UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ - UNIOESTE

CAMPUS CASCAVEL

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS - CCET

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA – PGEAGRI

REGIONALIZAÇÃO DE ÁREA AGRÍCOLA USANDO DADOS DE IMAGENS AÉREAS E COLETAS DE CAMPO

RODRIGO LORBIESKI

Cascavel – Paraná – Brasil

Fevereiro - 2020

RODRIGO LORBIESKI

REGIONALIZAÇÃO DE ÁREA AGRÍCOLA USANDO DADOS DE IMAGENS AÉREAS E COLETAS DE CAMPO

Projeto de dissertação apresentado ao Programa de Pós-Graduação Stricto Sensu em Engenharia Agrícola, em cumprimento parcial aos requisitos para obtenção do título de Mestre em Engenharia Agrícola, área de concentração Sistemas Biológicos e Agroindustriais e linha de pesquisa em Geoprocessamento, Estatística Espacial e Agricultura de Precisão da Universidade Estadual do Oeste do Paraná, UNIOESTE, Campus Cascavel

Orientadora: Profª. Drª. Luciana Pagliosa Carvalho Guedes

Cascavel – Paraná – Brasil
Fevereiro – 2020

Ficha de identificação da obra elaborada através do Formulário de Geração Automática do Sistema de Bibliotecas da Unioeste.

Lorbieski, Rodrigo
Regionalização de área agrícola usando dados de imagens aéreas e coletas de campo / Rodrigo Lorbieski; orientador(a), Luciana Pagliosa Carvalho Guedes, 2020.
104 f.

Dissertação (mestrado), Universidade Estadual do Oeste do Paraná, Campus de Cascavel, Centro de Ciências Exatas e Tecnológicas, Programa de Pós-Graduação em Engenharia Agrícola, 2020.

1. Agrupamentos. 2. Árvore de decisão. 3. Classificação supervisionada. 4. Função núcleo estimador. I. Guedes, Luciana Pagliosa Carvalho. II. Título.

Revisor de português, inglês e normas do PGEAGRI: Ana Maria M. A. Vasconcelos.
Maio de 2020.

RODRIGO LORBIESKI

Regionalização de uma área agrícola usando dados de imagens aéreas e de coletas de campo

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Agrícola em cumprimento parcial aos requisitos para obtenção do título de Mestre 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) - Luciana Pagliosa Carvalho Guedes
Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Gustavo Henrique Dalposso
Universidade Tecnológica Federal do Paraná (UTFPR)

Miguel Angel Uribe Opazo
Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Cascavel, 14 de fevereiro de 2020.

BIOGRAFIA RESUMIDA

Rodrigo Lorbieski, nascido em 16 de julho de 1983 no município de Campo Mourão, estado do Paraná, filho de Maria do Carmo Alves Lorbieski e Adolfo Lorbieski. Gradou-se em Licenciatura Plena em Ciências Biológicas pela Universidade Estadual do Oeste do Paraná (2008) e em Licenciatura Plena em Matemática também pela Universidade Estadual do Oeste do Paraná (2018). Possui especialização *lato sensu* em Docência do Ensino Superior pela Faculdade Assis Gurgaz (2012), e em Ensino de Ciências e Matemática pela Universidade Estadual do Oeste do Paraná (2012). Atualmente é discente de mestrado no Programa de Pós-Graduação em Engenharia Agrícola, sob a orientação da Prof.ª Dr.ª Luciana Pagliosa Carvalho Guedes.

AGRADECIMENTOS

Meus sinceros agradecimentos a todos aqueles que de alguma forma tornaram esse projeto possível. Entre eles, os amigos dos laboratórios de estatística aplicada (LEA), GEOSCIENCE e estatística espacial (LEE), por me fornecer os materiais e o suporte necessários para a realização deste trabalho. Meus agradecimentos também a todos os professores que fizeram parte da minha história e que de alguma forma compartilharam comigo seus conhecimentos, possibilitando, dessa forma, que me tornasse uma pessoa melhor e um profissional mais preparado.

Quero agradecer, de forma especial, a minha orientadora, professora Doutora Luciana, que sempre me auxiliou em todas as dúvidas que surgiram, apresentando os melhores caminhos para alcançar os resultados pretendidos.

Agradeço também à CAPES pelo apoio financeiro por meio da bolsa de estudo que me possibilitou dedicação exclusiva à pesquisa tornando possível a realização deste trabalho.

Por fim, e não menos importante, meu agradecimento a minha família, aos amigos e colegas que me deram força quando precisei e, assim, dividiram comigo os momentos bons e difíceis desses anos do curso de matemática.

A todos o meu muito OBRIGADO!

REGIONALIZAÇÃO DE ÁREA AGRÍCOLA USANDO DADOS DE IMAGENS AÉREAS E COLETAS DE CAMPO

RESUMO

A agricultura de precisão é uma importante ferramenta que visa à otimização da produtividade do setor agrícola, entretanto, seu custo da implementação pode ser um obstáculo para os pequenos agricultores. Uma alternativa seria, então, a regionalização da área agrícola dividindo-a em talhões ou zonas de manejo, que podem ser trabalhadas individualmente de acordo as características da área. Dados multivariados são comuns no processo de delineamento dessas zonas, assim, técnicas multivariadas para classificação e agrupamento podem ser aplicadas a esses dados, buscando considerar também a informação espacial destes. Dessa forma, essa pesquisa delineou zonas de manejo de uma área agrícola em quatro anos-safra consecutivos, com o uso de agrupamentos hierárquicos não paramétricos, levando em consideração a informação espacial de dados provenientes de atributos físicos e químicos do solo, índices vegetativos e dados de produtividade. Foram utilizadas técnicas para redução da dimensionalidade e agrupamentos dos dados, além de análises geoestatísticas e de estatísticas descritivas. Para melhor compreensão do comportamento das variáveis físico-químicas nas diferentes zonas de manejo formadas, construíram-se árvores de decisão, tendo como variável resposta as próprias zonas de manejo. Os subconjuntos que melhor formaram as zonas de manejo variaram de um ano-safra para o outro, e, a localização delas foi semelhante em três dos quatro anos-safra analisados. O ótimo número de zonas de manejo foi igual a dois em todos os anos-safra analisados. As árvores de decisão se mostraram importantes para caracterização das variáveis físico-químicas, pois auxiliaram a descrever a distribuição delas na formação de cada zona de manejo.

PALAVRAS-CHAVE: agrupamentos; árvore de decisão; classificação supervisionada; função núcleo-estimador.

REGIONALIZATION OF AGRICULTURAL AREA USING DATA FROM AERIAL IMAGES AND FIELD SAMPLINGS

ABSTRACT

Precision agriculture is an important tool that aims at optimizing the agricultural yield sector, however, its cost of implementation can be an obstacle for small farmers. An alternative would be regionalizing the agricultural area, and dividing it into plots or management zones, which can be worked individually according to the area characteristics. Multivariate data are common in designing these zones, thus, multivariate techniques for classification and grouping can be applied to these data, aiming at also taking into account their spatial information. Thus, this research outlined management areas for an agricultural area in four consecutive cropping years, using non-parametric hierarchical groupings, and considering spatial data information from physical and chemical attributes of soil, vegetative indexes and data of yield. Techniques were applied to reduce dimensionality and groupings of data, as well as geostatistical analyses and descriptive statistics. So, decision trees were built to better understand physical-chemical variables behavior in the different management zones formed, whose management zones are the response variable. The subsets that best formed the management zones varied from one cropping year to the next one, and their location was similar in three of the four cropping years analyzed. The excellent number of management zones was equal to two in all the studied cropping years. Decision trees proved to be important to characterize physical-chemical variables, as they helped to describe their distribution in the formation of each management zone.

KEYWORDS: groupings; decision tree; supervised classification; core-estimator function.

SUMÁRIO

LISTA DE TABELA

LISTA DE FIGURAS

1. INTRODUÇÃO

A necessidade crescente por produção de alimentos exige constantes estudos e pesquisas na área agrícola. Esse fato culmina em uma constante evolução do campo e isso pressiona o agricultor a adotar novos métodos que garantam aumento na produtividade e redução de custos.

O avanço nas áreas de tecnologia da informação permite que o agricultor tome decisões baseadas em dados cada vez mais precisos em relação a diversas variáveis que influenciam diretamente na produção agrícola. Essas tecnologias são responsáveis pela melhor análise das áreas produtivas tais como características físicas e químicas do solo, relevo, clima e desenvolvimento vegetativos das diversas culturas exploradas.

Neste cenário de tecnologia aplicada ao campo, surge a agricultura de precisão. Ela pode ser definida, de acordo com Balastreire (1998), como um conjunto de técnicas que permitem o gerenciamento localizado de culturas. Esse tipo de agricultura tem como fundamento a aplicação fixada de insumos e assim permite a diminuição do uso desses e, por consequência, a redução tanto de custos como dos impactos ambientais.

Entretanto, apesar das vantagens, a agricultura de precisão necessita de máquinas especializadas para sua aplicação, além de outros recursos tecnológicos que encarecem o processo e o tornam inviável, principalmente, para pequenos e médios produtores. Uma alternativa para esse problema seria a aplicação da técnica de zonas de manejo.

A criação de zonas de manejo se baseia na ideia de dividir talhões em unidades menores com características homogêneas, permitindo assim a aplicação de insumos a taxas constantes dentro de cada unidade e a taxas variadas entre as unidades, ou seja, com diferenças na quantidade de insumos aplicados entre as diferentes unidades.

A utilização das zonas de manejo permite que os pequenos e médios produtores agrícolas façam uso dos benefícios gerados pela agricultura de precisão, pois, podem usar suas máquinas tradicionais e adaptar a quantidade de insumos conforme a necessidade de cada unidade, tornando o processo mais viável economicamente.

Para o delineamento dessas unidades, é comum a consulta de conjunto de dados multivariados, que levam em consideração dados de produtividade, atributos físicos e químicos do solo, topografia, índices vegetativos e a combinação desses (RODRIGUES JR. *et al.*, 2011). Dessa forma, é importante a aplicação de técnicas de análise que considerem as associações espaciais entre essas variáveis e, assim, aproveitar o máximo do potencial de informações desses dados.

Uma maneira para se realizar a análise dos dados é empregar uma abordagem geoestatística multivariada, para a qual há construção e interpretação de semivariograma experimentais diretos e cruzados para todos os pares de variáveis que foram utilizadas no estudo (FOUEDJIO, 2016).

Além disso, o delineamento de zonas de manejo ocorre a partir de métodos de agrupamentos, supervisionados ou não. Os supervisionados supõem a existência de exemplos nos quais o analista ajuda na construção de um modelo baseado nas definições de classes e dos exemplos em cada método. Os métodos não supervisionados fazem a classificação baseada em observações e descobertas, e o sistema analisa os exemplos e reconhece os padrões automaticamente (FERREIRA, 2011).

Percebe-se então a necessidade de que sejam criadas áreas para manejo na agricultura, buscando nas tecnologias disponíveis formas de aliar praticidade e baixo custo para o delineamento dessas áreas, com aumento de produção e lucratividade. Assim, o desenvolvimento de pesquisas que visam ao aperfeiçoamento das técnicas de agricultura de precisão é fundamental para que tal ferramenta se torne acessível para um maior número de agricultores.

2. OBJETIVOS

2.1 Objetivo geral

Delinear zonas de manejo em uma área agrícola comercial utilizando métodos geoestatísticos e de análise multivariada e, avaliar o desempenho de conjuntos específicos de dados para a formação dessas unidades.

2.2 Objetivos específicos

- Delinear zonas de manejo por um agrupamento hierárquico aglomerativo não paramétrico, utilizando diferentes subconjuntos de variáveis formadas por dados referentes aos atributos físicos e químicos do solo, índices vegetativos e produtividade da soja de uma área agrícola comercial;
- Comparar e avaliar os resultados das diferentes zonas de manejo delineadas e verificar qual subconjunto foi mais efetivo na geração daquelas;
- Analisar as diferentes variáveis e avaliar comportamentos e influência entre as zonas de manejo formadas;
- Utilizar um classificador supervisionado para verificar quais variáveis físico-químicas do solo mais influenciaram a geração das zonas de manejo.

3. REVISÃO BIBLIOGRÁFICA

3.1 Agricultura de precisão

A demanda por alimentos aumentou com o crescimento populacional e, consequentemente, o incremento na produção agrícola tornou-se um dos maiores desafios do mundo atual (MOLIN, AMARAL & COLAÇO, 2015). A necessidade do aprimoramento do setor agrícola e da evolução de tecnologias como informática, geoprocessamento, sistemas de posicionamento global, entre outras, fez com que uma nova forma de se enxergar a agricultura fosse requerida. A aplicação dessas tecnologias tornou possível entender uma propriedade como algo não homogêneo, mas sim, composta por diferentes partes que poderiam ser tratadas conforme suas necessidades específicas e, ao mesmo tempo, proporcionar ao produtor maior conhecimento acerca da produtividade de sua área agrícola (TSCHIEDEL & FERREIRA, 2002).

Nesse sentido, a agricultura de precisão pode ser vista como ferramenta essencial para o produtor, pois baseia-se no gerenciamento localizado no campo com aplicação localizada de insumos. Economicamente, essa técnica traz maior retorno por priorizar áreas onde o potencial de produção seja mais efetivo. Do ponto de vista ambiental, a dedução do uso de insumos faz com que haja diminuição dos impactos ao meio ambiente (ANTUNIASSI *et al.*, 2007).

Uma forma de resolver esse problema seria dividir a área total do campo em unidades menores e homogêneas de forma que se possam aplicar insumos a taxas constantes dentro de cada sub-região, mas a partir de taxas diferentes entre elas. A essas unidades chamamos de zonas de manejo (DOERGE, 2000). Com isso, mesmo os pequenos produtores poderiam fazer uso da agricultura de precisão ao utilizarem as máquinas tradicionais e aplicarem somente a quantidade de insumos necessária para cada unidade de manejo identificada (BAZZI *et al.*, 2013).

3.2 Zonas de manejo

Alguns fatores naturais determinantes na produção agrícola sofrem variação espacial e temporal, e isto também refletirá na produção no campo. A base da agricultura de precisão é o conhecimento dessa variação para culminar na realização de um manejo adequado para uma cultura específica do local. Assim, espera-se que ao se realizar uma aplicação diferenciada local de insumos em toda a área, haja diferenças significativas na produtividade. A delimitação do campo em unidades em que um mesmo tratamento pode ser aplicado permite identificar requisitos específicos de insumos agrícolas e melhorar a eficiência e a proteção ao meio ambiente (CORDOBA *et al.*, 2013).

Define-se, então, zona de manejo como sub-regiões nas quais há uma combinação de fatores condicionantes da produtividade para a qual se pode aplicar uma dose uniforme de insumos, ou seja, cada uma dessas zonas é constituída de um conjunto de características semelhantes que podem ser tratadas como uma área homogênea (RODRIGUES JUNIOR *et al.*, 2011).

As zonas de manejo são vistas como uma abordagem economicamente viável, pois os equipamentos e procedimentos adotados são os mesmos utilizados na agricultura convencional já que a aplicação dos insumos em cada zona será feita de maneira homogênea (GAVIOLI *et al.*, 2019). Dessa forma, as zonas de manejo facilitam a aplicação das técnicas de agricultura de precisão, considerando que os mesmos sistemas utilizados na agricultura convencional podem ser aplicados no manejo das culturas (RODRIGUES JUNIOR *et al.*, 2011).

Segundo Santos & Saraiva (2015), quando definidas adequadamente, as zonas de manejo podem orientar os produtores em relação às áreas que possuam maior produtividade e qualidade do cultivo e ainda auxiliar no entendimento de quais são os fatores mais influentes na determinação dessas áreas. O estudo da variabilidade espacial de diversas variáveis é utilizado para investigar quais fatores determinantes estão sendo estudados na variabilidade do local e podem influenciar na produtividade. Depois de definidas as zonas de manejo, é possível gerar mapas de aplicação de insumo a taxas constantes que podem ser alteradas quando ocorre transição de uma zona para outra.

Os benefícios da aplicação de insumos a taxas adequadas às necessidades de cada região ressaltam a importância da tarefa de definição das zonas de manejo. Embora a delimitação dessas zonas seja uma tarefa complexa devido à grande inter-relação entre os atributos do solo e outros limitantes da produtividade, delimitar é de suma importância, pois pode resultar em benefícios para a lavoura e aumentar a lucratividade (SANTOS & SARAIVA, 2015).

3.3 Geração de zonas de manejo

Diversos métodos encontram-se na literatura para geração de zonas de manejo. Alguns autores geralmente dividem esses métodos em duas abordagens: métodos empíricos e métodos de análise de agrupamento. A primeira interpelação é a mais simples e sujeita a decisões subjetivas (FRAISSE *et al.*, 2001). Essa abordagem utiliza distribuição de frequência de produtividade para dividir o talhão geralmente em três ou quatro partes (XIANG, 2007).

A segunda abordagem corresponde a métodos mais complexos com utilização de técnicas de agrupamento que permitem maior grau de diferenciação entre classe por critérios mais objetivos. A interpelação ainda viabiliza a identificação de áreas com atributos

semelhantes a fim de quantificar os padrões de variabilidade e reduzir a natureza empírica no processo de delineamento de zonas de manejo (LI *et al.*, 2007). A utilização de métodos de agrupamentos é bastante sugerida na geração das zonas de manejo, pois permite o uso de um conjunto maior de dados (TAYLOR *et al.*, 2003). Segundo Doerge (2000), o ideal é que se utilize de fontes estáveis e previsíveis de informação espacial e que estejam correlacionadas com a produtividade.

O uso de um conjunto multivariado de dados com propriedades que podem influenciar na produtividade é vantajoso na definição das zonas de manejo. (GUASTAFERRO *et al.*, 2010). As variáveis normalmente utilizadas nesse processo são: dados de produtividade, atributos químicos e físicos do solo, índices de vegetação, características do relevo além da combinação desses conjuntos de variáveis (FRAISSE *et al.*, 2001).

Algumas etapas básicas no delineamento de zonas de manejo são seguidas por diversos pesquisadores, conforme o fluxograma apresentado na Figura 1. A primeira etapa ocorre pela entrada de dados que podem ser obtidos em coleta de campo (coleta direta), imagens de satélites ou fotografias aéreas. Na coleta direta, são obtidos os dados referentes às propriedades físicas e químicas do solo, produtividade, entre outros. Nas coletas indiretas são obtidos elementos referentes ao sensoriamento remoto, como os índices vegetativos, por exemplo.

Os dados podem ser coletados por imagens de satélite, fotografia aérea e radiometria de campo. A produtividade também pode ser obtida por coleta indireta quando estimada a partir da relação com o vigor da cultura determinado por índices vegetativos por sensoriamento remoto a partir de imagens multiespectrais (ARAÚJO, VETTORAZZI & MOLIN, 2005). Os dados de entrada ainda podem ser a combinação entre elementos obtidos por coleta indireta, como o índice NDVI (*Normalized Difference Vegetation Index*) combinado com dados obtidos por coleta direta, tais como as propriedades químicas e físicas do solo e produtividade (LI *et al.*, 2007; SALVADOR & ANTUNIASSI, 2011; CICORE *et al.*, 2016).

O passo seguinte é selecionar quais dessas variáveis serão utilizadas na pesquisa. Para isso, pode-se utilizar a dependência espacial nesse processo e, para tal, aplica-se o cálculo da dependência espacial relativa (EPR) na determinação de quais variáveis farão parte do processo e quais serão excluídas. O uso da dependência espacial no processo de seleção de variáveis para a geração de zonas de manejo também pode visto em trabalhos como o realizado por Alves *et al.* (2013), no qual definiram-se zonas de manejo com base na variabilidade espacial da condutividade elétrica aparente do solo e da matéria orgânica em áreas com plantio de soja e milho.

Recorrendo às técnicas geoestatísticas, mediu-se o grau de dependência espacial das variáveis, as que se apresentaram dependentes, foram utilizadas nas etapas seguintes

no processo de geração de zonas de manejo. Em seguida, foi utilizado um método de agrupamento para delinear as zonas de manejo. Os métodos de agrupamentos são divididos em dois grupos: os de aprendizagem supervisionada e os de aprendizagem não supervisionada (GOLDSCHMIDT, PASSOS & BEZERRA, 2015).

O método supervisionado se utiliza de um conjunto de dados do qual se conhece a categoria de algumas amostras e se constrói um modelo (regra) de classificação. Utiliza-se dessa regra para identificar as categorias que pertencem às novas amostras. O método não supervisionado agrupa as amostras em um número desconhecido de grupos por padrões associados às relações entre as variáveis. Esse tipo de agrupamento identifica grupos naturais das observações em um conjunto de dados (FERREIRA, 2011).

A avaliação desses agrupamentos é realizada com base nos grupos formados a partir de índices escolhidos de acordo com a metodologia utilizada na geração dos agrupamentos. Por fim, são gerados os mapas temáticos que expressam as zonas de manejo.

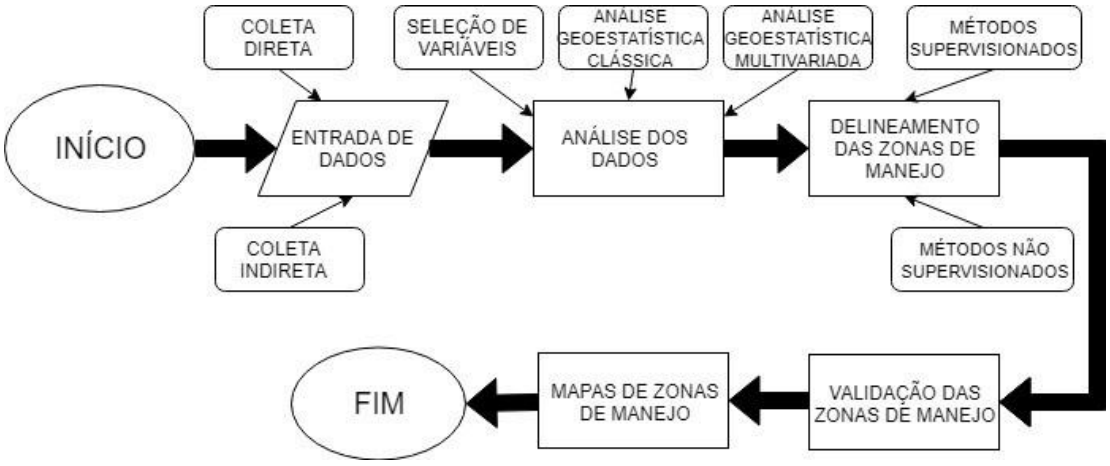

Figura 1 Fluxograma para delineamento de zonas de manejo (O autor, 2019)

3.3.1 Dados de entrada

Os dados utilizados na geração de zonas de manejo estão diretamente ligados ao geoprocessamento, assim, entende-se que, de alguma forma, esses dados estão vinculados a um lugar no espaço, por intermédio de um endereço ou de coordenadas. A coleta de informações ocorre de várias formas tais como: dados cartográficos (mapas), sensoriamento remoto (imagens de satélites e radares), fotogrametria (fotografias aéreas), topografia (levantamentos topográficos e geodésicos), dados alfanuméricos (tabelas) e a coleta direta (com dados georreferenciados) (SILVA, 2007).

Os dados oriundos de sensoriamento remoto podem fornecer informações mais precisas da variabilidade do campo em grandes áreas por apresentarem potencial para monitoramento de parâmetros biofísicos ligados à produtividade. Esse fato revela o sensoriamento remoto como alternativa para elaboração de mapas de produtividade (MOTOMIYA, *et al.,* 2012). A produtividade pode ser estimada mediante índices de vegetação, calculados com base em informações geradas a partir de imagens multiespectrais (ARAUJO, VETTORAZZI & MOLIN, 2005).

Os índices de vegetação são o resultado de combinações lineares de dados espectrais, ou seja, são operações algébricas que envolvem faixas de reflectância especificas e que permitem a determinação da cobertura vegetal e a sua densidade (CRUZ *et al.,* 2011). São, portanto, modelos matemáticos determinados por transformações de duas ou mais bandas espectrais, em geral, a do vermelho e a do infravermelho próximo, calculadas pela razão, diferença e combinação linear desses dados (BANNARI *et al.*, 1995).

Esses índices são indicadores de crescimento e vigor da vegetação e podem ser utilizados para diagnosticar vários parâmetros biofísicos tais como de: área foliar, biomassa, porcentagem de cobertura do solo, atividade fotossintética e produtividade (PONZONI, 2001). Muitos índices de vegetação são encontrados na literatura e, na Tabela 1, são apresentados alguns desses índices com as respectivas expressões matemáticas.

Tabela 1 Índices de vegetação

Índice	Expressão	Referência
NDVI	$NDVI = \frac{NIR - RED}{NIR + RED}$	ROUSE *et al.* (1973)
EVI2	$EVI2 = 2{,}5\frac{NIR - RED}{(NIR + \eta_1\ RED - \eta_2\ BLUE + L)}$	HUETE *et al.* (1997)
SAVI	$SAVI = \frac{NIR - RED}{(NIR + RED + L)}(1 + L)$	HUETE (1988)
OSAVI	$OSAVI = \frac{NIR - GREEN}{NIR + GREEN + 0{,}16}$	STEVEN (1998)
ARVI	$ARVI = \frac{NIR - \left(RED - \gamma(BLUE - RED)\right)}{NIR + \left(RED - \gamma(BLUE - RED)\right)}$	KAUFAMAN & TANRÉ (1992)
WDRVI	$WDRVI = \frac{\iota\ NIR - RED}{\iota\ NIR + RED}$	GITELSON (2004)

NIR: refletância da banda infravermelho próximo; RED: refletância da banda vermelha; BLUE: refletância da banda azul; GREEN: refletância da banda verde; η_1: coeficiente de correção dos efeitos atmosféricos para a banda vermelha; η_2: coeficiente de correção dos efeitos atmosféricos para a banda azul; L: fator de correção para a interferência do solo; γ: parâmetro que controla a correção atmosférica; ι: coeficiente de ponderação

3.3.2 Redução da dimensionalidade

3.3.2.1 Análise de componentes principais

A análise de componentes principais (ACP) também é uma técnica muito utilizada na agricultura de precisão e na geração de zonas de manejo. Os trabalhos realizados por Delalibera *et al.* (2012), Santi, *et al.* (2012), Cordoba *et al.* (2012), Tripathi *et al.* (2015), Behera *et al.* (2018), Reyes *et al.* (2019) ilustram a utilização desta ferramenta na seleção de variáveis.

A ACP é uma técnica multivariada da estrutura de covariância descrita inicialmente por Pearson e posteriormente por Hotelling tendo como propósito analisar a estrutura de correlação. Esta técnica transforma linearmente um conjunto original de variáveis correlacionadas em outro conjunto menor de variáveis não correlacionadas contendo a maior parte da informação do conjunto inicial (HONGYU *et al.*, 2015).

De acordo com esta técnica, o conjunto de dados consistirá em uma tabela na qual cada linha corresponderá aos valores das variáveis medidas sobre o indivíduo análogo a essa linha, ou seja, as linhas serão correspondentes aos indivíduos ou observações enquanto as colunas serão equivalentes às variáveis. Essa tabela consistirá, então, em uma matriz $\boldsymbol{X} = [x_{ij}]$, onde i,($i = 1, \dots, n$), é *i*-ésimo indivíduo, j,($j = 1, \dots, m$), é a *j*-ésima variável e x_{ij} é o valor da variável j no indivíduo i (FERNANDEZ & YOHAI, 2014).

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

Assim, a determinação da componente principal (CP) demanda primeiramente o cálculo da matriz de covariância, em seguida dos autovalores e dos autovetores da matriz de covariância ou de correlação (usada quando existe uma escala diferente de valores das variáveis), ordenando os autovalores de forma decrescente. Cada CP, então, será uma combinação linear de todas as variáveis originais estimadas com intuito de reter a máxima variância contida nos dados originais (MOURA *et al.*, 2018).

$$CP_i = e_{i1}X_{i1} + e_{i2}X_{i2} + \cdots + e_{im}X_{im}, \quad i = 1, \dots, n; \;\; j = 1, \dots, m \tag{1}$$

Em que, CP_i representa o *i*-ésimo componente principal (CP), as constantes e_{i1}, e_{i2}, ..., e_{im} são os elementos dos autovetores correspondentes aos autovetores ordenados, também chamados de coeficientes da componente principal (SANTO, 2012). Para cada CP, as variáveis com maior relevância são as que possuem os maiores coeficientes. As variáveis com os coeficientes positivos indicam uma influência direta dessas variáveis sobre o

componente, enquanto os coeficientes negativos exercem influência inversa (JOHNSON & WICHERN, 2007).

A contribuição de cada CP é expressa por porcentagem, sendo possível calcular individualmente o quanto cada componente explica da variação total do conjunto de dados. Por exemplo, para calcular o valor da contribuição do $i-ésimo$ componente utiliza-se a expressão (JOHNSON & WICHERN, 2007):

$$C_i = \frac{Var(Z_i)}{\sum_{i=1}^{j} Var(Z_i)} * 100 = \frac{\lambda_i}{\sum_{i=1}^{j} \lambda_i} * 100 \quad (2)$$

Em que Z_i é a variância de cada componente principal e λ_i são os autovalores associados a cada componente.

Na prática, ao aplicar a ACP, é comum a aplicação de poucas componentes no lugar das m variáveis originais, sem perda significativa de informação. Entre os critérios utilizados para selecionar as CPs em uma análise, encontra-se a descrita por Johnson & Wichern (2007), na qual devem ser consideradas as primeiras CPs que unidas expliquem ao menos 80% da variabilidade total das variáveis originais.

Além desse método acima descrito, o critério de Kaiser é muito utilizado para seleção das CPs. Por esse critério devem ser selecionados as CPs que possuam seus respectivos autovalores maiores que um $\lambda_i > 1$/ (HONGYU *et al.*, 2015).

3.3.2.2 Multispati-PCA

É possível reduzir a dimensão do conjunto de variáveis originais com a análise dos componentes principais. Entretanto, existe uma desvantagem no uso desta técnica quando se deseja trabalhar com variáveis espacialmente relacionadas, pois, não há uma diferenciação da estrutura espacial entre as componentes. Uma alternativa seria utilizar a técnica MULTISPATI-PCA, em que uma matriz de ponderação espacial é utilizada. Assim, adiciona-se uma restrição espacial à técnica de PCA tradicional e, a partir de então, as estruturas espaciais serão mais fortes nos primeiros componentes (MORAL & REBOLLO, 2017).

Tal análise permite estudar as relações entre as variáveis medidas e sua estrutura espacial pelo uso da matriz de ponderação espacial $\boldsymbol{W}_{nxn}$. A matriz pode ser considerada como uma representação matemática da distribuição geográficas dos pontos em estudos, pois os pesos atribuídos a cada ponto indicam ausência $(w_{ij} = 0)$ ou a intensidade $(w_{ij} > 0)$ das relações espaciais existentes entre os pontos da área em estudo (CORDOBA *et al.*, 2012).

Diferente do método de ACP tradicional, que é gerado sobre uma matriz $\boldsymbol{X}_{nxp}$ chamada de tabela de dados, o MULTISPATI-PCA tem como matriz de entrada a matriz

$\boldsymbol{Y}_{nxp}$, que é o produto da matriz $\boldsymbol{W}_{nxn}$ com a matriz $\boldsymbol{X}_{nxp}$, ou seja, $\boldsymbol{Y} = \boldsymbol{WX}$. Desta forma, cada elemento da matriz $\boldsymbol{Y}$ substitui um elemento da mesma posição na matriz $\boldsymbol{X}$ por um novo valor correspondente ao valor de uma variável p em um ponto amostral i (ARROUAYS *et al.*, 2011).

As variâncias explicadas pelas CPs apresentam diferenças entre as duas técnicas. Para o PCA, a variância de cada componente é igual ao seu autovalor associado enquanto para o MULTISPATI-PCA, os autovalores associados são equivalentes à variância espacial e não à variância total (CORDOBA *et al.*, 2012).

A vantagem da MULTISPATI-PCA em relação à ACP é que os escores gerados maximizam a autocorrelação espacial entre os pontos enquanto a ACP maximiza a variância total. Assim, os escores gerados pela MULTISPATI-PCA apresentam estruturas espaciais mais fortes nos primeiros CPs, enquanto os escores gerados pela PCA apresentaram estrutura espacial em qualquer CP, inclusive nos últimos que são geralmente descartados (ARROUAYS *et al.*, 2011).

O uso desta técnica na geração de zonas de manejo pode ser observado nos trabalhos publicados por Cordoba *et al.* (2012) (2013) (2016), Peralta *et al.* (2015), Gavioli *et al.* (2016) e Kurina *et al.* (2018). Nestes trabalhos, concluiu-se que a MULTISPATI-PCA representa uma ferramenta crucial para mapear a variabilidade dentro do campo e identificar zonas de manejo. O uso dessa ferramenta permite que, ao incorporar a relação espacial dos dados, haja um ganho de informação que, por sua vez, faz com que as zonas geradas apresentem maiores diferenças entre as distintas classes e sejam mais homogêneas dentro de cada classe gerada.

3.3.3 Análise geoestatística

O uso da análise geoestatística torna possível organizar dados distribuídos espacialmente de acordo com a semelhança entre seus vizinhos. Essa análise traz grande contribuição especialmente na definição de unidades de manejo a partir de mapas de produtividade (GREGO, OLIVEIRA & VIEIRA, 2014).

Uma importante etapa na geração de zonas de manejo é o uso de técnicas geoestatísticas para caracterizar a distribuição espacial das variáveis em estudo. A geoestatística pode ser vista, então, como uma ferramenta estatística que é utilizada para o estudo da variabilidade espacial (VIEIRA *et al.*, 2002). O seu uso permite que a estrutura da dependência espacial seja modelada e visualizada espacialmente (MENDES et al., 2008). Para isso, ela se apoia na hipótese de que dados vizinhos são mais semelhantes do que os dados distantes. A função semivariância pode ser utilizada para medir o grau de semelhança (GREGO, OLIVEIRA & VIEIRA, 2014). E, para o cálculo da semivariância,

pode-se utilizar, entre outros estimadores, o de Matheron, descrito por JOURNEL & HUIJBREGTS (2003), apresentado na equação abaixo:

$$\hat{\gamma}(\boldsymbol{h}) = \frac{1}{2N(\boldsymbol{h})} \sum_{i=1}^{N(\boldsymbol{h})} [Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i + \boldsymbol{h})]^2 \qquad (3)$$

Em que N($\boldsymbol{h}$) é o número de pares de valores medidos de dissimilaridade entre Z($\boldsymbol{s}_i$) e Z($\boldsymbol{s}_i$+ $\boldsymbol{h}$) separados por um vetor $\boldsymbol{h}$, sendo Z($\boldsymbol{s}_i$+ $\boldsymbol{h}$) o valor da variável regionalizada na amostra localizada em ($\boldsymbol{s}_i$+ $\boldsymbol{h}$), Z($\boldsymbol{s}_i$) é o valor da variável regionalizada na amostra localizada em ($\boldsymbol{s}_i$), e $h = \|\boldsymbol{h}\|$ é a distância euclidiana entre as duas localizações (JOURNEL & HUIJBREGTS, 2003).

Se o semivariograma gerado crescer conforme se aumenta a distância, significa que a variável avaliada possui dependência espacial. Além disso, é necessário também que, após o aumento, o semivariograma se estabilize no valor correspondente à variância dos dados. A maneira como o semivariograma cresce com a distância até se estabilizar define o comportamento espacial da variável estudada (VIEIRA, 2000).

Depois de verificados indícios da existência de dependência espacial, por intermédio da construção do semivariograma experimental, existe o interesse em se ajustar um modelo que descreva a semivariância em função da distância. De acordo com Isaaks & Srivastava (1989) e Cressie (1993), os modelos teóricos são divididos em transitivos, aqueles que possuem patamar, e não transitivos os que não possuem patamar. Os transitivos mais utilizados na Geoestatística são: exponencial, gaussiano e família *Matérn*. Para realizar o ajuste do modelo ao semivariograma um dos métodos mais usados é o da máxima verossimilhança (MELLO *et al.*, 2005).

Após estimar os parâmetros do modelo, é necessário observar os critérios de validação dos ajustes dos modelos para selecionar o modelo que melhor se ajusta ao semivariograma gerado. Entre os métodos de validação de modelos, os mais usuais são aqueles cujas técnicas são baseadas na comparação entre os valores teóricos de modelos geoestatísticos e valores empíricos (FARACO *et al.*, 2008).

Após a escolha do modelo que melhor se ajustou ao semivariograma, procede-se a estimativa de valores para locais não amostrados a partir da técnica de interpolação denominada krigagem ordinária. Com o uso da krigagem é possível gerar mapas da área de interesse (GONÇALVEZ, FOLEGATTI & MATA, 2001).

Tais mapas permitem descrever a variabilidade de atributos de interesse e utilizar essa informação como base para análise da produtividade agrícola (AMADO *et al*., 2007). Assim, a geoestatística auxilia efetivamente nas decisões em relação ao gerenciamento do

sistema de produção e contribui para a agricultura de precisão (GREGO, OLIVEIRA & VIEIRA, 2014).

3.3.3.1 Krigagem

Na literatura, diversos trabalhos defendem que a krigagem é utilizada como interpolador no delineamento de zonas de manejo. Como exemplo, pode-se verificar o trabalho desenvolvido por Oldoni & Bassoi (2016), no qual o objetivo foi delinear zonas de manejo de irrigação utilizando geoestatística e análise multivariada. A krigagem foi utilizada na interpolação de dados em locais não amostrados para obter um número mais denso de pontos por área e para posterior processo de agrupamento.

No artigo desenvolvido por Gavioli *et al.* (2019), foram testadas as seguintes técnicas de interpolação: krigagem ordinária, inverso da distância e inverso da distância ao quadrado. O objetivo era a geração de zonas de manejo com contornos mais suaves ao utilizarem-se os dois primeiros componentes principais espaciais (SPCs) para cada campo testado. A krigagem foi o interpolador que apresentou o melhor resultado e passou a ser utilizada para todos os outros SPCs considerados no trabalho.

A krigagem é baseada no conceito de variável regionalizada, desenvolvida por George Matheron. Trata-se de um método de interpolação geoestatístico que utiliza a estrutura de dependência espacial para estimar valores de variáveis para locais não amostrados, a partir de valores adjacentes conhecidos. Ela é determinada pelo ajuste de um modelo teórico à função semivariância (URIBE-OPAZO *et al.*, 2012). Este interpolador é considerado superior aos outros métodos de interpolação por uma característica que o diferencia dos demais que é o fato de permitir calcular o erro associado às estimativas, chamado de variância de estimação (MARTINS, 2017).

Na krigagem, considera-se que determinado ponto localizado no espaço se assemelha mais com os pontos que estão ao seu entorno do que aqueles que estão mais afastados. Ela se utiliza de métodos matemáticos para determinar maiores pesos para os pontos mais próximos e menores para os mais distantes (ISAAKS & SRIVASTAVA, 1989). Esses valores variam de 0, para os mais distantes, a 1 para os mais próximos, e assim permitem a interpolação de valores para locais não amostrados com bases em combinações lineares (MARTINS, 2017).

3.3.3.2 Krigagem ordinária

A krigagem ordinária baseia-se na suposição de que a variação é aleatória e de que há dependência espacial nos dados em análise. Além disso, o processo aleatório é

intrinsecamente estacionário, ou seja, a média é constante e a variância depende somente da distância e direção entre os pontos e não da sua posição absoluta (OLIVER & WEBSTER, 2015).

Assim sendo, considere $Z = \{Z(s), s \in D\}$ um processo aleatório estacionário de segunda ordem com média constante e desconhecida μ e uma função de covariância conhecida $C(h)$, sendo h a distância entre os pontos considerados. De acordo com essa suposição, as equações que determinarão os pesos necessários para a predição pela krigagem ordinária podem ser expressas em termos da função de covariância ou da semivariância. Em qualquer dos casos, $Z = \{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ é predito em um ponto não observado s_0 usando o preditor linear e, espera-se que o erro na predição seja zero e sua variância seja mínima (MONTERO *et al.*, 2015):

$$Z^*(\boldsymbol{s}_0) = \sum_{i=1}^{n} \lambda_i \, Z(\boldsymbol{s}_i), \qquad i = 1, \dots, n \tag{4}$$

Em que $Z^*(\boldsymbol{s}_0)$ é o valor a ser estimado de um ponto não amostrado; $\boldsymbol{s}_0$, $Z(\boldsymbol{s}_i)$ são os valores das variáveis nos pontos amostrados; λ_i são os pesos associados a cada valor $Z(\boldsymbol{s}_i)$ medido. Assim, para garantir que a estimativa não seja enviesada, a soma dos pesos deve ser igual a um (OLIVER & WEBSTER, 2015):

$$\begin{aligned} E\big(Z^*(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)\big) &= E\left(\sum_{i=1}^{n} \lambda_i Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_0)\right) = \sum_{i=1}^{n} \lambda_i \, E\big(Z(\boldsymbol{s}_i)\big) - E\big(Z(\boldsymbol{s}_0)\big) \\ &= \mu \sum_{i=1}^{n} \lambda_i - \mu = 0 \leftrightarrow \sum_{i=1}^{n} \lambda_i = 1 \end{aligned} \tag{5}$$

A diferença esperada entre o valor estimado e o valor amostrado deve ser igual a zero:

$$E[Z(\boldsymbol{s_0}) - Z^*(\boldsymbol{s_0})] = 0 \tag{6}$$

Para que isso ocorra, é organizado um sistema de equações com *n+1* incógnitas para se estimar um ponto s_0 utilizando-se os pesos λ_i combinados com os valores observados s_i (LANDIM, 2006). Estas equações constituem equações normais a *n+1* incógnitas que podem ser resolvidas por cálculo matricial, segundo:

$$[\mathbf{s}_i, \mathbf{s}_i][\lambda_i] = [\mathbf{s}_i, \mathbf{s}_0]$$

Multiplicando-se ambos os termos da equação pelo inverso de $[\mathbf{s}_i, \mathbf{s}_i]$, isto é, $[\mathbf{s}_i, \mathbf{s}_i]^{-1}$:

$$[\mathbf{s}_i, \mathbf{s}_i]^{-1}.[\mathbf{s}_i, \mathbf{s}_i].[\lambda_i] = [\mathbf{s}_i, \mathbf{s}_i]^{-1}.[\mathbf{s}_i, \mathbf{s}_0]$$

Como $[\mathbf{s}_\mathrm{i}, \mathbf{s}_\mathrm{i}]^{-1}.[\mathbf{s}_\mathrm{i}, \mathbf{s}_\mathrm{i}] = [\mathrm{I}]$ (matriz identidade) e $[I].[\lambda_\mathrm{i}] = [\lambda_\mathrm{i}]$,

$$[\lambda_\mathrm{i}] = [\mathbf{s}_\mathrm{i}, \mathbf{s}_\mathrm{i}]^{-1}.[\mathbf{s}_\mathrm{i}, \mathbf{s}_0]$$

Na matriz$[\mathbf{s}_\mathrm{i}, \mathbf{s}_\mathrm{i}]$**,** com dimensão **(***n+1)* x *(n+1)*, estão os valores obtidos da semivariância referentes às distâncias entre as amostras observadas, $\Upsilon(\mathbf{s}_\mathrm{i}, \mathbf{s}_\mathrm{i})$; no vetor $[\mathbf{s}_\mathrm{i}, \mathbf{s}_0]$, com dimensão*n+1* x *1,* estão os valores obtidos da semivariância referentes às distâncias entre cada amostra e o ponto a ser estimado, $\Upsilon(\mathbf{s}_\mathrm{i}, \mathbf{s}_0)$ e o vetor $[\lambda_\mathrm{i}]$, com dimensão **(***n+1)* x *1,* contêm os pesos a serem calculados e o multiplicador de Lagrange α, utilizado para balancear o sistema de equações (LANDIM, 2006).

Em notação matricial:

$$\begin{bmatrix} \Upsilon(\mathbf{s}_1, \mathbf{s}_1) & \Upsilon(\mathbf{s}_1, \mathbf{s}_2) & \dots & \Upsilon(\mathbf{s}_1, \mathbf{s}_\mathrm{n}) & 1 \\ \Upsilon(\mathbf{s}_2, \mathbf{s}_1) & \Upsilon(\mathbf{s}_2, \mathbf{s}_2) & \dots & \Upsilon(\mathbf{s}_2, \mathbf{s}_\mathrm{n}) & 1 \\ \vdots & \vdots & & \vdots\ \vdots\ \vdots & \\ \Upsilon(\mathbf{s}_\mathrm{n}, \mathbf{s}_1) & \Upsilon(\mathbf{s}_\mathrm{n}, \mathbf{s}_2) & \dots & \Upsilon(\mathbf{s}_\mathrm{n}, \mathbf{s}_\mathrm{n}) & 1 \\ 1 & 1 & \dots & 1 & 1 \end{bmatrix}_{n\mathrm{x}n} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_\mathrm{n} \\ \alpha \end{bmatrix}_{n\mathrm{x}1} \quad \begin{bmatrix} \Upsilon(\mathbf{s}_1, \mathbf{s}_0) \\ \Upsilon(\mathbf{s}_2, \mathbf{s}_0) \\ \vdots \\ \Upsilon(\mathbf{s}_\mathrm{n}, \mathbf{s}_0) \\ 1 \end{bmatrix}_{n\mathrm{x}1}$$

O erro associado ao estimador ε = $Z(\boldsymbol{s_0}) - Z^*(\boldsymbol{s_0})$ deve apresentar valores próximos a zero para que o estimador seja considerado de confiança e isso pode ser verificado pela distribuição desses valores. A maneira mais simples de se medir estatisticamente tal distribuição é por desvio padrão ou por variância (CAMARGO, 1998).

A correspondente variância minimizada do erro, denominada variância de krigagem ordinária (σ^2_{ko}), é dada pela expressão (MONTERO *et al.*, 2015):

$$\sigma^2_{\mathrm{ko}} = \mathrm{Var}[Z(\boldsymbol{s}_0) - Z^*(\boldsymbol{s}_0)] = C(0) - \sum_{i=1}^{n} \lambda_i C(\boldsymbol{s}_i, \boldsymbol{s}_0) - \alpha \qquad (07)$$

Em que, $C(\boldsymbol{s}_i, \boldsymbol{s}_0)$ é a semivariância entre os pontos $\boldsymbol{s}_i$e $\boldsymbol{s}_0$ e α é o multiplicador de Lagrange necessário para a minimização da variância do erro.

A krigagem ordinária é considerada um interpolador exato no sentido de que, quando um ponto alvo, ou seja, um ponto com valor gerado pelas equações acima, for um ponto já amostrado, isto é, seu valor já conhecido, os valores de ambos irão coincidir (OLIVER & WEBSTER, 2015).

3.3.4 Métodos não paramétricos de análise de dados

As zonas de manejos podem ser geradas seguindo abordagem de análise de dados, tanto paramétricas quanto não paramétricas. Na abordagem paramétrica assume-se que as

distribuições de probabilidades sejam conhecidas, o que não ocorre na abordagem não paramétrica (SILVERMAN, 1986). Segundo Vapnik (2000), é necessário o conhecimento da relação estrutural dos dados, pois eles fornecem informações iniciais importantes sobre os mesmos dados.

Assim, em abordagens não paramétricas, aplicam-se estimadores de densidades não-paramétricos, mais particularmente dos estimadores de densidade por *kernel* (KDE – *Kernel Density Estimador*) também conhecidos como função núcleo-estimador. Essa função objetiva estimar a densidade com base em informações locais e não estimar parâmetros globais para modelos de dados (CASELLA & BERGER, 2011).

Seja $\boldsymbol{X_1}, \dots, \boldsymbol{X_n}$ uma amostra de vetores aleatórios p-variados originários de uma distribuição comum descrita pela função de densidade conjunta f. A função núcleo-estimadora de densidade conjunta do vetor aleatório $\boldsymbol{x}$ é definida por:

$$\hat{f}_{n\Lambda}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\Lambda}(\boldsymbol{x} - \boldsymbol{X_i}) \qquad (08)$$

Em que $\boldsymbol{x} = \left(x_1, \dots, x_p\right)^T$ são os pontos onde a função núcleo-estimadora é definida; $\boldsymbol{X_i} = \left(X_{i1}, \dots, X_{ip}\right)^T, i = 1, \dots, n$ são variáveis aleatórias com densidade f desconhecida; $\boldsymbol{\Lambda}$ é a matriz de dimensão pxp, que representa a largura de banda ou suavização, sendo que $\boldsymbol{\Lambda}$ é simétrica e positiva definida; K é a função núcleo-estimador, que é uma densidade multivariada simétrica (SIMONOFF, 1996).

Deve-se ter um cuidado especial na escolha da matriz de largura de banda $\boldsymbol{\Lambda}$, pois, essa desempenha o papel de matriz de covariância, sendo o fator mais importante em relação à precisão da estimação (CHACÓN & DUONG, 2010).

3.3.4.1 Medida de dissimilaridade considerando a estrutura de dependência espacial e função de densidade não paramétrica

Considere um conjunto com p variáveis de interesse$\{Z_1, \dots, Z_p\}$ padronizadas, definidas em um domínio de estudo contínuo fixo $S \subset \mathbb{R}^d, d \geq 1$ e todas as medidas realizadas em pontos distintas $\{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n\}$. Uma função núcleo-estimador não paramétrica da estrutura de dependência espacial multivariada é descrita pelas semivariâncias diretas e cruzadas, em dois locais $\boldsymbol{x} \in S$ e $\boldsymbol{y} \in S$ é dado pela Equação (09) (FOUEDJIO, 2016):

$$\hat{\gamma}_{ij}(\boldsymbol{x}, \boldsymbol{y}, \lambda) = \frac{\sum_{k,l=1}^{n} K_{\lambda}\big((\boldsymbol{x}, \boldsymbol{y})(\boldsymbol{s}_k, \boldsymbol{s}_l)\big)\big(Z_i(\boldsymbol{s}_k) - Z_i(\boldsymbol{s}_l)\big)\left(Z_j(\boldsymbol{s}_k) - Z_j(\boldsymbol{s}_l)\right)}{2\sum_{k,l=1}^{n} K_{\lambda}\big((\boldsymbol{x}, \boldsymbol{y})(\boldsymbol{s}_k, \boldsymbol{s}_l)\big)} \mathbb{1}_{x \neq y} \qquad (09)$$

Em que $\hat{\gamma}_{ij}(\boldsymbol{x}, \boldsymbol{y}, \lambda)$ é o valor estimado da semivariância na amostra, tanto para os valores das semivariâncias diretas quanto cruzadas. Quando $i = j$, $\hat{\gamma}_{ij}$ apresenta valor da

semivariância direta para Z_i e quando $i \neq j$, $\hat{\gamma}_{ij}$ apresenta valor para semivariância cruzada entre Z_i e Z_j. Os índices i e j variam de 1 a p e representam as variáveis consideradas no estudo. $K_\lambda\big((\boldsymbol{x},\boldsymbol{y})(\boldsymbol{s}_k,\boldsymbol{s}_l)\big) = K_\lambda(||\boldsymbol{x}-\boldsymbol{s}_k||)K_\lambda(||\boldsymbol{y}-\boldsymbol{s}_l||)$, em que $K_\lambda(.)$ sendo uma função núcleo-estimador não negativa com parâmetro de largura de banda $\lambda > 0$ e $\mathbb{1}$ representa a função indicadora dada por (CASELLA & BERGER, 2011):

$$\mathbb{1}_{||\boldsymbol{x}-\boldsymbol{s}||<\lambda} = \begin{cases} 1, ||\boldsymbol{x}-\boldsymbol{s}|| \leq \boldsymbol{\lambda} \\ 0, ||\boldsymbol{x}-\boldsymbol{s}|| > \lambda \end{cases}$$

A função núcleo-estimador para os valores da semivariância direta e cruzada descrita na Equação (09) serve para qualquer par de observações dentro da área amostral. O objetivo dessa função é ponderar as localizações de forma que as que estiverem mais próximas do alvo receberão maior peso do que os dados em localizações mais distantes (FOUEDJIO, 2016).

Para determinar o parâmetro de largura de banda segue-se uma regra empírica de que a função núcleo-estimadora centrada em cada localização contenha ao menos 35 observações (JOURNEL & HUIJIBREGTD, 2003).

A não estacionariedade de segunda ordem nos dados pode ser bem capturada por este tipo de estimador de função de densidade (KLEIBER & NICHKA, 2012). Alguns outros tipos de função núcleo-estimador podem ser utilizados, tais como as funções Uniformes, Triangular, Epanechnikov e Gaussiana (THOMPSON & TAPIA, 1990). A função Epanechnikov pode ser descrita como:

- Epanechnikov:$K_\lambda(\|\boldsymbol{x}-\boldsymbol{s}\|)\ \alpha\ (\lambda^2 - \|\boldsymbol{x}-\boldsymbol{s}\|^2)\mathbb{1}_{\|\boldsymbol{x}-\boldsymbol{s}\|\leq\lambda}$.

Na função núcleo-estimador de Epanechnikov há pesos maiores para observações $\boldsymbol{x}$ e $\boldsymbol{s}$ mais próximos entre si. Assim, pontos $\boldsymbol{x}$ e $\boldsymbol{s}$ mais próximos entre si implicam em uma norma $\|\boldsymbol{x}-\boldsymbol{s}\|$ menor e, por consequência, a diferença $\lambda - \|\boldsymbol{x}-\boldsymbol{s}\|$ será maior. Dessa forma, mesmo considerando pontos tanto próximos quanto distantes no cálculo de $\hat{\gamma}_{ij}$, o valor de$\hat{\gamma}_{ij}$ será mais influenciado pelos pontos mais próximos.

Uma medida de dissimilaridade entre dois locais, $\boldsymbol{s}_k$ e $\boldsymbol{s}_l$ pode ser denotada por $d_\lambda(\boldsymbol{s}_k,\boldsymbol{s}_l)$ com o conjunto de valores estimados das semivariâncias diretas e cruzadas, a qual foi obtida de acordo com a seguinte expressão (THEODORIDIS & KOUTROUMBAS, 2009):

$$d_\lambda(\boldsymbol{s}_k,\boldsymbol{s}_l) = \sum_{i,j=1}^{p} |\hat{\gamma}_{ij}(\boldsymbol{s}_k,\boldsymbol{s}_l,\lambda)| \qquad (10)$$

A dissimilaridade entre esses dois locais podem ser definida, então, como a soma dos valores absolutos de todos os valores das semivariâncias diretas e cruzadas nesses dois pontos observados. A normalização dessa dissimilaridade é dada efetuando-se a

divisão de cada valor de dissimilaridade pela maior delas encontrada, conforme a seguinte expressão:

$$\tilde{d}_\lambda(\boldsymbol{s}_k, \boldsymbol{s}_l) = \frac{1}{D} d_\lambda(\boldsymbol{s}_k, \boldsymbol{s}_l) \tag{11}$$

Em que $D = máx_{(k,l)\in\{1,\dots,n\}^2}\{d_\lambda(\boldsymbol{s}_k, \boldsymbol{s}_l)\}$

A partir dessa medida de dissimilaridade normalizada, gera-se a matriz de dissimilaridade simétrica $\boldsymbol{D}_{nxn}$ para os locais amostrados (FOUEDJIO, 2016):

$$\boldsymbol{D} = \begin{bmatrix} \tilde{d}_\lambda(\boldsymbol{s}_1, \boldsymbol{s}_1) & \dots & \tilde{d}_\lambda(\boldsymbol{s}_1, \boldsymbol{s}_n) \\ \vdots & \ddots & \vdots \\ \tilde{d}_\lambda(\boldsymbol{s}_n, \boldsymbol{s}_1) & \dots & \tilde{d}_\lambda(\boldsymbol{s}_n, \boldsymbol{s}_n) \end{bmatrix} nxn$$

3.3.5 Métodos de agrupamentos, classificação e avaliação de classes

3.3.5.1 Método de agrupamento de dados

Os agrupamentos podem ser definidos como técnicas multivariadas que têm por finalidade principal, agregar variáveis com base em suas características (HAIR *et al.*, 2009). Para isso, as técnicas envolvidas têm por objetivo explorar os conjuntos de dados para avaliar se eles podem ou não ser resumidos significativamente em pequenos grupos ou em grupos de elementos que se assemelham entre si e, ao mesmo tempo, diferem em alguns aspectos de elementos de outros agrupamentos (EVERITT *et al.*, 2011).

Os dados em estudo são classificados, então, em diferentes combinações de variáveis, em classes discretas ou clusters, de maneira que a distância entre o centroide (centro do agrupamento) e a variáveis são minimizadas. Assim, essa técnica determinará o grau de similaridade de uma amostra a uma determinada classe (BAZZI *et al.*, 2015).

Muitos trabalhos publicados em análise de agrupamento envolvem em geral o uso de uma das duas classes de agrupamentos: algoritmos hierárquicos ou não hierárquicos (particionamento) (HENNIG *et al.*, 2016). Assim, essas classes de agrupamento serão discutidas com maiores detalhes, nos subcapítulos adiante.

3.3.5.2 Método de agrupamento hierárquico

No agrupamento hierárquico, os dados são particionados em uma série de divisões. As técnicas utilizadas nessa forma de agrupamento podem ser subdivididas em aglomerativo, na qual ocorrem fusões sucessivas dos n indivíduos em grupos ou divisivos que irão separar os n indivíduos sucessivamente em agrupamentos mais refinados. As

técnicas aglomerativas reduzirão o conjunto de dados em único cluster contendo todos os indivíduos no final do processo, enquanto as técnicas divisivas, de modo contrário ao aglomerativo, irão dividir todo o conjunto de dados formando n grupos com apenas um indivíduo no final do processo. Dessa forma, o pesquisador que decidirá em qual momento do agrupamento deverá parar para se ter um número de cluster com o qual se deseja trabalhar (EVERITT *et al.*, 2011).

As classificações hierárquicas podem ser representadas por um diagrama bidimensional também chamado de dendrograma (Figura 2), que permite a visualização das fusões ou divisões feitas em cada etapa da análise. O dendrograma é uma representação matemática do procedimento de agrupamento em que seus nós representam os clusters e os comprimentos das hastes (alturas) representam as distâncias nas quais esses clusters são unidos. Na maioria dos dendrogramas, duas bordas saem de cada nó (árvores binárias), e o arranjo de nós e hastes é chamado de topologia da árvore (EVERITT *et al.*, 2011).

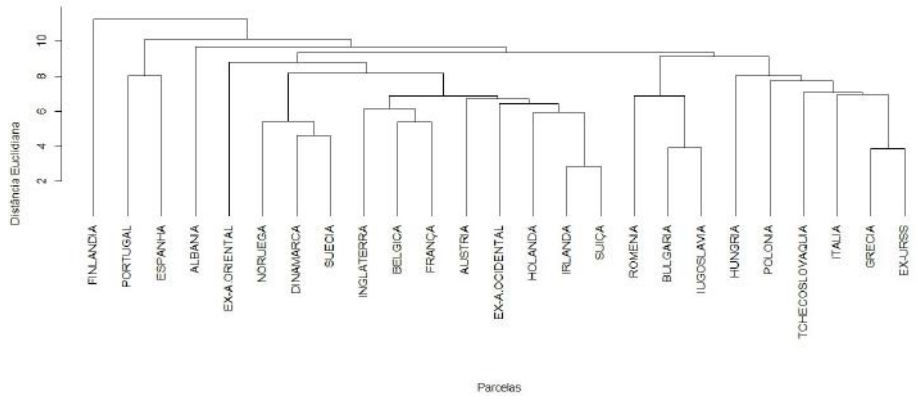

Figura 2 Exemplo de dendrograma construído pelo método de agrupamento hierárquico divisivo (O autor, 2019)

O método aglomerativo inicia a partir de n clusters nos quais cada elemento forma o próprio cluster, ou seja, cada cluster possui apenas um elemento. Em cada etapa, os dois clusters mais semelhantes são mesclados formando um único cluster. Os diferentes métodos hierárquicos aglomerativo se caracterizam por diferentes formas de se calcular a dissimilaridade. Entre esses clusters, os métodos mais conhecidos está o *"complete linkage"* (vizinho mais distante) (HENNIG *et al.*, 2016). Neste algoritmo, a distância entre dois grupos é dada pela maior distância entre um elemento de um grupo e o de outro grupo (JAIN & DUBES, 1988).

Os métodos divisivos, ao contrário do aglomerativo, começam como todos os elementos em um único cluster e prosseguem realizando divisões nos clusters existentes em cada etapa. Esse método é computacionalmente mais exigente devido à dificuldade em

encontrar divisões ótimas para realizar os agrupamentos e, por essa razão, os métodos aglomerativos são amplamente mais utilizados (HENNIG *et al.*, 2016).

Entretanto, se os dados utilizados consistirem em *p* variáveis binárias, métodos relativamente simples e com baixo custo computacional, conhecidos como métodos monotéticos divisivos, são uma opção. Estes métodos dividem os clusters de acordo com a presença ou ausência dessas *p* variáveis o que faz com que os dados precisem estar em uma matriz binária. Embora os métodos divisivos sejam menos utilizados, eles possuem uma vantagem, a estrutura principal de seus dados é apresentada desde o início do método, algo almejado pela maioria dos usuários (EVERITT *et al.*, 2011).

3.3.5.3 Índices de Avaliação dos Agrupamentos

A determinação do número ótimo de agrupamentos em um conjunto de dados tem sido a aplicação mais comum para os índices de validação de agrupamentos. De modo geral, esses índices se enquadram em uma das três categorias:

- Validação de agrupamento interno: com base apenas nas informações dos dados.
- Validação de agrupamento externo: baseado em conhecimento prévio dos dados.
- Validação de agrupamento relativo: Com base na análise repetida do mesmo algoritmo em diferentes parâmetros para obter resultados estáveis (MARY, SIVAGAMI & RANI, 2015).

Os agrupamentos têm por objetivos fazer com que os dados pertencentes ao mesmo grupo sejam semelhantes possíveis, enquanto dados pertencentes a grupos diferentes sejam o mais distinto possível. Dessa forma, a maioria dos índices de validação pressupõe que os agrupamentos devem ser os mais compactos e separados possível (CHENG, *et al*., 2018).

3.3.5.4 Índice Davies Bouldin (DB)

O índice Davies Bouldin (DB) é uma medida que avalia o desempenho dos agrupamentos. A proposta é avaliar a separação entre o i-ésimo e j-ésimo agrupamento, que deve ser a maior possível entre eles e a menor possível dentro do agrupamento. Esse índice tem correlação positiva para o caso "dentro da classe" e correlação negativa para o caso "entre classes", ou seja, quanto menor o valor do índice, melhor o resultado do agrupamento (XIAO, LU & LI, 2017).

$$DB = \frac{1}{K}\sum_{i=1}^{K}\max_{i;j\neq i}\frac{S_i + S_j}{d_{i,j}} \quad (12)$$

Onde $S_i = \frac{1}{C_i}\sum_{x_j \in C_i}\left\|x_j - v_i\right\|$ é uma medida de dispersão dentro do agrupamento i, K é o número de agrupamentos, x_j é um vetor n dimensional atribuído ao cluster i, v_i é o centro do cluster i, C_i representa o cluster i, $\|.\|$ é a distância euclidiana, $d_{i,j} = \left\|v_i - v_j\right\|$ é a distância entre o centro do cluster i e j.

3.3.5.5 Índice Dunn

O Índice Davies Bouldin (DBI) é uma medida para avaliar o desempenho do agrupamento que se baseia em avaliar a separação entre o i-ésimo e o j-ésimo cluster, que deve ser a maior possível entre os clusters e a menor possível dentro de um cluster (XIAO, LU & LI, 2017). No índice Dunn, a separação entre os agrupamentos é calculada com base no elemento central de cada grupo que é definido por:

$$D = \frac{d_{min}}{d_{max}} \quad (13)$$

Em que d_{min} é a distância mínima entre as amostras dos diferentes agrupamentos e d_{max} é a maior distância dentro do cluster.

Seja c_i e c_j diferentes agrupamentos. A distância entre as amostras mais próximas é:

$$d_{ij} = \min_{\forall x_k \in c_i, \forall x_{k'} \in c_j}\left\|x_k^{(i)} - x_{k'}^{(j)}\right\|_2$$

Onde x_k e $x_{k'}$ são elementos do conjunto de dados;

e

$$d_{min} = \min_{i\neq j} d_{ij},$$

É a menor distância entre dois elementos pertencentes a diferentes agrupamentos em todo o conjunto de dados.

A maior distância entre amostras distintas no agrupamento c_r é dado por:

$$d_{rr} = \max_{k\neq k'}\left\|x_k^{(r)} - x_{k'}^{(r)}\right\|_2$$

Sendo d_{max} a maior destas distâncias para todos os agrupamentos, temos:

$$d_{max} = \max_{r=1,\dots,c} d_{rr}$$

O número ideal de agrupamentos é aquele que der o maior valor para o índice. Assim, há indícios da existência de agrupamentos compactos e bem separados (MOTA, DAMASCENO & LEITE, 2018).

3.3.5.6 Índice C

O índice C é definido pela equação:

$$C = \frac{S_W - S_{min}}{S_{max} - S_{min}} \quad (14)$$

Em que, N_W: é o número total de pares de pontos dentro de um agrupamento.

N_T: é o total de pares distintos em todo o conjunto de dados;

S_W: é o somatório das distâncias de N_T entre todos os pares de pontos em todo o conjunto de dados;

S_{min}: é a soma das menores distâncias estabelecidas em N_W entre todos os pares de pontos em todo o conjunto de dados.

S_{max}: é a soma das maiores distancias de N_W entre todos os pares de pontos em todo o conjunto de dados.

Consideram-se as N_T distâncias entre pares de pontos como uma sequência de valores em ordem crescente. O índice *C* usa os menores e os maiores valores de N_W para calcular as somas S_{min} e S_{max} (DESGRAUPES, 2017).

3.3.5.7 Índice SD

A ideia do índice SD baseia-se na dispersão média e na separação total de clusters. A dispersão média para os clusters é definida da seguinte maneira:

Considere o vetor de variâncias para cada variável no conjunto de dados. É um vetor V de tamanho p definido por:

$$V = \left(Var\ (V_1), \dots, Var(V_p)\right)$$

De maneira semelhante, definem-se vetores de variância $V^{\{k\}}$ para cada agrupamento C_k, sendo k o número do agrupamento:

$$V^{\{k\}} = \left(Var\ ({V_1}^{\{k\}}), \dots, Var({V_p}^{\{k\}})\right)$$

Considere S a média da norma dos vetores $V^{\{k\}}$ dividida pela norma do vetor V:

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

Por outro lado, a total separação dos agrupamentos, denominada D, é definida como:

Seja D_{max} e D_{min} respectivamente a maior e a menor distância entre os baricentros (G) dos agrupamentos:

$$D_{max} = \max_{k \neq k'} \left\| G^k - G^{k'} \right\|$$

$$D_{min} = \min_{k \neq k'} \left\| G^k - G^{k'} \right\|$$

Podemos denotar

$$D = \frac{D_{max}}{D_{min}} \sum_{k=1}^{K} \frac{1}{\sum_{\substack{k'=1 \\ k' \neq k}}^{K} \| G^k - G^{k'} \|}$$

e finalmente define-se o índice SD como:

$$C = \alpha S + D \qquad (15)$$

Em que α é um peso com valor igual ao de D obtido para a partição com o maior número de agrupamentos (HALKIDI, VAZIRGIANNIS & BATISTAKIS, 2000).

3.3.5.8 Coeficiente de silhueta médio

O coeficiente de silhueta médio (CSM) é um critério de avaliação que mede a qualidade da formação interna e da separação externa de grupos. O valor desse coeficiente para um ponto p, denotado por CS_p, é calculado pela equação (ROUSSEEUW, 1987):

$$CS_p = \frac{b_p - a_p}{Max(a_p, b_p)} \qquad (16)$$

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

3.3.5.9 Coeficiente de correlação cofenética

O coeficiente de correlação cofenética mede o grau de ajuste entre uma matriz de dissimilaridade original (matriz S) e a matriz resultante da simplificação proporcionada pelo agrupamento (matriz C), ou seja, a matriz C é aquela obtida após a construção do dendrograma, a equação que representa a correlação é dada por (BUSSAB *et al.*, 1990):

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij} - \bar{c})(s_{ij} - \bar{s})}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij} - \bar{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(s_{ij} - \bar{s})^2}} \quad (17)$$

Em que:

c_{ij}: valor das similaridades entre os indivíduos i e j, obtidos a partir da matriz cofenética;

s_{ij}: valor das similaridades entre os indivíduos i e j, obtidos a partir da matriz de similaridade;

$\bar{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}$;

$\bar{s} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij}$.

Nota-se que essa correlação se equivale à correlação de Pearson entre a matriz original e à obtida após a construção do dendrograma, assim, quanto mais próximo a 1, menor será a distorção provocada pelo agrupamento (BUSSAB *et al.*, 1990).

3.4 Análise de classificação

A classificação é um tipo de tarefa para localizar um grupo de funções que descrevem e diferenciam classes ou conceitos com o objetivo de utilizar esse modelo para prever a classe de objetos que ainda não foi classificado. Essa tarefa se vale da aprendizagem supervisionada, que é um tipo de aprendizado indutivo referente à capacidade que alguns algoritmos têm de aprender de acordo com exemplos. Assim, os algoritmos aprendem com os relacionamentos que existem entre os dados e representam esse aprendizado no modelo de conhecimento. Dessa forma, na atividade de classificação, as variáveis ou atributos são separados em dois grupos: o atributo-alvo e os atributos de predição. Destaca-se que existe somente um atributo-alvo, o qual deve ser categórico (GOLDSCHMIDT, PASSOS & BEZERRA, 2015).

Um modelo de classificação é útil como modelagem descritiva, quando funciona como uma ferramenta explicativa de dados, e como modelagem preditiva, em que é utilizado para prever o rótulo de classes de registros conhecidos (TAN, STEINBACH & KUMAR, 2009).

O processo de classificação é realizado em três etapas: a criação do modelo de classificação; a verificação do modelo e a utilização deste em novos dados. A criação do modelo é gerada a partir de um banco de dados chamado treinamento, no qual os elementos são chamados de amostras ou exemplos, de acordo com regras que permitem

classificar as tuplas (sequência ordenada e finita de elementos) do banco de dados dentro de um número pré-determinado de classes.

A verificação do modelo é realizada mediante regras testadas em um banco de dados independente do que foi utilizado para o treinamento. A qualidade deste modelo é então medida conforme a porcentagem de tuplas que ele consegue classificar corretamente de acordo comas regras do modelo gerado. A etapa de utilização do modelo é a fase na qual o mesmo após ter passado pelas fases anteriores é aplicado em novos dados (HAN & KAMBER, 2006).

3.4.1 Árvore de decisão

Dentre as técnicas de classificação existentes há a árvore de decisão. Ela tem sido bastante utilizada por possuir estruturas gráficas hierárquicas de fácil entendimento e aplicação (CERVANTES et al., 2015; RAMYA et al., 2015). Essa técnica se utiliza de aprendizado supervisionado que fornece um modelo representado graficamente por nós e ramos, semelhante a uma árvore, mas no sentido invertido. Seus nós são constituídos por: nó raiz, localizado na parte superior da estrutura e nós internos, que formam as ramificações (WITTEN, FRANK & HALL, 2011).

A construção de uma árvore de decisão inicia-se por um conjunto de treinamento que é dividido de acordo com testes aplicados sobre as variáveis independentes, e assim, formando subconjuntos homogêneos em relação à variável dependente. Esse procedimento é repetido até se formar conjuntos de exemplos bem homogêneos sobre os quais seja possível atribuir um único valor referente à variável dependente (SOUZA *et al.*, 2010). Assim, de maneira geral, o procedimento de indução de uma árvore de decisão se baseia em uma sucessiva divisão do conjunto de exemplos utilizados para o treino até a formação de subconjuntos, que pertençam à mesma classe, ou até que uma das classes passe a ser predominante sobre esse subconjunto, fazendo com que novas divisões não sejam necessárias. Esses subconjuntos formados serão utilizados para classificar novos exemplos (QUINLAN, 1993).

É necessária a utilização de algoritmos específicos para se construir uma árvore de decisão. Há muitos algoritmos que possam ser manipulados para esse fim, e alguns apresentam desempenho melhor para determinadas situações, enquanto outros podem ser mais eficientes em outras. Os algoritmos mais utilizados para a indução de árvore de decisão são o C 4.5 (QUINLAN, 1993) e o CART (BREIMAN *et al.*, 1984).

O algoritmo CART primeiramente lê os dados de treinamento usados para a construção do modelo de classificação. Em seguida, conta as frequências de todas as possíveis combinações entre as variáveis independentes e as dependentes, a frequência dos valores da variável dependente também é contada. Essas frequências também são

utilizadas para se encontrar o melhor critério de divisão para cada um dos nós não-folhas. Como critério de divisão, o algoritmo baseia-se nos valores de impureza (chamado de ganho de Gini) dos nós gerados depois da divisão do nó-pai em dois sucessores. Na sequência, o algoritmo seleciona o valor da variável com melhor escore e divide o conjunto de dados em dois subconjuntos separados, gerando dois nós filhos. Por fim, o processo é repetido recursivamente até que os dados dos nós filhos atinjam um limite especificado (NARAYANAN, *et al.* 2007).

Uma das implementações desse algoritmo ocorre pelo pacote “Rpart” no software R. O pacote tem como critério de seleção de atributos a partição de nós, o índice de Gini ou o Índice de Informação. As árvores geradas são binárias e limitadas a trinta e um níveis de profundidade (THERNEAU & ATKINSON, 2011).

O pacote Rpart utiliza medidas de impureza como critério para a partição de nós. Seja f uma função para definir a impureza do nó A:

$$I(A) = \sum_{i=1}^{C} f(p_i A) \tag{18}$$

Em que $I(A)$ é o valor da impureza do nó A; $p_i A$ é a proporção dos elementos do nó A pertencer à classe i. Os dois candidatos para a função f são o $f(p) = -p\log(p)$ para o índice de informação e $f(p) = p(1-p)$ para o índice de Gini (THERNEAU & ATKINSON, 2011).

As duas medidas de impureza podem ser o índice de Gini ou o Índice de Informação. Para a separação do nó, é escolhido o índice I que maximiza a redução da impureza:

$$\Delta I = p(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R) \tag{19}$$

Em que p é a proporção de elementos em um nó; A_L e A_R são os nós da direita e da esquerda, candidatos à partição do nó A (THERNEAU & ATKINSON, 2011). Ao analisar a maneira como ocorre o treinamento de árvores de decisão, percebe-se que se o procedimento que cria novos nós for realizado indefinidamente, até que folhas com apenas uma única amostra formem um nó, haverá como resultado nós terminais com pureza máxima. Esse fato pode elevar muito a complexidade da árvore criada e fazer com que ela tenha muitas regras e com chances de ocorrer sobreajustes (*overfitting*), e assim reduzir sua capacidade de generalização. O sobreajuste é o ajuste demasiado da árvore podendo fazer com que ela se ajuste a peculiaridades dos dados que talvez não ocorram nos elementos ainda não vistos (FREITAS, 2000).

Uma maneira de contornar esse problema é utilizar técnicas de podas. Essas técnicas permitem detectar e excluir sub-árvores do modelo e assim melhorar a taxa de acerto de novas observações. As técnicas de poda em árvores de decisão podem ser divididas em dois grupos: pré-poda e pós-poda. Na pré-poda o particionamento de uma

árvore de decisão pode ser interrompido durante a fase de construção quando algum critério de parada ou condição pré-estabelecida for satisfeita. O pós-poda constrói toda a árvore para depois podá-la. As técnicas de pré-poda são mais rápidas, porém menos eficazes do que as de pós-poda, pelo fato de correrm os o risco de interromper o crescimento da árvore ao selecionar uma árvore subótima (BREIMAN *et al.*, 1984).

3.4.2 Avaliação do classificador

O potencial do modelo da árvore de decisão será avaliado com relação ao percentual de instâncias classificadas corretamente. Utiliza-se de uma matriz de confusão para avaliar o percentual de instâncias classificadas corretamente, que é gerada a partir dos elementos da matriz de contingência das classes dos mapas temáticos, também conhecida como matriz de erros (Tabela 2) (CONGALTON & GREEN, 1999).

Tabela 2 Matriz genérica dos erros de ordem c x c

	Mapa de referência		
			Total por linha
Classes	C_1	C_c	$n_{i.}$
C_1	n_{11}	n_{1c}	$n_{1.}$
	⋮	⋮	⋮
C_c	n_{c1}	n_{cc}	$n_{c.}$
Total por Coluna			Total geral
$n_{.i}$	$n_{.1}$	$n_{.c}$	n

c: número de classes; c_i: classe i; n_i: total de *pixels* na classe c_i no mapa modelo; $n_{.i}$: total de *pixels* na classe c_i no mapa de referência; n: número total de *pixels*.

Seja n o número total de *pixels*, os elementos da diagonal principal n_{ii} representam os casos nos quais os *pixels* tiveram a mesma classificação nos dois mapas. Os demais elementos dessa matriz representam as classificações incorretas dos *pixels*. Com o uso dos elementos da matriz de erros, é possível a obtenção de várias métricas para se comparar os mapas temáticos entre eles a exatidão global (CONGALTON & GREEN, 1999).

4. MATERIAL E MÉTODOS

4.1 Área de estudo

O estudo será realizado com dados coletados em uma área agrícola comercial de produção de grãos, com 167,35 ha (Figura 3), localizada no município de Cascavel, estado do Paraná. As coordenadas geográficas aproximadas da área são latitude 24,95º S, longitude de 53,57º O. Nesta área foram determinados 102 pontos para realização das coletas de dados. A altitude média da área é de 650 m (Mapa de localização da área de estudo). A região possui solo de classificação Latossolo Vermelho Distroférrico, com textura argilosa (EMBRAPA, 2013) e clima temperado mesotérmico superúmido, tipo climático Cfa (Koeppen), com temperatura anual média de 21ºC.

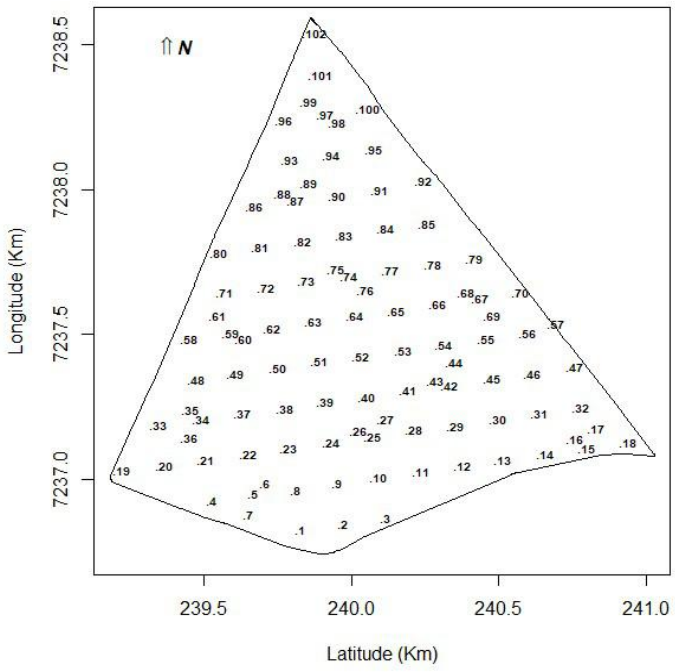

Figura 3 Representação do talhão com os respectivos números dos pontos amostrais para os anos-safra 2013/2014, 2015/2016 e 2016/2017

Para o ano de 2014/2015, não foi possível trabalhar com toda a área de estudo, pois a região ao norte do talhão não possuía valores para a produtividade, sendo então essa região desconsiderada do trabalho, conforme Figura 4.

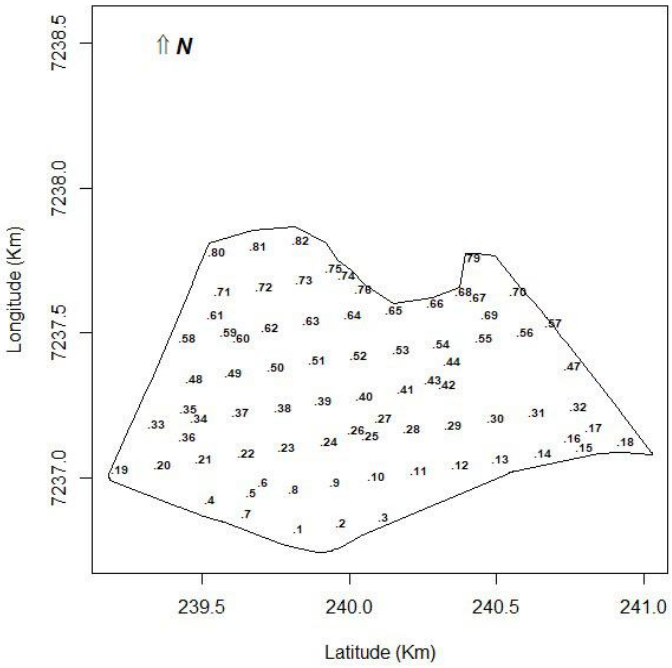

Figura 4 Representação do talhão com os respectivos números dos pontos amostrais para o ano-safra 2014/2015

4.2 Obtenção dos dados e seleção de variáveis

Serão utilizados dados obtidos durante a coleta em campo bem como os dados recolhidos por sensoriamento remoto. Os dados coletados em campo são referentes aos anos-safra 2013/2014, 2014/2015, 2015/2016 e 2016/2017. A área recebeu o plantio de soja, variedade AMS Tibagi RR. Um GPS (Global Position System) foi utilizado para localizar os pontos de amostragem no campo. Esse conjunto consta de observações de resistência do solo à penetração (RSP), densidade, umidade do solo, aos atributos físico-químicos do solo e à produtividade (kg ha^{-1}) (Tabela 3).

Observa-se, no entanto, que o conjunto de variáveis referentes às variáveis físico-químicas utilizadas em cada ano-safra sofreu variações baseadas na disponibilidade dessa informação e em relação ao valor da sua dependência espacial. Assim, as variáveis que apresentaram fraca dependência espacial, de acordo com o índice proposto por Cambardella *et al.* (1994), não foram consideradas no estudo, como mostrados nos Apêndices A, B C e D. A Tabela 3 apresenta quais variáveis foram utilizadas no estudo em cada ano-safra.

Tabela 3 Conjunto de variáveis referentes às variáveis físico-químicas do solo

	2013/2014	2014/2015	2015/2016	2016/2017
Alumínio (Al) (cmolcdm^{-3})	NÃO	SIM	NÃO	SIM
Carbono (C) (gdm^{-3})	SIM	SIM	SIM	SIM
Cálcio (Ca) (cmolcdm^{-3})	SIM	SIM	SIM	SIM
Cobre (Cu) (mgdm^{-3})	SIM	NÃO	SIM	NÃO
Ferro (Fe) (mgdm^{-3})	SIM	NÃO	SIM	NÃO
Acidez potencial (HAl_3) (cmolcdm^{-3})	NÃO	NÃO	NÃO	SIM
Potássio (K) (cmolcdm^{-3})	SIM	NÃO	SIM	SIM
Magnésio (Mg) (cmolcdm^{-3})	NÃO	SIM	NÃO	SIM
Manganês (Mn) (mgdm^{-3})	SIM	SIM	SIM	NÃO
Fósforo (P) (mgdm^{-3})	NÃO	SIM	NÃO	SIM
pH	NÃO	NÃO	NÃO	SIM
Zinco (Zn) (mgdm^{-3})	SIM	SIM	SIM	NÃO
Umidade 0-10 cm (%)	SIM	SIM	SIM	NÃO
Umidade 10-20 cm (%)	NÃO	SIM	NÃO	SIM
Umidade 20-30 cm (%)	NÃO	SIM	SIM	NÃO
Densidade 0-10 cm (gcm^{-3})	NÃO	NÃO	NÃO	NÃO
Densidade 10-20 cm (gcm^{-3})	NÃO	NÃO	SIM	SIM
Densidade 20-30 cm (gcm^{-3})	NÃO	SIM	SIM	NÃO
RSP 0-10 cm (kPa)	NÃO	NÃO	SIM	SIM
RSP 10-20 cm (kPa)	NÃO	SIM	SIM	NÃO
RSP 20-30 cm (kPa)	NÃO	SIM	SIM	NÃO
RSP 30-40 cm (kPa)	NÃO	SIM	SIM	NÃO

SIM: corresponde às variáveis utilizadas no estudo; NÃO: corresponde às variáveis não utilizadas no estudo; RSP: resistência do solo à penetração.

Os índices de vegetação coletados foram utilizados a partir das imagens do sensor OLI do satélite Landsat-8 para os dados obtidos por sensores. As informações recolhidas por esse sensor e utilizadas no trabalho são referentes ao ciclo total da soja. Os índices de vegetação empregados neste estudo são: ARVI, NDVI, EVI2, SAVI, OSAVI e WDRVI. Foram utilizadas somente imagens com percentual de nuvens abaixo de 10% da área total. Assim, para cada ano-safra, foi possível calcular os valores dos índices vegetativos para somente algumas datas dentro do ciclo vegetativo da soja, conforme Tabela 4. Para cada data, foi associado o estágio vegetativo aproximado de acordo com a época de plantio, conforme as informações disponíveis em Fehr e Caviness (1981).

O período escolhido abrange desde a emergência da plântula (VE) até a maturação plena (R8), de acordo com a Tabela 4. Assim, todos os estádios desde o surgimento até a maturação da planta são compreendidos.

Tabela 4 Data das imagens de satélite para a qual foi realizado o cálculo dos índices vegetativos para cada ano-safra

Ano-safra 2013/2014		**Ano-safra 2014/2015**		**Ano-safra 2015/2016**		**Ano-safra 2016/2017**	
Data	E.V.	Data	E.V.	Data	E.V.	Data	E.V.
13/09/2013	VE	16/09/2014	VE	21/10/2015	VE	07/10/2016	VE
31/10/2013	R2	05/12/2014	R4	06/11/2015	R1	23/10/2016	R1
18/12/2013	R6	22/01/2015	R6			24/11/2016	R3
19/01/2014	R7					26/12/2016	R5

E.V.: estágio vegetativo; VE: fase de emergência; R1: início do florescimento; R2: pleno florescimento; R3: início da formação das vagens; R4: plena formação das vagens; R5: início do enchimento das sementes; R6: pleno enchimento das sementes; R7: início da maturação; R8: maturação plena.

Realizou-se a análise geoestatística para cada ano-safra com destaque para cada variável a fim de se observar a dependência espacial e posteriormente gerar os mapas temáticos por krigagem. Inicialmente, para realizar essa análise, foi necessário observar a existência ou não de tendência direcional e anisotropia no conjunto de dados e, em caso afirmativo, ajustar o modelo para correção destes fatores.

Os modelos teóricos utilizados foram os correspondentes à família *Matérn*, propostos por Diggle & Ribeiro Junior (2000), que apresentam um parâmetro k chamado de ordem do modelo *Matérn*. Os valores escolhidos para esse parâmetro foram 0,5 (modelo exponencial), 1, 2 e com $k \to \infty$ (modelo gaussiano). O método estatístico utilizado para o ajuste dos modelos e estimação de seus parâmetros foi o de máxima verossimilhança. E a escolha do melhor modelo ocorreu por validação cruzada e pelo critério de Akaike (FARACO *et al.*, 2008).

O índice proposto por Cambardella *et al.* (1994) foi utilizado para se avaliar o grau de dependência espacial (EPR), no qual valores do índice abaixo de 0,25 são considerados como forte dependência espacial, entre 0,25 e 0,75 como média dependência espacial e acima de 0,75 como fraca dependência espacial. As variáveis que apresentaram EPR forte e médio foram selecionadas com base no conhecimento do grau de dependência espacial, excluindo, portanto, os que apresentaram baixa dependência ou efeito pepita puro, Tabela 3.

As variáveis que sobraram desse processo foram agrupadas em cinco subconjuntos:

1. todas as variáveis;
2. variáveis relacionadas aos índices vegetativos;
3. variáveis relacionadas às propriedades químicas e físicas do solo;
4. variáveis referentes aos índices vegetativos mais a produtividade;
5. variáveis relacionadas às propriedades físicas e químicas do solo mais a produtividade.

A redução da dimensionalidade foi realizada usando a técnica de MULTISPATI-PCA PCA para cada subconjunto (GAVIOLI *et al.*, 2016). A partir da técnica de MULTISPATI-PCA, cada subconjunto de variáveis foi transformado em variáveis sintéticas chamadas de componentes principais espaciais (CPEs). Utilizou-se a quantidade de CPEs que representasse no mínimo 70% da variabilidade total (GAVIOLI *et al.*, 2016). As CPEs foram empregadas posteriormente para a construção das matrizes de dissimilaridades por meio da função semivariância direta e cruzada com núcleo estimador não paramétrico, descrita por Fouedjio (2006).

4.3 Delineamento e análise das zonas de manejo

Algumas etapas descritas no fluxograma da Figura 5 foram seguidas para geração e análise das zonas de manejo:

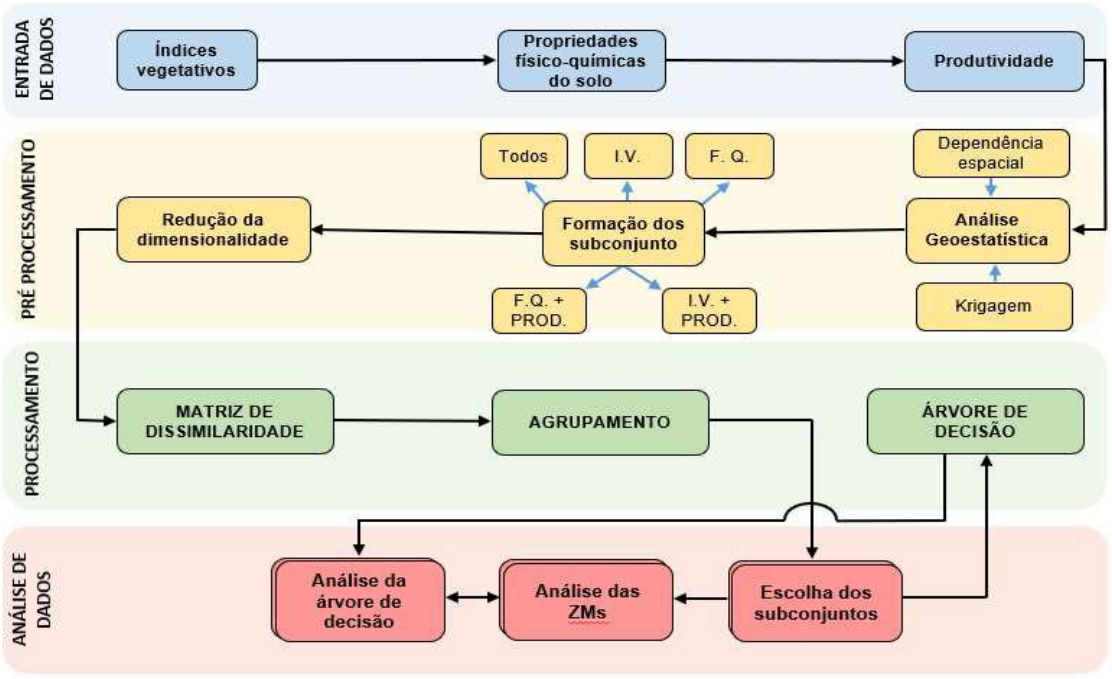

Figura 5 Fluxograma para o delineamento de zonas de manejo baseado em um método hierárquico aglomerativo (O autor, 2019)

As zonas de manejo foram geradas com a proposta de FOUEDJIO (2016), (Figura 4). Para isso, utilizou-se uma técnica de agrupamento hierárquico aglomerativo, que leva em consideração a dependência espacial entre as observações. A informação espacial foi incluída no agrupamento por uma função núcleo-estimador não paramétrica (Equação 09), a qual serviu para construir uma medida de dissimilaridade entre cada par de pontos amostrais.

O parâmetro largura de banda λ foi escolhido de acordo com a regra empírica em que o valor de λ será ampliado em 10% até que pelo menos 35 locais amostrados estejam incluídos em um círculo com este raio. A função núcleo-estimador usada nesse trabalho foi a de Epanechnikov, ou seja, $K_\lambda(||\boldsymbol{x}-\boldsymbol{s}||)\ \alpha\ (\lambda^2-||\boldsymbol{x}-\boldsymbol{s}||^2)\mathbb{1}_{||\boldsymbol{x}-\boldsymbol{s}||<\lambda}$, sugerida por FOUEDJIO (2016), por apresentar suporte e mostrar propriedade de otimalidade na estimativa de densidade (WAND & JONES, 1995).

Um algoritmo aglomerativo hierárquico foi utilizado para realizar os agrupamentos de acordo com os locais amostrados, e sobre ele foi aplicada a matriz de dissimilaridade (Equação 11). O algoritmo utilizado foi o de aglomerativo de ligação completa, indicado por Fouedjio (2016), como o mais adequado, dado que a medida de dissimilaridade indicada leva em consideração tanto a proximidade no espaço do atributo quanto no espaço geográfico gerando *clusters* compactos e conectados.

Mapas foram gerados com 2, 3 e 4 zonas de manejo. O coeficiente de correlação cofenética (Equação 17) foi utilizado para escolha do subconjunto que melhor agrupou os dados. Esse coeficiente faz uma análise da correlação entre a matriz de dissimilaridade com o dendrograma gerado por meio desta, assim, quanto maior o valor do coeficiente mais ajustado estará o agrupamento.

A análise de diversos índices de avaliação de agrupamento foi utilizada como critério para a escolha da quantidade ótima de zonas para a área em estudo. São eles: o coeficiente de silhueta médio (C.S.M.) (Equação 16), índice C (Equação 14), índice SD (Equação 15), índice Dunn (Equação 13) e índice DB (Equação 12). Para os índices C, SD e DB, os menores valores indicam melhor formação de grupo enquanto para os coeficientes de silhueta médio e o índice Dunn, quanto maior o valor melhor é a formação do grupo (XIAO & LU & LI, 2017; MOTA, DAMASCENO & LEITE, 2018; DESGRAUPES, 2017; HALKIDI, VAZIRGIANNIS & BATISTAKIS, 2000; ROUSSEEUW, 1987). O método de classificação de árvore de decisão foi utilizado para melhor entendimento da distribuição das variáveis na geração das zonas de manejo. As zonas de manejo geradas pelo subconjunto que melhor agrupou os dados foram empregadas como variáveis dependentes (resposta) na árvore; as variáveis independentes (explicativas) escolhidas foram as que faziam parte do subconjunto escolhido para gerar as zonas de manejo.

A árvore foi gerada por algoritmo disponível no pacote "*rpart*" do software R. Os dados foram divididos de maneira aleatória em 70% para treinamento e os outros 30% para testes. Devido à pequena quantidade de pontos amostrados foi necessária a interpolação dos dados para gerar uma grande quantidade de elementos e maior precisão na classificação. Além disso, utilizou-se também a função "*ovun.sample*" do pacote "*ROSE*" do software R para balancear as classes, pois havia predominância de uma das classes sobre a outra, e estava interferindo na classificação dos dados. Foram realizadas, quando necessário, podas nas árvores de decisão para gerar árvores menores e de fácil

interpretação. A avaliação da classificação ocorreu pela análise da matriz de confusão entre os elementos classificados e os dados de teste e de treinamento.

As medidas de desempenho são extraídas da matriz de confusão, e são utilizadas para avaliar o classificador. As medidas de desempenho utilizadas foram: Acurácia, sensibilidade, especificidade, valor da predição positiva, valor da predição negativa e a acurácia balanceada.

A acurácia descreve a precisão do classificador, e seu valor é calculado pela soma da diagonal principal da matriz de confusão dividido pelo total, assim como o cálculo da exatidão global descrita por Congalton (1991).

A sensibilidade indica o quanto dos valores preditos da classe positiva foi classificado corretamente. A especificidade, por sua vez, mede o quanto negativa foi classificado corretamente dos elementos preditos da classe. Essas medidas são semelhantes à acurácia do produtor descrita por Congalton (1991).

O valor da predição positiva indica o quanto os valores pertencentes originalmente à classe ZM1 (Classe positiva) foram classificados corretamente pelo modelo preditor, enquanto o valor da predição negativa indica o quanto foi classificado corretamente dos elementos pertencentes originalmente à classe ZM2 como sendo desta classe. Essas medidas são semelhantes à acurácia do usuário descrita por Congalton (1991). A acurácia balanceada é a média entre os valores de sensibilidade e especificidade, ou seja, leva em conta o percentual de predição de cada classe.

5. RESULTADOS E DISCUSSÃO

5.1 Geração e análise das zonas de manejo

Um total de 32 variáveis foi utilizado para o ano-safra de 2013/2014, com 24 correspondentes aos índices vegetativos (seis índices calculados para cada uma das cinco imagens disponíveis) (Tabela 4), sete delas relacionadas às variáveis físico-químicas do solo (Tabela 3) e uma relacionada à produtividade. A análise da dependência espacial assim como do raio de dependência espacial (alcance) foi determinante na escolha dessas variáveis. Assim, todas as variáveis relacionadas aos índices vegetativos apresentaram dependência espacial de média à forte. Enquanto as relacionadas às variáveis físico-químicas do solo, somente sete delas atenderam aos critérios estabelecidos nesse trabalho quanto aos valores para a dependência espacial e o alcance (Apêndice A).

No ano-safra de 2014/2015, foi possível obter dados referentes aos índices vegetativos somente em três dias de todo o ciclo da soja, devido à alta porcentagem de nuvens nas imagens de satélite, conforme Tabela 4. Por outro lado, de acordo com a Tabela 3, foi possível trabalhar com mais variáveis físico-químicas do solo do que no ano-safra anterior, pois a maioria dessas apresentou dependência espacial média à forte (Apêndice B).

Para o ano-safra de 2015/2016, foram utilizadas, para a formação dos subconjuntos, doze variáveis relacionadas aos índices vegetativos (seis índices calculados para as duas imagens), conforme Tabela 3. As variáveis referentes às variáveis físico-químicas do solo foram Cobre (Cu)(mg/dm^3), Zinco (Zn)(mg/dm^3), Ferro (Fe)(mg/dm^3), Manganês (Mn)(mg/dm^3), Carbono (C)(g/dm^3), Cálcio (Ca)(cmolc/dm^3), Potássio (K)(cmolc/dm^3), umidade(%) nas camadas de 0 a 10 cm e 10 a 20 cm, densidade (g/cm^3) nas camadas de 10 a 20 cm e 20 a 30 cm e resistência do solo à penetração (RSP) (kPa) nas camadas de 0 a 10 cm, 10 a 20 cm, 20 a 30 cm e 30 a 40 cm. Além dessas, os valores para a produtividade também fizeram parte do conjunto de variáveis.

O ano-safra de 2015/2016 teve o menor número de variáveis relacionadas aos índices vegetativos do que os outros anos-safra devido à alta porcentagem de nuvens sobre o talhão nas datas em que o satélite passou sobre a área que impossibilitou o uso dessas. Por outro lado, foi o que teve maior disponibilidade de variáveis físico-químicas para o estudo, ou seja, foi o ano em que mais variáveis relacionadas a esse conjunto de dados apresentaram dependência espacial média ou forte, conforme o Apêndice C.

Para o ano-safra de 2016/2017, foram utilizadas variáveis relacionadas aos índices vegetativos calculados em quatro diferentes datas dentro do ciclo vegetativo da soja, conforme Tabela 4. Foram empregadas também onze variáveis concernentes às variáveis físico-químicas do solo, sendo oito delas variáveis químicas e três físicas, conforme Tabela

3. Além dessas, a variável relacionada à produtividade também foi aplicada. Os valores da dependência espacial de cada variável assim como o raio de dependência espacial podem ser vistos no Apêndice D.

Para cada ano-safra e para cada um dos cinco subconjuntos utilizados no estudo, foi realizada a redução da dimensionalidade gerando duas componentes principais espaciais (CPEs). A partir dessas componentes, foi efetivado o agrupamento gerando mapas com 4, 3 e 2 zonas de manejo.

Observa-se que, para os anos-safra de 2013/2014, 2014/2015 e 2016/2017, ao serem comparadas essas zonas de manejo nos quatro anos-safra pesquisados, há semelhanças quanto à localização dessas zonas (Figuras 6, 7 e 9), ou seja, nesses casos a ZM2 se encontra ao leste do talhão, mais especificamente ao sudeste nos casos de 2013/2014 e 2016/2017. No ano-safra de 2014/2015, a divisão do talhão está entre leste e o Oeste, entretanto, somente a parte pôde ser avaliada. O ano-safra que apresentou diferença quanto à localização foi somente o de 2015/2016 em que o talhão ficou divido em norte (ZM2) e sul (ZM1) (Figura 8).

Além disso, na maioria dos mapas gerados para todos os anos-safra há sempre o predomínio de uma zona de manejo sobre as demais, isto é, o grupo representado pelos quadrados pretos geralmente ocupa uma porcentagem de área e de pontos amostrais maiores que os outros grupos. De modo geral, os anos-safra de 2013/2014, 2014/2015 e 2016/2017 foram os que tiveram maior diferença no tamanho entre as zonas de manejo.

Em relação ao ano-safra de 2015/2016 é possível ainda observar que nos três primeiros subconjuntos quando há a formação de mais de duas zonas de manejo, um agrupamento na região mais ao sul e leste do mapa (canto inferior direito), formado pelos quadrados azuis, semelhante ao ocorrido nos outros anos-safra, entretanto essa região se associou ao grupo maior, formado pelos quadrados pretos. Percebe-se então, que para esse ano-safra a região ao norte do mapa apresentou maior divergência do que nos outros anos-safra.

Outra questão importante que pode ser observada analisando os mapas é a de que, de modo geral, a divisão do talhão em mais do que três grupos pode ser inviável para o agricultor, pois, na maioria dos casos, há formação de pelo menos um grupo muito pequeno o que dificultaria o manejo sem o uso de maquinário especializado o que contradiz as definições de zona de manejo.

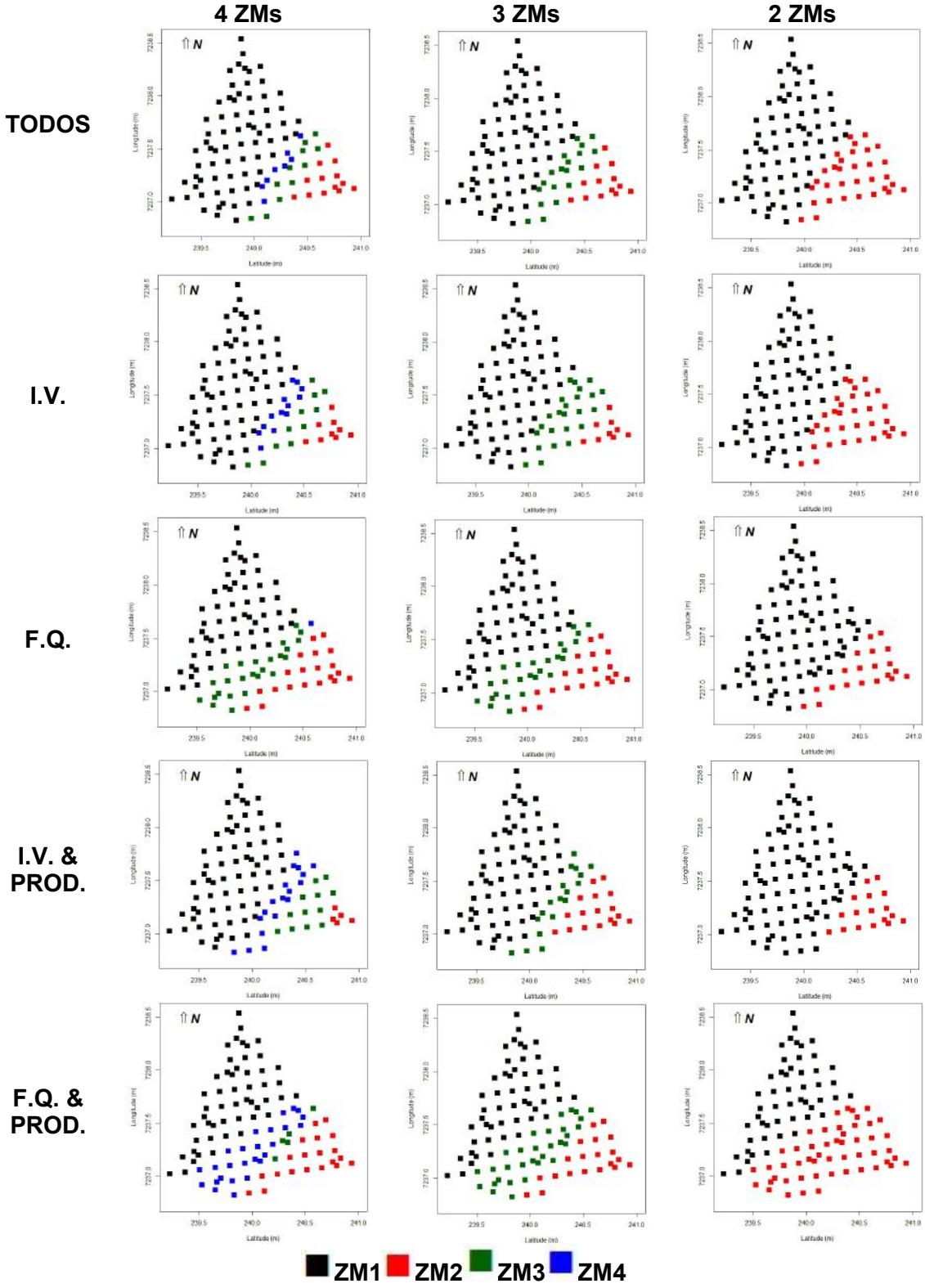

Figura 6 Mapas de Zonas de Manejo para os cinco subconjuntos de variáveis para o ano-safra de 2013/2014 (TODOS: subconjunto com todas as variáveis; I.V.: subconjunto formado somente pelas variáveis relacionadas aos índices vegetativos; F.Q.: subconjunto formado somente pelas variáveis relacionadas às variáveis físico-químicas do solo; I.V. & PROD.: subconjunto formado pelas variáveis relacionadas aos índices vegetativos e à produtividade; F.Q. & PROD.: subconjunto formado pelas variáveis relacionadas às variáveis físico-químicas do solo e produtividade; ZM: zona de manejo)

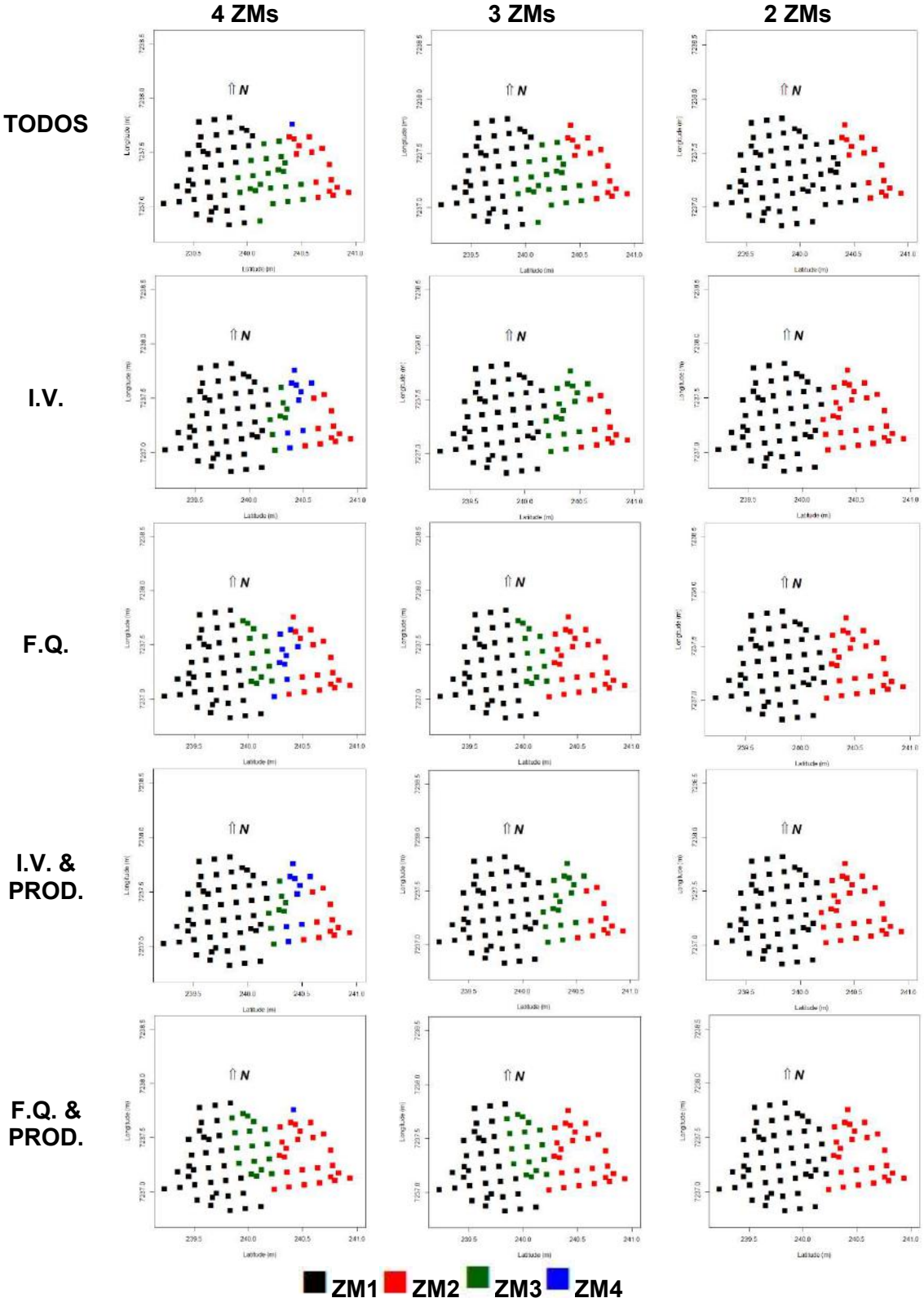

Figura 7 Mapas de Zonas de Manejo para os cinco subconjuntos de variáveis para o ano-safra de 2014/2015 (TODOS: subconjunto com todas as variáveis; I.V.: subconjunto formado somente pelas variáveis relacionadas aos índices vegetativos; F.Q.: subconjunto formado somente pelas variáveis relacionadas às variáveis físico-químicas do solo; I.V. & PROD.: subconjunto formado pelas variáveis relacionadas aos índices vegetativos e à produtividade; F.Q. & PROD.: subconjunto formado pelas variáveis relacionadas às variáveis físico-químicas do solo e produtividade; ZM: zona de manejo)

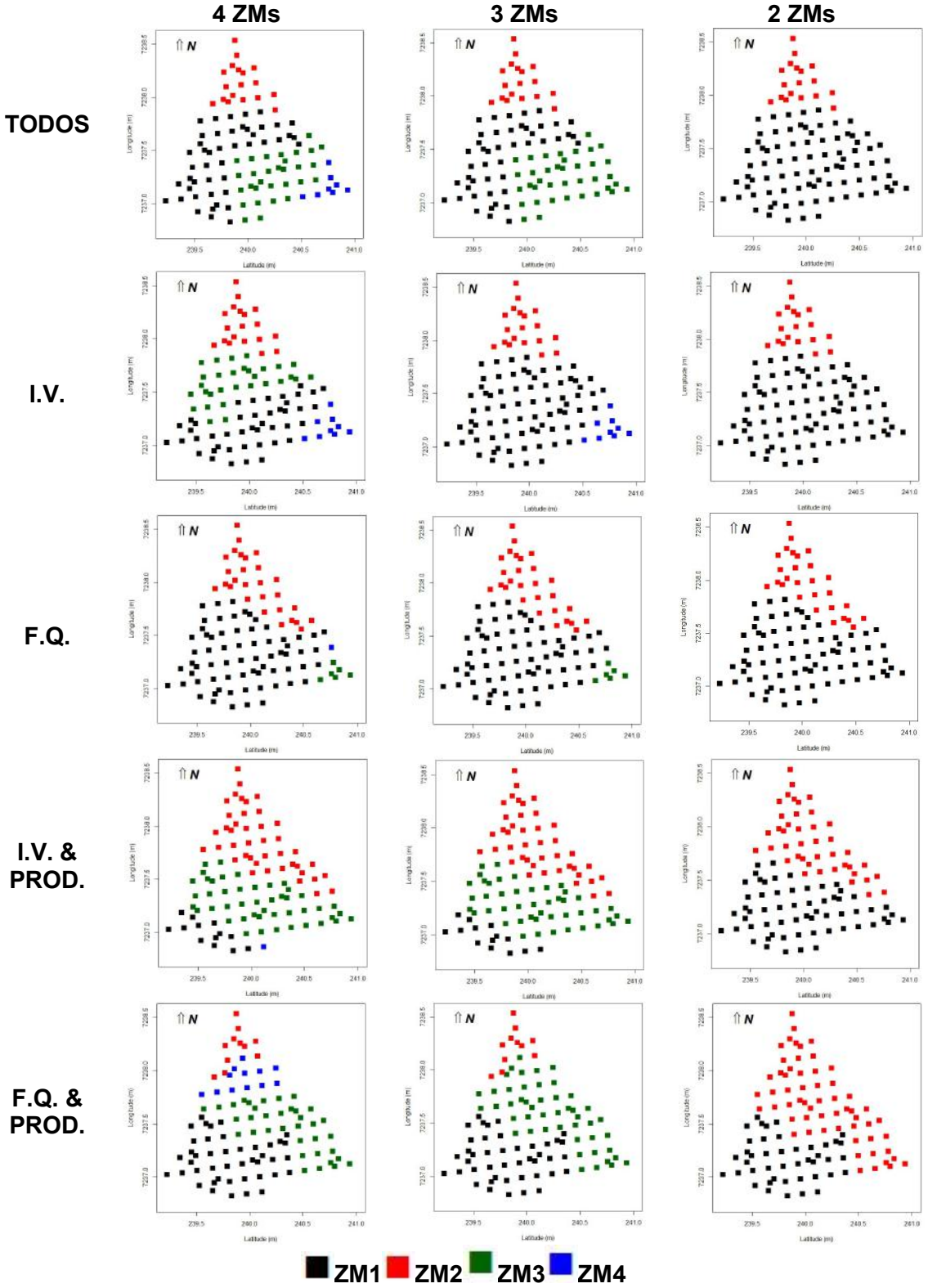

Figura 8 Mapas de Zonas de Manejo para os cinco subconjuntos de variáveis para o ano-safra de 2015/2016 (TODOS: subconjunto com todas as variáveis; I.V.: subconjunto formado somente pelas variáveis relacionadas aos índices vegetativos; F.Q.: subconjunto formado somente pelas variáveis relacionadas às variáveis físico-químicas do solo; I.V. & PROD.: subconjunto formado pelas variáveis relacionadas aos índices vegetativos e à produtividade; F.Q. & PROD.: subconjunto formado pelas variáveis relacionadas às variáveis físico-químicas do solo e produtividade; ZM: zona de manejo)

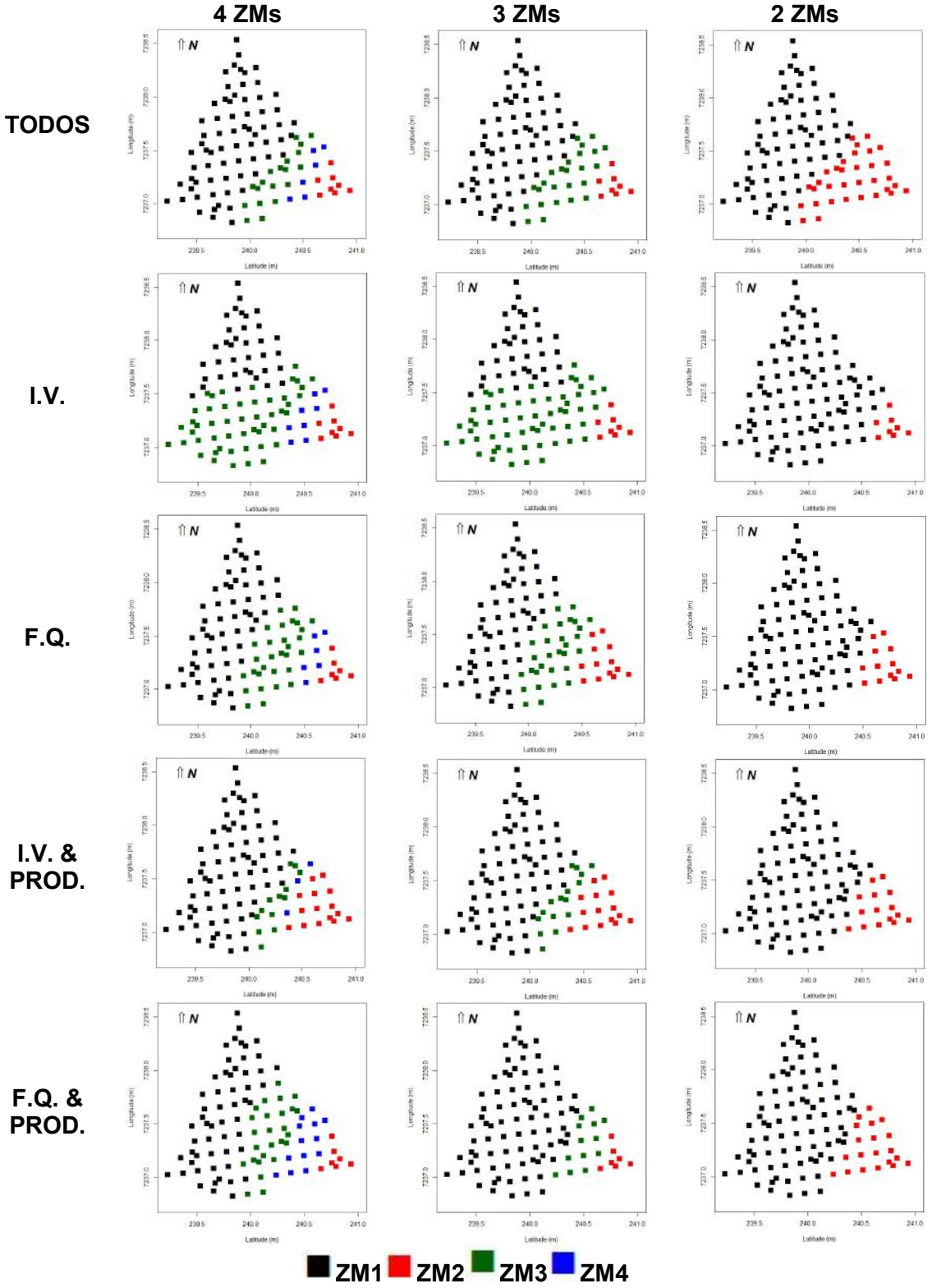

Figura 9 Mapas de Zonas de Manejo para os cinco subconjuntos de variáveis para o ano-safra de 2016/2017 (TODOS: subconjunto com todas as variáveis; I.V.: subconjunto formado somente pelas variáveis relacionadas aos índices vegetativos; F.Q.: subconjunto formado somente pelas variáveis relacionadas às variáveis físico-químicas do solo; I.V. & PROD.: subconjunto formado pelas variáveis relacionadas aos índices vegetativos e à produtividade; F.Q.&PROD.: subconjunto formada pelas variáveis relacionadas às variáveis físico-químicas do solo e produtividade; ZM: zona de manejo)

5.2 Análise dos subconjuntos e escolha das zonas de manejo

Os índices C, SD, DUNN e DB, bem como os coeficientes de silhueta médio e a cofenética foram utilizados para avaliar quais dos subconjuntos de variáveis formados em cada ano-safra geraram os melhores agrupamentos e a quantidade ótima de grupos que deveriam ser utilizados em cada situação.

Os subgrupos escolhidos de acordo com os índices propostos foram semelhantes somente para os anos-safra de 2015/2016 e 2016/2017, sendo diferentes nos outros anos. Assim, de modo geral, percebe-se que não houve um conjunto específico de variáveis que se sobressaiu sobre o outro em termos de definição das zonas de manejo, ou seja, o subconjunto que melhor agrupou os dados variou conforme o ano-safra. Além disso, o número ótimo de zonas de manejo para a área em estudo foi igual a dois em todos os anos-safra que fizeram parte desta pesquisa.

Para o ano-safra 2013/2014, as zonas de manejo formadas pelo subconjunto de variáveis relacionadas aos índices vegetativos e à produtividade apresentaram os melhores resultados para a maioria dos índices requeridos na escolha do subconjunto que melhor agrupou os dados (Tabela 5).

Tabela 5 Avaliação de formação de grupos pelos índices de ajuste de agrupamentos para o ano-safra 2013/2014

			SUBCONJUNTOS			
ÍNDICES	**Nº ZM**	**TODOS**	**I.V.**	**F.Q.**	**I.V. PROD.**	**F.Q. PROD.**
C.S.M.	**4 ZM**	0,139	0,135	0,083	0,127	0,069
	3 ZM	0,158	0,158	0,084	0,139	0,074
	2 ZM	0,217	0,209	0,168	**0,233**	0,104
C	**4 ZM**	0,705	0,726	**0,360**	0,716	0,743
	3 ZM	0,701	0,712	0,735	0,665	0,750
	2 ZM	0,572	0,591	0,695	0,678	0,654
SD	**4 ZM**	9,553	7,165	4,352	4,422	12,703
	3 ZM	4,601	3,795	4,574	3,402	7,701
	2 ZM	4,080	3,406	3,303	**2,535**	6,029
DUNN	**4 ZM**	**0,905**	0,883	0,803	0,839	0,896
	3 ZM	0,894	0,860	0,797	0,795	0,891
	2 ZM	0,755	0,759	0,763	0,796	0,804
DB	**4 ZM**	1,852	1,583	1,298	1,210	2,415
	3 ZM	1,022	0,941	1,459	1,143	1,623
	2 ZM	0,888	0,904	0,991	**0,782**	1,298
C.C.C.	**-**	0,787	0,780	0,751	**0,830**	0,622

C.S.M.: Coeficiente de Silhueta Médio; C.C.C.: Coeficiente de Correlação Cofenética; I.V.: índice vegetativo; F.Q.: variáveis físico-químicas do solo; PROD.: produtividade da soja; ZM: Zonas de Manejo; Valores em vermelhos representam os melhores resultados.

Para o ano-safra de 2014/2015, (Tabela 6), o subconjunto que apresentou os melhores resultados de acordo com as métricas estabelecidas foi o que utilizou todas as variáveis selecionadas, sendo este um resultado que difere do ano-safra anterior.

Tabela 6 Avaliação de formação de grupos pelos índices de ajuste de agrupamentos para o ano-safra 2014/2015

ÍNDICES	Nº ZM	TODOS	SUBCONJUNTOS I.V.	F.Q.	I.V. PROD.	F.Q. PROD.
C.S.M.	4 ZM	0,14	0,13	0,03	0,13	0,05
	3 ZM	0,13	0,15	0,04	0,15	0,04
	2 ZM	**0,25**	0,20	0,09	0,20	0,10
C	4 ZM	0,22	0,60	0,50	0,61	**0,21**
	3 ZM	0,48	0,59	0,52	0,59	0,54
	2 ZM	0,46	0,51	0,52	0,52	0,52
SD	4 ZM	3,40	8,77	12,58	8,69	8,09
	3 ZM	3,15	5,14	10,96	5,09	9,79
	2 ZM	**2,30**	3,82	7,47	3,77	6,46
DUNN	4 ZM	0,63	**0,91**	0,87	0,90	0,87
	3 ZM	0,62	0,88	0,87	0,87	0,86
	2 ZM	0,62	0,78	0,85	0,78	0,83
DB	4 ZM	1,02	1,90	3,05	1,89	2,33
	3 ZM	1,12	1,19	2,79	1,18	2,79
	2 ZM	**0,75**	0,91	1,69	0,90	1,65
C.C.C.	-	0,814	0,827	0,733	**0,828**	0,723

C.S.M.: Coeficiente de Silhueta Médio; C.C.C.: Coeficiente de Correlação Cofenética; I.V.: índice vegetativo; F.Q.: variáveis físico-químicas do solo; PROD.: produtividade da soja; ZM: Zonas de Manejo; Valores em vermelhos representam os melhores resultados.

Tabela 7 Avaliação de formação de grupos pelos índices de ajuste de agrupamentos para o ano-safra 2015/2016

ÍNDICES	Nº ZM	TODOS	SUBCONJUNTOS I.V.	F.Q.	I.V. PROD.	F.Q. PROD.
C.S.M.	4 ZM	0,05	0,07	0,06	0,08	0,03
	3 ZM	0,06	0,10	0,05	0,07	0,05
	2 ZM	**0,25**	0,15	0,09	0,10	0,07
C	4 ZM	0,53	0,54	**0,39**	0,29	0,50
	3 ZM	0,52	0,55	0,56	0,63	0,52
	2 ZM	0,42	0,43	0,55	0,63	0,43
SD	4 ZM	5,64	4,92	4,87	4,30	7,98
	3 ZM	4,54	3,33	6,22	4,39	7,09
	2 ZM	3,45	**3,11**	4,41	3,37	5,78
DUNN	4 ZM	0,37	0,55	0,80	0,79	0,72
	3 ZM	0,35	0,53	0,78	0,77	0,71
	2 ZM	0,52	0,42	**0,82**	0,73	0,65
DB	4 ZM	0,53	0,54	1,75	**0,29**	2,25
	3 ZM	0,52	0,55	2,10	0,63	1,85
	2 ZM	0,42	0,43	1,66	0,63	1,90
C.C.C.	-	0,61	0,51	**0,78**	0,49	0,44

C.S.M.: Coeficiente de Silhueta Médio; C.C.C.: Coeficiente de Correlação Cofenética; I.V.: índice vegetativo; F.Q.: variáveis físico-químicas do solo; PROD.: produtividade da soja; ZM: Zonas de Manejo; Valores em vermelhos representam os melhores resultados.

De acordo com os resultados apresentados na Tabela 7, para o ano-safra de 2015/2016, diferente dos dois anos-safra anteriores, o subconjunto formado pelas variáveis relacionadas às variáveis físico-químicas do solo, foram as que melhor agruparam os dados disponíveis.

O ano-safra de 2016/2017 assim como o ano-safra anterior tiveram os subconjuntos formados pelas variáveis referentes aos atributos físico-químicos como os que apresentaram as melhores formações de zonas de manejo (Tabela 8).

Tabela 8 Avaliação de formação de grupos pelos índices de ajuste de agrupamentos para o ano-safra 2016/2017

			SUBCONJUNTOS			
ÍNDICES	**Nº ZM**	**TODOS**	**I.V.**	**F.Q.**	**I.V. PROD.**	**F.Q. PROD.**
C.S.M.	**4 ZM**	0,21	0,06	0,13	0,12	0,10
	3 ZM	0,24	0,07	0,13	0,13	0,19
	2 ZM	0,28	**0,39**	0,27	0,23	0,24
C	**4 ZM**	0,51	0,56	0,52	0,53	0,53
	3 ZM	0,50	0,58	0,47	0,53	0,56
	2 ZM	0,41	**0,40**	0,49	0,54	0,54
SD	**4 ZM**	3,66	4,92	3,83	8,50	4,15
	3 ZM	2,50	4,26	3,28	5,14	3,50
	2 ZM	2,37	2,53	**2,33**	3,95	2,53
DUNN	**4 ZM**	0,65	0,40	0,77	0,79	0,72
	3 ZM	0,62	0,37	0,72	0,79	0,84
	2 ZM	0,49	0,66	0,81	**0,85**	0,84
DB	**4 ZM**	0,97	1,30	1,10	1,70	1,19
	3 ZM	0,78	1,21	0,98	1,19	0,94
	2 ZM	0,83	0,73	**0,67**	0,75	0,78
C.C.C.	**-**	0,74	0,65	**0,88**	**0,89**	0,85

C.S.M.: Coeficiente de Silhueta Médio; C.C.C.: Coeficiente de Correlação Cofenética; I.V.: índice vegetativo; F.Q.: variáveis físico-químicas do solo; PROD.: produtividade da soja; ZM: Zonas de Manejo; Valores em vermelhos representam os melhores resultados.

O uso de conjunto específico de variáveis é comum em estudos de zonas de manejo. Como exemplo, podemos citar Kuiawski et al. (2017), que se valeram de índices vegetativos obtidos em diferentes estádios fenológicos da soja para gerar mapas com zonas de manejo e, assim, determinar em qual estádio vegetativo da soja o mapa melhor concordou com o mapa da produtividade. Em seguida, os autores estudaram as propriedades químicas do solo nas zonas de manejo selecionadas. Nesse estudo, assim como em nosso estudo, o número ótimo de zonas de manejo geradas também foi igual a dois.

De acordo com os resultados apresentados anteriormente, os mapas escolhidos com suas respectivas zonas de manejo para cada ano-safra são apresentados na Figura 10.

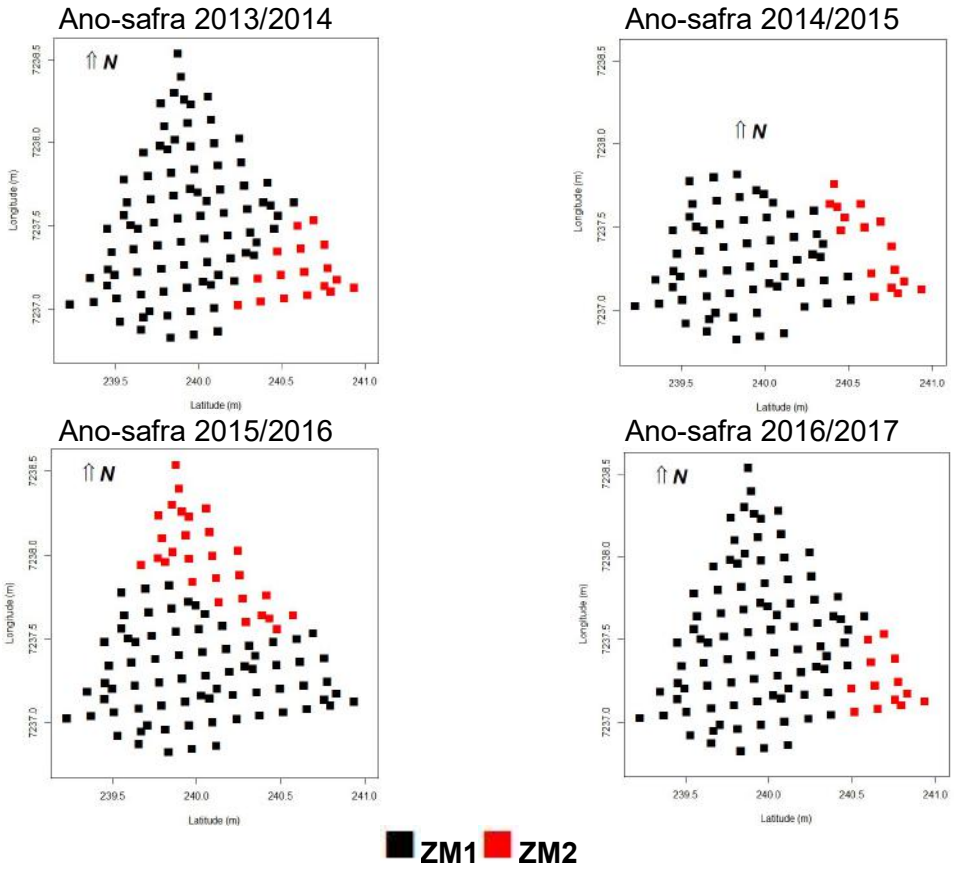

Figura 10 Mapas com suas respectivas zonas de manejo para os anos-safra de 2013/2014, 2014/2015, 2015/2016 e 2016/2017

Ressalta-se, novamente, que há formação de uma zona de manejo em todos os anos-safra ocupando uma região muito maior que a outra (cor preta, Figura 10). A ZM1 foi maior, proporcionalmente, para o ano-safra de 2016/2017, ao ocupar 149,94 hectares (90% do talhão ou 89 pontos amostrais). Em seguida, vieram os anos-safra de 2014/2015 com 103,72 hectares (84% da área total ou 61 pontos amostrais) e o ano-safra de 2013/2014 com 139,18 hectares (83% do talhão ou 85 pontos amostrais). A menor área ocupada foi para o ano-safra de 2015/2016 com 123,38 hectares (73,56% da área total ou 74 pontos amostrais).

A outra região (cor vermelha, Figura 10) tem como característica a ocupação da menor área do talhão. Para os anos-safra de 2013/2014, 2014/2015 e 2016/2017, a ZM2 abrangeu respectivamente 28,54 hectares (17% da área total ou 17 pontos amostrais), 20,40 hectares (16% da área total ou 16 pontos amostrais) e 17,44 hectares (10% da área total ou 13 pontos amostrais), todas localizadas na região sudeste do talhão. Já para o ano-safra de 2015/2016 a ZM2, ocupou uma área um pouco maior do que os outros anos-safra com 44,34 hectares, o que representa 26% da região em estudo (28 pontos amostrais). Além disso, a localização dessa zona de manejo também se distinguiu dos outros anos-safra, pois se localizava ao norte do talhão.

5.3 Perfil das zonas de manejo formadas em cada ano-safra

Analisou-se então para cada ano-safra o comportamento das diversas variáveis que fizeram parte do estudo, utilizando-se para isso o gráfico boxplot mapas temáticos. Assim, foi possível traçar um perfil para cada uma das zonas de manejo delineadas.

Em relação às propriedades químicas, nota-se que algumas variáveis foram importantes na diferenciação das zonas de manejos em, pelo menos, mais de um ano-safra, são eles: Zinco, Manganês, Potássio e Carbono. O Zinco foi o que mais se destacou por apresentar características importantes para a diferenciação das zonas de manejo em três anos-safra, 2013/2014, 2014/2015 e 2015/2016, (Figuras 11, 12, 13, 14, 15 e 16), lembrando ainda, que no último ano-safra os valores dessa variável não estavam disponíveis para o estudo. O Manganês, o Potássio, o Carbono e o Cobre apresentam diferenças importantes entre as zonas de manejo em dois anos-safra 2013/2014 e 2015/2016 para o Manganês (Figuras 11, 12, 15 e 16), 2013/2014 e 2016/2017 para o Potássio (Figuras 11, 12, 17 e 18), 2015/2016 e 2016/2017 para o Carbono (Figuras 15, 16, 17 e 18) e 2013/2014 e 2015/2016 para o Cobre (Figuras 11, 12, 15 e 16).

De acordo com as Figuras 11 e 12, para o ano-safra de 2013/2014, a ZM2 apresentou valores menores de Manganês (Mn) e Zinco (Zn) e uma grande porcentagem de área com valores altos de Potássio (K) e Cobre (Cu).

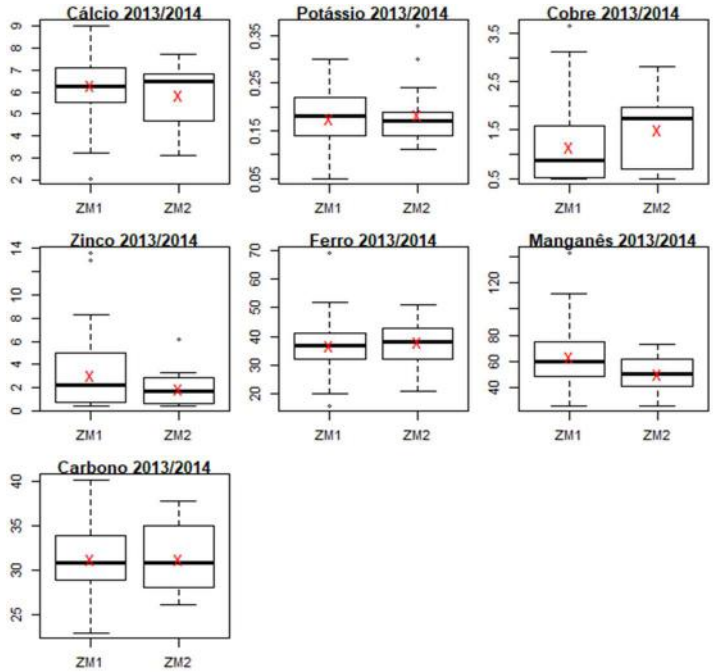

Figura 11 Boxplot para as zonas de manejo em relação às variáveis químicas (ano-safra 2013/2014) (X em vermelho representa a média)

A disponibilidade das variáveis Potássio (entre 0,11 e 0,30 cmol/dm^3), Zinco (> 1,60 mg/dm^3) e Manganês (> 5,10 mg/dm^3) no solo são consideradas como média para ambas as zonas de manejo, com exceção da variável Potássio na qual há regiões dentro da ZM1 com valores abaixo de 0,10 cmol/dm^3, o que é considerado baixo. A variável Cobre apresenta valores considerados baixos (entre 0,31 a 0,99 mg/dm^3) e médios (> 1 mg/dm^3) nas duas zonas de manejo, entretanto, é possível notar tanto pelo boxplot (Figura 11) quanto pelo mapa (Figura 12) que os valores mais baixos para essa variável pertencem à ZM1.

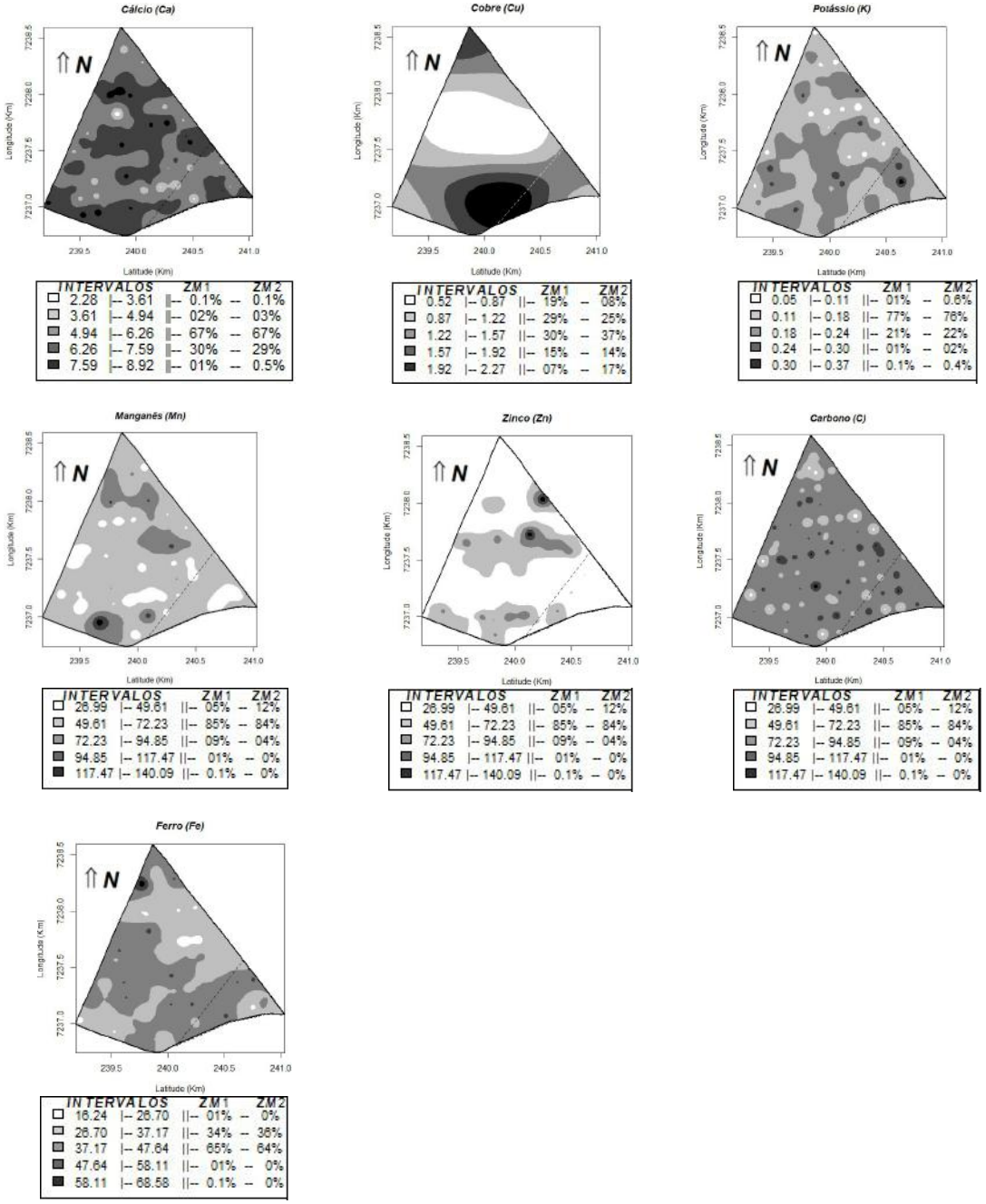

Figura 12 Mapas das variáveis químicas com as respectivas zonas de manejo para o ano-safra de 2013/2014 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Em relação ao ano-safra de 2014/2015, nota-se que as principais diferenças entre as duas zonas de manejo foram em relação a algumas variáveis relacionadas às propriedades químicas do solo como o teor de Zinco (Zn), Cálcio (Ca), Magnésio (Mg) e Alumínio (Al) (Figura 13).

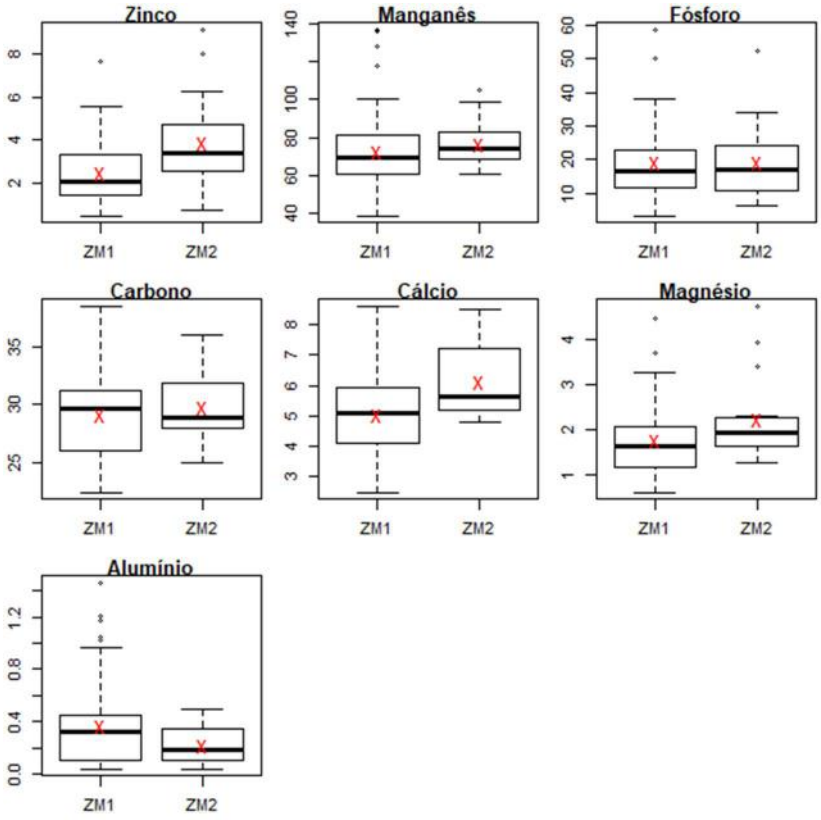

Figura 13 Boxplot para variáveis relacionadas às propriedades químicas do solo (ano-safra 2014/2015) (X em vermelho representa a média)

A concentração do teor de Alumínio no solo apresenta, em sua maioria, valores abaixo de 0,5 cmol/dm^3 o que é considerado baixo, no entanto, há regiões dentro da ZM1 que possuem concentrações cujas respostas variam entre média e alta. As concentrações da variável Zinco (> 1,6 mg/dm^3) são consideradas médias nas duas zonas de manejo. Os teores de Magnésio (> 0,80 cmolc/dm^3) e Cálcio (> 4 cmolc/dm^3) possuem concentrações consideradas entre alta e muito alta em ambas as zonas de manejo, segundo a classificação de (SEAB, 1989). Esses resultados podem ser visualizados com a análise dos mapas da Figura 14.

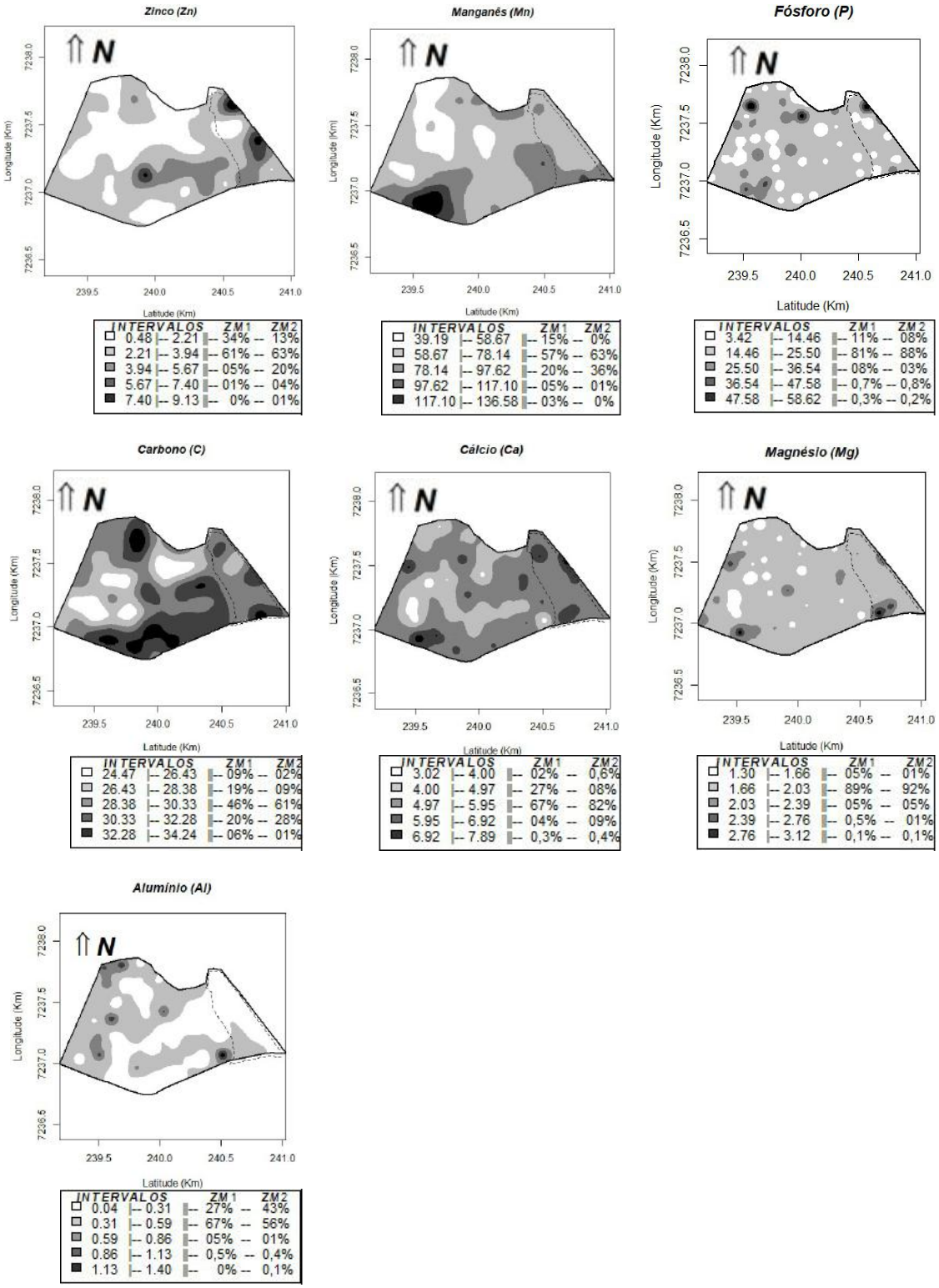

Figura 14 Mapas das variáveis químicas com as respectivas zonas de manejo para o ano-safra de 2014/2015 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

De modo geral, para o ano-safra de 2015/2016 observa-se que a ZM1 possui valores menores quando comparadas à ZM2, em relação ao teor de Cobre (Cu), Zinco (Zn) e de Manganês (Mn) no solo, entretanto, possui valores um pouco maiores para o teor de Carbono (C) no solo (Figura 15).

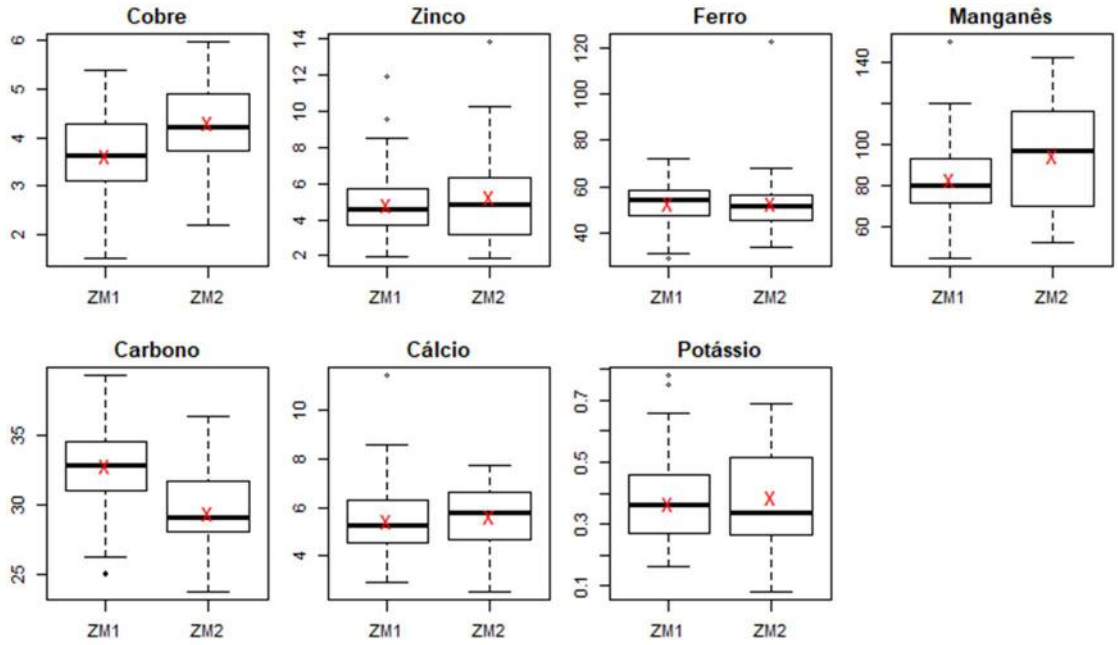

Figura 15 Boxplot para variáveis químicas para o ano-safra 2015/2016 (X em vermelho representa a média)

De acordo com os dados do SEAB (1989), a área em estudo apresentou concentrações médias para as variáveis Cobre (> 1 mg/dm^3), Zinco (> 1,60 mg/dm^3) e Manganês (> 5,10 mg/dm^3) em ambas zonas de manejo. A variável Zinco está dentro da faixa considerada boa para o cultivo em quase toda a área de estudo (entre 1,5 a 6 mg/dm^3); já as variáveis Cobre e Manganês estão com valores acima da faixa considerada boa para o cultivo (1,5 a 2 mg/dm^3 para o Cobre e 5 a 30 mg/dm^3 para o Manganês) e próximo do valor considerado excessivo que é de 8 mg/dm^3 para o Cobre e de 150 mg/dm^3 para o Manganês. O teor de Carbono possui concentrações consideradas elevadas em ambas as zonas de manejo, entretanto, existem áreas dentro dessas zonas que possuem concentrações consideradas médias para essa variável, conforme Figura 16.

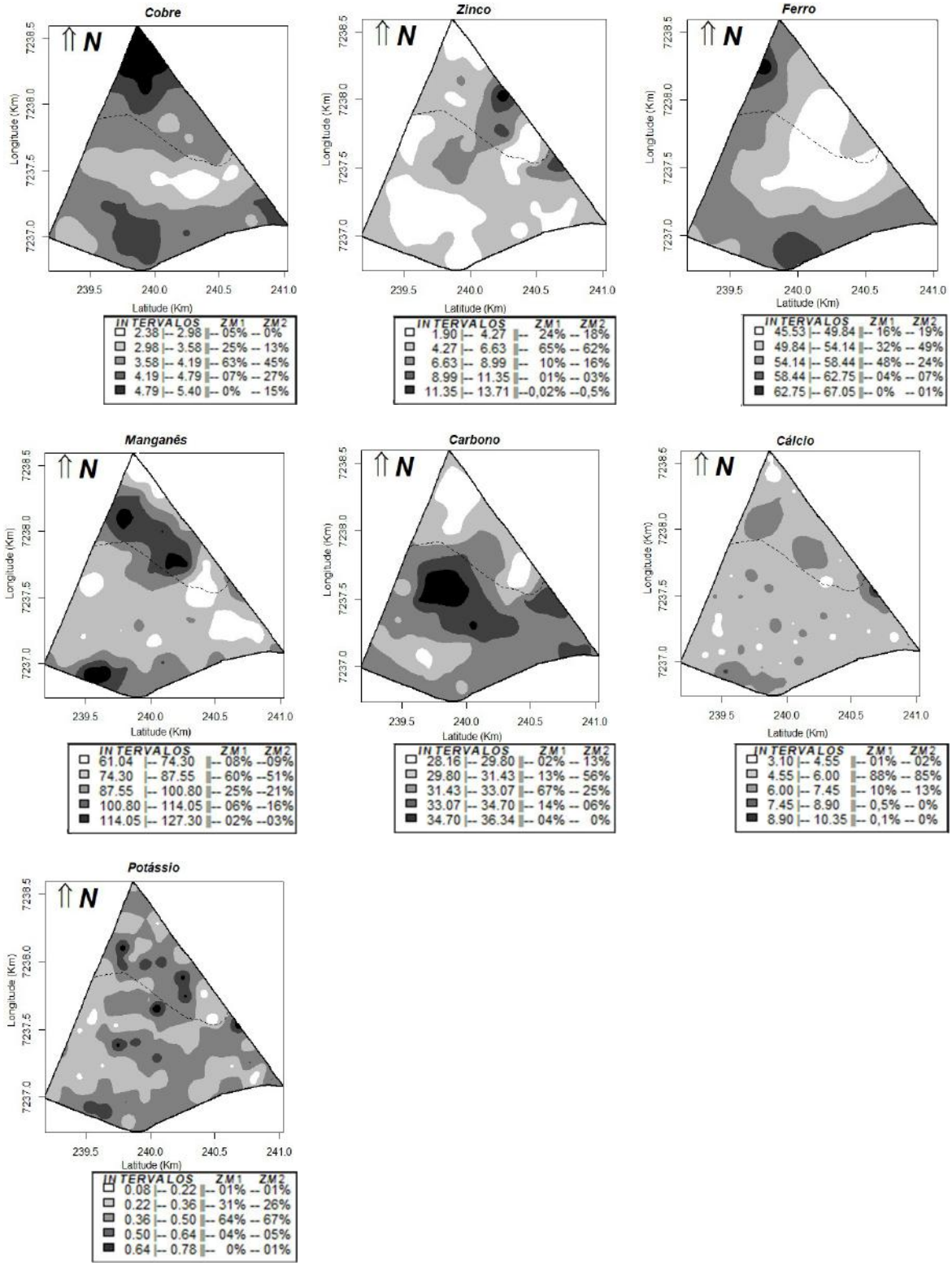

Figura 16 Mapas das variáveis químicas com as respectivas zonas de manejo para o ano-safra de 2015/2016 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Para o ano-safra de 2016/2017, observa-se que, de modo geral, a ZM1 apresentou valores maiores para os teores de Carbono (C), Fósforo (P), Potássio (K) no solo em relação

à ZM2. As demais variáveis não apresentaram diferenças relevantes entre as zonas de manejo (Figura 17).

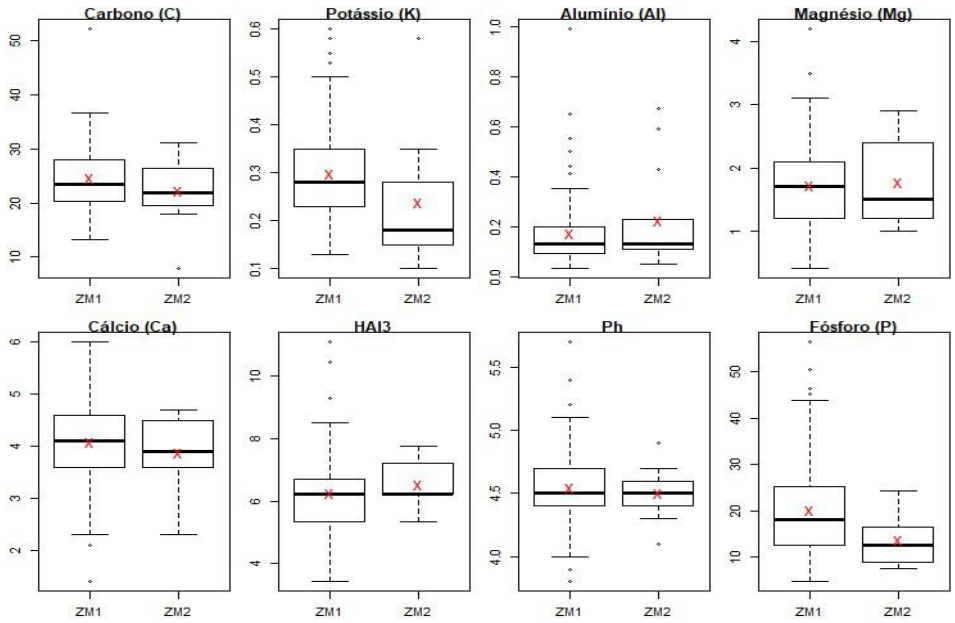

Figura 17 Boxplot para variáveis químicas para o ano-safra 2016/2017 (X em vermelho representa a média)

De acordo com a classificação de SEAB (1989), a variável Potássio apresenta regiões onde suas concentrações podem ser consideradas média (0,11 – 0,20 cmolc/dm^3), elevada (0,21 – 0,30 cmolc/dm^3) e muito elevada (> 0,30 cmolc/dm^3). As concentrações mais elevadas são encontradas na ZM1 e as menores estão na ZM2. O Carbono apresenta valores considerados elevados (20 - 35 g/dm^3) na maior parte das duas zonas de manejo, entretanto, em partes das duas zonas de manejo, podem ser encontrados valores abaixo desse intervalo (concentração média) e, em parte da ZM1, podem ser encontrados valores acima desse intervalo. A concentração para a variável Fósforo é considerada muito elevada (> 9 mg/dm^3) para as duas zonas de manejo, e valores maiores podem ser vistos na ZM1. Esses resultados também podem ser visualizados nas Figuras 17 e 18.

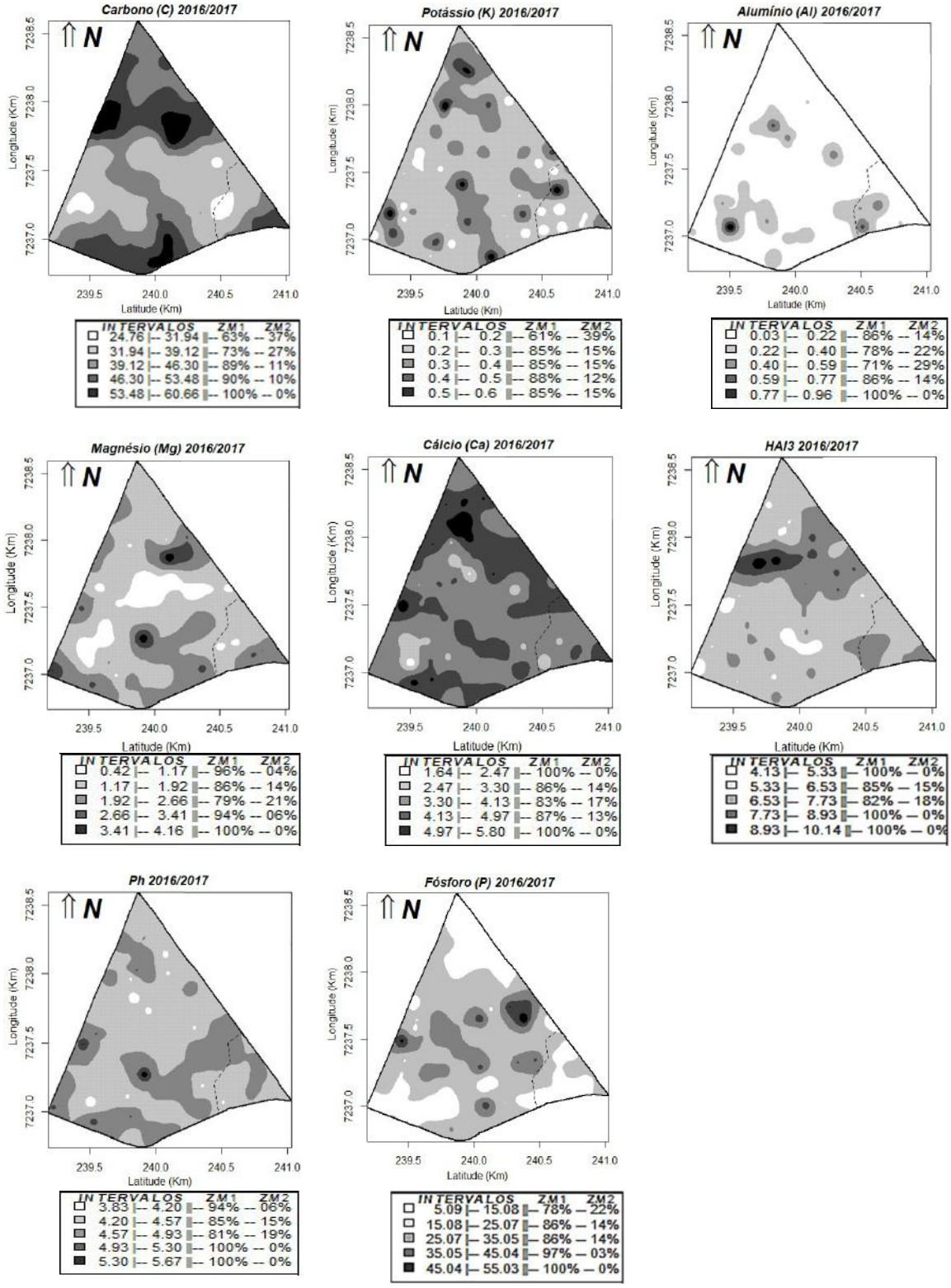

Figura 18 Mapas das variáveis químicas com as respectivas zonas de manejo para o ano-safra de 2016/2017 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Em relação às propriedades físicas do solo, nota-se que, de modo geral, a maioria das variáveis não apresentou dependência espacial média ou forte e das que apresentaram e foram utilizadas na pesquisa, poucas contribuíram efetivamente para a diferenciação das

zonas em todos os anos-safra. As exceções foram para os anos-safra de 2015/2016 e 2016/2017. Entretanto, a principal contribuição dessas variáveis foi em relação à RSP na camada entre 0 a 10 cm de profundidade do ano-safra de 2016/2017. Em geral, os mapas para essas variáveis mostram que há pouca diferenciação de seus valores em todo o talhão.

De acordo com esses resultados para cada ano-safra, o ano-safra de 2013/2014 registrou umidade na camada entre 0 e 10 cm de profundidade um pouco maior na ZM2 em relação à ZM1.

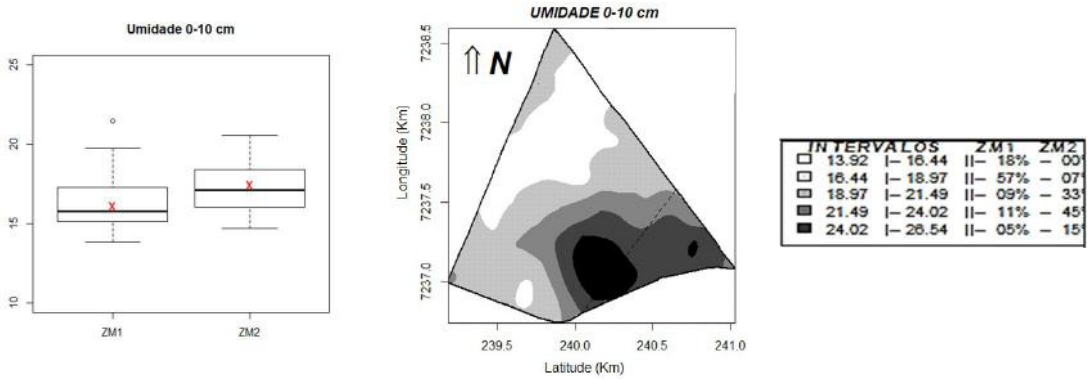

Figura 19 Boxplot e mapa para a variável umidade na camada entre 0 e 10 cm (ano-safra 2013/2014) (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Para o ano-safra de 2014/2015, a RSP na camada entre 20 a 30 cm de profundidade apresentou valores maiores na ZM2 em relação à ZM1. As demais variáveis relacionadas às propriedades físicas do solo não apresentaram diferenças relevantes entre as duas zonas de manejo, como pode ser visto nos gráficos da Figura 20 e nos mapas da Figura 21.

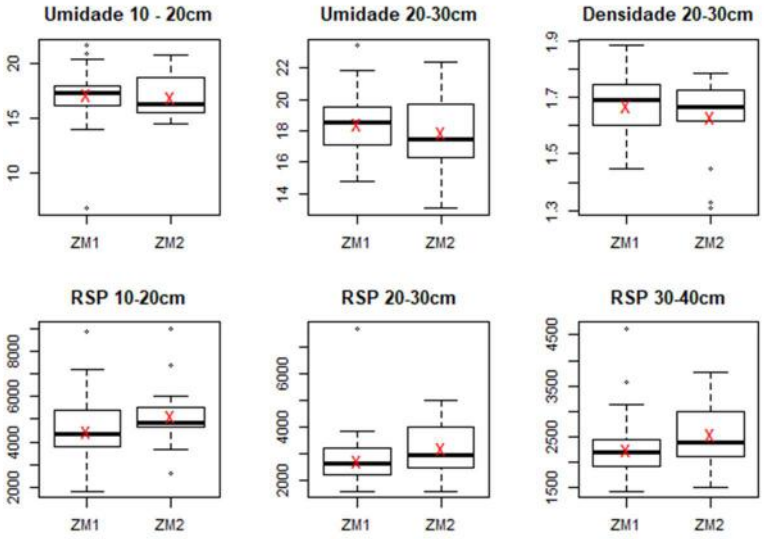

Figura 20 Boxplot para a variável umidade na camada entre 0 e 10 cm(ano-safra 2014/2015) (X em vermelho representa a média)

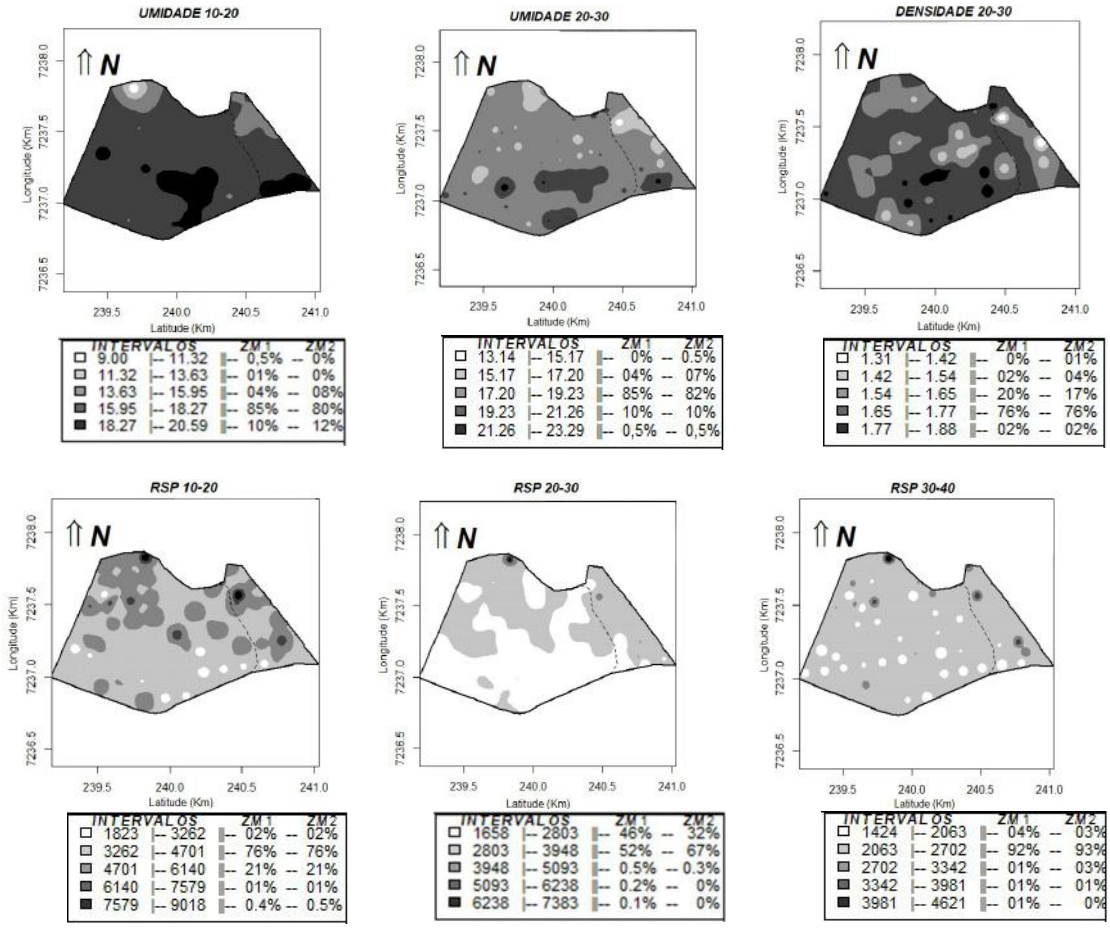

Figura 21 Mapas das variáveis físicas com as respectivas zonas de manejo para o ano-safra de 2014/2015 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Em geral, para o ano-safra de 2015/2016 observa-se que a ZM1 possui valores um pouco menores quando comparadas a ZM2, em relação à umidade na camada de 0 a 10 cm e de densidade na camada entre 21 a 30 cm de profundidade (Figuras 22 e 23).

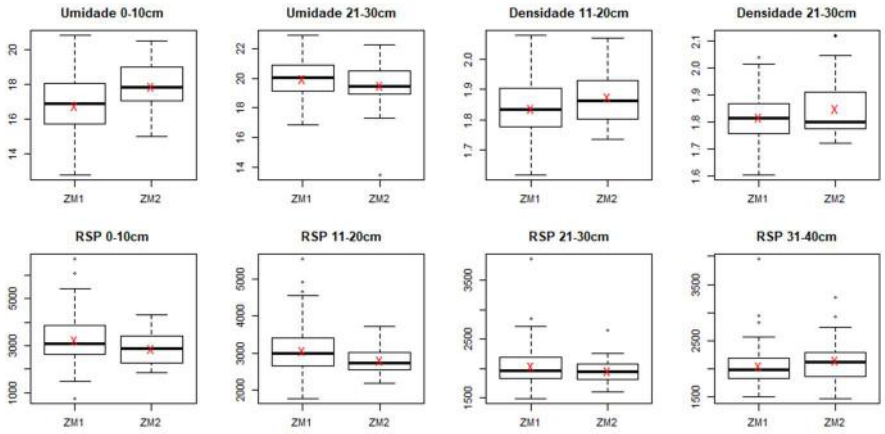

Figura 22 Boxplot para variáveis físicas para o ano-safra 2015/2016 (X em vermelho representa a média)

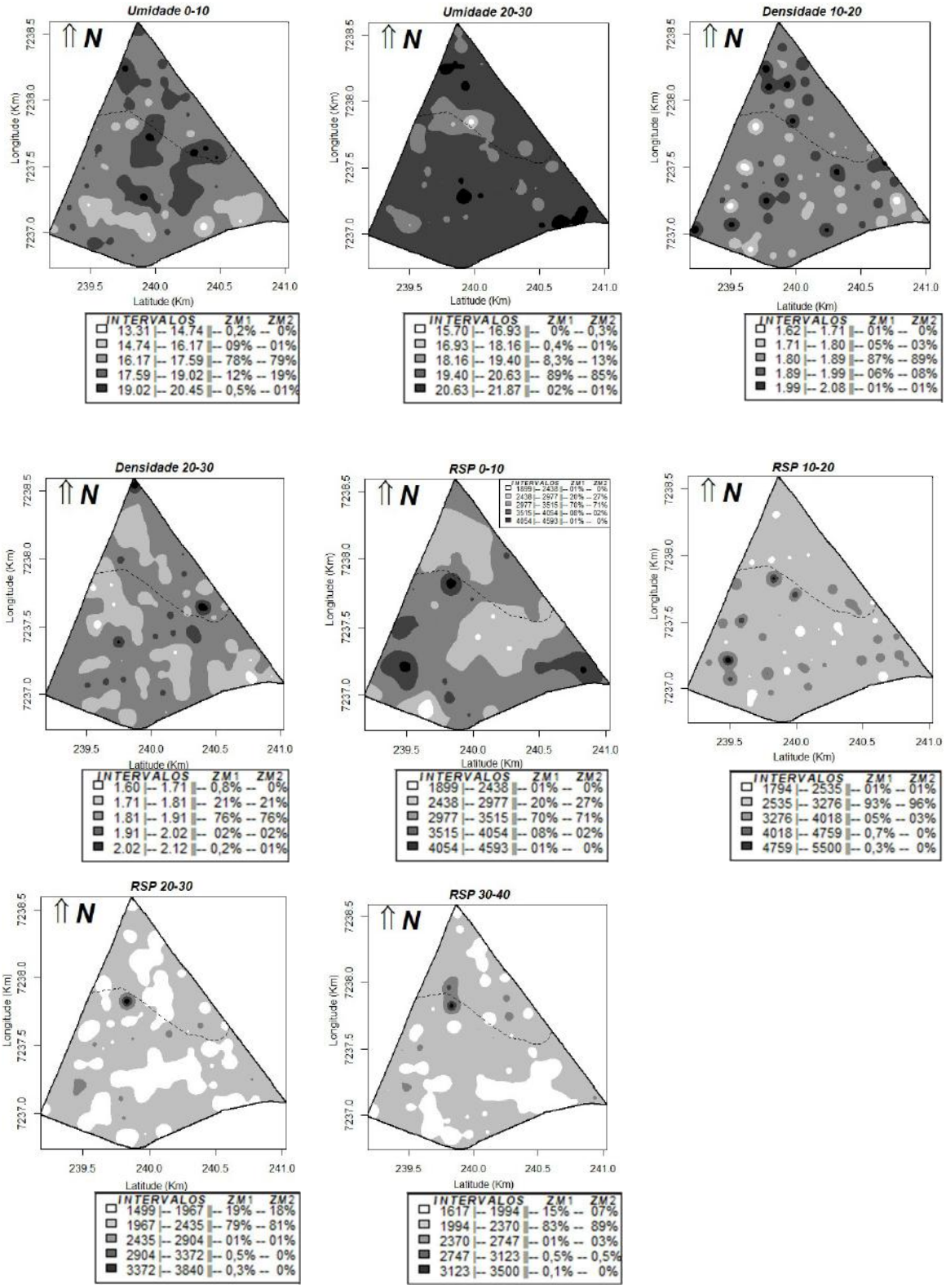

Figura 23 Mapas das variáveis físicas com as respectivas zonas de manejo para o ano-safra de 2015/2016 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Para o ano-safra de 2016/2017, percebe-se que a ZM1 apresentou valores maiores para a densidade na camada entre 10 a 20 cm e menores para a RSP na camada entre 0 a

10 cm de profundidade, já a umidade nas camadas de 10 a 20 cm não mostrou muita diferença entre as zonas de manejo. Pode-se dizer, então, que a ZM2 apresentou maior resistência do solo à penetração na camada mais superficial e menor densidade do solo na camada intermediária (Figuras 24 e 25).

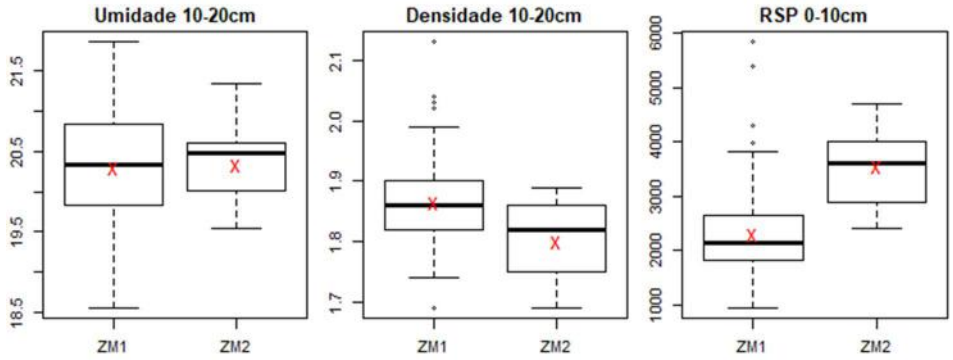

Figura 24 Boxplot para variáveis físicas do solo para o ano-safra 2016/2017 (X em vermelho representa a média)

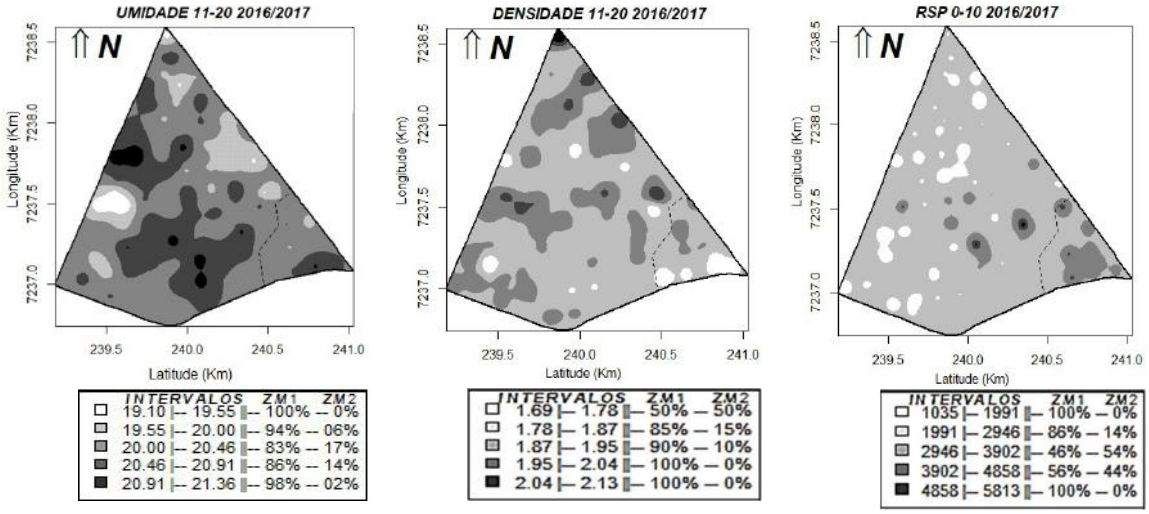

Figura 25 Mapas das variáveis físicas com as respectivas zonas de manejo para o ano-safra de 2016/2017 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Das variáveis relacionadas aos índices vegetativos, a fase de emergência (VE) foi a que apresentou diferenças entre as zonas de manejo em mais anos-safra, no qual em todos os casos apresentou valores maiores para a ZM1, indicando maior vigor da planta nesse período para essas regiões. Para as outras fases do ciclo reprodutivo da soja que apresentaram diferenças relevantes entre as zonas de manejo, a ZM1 apresentou valor maior para os índices calculados, com exceção da fase R2 do ano-safra de 2013/2014.

A partir da análise para cada ano-safra, é possível observar que, para o ano-safra de 2013/2014, a ZM2 apresentou valores menores em relação aos índices vegetativos nas

fases VE (emergência), R6 (pleno enchimento das vagens) e R7 (início da maturação). Isso pode representar menor vigor vegetativo nas plantas ou menor área de cobertura vegetal nessa região e valores maiores na R2 (pleno florescimento) (Figuras 26, 27 e 28).

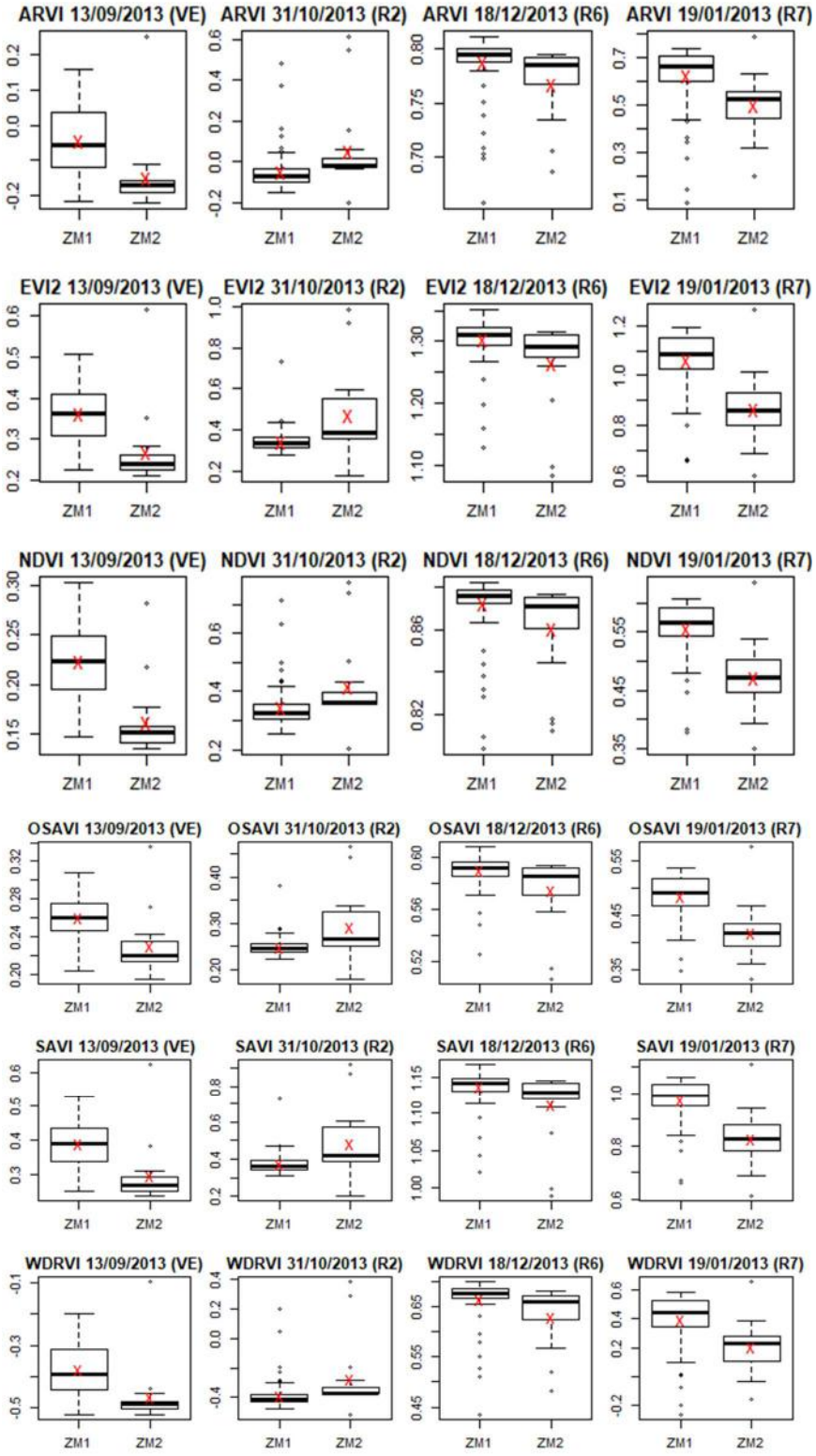

Figura 26 Boxplot para as zonas de manejo em relação aos índices vegetativos (X em vermelho representa a média)

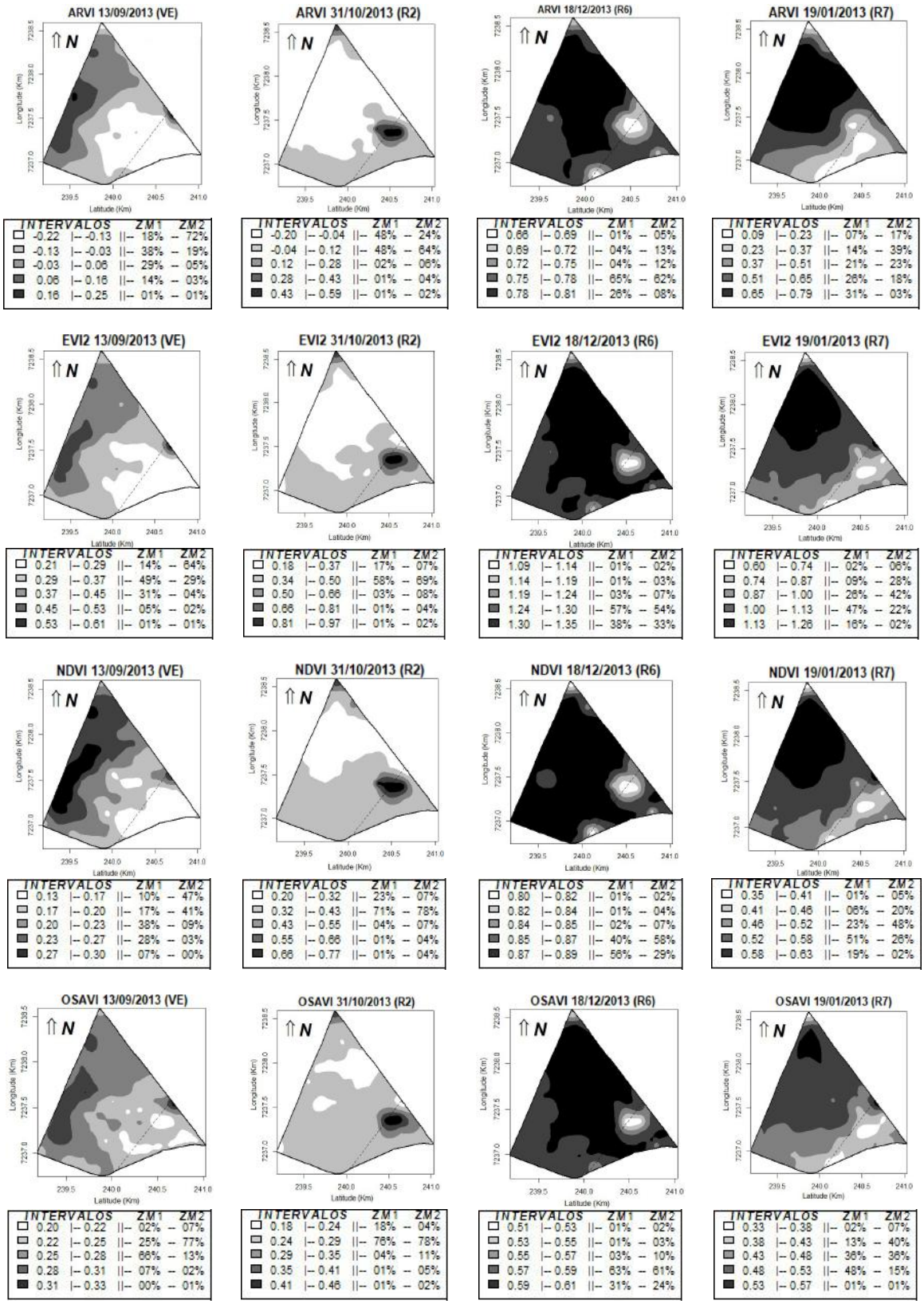

Figura 27 Mapas dos índices vegetativos com as respectivas zonas de manejo para o ano-safra de 2013/2014 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

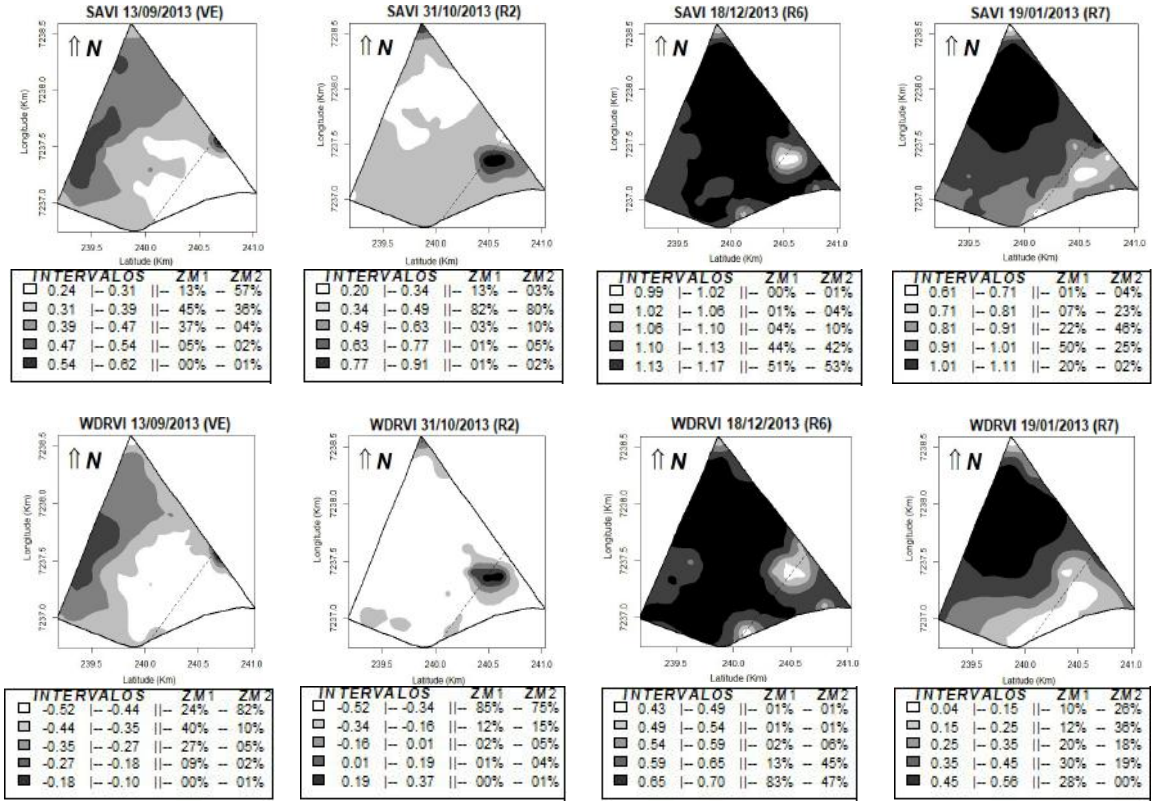

Figura 28 Mapas dos índices vegetativos com as respectivas zonas de manejo para o ano-safra de 2013/2014 (continuação) (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Para o ano-safra de 2014/2015, observa-se que a ZM1 possui valores maiores que a ZM2 na fase de emergência da planta (VE) e no estádio R4 (plena formação das vagens) (Figuras 29, 30 e 31).

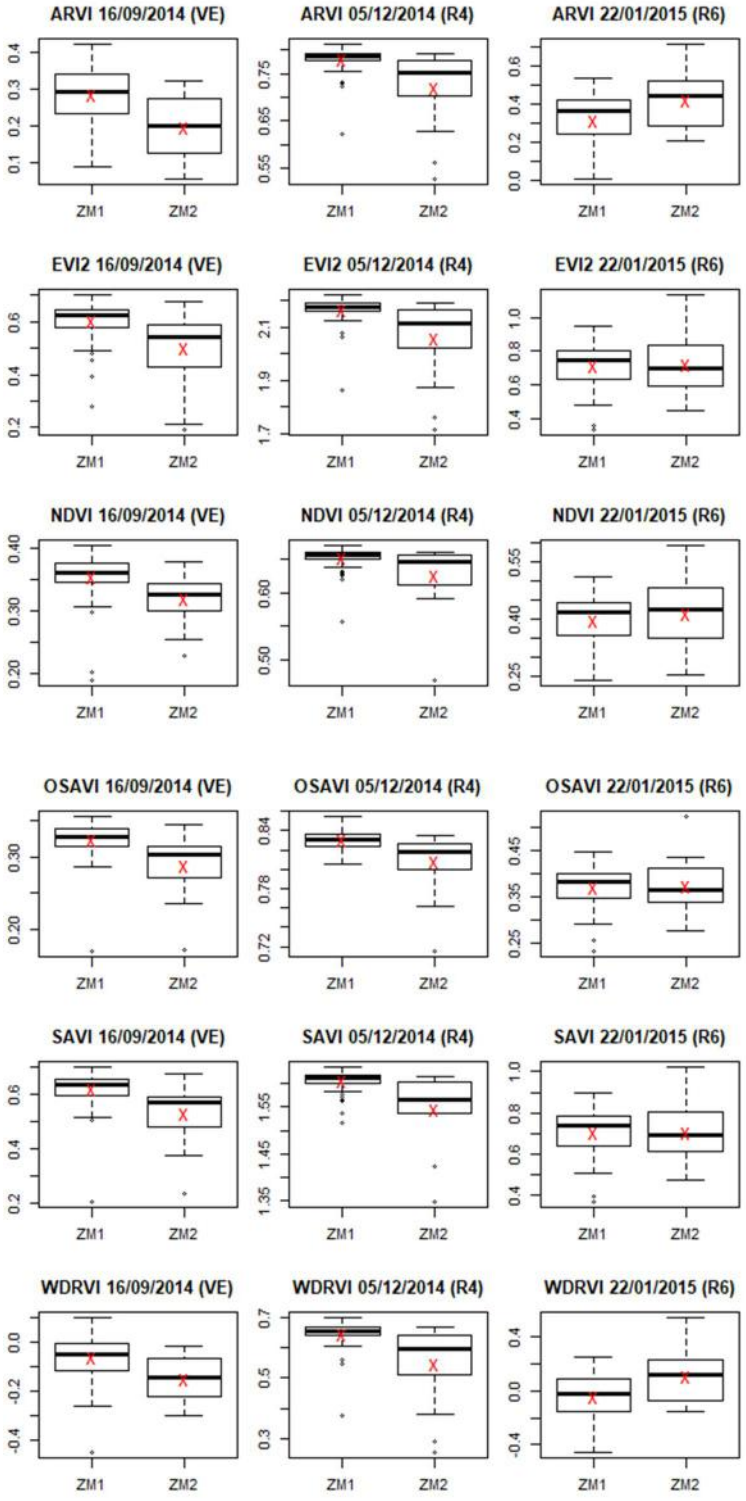

Figura 29 Boxplot para as zonas de manejo em relação aos índices vegetativos (ano-safra 2014/2015) (X em vermelho representa a média)

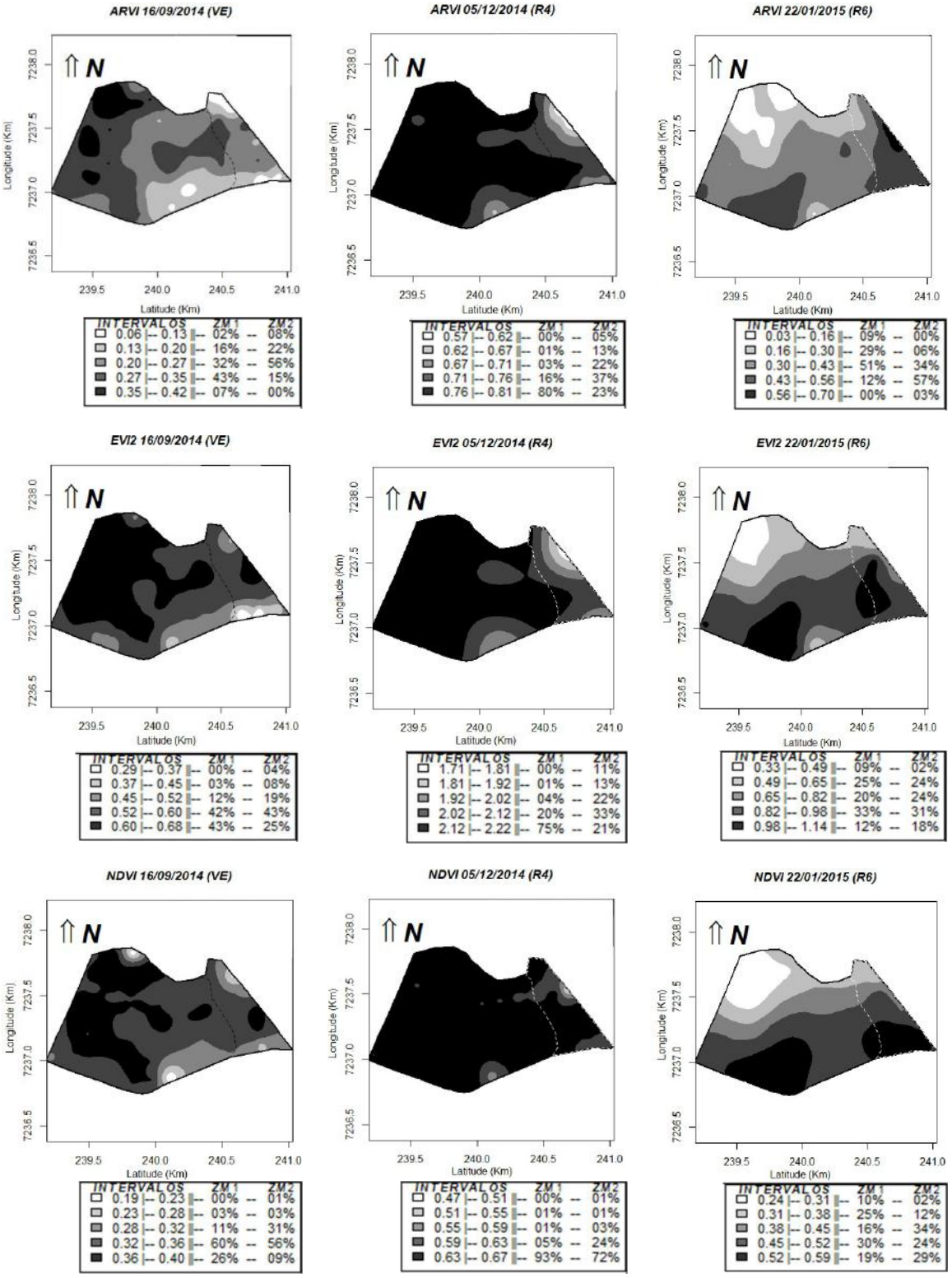

Figura 30 Mapas dos índices vegetativos com as respectivas zonas de manejo para o ano-safra de 2014/2015 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

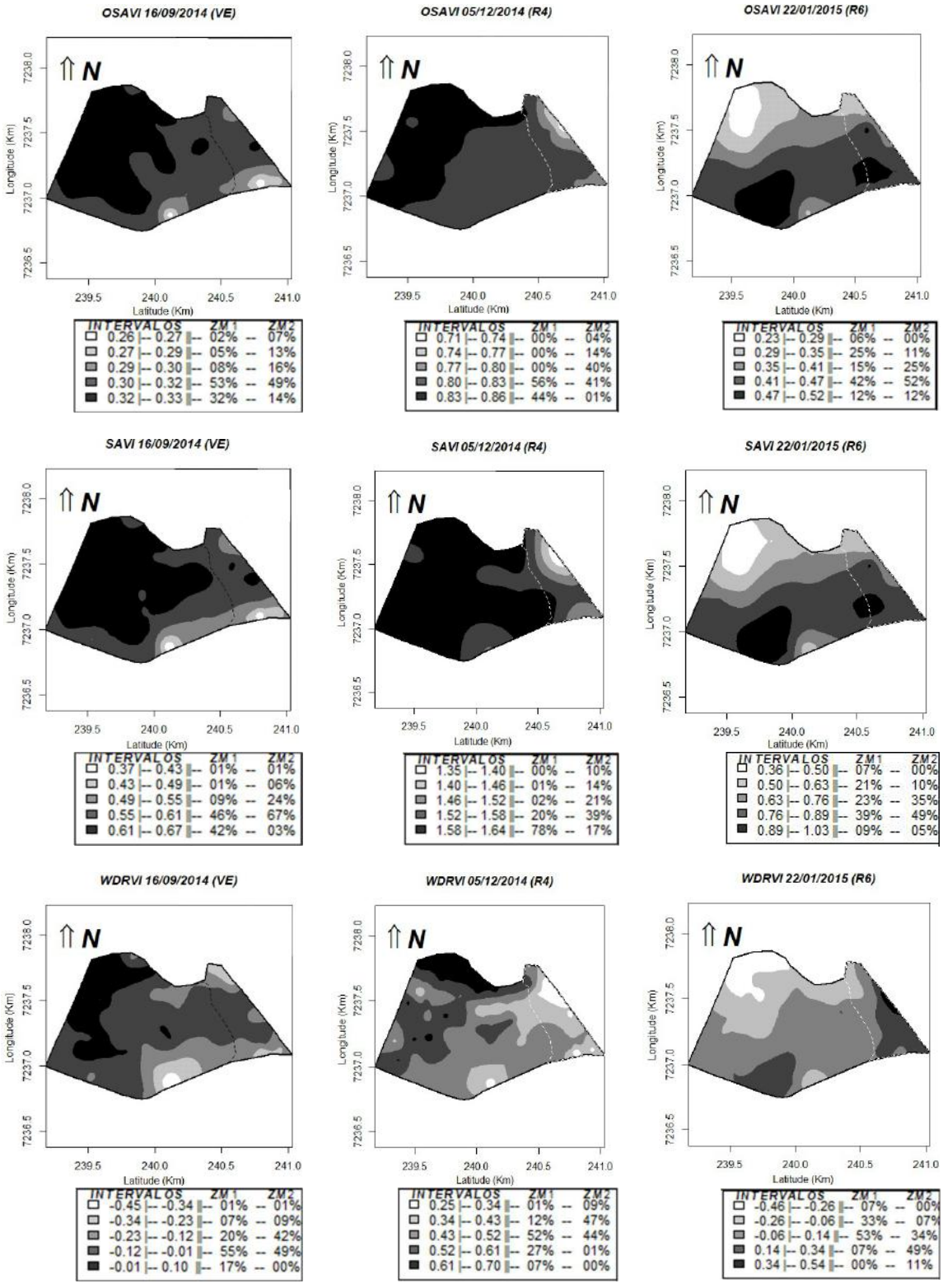

Figura 31 Mapas dos índices vegetativos com as respectivas zonas de manejo para o ano-safra de 2014/2015 (continuação) (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Para o ano-safra de 2015/2016, os índices vegetativos apresentam concentração maior para valores menores na ZM2 em comparação à ZM1 nos dois estádios analisados, indicando que o cultivar na ZM1 apresentou maior vigor vegetativo em relação à ZM2 (Figuras 32, 33 e 34).

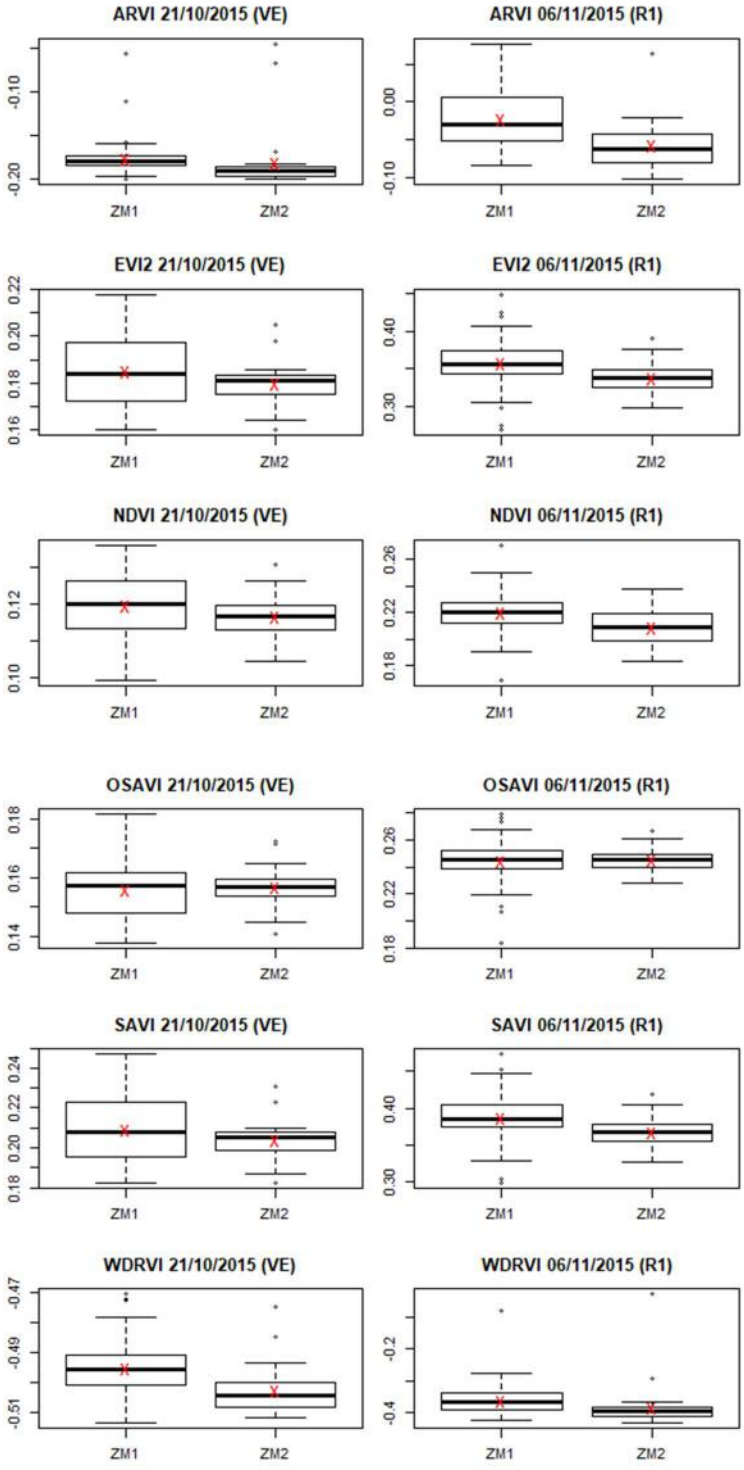

Figura 32 Boxplot para as zonas de manejo em relação aos índices vegetativos (ano-safra 2015/2016) (X em vermelho representa a média)

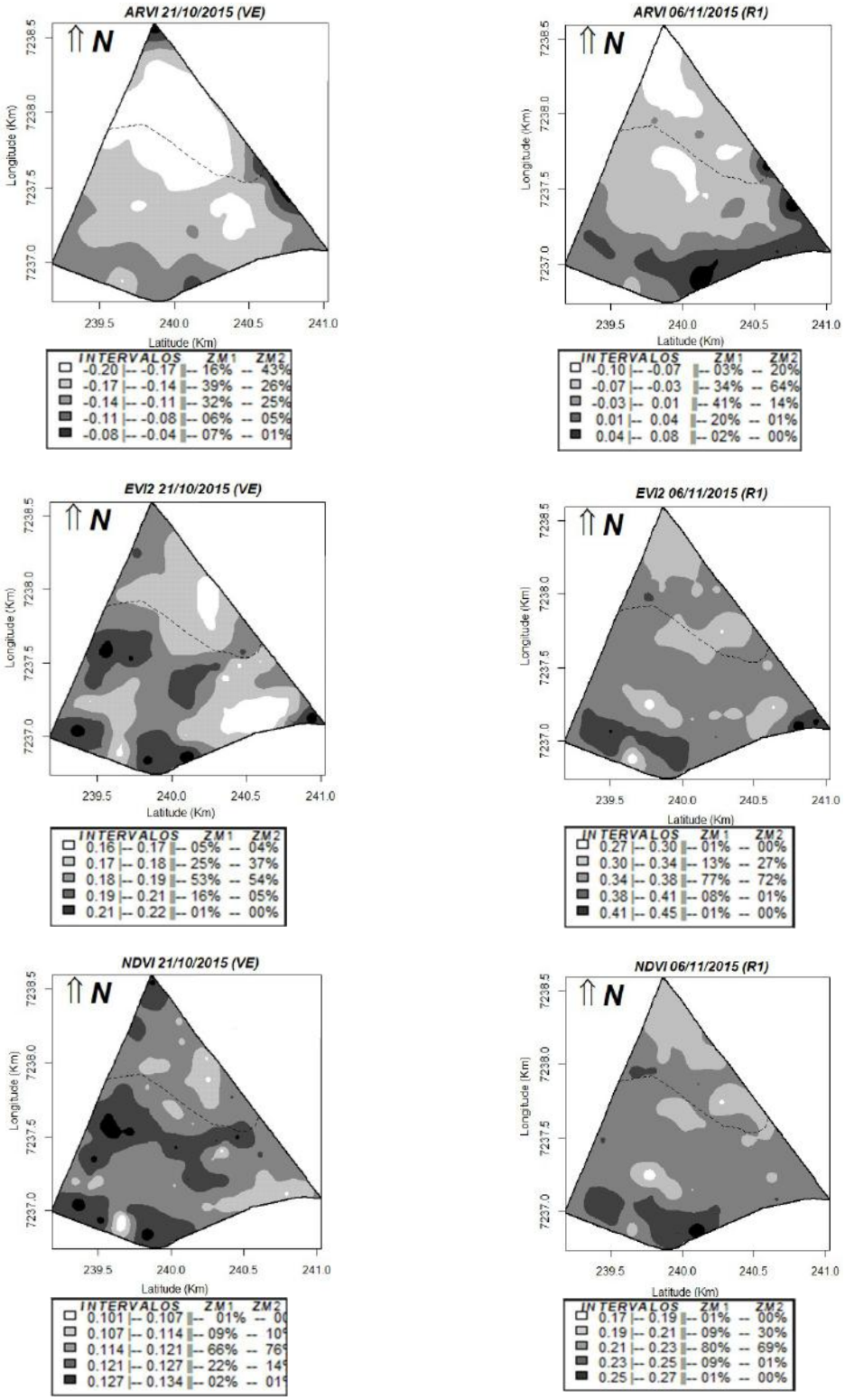

Figura 33 Mapas dos índices vegetativos com as respectivas zonas de manejo para o ano-safra de 2015/2016 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

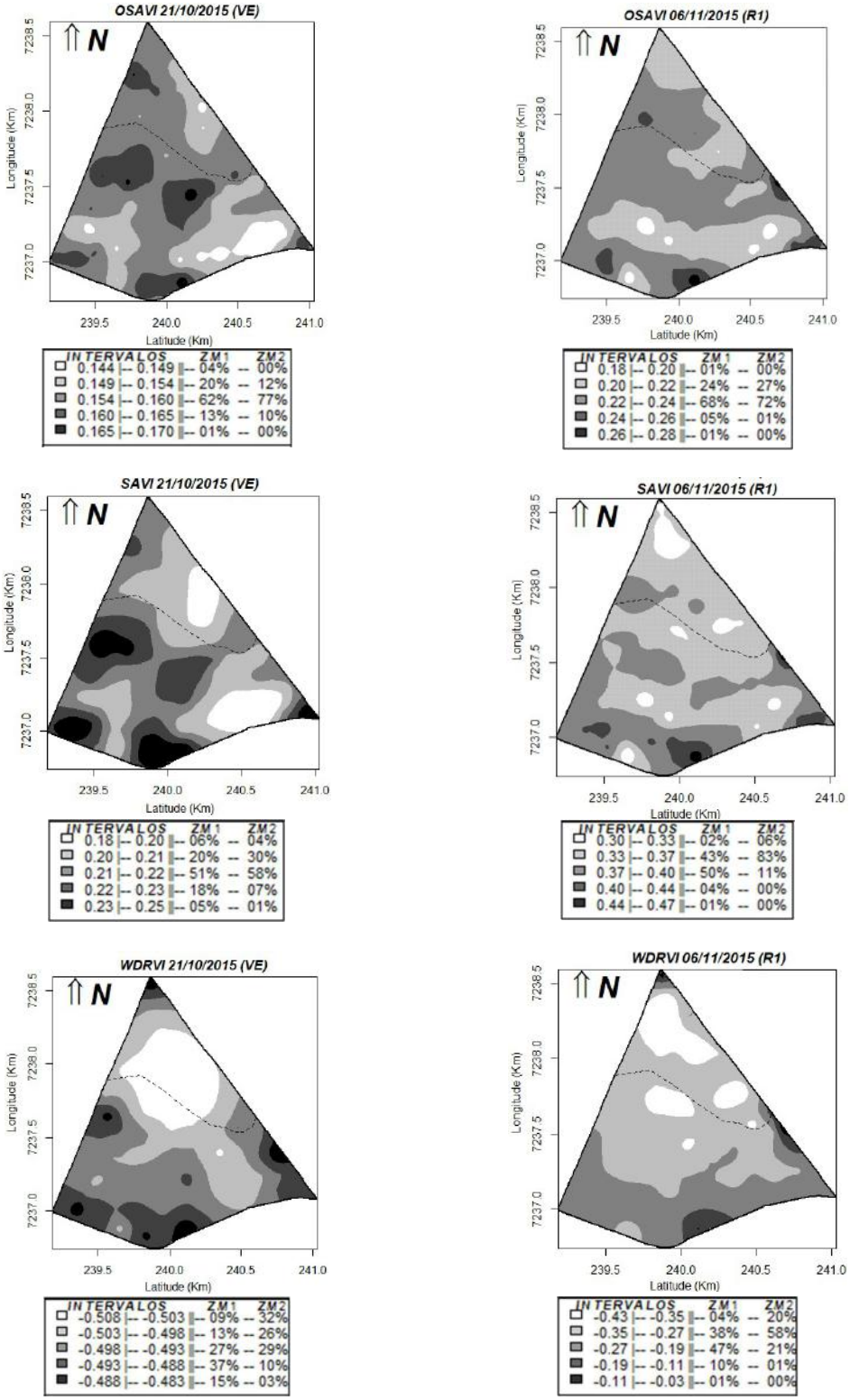

Figura 34 Mapas dos índices vegetativos com as respectivas zonas de manejo para o ano-safra de 2015/2016 (continuação) (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Para o ano-safra de 2016/2017, nota-se que a ZM1 apresenta valores maiores nas fases R3 e R5 do ciclo vegetativo da soja (Figuras 35, 36, 37 e 38).

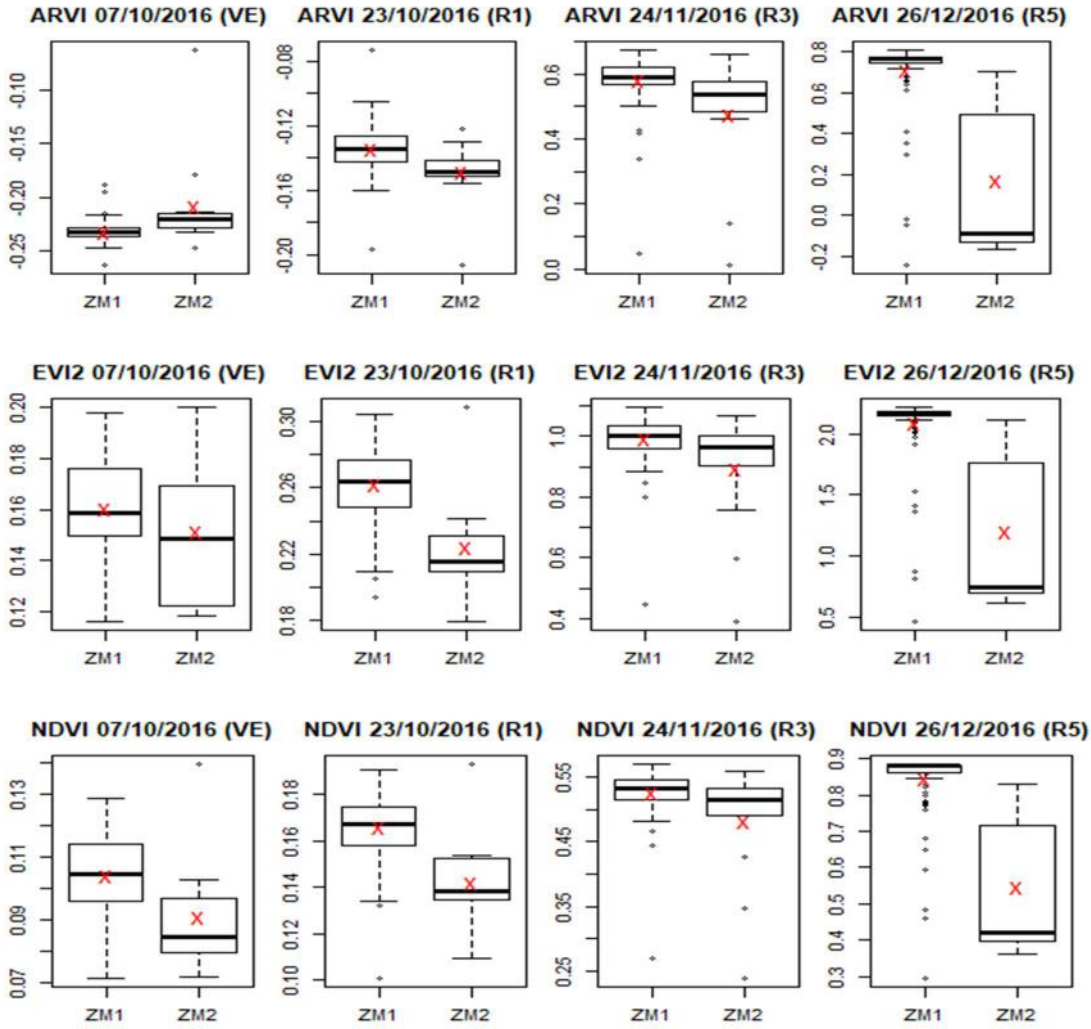

Figura 35 Boxplot para as zonas de manejo em relação aos índices vegetativos (ano-safra 2016/2017) (X em vermelho representa a média)

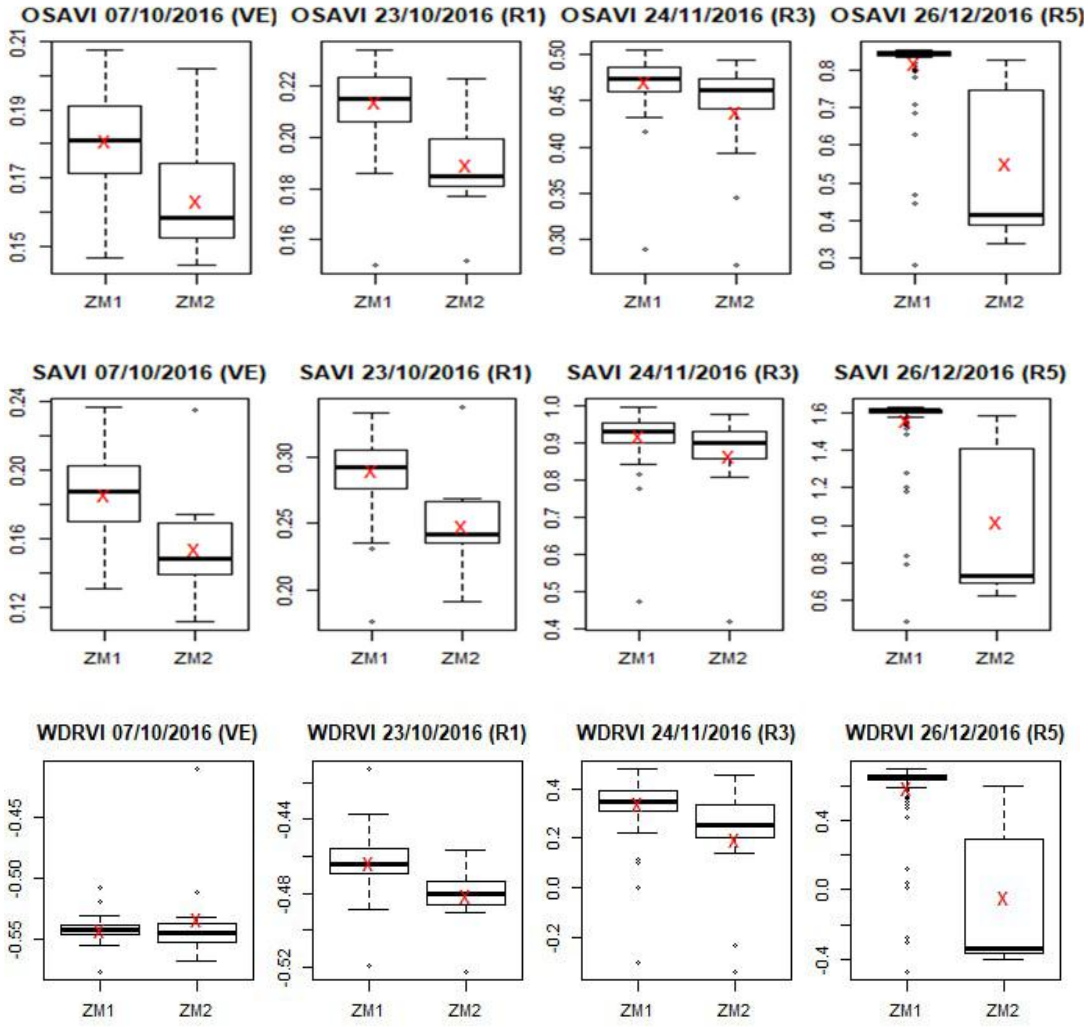

Figura 36 Boxplot para as zonas de manejo em relação aos índices vegetativos (ano-safra 2016/2017) (Continuação)

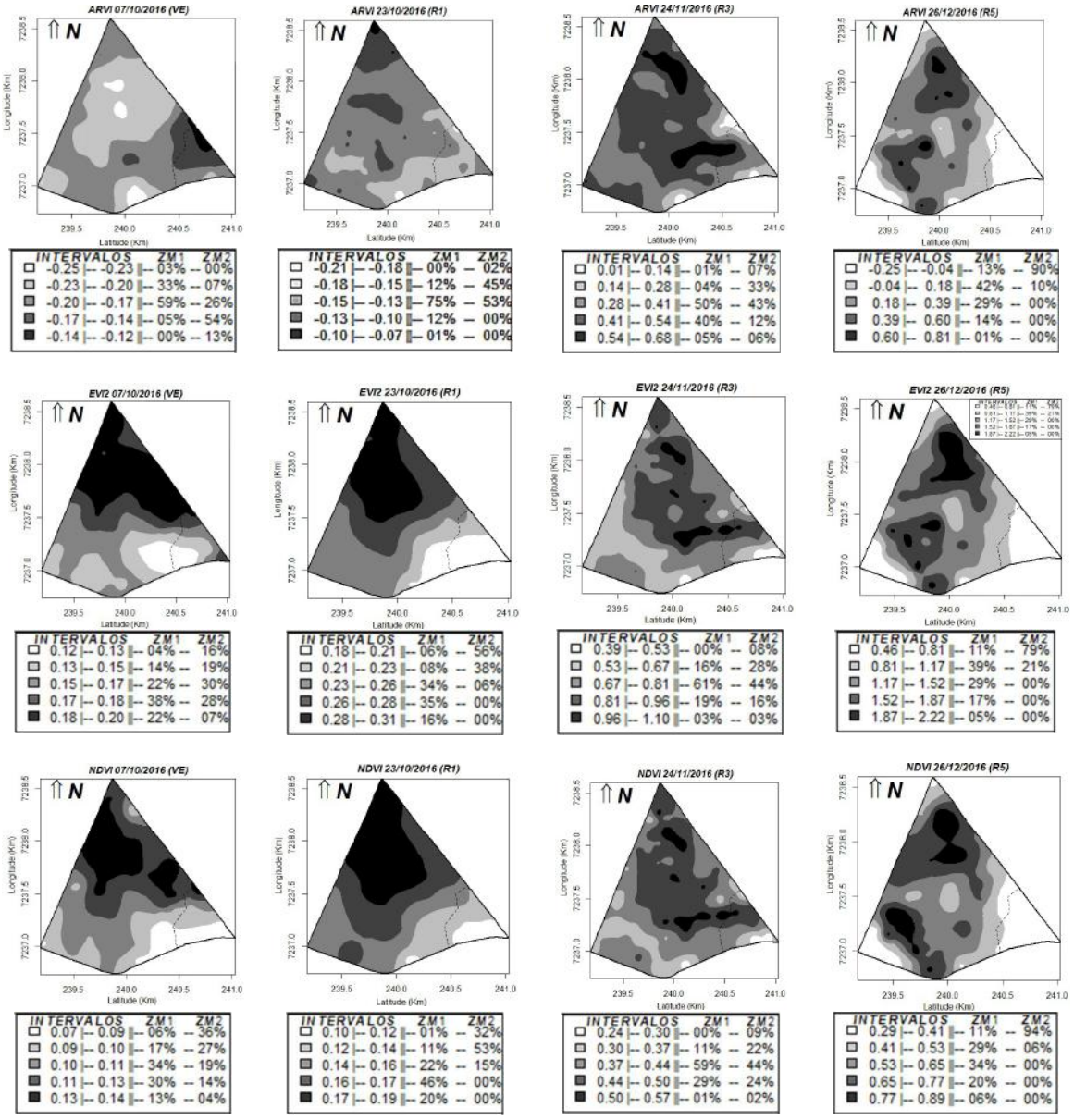

Figura 37 Mapas dos índices vegetativos com as respectivas zonas de manejo para o ano-safra de 2016/2017 (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

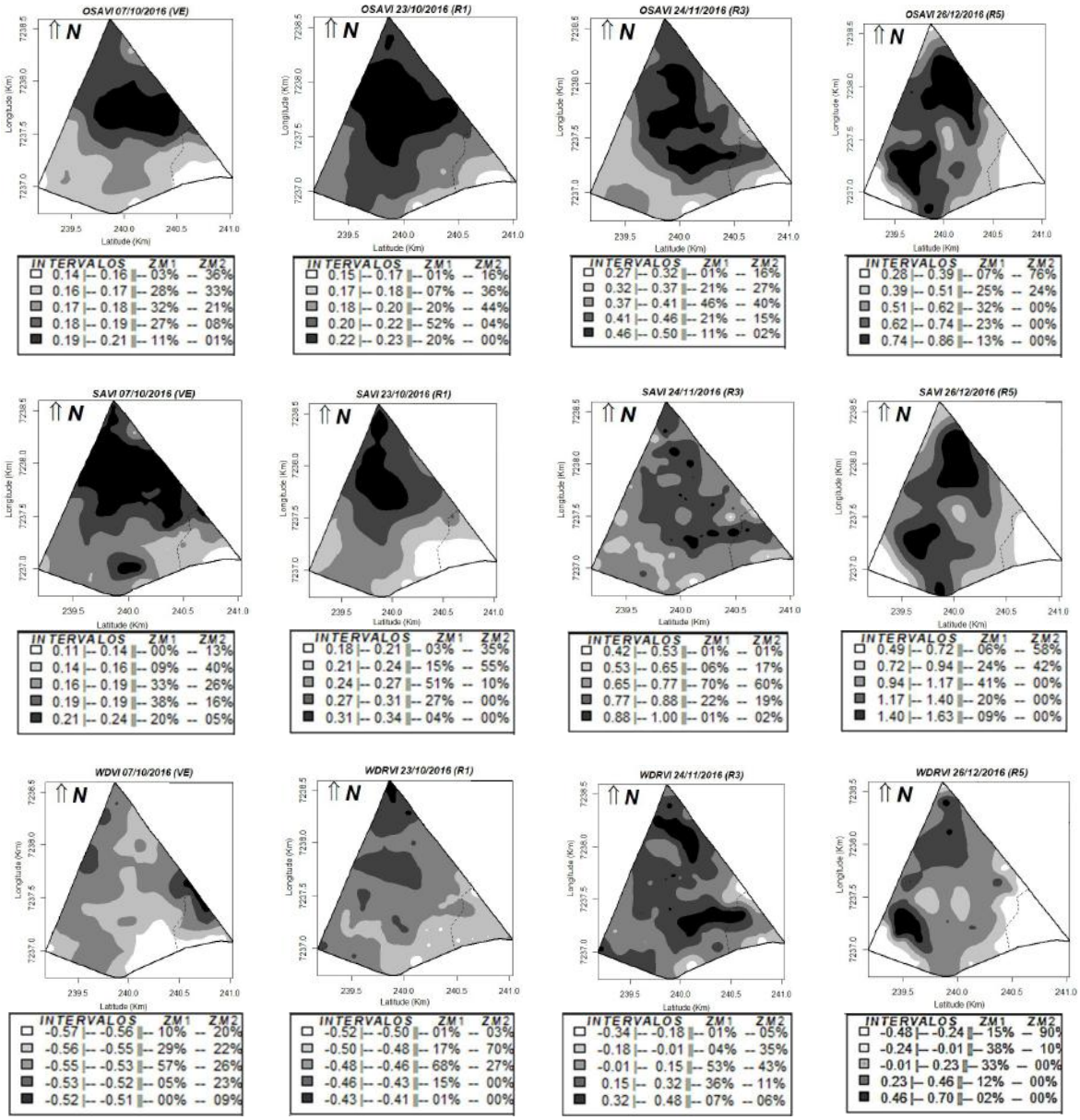

Figura 38 Mapas dos índices vegetativos com as respectivas zonas de manejo para o ano-safra de 2016/2017 (continuação) (as porcentagens indicam as proporções de cada intervalo nas zonas de manejo)

Verifica-se que os anos-safra 2013/2014 (Figura 39), 2014/2015 (Figura 40) e 2016/2017 (Figura 42) apresentam, com relação à produtividade, pouca diferença entre as zonas de manejo. Nos três casos, a ZM2 apresentou valores intermediários dessa variável. A partir dos mapas, é possível ver também bastante homogeneidade na destruição dos valores, principalmente para o primeiro e o último ano-safra. Além disso, no mapa de produtividade para o ano-safra de 2013/2014, observa-se a formação de algumas microrregiões circulares centradas nos pontos amostrais combinadas à predominância de pixels em uma única classe, e estes fatores caracterizam o fenômeno *bull eyes effect* (MENEZES *et al*., 2016).

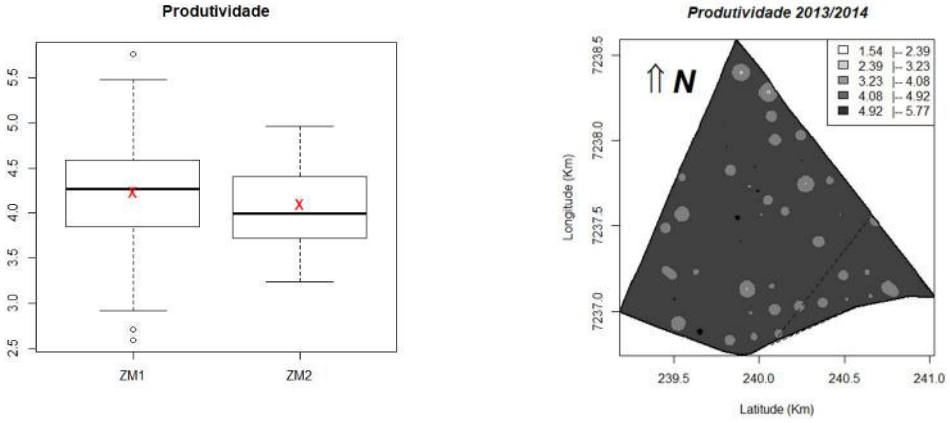

Figura 39 Boxplot das ZM1 e ZM2 e mapa temático para a produtividade (ano-safra 2013/2014) (X em vermelho representa a média)

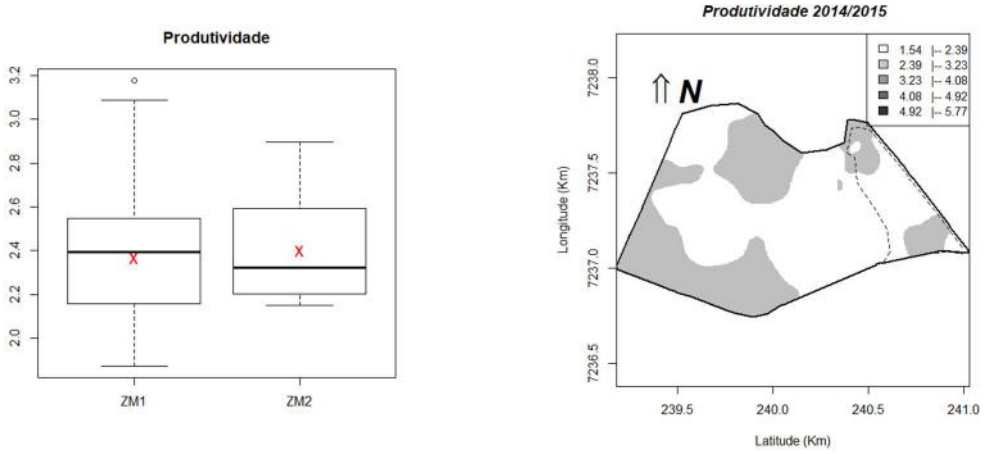

Figura 40 Boxplot das ZM1 e ZM2 e mapa temático para a produtividade (ano-safra 2014/2015) (X em vermelho representa a média)

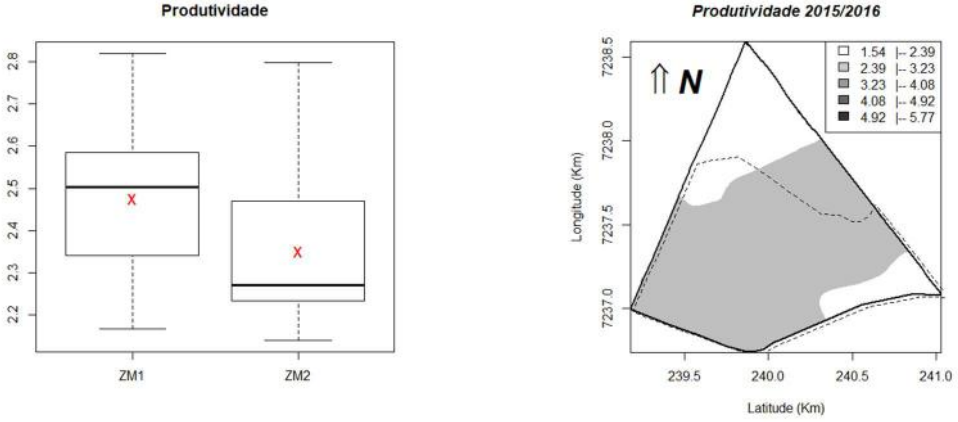

Figura 41 Boxplot das ZM1 e ZM2 e mapa temático para a produtividade (ano-safra 2015/2016) (X em vermelho representa a média)

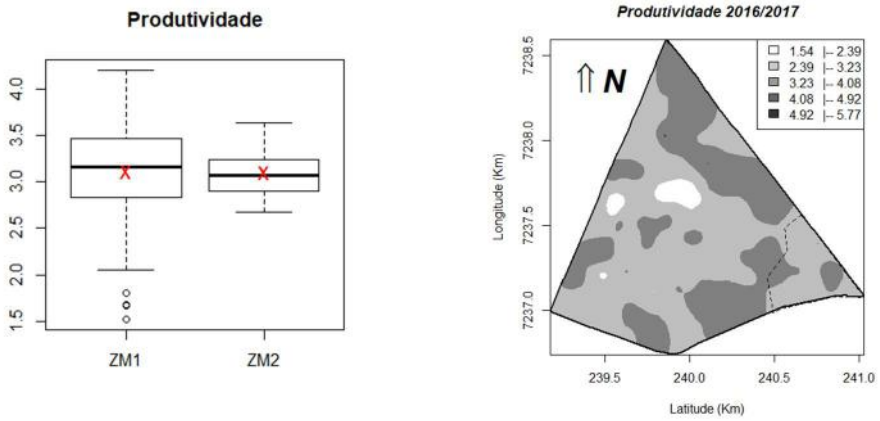

Figura 42 Boxplot das ZM1 e ZM2 e mapa temático para a produtividade (ano-safra 2016/2017) (X em vermelho representa a média)

As zonas de manejo geradas para o ano-safra de 2015/2016 (Figura 41) foram as que mais se diferenciaram das outras. A partir da análise dos mapas e do boxplot, foi possível observar elevada concentração para as variáveis Manganês, Cobre e Zinco na ZM2 associadas aos menores valores de Carbono e da produtividade no talhão, ou seja, parece haver uma maior associação da produtividade com as propriedades químicas do solo o que não foi tão evidente nos outros anos-safra.

5.4 Análise das árvores de decisão

Foram geradas árvores de decisão com os atributos que formaram os subconjuntos que melhor agruparam os dados para cada ano-safra. O intuito é avaliar o comportamento e a contribuição desses atributos na formação das zonas de manejo. Para isso, inicialmente, avaliou-se a precisão dos classificadores, conforme Tabela 09:

Tabela 9 Medidas de Desempenho do Classificador

Medidas de Desempenho	**ANO-SAFRA**			
	2013/2014	**2014/2015**	**2015/2016**	**2016/2017**
Acurácia	0,88	0,88	0,85	0,89
Kappa	0,79	0,79	0,78	0,79
Sensibilidade	0,97	0,91	0,86	0,93
Especificidade	0,82	0,82	0,85	0,84
Valor da Predição Positiva	0,82	0,83	0,92	0,89
Valor da Predição Negativa	0,96	0,92	0,74	0,90
Acurácia Balanceada	0,90	0,86	0,85	0,89

Assim, de acordo com essas medidas de desempenho, a acurácia geral do modelo e a acurácia balanceada foram superiores a 80% de valores classificados corretamente em todos os anos-safra. Isso indica que, de modo geral, o modelo apresenta pouco erro de

acordo com Congalton (1991). As medidas de sensibilidade e valor da predição positiva com valores acima de 80% indicam um elevado índice de acerto na classificação da classe positiva, e que de acordo com os dados obtidos nesta pesquisa, tais respostas correspondem à classe ZM1 em todos os casos analisados. Em contrapartida, as medidas de especificidade e valor da predição negativa indicam a acurácia do modelo em classificar os elementos da classe negativa. E para esta pesquisa, ZM2, os valores para as medidas na ZM2 apresentam bom índice de acerto para os dados dessa classe. Portanto, no geral, pode-se dizer que os modelos classificadores possuem confiabilidade.

A árvore de decisão gerada para o ano-safra de 2013/2014 com os dados referentes aos índices vegetativos mais produtividade pode ser visualizada na Figura 43. De acordo com essa figura, nota-se que, ao final do processo de classificação, foram geradas cinco folhas, sendo três delas pertencentes à ZM1 e duas pertencentes à ZM2.

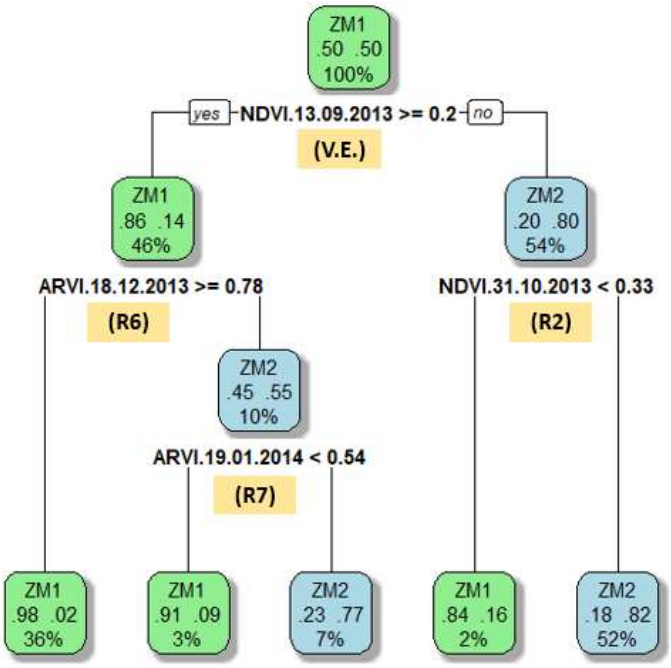

Figura 43 Árvore de Decisão para o ano-safra 2013/2014 (V.E.: fase de emergência; R2: pleno florescimento; R6: pleno enchimento das sementes; R7: início da maturação; ZM1: zona de manejo 1; ZM2: zona de manejo 2)

De acordo com a árvore da Figura 43, os atributos que mais contribuíram para a diferenciação das zonas de manejo foram os índices: NDVI calculado para os dias 13/09/2013 e 31/10/2013 e ARVI para o dia 18/12/2013. De modo geral, a ZM1 apresentou valores maiores para os índices NDVI (13/09/2013) e ARVI (18/12/2013), ou seja, no estádio de emergência da planta (VE) e no de pleno enchimento dos grãos (R6), respectivamente. Já a ZM2 apresentou valores baixos para o índice NDVI (13/09/2013), no estádio de emergência da planta (VE). Entretanto, foram apresentados valores mais elevados para

esse mesmo índice no estádio de pleno florescimento da soja (R2), isto é, para o NDVI calculado para o dia 31/10/2013. As variáveis mais importantes para classificação das zonas de manejo, de acordo com o classificador, foram os índices calculados para o dia 13/09/2013 (estádio de emergência da planta (VE)), seguidos dos índices calculados para o dia 18/12/2013 (estádio de pleno enchimento dos grãos (R6)). Os resultados também podem ser vistos quando analisados os mapas temáticos e os gráficos boxplots para esses atributos.

Para o ano-safra de 2014/2015, a árvore de decisão gerou cinco folhas ao final do processo: duas delas referentes à ZM1 e três referentes à ZM2, conforme Figura 44.

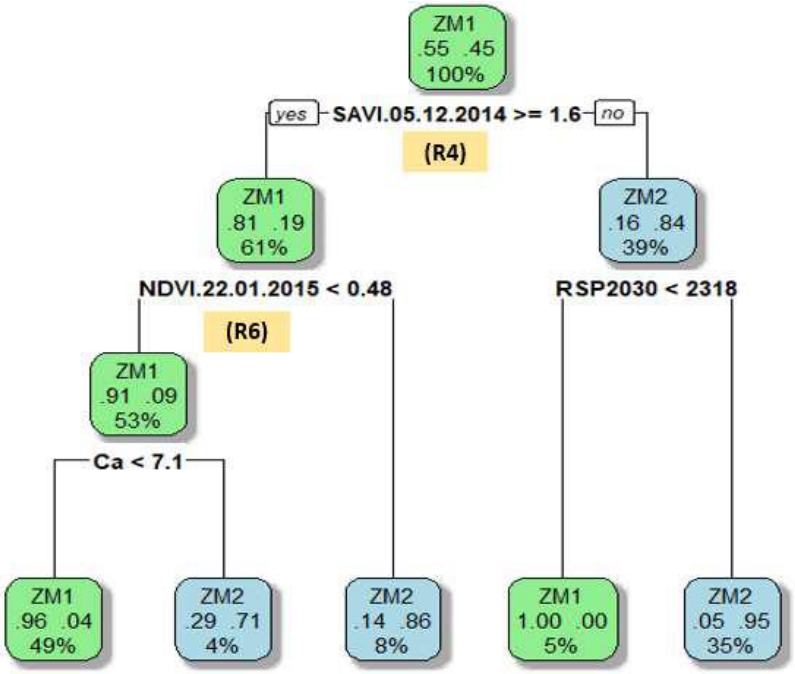

Figura 44 Árvore de Decisão para o ano-safra 2014/2015 (R4: plena formação das vagens; R6: pleno enchimento das sementes; ZM1: zona de manejo 1; ZM2: zona de manejo 2)

A análise dos resultados gerados pela árvore de decisão informa que os índices vegetativos que melhor diferenciaram as zonas de manejo foram o SAVI calculado para o dia 05/12/2014 e o NDVI para o dia 22/01/2015. O Cálcio (Ca) e a RSP 20-30 foram os atributos físico-químicos do solo que mais diferenciaram.

Como características principais, a ZM1 apresenta valores mais elevados para o índice SAVI calculado para o dia 05/12/2014, correspondente à fase de formação plena das vagens (R4) e os valores menores para o NDVI calculado para o dia 22/01/2015 e Cálcio. Além disso, os valores obtidos para a RSP nas camadas entre 20-30 cm de profundidade também foram baixos e o mesmo ocorreu nas regiões cujos valores para SAVI (05/12/2014) foram baixos. A ZM2, de modo geral, apresentou valores mais baixos para SAVI (05/12/2014) e maiores para a RSP nas camadas entre 20-30 cm de profundidade. As

variáveis mais importantes para a geração do modelo classificador foram os índices vegetativos calculados para o dia 05/12/2014, ou seja, de acordo com o classificador, a maior diferenciação entre as zonas de manejo ocorreu no período de formação das vagens. Os resultados estão de acordo com o que foi discutido anteriormente ao se analisarem os boxplots e os mapas dessas variáveis.

Assim, sete folhas foram obtidas ao se gerar a árvore de decisão para o ano-safra de 2015/2016 ao final da classificação. Quatro delas são referentes à ZM1 e três referentes à ZM2, conforme Figura 45.

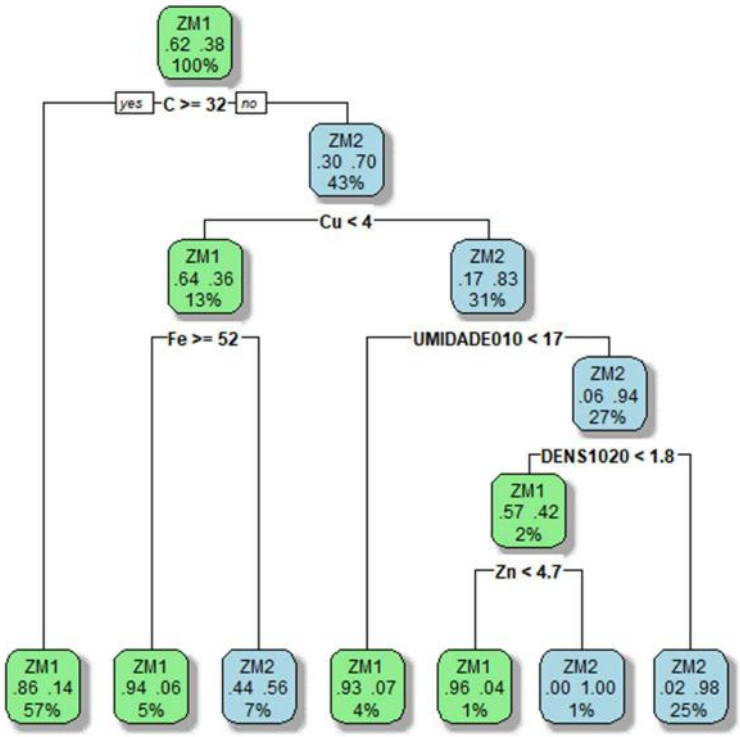

Figura 45 Árvore de decisão para as variáveis físico-químicas (ano-safra 2015/2016) (ZM1: zona de manejo 1; ZM2: zona de manejo 2)

De acordo com a árvore de decisão, os atributos que mais influenciaram para diferenciar as zonas de manejo foram o Carbono (C) e o Cobre (Cu). A ZM1 apresenta como característica principal valores de Carbono maiores ou iguais a 32 g/dm^3, e são os maiores valores para essa variável em todo talhão. Nas regiões dentro dessa zona em que o valor do Carbono é menor do que 32 g/dm^3 também encontram-se valores baixos para o Cobre (Cu) e elevados para o Ferro (Fe). A ZM2, por sua vez, caracteriza-se por possuir valores mais baixos para o Carbono (C) e mais elevados para o Cobre (Cu), cujas maiores concentrações dessa variável estão no talhão. Estes resultados estão de acordo com as análises dos boxplots e mapas apresentados anteriormente para essas variáveis. De acordo com o classificador, o Carbono e o Cobre foram as variáveis mais importantes no processo de classificação das variáveis nas classes ZM1 e ZM2.

Para o ano-safra de 2016/2017, a árvore de decisão gerou ao final do processo cinco folhas, sendo três referentes à ZM1 e duas referentes à ZM2, conforme Figura 46.

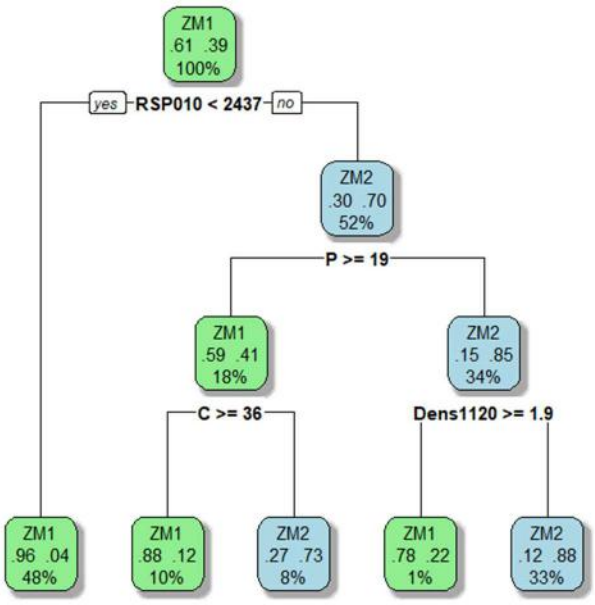

Figura 46 Árvore de decisão para as variáveis físico-químicas (ano-safra 2016/2017) (ZM1: zona de manejo 1; ZM2: zona de manejo 2)

A análise das regras geradas pela árvore aponta que a RSP entre 0 e 10 cm e o Fósforo (P) foram as variáveis que mais influenciaram na classificação das zonas de manejo. De modo geral, a ZM1 possui os menores valores da RSP nas camadas de 0 a 10 cm de profundidade com valores menores que 2437 kPa. Nos locais dentro dessa mesma zona de manejo, onde o valor dessa variável é maior do que 2437 kPa, os valores para o teor de Fósforo (P) é maior do que 19 mg /dm^{-3} e do Carbono (C) é maior do que 36 g/dm^{-3}. A ZM2 caracteriza-se por possuir valores mais elevados para a RSP nas camadas entre 0 e 10 cm de profundidade do solo, e valores mais baixos para a variável Fósforo (P) e densidade na camada entre 11 e 20 cm de profundidade. Os resultados condizem com os encontrados nos boxplots e mapas apresentados anteriormente para esse ano-safra.

De acordo com o classificador, as variáveis mais importantes para a classificação dos dados em ZM1 e ZM2 foram a resistência do solo à penetração (RSP) na camada entre 0 e 10 cm de profundidade, o Carbono (C) e o Fósforo (P), sendo que a RSP teve relevância bem maior do que as demais variáveis no processo de classificação.

A utilização de árvores de decisão como ferramenta auxiliar na agricultura de precisão pode ser vista em vários trabalhos, por exemplo, o trabalho realizado por souza, *et al.* (2010), no qual a partir de experimentos em uma área de cana de açúcar no interior do estado de São Paulo, com o objetivo de avaliar a produtividade com base nas variáveis químicas do solo e da altitude, utilizou-se da árvore de decisão. Para isso, foram

selecionadas as variáveis mais correlacionadas à produtividade e utilizadas como variáveis dependentes no processo de classificação. A produtividade, por sua vez, foi discretizada e requerida como variável impendente nesse mesmo processo. Como resultado, verificou-se que a altitude foi a que teve melhor potencial de interpretar a produtividade e a importância da utilização da árvore de decisão em processo de geração de zonas de manejo.

6. CONCLUSÕES

De acordo com os resultados obtidos para os quatro anos-safra analisados, constata-se que os subconjuntos que apresentam os melhores ajustes para agrupar os dados variaram entre os anos-safra. O subconjunto formado pelas variáveis físico-químicas foi o que obteve os melhores resultados para os anos-safra de 2015/2016 e 2016/2017. E para o ano-safra de 2013/2014, os melhores resultados foram obtidos no subconjunto formado pelos índices vegetativos somados à produtividade. Porém, para o ano-safra de 2014/2015, os melhores resultados foram obtidos no subconjunto formado por todas as variáveis disponíveis.

A melhor divisão da área em estudo, de acordo com os índices utilizados para avaliação de agrupamentos, foi com duas zonas de manejo. E, o resultado se repetiu em todos os anos-safra e para todos os subconjuntos testados. Assim, a utilização de duas zonas de manejo é a indicada para essa área estudada.

Observou-se nos anos-safra de 2013/2014, 2014/2015 e 2016/2017 a formação de duas zonas de manejo em regiões semelhantes no mapa, com um grupo menor situado à sudeste do mapa e outro maior que ocupou o restante da área. A única diferença quanto à localização das zonas de manejo foi para o ano-safra de 2015/2016, no qual houve a formação de um grupo menor ao norte do mapa e um grupo maior que ocupou o restante da área.

A utilização da árvore de decisão foi importante na caracterização das variáveis físico-químicas que melhor dividiram as duas zonas de manejo geradas, descrevendo seus comportamentos no processo de formação de cada classe.

As variáveis físico-químicas mais importantes para o processo de classificação alternaram-se conforme o ano-safra, e entre os mais relevantes estão: Manganês, Zinco, Cobre e Carbono. As variáveis relacionadas às propriedades físicas do solo tiveram pouca contribuição na formação das zonas de manejo, e a exceção mais importante foi a RSP nas camadas entre 0 e 10 cm de profundidade do solo.

7. REFERÊNCIAS

ALVES, S.M.F; ALCÂNTARA, G.R.; REIS, E.F.; QUEIROZ, D.M.; VALENTE, D.S.M. Definição de zonas de manejo a partir de mapas de Condutividade elétrica e matéria orgânica. **Bioscience Journal**. Uberlândia, v. 29, n. 1, p. 104-114, Jan./Feb. 2013.

AMADO, T. J.C.; PONTELLI, C.B.; SANTI, A. L.; VIANA, J.H.M.; SULZBACH, L.A.S. Variabilidade espacial e temporal da produtividade de culturas sob sistema plantio direto. **Pesquisa Agropecuária Brasileira**, v. 42, n. 8, p. 1101-1110, 2007.

ANTUNIASSI, U.R.; BAIO, F.H.R.; SHARP, T.C. **Agricultura de Precisão**. In: ELEUSIO, C. F. (Org.). Algodão no Cerrado do Brasil. 2. ed. Brasília/DF. 2007.

ARAÚJO, J.C.; VETTORAZZI, C. A.; MOLIN, J.P. Estimativa da produtividade e determinação de zonas de manejo, em culturas de grãos, por meio de videografia aérea multispectral. **Acta Sci. Agron**. Maringá, v. 27, n. 3, p. 437-447, 2005.

ARROUAYS, D.; SABY, N.P.A.; THIOULOUSE, J.; JOLIVET, C.; BOULONNE, L.; RATIÉ, C. Large trends in French top soil characteristics are revealed by spatially constrained multivariate analysis. **Geoderma**, v.161. 2011. Disponível em <https://www.sciencedirect.com/science/article/pii/S001670611000368X?via%3Dihub> Acesso em Fev. 2019.

BANNARI, A.; MORIN, D.; BONN, F.; HUETE, A. R. A review of vegetation indices. **Remote Sensing Reviews**, v. 13, p. 95-120, 1995.

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. **Revista Engenharia Agrícola**, Jaboticabal, v. 34, n. 5, p. 952-964, 2013.

BAZZI, C.L; SOUZA, E.G.; BETZEK, N.M. **SDUM – Software Para Definição de Unidades de Manejo**: Teoria e Prática, 1 ed. 2015.

BEHERA, S.K.; MATHUR, R.K.; SHUKLA, A.K.; SURESH, K.; PRAKASH, C. Spatial variability of soil properties and delineation of soil management zones of oil palm plantations grown in a hot and humid tropical region of southern India. **Catena**, v.165. 2018. Disponível em <https://www.sciencedirect.com/science/article/pii/S0341816218300444> Acesso em Jan. 2019.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R.A.; STONE, C.J. **Classification and Regression Trees**, v. 19, 1984.

BUSSAB, W.Q.; MORETTIN, P.A. **Estatística básica.** 5 ed. São Paulo: Saraiva, 2004, 526p.

CAMARGO, E.C.G. Geoestatística: fundamentos e aplicações. In: Câmara, G. & Medeiros, J.S. eds. **Geoprocessamento para projetos ambientais**. São José dos Campos: INPE, 1998.

CASELLA, G.; BERGER, R.L. **Inferência estatística**. Trad. da 2. ed. São Paulo: Cengage Learning, 2011. 612p.

CERVANTES, J.; LAMONT, F.G.; LÓPEZ-CHAU, A.; MAZAHUA, L.R.; RUÍZ, J.S. Data selection based on decision tree for svm classification on large data sets, Applied Soft Computing. **Applied Soft Computing**. v. 37, 2015.

CHACÓN, J.E.; DUONG, T. Multivariate plug-in band width selection with unconstrained pilot band width matrices. **Test**, v. 19, p. 375-398, 2010.

CHENG, D.; ZHU, Q.; HUANG, J.; WU, Q.; YANG, L.A Novel Cluster Validity Index Based on Local Cores. **IEEE Transactions on Neural Networks and Learning Systems**. July, v. 30 e p. 985 - 999, 2018.

CICORE, P.; SERRANO, J.; SHAHIDIAN, S.; SOUSA, A.; COSTA, J.L.; SILVA, J.R. M. Assessment of the spatial variability in tall wheat grass forage using LANDSAT 8 satellite imagery to delineate potential management zones. **Environmental Monitoring and Assesment**, v. 188, p. 513, 2016.

CLEVERS, J.G.P.W. The derivation of simplified reflectance model for the estimation of leaf area index, **Remote Sensing of Environment**, v. 25, n. 1, p. 53-69, 1988.

CONGALTON, R. G.; GREEN, K. **Assessing the accuracy of remotely sensed data**: principles and practices. New York: Lewis Publisher, 1999, 130p.

CORDOBA, M.; BALZARINI, M.; BRUNO, C.; COSTA, J.L. Análisis de componentes principales com datos georreferenciados: Una aplicación en agricultura de precisión. **Revista de la Facultad de Ciencias Agrarias**, v. 44, n. 1, 2012. Disponível em <http://revista.fca.uncu.edu.ar/> Acesso em Fev. 2019.

CORDOBA, M.; BALZARINI, M.; BRUNO, C.; COSTA, J.L. Subfield management cluster analysis from spatial principal components of soil variables. **Computers and Electronics in Agriculture**. 2013. Disponível em <https://www.sciencedirect.com/science/article/pii/S0168169913001282?via%3Dihub> Acesso: Mar. 2019.

CORDOBA, M.; BALZARINI, M.; BRUNO, C.; COSTA, J.L. Variabilidad espacial de suelo a escala de lote y surelación con los rendimientos. **Revista investigaciones agropecuárias**, v. 42, n. 1. 2016. Disponível em <http://www.scielo.org.ar/scielo.php?script=sci_abstract&pid=S1669231420160001000008&lng=es&nrm=iso> Acesso em: Mar. 2019.

CRESSIE, N. A. C. **Statistic for spatial data**. New York: J. Wiley, 1993. 900p.

CRESSIE, N. A. C. **Statistics for spatial data**. Hoboken, NJ: John Wiley Sons Inc. 2015.

CRUZ M.A.S.; SOUZA A.M.B.; JESUS J.S. Avaliação da cobertura vegetal por meio dos Índices de Vegetação SR, NDVI, SAVI e EVI na bacia do rio Japaratuba-Mirim em Sergipe. In: **Simp. Bras. Sens. Remoto**, 15. (SBSR). 2011, Curitiba. Anais. São José dos Campos: INPE, 2011.

DELALIBERA, H.C.; WEIRICH NETO, P.H.; NAGATA, N. Management zones in agriculture according to the soil and landscape variables. **Engenharia Agrícola**, Jabuticabal, v. 32, n. 6, 2012. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-69162012000600021> Acesso em Fev. 2019

DESGRAUPES, B. Clustering Indices. **Package cluster Crit for R.** 2017. Disponível em: <https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf> Acesso em: 12 Nov. 2019.

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

EMBRAPA – Centro Nacional de Pesquisa de Solos. Sistema brasileiro de classificação de solos, 3ª ed. Rio de Janeiro: **EMBRAPA** – SPI, 2013, 412p.

EVERITT, B.S.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster Analysis**. 5ª ed. London, UK: Wiley, 2011.

FARACO, M.A.; URIBE-OPAZO, M.A.; SILVA, E.A.A.; JOHANN, J.A.; BORSSOI, J.A. Seleção de modelos de variabilidade espacial para a elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. **Revista brasileira de ciências do solo**, v. 32, p. 463-476, Viçosa-MG, 2008.

FEHR, W.R. & CAVINESS, C.E. **Stage of soybean development**. Iowa State University. Special report 80, March, 1981.

FERNANDEZ, P.J.; YOHAI, V. **Introdução à Análise Exploratória de Dados Multivariados**. IMPA. 2014.

FERREIRA, D. F. **Estatística multivariada**. 2 ed. Lavras: Ed UFLA, 2011. 676p.

FOUEDJIO, F. A hierarchical clustering method for multivariate geostatistical data. **Spatial Statistics**, n. 18, p. 333-351, 2016.

FRAISSE, C.W.; SUDDUTH, K.A.; KITCHEN, N.R. Calibration of the CERES-Maize model for simulating site specific crop development and yield on clay pan soils. **Applied Engineering in Agriculture**, v. 17, n. 4, 2001. Disponível em <https://naldc.nal.usda.gov/download/25976/PDF> Acesso em: Mar 2019.

FREITAS, A.A. Uma Introdução a Data Mining. **Informática Brasileira em Análise**. CESAR - Centro de Estudos e Sistemas Avançados do Recife. Ano II, n. 32, mai./jun., 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. 2016. Disponível em <https://www.sciencedirect.com/science/article/pii/S016816991630432X?via%3Dihub> Acesso em: Fev. 2019.

GAVIOLI, A; SOUZA, E.G.; BAZZI, C.L.; SCHENATTO, K.; BETZEK, N.M. Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods. **Biosystems Engineering**. v. 181, 2019. Disponível em <https://www.sciencedirect.com/science/article/pii/S1537511018303295?via%3Dihub> Acesso em: Mar. 2019.

GITELSON, A.A. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation. **Jounal of Plant Physiology**, v. 161, ed. 2, p. 165-173, 2004.

GOLDSCHMIDT, R.; PASSOS, E. BEZERRA, E. **Data mining: Conceitos, técnicas, algoritmos, orientações e aplicações**. 2 ed, Rio de Janeiro: Elsevier, 2015. 276 p.

GONÇALVES, A.C.A.; FOLEGATTI, M.V.; MATA, J.D.V. Análises exploratória e geoestatística da variabilidade de propriedades físicas de um argissolo vermelho. **Acta Scientiarum**, Maringá, v. 23, n. 5, p. 1149-1157, 2001.

GREGO, C.R.; OLIVEIRA, R.P. de; VIEIRA, S.R. Geoestatística aplicada à agricultura de precisão. In: BERNARDI, A.C. de C.; NAIME, J. de M.; RESENDE, A.V. de; BASSOI, L.H.; INAMASU, R.Y. (Ed.). **Agricultura de precisão**: resultados de um novo olhar. Brasília, DF: Embrapa, 2014b. cap. 5, p. 74-83.

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

HAIR, J.F.; BLACK, W.C.; BABIN, B.J.; ANDERSON, R.E.; TATHAN, R.L. **Análise Multivariada de Dados**. 6 ed. Porto Alegre: Bookman, 2009.

HALKIDI, M., VAZIRGIANNIS, M., BATISTAKIS, I. Quality Scheme Assessment in the Clustering Process. **In Proceedings of PKDD**, Lyon, France. 2000.

HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. San Diego: Academic, 2001. 550p.

HENNIG, C.; MEILA, M.; MURTAGH, F.; ROCCI, R. **Handbook of Cluster Analysis**. Boston, USA: CRC Press, 2016.

HONGYU, K.; SANDANIELO, V.L.M.; OLIVEIRA JUNIOR, G.J. Análise de Componentes Principais: Resumo Teórico, Aplicação e Interpretação. **Engineering and Science**, v. 5, n. 1. 2015. Disponível em <http://periodicoscientificos.ufmt.br/ojs/index.php/eng/article/view/3398/2623> Acesso: Jan. 2019.

HUETE, A.R.; LIU, H.Q.; BATCHILY, K.; VAN LEEUWEN, W.A comparison of vegetation indices over a global set of TM images for EOS-MODIS. **Remote Sensing of Environment**. v. 59, n. 3, p. 440-451, 1997.

ISAAKS, E.H. & SRIVASTAVA, R.M. **An introduction to applied geostatistics**. New York, Oxford University Press, 1989.

JOHNSON, R.A. & WICHERN, D.W. **Applied Multivariate Statistical Analysis.** 6 ed, New Jersey: Pearson Prentice Hall, 2007.

JOURNEL, A.G. & HUIJBREGTS, C.H. **Mining Geostatistics**, Academic Press, New York, 2003.

KLEIBER, W. & NYCHKA, D. Non stationary modeling for multivariate spatial processes. **J. Multivariate Anal**, v. 112, p. 76–91, 2012.

KUIAWSKI, A.C.M.B.; SAFANELLI, J.L.; BOTTEGA, E.L.; OLIVEIRA NETO, A.M.; GUERRA, N. Vegetation indexes and delineation of management zones for soybean. **Pesquisa Agropecuária Tropical.** v. 47, n. 2, 2017. Disponível em <http://www.scielo.br/scielo.php?pid=S198340632017000200168&script=sci_abstract> Acesso em: Fev. 2019.

KURINA, F.G.; HANG, S.; CORDOBA, M.A.; NEGRO, G.J.; BALZARINI, M.G. Enhancing edapho climatic zoning by adding multivariate spatial statistics to regional data. **Geoderma.** v. 310. 2018**.** Disponível em <https://www.sciencedirect.com/science/article/pii/S001670611730232X> Acesso em Fev. 2019.

LANDIM, P.M.B. Sobre Geoestatística e Mapas. **Revista Terra e Didática.** v. 2, n. 1, p. 19-33, 2006.

LI, X.; PAN, Y.C.; GE, Z.Q.; ZHAO, C.J. 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, 2007a.

MARTINS, R.C. **Definição de zonas de manejo por índices de vegetação obtidos por sensoriamento remoto e mapas de produtividade**. Dissertação (Mestrado em Agricultura de Precisão), Universidade Federal de Santa Maria, RS, 2017. Disponível em <https://repositorio.ufsm.br/handle/1/11346> Acesso em: Jan 2019.

MARY, S.A.L.; SIVAGAMI, A.N.; RANI, M.U. Cluster Validity Measures Dynamic Clustering Algorithms. **ARPN Journal of Engineering and Applied Sciences**. v. 10, n. 9, 2015.

MELLO, J.M. de; BATISTIA, J.L.F.; RIBEIRO JÚNIOR, P.J.; OLIVEIRA, M.S. Ajuste e seleção de modelos espaciais de semivariograma visando à estimativa volumétrica de *Eucaliptus grandis*. **Scientia Florestalis**, v. 1, n. 69, p. 25-37, 2005. Disponível em: <http://www.ipef.br/publicacoes/scientia/nr69/cap02.pdf> Acesso em: 16 Abr. 2019.

MENDES, A.M.S.; FONTES, R. L.F.; OLIVEIRA, M. Variabilidade espacial da textura de dois solos do deserto salino, no Estado do Rio Grande do Norte. **Revista Ciência Agronômica**, v. 39, n. 1, p. 19-27, 2008.

MOLIN, J.P.; AMARAL, L.R.; COLAÇO, A.F. **Agricultura de Precisão**. São Paulo: Oficina de Textos, 2015. 238 p.

MONTERO, J.M.; AVILÉS, G.F.; MATEU, J. **Spatial and Spatio-Temporal Geostatistical Modelling and Kriging**. Wiley: UK. 2015.

MORAL, F.J.; REBOLLO, F.J. Characterization of soil fertility using the Rasch model. **Journal of Soil Science and Plant Nutrition.** v. 17, n. 2. 2017. Disponível em <https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S071895162017000200016> Acesso em: Mar. 2019.

MOTA, V.C.; DAMASCENO, F.A.; LEITE, D.F. Fuzzy clustering and fuzzy validity measures for knowledge discovery and decision making in agricultural engineering. **Computers and Electronics in Agriculture.** v. 150, p.118-124. 2018

MOTOMIYA, A.V.A.; MOLIN, J.P.; MOTOMIYA, W.R.; BAIO, H.R. Mapeamento do índice de vegetação da diferença normalizada em lavoura de algodão. **Pesq. Agropec. Trop**., Goiânia, n. 1, p. 112-118, 2012.

MOURA, M.N.; VITORINO, M.I.; ADAMI, M. Análise de Componentes Principais da Precipitação Pluvial Associada à Produtividade de Soja na Amazônia Legal. **Revista Brasileira de Climatologia**, v. 22, 2018. Disponível em <https://revistas.ufpr.br/revistaabclima/article/view/55109/35656> Acesso em: Fev. 2019.

NARAYANAN, R., HONBO, D., MEMIK, G., CHOUDHARY, A., ZAMBRENO, J. An FPGA implementation of decision tree classification. **In Proceedings of the Design, Automation and Test in Europe Conference and Exhibition.** 2007

OLDONI, H. & BASSOI, L.H. Delineation of irrigation management zones in a Quartz ipsamment of the Brazilian semiarid region. **Pesquisa Agropecuária Brasileira**. v. 51, n. 9. 2016. Disponível em <http://www.scielo.br/scielo.php?script=sci_abstract&pid=S0100204X2016000901283&lng=en&nrm=iso> Acesso em: Fev. 2019.

OLIVER, M.A.; WEBSTER, R. **Basic Steps in Geostatistics**: The Variogram and Kriging. Springer. 2015.

PERALTA, N.R.; COSTA, J.L.; BALZARINI, M.; FRANCO, M.C.; CORDOBA, M.; BULLOCK, D. Delineation of management zones to improve nitrogen management of wheat. **Computers and Electronics in Agriculture**, v. 110, 2015. Disponível em <https://www.sciencedirect.com/science/article/pii/S0168169914002786> Acesso em: Fev. 2019.

PONZONI, F.J. Comportamento Espectral da Vegetação. In: MENESES, P. R., NETTO, J. S. M. (Org.) **Sensoriamento remoto, reflectância dos alvos naturais.** Brasília – DF: Editora Universidade de Brasília - UNB, Embrapa Cerrados, p 157-199, 2001.

QUINLAN, J.R. **C4.5**: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 1993.

RAMYA, M.; LOKESH, V.; MANJUNATH, T.; HEGADI, R.S. A predictive model construction for mulberry crop productivity, **Procedia Computer Science** v. 45, p. 156-165, 2015.

REYES, J.; WENDROTH, O.; MATOCHA, C.; ZHU, J. Delineating site-specific management zones and evaluating soil water temporal dynamics in a farmer's field in Kentucky. **Vadose Zone Journal**, v.18, n. 1. 2019. Disponível em <> Acesso em: Fev. 2019.

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

ROUSE, J.W.; HAAS, R.H.; SCHELL, J.A.; DEERING, D.W. Monitoring vegetation systems in the Great Plains with ERTS. In 3rd ERTS Symposium, **NASA** SP-351 I, p. 309-317, 1973.

ROUSSEW, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **J. Comput. Appl. Math**. v. 20, p. 53–65, 1987.

SALVADOR, A.; ANTUNIASSI, U.R. Imagens aéreas multiespectrais na identificação de zonas de manejo em áreas de algodão para aplicação localizada de insumos. **Botucatu, Energia na agricultura**, v. 26, n. 2, p. 1-19, 2011.

SANTI, A.L.; AMADO, T.J.C.; CHERUBIN, M.R.; MARTIN, T.N.; PIRES, J.L.; FLORA, L.P.D.; BASSO, C.J. Análise de componentes principais de atributos químicos e físicos do solo limitantes à produtividade de grãos. **Pesquisa Agropecuária Brasileira**, v. 47, n. 9, 2012. Disponível em <https://seer.sct.embrapa.br/index.php/pab/article/view/11238/7998> Acesso em: Mar 2019.

SANTO, R.E. Utilização da Análise de Componentes Principais na compressão de imagens digitais. **Einstein**, v. 10, n. 2, 2012. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S167945082012000200004&nrm=iso&tlng=pt> Acesso em: Fev. 2019.

SANTOS, R.T.; SARAIVA, A.M. Reference Process for Management Zones Delineation in Precision Agriculture. **IEEE Latin America Transactions**. v. 13, n. 3, 2015. Disponível em <http://www.revistaieeela.pea.usp.br/issues/vol13issue3March2015/13TLA3_25TeruelSantos.pdf> Acesso em: Mar. 2019.

SEAB. (Org.). Manual Técnico do Subprograma de Manejo e Conservação do Solo. 1ª ed. Curitiba: **Secretaria de Estado da Agricultura e do Abastecimento do Paraná**, 1989, v. 1, p. 41-50.

SILVA, R.M. **Introdução ao Geoprocessamento**: conceitos, técnicas e aplicações. Novo Hamburgo: FEEVALE, 2007. 176p.

SILVERMAN, B. Density Estimation for Statistics and Data Analysis. **Monographs on Statistics and Applied Probability**, London, 1986.

SIMONOFF, J.S. **Smoothing methods in statistics**. Springer, 1996.

STEVEN, M.D. The sensitivity of the OSAVI vegetation index to observational parameters. **Remote Sensing of Environmental**, v. 63, p. 49-60, 1998.

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

TAN, P.N.; STEINBACH, M.; KUMAR, V. **Introdução ao data mining**. Ciência Moderna, Rio de Janeiro, 2009.

TAYLOR, J.C.; WOOD, G.A.; EARL, R.; GODWIN, R.J. Soil Factors and their Influence on Within-field Crop Variability, Part II: Spatial Analysis and Determination of Management Zones. **Biosystems Engineering**, Amsterdam, v. 4, n. 84, p. 441-453, 2003.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**, fourth ed. Academic Press. 2009.

THERNEAU, T.M.; ATKINSON, E.J. An Introduction to Recursive Partitioning Using the RPART Routines. **R package**. 2019. Disponível em: <https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf> Acesso em: 14 Nov. 2019.

THOMPSON, J.R.; TAPIA, R.A. **Non parametric function estimation, modeling and simulation**. 1 ed. [S.l.]: Ed. Siam - Society for Industrial and Applied Mathematics, 1990.

TRIPATHI, R.; NAYAK, A.K.; SHAHID, M.; LAL, B.; GAUTAM, P.; RAJA, R.; MOHANTY, S.; KUMAR, A.; PANDA, B.B.; SAHOO, R.N. Delineation of soil management zones for a rice cultivated area in Eastern India using fuzzy clustering. **Catena**, v.133, 2015. Disponível em <https://www.sciencedirect.com/science/article/pii/S0341816215300175> Acesso em: Mar. 2019.

TSCHIEDEL, M.; FERREIRA, M.F. Introdução à Agricultura de Precisão: Conceitos e Vantagens. **Ciência Rural**, v. 32, n. 1. Santa Maria, RS. 2002. Disponível em <http://www.scielo.br/pdf/cr/v32n1/a27v32n1.pdf> Acesso em: Jan. 2019.

VAPNIK, V.N. The nature of statistical learning theory. Statistics for engineering and information Science. **Springer**, 2000. 314 p.

VIEIRA, S.R. Geoestatística em estudos de variabilidade espacial do solo. In: NOVAIS, R.F.; ALVAREZ, V.H.; SCHAEFER, G.R. (Ed.). **Tópicos em ciência do solo**. Viçosa: Sociedade Brasileira de Ciência do Solo, 2000.

VIEIRA, S.R.; MILLETE, J.; TOPP, G.C.; REYNOLDS, W.D. **Handbook for geostatistical analysis of variability in soil and climate data**. In: ALVAREZ V., V.H.; SCHAEFER, C.E.G.R.; BARROS, N.F.; MELLO, J.W.V. & COSTA, L.M., eds. Tópicos em ciência do solo. Viçosa, Sociedade Brasileira de Ciência do Solo, v. 2, p.1-45, 2002.

XIANG, L. Delineation and Scale Effect of Precision Agriculture Management Zones Using Yield Monitor Data Over Four Years. **Agriculture Sciences**, Maryland, v. 6, n. 2, p. 180-188, 2007.

XIAO, J.; LU, J.; LI, X. Davies Bouldin Index based hierarchical initialization K-means. **Intelligent Data Analysis,** China, v.21, n.6, p. 1327 - 1338, 2017.

WAND, M.P.; JONES, M.C. Kern Smooth: Functions for kernel smoothing for Wand & Jones. **R package version** 2.22-15, 1995.

WITTEN, I.H.; FRANK, E.; HALL, M.A. **Data mining**: practical machine learning tools and techniques. São Francisco, CA: The Morgan Kaufmann series in data management systems, 2011.

APÊNDICE A

Tabela 10 Análise da dependência espacial para o ano-safra de 2013/2014

Variável	a	EPR	D.E.
ARVI 13/09/2013	0,371	0,202	FORTE
EVI2 13/09/2013	0,242	0,189	FORTE
NDVI 13/09/2013	0,218	0,111	FORTE
OSAVI 13/09/2013	0,216	0,250	MÉDIA
SAVI 13/09/2013	0,243	0,188	FORTE
WDRVI 13/09/2013	0,366	0,192	FORTE
ARVI 31/10/2013	0,495	0,156	FORTE
EVI2 31/10/2013	0,382	0,092	FORTE
NDVI 31/10/2013	0,443	0,024	FORTE
OSAVI 31/10/2013	0,487	0,058	FORTE
SAVI 31/10/2013	1,089	0,107	FORTE
WDRVI 31/10/2013	0,493	0,152	FORTE
ARVI 18/12/2013	0,365	0,163	FORTE
EVI2 18/12/2013	0,659	0,042	FORTE
NDVI 18/12/2013	0,788	0,464	MÉDIA
OSAVI 18/12/2013	0,795	0,113	FORTE
SAVI 18/12/2013	0,657	0,047	FORTE
WDRVI 18/12/2013	0,376	0,177	FORTE
ARVI 19/01/2014	0,352	0,216	FORTE
EVI2 19/01/2014	0,710	0,072	FORTE
NDVI 19/01/2014	0,268	0,201	FORTE
OSAVI 19/01/2014	0,400	0,200	FORTE
SAVI 19/01/2014	0,634	0,292	MÉDIO
WDRVI 19/01/2014	0,397	0,192	FORTE
Fósforo	0,201	0,855	FRACO
pH	0,000	0,000	SDE
HAl3	0,192	0,766	FRACO
Cálcio	0,179	0,553	MÉDIA
Magnésio	0,000	0,000	SDE
Potássio	0,143	0,541	MÉDIA
Zinco	0,187	0,055	FORTE
Ferro	0,218	0,517	MÉDIA
Manganês	0,201	0,484	MÉDIA
Cobre	0,835	0,569	MÉDIA
Carbono	0,122	0,531	MÉDIA
Alumínio	0,000	0,000	SDE
umidade 0-10	0,239	0,715	MÉDIA
umidade 10-20	0,075	0,737	FRACO
umidade 20-30	0,000	0,000	SDE
Densidade 0-10	0,292	0,883	FRACO
Densidade 11-20	0,505	0,851	FRACO
Densidade 21-30	0,000	0,000	SDE
RSP 0-10	0,000	0,000	SDE
RSP 11-20	0,000	0,000	SDE
RSP 21-30	0,000	0,000	SDE
RSP 31-40	0,059	0,726	FRACO
Produtividade	0,068	0,424	MÉDIA

a: alcance; EPR: efeito pepita relativo; D.E.: dependência espacial

APÊNDICE B

Tabela 11 Análise da dependência espacial para o ano-safra de 2014/2015

Variável	a	EPR	D.E.
ARVI 16/09/2014	0,245	0,139	FORTE
EVI2 16/09/2014	0,253	0,111	FORTE
NDVI 16/09/2014	0,311	0,170	FORTE
OSAVI 16/09/2014	0,170	0,010	FORTE
SAVI 16/09/2014	0,203	0,200	FORTE
WDRVI 16/09/2014	0,350	0,012	FORTE
ARVI 05/12/2014	0,234	0,371	MEDIO
EVI2 05/12/2014	0,580	0,027	FORTE
NDVI 05/12/2014	0,243	0,084	FORTE
OSAVI 05/12/2014	0,623	0,225	FORTE
SAVI 05/12/2014	0,585	0,712	MEDIO
WDRVI 05/12/2014	0,282	0,016	FORTE
ARVI 22/01/2015	0,325	0,321	MEDIO
EVI2 22/01/2015	0,456	0,677	MEDIO
NDVI 22/01/2015	0,687	0,687	MEDIO
OSAVI 22/01/2015	0,443	0,727	MEDIO
SAVI 22/01/2015	0,682	0,356	MEDIO
WDRVI 22/01/2015	0,642	0,178	FORTE
Cobre	0,623	0,815	FRACO
Zinco	0,233	0,217	FORTE
Ferro	0,238	0,866	FRACO
Manganês	0,443	0,149	FORTE
Fosforo	0,098	0,158	FORTE
Carbono	0,716	0,720	MEDIO
pH	0,192	0,815	FRACO
HAl3	0,172	0,909	FRACO
Cálcio	0,143	0,543	MEDIO
Magnésio	0,161	0,459	MEDIO
Alumínio	0,171	0,394	MEDIO
Potássio	0,432	0,945	FRACO
umidade 0-10	1,542	0,601	MEDIO
umidade 10-20	0,358	0,642	MEDIO
umidade 20-30	0,158	0,491	MEDIO
Densidade 0-10	0,882	0,876	FRACO
Densidade 11-20	1,404	0,863	FRACO
Densidade 21-30	0,097	0,413	MEDIO
RSP 0-10	0,494	0,747	FRACO
RSP 11-20	1,063	0,433	MEDIO
RSP 21-30	1,152	0,667	MEDIO
RSP 31-40	1,396	0,675	MEDIO
Produtividade	0,264	0,476	MEDIO

a: alcance; EPR: efeito pepita relativo; D.E.: dependência espacial

APÊNDICE C

Tabela 12 Análise da dependência espacial para o ano-safra de 2015/2016

Variável	a	EPR	D.E.
ARVI 21/10/2015	1,797	0,035	FORTE
EVI2 21/10/2015	0,390	0,429	MEDIO
NDVI 21/10/2015	0,335	0,518	MEDIO
OSAVI 21/10/2015	0,342	0,585	MEDIO
SAVI 21/10/2015	0,340	0,190	FORTE
WDRVI 21/10/2015	1,053	0,094	FORTE
ARVI 06/11/2015	0,441	0,145	FORTE
EVI2 06/11/2015	0,279	0,110	FORTE
NDVI 06/11/2015	0,191	0,117	FORTE
OSAVI 06/11/2015	0,318	0,081	FORTE
SAVI 06/11/2015	0,286	0,095	FORTE
WDRVI 06/11/2015	0,678	0,082	FORTE
Cobre	0,624	0,403	MEDIO
Zinco	0,366	0,456	MEDIO
Ferro	0,724	0,619	MEDIO
Manganês	0,367	0,543	MEDIO
Fósforo	0,000	0,000	EFP
Carbono	0,539	0,703	MEDIO
pH	0,000	0,000	EFP
HAl3	0,000	0,000	EFP
Cálcio	0,240	0,453	MEDIO
Magnésio	0,000	0,000	EFP
Alumínio	0,000	0,000	EFP
Potássio	0,126	0,083	FORTE
umidade 0-10	0,163	0,073	FORTE
umidade 10-20	0,542	0,909	FRACA
umidade 20-30	0,199	0,627	MEDIO
Densidade 0-10	0,000	0,000	EFP
Densidade 11-20	0,116	0,280	MEDIO
Densidade 21-30	0,114	0,916	FORTE
RSP 0-10	0,253	0,699	MEDIO
RSP 11-20	0,173	0,678	MEDIO
RSP 21-30	0,088	0,107	FORTE
RSP 31-40	0,126	0,402	MEDIO
Produtividade	0,819	0,681	MEDIO

a: alcance; EPR: efeito pepita relativo; D.E.: dependência espacial

APÊNDICE D

Tabela 13. Análise da dependência espacial para o ano safra de 2016/2017

Variável	a	EPR	D.E.
ARVI 07/10/2016	0,487	0,424	MEDIO
EVI2 07/10/2016	0,479	0,667	MEDIO
NDVI 07/10/2016	0,898	0,500	MEDIO
OSAVI 07/10/2016	0,346	0,443	MEDIO
SAVI 07/10/2016	0,498	0,200	FORTE
WDRVI 07/10/2016	1,351	0,289	MEDIO
ARVI 23/10/2016	0,623	0,412	MEDIO
EVI2 23/10/2016	0,738	0,500	MEDIO
NDVI 23/10/2016	0,375	0,384	MEDIO
OSAVI 23/10/2016	0,516	0,500	MEDIO
SAVI 23/10/2016	1,281	0,223	FORTE
WDRVI 23/10/2016	0,269	0,574	MEDIO
ARVI 24/11/2016	0,325	0,289	MEDIO
EVI2 24/11/2016	0,843	0,395	MEDIO
NDVI 24/11/2016	0,788	0,464	MEDIO
OSAVI 24/11/2016	0,795	0,113	FORTE
SAVI 24/11/2016	0,866	0,571	MEDIO
WDRVI 24/11/2016	0,310	0,284	MEDIO
ARVI 26/12/2016	0,836	0,007	FORTE
EVI2 26/12/2016	1,010	0,020	FORTE
NDVI 26/12/2016	0,418	0,280	MEDIO
OSAVI 26/12/2016	1,153	0,065	FORTE
SAVI 26/12/2016	1,042	0,019	FORTE
WDRVI 26/12/2016	0,283	0,371	MEDIO
Carbono	0,399	0,390	MEDIO
Fósforo	0,312	0,150	FORTE
Ph	0,197	0,500	MEDIO
HAl3	0,244	0,549	MEDIO
Cálcio	0,223	0,508	MEDIO
Magnésio	0,258	0,413	MEDIO
Alumínio	0,215	0,537	MEDIO
Potássio	0,141	0,116	FORTE
umidade 0-10	0,466	0,722	FRACA
umidade 10-20	0,209	0,343	MEDIO
umidade 20-30	0,335	0,956	FRACA
Densidade 0-10	0,138	0,743	FRACA
Densidade 11-20	0,214	0,500	MEDIO
Densidade 21-30	0,215	0,721	FRACA
RSP 0-10	0,169	0,501	MEDIO
RSP 11-20	0,000	0,000	EFP
RSP 21-30	0,000	0,000	EFP
RSP 31-40	0,000	0,000	EFP
Produtividade	0,172	0,449	MEDIO

a: alcance; EPR: efeito pepita relativo; D.E.: dependência espacial