as well as the association between them. On the first scientific paper, using data corresponding to four distinct areas, the three metrics with original and normalized data associated with soybean yield were evaluated. For area A, the Diagonal and Mahalanobis distances exempted the need for normalization of the variables, presenting areas that were identical for both versions. After the normalization of the data, the Euclidian distance presented a better delineation in its MZs for area A. For areas B, C, and D it was not possible to reach conclusions regarding the best performance, since only one variable was used for the process of MZs, and that has directly influenced the results. On the second scientific paper, data corresponding to three distinct areas were applied to analyze the use of soybean and corn yields, as well as the association between them, in the selection of variables to define MZs. Based on the variables available for each of the areas, the selection was carried out using the spatial correlation method, considering, for each one of the areas, the three target yields (soybean, corn, and soybean+corn). The type of productivity used demonstrated two different outcomes: first in the variable selection process, where its alternation resulted in different selections for the same area, and second, in the evaluation of the defined MZs, where even when the same variables were selected in the definition of the MZs, the performances of the MZs were different. After the validation methods applied, it was verified that the best target yield was soybean+corn, reasserting the idea of being better to use these two cultures, together, when defining the MZs of an area with rotating crops of soybean and corn.

Keywords: Precision Agriculture, Management Units, Data Grouping Methods, Clustering.

SUMÁRIO

LISTA DE FIGURAS

REVISÃO BIBLIOGRÁFICA

ARTIGO 1

ARTIGO 2

LISTA DE TABELAS

ARTIGO 1

ARTIGO 2

1 INTRODUÇÃO

A utilização da tecnologia na agricultura é cada vez mais pertinente devido à necessidade crescente do aumento de produção e lucratividade, a diminuição do uso de defensivos e do impacto ambiental nos mais variados ramos rurais, visando sempre beneficiar estes aspectos (MOLIN, 2015).

A Agricultura de Precisão (AP) tem como principal objetivo o fornecimento da necessidade local de insumos agrícolas. Ela apresenta o potencial de aumentar a eficiência do uso de insumos como fertilizantes, água, sementes e herbicidas sem prejudicar o meio ambiente. Entretanto, o custo para a implantação e a manutenção da AP é usualmente um problema para pequenos produtores. Assim, a divisão de áreas agrícolas em unidades homogêneas menores, conhecidas como zonas de manejo (ZMs), é tida como alternativa para a aplicação da AP (DOERGE, 2000), por possibilitar uso de equipamentos convencionais, além de reduzir o número de análises de solo necessário para a definição das recomendações de insumos.

A definição de ZMs pode ocorrer de diversas formas. Johannsen et al. (2000) apresentam uma abordagem com uso de sensoriamento remoto a fim de obter índices de vegetação e associá-los a grades de amostragem de solo. Outras abordagens consideram a sensibilidade do produtor por meio do conhecimento empírico, embora o método mais difundido na literatura consista em agrupar parâmetros químicos e físicos de solo coletados em pontos estratégicos georreferenciados (FRAISSE et al., 2001; MOLIN; FAULIN, 2013).

Com aplicação da mineração de dados associada a conceitos de AP é possível detectar as necessidades (por exemplo, excesso ou falta de nutriente) em uma determinada amostra ou área da plantação e definir o que deve ser acrescentado ou neutralizado para que esta amostra seja corrigida. Atualmente existem diversas ferramentas computacionais disponíveis publicamente que auxiliam na execução de algoritmos de mineração de dados, porém os resultados gerados por meio dessas ferramentas devem ser analisados por especialistas do domínio da aplicação, para que possam ser interpretados e transformados em conhecimento útil.

O software SDUM (Software para a definição de unidades de manejo), que possui interface trilíngue (português, espanhol e inglês), com download gratuito na internet (disponível em: <https://ftp.unioeste.br/SDUM/>), apresenta em sua estrutura a possibilidade de definição de ZMs, com aplicação do algoritmo fuzzy c-means, dentre outros, associado apenas à métrica de distância Euclidiana.

Para Vicini (2005), a distância euclidiana é, sem dúvida, a medida de distância mais utilizada para a análise de agrupamentos. Além disso, Seidl et al. (2008) diz que a distância euclidiana é a medida de distância mais frequentemente empregada quando todas as variáveis são quantitativas.

Segundo Manly (1986), a distância euclidiana, quando estimada a partir de variáveis originais, apresenta a inconveniência de ser influenciada pela escala destas, sendo necessário a normalização das variáveis em estudo, para que possuam a variância igual à unidade.

Entretanto, outras métricas podem ser utilizadas para a mesma finalidade de delineamento de agrupamento, como as distâncias Chebshev, Diagonal, Manhattan e Mahalanobis.

2 OBJETIVOS

2.1 GERAL

Avaliar três métricas de distâncias (medida de similaridade) junto aos Softwares FUZME e SDUM.

2.2 ESPECÍFICOS

- Avaliar as zonas de manejo definidas utilizando três métricas de distâncias: Diagonal, Euclidiana e Mahalanobis.

3 REVISÃO BIBLIOGRÁFICA

3.1 MINERAÇÃO DE DADOS E APRENDIZADO DE MÁQUINA

Segundo Hand (2001) mineração de dados (MD) é “a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de um modo que eles sejam tanto úteis quanto compreensíveis ao dono dos dados”. Fayyad (1996) apresenta uma definição a partir da perspectiva do aprendizado de máquina, na qual “MD é um passo no processo de descoberta de conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados”.

Aprendizagem de máquina (AM) é uma subárea da inteligência artificial responsável pelo desenvolvimento de teorias, técnicas e sistemas computacionais capazes de adquirir conhecimento de maneira automática e imitar o comportamento inteligente, descobrindo regras e/ou relacionamentos que permitem simular o processo de aprendizagem (MITCHELL, 1997). Seu principal objetivo é a construção de sistemas capazes de inferir conhecimento novo de maneira automática ou semi-automática a partir de um conjunto de dados que contém casos do problema analisado (MITCHELL, 1997). Softwares desenvolvidos com esta técnica auxiliam no processo de tomada de decisões com base no conhecimento prévio e acumulado por meio da interação com o ambiente e em experiências anteriores bem sucedidas (REZENDE, 2003). De maneira geral, podemos dizer que o aprendizado representa a capacidade que um sistema tem de se adequar e melhorar o seu desempenho na segunda vez que repetir a mesma tarefa, ou outra tarefa da mesma população (SIMON, 1983).

Três modos de aprendizado são considerados: supervisionado, semi-supervisionado e não-supervisionado, os quais diferenciam-se, entre outros, pela presença do atributo classe que rotula os exemplos do conjunto de dados. No aprendizado supervisionado o objetivo é induzir um classificador, por meio de um conjunto expressivo de dados previamente rotulados, para então classificar novos exemplos cuja classe não é conhecida. O modo de aprendizado semi-supervisionado, por outro lado, faz uso de poucos exemplos rotulados e muitos exemplos não rotulados, sendo que o objetivo é rotular um maior número de exemplos para os quais a classe não é conhecida. No aprendizado não-supervisionado, também conhecido como análise exploratória de dados e/ou clustering[1], o conjunto de dados é composto por exemplos não rotulados. Nesse caso, são utilizados algoritmos para descobrir padrões nos dados a partir de alguma caracterização de regularidade, sendo

[1]Apesar de existirem outras técnicas de aprendizado não-supervisionado, como sumarização e regras de associação, neste trabalho o termo clustering é utilizado como sinônimo de aprendizado não-supervisionado.

esses padrões denominados clusters. A ideia é agrupar uma coleção de exemplos segundo alguma medida de similaridade, de modo que exemplos pertencentes ao mesmo cluster devem ser mais similares entre si e menos similares aos exemplos que pertencem a outros clusters (EVERITT, 1993).

3.2. AGRUPAMENTOS DE DADOS

Existem diversas abordagens de clustering, como particional, probabilística, otimização e hierárquica (JAIN, 1999). Cada abordagem usa uma estratégia diferente para identificar e representar os clusters (ou grupos de instâncias). Os algoritmos pertencem à abordagem particional e representam os clusters por meio de elementos representativos (protótipos) chamados centróides (BERKHIN, 2002).

Segundo Linden (2009), a análise de agrupamento, ou clustering, tem o propósito de separar objetos em grupos, baseando-se nas características que os objetos possuem. Um cluster é considerado como uma coleção de instâncias similares entre si, porém com diferenças significativas das instâncias presentes nos demais grupos. O objetivo da criação destes grupos é maximizar a homogeneidade intragrupo e maximizar a heterogeneidade intergrupo. Os conjuntos de dados utilizados por algoritmos de clustering são representados por meio de uma estrutura de matriz, denominada tabela atributo-valor, na qual é disposto um conjunto E de N exemplos (ou casos) de treinamento E =$\{x_1, \ldots, x_N\}$ não rotulados, onde os x_I são vetores da forma $(x_{i1}, x_{i2}, \ldots, x_{iM})$ cujos componentes são valores discretos ou contínuos relacionados ao conjunto dos M atributos $X = \{X_1, X_2, \ldots, X_M\}$.

Em outras palavras, x_{il} denota o valor do atributo x_l do exemplo i. Um exemplo de clusterização é mostrado na Figura 1, na qual os pontos foram agrupados em três clusters onde os pontos pertencentes ao mesmo cluster recebem o mesmo rótulo.

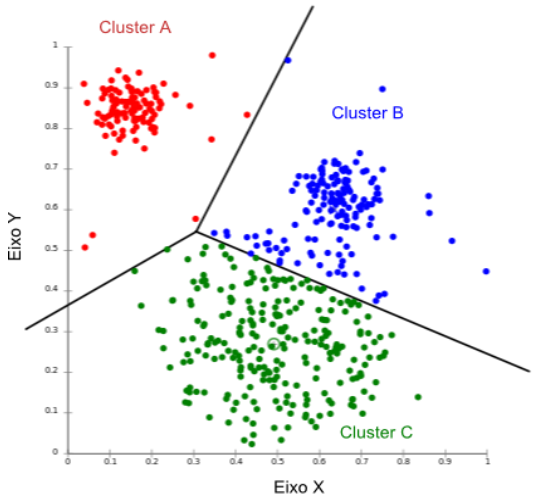


Figura 1 Exemplo de Clusterização de Dados.
Fonte: Adaptado Wikipedia - Cluster analysis (2016)

Um dos principais desafios em tarefas de agrupamento é a análise e a definição do número ideal de grupos. Uma forma alternativa de análise de agrupamentos é dar uma nota à qualidade do resultado do agrupamento, levando em conta distâncias intra e inter grupos em busca de grupos compactos e bem separados. A fim de decidir quantos clusters (neste caso, ZMs) devem ser gerados, considera-se que, quanto menor a quantidade, mais facilidade o produtor terá para realizar aplicação a taxa variável em sua lavoura (XIANG, 2007; PEDROSO et al., 2010; BAZZI et al., 2013). A utilização de métodos empíricos para aferir quantidade ideal direcionam em sua maioria para três ou quatro ZMs (SUSZEK et al., 2012), não descartando a opção por duas ZMs.

Modelos fuzzy têm sido utilizados para detectar a similaridade entre membros de uma coleção de objetos (WINDHAM, 1982; ATECA et al., 2001). Fridgen et al. (2000) utilizou o método fuzzy c-means para definição de ZMs, em um conjunto de dados que incluía condutividade elétrica aparente, elevação e declividade de duas áreas.

O algoritmo fuzzy c-means tem como objetivo encontrar grupos fuzzy para um conjunto de dados. Para alcançar este objetivo, o algoritmo precisa minimizar uma função que diz respeito à minimização das distâncias entre os dados e os centros dos grupos (centroides) aos quais tais dados pertencem com algum grau de pertinência (XU; WUNSCH, 2005).

Oliveira (2004) utilizou o método fuzzy c-means para definição de zonas de manejo para a cultura do mamoeiro, em um plantio comercial localizado em São Mateus, ES, com base em determinações realizadas através de amostragens e análises químicas do solo.

3.3. MÉTRICAS DE DISTÂNCIAS

As medidas de distância de uma maneira geral podem ser definidas como medidas de similaridade e dissimilaridade; na qual a primeira é para definir o grau de semelhança entre as instâncias e realizam o agrupamento de acordo com a sua coesão, e a segunda de acordo com as diferenças dos atributos das instâncias (WINTER, 2005).

Para particionar os dados em grupos são utilizadas algumas medidas de similaridade entre vetores, as quais servem para guiar o processo de construção da superfície de decisão que determinará a distribuição dos dados nos respectivos grupos. Existem diversas medidas utilizadas para o cálculo da similaridade, entre as quais estão as medidas de distância, de correlação e de associação. Quando o conjunto de dados é composto por atributos numéricos (quantitativos), as métricas de distância podem ser aplicadas para o cálculo da similaridade entre os exemplos (METZ; MONARD, 2006).

Para que uma distância seja validada como uma métrica ela deve atender a três propriedades: positividade, identidade e simetria, e também a propriedade da desigualdade

triangular. A similaridade é calculada com base em uma medida de distância entre dois vetores, usando o princípio de que quanto mais próximos estiverem duas instâncias, mais similares elas serão. Na análise de agrupamentos a similaridade entre duas amostras pode ser expressa como uma função da distância entre os dois pontos representativos destas amostras no espaço n-dimensional (NETO; MOITA, 1998).

Existem diferentes medidas de distância que podem ser aplicadas no processo de agrupamento de dados. Neste trabalho são utilizadas a distância Diagonal, Euclidiana e Mahalanobis.

A similaridade entre duas instâncias pode ser calculada como o complemento da distância $\mathrm{sim}(\mathrm{Ei}, \mathrm{Ej}) = 1 - \mathrm{dist}(\mathrm{Ei}, \mathrm{Ej})$, onde $\mathrm{dist}(\mathrm{Ei}, \mathrm{Ej})$ é a distância entre os exemplos Ei e Ej calculada a partir de uma métrica de distância pré-definida (METZ; MONARD, 2006).

Segundo Wintten (2005), na aprendizagem baseada em exemplo, cada nova instância é comparada com as já existentes, utilizando uma métrica de distância, e a instância mais próxima existente é usada para atribuir a classe para o novo exemplo. Isso é chamado de método de classificação do vizinho mais próximo.

Guerreiro e Breve (2015) apresentam em seu trabalho um estudo comparativo sobre o uso de diferentes métricas de distâncias: Euclidiana, Mahalanobis, City Block, Chebyshev, Minkowski, Bray-Custis e Canbera na formação de grafos, visando verificar a influência destas em bases de dados diferentes. Como resultados, constatou-se que Mahalanobis é a melhor métrica para a base Iris, City Block Normalizada para a base Wine e Euclidiana Normalizada para a base Banknote Authentication.

3.3.2 DISTÂNCIA DIAGONAL

A distância Diagonal é utilizada para padronizar medições no momento em que é detectada à igualdade de variância no processo de agrupamento (MCBRATNEY et al., 1985; ODEH et al., 1992). Para isso uma matriz diagonal é definida (CHITTLEBOROUGH; MCBRATNEY; ODEH, 1992):

$$A_D = \begin{vmatrix} (1/\sigma)^2 & 0 & \cdots & 0 \\ 0 & (1/\sigma)^2 & \cdots & 0 \\ \cdot & & \cdot & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ 0 & 0 & & (1/\sigma)^2 \end{vmatrix}$$

em que: A_D é a matriz diagonal e σ é o desvio padrão.

Com a matriz diagonal calculada aplica-se a Equação 2 para obtenção da distância (CHITTLEBOROUGH; MCBRATNEY; ODEH, 1992).

$$Dist\,(E_i, E_j) = \sqrt{(x_{il} - x_{jl}).\ A_D.(x_{il} - x_{jl})} \qquad \text{Eq.(2)}$$

em que: dist será a distância entre os pontos; e são os valores do atributo de cada ponto e AD é a matriz diagonal.

A distância Diagonal (MCBRATNEY et al., 1985; ODEH et al., 1992), também é insensível a variáveis correlacionadas, como distância Euclidiana. No entanto, ela compensa distorções que ocorrem ao se considerar os aglomerados esféricos, por ponderar por meio das variâncias das variáveis medidas.

Zang et al. (2010) apresentaram em seu trabalho um comparativo entre as métricas de distância Diagonal e Mahalanobis, em que definiram ZMs com uma ferramenta (ZoneMAP) de suporte à decisão para aplicação de taxa variável. Uma de suas conclusões foi que a distância Diagonal apresenta um computo mais rápido que a de Mahalanobis.

Já Bezdek, Ehrlich e Full (1984) utilizaram em seu trabalho as métricas de distância Euclidiana, Diagonal e Mahalanobis a fim de análise de dados geoestatísticos, evidenciando a detecção de outliers quando a distância de Mahalanobis é utilizada.

3.3.3 DISTÂNCIA EUCLIDIANA

A distância Euclidiana é definida como a raiz quadrada da soma das diferenças entre x_{il} e x_{jl} elevadas ao quadrado (Equação 3). Para sua aplicação recomenda-se a normalização das variáveis (normalização dos dados) antes de se obter o valor da distância Euclidiana, devido ao fato dos dados não se apresentarem na mesma escala de medidas (LIU et al., 2014).

$$Dist\,(E_i, E_j) = \sqrt{\sum_{l=1}^{M} \left(x_{il} - x_{jl}\right)^2} \qquad \text{Eq.(3)}$$

em que: *dist* será a distância entre os pontos; x_{il} e x_{jl} são os valores do atributo de cada ponto.

A distância Euclidiana é calculada a partir do centro da célula de origem (centróide) para o centro de cada uma das células vizinhas (LIU et al., 2014).

Associada aos métodos de agrupamento K-means e fuzzy c-means, a distância foi utilizada em um experimento em que buscava-se definir zonas de manejo para cafeicultura, com base em determinações realizadas com sensor de clorofila e por análise foliar. Como resultado, destacou-se que os métodos de agrupamento de dados K-Means e fuzzy c-means, ambos com distância Euclidiana, não apresentaram diferenças significativas na definição das zonas de manejo (RODRIGUES et al., 2011).

3.3.5 DISTÂNCIA MAHALANOBIS

A distância de Mahalanobis é uma métrica que difere da distância Euclidiana por levar em consideração a covariância entre os dados analisados. Sendo assim, não depende

da escala das informações, não necessitando a normalização dos dados para se obter escalas equivalentes. A fórmula analisa a distância de Mahalanobis entre dois vetores da mesma distribuição que possuam uma matriz de covariância S. Ela considera em seu desenvolvimento a raiz quadrada das diferenças entre x_{il} e x_{jl} transpostas, vezes a matriz de variância amostral elevado a menos um vezes a diferença entre x_{il} e x_{jl} (Equação 5) (LIU et al., 2014).

$$Dist\,(E_i, E_j) = \sqrt{\left(x_{il} - x_{jl}\right)^T .\ S^{-1}.\left(x_{il} - x_{jl}\right)} \qquad \text{Eq.(5)}$$

em que: *dist* será a distância entre os pontos; x_{il} e x_{jl} são os valores do atributo de cada ponto; e S é a matriz de covariância.

A matriz de covariâncias entre grupos é calculada com base nos atributos selecionados. Se a matriz de covariância é a matriz identidade, a distância de Mahalanobis coincide com a distância Euclidiana. Se a matriz de covariância é diagonal, então a medida de distância resultante é chamada distância Euclidiana normalizada (Equação 6) (LIU et al., 2014).

$$\text{Dist}\,(E_i, E_j) = \sqrt{\sum_{l=1}^{M} \frac{\left(x_{il} - x_{jl}\right)^2}{\sigma^2}} \qquad \text{Eq. (6)}$$

em que: *dist* será a distância entre os pontos; x_{il} e x_{jl} são os valores do atributo de cada ponto; σ é o desvio padrão.

O uso da distância de Mahalanobis corrige algumas das limitações da distância Euclidiana, pois leva em consideração automaticamente a escala dos eixos das coordenadas e também a covariância entre as características (LIU et al., 2014).

A detecção de outliers[2] é um dos usos mais comuns da distância de Mahalanobis, pois um valor alto determina que um elemento está a vários desvios padrões da média e, por consequência, é provavelmente um outlier (PENNY, 1987).

3.4. SELEÇÃO DE ATRIBUTOS PARA DEFINIÇÃO DE ZONAS DE MANEJO

A utilização de um grande número de atributos pode contribuir positiva ou negativamente para o processo de definição de ZMs. Se em pequena quantidade, pode não fornecer as características suficientes para a definição de agrupamentos distintos; se em grande quantidade, pode confundir as informações correspondentes aos atributos e ocasionar má distribuição dos agrupamentos. Para indicação da quantidade ideal de atributos utilizados na definição das ZMs podem ser empregados dois métodos: Correlação Espacial Cruzada e Análise dos Componentes Principais – ACP.

[2]Valor atípico, é uma observação que apresenta um grande afastamento das demais séries.

3.4.1 CORRELAÇÃO ESPACIAL CRUZADA

É uma técnica que tem como objetivo avaliar se amostras possuem correlação espacial (autocorrelação) e também se este tipo de correlação é significativa entre duas variáveis. Pode-se calcular a correlação espacial cruzada I_{XY} por meio da Equação 7, o que resultará em um valor pertencente ao intervalo *[-1, 1]*, que representa o nível de associação entre as variáveis *X* e *Y* (REICH, 2008):

$$I_{XY} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(W_{ij}.\ X_i\ .\ Y_j)}{W\sqrt{m_X{}^2.\ m_Y{}^2}} \qquad \text{Eq.(7)}$$

em que: W_{ij} é chamada *matriz de associação espacial*, sendo 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* transformada, no ponto *i*; Y_j é o valor da variável *Y* transformada, no ponto *j*; *W* corresponde à soma dos graus de associação espacial, obtidos da matriz $W_{ij,}$ para $i \neq j$; $m_X{}^2$ corresponde à variância amostral de *X*; $m_Y{}^2$ corresponde à variância amostral de *Y*. Neste ponto, deve-se interpretar a transformação de uma variável *Z* como o procedimento executado sobre seus valores para que ela fique com média igual a zero, aplicando-se a equação $Z_k = (Z_k - \bar{Z})$, em que $\bar{Z}$ representa a média amostral de *Z*. Quando se obtém o valor da correlação espacial cruzada I_{XY}, diz-se que há correlação positiva se $I_{XY} > 0$, ou que há correlação negativa se $I_{XY} < 0$.

Após a determinação da autocorrelação para as amostras de cada variável e da correlação cruzada para cada possível par de variáveis *(X, Y)*, pode-se então gerar a matriz de correlação espacial, uma matriz simétrica que apresentará todos os valores I_{XY} calculados. Finalmente, essa matriz poderá ser utilizada para a identificação dos atributos necessários para o delineamento de ZMs, seguindo o procedimento proposto por Bazzi (2011):

1. Eliminar os atributos com autocorrelação espacial não significativa a 95% de confiança (ou 5% de significância);
2. Remover os atributos que não possuam correlação espacial com a produtividade;
3. Ordenar de modo decrescente os atributos restantes, considerando o módulo do grau de correlação espacial cruzada com a produtividade;
4. Eliminar os atributos redundantes (que se correlacionem entre si), dando preferência para a retirada dos que possuam menor correlação espacial com a produtividade;
5. Finalmente, os atributos restantes poderão ser utilizados na definição das ZMs.

3.4.2 ANÁLISE DE COMPONENTES PRINCIPAIS

A Análise de Componentes Principais (ACP) é um método que tem por finalidade básica a análise dos dados usados visando sua redução, eliminação de sobreposições e a escolha das formas mais representativas de dados a partir de combinações lineares de variáveis originais (JOLLIFFE, 2002).

A ACP é uma técnica de análise multivariada que tem como principal objetivo reduzir a dimensão de análise de conjuntos de dados associados a variáveis quantitativas correlacionadas entre si. A ACP permite identificar as variáveis que explicam a maior parte da variância total contida em conjuntos de dados, além de explicitar relacionamentos que podem existir entre variáveis ou entre observações de uma população ou de uma amostra (JOLLIFFE, 2002).

Jolliffe (1972) recomenda que sejam descartadas as variáveis quando a análise de componentes principais utiliza a matriz de correlação, estabelece-se que o número de variáveis descartadas deve ser igual ao número de componentes cuja variância (autovalor) é inferior a 0,7.

3.5 MÉTODOS PARA AVALIAÇÃO DE ZONAS DE MANEJO

A identificação da quantidade ideal de agrupamentos formados através da aplicação de algoritmos fuzzy pode ser avaliada através de quatro índices: Redução de variância, Índice de Desempenho Fuzzy, Índice de partição da entropia modificada e Índice de validação de cluster (Equações de 8 a 11).

- Redução de variância – RV (Equação 8) (DOBERMANN et al., 2003; XIANG et al., 2007)

$$RV = 1 - \frac{\sum_{i=1}^{n} W_i * V_{um_i}}{V_{área}} * 100 \qquad \text{Eq.(8)}$$

em que: n corresponde ao tamanho da amostra para toda a área, W_i é a proporção da área em cada unidade de manejo, V_{um_i} é a variância dos dados de cada unidade de manejo e $V_{área}$ é a variância da amostra dos dados para toda a área.

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 (GAVIOLI et al., 2016).

- Índice de desempenho Fuzzy – FPI (Equação 9) (FRIDGEN et al., 2004)

$$FPI = 1 - \frac{c}{(c-1)}\left[1 - \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^2 / n\right] \quad \text{Eq.(9)}$$

em que: c corresponde ao número de agrupamentos, n é o tamanho da amostra para toda a área (número de observações), u_{ij} é o elemento *ij* da matriz de pertinência Fuzzy.

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 (GAVIOLI et al., 2016).

- Índice da partição da entropia modificada – *MPE* (Equação 10) (FRIDGEN et al., 2004)

$$MPE = \frac{-\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\log(u_{ij})/n}{\log c} \quad \text{Eq.(10)}$$

em que: c corresponde ao número de agrupamentos, n é o tamanho da amostra para toda a área (número de observações), u_{ij} é o elemento *ij* da matriz de pertinência Fuzzy.

O índice 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 (GAVIOLI et al., 2016).

- Índice de Validação de Cluster Melhorado – ICVI (Equação 11) (GAVIOLI et al., 2016)

$$ICVI_i = \frac{1}{3} * \left(\frac{FPI_i}{Max\{FPI\}} + \frac{MPE_i}{Max\{MPE\}} + \left(1 - \frac{VR_i}{Max\{RV\}}\right)\right) \quad \text{Eq.(11)}$$

em que: *i* corresponde aos índices de todas as métricas de distância utilizadas no experimento.

O número ideal de agrupamentos de um conjunto de dados baseia-se no valor mínimo de FPI e MPE, e no valor máximo do RV. Para evitar a situação em que estas estimativas apontem para diferentes agrupamentos, quando analisados de forma individual, o *ICVI* pode ser utilizado unindo os conceitos e considerando o agrupamento que apresentar o menor *ICVI* como o melhor (GAVIOLI et al., 2016).

Após a construção das ZMs, é importante fazer a avaliação destas com intuito de constatar se o delineamento representa diferença significativa de potencial produtivo da cultura, independentemente do método que tenha sido empregado para este fim. Esta avaliação é importante para verificar se cada unidade pode ser tratada como subárea de gerenciamento diferenciado do restante do talhão e se pode ser utilizada como fonte de recomendação e de análise para atributos físicos (MORAL et al., 2010; BAZZI, 2011; SALEH; BELAL, 2014).

Durante o processo de avaliação, qualquer atributo amostral pode ser utilizado; porém, a produtividade geralmente é o atributo indicado (BAZZI, 2011). Também é importante definir a melhor quantidade de ZMs a serem efetivamente implantadas em um talhão, sabendo-se que quanto menor for essa quantidade, mais fácil será executar operações em campo (FRAISSE et al., 2001).

Dentre os vários métodos existentes para a realização da avaliação das ZMs, duas serão citadas neste trabalho: Análise da Variância - ANOVA e Índice de Suavidade.

- Análise de Variância - ANOVA:

Análise de variância é uma técnica estatística que permite avaliar afirmações sobre as médias de populações (MILONE, 2009). A análise visa, primordialmente, verificar se existe uma diferença significativa entre as médias e se os fatores exercem influência em alguma variável dependente.

Após a identificação de que as médias são estatisticamente diferentes, é necessário um método que forneça a diferença mínima significante entre duas médias. Essa diferença seria o instrumento de medida. Toda vez que o valor absoluto da diferença entre duas médias é igual ou maior do que a diferença mínima significante, as médias são consideradas estatisticamente diferentes, ao nível de significância estabelecida (VIEIRA et al., 1989).

De acordo com Bazzi (2011), o teste de comparação de médias da ANOVA pode ser aplicado para verificar se as subáreas definidas realmente representam grupos diferentes a determinado nível de significância. Porém, esse teste considera que as amostras são independentes dentro de cada unidade.

- Índice de Suavidade (Smoothness index) - *SI*

A avaliação dos melhores métodos de definição de agrupamento deve também incluir o aspecto visual do agrupamento criado e, portanto, deve-se levar em conta a suavidade das curvas de contorno, pois facilita a interpretação visual e a aplicação em taxa variada de insumos agrícolas. O índice de suavidade (*SI*, Equação 12), proposto por Bazzi et al. (2010), calcula a frequência da mudança de classes nos mapas temáticos nas direções horizontais, verticais e diagonais, pixel por pixel. Na hipótese de que o mapa seja uma única área totalmente homogênea, um índice de suavidade de 100% será obtido, devido à ausência de mudança de classe. Da mesma forma, se o mapa foi gerado com valores aleatórios, o índice SI apresentaria um valor próximo de zero.

$$SI = 100 - \left(\left(\frac{\sum_{i=1}^{l} NM_{Hi}}{4P_H} + \frac{\sum_{i=1}^{c} NM_{Vi}}{4P_V} + \frac{\sum_{i=1}^{n} NM_{Ddi}}{4P_{Dd}} + \frac{\sum_{i=1}^{n} NM_{Dei}}{4P_{De}} \right) * 100 \right) \qquad \text{Eq. (12)}$$

em que: NM_{Hi} corresponde ao número de mudanças na linha horizontal i; $NM_{V\,i}$ é o número de mudanças na linha vertical j; NM_{Ddi} é o número de mudanças na diagonal direita; NM_{Dei} é

o número de mudanças na diagonal esquerda; P_H é a possibilidade de mudança na horizontal; P_V é a possibilidade de mudança na vertical; P_{Dd} é a possibilidade de mudança na diagonal direita e P_{De} é a possibilidade de mudança na diagonal esquerda.

4 REFERÊNCIAS

ATECA, M. R.; SERENO, R.; APEZTEGUÍA, H. Zonificacion de uma superficie cultivada com soja segun aspectos fenométricos y consumo de agua del suelo. **Revista Brasileira de Agrometeorologia**, Santa Maria, v. 9, n. 1, p.111-116, 2001.

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

BAZZI, C. L. **Software para definição e avaliação de unidades de manejo em agricultura de precisão**. 2011. 123f. Tese (Doutorado em Engenharia Agrícola). Programa de Pós-Graduação em Engenharia Agrícola. Universidade Estadual do Oeste do Paraná. Cascavel, 2011.

BAZZI, C. L.; SOUZA, E. G. de; QUEIROZ, F. N. de; SANTOS, D. dos; KONOPATZKI, M. R. S. Influência do tipo de interpolador em mapas de resistência a penetração. Congresso Brasileiro de Agricultura de Precisão, Ribeirão Preto. **Anais...** 2010.

BERKHIN, P. **Survey of clustering data mining techniques.** San Jose: Accrue Software, 2002.

BEZDEK, J.; EHRLICH, R.; FULL, W. FCM: The Fuzzy c-Means Clustering Algorithm. **Computers & Geosciences**, v. 10, n. 2-3, p. 191-203, 1984.

CHITTLEBOROUGH, D.; MCBRATNEY, A.; ODEH, O. Soil Pattern Recognition with Fuzzy-c-means: Application to Classification and Soil-Landform Interrelationships. **Soil Science Society of America Journal,** n. 56, p. 505-516, 1982.

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. 5, p. 1105-1120, 2003.

DOERGE, T. A. **Management Zone Concepts.** Site-Specific Management Guidelines, 2000.

EVERITT, B. S. **Cluster Analysis.** London: Edward Arnold, 1993.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-54, 1996.

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 Engineers,** v. 44, n. 1, p. 155-166, 2001.

FRIDGEN, J. J.; KITCHEN, N. R.; SUDDUTH, K. A. Variability of soil and landscape attributes within sub-field management zones. In: International Conference on Precision Agriculture. **Anais…** Bloomington: Madison, 2000.

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.

GUERREIRO, L.; BREVE, F. A. **Analysis of the Influence of Distance Metrics on the Semi-supervised Algorithm of Particle Competition and Cooperation.** Rio Claro - SP: Universidade Estadual Paulista, 2015.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining.** Chicago: MIT Press, 2001.

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

JOHANNSEN, C. J.; CARTER, P. J.; ERICKSON, B. J.; MORRIS, D. K.; WILLIS, P. R. A cornucopia of agricultural applications. **Space Imaging**, Thornton, Jan/Fev, p.22-23, 2000.

JOLLIFE, I. T. Discarding variables in a principal component analysis. **Journal of the Royal Statistical Society. Series C (Applied Statistics),** v. 21, n. 2., p. 160-173, 1972.

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

LINDEN, R. Técnicas de Agrupamento. **Revista de Sistemas de Informação da FSMA**, n. 4, p. 18-36, 2009.

LIU, Q.; CHU, X.; XIAO, J.; ZHU, H. Optimizing Non-orthogonal Space Distance Using PSO in Software Cost Estimation. In: IEEE Computer Software Applications Conference (COMPSAC). **Anais...,** 2014.

MANLY, B. F. J. **Multivariate statistical methods:** a primer. London: Chapman and Hall, 1986.

METZ, J.; MONARD, M. C. **Projeto e implementação do módulo de clustering hierárquico do discover**. ICMC-USP, 2006.

MILONE, G. **Estatística geral e aplicada.** São Paulo: Centage Learning, 2009.

MITCHELL, T. M. **Machine learning.** Mcgraw-Hill series in computer science, 1997.

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.

MOLIN, J. P.; AMARAL, L. R.; COLAÇO, A. **Agricultura de Precisão.** São Paulo: Oficina de Textos, 2015.

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.

NETO, JM Moita; MOITA, GraziellaCiaramella. **Uma introdução à análise exploratória de dados multivariados.** Química nova, v. 21, n. 4, p. 467-469, 1998.

ODEH, I.O.A.; MCBRATNEY, A.B.; CHITTLEBOROUGH, D.J. Soil pattern recognition with fuzzy-c-means: application to classification and soil –landform interrelationships. **Soil Science Society of America Journal**, n. 56, p. 505-516, 1992.

OLIVEIRA, F. B. **Utilização de lógica Fuzzy na geração de zonas de manejo**. Alegre - ES: Universidade Federal do Espírito Santo, 2014.

PEDROSO, M.; TAYLOR, J.; TISSEYRE, B.; CHARNOMORDIC, B.; GUILLAUME, S. A segmentation algorithm for the delineation of agricultural management zones. **Computers and Electronics in Agriculture,** Netherlands, v. 70, n. 1, p.199-208, 2010.

PENNY, K. I. Appropriate Critical Values when Testing for a Single Multivariate Outlier by using the Mahalanobis Distance. In: **Applied Statistics**. Royal Statistical Society, UK, 1987.

REICH, R. M. **Spatial Statistical modeling of natural resources.** Colorado State University: Fort Collins, 2008.

REZENDE, S. O. **Sistemas Inteligentes**: fundamentos e aplicações. Barueri: Manole, 2003.

RODRIGUES, 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**. São Paulo: UNICAMP, 2011.

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, 2014. doi: 10.1088/1755-1315/18/1/012046.

SEIDL, E. J.; MOREIRA JÚNIOR, F. de J.; ANSUJ, A. P.; NOAL, M. R. C. Comparação entre o método ward e o método k-médias no agrupamento de produtores de leite. **Ciência e Natura,** v. 30, n. 1, p. 7-15, 2008.

SILVA, S.; ALMEIDA, L. A.; DIAS, L. C. Processo de decisão espacial multicritério e modelo multiobjeto para a localização de centrais de biogás. Simpósio Brasileiro de Pesquisa Operacional, Porto de Galinhas, PE. **Anais...** 2015.

SIMON, H. A. Why should machines learn?. In: MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. (Eds.). **Machine Learning I**, p. 25-37. Los Altos, CA: Morgan Kaufmann, 1983.

SUSZEK, G.; SOUZA, E. G.; URIBE-OPAZO, M. A.; NÓBREGA, L. H. P. Determination of management zones from normalized and standardized equivalent productivity maps in the soybean culture. **Engenharia Agrícola,** Jaboticabal, v. 32, 2012.

VICINI, L. **Analise multivariada da teoria à prática.** Universidade Federal de Santa Maria. Santa Maria, UFSM: CCNE, 2005.

VIEIRA, S.; HOFFMANN, R. **Estatística experimental**. São Paulo: Atlas, 1989.

WINDHAM, M. P. Cluster validity C-means for Fuzzy clustering algorithm. **IEEE Transactions on Pattern Analyses Machine Intelligence**, v. 11, p. 357-363, 1982.

WITTEN, I. H.; FRANK, E. **Data Mining**: practical machine learning tools and techniques. San Francisco: Morgan Kaufmann, 2005.

XIANG, L.; YU-CHUN, P.; ZHONG-QIANG, G.; CHIN-JIANG, Z. 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.

XU, R.; WUNSCH, D. Survey of clustering algorithms. **IEEE Transactions on Neural Networks**, v. 16, n. 3, p. 645-678, 2005.

WIKIPEDIA. **Cluster analysis.** Disponível em: <https://en.wikipedia.org/wiki/Cluster_ analysis>. Acesso em: 26 out. 2016.

YANG, Z.; SHUFAN, Y.; YANG, X.; LIQUN, G. High-Dimensional Statistical Distance for Object Tracking. International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). **Anais…** 2010.

ZHANG, X.; SHI, L.; JIA, X.; SEIELSTAD, G.; HELGASON, C. Zone mapping application for precision-farming: a decision support tool for variable rate application. **Precision Agriculture**, v. 11, n.2, p. 103-114, 2010.

5 ARTIGO 1 – ANÁLISE COMPARATIVA ENTRE MÉTRICAS DE DISTÂNCIAS UTILIZANDO O ALGORITMO DE AGRUPAMENTO FUZZY C-MEANS PARA DEFINIÇÃO DE ZONAS DE MANEJO

Resumo

A Agricultura de Precisão faz uso de tecnologias objetivando o aumento da produtividade e a redução do impacto ambiental, sendo uma de suas técnicas a definição de zonas de manejo (ZMs). As ZMs consistem-se de subáreas com características topográficas, de solo e/ou de plantas cultivadas similares, onde um único tratamento pode ser utilizado. Na definição das ZMs é comum a aplicação de algoritmos de agrupamento como o fuzzy c-means, que usualmente associa aos seus cálculos a métrica de distância euclidiana. No entanto, existem outras métricas que podem ser associadas ao algoritmo. Nesse contexto, o objetivo deste trabalho foi avaliar a utilização de outras duas métricas de distâncias para a definição de ZMs: Diagonal e Mahalanobis. A avaliação foi realizada com dados obtidos entre os anos de 2012 e 2015 em quatro áreas agrícolas, localizadas no estado do Paraná, cultivadas com soja. A partir das variáveis disponíveis para cada uma das áreas foi realizada a seleção de variáveis através do método da correlação espacial. Para a definição das ZMs foram definidos dois conjuntos de dados, sendo que o primeiro foi criado a partir dos dados originais das áreas, e no outro se utilizou a normalização destes através do método da amplitude. Em todos os casos, as distâncias Diagonal e Mahalanobis dispensaram a necessidade de normalização das variáveis, apresentando áreas idênticas para as duas versões. Após a normalização dos dados, apenas a distância Euclidiana apresentou um melhor delineamento em suas ZMs. Na área A, o melhor desempenho foi para a distância Euclidiana utilizando dados normalizados. Para as áreas B, C e D não foi possível obter conclusões quanto ao melhor desempenho, visto que o fato de ser utilizado apenas uma variável para o processo de definição de ZMs influenciou diretamente no processo, ocasionando ZMs idênticas entre todas as distâncias utilizadas.

Palavras-chave: Agricultura de Precisão, Clusterização, Distância Diagonal, Distância Euclidiana, Distância Mahalanobis, Soja.

COMPARATIVE ANALYSIS AMONG SIMILARITY MEASURES USING FUZZY C-MEANS CLUSTERING ALGORITHM FOR MANAGEMENT ZONES DEFINITION

Abstract

Precision Agriculture uses technologies aimed to increasing productivity and reducing environmental impact, one of its techniques being the definition of management zones (MZs). MZs consist of subareas with similar topographic, soil and/or cultivated plant characteristics, noting that a single treatment can be used. In the definition of MZs it is common to apply clustering algorithms such as fuzzy c-means, which usually associates the Euclidean distance metric to its calculations. However, there are other metrics that can be associated with the algorithm. In this context, the objective of this research was to evaluate the use of two other distance metrics for the definition of MZs: Diagonal and Mahalanobis. The evaluation was carried out with data gathered between the years of 2012 and 2015 in four agricultural areas, located in the State of Paraná, cultivated with soybean. From the variables available for each of the areas, the selection of variables was carried out through the spatial correlation method. Two sets of data were defined for the definition of the MZs, considering that the first one was created from the original data from the areas, and in the other, the normalization of the data was used through the amplitude method. In all cases, the Diagonal and Mahalanobis distances exempted the need for normalization of the variables, presenting areas that were identical for both versions. After the normalization of the data, only the Euclidian distance presented a better delineation in its MZs. In area A, the best performance was for Euclidean distance using normalized data. For the areas B, C, and D it was not possible to reach conclusions regarding the best performance, since only one variable was used for the process of MZ definition and that has directly influenced the output, causing identical MZs among all the distances used.

Keywords: Precision Agriculture, Clustering, Diagonal Distance, Euclidian Distance, Mahalanobis Distance, Soybean.

5.1 INTRODUÇÃO

A utilização da tecnologia na agricultura é cada vez mais pertinente, devido à necessidade crescente do aumento de produção e lucratividade, a diminuição do uso de defensivos e o impacto ambiental nos mais variados ramos rurais, visando sempre beneficiar estes aspectos (MOLIN, 2015).

A agricultura de precisão (AP) tem como principal objetivo o fornecimento da necessidade local de insumos agrícolas. Ela apresenta o potencial de aumentar a eficiência do uso de insumos como fertilizantes, água, sementes e herbicidas sem prejudicar o meio ambiente. Entretanto, o custo para a implantação e a manutenção da AP é usualmente um problema para pequenos produtores. Assim, a divisão de áreas agrícolas em unidades homogêneas menores, conhecidas como zonas de manejo (ZMs), é tida como alternativa para a aplicação da AP (DOERGE, 2000), por possibilitar o uso de equipamentos convencionais, além de reduzir o número de análises de solo necessário para a definição das recomendações de insumos.

A definição ZMs pode ocorrer de diversas formas, embora o método mais difundido na literatura consista em agrupar parâmetros químicos e físicos de solo coletados em pontos estratégicos georreferenciados (FRAISSE et al., 2001; MOLIN; FAULIN, 2013), utilizando-se o algoritmo fuzzy c-means (FRIDGEN et al., 2004).

Para gerar os clusters, os algoritmos utilizam medidas de similaridade, que guiam o processo de decisão que determinará a distribuição dos dados nos respectivos grupos. Há diversas medidas para o cálculo de similaridade, dentre as quais estão distância, correlação e associação. Quando o conjunto de dados é composto por variáveis quantitativas, as métricas de distância podem ser aplicadas para o cálculo da similaridade entre os dados (METZ; MORNARD, 2006).

O software SDUM (Software para a definição de unidades de manejo) possui interface trilíngue (português, espanhol e inglês), com download gratuito na internet (disponível em: <https://ftp.unioeste.br/SDUM/>). Também apresenta em sua estrutura a possibilidade de definição de ZMs, com aplicação do algoritmo fuzzy c-means, dentre outros, associado apenas à métrica de distância Euclidiana.

Para Vicini (2005), a distância euclidiana é, sem dúvida, a medida de distância mais utilizada para a análise de agrupamentos. Além disso, Seidl et al. (2008) dizem que a distância euclidiana é a medida de distância mais frequentemente empregada quando todas as variáveis são quantitativas.

Segundo Manly (1986), a distância euclidiana, quando estimada a partir de variáveis originais, apresenta a inconveniência de ser influenciada pela escala destas, sendo necessário a normalização das variáveis em estudo, para que possuam a variância igual à

unidade. Schenatto et al. (2017) conduziram um experimento em três áreas comerciais (com 9,9, 15,0 e 19,8 ha), localizado no estado do Paraná, no sul do Brasil. As variáveis usadas para definir as MZs foram selecionadas usando estatísticas de correlação espacial e os dados foram normalizados usando três métodos (standard score, amplitude e média aritmética. As MZs foram definidas usando o algoritmo fuzzy c-means, que criou dois, três e quatro clusters. Verificou-se que, quando a delimitação das MZs usa mais de uma variável com diferentes escalas e no processo de agrupamento é usado a distância Euclidiana, é necessária uma normalização. O método da amplitude foi considerado o melhor método de normalização.

Entretanto, outras métricas podem ser utilizadas para a mesma finalidade de definição de ZMs, como a distância Diagonal e Mahalanobis. Este trabalho teve como objetivo, portanto, avaliar a influência das métricas de distâncias Diagonal, Euclidiana e Mahalanobis, dentro do algoritmo fuzzy c-means, na qualidade e na representatividade das ZMs.

5.2 MATERIAL E MÉTODOS

Um fluxograma (Figura 1) foi criado para apresentar as etapas seguidas durante a definição e a avaliação das ZMs.

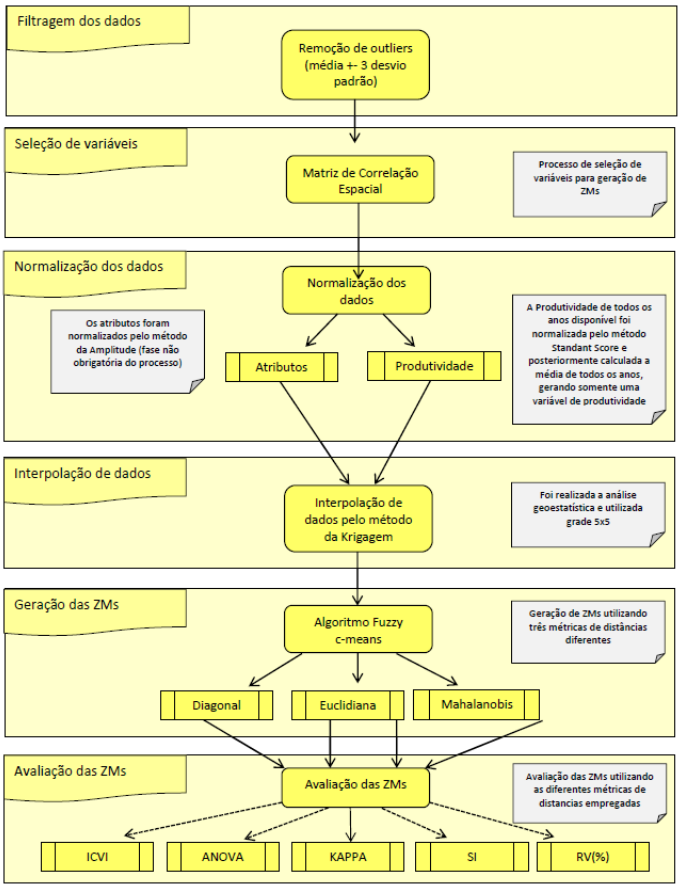


Figura 1 Fluxograma de Análise de Processo.

5.2.1 CONJUNTO DE DADOS

O conjunto de dados reais que foi utilizado neste artigo pertence a quatro áreas agrícolas no estado do Paraná (Figura 2), sendo que o tamanho e a localização das áreas, bem como o número de amostras realizadas estão apresentados na Tabela 1.

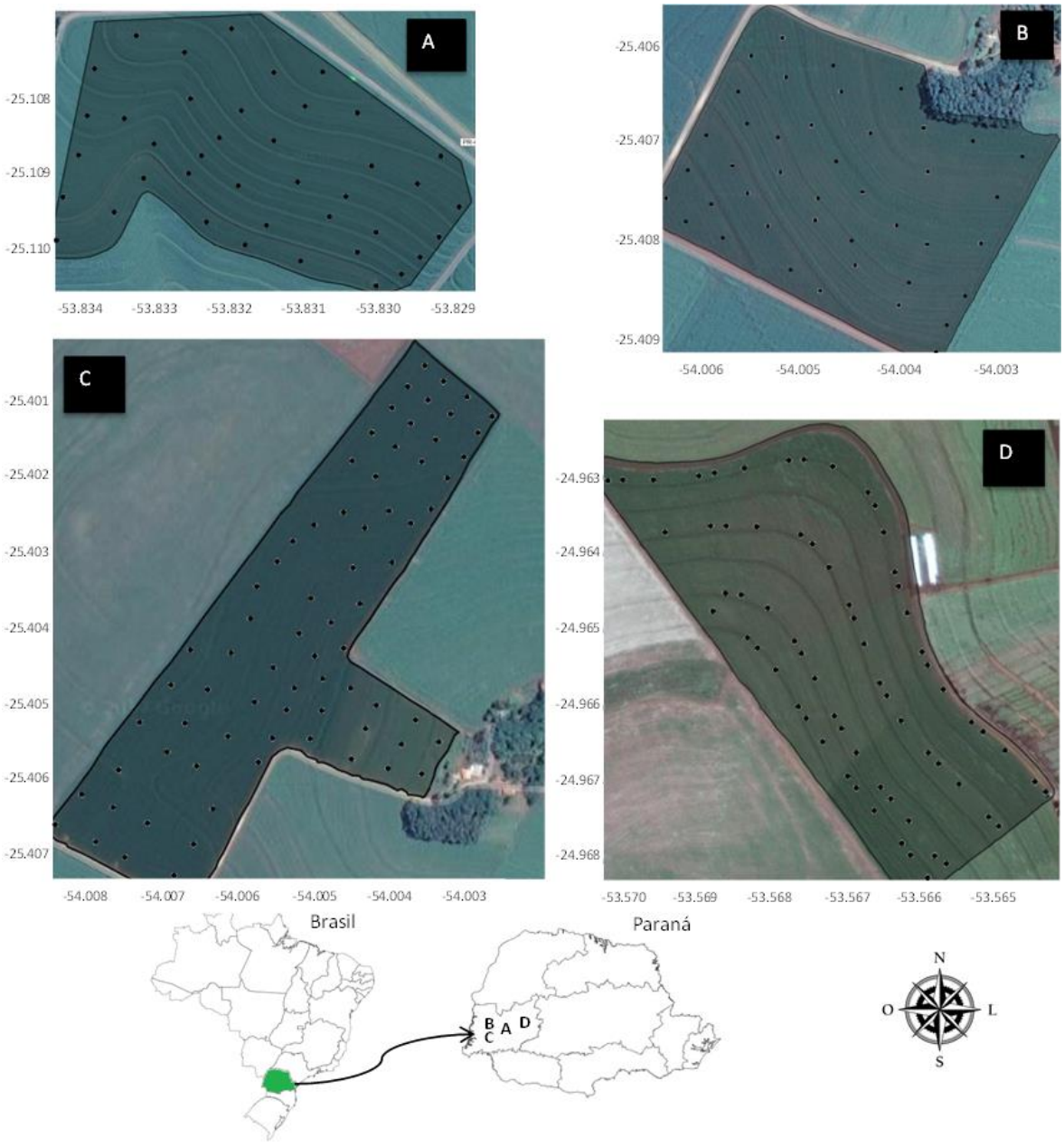


Figura 2 Localização das áreas experimentais na Região Oeste do Paraná.

Tabela 1 Identificação, tamanho, localização e altitude média das áreas de estudo

ID	**Cidade**	**Área** (ha)	**Coordenadas** (SAD 69 - WGS 1984)	**Elevação**(m)	**Números de Pontos**	**Pontos por hectare**
A	Céu Azul	15,0	25º06'32" S e 53º49'55" O	460	40	2,7
B	Serranópolis do Iguaçu	9,9	25º24'28" S e 54º00'17" O	355	42	4,2
C	Serranópolis do Iguaçu	23,8	25º24'28" S e 54º00'17" O	355	73	3,1
D	Cascavel	19,8	24º57'08" S e 53º33'59" O	650	68	3,4

A primeira etapa executada foi a “limpeza” dos dados, em busca da redução de discrepâncias de ruídos e correção inconsistências. Nessa fase os dados são modificados de que acordo com formatos apropriados à definição de ZMs. Na fase de filtragem/processamento dos dados foram considerados como outliers os valores superiores ou inferiores a três vezes o desvio padrão, tendo como referência a média da variável (CÓRDOBA et al., 2016). Quando identificados, estes foram removidos do conjunto de dados. As quatro áreas estudadas fornecem oito conjuntos de dados, sendo que em uma versão teremos quatro conjuntos com dados normalizados e uma segunda versão com quatro conjuntos de dados não normalizados. Com o experimento envolvendo as duas versões foi possível identificar sobre quais aspectos a normalização influencia ou não na definição das ZM.

Para a definição dos agrupamentos foram utilizados somente variáveis consideradas estáveis, excluindo, portanto, os atributos químicos do solo, satisfazendo recomendação geral de literatura (DOERGE, 2000). Grades de amostragem densas foram utilizadas a fim de observar as restrições de análise geoestatística (JOURNEL; HUIJBREGTS, 1978), com pelo menos 2,5 pontos ha^{-1}. As grades são irregulares e foram definidas considerando a linha imaginária central entre as curvas de nível de cada área. A Tabela 2 apresenta as variáveis que foram coletadas e usadas como variáveis (atributos) candidatas a serem utilizadas na definição dos agrupamentos: resistência mecânica do solo à penetração (0 - 10 cm, SRP 0.0 - 0.1 m; 10 - 20 cm, SRP 0.1 - 0.2 m; e 20 - 30 cm, SRP 0.2 - 0.3 m); elevação (%); declividade (°); densidade (g cm^{-3}); argila (%); silte (%); areia (%); matéria orgânica (%); e produtividade de soja (t ha^{-1}). Os dados de produtividade da cultura para a zona A foram determinados utilizando monitor de colheita CASE AFS PRO 600 acoplado a uma colhedora CASE IH 2388. Para as áreas B, C e D, a produtividade foi determinada por meio da colheita de uma área de amostragem de 1 m^2 em cada um dos pontos de amostragem. A produtividade foi então corrigida para o conteúdo de água de 13% do grão.

Tabela 2 Variáveis coletadas em função do ano agrícola e área experimental

Atributos	Área A					Área B					Área C					Área D	
	2012	2013	2014	2015	2016	2012	2013	2014	2015	2016	2012	2013	2014	2015	2016	2010	2011
SRP 0.0 - 0.1 m (MPa)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
SRP 0.1 - 0.2 m (MPa)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
SRP 0.2 - 0.3 m (MPa)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
Elevação (m)	X					X					X					X	
Declividade (˚)	X					X					X					X	
Densidade (g cm^{-3})	X					X					X					X	
Argila (%)	X					X					X					X	
Silte (%)	X					X					X					X	
Areia (%)	X					X					X					X	
Matéria Orgânica (%)	X					X					X					X	
Produtividade da soja (t ha^{-1})	X	X	X			X	X	X			X	X	X			X	X

*SPR – resistência mecânica à penetração.

5.2.2 SELEÇÃO DE VARIÁVEIS

O processo de seleção de variáveis buscou encontrar as variáveis ideais para a definição das ZMs. A técnica utilizada neste trabalho foi a Correlação Espacial Cruzada, que utiliza-se de uma matriz de correlação espacial como objeto de análise e na sequência aplicam-se os procedimentos propostos por Bazzi et al. (2013) e Schenatto et al. (2016), em que: eliminam-se as variáveis com autocorrelação espacial não significativa a 95% de confiança (ou 5% de significância); removem-se as variáveis que não possuam correlação com a produtividade; ordenam-se de modo decrescente as variáveis restantes, considerando o módulo do grau de correlação cruzada com a produtividade; eliminam-se as variáveis redundantes (que se correlacionem entre si), dando preferência para a retirada dos que possuam menor correlação com a produtividade; e as variáveis restantes deverão ser utilizadas na definição das ZMs.

5.2.3 NORMALIZAÇÃO DOS DADOS

As variáveis foram normalizadas visando estabelecer uma mesma unidade de medidas para todas elas. Seguindo o procedimento utilizado por Gavioli et al. (2016), dois métodos de normalização foram utilizados. Para as variáveis consideradas temporalmente estáveis, foi utilizado o método da amplitude (Equação 1), seguindo recomendação de Schenatto et al. (2017).

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

em que:, P_{iN} - variável normalizada pelo método da amplitude no ponto i; P_i - variável no ponto i; Mediana - mediana das amostras; Amplitude - amplitude das amostras.

Para o caso da produtividade, variável que tem variabilidade temporal devido a fatores como clima e/ou plantas invasoras, foi realizada a normalização pelo método Standard Score (Equação 2).

$$Z_i = \frac{(X_i - \overline{X})}{s} \qquad \text{Eq.(2)}$$

em que: Z_i - variável normalizada pelo método Standard Score no ponto i; X - variável no ponto i; $\overline{X}$ - média das amostras; s - desvio padrão das amostras.

5.2.4 INTERPOLAÇÃO DOS DADOS

A interpolação dos dados foi realizada sobre todas as variáveis selecionadas, a fim de possuir uma melhor representatividade dos conjuntos de dados. Para este processo foi utilizado o método da Krigagem, a fim de criar uma grade de 5 x 5 m, procurando um número mais denso de pontos por área e, portanto, definindo MZs mais suaves.

5.2.5 AGRUPAMENTO DE DADOS E MÉTRICAS DE DISTÂNCIAS

Para cada um dos oito conjuntos de dados foi aplicado o algoritmo de agrupamento fuzzy c-means associado a cada uma das três métricas de distâncias, que são objetivos de estudo do presente trabalho.

A distância Diagonal é utilizada para padronizar medições no momento em que é detectada a igualdade de variância no processo de agrupamento (MCBRATNEY et al., 1985; ODEH et al., 1992). Para isso, uma matriz diagonal é definida (CHITTLEBOROUGH; MCBRATNEY; ODEH, 1992):

$$A_D = \begin{vmatrix} (1/\sigma)^2 & 0 & \dots & 0 \\ 0 & (1/\sigma)^2 & \dots & 0 \\ \cdot & & \cdot & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ 0 & 0 & & (1/\sigma)^2 \end{vmatrix}$$

em que: A_D é a matriz diagonal e σ é o desvio padrão.

Com a matriz diagonal calculada aplica-se a Equação 3 para obtenção da distância (CHITTLEBOROUGH; MCBRATNEY; ODEH, 1992).

$$Dist\,(E_i, E_j) = \sqrt{(x_{il} - x_{jl}).\ A_D.(x_{il} - x_{jl})} \qquad \text{Eq.(3)}$$

em que: dist será a distância entre os pontos; e são os valores do atributo de cada ponto e AD é a matriz diagonal.

A distância diagonal (MCBRATNEY et al., 1985; ODEH et al., 1992), também é insensível a variáveis correlacionadas, como distância Euclidiana. No entanto, ela compensa distorções que ocorrem ao se considerar os aglomerados esféricos, por ponderar por meio das variâncias das variáveis medidas.

Zang et al. (2010) apresentaram em seu trabalho um comparativo entre as métricas de distância Diagonal e Mahalanobis, em que definiram ZMs com uma ferramenta (ZoneMAP) de suporte à decisão para aplicação de taxa variável. Uma de suas conclusões foi que a distância Diagonal apresenta um computo mais rápido que a de Mahalanobis.

Já Bezdek, Ehrlich e Full (1984) utilizaram em seu trabalho as métricas de distância Euclidiana, Diagonal e Mahalanobis a fim de análise de dados geoestatísticos.

A Distância Euclidiana (Equação 4) é definida como a raiz quadrada da soma das diferenças entre x_{il}e x_{jl} elevadas ao quadrado. Para sua aplicação recomenda-se a normalização das variáveis (normalização dos dados) antes de se obter o valor da distância Euclidiana quando os dados não se apresentam na mesma escala de medidas (LIU et al., 2014).

$$Dist\,(E_i, E_j) = \sqrt{\sum_{l=1}^{M}\left(x_{il} - x_{jl}\right)^2} \qquad \text{Eq.(4)}$$

em que: *dist* é a distância entre os pontos e x_{il}e x_{jl} são os valores da variável de cada ponto, sendo calculadas a partir do centro da célula de origem (centróide) para o centro de cada uma das células vizinhas.

A Distância Mahalanobis (Equação 5) é uma métrica que se difere da distância Euclidiana por levar em consideração a correlação entre os dados analisados; sendo assim, não depende da escala das informações, não necessitando a normalização dos dados para se obter escalas equivalentes. A fórmula analisa a distância de Mahalanobis entre dois vetores da mesma distribuição que possuam uma matriz de covariância S. Ela considera em seu desenvolvimento a raiz quadrada das diferenças entre x_{il} e x_{jl} transpostas, vezes a matriz de variância amostral elevado a menos uma vez a diferença entre x_{il} e x_{jl} (LIU et al., 2014).

$$Dist\,(E_i, E_j) = \sqrt{\left(x_{il} - x_{jl}\right)^T . S^{-1} . \left(x_{il} - x_{jl}\right)} \qquad \text{Eq.(5)}$$

em que: *dist* será a distância entre os pontos; x_{il}e x_{jl} são os valores da variável de cada ponto e S é a matriz de covariância.

Se a matriz de covariância é a matriz identidade, a distância de Mahalanobis coincide com a distância Euclidiana. Se a matriz de covariância é diagonal, então a medida de distância resultante é chamada distância Euclidiana normalizada (Equação 6).

$$Dist\,(E_i, E_j) = \sqrt{\sum_{l=1}^{M} \frac{\left(x_{il} - x_{jl}\right)^2}{\sigma^2}} \qquad \text{Eq.(6)}$$

em que: *dist* será a distância entre os pontos; x_{il}e x_{jl} são os valores da variável de cada ponto; σ é o desvio padrão.

O uso da distância de Mahalanobis corrige algumas das limitações da distância Euclidiana, pois leva em consideração automaticamente a escala dos eixos das coordenadas e também a correlação entre as características.

A detecção de outliers é um dos usos mais comuns da distância de Mahalanobis, pois um valor alto determina que um elemento está a vários desvios padrões do centro e, por consequência, é provavelmente um outlier (PENNY, 1987).

Para a definição das ZMs foi utilizado o Software SDUM (Software para definição de unidades de manejo; BAZZI et al., 2013) e o FuzME (MINASNY; MCBRATNEY, 2002).

5.2.6 AVALIAÇÃO DOS AGRUPAMENTOS E DA ZONAS DE MANEJO

Para a identificação do número ideal de agrupamentos formados através da aplicação de algoritmos foram utilizadas as técnicas que seguem (Equações 8 a 11):

- Redução de variância – RV: é calculada 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 8) (DOBERMANN et al., 2003; XIANG et al., 2007).

$$RV = 1 - \frac{\sum_{i=1}^{n} W_i * V_{um_i}}{V_{área}} * 100 \qquad \text{Eq.(8)}$$

em que: n corresponde ao tamanho da amostra para toda a área, W_i é a proporção da área em cada unidade de manejo, V_{um_i} é a variância dos dados de cada unidade de manejo e $V_{área}$ é a variância da amostra dos dados para toda a área.

- Índice de desempenho Fuzzy – FPI: 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 9) (FRIDGEN et al., 2004).

$$FPI = 1 - \frac{c}{(c-1)}\left[1 - \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^2 / n\right] \qquad \text{Eq.(9)}$$

em que: c corresponde ao número de agrupamentos, n é o tamanho da amostra para toda a área (número de observações), u_{ij} é o elemento *ij* da matriz de pertinência Fuzzy.

- Índice da partição da entropia modificada – *MPE*: 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 10) (FRIDGEN et al., 2004).

$$MPE = \frac{-\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}\log(u_{ij}) / n}{\log c} \qquad \text{Eq.(10)}$$

em que:c corresponde ao número de agrupamentos, n é o tamanho da amostra para toda a área (número de observações), u_{ij} é o elemento *ij* da matriz de pertinência Fuzzy.

- Índice de Validação de Cluster Melhorado – ICVI: O número ideal de agrupamentos de um conjunto de dados baseia-se no valor mínimo de FPI, MPE e máximo RV. Para evitar a situação em que estas estimativas apontem para diferentes agrupamentos, quando analisados de forma individual, o *ICVI* pode ser utilizado unindo os conceitos e considerando o agrupamento que apresentar o menor *ICVI* como o melhor

(Equação 11) (GAVIOLI et al., 2016)

$$ICVI_i = \frac{1}{3} * \left(\frac{FPI_i}{Max\{FPI\}} + \frac{MPE_i}{Max\{MPE\}} + \left(1 - \frac{VR_i}{Max\{RV\}}\right)\right) \quad \text{Eq.(11)}$$

em que: *i* corresponde aos índices de todas as métricas de distância utilizadas no experimento.

Para a identificação de quais foram as melhores ZMs definidas através da aplicação dos algoritmos de agrupamentos e suas respectivas distâncias utilizaram-se as seguintes técnicas:

- Análise de Variância - ANOVA:

Análise de variância é uma técnica estatística que permite avaliar afirmações sobre as médias de populações (MILONE, 2009). A análise visa, primordialmente, verificar se existe uma diferença significativa entre as médias e se os fatores exercem influência em alguma variável dependente.

Após a identificação de que as médias são estatisticamente diferentes, é necessário um método que forneça a diferença mínima significante entre duas médias. Essa diferença seria o instrumento de medida. Toda vez que o valor absoluto da diferença entre duas médias é igual ou maior do que a diferença mínima significante, as médias são consideradas estatisticamente diferentes, ao nível de significância estabelecida (VIEIRA et al., 1989). Neste trabalho, a comparação de médias foi feita pelo Teste de Tukey.

De acordo com Bazzi (2011), o teste de comparação de médias da ANOVA pode ser aplicado para verificar se as subáreas definidas realmente representam grupos diferentes a determinado nível de significância. Porém, esse teste considera que as amostras são independentes dentro de cada ZM. Por isso é necessário verificar primeiramente se em cada ZM não existe dependência espacial dos dados.

- Índice de Suavidade (Smoothness index) – *SI*: A avaliação dos melhores métodos de definição de agrupamento deve também incluir o aspecto visual do agrupamento criado e, portanto, deve-se levar em conta a suavidade das curvas de contorno, pois facilita a interpretação visual e a aplicação em taxa variada de insumos agrícolas. O índice de suavidade (*SI*) calcula a frequência da mudança de classes nos mapas temáticos nas direções horizontais, verticais e diagonais, pixel por pixel. Na hipótese de que o mapa seja uma única área totalmente homogênea, um índice de suavidade de 100% será obtido, devido à ausência de mudança de classe. Da mesma forma, se o mapa foi gerado com valores aleatórios, o índice SI apresentaria um valor próximo de zero *(*Equação 12) (GAVIOLI et al., 2016).

$$SI = 100 - \left(\left(\frac{\sum_{i=1}^{l} NM_{Hi}}{4P_H} + \frac{\sum_{i=1}^{c} NM_{Vi}}{4P_V} + \frac{\sum_{i=1}^{n} NM_{Ddi}}{4P_{Dd}} + \frac{\sum_{i=1}^{n} NM_{Dei}}{4P_{De}} \right) * 100 \right) \quad \text{Eq.(12)}$$

em que: NM_{Hi} corresponde ao número de mudanças na linha horizontal i; NM_{Vi} é o número de mudanças na linha vertical j; NM_{Ddi} é o número de mudanças na diagonal direita; NM_{Dei} é o número de mudanças na diagonal esquerda; P_H é a possibilidade de mudança na horizontal; P_V é a possibilidade de mudança na vertical; P_{Dd} é a possibilidade de mudança na diagonal direita e P_{De} é a possibilidade de mudança na diagonal esquerda.

- Índice Kappa: Para avaliar a concordância entre as ZMs definidas utilizando diferentes métricas de distâncias foi utilizado o índice Kappa *(*Equação 13) (COHEN, 1960). O Kappa avalia o nível de concordância entre duas classificações, sendo considerado que 0 < K ≤ 0,2: não há concordância; 0,2 < K ≤ 0,4: fraca; 0,4 < K ≤ 0,6: moderada; 0,6 < K ≤ 0,8: forte; 0,8 < K ≤ 1: muito forte (LANDIS; KOCH, 1977).

$$K = \frac{\left\{ n \sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} \left(x_{i+} * x_{+i} \right) \right\}}{\left\{ n^2 - \sum_{i=1}^{r} \left(x_{i+} * x_{+i} \right) \right\}} \quad \text{q.(13)}$$

em que: K - índice Kappa de concordância; n - o número total de observações (pontos amostrais); r - número de classes da matriz de erro; x_{ii} - número de combinações na diagonal; x_{i+} - total de observações na linha i; x_{+i} - total de observações na coluna i.

5.3 RESULTADOS E DISCUSSÃO

A análise descritiva da produtividade média de soja de cada safra (Tabela 3) mostrou que, para a área A, a safra 2014 foi a que apresentou a maior média (4.525) entre os anos analisados, o que justifica salientar também os menores valores de SD e CV (0,281 e 6,2%, respectivamente), indicando uma homogeneidade de produtividade. A área B apresenta para todas as suas safras uma média dispersão em sua produtividade que pode ser identificada através do CV que se encontra na faixa entre 10 e 20% para todos os anos. Para a área C, identificou-se baixa dispersão em sua produtividade com análise do CV, e sua safra de 2012 foi a que apresentou o maior ponto de produtividade, média e mediana dentre os conjuntos de dados (7.436, 5.348 e 5.277, respectivamente). Já para a área D, a safra de 2010 foi a que apresentou uma maior dispersão entre todo o conjunto de dados analisados, apresentando o menor ponto de produtividade (1.550).

Tabela 3 Análise descritiva da produtividade média de cada safra de soja e da produtividade média

Área	Safra	Media	Mediana	Maximo	Mínimo	SD	CV (%)
Dados de Produtividade Originais (amostragem de dados) (t ha^{-1})							
A	2012	3,984	4,067	5,068	2,541	0,472	11,8
	2013	3,941	4,012	6,166	2,057	0,518	13,1
	2014	4,525	4,553	5,354	3,635	0,281	6,2
B	2012	5,255	5,255	6,980	2,563	0,749	14,2
	2013	5,079	5,107	6,808	3,366	0,586	11,5
	2014	3,888	3,819	4,759	2,934	0,496	12,8
C	2012	5,348	5,277	7,436	4,029	0,764	6,9
	2013	4,675	4,711	5,817	3,357	0,518	9,0
	2014	4,505	4,557	5,610	2,579	0,584	7,7
D	2010	2,638	2,565	4,340	1,550	0,606	23,0
	2011	3,243	3,263	4,644	2,300	0,484	14,9

CV - coeficiente de variação; SD - desvio padrão.

A variável elevação foi selecionada para definição das ZMs para três áreas (A, C e D) (Tabela 2). Já a variável SRP 0.0 - 0.1 m foi escolhida em conjunto com a elevação para a área A e exclusivo para área B. A Figura 3 apresenta os mapas temáticos das variáveis produtividade normalizada de soja, elevação e resistência à penetração no solo (SRP), com duas, três e quatro classificações. Para todas as áreas, os mapas temáticos que correspondem ao atributo de elevação são os que apresentam uma melhor separação entre as classes mesmo quando o número de agrupamentos é aumentado.

Tabela 4 Seleção de variáveis para o processo de definição das ZMs

Variáveis	**Área A**			**Área B**			**Área C**			**Área D**
	2012	**2013**	**2014**	**2012**	**2013**	**2014**	**2012**	**2013**	**2014**	**2010**
SRP 0.0 - 0.1 m (MPa)		X	X	X	X	X		X	X	X
SRP 0.1 - 0.2 m (MPa)		X	X	X	X	X		X	X	X
SRP 0.2 - 0.3 m (MPa)		X	X	X	X	X		X	X	X
Elevação (m)	X			X			X			X
Declividade (˚)	X			X						X
Densidade (g cm^{-3})	X			X						X
Argila (%)	X			X			X			X
Silte (%)	X			X			X			X
Areia (%)	X			X			X			X
Matéria Orgânica (%)	X			X			X			

[ ] - Eliminado por não ter autocorrelação espacial; [ ] - Eliminado por não ter correlação espacial com a produtividade; [ ] - Eliminado por ser redundante; [ ] - Selecionado para gerar as MZs.

Após a seleção das variáveis, iniciou-se o processo de definição das ZMs e criação dos mapas temáticos para os agrupamentos com 2, 3 e 4 clusters para cada uma das métricas de distâncias avaliadas, utilizando-se dados originais e normalizados (pelo método da amplitude) obtidos em quatros áreas.

ÁREA A (Figura 4): As ZMs definidas com dados sem (originais) e com normalização foram idênticas quando se utilizaram as métricas Diagonal e Mahalanobis. Já quando se utilizou a

métrica euclidiana as ZMs foram diferentes e a influência da normalização foi maior com o aumento do número de ZMs.

ÁREAS B, C e D (Figura 5): As ZMs definidas com dados sem (originais) e com normalização foram idênticas para as três métricas utilizadas (Diagonal, Euclidiana e Mahalanobis). Este fato, em contraste ao observado na área A, se deve à utilização de somente uma variável na definição das ZMs das áreas, enquanto na área A foram utilizadas duas. Em outras palavras, quando se tem somente uma variável, o seu desvio padrão se torna irrelevante na definição das ZMs. Por isso, para cada número de ZMs, as ZMs foram idênticas. Ressalte-se que quando se tem mais de uma variável na definição de ZM, mesmo que os dados sejam normalizados, as ZMs de manejo para cada métrica são semelhantes, mas não iguais.

Outro aspecto a ser ressaltado é que com a utilização da variável de resistência à penetração no solo (SRP 0-0.1 m (2013)), na área B, não foi possível obter ZMs contínuas com nenhuma das métricas, sendo que o fracionamento das ZMs aumentou conforme o acréscimo do número de agrupamentos ocorreu. Este comportamento segue o padrão da SRP que pode ser visto na Figura 3.

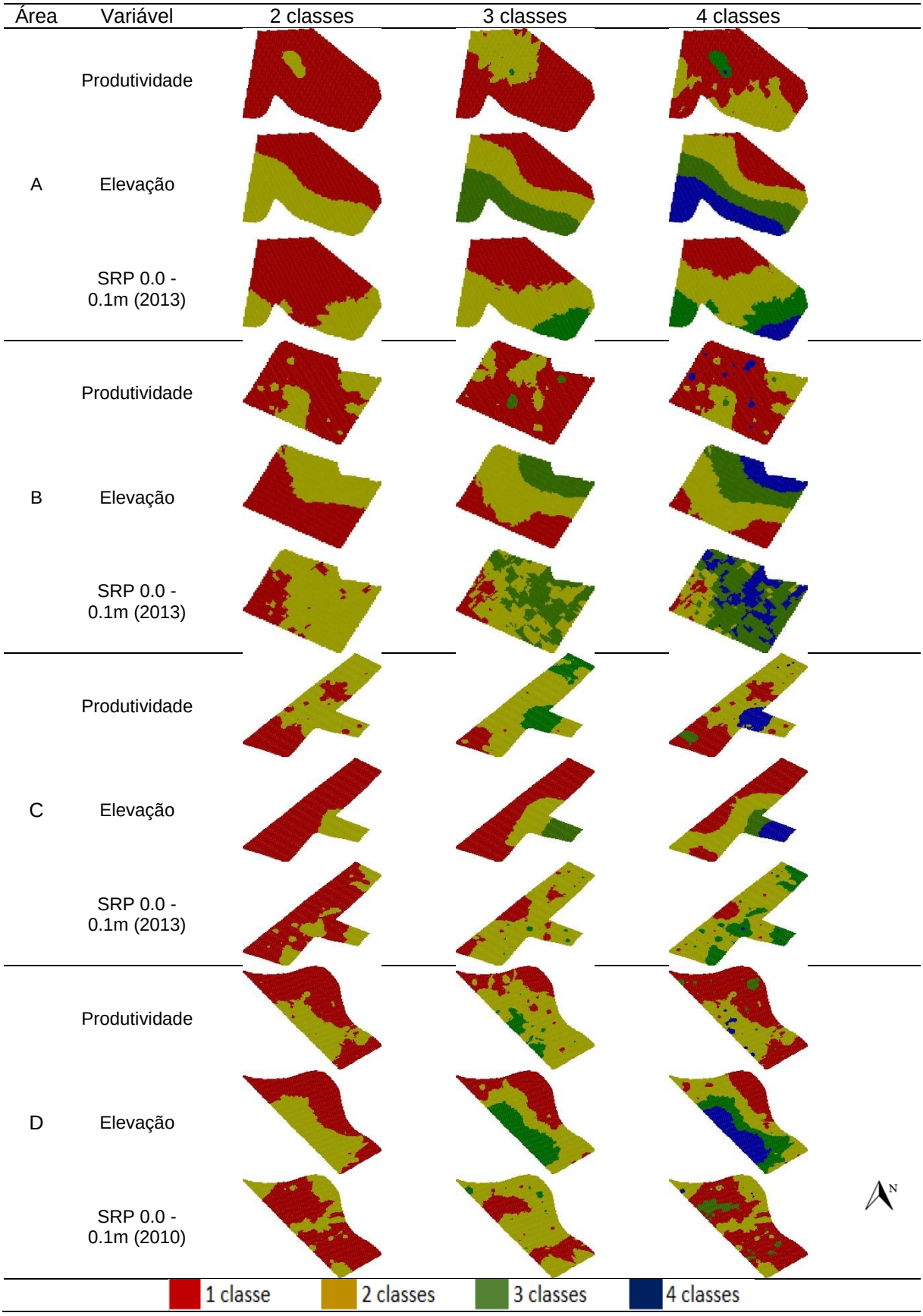


Figura 3 Mapas temáticos das variáveis produtividade normalizada de soja, elevação e resistência à penetração no solo (SRP), com duas, três e quatro classificações.

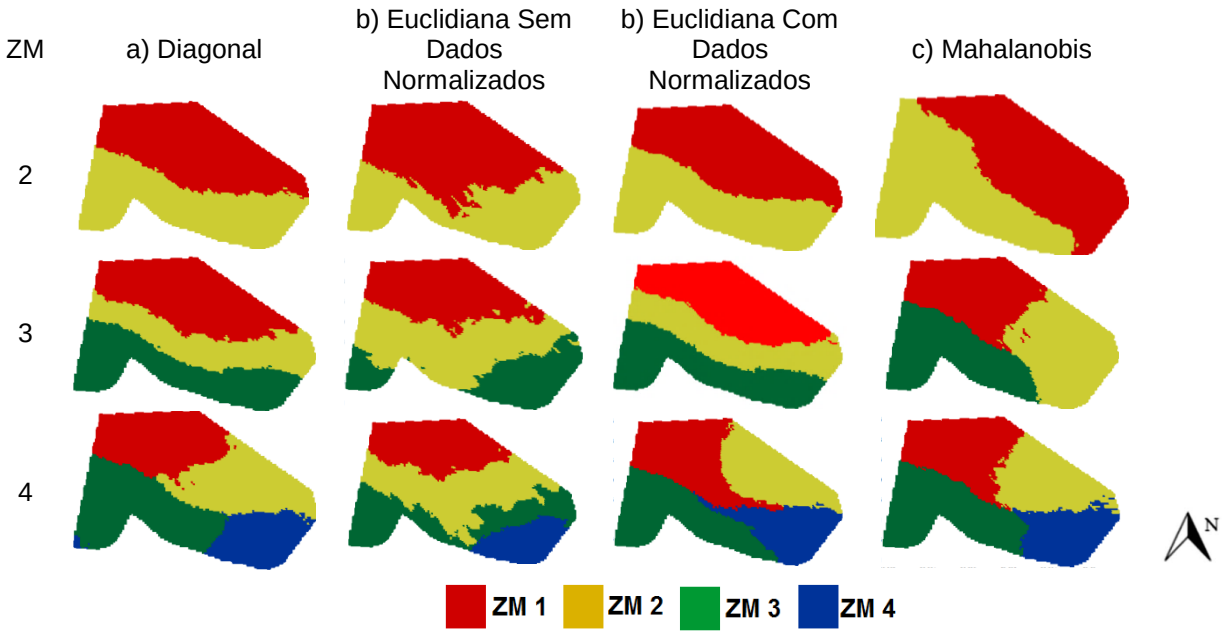


Figura 4 Zonas de manejo para a área A, definidas com as variáveis elevação e RSP 0-0,1m (2013), considerando as métricas de distância Diagonal (a); Euclidiana sem dados Normalizados (b), Euclidiana com dados normalizados (c) e Mahalanobis (d).

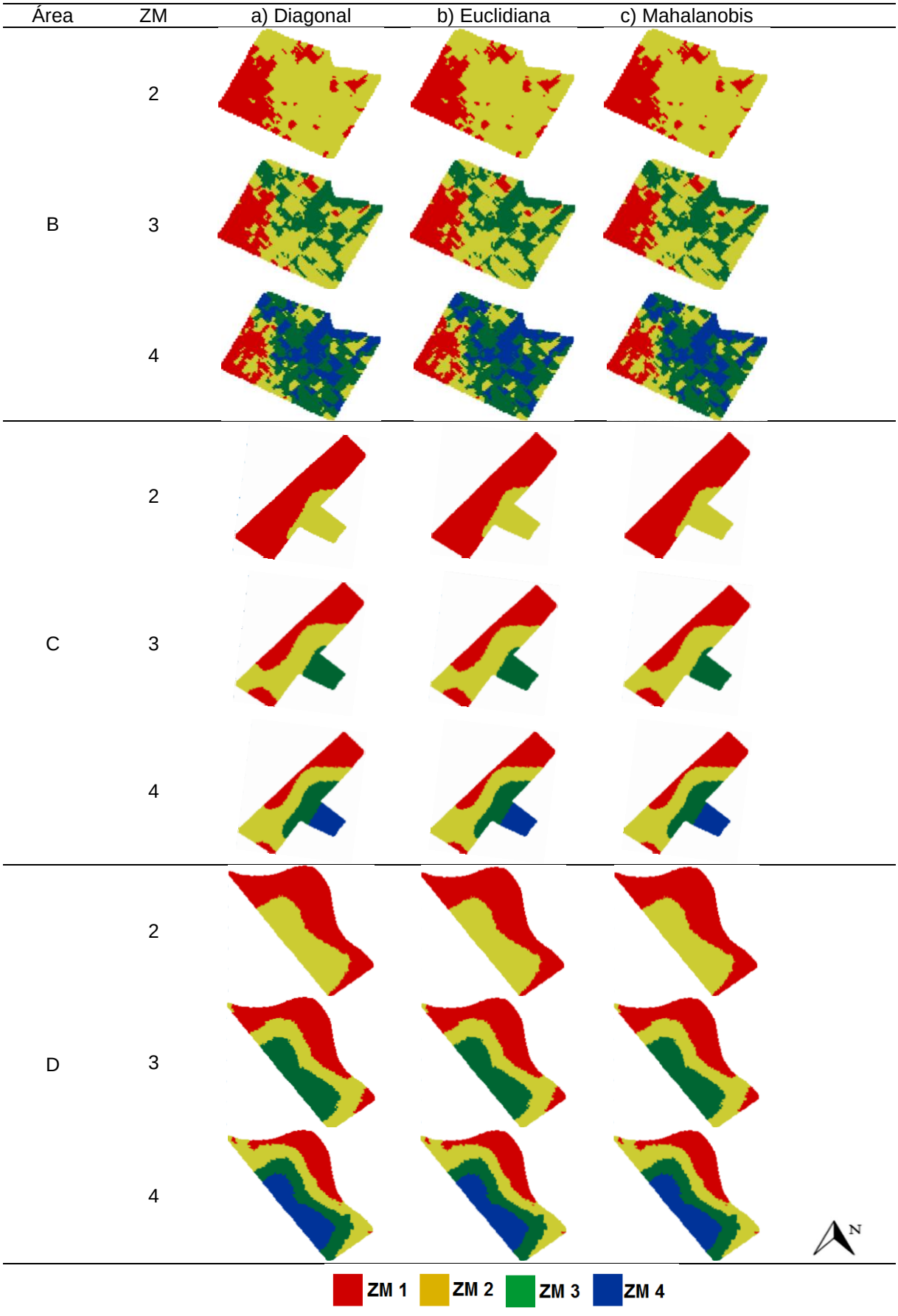


Figura 5 Zonas de manejo (ZMs) para a área B, C e D, definidas com as variáveis RSP 0-0,1 m (2013)((área B) elevação (áreas C e D), considerando as métricas de distância Diagonal (a); Euclidiana (b) e Mahalanobis (c).

Os valores dos índices RV, FPI, MPE, ICVI e SI, além do teste Tukey das médias de produtividade, estão apresentados na Tabela 5. São apresentados os resultados obtidos pelas três métricas de distâncias, sendo descriminados os dados normalizados dos não normalizados para a métrica Euclidiana, em que a normalização dos dados interferiu na definição das zonas de manejo.

Tabela 5 Índices de avaliação das zonas de manejo definidas utilizando diferentes métricas de distância – Dados originais

Área	Nº Zonas	Dados normalizados	Métrica de Distância	ANOVA e Teste de Tukey*				RV%	FPI	MPE	ICVI	SI%
				ZM1	ZM2	ZM3	ZM4					
A	2	£	Diagonal	a	b			7,7	0,124	0,149	0,720	98,2
		Não	Euclidiana	a	a			-6,0	0,104	0,123	0,842	97,2
		Sim	Euclidiana	a	b			7,8	0,117	0,141	0,691	98,4
		£	Mahalanobis	a	a			-2,3	0,171	0,190	1,035	98,3
	3	£	Diagonal	a	a	b		4,0	0,156	0,167	0,870	96,7
		Não	Euclidiana	a	a	b		7,9	0,099	0,108	0,597	96,3
		Sim	Euclidiana	a	a	b		22,1	0,152	0,163	0,582	96,9
		£	Mahalanobis	a	a	b		11,7	0,148	0,162	0,730	97,7
	4	£	Diagonal	a	b	c	ab	13,6	0,150	0,154	0,691	96,9
		Não	Euclidiana	a	abc	bc	c	-22,8	0,095	0,099	1,036	93,4
		Sim	Euclidiana	a	a	a	b	11,8	0,128	0,138	0,647	96,6
		£	Mahalanobis	a	a	a	b	11,8	0,128	0,138	0,647	96,6
B	2	£	Diagonal	a	a			-2,6	0,089	0,109	1,114	92,3
		£	Euclidiana	a	a			-2,6	0,089	0,109	1,114	92,3
		£	Mahalanobis	a	a			-2,6	0,089	0,109	1,114	92,3
	3	£	Diagonal	a	ab	ab		6,4	0,095	0,104	0,651	83,6
		£	Euclidiana	a	ab	ab		6,4	0,095	0,104	0,651	83,6
		£	Mahalanobis	a	ab	ab		6,4	0,095	0,104	0,651	83,6
	4	£	Diagonal	a	ab	abc	abc	-6,3	0,090	0,093	1,262	78,6
		£	Euclidiana	a	ab	abc	abc	-6,3	0,090	0,093	1,262	78,6
		£	Mahalanobis	a	ab	abc	abc	-6,3	0,090	0,093	1,262	78,6
C	2	£	Diagonal	a	b			12,5	0,072	0,090	0,632	98,9
		£	Euclidiana	a	b			12,5	0,072	0,090	0,632	98,9
		£	Mahalanobis	a	b			12,5	0,072	0,090	0,632	98,9
	3	£	Diagonal	a	ab	ab		-8,9	0,095	0,102	1,177	97,5
		£	Euclidiana	a	ab	ab		-8,9	0,095	0,102	1,177	97,5
		£	Mahalanobis	a	ab	ab		-8,9	0,095	0,102	1,177	97,5
	4	£	Diagonal	a	ab	c	abc	16,8	0,089	0,091	0,610	96,8
		£	Euclidiana	a	ab	c	abc	16,8	0,089	0,091	0,610	96,8
		£	Mahalanobis	a	ab	c	abc	16,8	0,089	0,091	0,610	96,8
D	2	£	Diagonal	a	b			18,8	0,086	0,104	0,741	98,0
		£	Euclidiana	a	b			18,8	0,086	0,104	0,741	98,0
		£	Mahalanobis	a	b			18,8	0,086	0,104	0,741	98,0
	3	£	Diagonal	a	b	a		28,3	0,088	0,098	0,617	96,5
		£	Euclidiana	a	b	a		28,3	0,086	0,104	0,629	96,5
		£	Mahalanobis	a	b	a		28,3	0,086	0,104	0,629	96,5
	4	£	Diagonal	a	ab	ab	c	23,9	0,097	0,100	0,706	94,4
		£	Euclidiana	a	ab	ab	c	23,9	0,086	0,104	0,681	94,4
		£	Mahalanobis	a	ab	ab	c	23,9	0,086	0,104	0,681	94,4

* Significativo ao nível de 0.05. £ Mesmo resultado para dados normalizados ou não. Casos sombreados em cinza são significativamente diferentes.

Considerando que somente é interessante dividir a área total em ZMs que possuam variável-alvo (neste caso a produtividade) estatisticamente distintas, a primeira análise a ser feita é o Teste Tukey de médias. Como resultado, tem-se que somente é aconselhável se dividir em duas ZMs (casos sombreados em cinza) e que a área B não deve ser subdividida. Com relação aos índices, tanto melhor quanto maior for RV e SI e menor for FPI, MPE e

ICVI, o que resulta na área A em vantagem da distância Euclidiana utilizando dados normalizados sobre as demais. Ressalte-se que, como já era esperado, os resultados para a métrica Euclidiana utilizando dados normalizados sempre foi superior aos obtidos com dados normalizados, pois a normalização dos dados resolve a deficiência desta métrica não observar o desvio padrão de cada variável. Lamentavelmente, para as áreas B, C e D todas as métricas conduziram às mesmas divisões, portanto resultando em índices iguais.

Considerando que o índice ICVI corresponde a uma composição dos índices RV, FPI e MPE, pode-se restringir a análise dos índices somente analisando os índices ICVI e SI (Tabela 5 e Figura 6). O SI, que caracteriza a suavidade das curvas de contorno das ZMs (facilita a interpretação visual e a aplicação em taxa variada de insumos agrícolas), praticamente manteve-se constante dentro de cada número de ZMs, tendendo a diminuir suavemente com o aumento do número de ZMs, mas com exceção para a área B, onde o decréscimo foi muito mais forte.

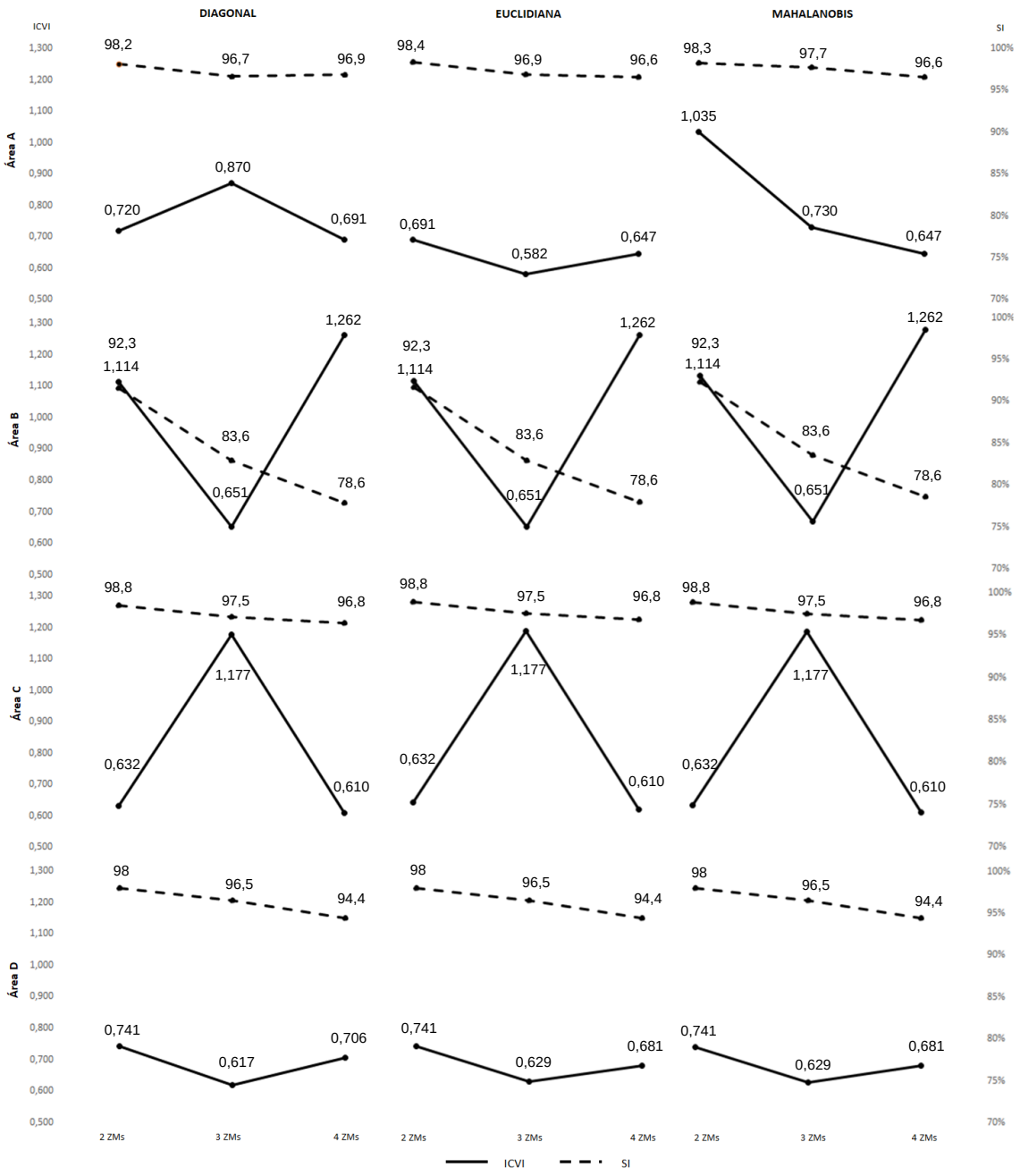


Figura 6 Índice de Validação de Cluster Melhorado (ICVI) e Índice de Suavidade (SI) obtidos com a utilização dos dados normalizados, nos agrupamentos para as áreas A, B, C e D.

O ICVI, que é tanto melhor quanto menor, teve o seguinte comportamento em função da área:

ÁREA A: O menor ICVI correspondeu à distância Euclidiana (dados normalizados) com três ZMs. Infelizmente, para esta divisão das ZMs não foram significativamente distintas.

ÁREAS B, C e D: Sempre a área B em discordância com as outras duas áreas. Menores ICVIs: três ZMs para áreas B e D; quatro ZM para área C.

Para definição do Índice Kappa, foi realizada uma comparação entre pares de distâncias: Diagonal X Euclidiana, Diagonal X Mahalanobis, Euclidiana X Mahalanobis. Para

cada par analisado foram considerados os agrupamentos com 2, 3 e 4 ZMs, conforme mostra a Figura 7.

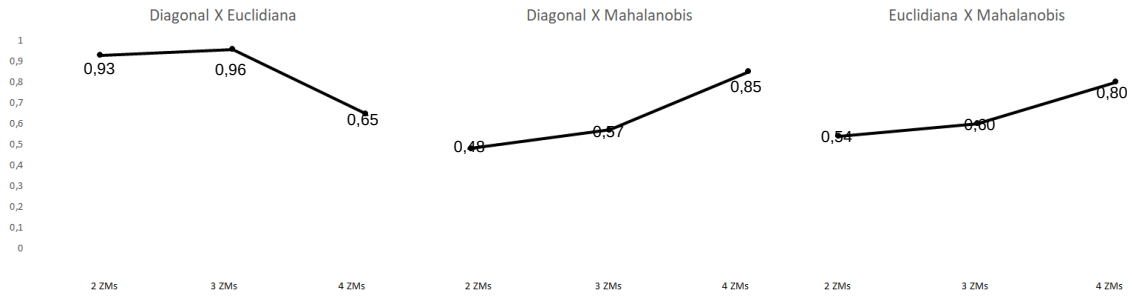


Figura 7 Índice Kappa entre métricas de distância (Euclidiana, Diagonal e Mahalanobis) em função do número de zonas de manejo (ZMs).

O Índice Kappa foi calculado apenas para a área A, visto que as demais áreas estudadas não apresentaram diferenças entre seus mapas temáticos. Através da sua análise pode observar-se que existe em todas as situações uma concordância entre as distâncias analisadas, sendo que: a Diagonal X Euclidiana variou de 0,65 a 0,96, ou seja, de forte a muito forte; a combinação Diagonal X Mahalanobis variou de 0,48 a 0,85, ou seja, de moderada a muito forte; e a concordância entre as distâncias Euclidiana X Mahalanobis apresentou uma variação de 0,54 a 0,80, ou seja, de moderada a forte.

5.4 CONCLUSÕES

Somente foi aconselhável se dividir em duas zonas de manejo (ZMs) as áreas A, C e D. A área B não deve ser subdividida.

Como somente a área A foi definida com mais de uma variável de entrada no processo de clusterização, somente nela pode-se avaliar a eficiência das métricas de distância (Diagonal, Euclidiana e Mahalanobis). Neste caso, o melhor desempenho foi para distância Euclidiana utilizando dados normalizados.

As métricas Diagonal e Mahalanobis produziram as mesmas ZMs, independente da normalização dos dados.

5.5 REFERÊNCIAS

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

BAZZI, C. L. **Software para definição e avaliação de unidades de manejo em agricultura de precisão**. 2011. 123f. Tese (Doutorado em Engenharia Agrícola). Programa de Pós-Graduação em Engenharia Agrícola. Universidade Estadual do Oeste do Paraná. Cascavel, 2011.

BEZDEK, J.; EHRLICH, R.; FULL, W. FCM: The Fuzzy c-Means Clustering Algorithm. **Computers & Geosciences,** v. 10, n. 2-3, p. 191-203, 1984.

BUNSELMEYER, H. A.; LAUER, J. G. Using Corn and Soybean Yield History to Predict Subfield Yield Response. **Agronomy Journal,** v.107, p. 558-562, 2015.

ODEH, I.O.A.; MCBRATNEY, A.B.; CHITTLEBOROUGH, D.J. Soil pattern recognition with fuzzy-c-means: application to classification and soil –landform interrelationships. **Soil Science Society of America Journal**, n. 56, p. 505-516, 1992

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

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

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. 5, p. 1105-1120, 2003.

DOERGE, T. A. **Management Zone Concepts**. Site-Specific Management Guidelines. Potash & Phosphate Institute (PPI), South Dakota State University (SDSU), 2000.

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 Engineers,** v. 44, n. 1, p. 155-166, 2001.

FRIDGEN, J. J.; KITCHEN, N. R.; SUDDUTH, K. A. Variability of soil and landscape attributes within sub-field management zones. In: International Conference on Precision Agriculture. **Anais…** Bloomington: Madison, 2000.

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.

JAYNES, E. T. Probability Theory. **The logic of Science.** New York: Cambridge University Press, 2003.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining Geostatistics.** London: Academic Press, 1978.

LIU, Q.; CHU, X.; XIAO, J.; ZHU, H. Optimizing Non-orthogonal Space Distance Using PSO in Software Cost Estimation. In: IEEE Computer Software Applications Conference (COMPSAC). **Anais...,** 2014.

MANLY, B. F. J. **Multivariate statistical methods:** a primer. London: Chapman and Hall, 1986.

METZ, J.; MONARD, M. C. **Projeto e implementação do módulo de clustering hierárquico do discover**. ICMC-USP, 2006.

MILONE, G. **Estatística geral e aplicada.** São Paulo: Centage Learning, 2009.

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

MOLIN, J. P. Agricultura de Precisão. Parte 1: O que é estado-da-arte em sensoriamento. **Engenharia Agrícola**, v.17, p. 97- 107, 1997.

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.

MOLIN, J. P.; AMARAL, L. R.; COLAÇO, A. **Agricultura de Precisão.** São Paulo: Oficina de Textos, 2015.

PENNY, K. I. Appropriate Critical Values when Testing for a Single Multivariate Outlier by using the Mahalanobis Distance. In: **Applied Statistics**. Royal Statistical Society, UK, 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.

SCHENATTO, K.; SOUZA, E. G.; BAZZI, C. L.; GAVIOLI, A.; BETZEK, N. M.; BENEDUZZI, H. M. Normalization of data for delineating management zones. **Computers and Electronics in Agriculture**, v. 143, p. 238-248, 2017.

SEIDL, E. J.; MOREIRA JÚNIOR, F. de J.; ANSUJ, A. P.; NOAL, M. R. C. Comparação entre o método ward e o método k-médias no agrupamento de produtores de leite. **Ciência e Natura,** v. 30, n. 1, p. 7-15, 2008.

VICINI, L. **Analise multivariada da teoria à prática.** Universidade Federal de Santa Maria. Santa Maria, UFSM: CCNE, 2005.

VIEIRA, S.; HOFFMANN, R. **Estatística experimental.** São Paulo: Atlas, 1989. 175p.

XIANG, L.; YU-CHUN, P.; ZHONG-QIANG, G.; CHIN-JIANG, Z. 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.

YANG, Z.; SHUFAN, Y.; YANG, X.; LIQUN, G. High-Dimensional Statistical Distance for Object Tracking. International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). **Anais…** 2010.

ZHANG, X.; SHI, L.; JIA, X.; SEIELSTAD, G.; HELGASON, C. Zone mapping application for precision-farming: a decision support tool for variable rate application. **Precision Agriculture**, v. 11, n.2, p. 103-114, 2010.

6 ARTIGO 2 – SELEÇÃO DE VARIÁVEIS PARA DEFINIÇÃO DE ZONAS DE MANEJO COM PRODUTIVIDADES DE SOJA E MILHO

Resumo

A utilização da tecnologia na agricultura está cada vez mais pertinente devido à necessidade crescente do aumento de produção e lucratividade por meio de tratamentos específicos para determinadas áreas agrícolas, chamadas de zonas de manejo (ZMs). A busca da melhor produtividade é um fato crescente dentre os produtores rurais, que utilizam em uma mesma propriedade, de forma intercalada, duas ou mais culturas. Nesse contexto, o objetivo deste trabalho foi avaliar o uso de produtividades de soja e milho (bem como a sua associação) na seleção de variáveis para definição de ZMs, assim como as ZMs resultantes. Foram utilizados dados experimentais de três áreas agrícolas obtidos entre os anos de 2012 e 2015, localizadas no estado do Paraná. A partir das variáveis disponíveis para cada uma das áreas foi realizada a seleção através do método da correlação espacial levando em consideração, para cada uma das áreas, três produtividades-alvo (soja, milho e soja+milho). Para a definição das ZMs foi utilizado o algoritmo fuzzy c-means, associado à métrica de distância euclidiana. Concluiu-se que a melhor produtividade-alvo foi soja+milho, reforçando a ideia de ser útil a utilização destas duas culturas na definição das ZMs de uma área com alternância de produção de soja e milho.

Palavras-chave: Agricultura de Precisão, Clusterização, Milho, Mineração de Dados, Seleção de Variáveis, Soja.

SELECTION OF VARIABLES FOR THE DEFINITION OF MANAGEMENT ZONES WITH SOYBEANS AND CORN YIELD

Abstract

The use of technology in agriculture is increasingly relevant due to the growing need for higher yield and profitability through specific treatments for determined agricultural areas, called management zones (MZs). The search for better yield is an increasing fact among rural producers, who grow in the same property, in rotation, two or more crops. In this context, the objective of this research was to evaluate the use of soy and corn yield (as well as their association) in the selection of variables to define MZs, as well as the resulting MZs. Experimental data from three agricultural areas gathered between the years of 2012 and 2015, located in the State of Paraná, were used. From the available variables for each of the areas, the selection was performed using the spatial correlation method, considering, for each of the areas, three target yields (soybean, corn, and soybean+corn). For the definition of the MZs, the fuzzy c-means algorithm was used, associated with the Euclidean distance metric. It was concluded that the best target yield was soybean+corn, reasserting the idea of being better to use these two cultures in the definition of MZs of an area with rotating crops of soybean and corn.

Keywords: Precision Agriculture, Clustering, Corn, Data Mining, Soybean, Variable Selection.

6.1 INTRODUÇÃO

A associação de tecnologia e agricultura é cada vez mais pertinente devido à necessidade do aumento da produtividade e lucratividade, tendo como consequência a

diminuição do uso de defensivos e o impacto ambiental nos mais variados ramos rurais (MOLIN, 2015). Essa associação é o que rege a Agricultura de Precisão (AP).

A produtividade agrícola depende criticamente do clima e sua variabilidade, sendo de grande importância durante o ciclo de vida de determinadas culturas, e responsável pela alternância das produções anuais destas (FERREIRA, 2005).

O regime de precipitação é a principal característica climática que determina a duração da estação de crescimento das plantas em regiões tropicais, em contraste com as regiões temperadas, nas quais o início e o fim da estação de crescimento são definidos pelo regime sazonal da temperatura do ar. A época de plantio de uma cultura está diretamente condicionada ao regime de chuvas de uma determinada região e a fertilidade do solo explorado (OLIVEIRA et al., 2000).

A AP é um paradigma na gestão de atividades agrícolas no qual subáreas de produção não são tratadas como equivalentes, proporcionando a administração por meio de zonas de manejo (ZMs) que, com características distintas, são guiadas por georreferenciamento (MOLIN, 1997).

A definição de ZMs pode ocorrer de diversas formas. Johannsen et al. (2000) apresentam uma abordagem com uso de sensoriamento remoto a fim de obter índices de vegetação e associá-los a grades de amostragem de solo. Outras abordagens consideram a sensibilidade do produtor por meio do conhecimento empírico, embora o método mais difundido na literatura consista em agrupar parâmetros químicos e físicos de solo coletados em pontos estratégicos georreferenciados (FRAISSE et al., 2001; MOLIN; FAULIN, 2013). Usualmente, esta definição de ZMs utiliza o algoritmo fuzzy c-means (FRIDGEN et al., 2004).

Para gerar os clusters, os algoritmos utilizam medidas de similaridade, que guiam o processo de decisão que determinará a distribuição dos dados nos respectivos grupos. Há diversas medidas para o cálculo de similaridade, dentre as quais estão de distância, de correlação e de associação. Quando o conjunto de dados é composto por variáveis quantitativas, as métricas de distância podem ser aplicadas para o cálculo da similaridade entre os dados (METZ; MORNARD, 2006).O software SDUM (Software para a definição de unidades de manejo), que possui interface trilíngue (português, espanhol e inglês), com download gratuito na internet (disponível em: <https://ftp.unioeste.br/SDUM/>), apresenta em sua estrutura a possibilidade de definição de ZMs, com aplicação do algoritmo fuzzy c-means associado à métrica de distância Euclidiana.

De acordo com Bunselmeyer e Lauer (2015), a idéia é usar o número máximo de safras e seus respectivos dados de produtividade. Além disso, Jaynes (2003) opina que a soja e o milho apresentam comportamento semelhante no campo em relação ao seu potencial produtivo.

O presente trabalho teve como objetivo avaliar se a definição de ZMs a partir de produtividades de soja e milho deve ser feita separadamente para cada cultura ou com os dados agrupados.

6.2 MATERIAL E MÉTODOS

Um fluxograma (Figura 1) foi criado para apresentar as etapas seguidas durante a definição e a avaliação das ZMs.

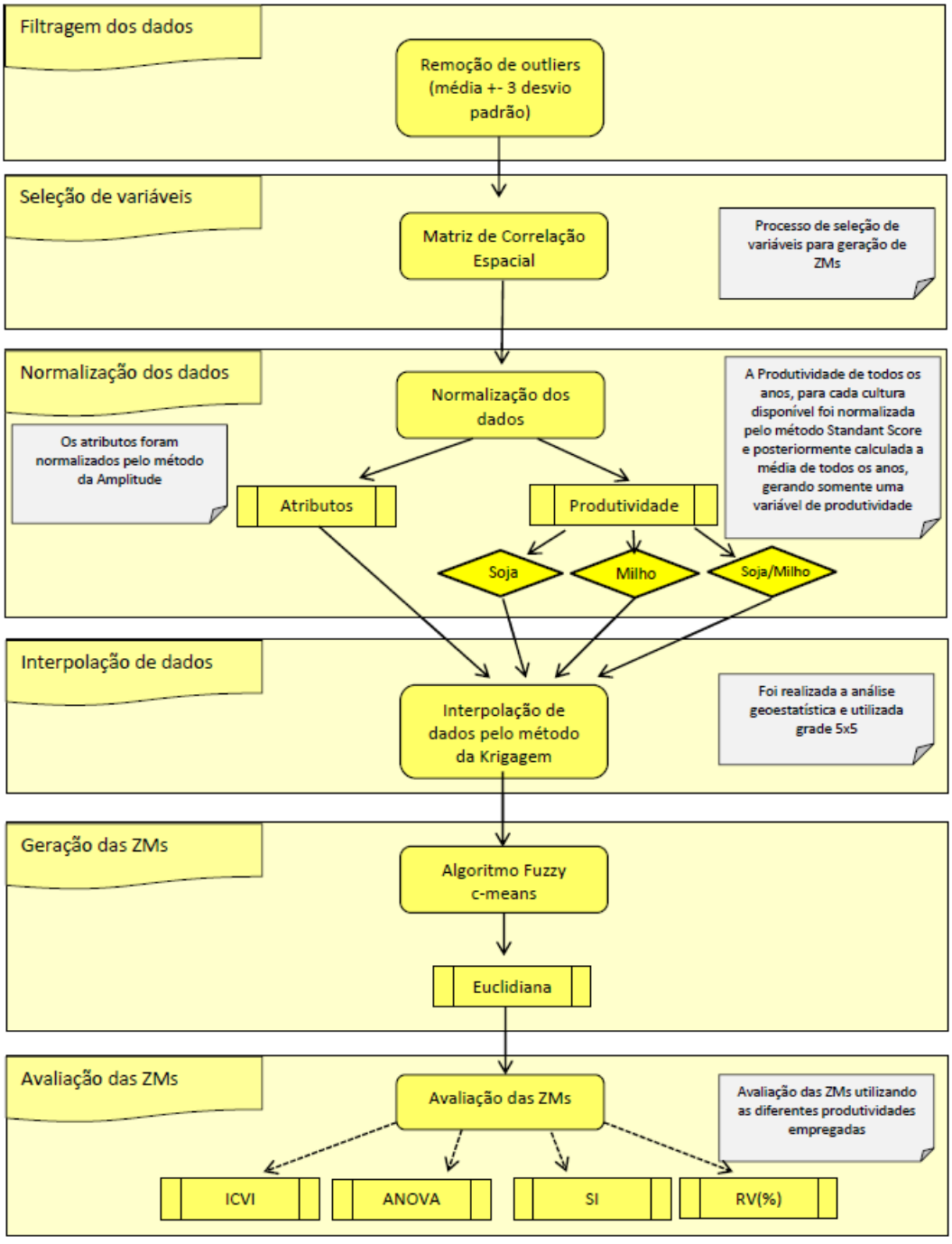


Figura 1 Fluxograma de Análise de Processo.

5.2.1 CONJUNTO DE DADOS

O conjunto de dados reais que foi utilizado neste artigo pertencem a três áreas agrícolas no estado do Paraná (Figura 2), sendo que o tamanho e a localização das áreas, bem como o número de amostras realizadas estão apresentados na Tabela 1.

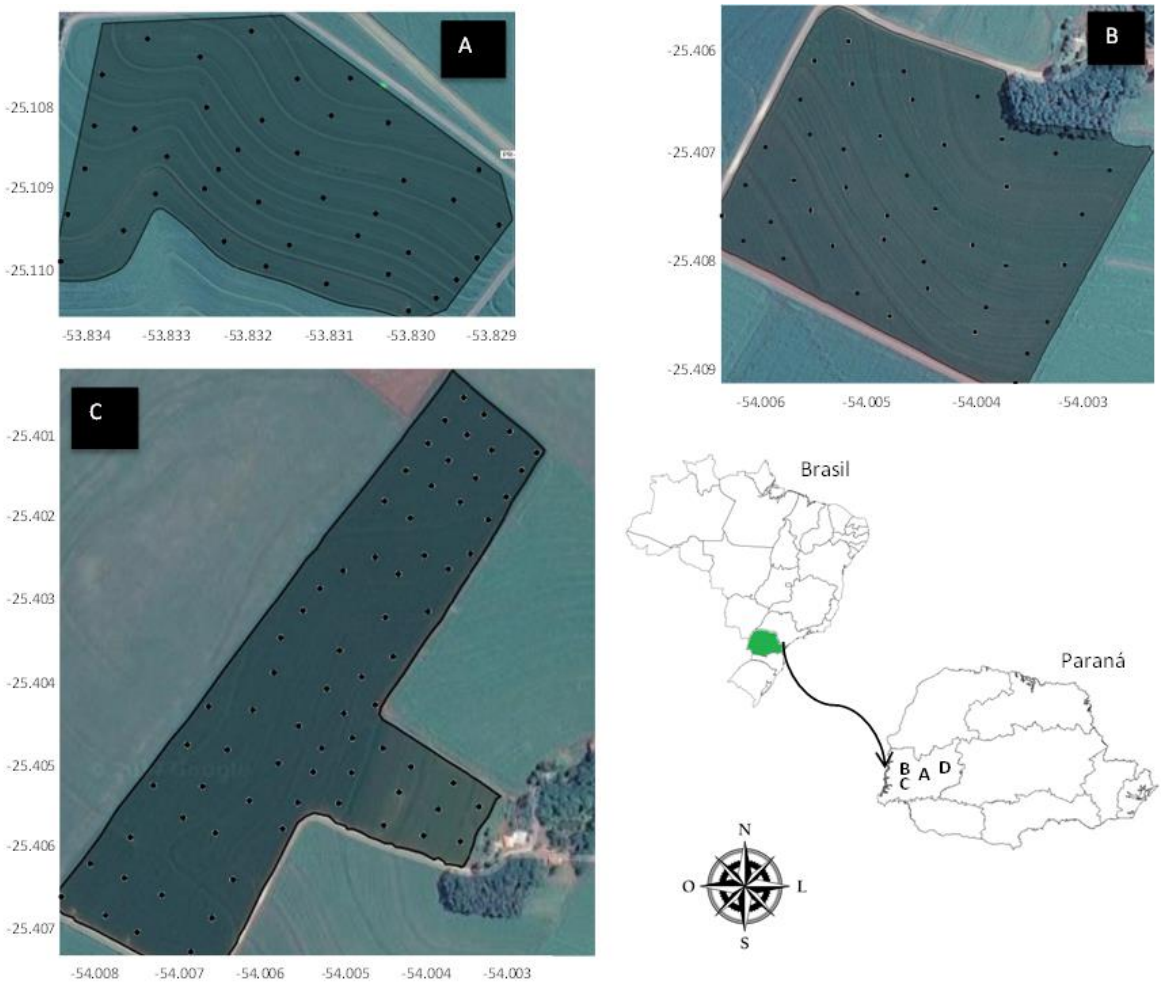


Figura 2 Localização das áreas experimentais na Região Oeste do Paraná.

Tabela 1 Identificação, tamanho, localização e altitude média das áreas de estudo

ID	Cidade	Área (ha)	Coordenadas (SAD 69 - WGS 1984)	Elevação (m)	Números de Pontos	Pontos por hectare
A	Céu Azul	15,0	25º06'32" S e 53º49'55" O	460	40	2,7
B	Serranópolis do Iguaçu	9,9	25º24'28" S e 54º00'17" O	355	42	4,2
C	Serranópolis do Iguaçu	23,8	25º24'28" S e 54º00'17" O	355	73	3,1

A primeira etapa executada foi a “limpeza” dos dados, em busca de reduzir discrepâncias de ruídos e corrigir inconsistências. Nessa fase os dados foram modificados de acordo com formatos apropriados à definição de ZMs. Na fase de filtragem/processamento dos dados foram considerados como *outliers* os valores superiores

ou inferiores a três vezes o desvio padrão, tendo como referência a média da variável (CORDOBA et al., 2016). Quando identificados, estes foram removidos do conjunto de dados.

Foram utilizadas as culturas de soja e milho, sendo obtida a média das produtividades por área, baseada nos anos indicados na Tabela 2, e, logo após, realizada a normalização dos dados pelo método standard score, por se tratarem de variáveis temporais. Para a união de soja+milho foi obtida a média de produtividade baseada em todos os anos das respectivas culturas, sendo posteriormente normalizada.

Para a definição dos agrupamentos foram utilizados somente variáveis consideradas estáveis, excluindo, portanto, os atributos químicos do solo, satisfazendo recomendação geral de literatura (DOERGE, 2000). Grades de amostragem densas foram utilizadas a fim de observar as restrições de análise geoestatística (JOURNEL; HUIJBREGTS, 1978), com pelo menos 2,5 pontos ha^{-1}. As grades são irregulares e foram definidas considerando a linha imaginária central, entre as curvas de nível de cada área. A Tabela 2 apresenta os atributos que foram coletados e usados como variáveis (atributos) candidatas a serem utilizadas na definição dos agrupamentos: resistência mecânica do solo à penetração (0 - 10 cm, SRP 0.0 - 0.1 m; 10 - 20 cm, SRP 0.1 - 0.2 m; e 20 - 30 cm, SRP 0.2 - 0.3 m); elevação (%); declividade(°); densidade (g cm^{-3}); argila (%); silte (%); areia (%); matéria orgânica (%); produtividade de soja e milho (t ha^{-1}). Os dados de produtividade da cultura para a zona A foram determinados utilizando 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 de uma área de amostragem de 1 m^2 em cada um dos pontos de amostragem. A produtividade foi então corrigida para o conteúdo de água de 13%.

Tabela 2 Variáveis (atributos) coletados em função do ano agrícola e área experimental

Atributos	Área A					Área B					Área C					Área D	
	2012	2013	2014	2015	2016	2012	2013	2014	2015	2016	2012	2013	2014	2015	2016	2010	2011
SRP 0.0 - 0.1 m (MPa)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
SRP 0.1 - 0.2 m (MPa)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
SRP 0.2 - 0.3 m (MPa)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
Elevação (m)	X					X					X					X	
Declividade (°)	X					X					X					X	
Densidade (g cm^{-3})	X					X					X					X	
Argila (%)	X					X					X					X	
Silte (%)	X					X					X					X	
Areia (%)	X					X					X					X	
Matéria Orgânica (%)	X					X					X					X	
Produtividade da soja (t ha^{-1})	X	X	X			X	X	X			X	X	X			X	X

*SPR – resistência mecânica à penetração

As três áreas estudadas fornecem três conjuntos de dados, sendo que em uma versão foi utilizada a cultura de soja; em uma segunda, a de milho; e a terceira com a junção de ambas as culturas. Com o experimento envolvendo as três versões foi possível identificar sobre quais aspectos a cultura influencia ou não na definição das ZM.

5.2.2 SELEÇÃO DE VARIÁVEIS

O processo de seleção de variáveis busca encontrar os atributos ideais para a definição das ZMs. A técnica utilizada neste trabalho foi a Correlação Espacial Cruzada, que utiliza-se de uma matriz de correlação espacial como objeto de análise e, na sequência, aplicam-se os procedimentos propostos por Bazzi et al. (2013) e Schenatto et al. (2016), em que: eliminam-se as variáveis com autocorrelação espacial não significativas a 95% de confiança (ou 5% de significância); removem-se as variáveis que não possuam correlação com a produtividade; ordenam-se de modo decrescente as variáveis restantes, considerando o módulo do grau de correlação cruzada com a produtividade; eliminam-se as variáveis redundantes (que se correlacionem entre si), dando preferência para a retirada dos que possuam menor correlação com a produtividade; e as variáveis restantes deverão ser utilizados na definição das ZMs.

5.2.3 NORMALIZAÇÃO DOS DADOS

As variáveis foram normalizadas visando estabelecer uma mesma unidade de medidas para todas elas. Seguindo o procedimento utilizado por Gavioli et al. (2016), dois métodos de normalização foram utilizados. Para as variáveis consideradas temporalmente estáveis, foi utilizado o método da amplitude (Equação 1).

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

em que:, P_{iN} - variável normalizada pelo método da amplitude no ponto i; P_i - variável no ponto i; Mediana - mediana das amostras; Amplitude -amplitude das amostras.

Para o caso da produtividade, variável que tem variabilidade temporal devido a fatores como clima e/ou plantas invasoras, foi realizada a normalização pelo método Standard Score (Equação 2).

$$Z = \frac{(X - \overline{X})}{s} \qquad \text{Eq.(2)}$$

em que: Z_i - variável normalizada pelo método Standard Score no ponto i; X - variável no ponto i; $\overline{X}$ - média das amostras; s - desvio padrão das amostras.

5.2.4 INTERPOLAÇÃO DOS DADOS

A interpolação dos dados foi realizada sobre todas as variáveis selecionadas, a fim de possuir uma melhor representatividade dos conjuntos de dados. Para este processo foi utilizado o método da Krigagem, a fim de criar uma grade de 5 x 5 m procurando um número mais denso de pontos por área e, portanto, definindo MZs mais suaves.

5.2.5 AGRUPAMENTO DE DADOS E MÉTRICAS DE DISTÂNCIAS

Para cada uma das três versões dos conjuntos de dados foi aplicado o algoritmo de agrupamento fuzzy c-means associado à métrica de distâncias Euclidiana, que é definida como a raiz quadrada da soma das diferenças entre x_{il}e x_{jl} elevadas ao quadrado (Equação 3). Para aplicação desta recomenda-se a normalização das variáveis (normalização dos dados) antes de se obter o valor da distância Euclidiana quando os dados não se apresentam na mesma escala de medidas (LIU et al., 2014).

$$Dist\,(E_i, E_j) = \sqrt{\sum_{l=1}^{M}\left(x_{il} - x_{jl}\right)^2} \qquad \text{Eq.(3)}$$

em que: *dist* é a distância entre os pontos; x_{il}e x_{jl} são os valores davariável de cada ponto.

A distância Euclidiana é calculada a partir do centro da célula de origem (centróide) para o centro de cada uma das células vizinhas.

Para a definição das ZMs foi utilizado o Software SDUM (Software para definição de unidades de manejo; BAZZI et al., 2013) e FuzME (MINASNY; MCBRATNEY, 2002)

5.2.6 AVALIAÇÃO DOS AGRUPAMENTOS E DA ZONAS DE MANEJO

Para a identificação do número ideal de agrupamentos formados através da aplicação de algoritmos foram utilizadas as técnicas que seguem (Equações de 4 a 7):

- Redução de variância – RV: é calculada 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 4) (DOBERMANN et al., 2003; XIANG et al., 2007).

$$RV = 1 - \frac{\sum_{i=1}^{n} W_i * V_{um_i}}{V_{área}} * 100 \qquad \text{Eq.(4)}$$

em que: n corresponde ao tamanho da amostra para toda a área, W_i é a proporção da área em cada unidade de manejo, V_{um_i} é a variância dos dados de cada unidade de manejo e $V_{área}$ é a variância da amostra dos dados para toda a área.

- Índice de desempenho Fuzzy – FPI: 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 5) (FRIDGEN et al., 2004).

$$FPI = 1 - \frac{c}{(c-1)}\left[1 - \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^2 / n\right] \qquad \text{Eq.(5)}$$

em que: c corresponde ao número de agrupamentos, n é o tamanho da amostra para toda a área (número de observações), u_{ij} é o elemento *ij* da matriz de pertinência Fuzzy.

- Índice da partição da entropia modificada – *MPE*: 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 6) (FRIDGEN et al., 2004).

$$MPE = \frac{-\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}\log(u_{ij})/n}{\log c} \qquad \text{Eq.(6)}$$

em que:c corresponde ao número de agrupamentos, n é o tamanho da amostra para toda a área (número de observações), u_{ij} é o elemento *ij* da matriz de pertinência Fuzzy.

- Índice de Validação de Cluster Melhorado – ICVI: O número ideal de agrupamentos de um conjunto de dados baseia-se no valor mínimo de FPI, MPE e máximo RV. Para evitar a situação em que estas estimativas apontem para diferentes agrupamentos, quando analisados de forma individual, o *ICVI* pode ser utilizado unindo os conceitos e considerando o agrupamento que apresentar o menor *ICVI* como o melhor (Equação 7) (GAVIOLI et al., 2016)

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

em que: *i* corresponde aos índices de todas as métricas de distância utilizadas no experimento.

Para a identificação de quais foram as melhores ZMs definidas através da aplicação dos algoritmos de agrupamentos e suas respectivas distâncias utilizaram-se as seguintes técnicas:

- Análise de Variância - ANOVA:

Análise de variância é uma técnica estatística que permite avaliar afirmações sobre as médias de populações (MILONE, 2009). A análise visa, primordialmente, verificar se existe uma diferença significativa entre as médias e se os fatores exercem influência em alguma variável dependente.

Após a identificação de que as médias são estatisticamente diferentes, é necessário um método que forneça a diferença mínima significante entre duas médias. Essa diferença seria o instrumento de medida. Toda vez que o valor absoluto da diferença entre duas médias é igual ou maior do que a diferença mínima significante, as médias são consideradas estatisticamente diferentes, ao nível de significância estabelecida (VIEIRA et al., 1989). Neste trabalho a comparação de médias foi feita pelo Teste de Tukey.

De acordo com Bazzi (2011), o teste de comparação de médias (Teste de Tukey) da ANOVA pode ser aplicado para verificar se as subáreas definidas realmente representam grupos diferentes a determinado nível de significância. Porém, esse teste considera que as amostras são independentes dentro de cada ZM e, por isso, é necessário verificar primeiramente que se em cada ZM não existe dependência espacial dos dados.

- Índice de Suavidade (Smoothness index) –*SI (*Equação 8, GAVIOLI et al., 2016)

A avaliação dos melhores métodos de definição de agrupamento deve também incluir o aspecto visual do agrupamento criado e, portanto, deve-se levar em conta a suavidade das curvas de contorno, pois facilita a interpretação visual e a aplicação em taxa variada de insumos agrícolas. O índice de suavidade (*SI*) calcula a frequência da mudança de classes nos mapas temáticos nas direções horizontais, verticais e diagonais, pixel por pixel. Na hipótese de que o mapa seja uma única área totalmente homogênea, um índice de suavidade de 100% será obtido, devido à ausência de mudança de classe. Da mesma forma, se o mapa foi gerado com valores aleatórios, o índice SI apresentaria um valor próximo de zero.

$$SI = 100 - \left(\left(\frac{\sum_{i=1}^{l} NM_{Hi}}{4P_H} + \frac{\sum_{i=1}^{c} NM_{Vi}}{4P_V} + \frac{\sum_{i=1}^{n} NM_{Ddi}}{4P_{Dd}} + \frac{\sum_{i=1}^{n} NM_{Dei}}{4P_{De}} \right) * 100 \right) \qquad \text{Eq.(8)}$$

em que: NM_{Hi} corresponde ao número de mudanças na linha horizontal i; NM_{Vi} é o número de mudanças na linha vertical j; NM_{Ddi} é o número de mudanças na diagonal direita; NM_{Dei} é o número de mudanças na diagonal esquerda; P_H é a possibilidade de mudança na horizontal; P_V é a possibilidade de mudança na vertical; P_{Dd} é a possibilidade de mudança na diagonal direita e P_{De} é a possibilidade de mudança na diagonal esquerda.

- Índice Kappa: para avaliar a concordância entre as ZMs definidas utilizando diferentes métricas de distâncias, foi utilizado o índice Kappa. Para isso, utilizaram-se os graus de concordância Kp sendo classificados conforme proposto por Landis e Koch (1977):

0 <Kp ≤ 0,2: não há concordância; 0,2 <Kp ≤ 0,4: fraca; 0,4 <Kp ≤ 0,6: moderada; 0,6 <Kp ≤ 0,8: forte; 0,8 <Kp ≤ 1: muito forte *(*Equação 9) (COHEN, 1960).

$$K = \frac{\left\{ n\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}\left(x_{i+} * x_{+i}\right)\right\}}{\left\{ n^2 - \sum_{i=1}^{r}\left(x_{i+} * x_{+i}\right)\right\}} \qquad \text{q.(9)}$$

em que: K - índice Kappa de concordância; n - o número total de observações (pontos amostrais); r - número de classes da matriz de erro; x_{ii} - número de combinações na diagonal; x_{i+} - total de observações na linha i; x_{+i} - total de observações na coluna i.

6.3 RESULTADOS E DISCUSSÃO

A análise descritiva da produtividade média de soja e milho de cada safra (Tabela 3) mostrou que, para a área A, a safra de soja em 2014 foi a que apresentou a melhor média (4.525) entre os anos analisados, com os menores SD e CV (0,281 t ha^{-1} e 6,2%, respectivamente), indicando homogeneidade na cultura. A área B apresenta para todas as suas safras de soja uma média dispersão em sua produtividade, que pode ser identificada através do CV que se encontra na faixa entre 10 e 20% para todos os anos, o que não se aplica para a cultura do milho, que em suas safras expressa uma baixa dispersão. Para a área C, identificou-se baixa dispersão em sua produtividade de soja e milho com análise do CV, e sua safra de soja em 2012 foi a que apresentou a maior produtividade, média e mediana dentre os conjuntos de dados (7,44; 5,35 e 5,28 t ha^{-1} respectivamente). Apesar da área A ter apresentado o maior valor de produtividade de milho em sua safra de 2014 (14,71 t ha^{-1}), foi a área C, em mesmo ano de safra, que apresentou a maior média e mediana (10,28 e 40,48 t ha^{-1}, respectivamente) com a referida cultura.

Tabela 3 Análise descritiva da produtividade média de cada safra por cultura

Área	Safra	Media	Mediana	Maximo	Minimo	SD	CV (%)
Dados de Produtividade Soja- Originais (amostragem de dados) (t ha^{-1})							
A	2012	3,984	4,067	5,068	2,541	0,472	11,8
	2013	3,941	4,012	6,166	2,057	0,518	13,1
	2014	4,525	4,553	5,354	3,635	0,281	6,2
B	2012	5,255	5,255	6,980	2,563	0,749	14,2
	2013	5,079	5,107	6,808	3,366	0,586	11,5
	2014	3,888	3,819	4,759	2,934	0,496	12,8
C	2012	5,348	5,277	7,436	4,029	0,764	6,9
	2013	4,675	4,711	5,817	3,357	0,518	9,0
	2014	4,505	4,557	5,610	2,579	0,584	7,7
Dados de Produtividade Milho - Originais (amostragem de dados) (t ha^{-1})							
A	2014	8,950	9,147	14,707	2,804	1,741	19,4
B	2014	8,945	9,195	10,798	6,678	0,991	11,0
	2015	10,276	10,415	13,342	6,762	1,094	10,6
C	2013	4,813	4,750	6,692	3,626	0,687	14,2
	2014	10,284	10,479	13,178	7,370	1,125	10,9
	2015	8,420	8,425	10,772	5,262	0,975	11,5

CV: coeficiente de variação; SD: desvio padrão.

A Tabela 4 apresenta as produtividades normalizadas pelo método standard score, em que se destaca a média da área A para as produtividades de milho e soja/milho, que ficaram com suas médias tendendo a zero. A área B mostra um coeficiente de variação negativo em função das suas médias em todas as produtividades serem negativas também. Ainda sobre a área B, pode-se identificar que ela apresenta o menor SD entre as produtividades (0,505), indicando uma maior estabilidade que as demais.

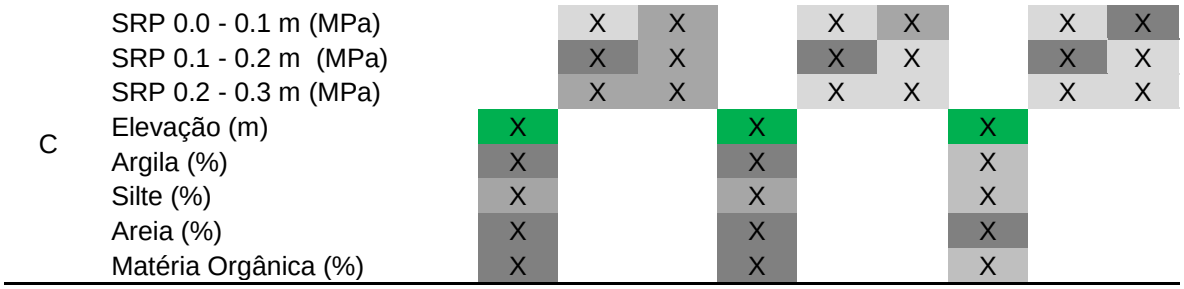
Tabela 4 Análise descritiva da produtividade média normalizada (média zero) de cada safra por cultura

Cultura	Área	Mediana	Maximo	Minimo	SD
Soja	A	-0,044	1,812	-1,184	0,61
	B	0,065	0,905	-1,224	0,50
	C	-0,035	1,448	-1,805	0,69
Milho	A	0,112	3,307	-3,533	1,0
	B	0,060	1,954	-2,653	0,74
	C	-0,131	1,662	-1,245	0,69
Soja + Milho	A	-0,007	1,218	-1,073	0,50
	B	-0,017	0,635	-1,538	0,40
	C	-0,080	0,609	-1,189	0,61

Durante o processo de seleção de variáveis, onde cada uma das produtividades foi utilizada para o processo de Correlação Espacial, obtiveram-se as variáveis apresentadas na Tabela 5, para cada uma das áreas em estudo.

Tabela 5 Seleção de variáveis para o processo de definição das ZMs das áreas A, B e C

Área	**Produtividade**	**Soja**			**Milho**			**Soja + Milho**		
	Variáveis	**2012**	**2013**	**2014**	**2012**	**2013**	**2014**	**2012**	**2013**	**2014**
A	SRP 0.0 - 0.1 m (MPa)		X	X		X	X		X	X
	SRP 0.1 - 0.2 m (MPa)		X	X		X	X		X	X
	SRP 0.2 - 0.3 m (MPa)		X	X		X	X		X	X
	Elevação (m)	X			X			X		
	Declividade (˚)	X			X			X		
	Densidade (g cm^{-3})	X			X			X		
	Argila (%)	X			X			X		
	Silte (%)	X			X			X		
	Areia (%)	X			X			X		
	Matéria Orgânica (%)	X			X			X		
B	SRP 0.0 - 0.1 m (MPa)	X	X	X	X	X	X	X	X	X
	SRP 0.1 - 0.2 m (MPa)	X	X	X	X	X	X	X	X	X
	SRP 0.2 - 0.3 m (MPa)	X	X	X	X	X	X	X	X	X
	Elevação (m)	X			X			X		
	Argila (%)	X			X			X		
	Silte (%)	X			X			X		
	Areia (%)	X			X			X		
	Matéria Orgânica (%)	X			X			X		



[ ] - Eliminado por não ter autocorrelação espacial; [ ] - Eliminado por não ter correlação espacial com a produtividade; [ ] - Eliminado por ser redundante; [ ] - Selecionado para gerar as MZs.

A variável elevação foi selecionada na área A para definição das ZMs para as duas culturas e na sua associação também. Já a variável de resistência à penetração ao solo (SRP 0.0 - 0.1 m (2013)) foi escolhida em conjunto com a elevação para a cultura de soja. Já na área B a variável SPR foi selecionada para definição das ZMs de todas culturas, variando apenas profundidade de medição (SRP 0.0 - 0.1 m e 0,2 – 0,3 m). A variável elevação foi também selecionada, junto à correlação da produtividade de soja + milho para a definição dos agrupamentos. Para a área C, apenas a variável elevação foi selecionada através da análise da correlação espacial para as produtividades de soja, milho e soja + milho (Tabela 5).

A Figura 3 apresenta os mapas temáticos das variáveis produtividade normalizada de soja, milho e soja+milho, elevação e resistência à penetração no solo (SRP), com duas, três e quatro classificações. Para todas as áreas, os mapas temáticos que correspondem ao atributo de elevação são os que apresentam uma melhor separação entre as classes, mesmo quando o número de agrupamentos é aumentado.

Já o atributo de resistência à penetração do solo nos apresenta mapas com classes descontínuas, que se acentuam conforme o seu número é incrementado, se comparado com a elevação. Porém, pode-se identificar que os níveis de penetração também se diferem em seus mapas temáticos, onde o atributo SRP 0.2 – 0.3 apresentou gráficos com classes melhores delineadas do que SRP 0.0 – 0.1 (Área B). No que se refere aos mapas temáticos das produtividades, observa-se que o tipo da cultura (soja ou milho), assim como a associação entre elas (soja+milho), geram classes com diferenças significativas entre si, expressando a importância do estudo sobre estas variações.

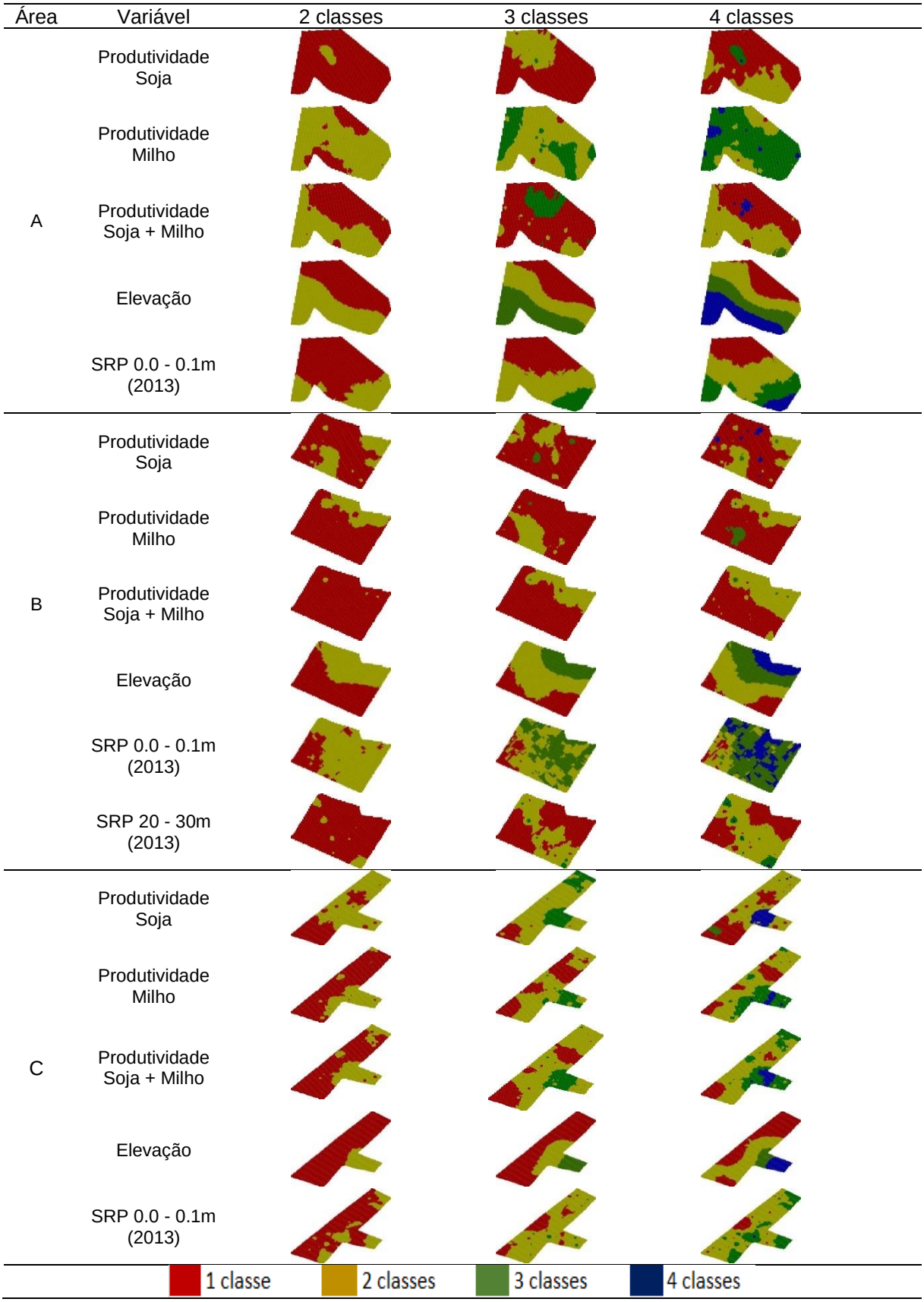


Figura 3 Mapas temáticos das variáveis produtividade normalizada de soja, milho e soja+milho, elevação e resistência à penetração no solo (SRP), com duas, três e quatro classificações.

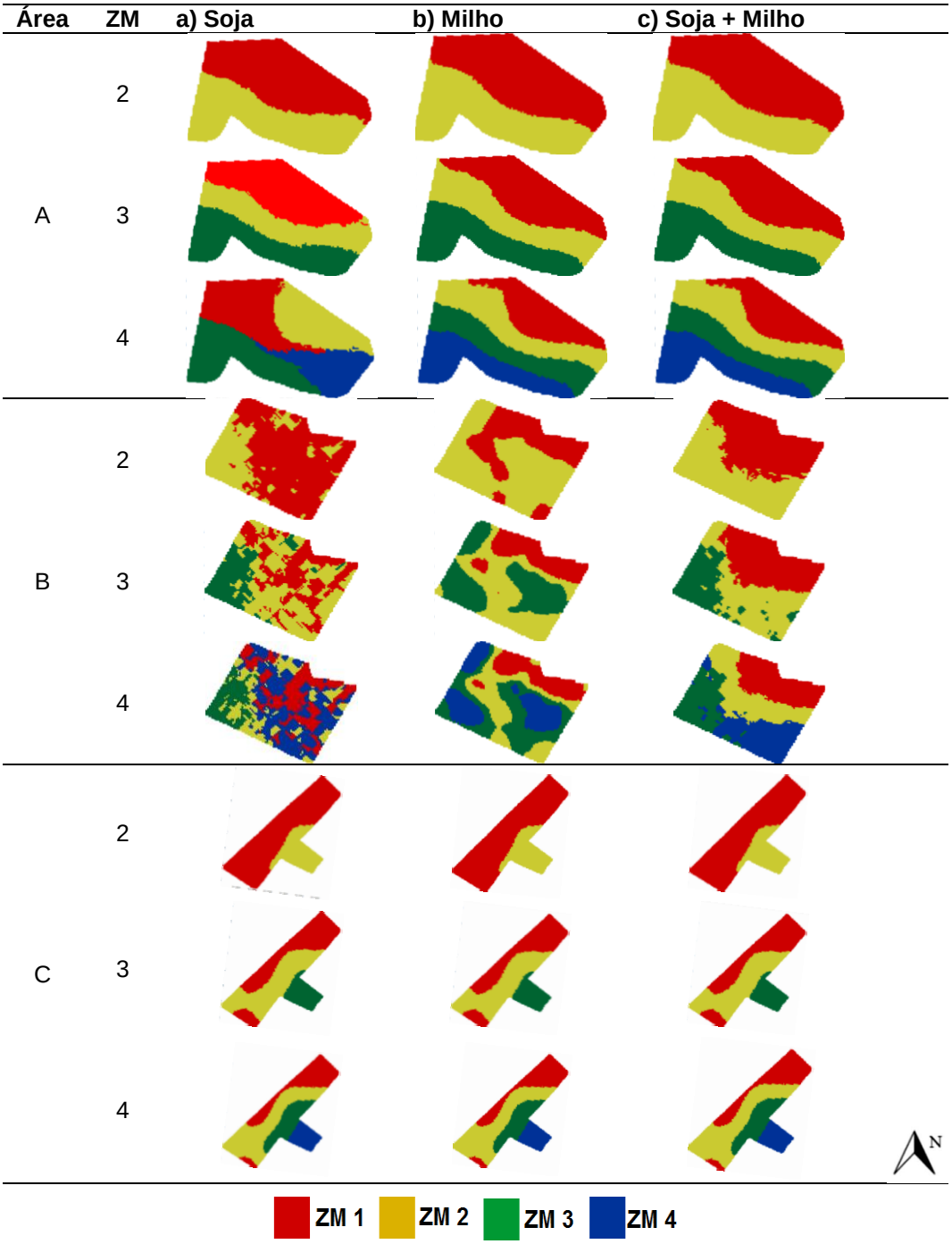


Figura 4 Zonas de manejo para as áreas A, B e C, considerando com a seleção de variáveis baseadas na produtividade de Soja (a); Milho (b) e Soja + Milho (c).

Para a área A notamos que houve diferenças visualmente expressas apenas nas ZMs com quatro agrupamentos na produtividade de soja, o que pode ser justificado pelo fato de, apenas nesta situação, o índice de resistência de penetração ao solo 0.0 - 0.1 ter sido utilizado junto ao atributo elevação.

Na área B pode-se observar que as ZMs se diferem entre elas, independentemente do número de agrupamentos utilizado, visto que para cada situação de análise de correlação espacial, onde a produtividade é utilizada no processo, variáveis diferentes foram

selecionadas para cada uma das situações, originando resultados distintos para cada uma delas.

Já para a área C, independente da produtividade utilizada no processo de correlação espacial, todas as situações indicaram o atributo elevação como sendo o mais indicado para a definição de ZM, o que justifica os mapas temáticos iguais para todas as culturas, variando apenas com o acréscimo do número de agrupamentos.

Os valores dos índices RV, FPI, MPE, ICVI e SI, além do teste Tukey das médias de produtividade, estão apresentados na Tabela 6. São apresentados os resultados obtidos pelo uso de três produtividades (soja, milho e soja+milho) no processo de seleção de variáveis que foram utilizadas na definição das ZMs.

Tabela 6 Índices de avaliação das zonas de manejo definidas utilizando diferentes produtividades – Dados normalizados pelo método da Amplitude

Área	Nº Zonas	Produtividade	Anova e Teste de Tukey*				RV%	FPI	MPE	ICVI	SI%
			ZM1	ZM2	ZM3	ZM4					
A	2	Soja	a	b			7,8	0,117	0,141	0,780	98,4
		Milho	a	a			-4,2	0,081	0,099	0,762	98,5
		Soja + Milho	a	b			27,5	0,081	0,099	0,378	98,5
	3	Soja	a	a	b		22,1	0,152	0,163	0,728	96,9
		Milho	a	ab	ab		-4,7	0,090	0,100	0,790	97,0
		Soja + Milho	a	b	a		26,0	0,090	0,100	0,417	97,0
	4	Soja	a	a	ab	b	17,2	0,154	0,158	0,781	96,4
		Milho	a	ab	bc	abc	11,5	0,092	0,095	0,587	95,7
		Soja + Milho	a	b	a	b	27,2	0,092	0,095	0,397	95,7
B	2	Soja	a	a			-2,6	0,089	0,109	0,861	92,3
		Milho	a	a			1,3	0,097	0,115	0,686	96,2
		Soja + Milho	a	b			6,3	0,151	0,175	0,651	98,1
	3	Soja	a	ab	ab		6,4	0,095	0,104	0,395	83,6
		Milho	a	a	a		-15,5	0,093	0,103	1,529	93,6
		Soja + Milho	a	ab	ab		-4,7	0,159	0,171	1,233	96,7
	4	Soja	a	ab	abc	abc	-6,3	0,090	0,093	1,025	78,6
		Milho	a	ab	ab	b	-10,8	0,098	0,101	1,291	90,3
		Soja + Milho	a	a	ab	b	4,0	0,161	0,166	0,775	95,0
C	2	Soja	a	b			12,5	0,072	0,090	0,722	98,9
		Milho	a	b			26,4	0,072	0,090	0,547	98,9
		Soja + Milho	a	b			23,8	0,072	0,090	0,580	98,9
	3	Soja	a	ab	ab		-8,9	0,095	0,102	1,112	97,5
		Milho	a	a	b		9,7	0,095	0,102	0,878	97,5
		Soja + Milho	a	ab	a		-1,3	0,095	0,102	1,016	97,5
	4	Soja	a	ab	c	abc	16,8	0,089	0,091	0,731	96,8
		Milho	a	b	b	a	21,6	0,089	0,091	0,670	96,8
		Soja + Milho	a	a	b	ac	18,7	0,089	0,091	0,707	96,8

* Significativo ao nível de 0,05. Casos sombreados em cinza são significativamente diferentes.

Considerando que somente é interessante dividir a área total em ZMs que possuam variável-alvo (neste caso a produtividade) estatisticamente distintas, a primeira análise a ser feita é o Teste Tukey de médias. Como resultado, tem-se que somente é aconselhável se dividir em duas ZMs (casos sombreados em cinza). A área A pode ser dividida usando a variável-alvo produtividade de soja e produtividade de soja+milho; a área B, somente produtividade de soja+milho, a área C, com as três produtividades (soja, milho e soja+milho). Com relação aos índices, tanto melhor quanto maior for RV e SI e menor for

FPI, MPE e ICVI, o que resulta na área A em vantagem da produtividade soja+milho sobre as demais. Já para a área B a produtividade de soja+milho foi a que se destacou com 2 e 4 ZMs, sendo para 3 ZMs a produtividade de soja a mais indicada. Para a área C apenas os índices de RV e SI apresentaram valores diferenciados, em função da mesma variável ter sido selecionada no processo de análise de correlação espacial, indicando, assim, a cultura de milho com melhores resultados.

Considerando que o índice ICVI corresponde a uma composição dos índices RV, FPI e MPE, pode-se restringir a análise dos índices somente analisando os índices ICVI e SI (Tabela 6 e Figura 5). O SI, que caracteriza a suavidade das curvas de contorno das ZMs (facilita a interpretação visual e a aplicação em taxa variada de insumos agrícolas) praticamente manteve-se constante dentro de cada número de ZMs, tendendo a diminuir suavemente com o aumento do número de ZMs.

O ICVI, que é tanto melhor quanto menor, teve o seguinte comportamento em função da área:

ÁREA A: O menor ICVI correspondeu à variável produtividade-alvo soja com três ZMs. Infelizmente para esta divisão das ZMs não foram significativamente distintas.

ÁREAS B, C: Uma área em discordância com a outra. Menores ICVIs: variável produtividade-alvo soja com três ZMs para a área B e variável produtividade-alvo milho com duas ZMs para a área C.

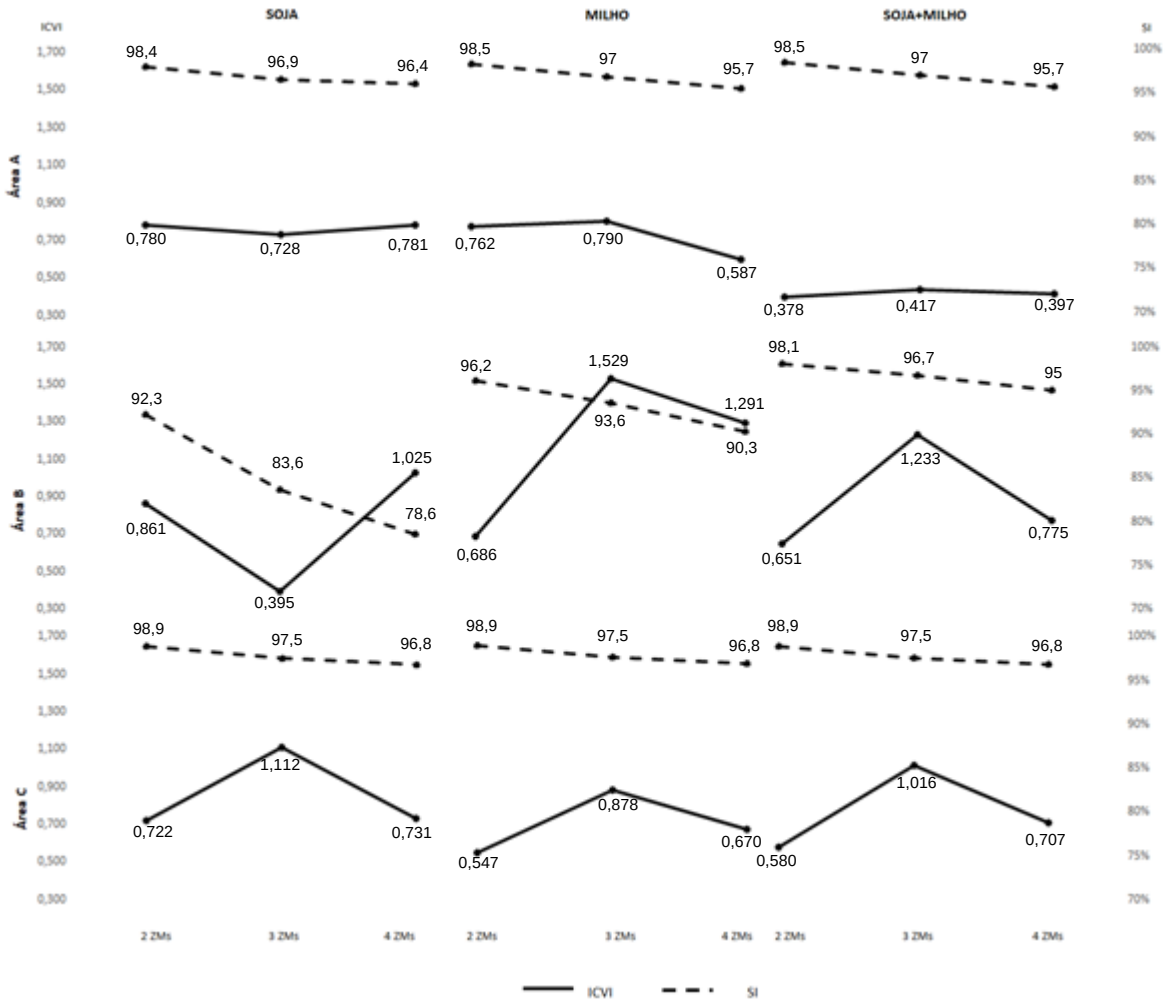


Figura 5 Índice de Validação de Cluster Melhorado (ICVI) e Índice de Suavidade (SI) obtidos com a utilização dos dados normalizados, nos agrupamentos para as áreas A, B e C em cada uma das culturas estudadas.

Para definição do Índice Kappa, foi realizada uma comparação entre pares de culturas: Milho X Soja, Milho X Soja+Milho, Soja X Soja+Milho. Para cada par analisado foram considerados os agrupamentos com duas, três e quatro ZMs (Figura 6).

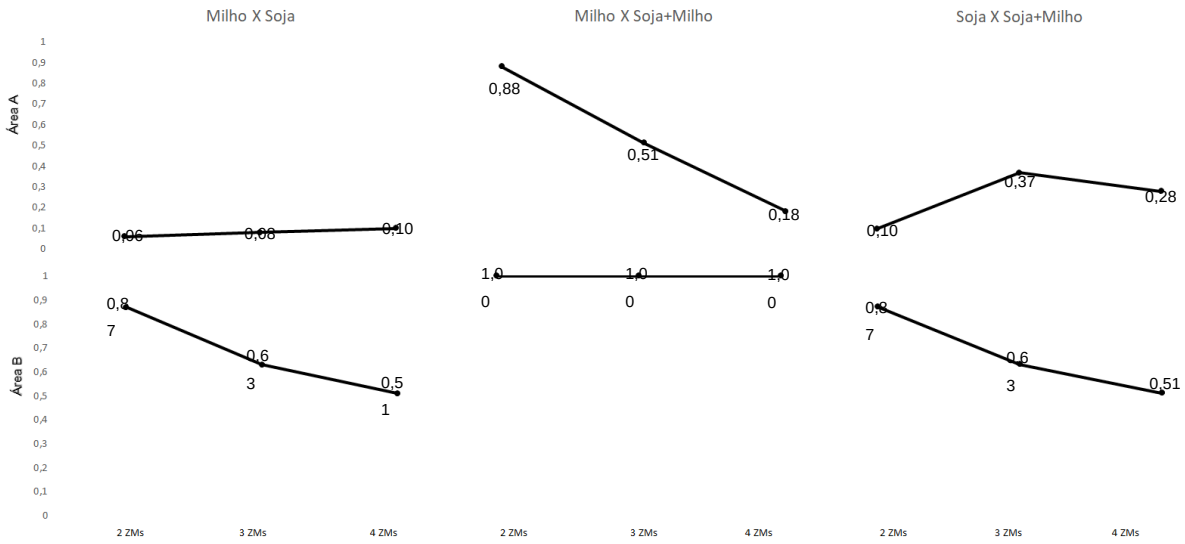


Figura 6 Índice Kappa entre variáveis-alvos (produtividade de soja, milho e soja+milho) em função do número de zonas de manejo (ZMs).

O Índice Kappa foi calculado apenas para as áreas A e B, visto que a área C não apresentou diferenças entre seus mapas temáticos. Observa-se que na área A as ZMs que foram mais concordantes foram definidas para duas ZMs e com as variáveis alvo produtividade de milho e produtividade de soja+milho) consideradas muito fortes (Landis e Koch, 1977). Já para área B a concordância foi muito forte para duas combinações de produtividade alvo: milho x soja, soja x soja+milho.

6.4 CONCLUSÕES

A utilização de diferentes produtividades-alvo (soja, milho e soja+milho) influenciou a seleção de variáveis a serem utilizadas na definição das zonas de manejo (ZMs) em duas das três áreas), bem como no desempenho das ZMs selecionadas para cada área, mesmo quando as variáveis utilizadas na definição das ZMs foram as mesmas (área C).

Para todas as áreas é aconselhável a divisão em duas ZMs.

A melhor produtividade-alvo foi soja+milho, reforçando a ideia de ser útil a utilização destas duas culturas, em conjunto, na definição das ZMs de uma área com alternância de produção de soja e milho.

6.5 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.

BAZZI, C. L. **Software para definição e avaliação de unidades de manejo em agricultura de precisão**. 2011. 123f. Tese (Doutorado em Engenharia Agrícola). Programa de Pós-Graduação em Engenharia Agrícola. Universidade Estadual do Oeste do Paraná. Cascavel, 2011.

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

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

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. 5, p. 1105-1120, 2003.

DOERGE, T. A. **Management Zone Concepts**. Site-Specific Management Guidelines. Potash & Phosphate Institute (PPI), South Dakota State University (SDSU), 2000.

FERREIRA, D. B. **Relações entre a variabilidade da precipitação e a produtividade agrícola de soja e milho nas regiões sul e sudeste do Brasil.** São José dos Campos: INPE, 2005.

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 Engineers,** v. 44, n. 1, p. 155-166, 2001.

FRIDGEN, J. J.; KITCHEN, N. R.; SUDDUTH, K. A. Variability of soil and landscape attributes within sub-field management zones. In: International Conference on Precision Agriculture. **Anais…** Bloomington: Madison, 2000.

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.

JOHANNSEN, C. J.; CARTER, P. J.; ERICKSON, B. J.; MORRIS, D. K.; WILLIS, P. R. A cornucopia of agricultural applications. **Space Imaging**, Thornton, Jan/Fev, p.22-23, 2000.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining Geostatistics.** London: Academic Press, 1978.

LIU, Q.; CHU, X.; XIAO, J.; ZHU, H. Optimizing Non-orthogonal Space Distance Using PSO in Software Cost Estimation. In: IEEE Computer Software Applications Conference (COMPSAC). **Anais...,** 2014.

METZ, J.; MONARD, M. C. **Projeto e implementação do módulo de clustering hierárquico do discover**. ICMC-USP, 2006.

MILONE, G. **Estatística geral e aplicada.** São Paulo: Centage Learning, 2009.

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

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.

MOLIN, J. P.; AMARAL, L. R.; COLAÇO, A. **Agricultura de Precisão.** São Paulo: Oficina de Textos, 2015.

OLIVEIRA, A. D.; COSTA, J. M. N.; LEITE, R. A.; SOARES, P. C.; SOARES, A. A. Probabilidade de chuvas e estimativas de épocas de semeadura para cultivares de arroz de sequeiro, em diferentes regiões do Estado de Minas Gerais, Brasil. **Revista Brasileira de Agromeorologia**, v. 8, n. 2, p. 295-309, 2000.

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.

VIEIRA, S.; HOFFMANN, R. **Estatística experimental.** São Paulo: Atlas, 1989. 175p.

XIANG, L.; YU-CHUN, P.; ZHONG-QIANG, G.; CHIN-JIANG, Z. 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.

7 CONSIDERAÇÕES FINAIS

7.1 CONCLUSÕES

Os experimentos realizados para avaliar de forma comparativa as três métricas de distâncias empregadas junto ao algoritmo fuzzy c-means possibilitaram concluir que as distâncias Diagonal e Mahalanobis se mostram eficientes mesmo em um conjunto de dados não normalizados, o que mostra que o fato do desvio padrão ser utilizado em seu cálculo reproduz o mesmo efeito da normalização no conjunto de dados. Porém, a partir do momento que os dados normalizados foram analisados, foi possível notar a evolução da distância Euclidiana tanto em aspectos visuais quanto em seus índices calculados.

Quanto às culturas utilizadas pode-se referenciar suas influências de forma significativa em dois aspectos: primeiro sobre o processo de seleção de variáveis em que a produtividade é o elemento chave para a seleção destas; e, em segundo momento, não menos importante, no processo de avaliação dos agrupamentos e ZMs definidas. Aqui, a mais importante conclusão foi que a melhor produtividade-alvo foi soja+milho, reforçando a ideia de ser útil a utilização destas duas culturas, em conjunto, na definição das ZMs de uma área com alternância de produção de soja e milho.

Outro fato que pode ser identificado é o poder de influência que o atributo elevação tem sobre a produtividade, onde em 76% dos processos de seleção de variáveis esta esteve presente entre as eleitas, sempre apresentando com ela o maior índice de correlação com a produtividade.

7.2 TRABALHOS FUTUROS

Inicialmente, pretende-se implementar e incluir as métricas de distâncias de Diagonal e Mahalanobis junto ao SDUM, que em conjunto com a Euclidiana (já implementada) comporão o conjunto de três distâncias que poderão ser associadas aos métodos de agrupamentos já existentes no mesmo (fuzzy c-means e k-means). Estas implementações serão utilizadas na nova plataforma web que está em construção. Com isso, as métricas implementadas nesses módulos poderão ser utilizadas diretamente na Internet, por meio do uso de um software de navegação (web browser).