ANÁLISE DE AGRUPAMENTO DA VARIABILIDADE ESPACIAL DA PRODUTIVIDADE DA SOJA E VARIÁVEIS AGROMETEOROLÓGICAS NA REGIÃO OESTE DO PARANÁ

EVERTON C. DE ARAÚJO[1], MIGUEL A. URIBE-OPAZO[2], JERRY A. JOHANN[3]

RESUMO: O presente trabalho realizou uma análise de agrupamentos espacial por meio da estatística multivariada, no intuito de investigar a relação entre a produtividade da soja e as seguintes variáveis agrometeorológicas: precipitação pluvial, temperatura média do ar, radiação solar global e índice local de Moran (*LISA*) da produtividade. O estudo foi realizado com os dados das safras dos anos agrícolas de 2000/2001 a 2007/2008 da região oeste do Estado do Paraná. A identificação do número adequado de *clusters* para cada ano-safra foi obtida utilizando a minimização de desvios. O estudo mostrou a formação de grupos de municípios utilizando as similaridades das variáveis em análise. A análise de agrupamento foi um instrumento útil para melhor gestão das atividades de produção da agricultura, em função de que, com o agrupamento, foi possível estabelecer similaridades que proporcionem parâmetros para melhor gestão dos processos de produção que traga, quantitativa e qualitativamente, resultados almejados pelo agricultor.

PALAVRAS-CHAVE: Estatística espacial de área; similaridade espacial; estatística multivariada.

CLUSTER ANALYSIS OF SPATIAL VARIABILITY OF SOYBEAN PRODUCTIVITY AND AGROMETEOROLOGICAL VARIABLES FROM WESTERN REGION OF THE STATE OF PARANA

ABSTRACT: This study addresses a spatial cluster analysis by multivariate statistics to investigate the relation between soybean yield and the following meteorological variables: precipitation, average air temperature, solar radiation and Moran's local index (*LISA*). The study was conducted with harvesting data from the crop calendars from 2000/2001 to 2007/2008 in the Western region of the State of Parana. The identification of the appropriate number of clusters for each crop year was obtained using the minimization of deviations. The study showed that it is possible to form groups of municipalities using the similarities of the variables under consideration. Cluster analysis is a useful tool for better management of production activities of agriculture, according to which the grouping is possible to establish similarities that provide parameters for better management of production processes that bring both quantitatively and qualitatively, results sought by the farmer.

KEYWORDS: Spatial statistics area; spatial similarity; multivariate analysis.

INTRODUÇÃO

O estudo da correlação de dados agrometeorológicos em relação à produtividade da soja tem sido um grande desafio devido à complexidade das inter-relações existentes entre estes fatores. O emprego de métodos estatísticos multidimensionais torna-se, portanto, uma técnica fundamental na análise dessas inter-relações, já que é considerada também a localização dos dados.

A análise multivariada conta com a análise de agrupamento (*cluster analysis*) que identifica grupos em objetos de dados multivariados, cujo objetivo é formar grupos com propriedades homogêneas entre os elementos amostrais (HÄRDLE & SIMAR, 2007). A análise de agrupamentos é utilizada quando se deseja explorar as similaridades entre indivíduos, definindo-os em grupos,

[1] Computata, Dr. em Engenharia Agrícola, Prof. da UTFPR, Campus Medianeira-PR, Fone: (45) 9975-0861; everton@utfpr.edu.br.
[2] Estatístico, Dr. em Estatística, Pesquisador de Produtividade do CNPq, Prof. Associado do PGEAGRI, UNIOESTE, Cascavel- PR, Fone: (45) 3220-3228; miguel.opazo@unioeste.br.
[3] Engenheiro Agrícola, Dr. em Engenharia Agrícola, Prof. Adjunto do PGEAGRI, UNIOESTE, Cascavel- PR, Fone: (45) 3220-7320; jerry.johann@unioeste.br.
Recebido pelo Conselho Editorial em: 28-8-2012
Aprovado pelo Conselho Editorial em: 4-3-2013

considerando, simultaneamente, todas as variáveis observadas em cada indivíduo. Segundo esse método, aplicado por KUNZ et al. (2008), procura-se por agrupamentos homogêneos de itens representados por pontos em um espaço *n*-dimensional em um número conveniente de grupos, relacionando-os por meio de coeficientes de similaridade ou de distâncias (JOHNSON & WICHERN, 1992).

Segundo CORRAR et al. (2007), o princípio da análise de agrupamento consiste em que cada observação de uma amostra multivariada corresponda a um ponto em um espaço euclidiano multidimensional. Os processos de classificação resultam em agrupar os pontos em conjuntos que evidenciam aspectos marcantes da amostra. O resultado final pode ser apresentado em forma de um gráfico de esquema hierárquico denominado dendograma, contendo uma síntese dos resultados.

Segundo OLIVEIRA & BERGAMASCO (2003), a decisão do número de *clusters* é tomada, geralmente, a partir do exame do dendograma, onde podem ser lidos os índices de similaridade, que correspondem às distâncias euclidianas em que ocorrem as junções dos pontos observados para formar grupos. Um grande salto nesses índices, que equivale a uma grande distância no dendograma, indica que a agregação reuniu dois grupos muito dissimilares e, em razão disso, deve-se definir o número de grupos anterior a esse salto. Na definição do número de grupos a ser utilizado, KÓVACS et al. (2005) e KUNZ et al. (2008) apresentam o procedimento de agrupamento hierárquico para obter o número ótimo de *clusters*.

O objetivo deste trabalho foi realizar uma análise de agrupamento da variabilidade espacial da produtividade da soja e das seguintes variáveis agrometeorológicas: precipitação pluvial (mm), temperatura média do ar (ºC), radiação solar global média (W m^{-2}) e índice de Moran Local (*LISA - Local Indicator of Spatial Association*) univariado para a produtividade da soja da região oeste do Estado do Paraná.

MATERIAL E MÉTODOS

A área de estudo deste trabalho é apresentada na Figura 1 e compreende 48 municípios da região oeste do Estado do Paraná. Foram utilizados dados dos anos-safra de 2000/2001 a 2007/2008 das variáveis produtividade da soja [Prod] (t ha^{-1}), precipitação pluvial [Prec] (mm), temperatura média do ar [TMed] (ºC), radiação solar global média [Rs] (W m^{-2}) e índice de Moran Local da produtividade da soja [*LISA*].

O período das safras utilizado para a obtenção dos dados agrometeorológicos diários foi de 1º de outubro do ano inicial da safra até 28 de fevereiro de seu ano final.

A precipitação pluvial utilizada foi obtida por meio da soma dos dados do período de cada safra e da temperatura média do ar e da radiação solar global média pela média aritmética. Os dados referentes à produtividade da soja foram fornecidos pela SEAB (2010), e os dados agrometeorológicos, pelo SIMEPAR (2010). Os dados agrometeorológicos (temperatura média do ar, radiação solar global média e precipitação pluvial) estavam disponíveis apenas para oito municípios da região em estudo. Para os dados de precipitação pluvial, houve a situação de inexistência de medição para alguns dias, para os períodos do estudo no conjunto dos municípios com estações meteorológicas. A estimação da precipitação pluvial para os dias e municípios sem medição foi obtida por meio do uso de Polígonos de Thiessen (ANDRADE et al., 2008) e *Spatial Join* (JACOX & SAMET, 2007).

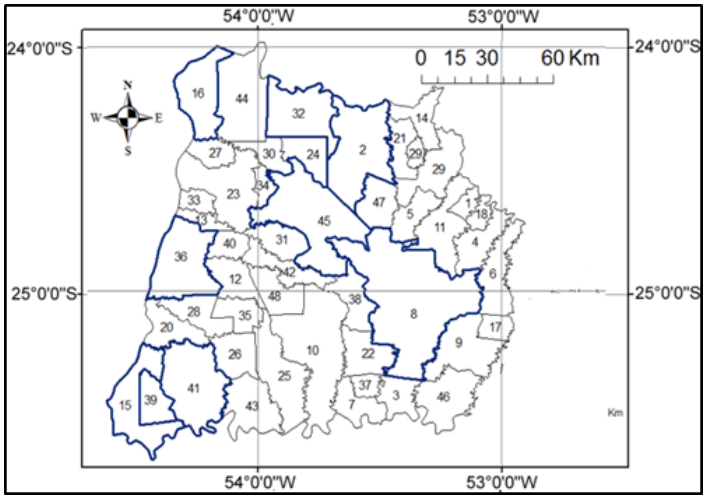


FIGURA 1. Região oeste do Paraná, com destaque para os municípios com estações meteorológicas. **Western Paraná, especially for municipalities with weather stations**.

(1) Anahy, (2) Assis Chateaubriand, (3) Boa Vista da Aparecida, (4) Braganey, (5) Cafelândia, (6) Campo Bonito, (7) Capitão Leônidas Marques, (8) Cascavel, (9) Catanduvas, (10) Céu Azul, (11) Corbélia, (12) Diamante D'Oeste, (13) Entre Rios do Oeste, (14) Formosa do Oeste, (15) Foz do Iguaçu, (16) Guaíra, (17) Ibema, (18) Iguatu, (19) Iracema do Oeste, (20) Itaipulândia, (21) Jesuítas, (22) Lindoeste, (23) Marechal Cândido Rondon, (24) Maripá, (25) Matelândia, (26) Medianeira, (26) Mercedes, (28) Missal, (29) Nova Aurora, (30) Nova Santa Rosa, (31) Ouro Verde do Oeste, (32) Palotina, (33) Pato Bragado, (34) Quatro Pontes, (35) Ramilândia, (36) Santa Helena, (37) Santa Lúcia, (38) Santa Tereza do Oeste, (39) Santa Terezinha de Itaipu, (40) São José das Palmeiras, (41) São Miguel do Iguaçu, (42) São Pedro do Iguaçu, (43) Serranópolis do Iguaçu, (44) Terra Roxa, (45) Toledo, (46) Três Barras do Paraná, (47) Tupãssi e (48) Vera Cruz do Oeste.

Para o desenvolvimento da análise multivariada espacial de agrupamentos, foram utilizadas técnicas de estatística multivariada, dendograma e mapa temático. Uma observação multivariada de *p*-variada é da forma representada na Equação (1), cujos elementos X_{i1} a X_{ip} são variáveis aleatórias oriundas de várias medidas de um mesmo elemento amostral *i*, *i* = 1,..., *n*, sendo *n* o número de elementos da população e *p* o número de variáveis em estudo.

$$Xi = (Xi1, Xi2, ..., Xip), \text{ para } i = 1, ..., n, \quad (1)$$

Seja *X* uma matriz de observações de *n* x *p* de *n* elementos amostrais em *p* variáveis, escrita da forma especificada na Equação (2):

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{nj} & \cdots & X_{np} \end{bmatrix} = [X_1 X_2 \ldots X_n]^T \quad (2)$$

A medida mais utilizada na indicação da proximidade entre dois objetos *i* e *k* é a distância euclidiana, representada por JOHNSON & WICHERN (1992) pela Equação (3).

$$d_{ik} = d(i,k) = \left[\sum_{j=1}^{p} (X_{ij} - X_{kj})^2\right]^{1/2} \quad (3)$$

em que: $i \neq k = 1, ..., n$ (total de elementos amostrais); X_{ij} é o elemento observado da *j*-ésima variável do elemento amostral *i*; X_{kj} é o elemento observado da *j*-ésima variável do elemento amostral *k*.

Quando se trabalha com variáveis quantitativas não comparáveis (cm, kg, anos ou milhões, dentre outras), a mudança de uma das unidades pode alterar completamente o significado e o valor do coeficiente; assim, deve-se proceder à padronização das variáveis dos elementos $X_{i1},..., X_{ip}$ do vetor X_i, usando a transformação descrita na Equação (4).

$$z_{ij} = \frac{X_{ij} - \bar{X}_{.j}}{s_{.j}} \quad (4)$$

em que: *i =1, ..., n; j =1, ..., p;* e e indicam, respectivamente, a média e o desvio-padrão amostral de *j*-ésima variável.

Feita a transformação, a distância euclidiana entre os municípios (objetos) foi determinada pela Equação (5), que é a soma dos desvios padronizados.

$$d_{ik} = d(i,k) = [\sum_{j=1}^{p} (z_{ij} - z_{kj})^2]^{1/2} \quad (5)$$

De acordo com BOSCHI et al. (2011), a técnica de agrupamento apresenta grande eficiência, e BUSSAB et al. (1990) sugeriram duas ideias básicas: coesão interna dos dados e isolamento externo entre os grupos.

A similaridade entre grupos pode ser classificada em categorias, nas quais as técnicas hierárquicas são as mais utilizadas na literatura. Por meio dessas técnicas hierárquicas, os objetos são classificados em grupos, em diferentes etapas, de modo hierárquico, produzindo uma árvore de classificação. Para essa análise, utilizou-se o algoritmo hierárquico de Mcquitty (GIMENES et al., 2004), a qual é definida pela Equação (6):

$$d_{(kl)j} = \frac{(d_{kl} + d_{lj})}{2} \quad (6)$$

em que: a distância entre o agrupamento (*kl*) e o agrupamento *j*; e são as distâncias entre a maior distância dos membros dos agrupamentos *k* e *j* e dos agrupamentos *l* e *j*. Desta maneira, define-se a matriz de distância MD = [(d_{ij})], *n x n*, que informa a distância entre as observações *i* e *j*, sendo *n* o número de elementos amostrais em estudo, e o nível de similaridade *s(ij)* entre dois grupos *i* e *j* é dado de acordo com o descrito na Equação (7):

$$s(ij) = 100(1 - (\frac{d_{ij}}{d_{(max)}}) \quad (7)$$

em que: $d_{(max)}$ é o valor máximo da matriz da distância MD.

O *LISA* busca captar padrões de associação local. A autocorrelação local pode ser calculada pela estatística *I* de Moran local (ANSELIN, 1995), que é uma estatística que deve possuir para cada observação uma indicação de grupos espaciais significantes de valores similares em torno da observação (e.g. região).

Segundo LE GALLO & ERTHUR (2003), o índice *LISA* univariado, baseado no *I* de Moran local, pode ser especificado para uma determinada variável X_j, *j* = 1,...,*p*, da forma descrita na Equação (8):

$$I_{ij} = \frac{X_{ij} - \overline{X}_{.j}}{\sigma_j^2} \sum_{k=1}^{n} w_{ik}^{(j)} (X_{kj} - \overline{X_{.j}}), \; i = 1,...,n \text{ para cada j=1, ...,p,} \quad (8)$$

sendo $w_{ik}^{(j)}$ o elemento da matriz proximidade *W*, *n* x *n*, da variável fixa X_j, j = 1,...,*p* e σ_j^2 a variância populacional da variável X_j em estudo das *n* populações. O método de contiguidade utilizado foi Torre (TEIXEIRA & BERTELLA, 2010).

O índice *LISA* I_{ij}, para *i =1,..., n, j =1,...,p*, pode ser interpretado da seguinte maneira: valores positivos de I_{ij} significam que existem *clusters* espaciais com valores similares (alto ou baixo); valores negativos significam que existem *clusters* espaciais com valores diferentes entre as regiões e seus vizinhos da *j*-ésima variável.

Um método de agrupamento hierárquico produz uma solução de agrupamento com qualquer número (*c*) de *clusters*, entre *1* e *n*. Para uma avaliação do número ideal de *clusters*, além da

definição empírica, algumas estatísticas estão disponíveis para a determinação do melhor número de *clusters*. Neste trabalho, foram consideradas as estatísticas conhecidas como *Root Mean Square Standard Deviation* (*RMSSTD*) e *R-square* (*RS*). Essa família de índices é aplicável nos casos em que os algoritmos hierárquicos são usados para agrupar os conjuntos de dados. Em um processo de simulação de *n* etapas, onde cada etapa gera um conjunto de *clusters*, o uso desses dois índices auxilia na determinação do número de grupos ótimo para um conjunto de dados (FIORINI et al., 2010).

O *RMSSTD* é uma medida da homogeneidade dentro dos clusters (KOVÁCS et al., 2005) e é definido pela Equação (9).

O *RS* pode ser considerado uma medida da similaridade entre os agrupamentos. Além disso, mede o grau de homogeneidade entre os grupos. Os valores de *RS* variam entre 0 e 1. No caso em que o valor de *RS* seja zero (0), há indicação de inexistência de diferença entre os grupos. Por outro lado, quando *RS* é igual a 1, existe indicação da diferença entre os grupos (KOVÁCS et al., 2005). Como resultado, quanto maiores as diferenças entre os grupos, mais homogêneos serão cada grupo, e vice-versa. O *RS* é definido pela Equação (10).

$$RMSSTD = \left[\frac{\sum_{k=1}^{c} \sum_{i=1}^{n_k} \sum_{j=1}^{p} (X_{ij}^{(k)} - \overline{X_{.j}^{(k)}})^2}{p \sum_{k=1}^{c} (n_k - 1)} \right]^{1/2} \quad (9)$$

$$RS = 1 - \frac{\sum_{k=1}^{c} \sum_{i=1}^{n_k} \sum_{j=1}^{p} (X_{ij}^{(k)} - \overline{X_{.j}^{(k)}})^2}{\sum_{k=1}^{c} \sum_{i=1}^{n_k} \sum_{j=1}^{p} (X_{ij}^{(k)} - \overline{X}_{..})^2} \quad (10)$$

em que: *c* é o número de *clusters*; n_c é o número de elementos de cada *cluster; p* o número de variáveis: $X_{ij}^{(k)}$ é o valor do *i*-ésimo elemento da população na *j*-ésima variável locado no *k*-ésimo *cluster*; $\overline{X_{.j}^{(k)}}$ e a média da *j*-ésima variável no *k*-ésimo *cluster,* e $\overline{X}_{..}$ é a média geral, $k = 1,...,c$; $i = 1,..,n_k$ e $j = 1,...,p$.

Para cada ano-safra (2000/2001 a 2007/2008), tendo como dados todas as variáveis do estudo (produtividade da soja, precipitação pluvial, temperatura média do ar, radiação solar global média e *LISA)*, foram geradas as estatísticas *RMSSTD* e *RS* para 10 grupos de *clusters*, com o objetivo de identificação do melhor número de *clusters* para cada ano-safra. Em um segundo estudo, também com 10 grupos de *clusters* e com as mesmas variáveis, foram geradas as estatísticas *RMSSTD* e *RS* em um único *cluster,* tendo todos os anos-safras com uma única medida.

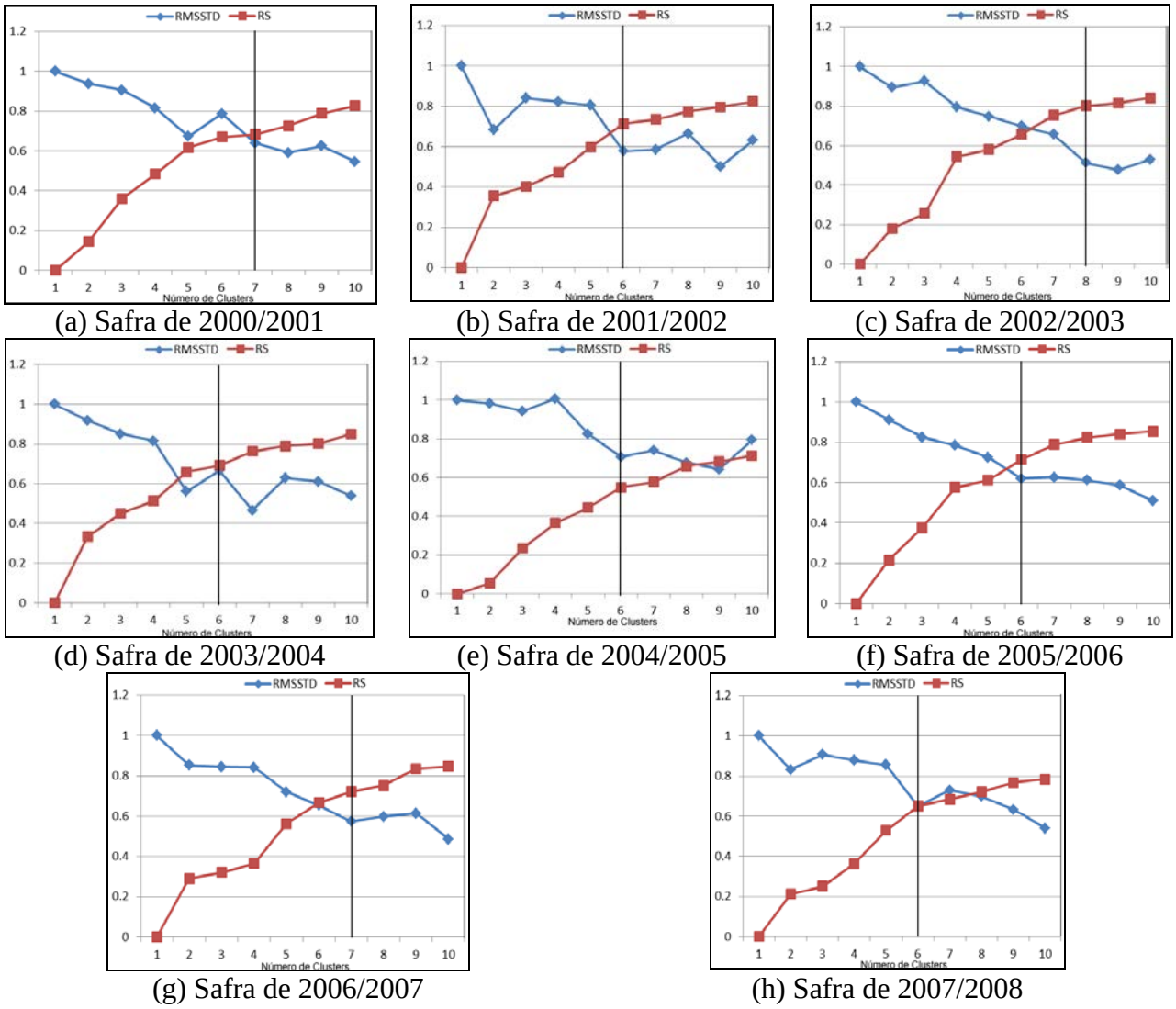
Para desenvolver a análise espacial de área, foram utilizados os *softwares* Minitab 15.0 (MINITAB, 2011), SAS® (SAS, 2011), ArcMap 9.3 (ESRI, 2011) e OpenGeoda 0.9.9.6 (OPENGEODA, 2011).

RESULTADOS E DISCUSSÃO

Na busca por um número ótimo para a quantidade de *clusters*, determinaram-se as estatísticas *RMSSTD* e *RS* para cada ano-safra estudado (Figura 2). O primeiro critério foi a escolha dos pontos de máxima curvatura (WANG et al., 2009) e, caso essa escolha não fosse viável para o ano-safra em estudo, optou-se pelo menor valor de *RMSSTD* em um ponto em que o valor de *RS*, que representa a heterogeneidade, não fosse alto (KOVÁCS et al., 2005).

Na execução das estatísticas (Figura 2), em algumas safras, o número de *clusters* identificado como ótimo causou a existência de municípios isolados, sem pertencer a nenhum *cluster*. Os

resultados dos agrupamentos por nível de similaridade são apresentados na Tabela 1 e podem ser visualizados na Figura 3.



(a) Safra de 2000/2001 (b) Safra de 2001/2002 (c) Safra de 2002/2003

(d) Safra de 2003/2004 (e) Safra de 2004/2005 (f) Safra de 2005/2006

(g) Safra de 2006/2007 (h) Safra de 2007/2008

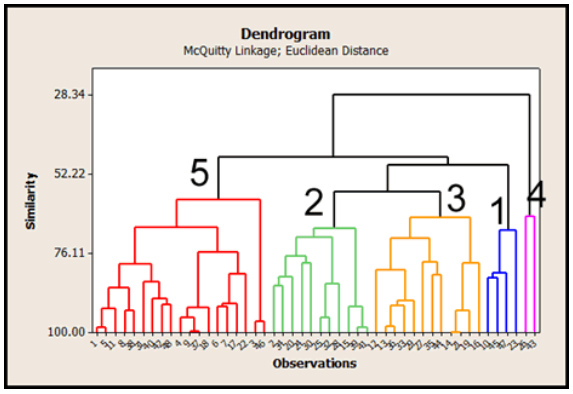
FIGURA 2. Gráfico de estimação do número ótimo de *clusters* para os anos-safra em estudo por meio das estatísticas RMSSTD e RS **Chart for estimating the optimal number of clusters for the crops under study by means of the statistics RS and RMSSTD**.

Para os anos-safra de 2001/2002, 2003/2004, 2004/2005 e 2007/2008, um total de seis *clusters* foram identificados como o número ótimo (Figura 2). Para cada um desses anos-safras, houve um município isolado dos demais clusters (3: Boa Vista da Aparecida, 1: Anahy, 40: São José das Palmeiras e 40: São José das Palmeiras, respectivamente, para os anos-safras). O ano-safra de 2005/2006 também teve seis *clusters*, entretanto sem a existência de municípios isolados. Para os anos-safras de 2000/2001 e 2006/2007, o número ótimo indicado foi de sete *clusters*, com dois municípios isolados (2000/2001 – 26: Medianeira e 43: Serranópolis do Iguaçu; 2006/2007-10: Céu Azul e 23: Marechal Cândido Rondon). A identificação do número ótimo de *clusters* para o ano--safra de 2002/2003 foi de oito, porém dois municípios (10: Céu Azul e 48: Vera Cruz do Oeste) ficaram isolados.

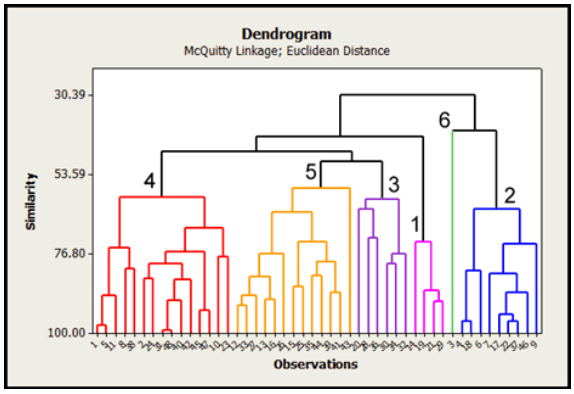
TABELA 1. Processo de agrupamento por similaridade e distância euclidiana dos municípios da área em estudo, considerando as variáveis Prod, Prec, TMed, Rs, *LISA*. **Process of grouping by similarity and Euclidean distance of the municipalities in the study area, considering the variables Prod, Prec, TMED, Rs, *LISA***.

Safra	Cluster	Nível de Similaridade	Nível de Distância	Quantidade Municípios Agrupados	Municípios Agrupados
2000/2001	1	96,60	0,24	2	3 e 46
	2	79,41	2,25	17	1, 4, 5, 6, 7, 8, 9, 11, 17, 18, 22, 31, 37, 38, 40, 42 e 48
	3	69,26	2,19	4	10, 23, 45 e 47
	4	68,76	2,22	11	2, 15, 20, 24, 25, 28, 30, 32, 34, 39 e 41
	5	65,51	2,45	12	12, 13, 14, 16, 19, 21, 27, 29, 33, 35, 36 e 44
	6	**65,05**	**2,49**	**1**	**26**
	7	**65,05**	**2,49**	**1**	**43**
2001/2002	1	73,38	1,56	4	14, 19, 21 e 29
	2	63,69	2,13	9	4, 6, 7, 9, 17, 18, 22, 37 e 46
	3	60,88	2,29	6	20, 28, 30, 32, 34 e 36
	4	60,21	2,33	15	1, 2, 5, 8, 10, 11, 23, 24, 31, 38, 40, 42, 45, 47 e 48
	5	57,57	2,49	13	12, 13, 15, 16, 25, 26, 27, 33, 35, 39, 41, 43 e 44
	6	**40,92**	**3,46**	**1**	**3**
2002/2003	1	80,39	1,46	4	15, 20, 39 e 41
	2	77,29	1,69	5	2, 23, 24, 30 e 47
	3	77,18	1,69	7	13, 14, 16, 19, 21, 27 e 33
	4	72,51	2,04	13	12, 25, 26, 28, 29, 31, 32, 35, 36, 40, 42, 43 e 44
	5	70,64	2,18	10	3, 4, 6, 7, 9, 17, 18, 22, 37 e 46
	6	70,06	2,22	7	1, 5, 8, 11, 34, 38 e 45
	7	**55,78**	**3,28**	**1**	**10**
	8	**45,33**	**4,06**	**1**	**48**
2003/2004	1	71,30	1,64	3	16, 27 e 44
	2	63,32	2,10	11	12, 13, 15, 25, 28, 33, 35, 36, 39, 40 e 41
	3	63,15	2,11	3	10, 20 e 23
	4	62,23	2,16	14	3, 4, 5, 6, 7, 8, 9, 11, 17, 18, 22, 37, 38 e 46
	5	57,59	2,43	16	2, 14, 19, 21, 24, 26, 29, 30, 31, 32, 34, 42, 43, 45, 47 e 48
	6	**57,14**	**2,45**	**1**	**1**
2004/2005	1	69,09	1,91	4	15, 20, 39 e 41
	2	59,39	2,51	2	10 e 23
	3	55,66	2,74	6	1, 24, 30, 32, 34 e 43
	4	51,42	3,01	24	2, 3, 4, 5, 6, 7, 8, 9, 11, 14, 17, 18, 19, 21, 22, 29, 31, 37, 38, 42, 45, 46, 47 e 48
	5	50,94	3,04	11	12, 13, 16, 25, 26, 27, 28, 33, 35, 36 e 44
	6	**27,77**	**4,47**	**1**	**40**
2005/2006	1	86,29	0,89	2	27 e 33
	2	82,44	1,14	7	3, 4, 7, 18, 22, 37 e 46
	3	78,36	1,40	3	6, 9 e 17
	4	67,28	2,12	3	23, 30 e 34
	5	63,95	2,33	10	1, 2, 5, 8, 10, 11, 24, 38, 45 e 47
	6	61,34	2,50	23	12, 13, 14, 15, 16, 19, 20, 21, 25, 26, 28, 29, 31, 32, 35, 36, 39, 40, 41, 42, 43, 44 e 48
2006/2007	1	80,60	1,15	4	14, 19, 21 e 29
	2	66,94	1,97	9	2, 15, 20, 24, 30, 32, 39, 41 e 44
	3	63,60	2,17	5	13, 16, 27, 33 e 40
	4	60,94	2,32	18	1, 3, 4, 5, 6, 7, 8, 9, 11, 17, 18, 22, 34, 37, 38, 45, 46 e 47
	5	60,51	2,35	10	12, 25, 26, 28, 31, 35, 36, 42, 43 e 48
	6	**42,70**	**3,41**	**1**	**23**
	7	**41,24**	**3,50**	**1**	**10**
2007/2008	1	79,61	1,34	10	3, 4, 6, 7, 9, 17, 18, 22, 37 e 46
	2	64,10	2,36	5	13, 16, 27, 32 e 44
	3	61,18	2,55	13	1, 5, 8, 11, 14, 15, 19, 20, 21, 29, 39, 41 e 47
	4	58,87	2,71	11	2, 12, 23, 24, 30, 31, 33, 34, 36, 42 e 45
	5	57,10	2,82	8	10, 25, 26, 28, 35, 38, 43 e 48
	6	**40,82**	**3,89**	**1**	**40**

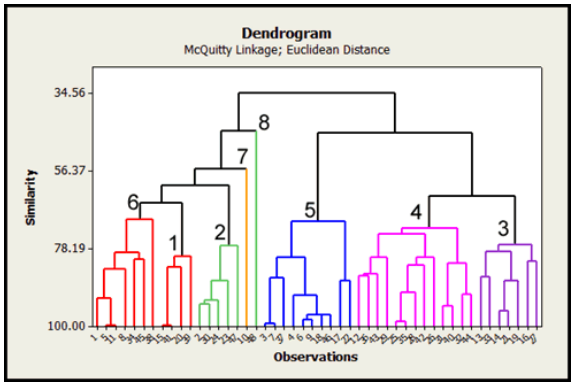
Em negrito estão os municípios que ficaram isolados, sem associação a *clusters*.



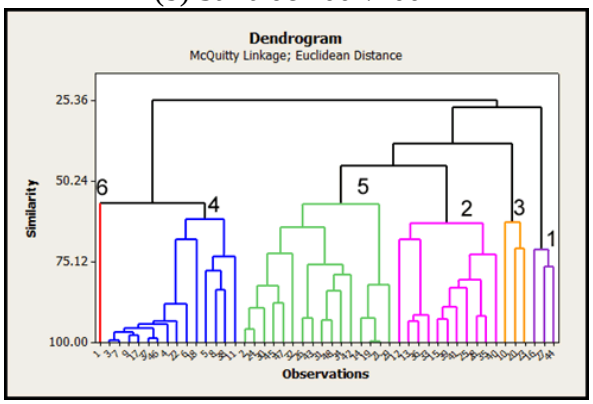
(a) Safra de 2000/2001



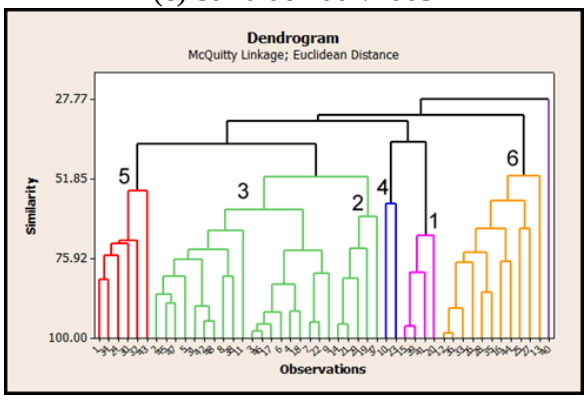
(b) Safra de 2001/2002



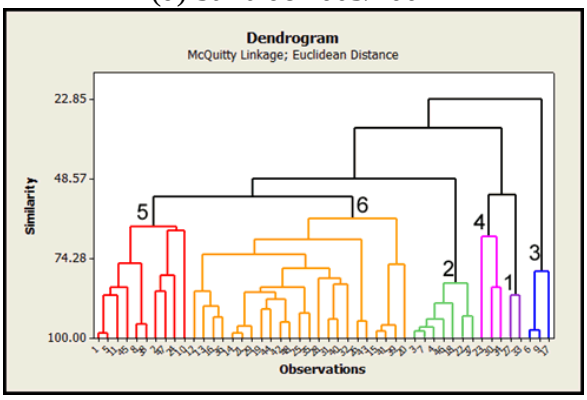
(c) Safra de 2002/2003



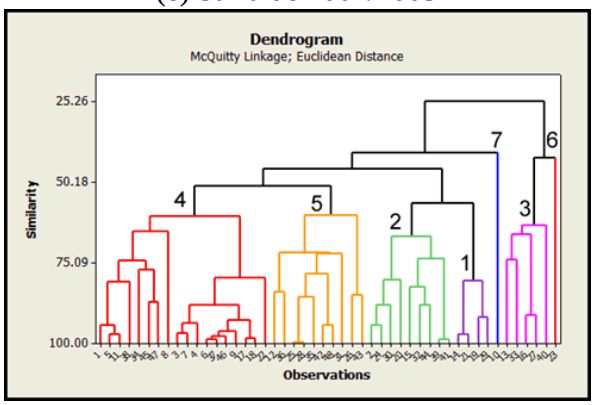
(d) Safra de 2003/2004



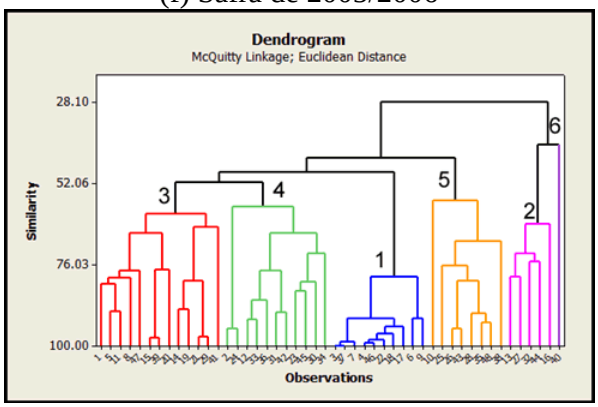
(e) Safra de 2004/2005



(f) Safra de 2005/2006



(g) Safra de 2006/2007



(h) Safra de 2007/2008

FIGURA 3. Dendogramas gerados com as variáveis produtividade da soja (t ha^{-1}), precipitação pluvial (mm), temperatura média do ar (ºC), radiação solar global média (W m^{-2}) e índice *LISA* para os 48 municípios da área de estudo, em oito anos. **Dendogram generated with variables soybean yield (t ha^{-1}), rainfall (mm), average air temperature (° C), global irradiance average (W m^{-2}) and index *LISA* for the 48 municipalities of the study area in eight years**.

Na análise de similaridade dos municípios apresentada na Tabela 1 e Figura 3, identificou-se a quantidade de *clusters* indicada pelas estatísticas *RMSSTD* e *RS* para cada um dos anos-safras estudados. A similaridade variou entre 27,77% (*cluster* 6 do ano-safra de 2004/2005) e 96,60% (*cluster* 1 do ano-safra de 2005/2006) com média de 63,89% para os clusters e anos-safras avaliados. Verificou-se também que, na medida em que se aumenta o número de clusters dentro de cada ano-safra, menor é o nível de similaridade, sendo esse fato comprovado pela característica do índice *RS*, que mede a heterogeneidade e que tem seu valor maior de acordo com o aumento do número de clusters. Constatou-se que, em cada ano-safra, de acordo com a diminuição da similaridade dos clusters, há um aumento da distância entre eles, fato também observado por FERREIRA et al. (2008) e em concordância com as ideias de CONDIT (1996), segundo o qual a proximidade geográfica seria o único fator confiável para se prever a similaridade entre áreas.

Também foi realizada uma simulação de agrupamentos, semelhante à que foi apresentada na Tabela 1, entretanto sem a variável *LISA*. Observa-se que o número ótimo de *clusters* sofreu variações, o que era esperado, uma vez que a variável *LISA* representa o nível de autocorrelação da produtividade entre os municípios. Para uma comparação de similaridade, as variáveis, sem a *LISA*, foram submetidas à geração de dendogramas. Comparando os resultados, verificou-se que os níveis de similaridade foram sempre maiores quando não se utilizou a variável de autocorrelação *LISA*.

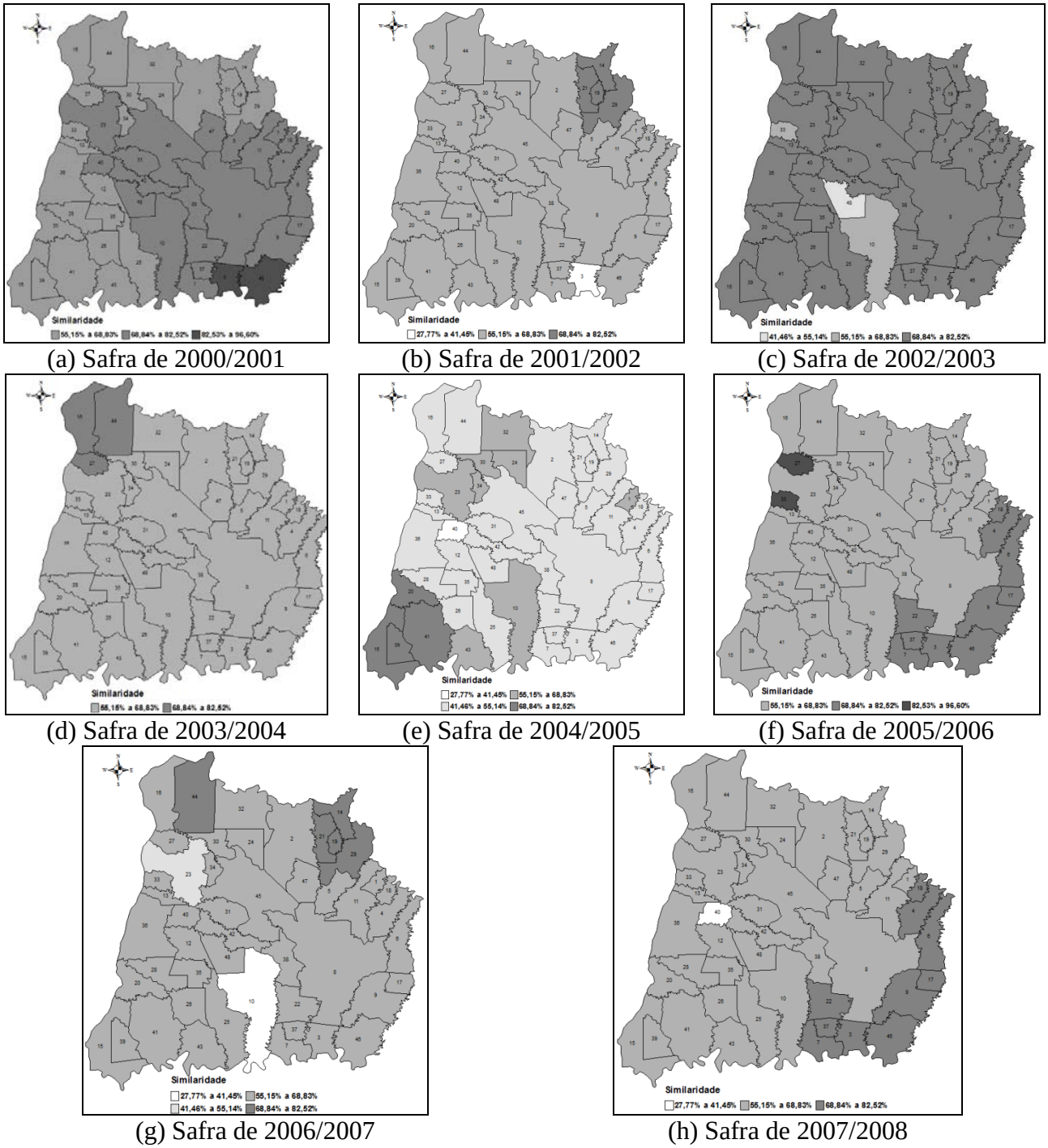
A Figura 3 apresenta os dendogramas de similaridade para os anos-safras e variáveis estudadas. A distância euclidiana máxima entre todos os municípios, como um único *cluster*, onde todos os municípios fariam parte de um único grupo, está próxima a 5,10% na safra de 2000/2001 e a maior similaridade encontrada está próxima a 34,56% na safra de 2002/2003.

Na Figura 4, foram construídos mapas temáticos para representar os agrupamentos obtidos segundo a análise de agrupamento. Uma vez que cada safra pode ter seu número ótimo para a quantidade de *clusters*, as classes para os mapas foram obtidas por meio de divisão do intervalo entre a menor similaridade (ano-safra de 2004/2005, com 27,77%) e a maior similaridade (ano-safra de 2000/2001, com 96,60%). Esse valor foi então dividido em cinco classes de intervalos iguais (27,77% a 41,45%; 41,46% a 55,14%; 55,15% a 68,83%; 68,84% a 82,52%; e 82,53% a 96,60%), formando cinco *clusters*.

É possível verificar, na Figura 4, que dois anos-safras (2000/2001 e 2005/2006) apresentaram seus municípios agrupados entre 55,15% e 96,60% de similaridade, e no ano-safra de 2000/2001 ocorreu a maior similaridade do período em estudo (96,60% para os municípios 3: Boa Vista da Aparecida e 46: Três Barras do Paraná). No ano-safra de 2004/2005, que teve o menor índice de similaridade (27,77% para o município 40: São José das Palmeiras, que ficou isolado), seu intervalo foi definido entre 27,77% e 82,52% de similaridade entre quatro agrupamentos. Nos anos-safras de 2001/2002 e 2007/2008 foram identificados três *clusters* com intervalos semelhantes de 27,77% a 82,52%. O ano-safra de 2002/2003 também foi organizado em três *clusters*, porém com o intervalo de 41,46% a 82,52%. A safra de 2006/2007 teve sua similaridade identificada em quatro *clusters*, no intervalo de 27,77% a 82,52%.

A repetição dos agrupamentos nem sempre ocorre na mesma faixa de similaridade/distância, entretanto os municípios, em boa quantidade, repetem-se nestes agrupamentos. Como exemplos, nas safras de 2000/2001 (5a) e 2002/2003 (5c), os municípios 23: Marechal Cândido Rondon, 45: Toledo e 47: Tupãssi fazem parte do *cluster* que representa o intervalo de 68,84% a 82,52% de similaridade; nas safras de 2001/2002 (5b), 2003/2004 (5d), 2005/2006 (5f) e 2007/2008 (5h), estes mesmos municípios encontram-se no *cluster* de faixa entre 55,15% e 68,83%. Na safra de 2004/2005 (5e), destes municípios, apenas 45: Toledo e 47: Tupãssi encontram-se no mesmo *cluster*, de 41,46% a 55,14%, enquanto o município 23: Marechal Cândido Rondon se encontra no intervalo de 55,15% a 68,83%, mantendo a similaridade da safra de 2001/2002. Na safra de 2006/2007 (5g), novamente, apenas 45: Toledo e 47: Tupãssi encontram-se no mesmo *cluster* (de 55,15% a 68,83%), enquanto o município 23: Marechal Cândido Rondon encontra-se no intervalo de 41,46% a 55,14%, isolado dos demais. É possível, também, verificar, visualmente nos mapas, que existem municípios, em todas as regiões, que se mantêm agrupados em todas as safras.

Outra estratégia de análise consistiu em gerar *clusters* considerando conjuntamente todas as safras (2000/2001 a 2007/2008) e variáveis (Prod, Prec, Tmed, Rs, *LISA*) em estudo. A Figura 5 apresenta os resultados das estatísticas *RMSSTD* e *RS* para esse conjunto de dados, que identificou sete como número ótimo para a quantidade de *clusters*, o que gerou três municípios isolados. A Tabela 2 apresenta a análise de similaridade dos municípios para os sete *clusters*, e o resultado em forma de mapa temático pode ser visualizado na Figura 6, juntamente com o dendograma.



(a) Safra de 2000/2001 (b) Safra de 2001/2002 (c) Safra de 2002/2003

(d) Safra de 2003/2004 (e) Safra de 2004/2005 (f) Safra de 2005/2006

(g) Safra de 2006/2007 (h) Safra de 2007/2008

FIGURA 4. Mapa temático de análise dos agrupamentos dos municípios da pesquisa com base no índice de similaridade considerando as variáveis na Produtividade da soja (t ha^{-1}), Precipitação pluvial (mm), Temperatura Média do ar (ºC), Radiação Solar Global Média (W m^{-2}) e Índice *LISA* Univariado. **Thematic map analysis of groups of municipalities in the search based on the similarity index considering the variables in the productivity of soybean (t ha^{-1}), Rainfall (mm) Average temperature of air (° C), Global Solar Radiation Average (W m^{-2}) and univariate *LISA* Index**.

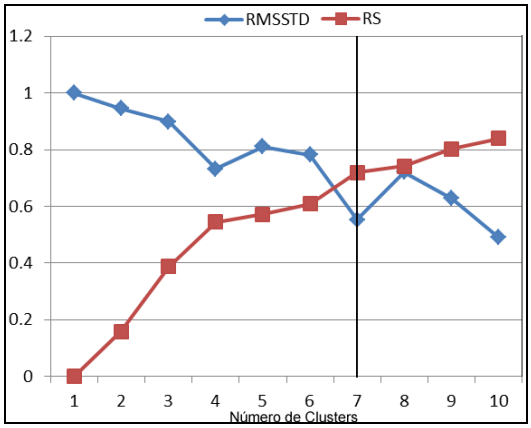


FIGURA 5. Gráfico de estimação do número ótimo de *clusters* para todas as safras em estudo, como um único conjunto, por meio das estatísticas RMSSTD e RS. **Chart for estimating the optimal number of clusters for all crops under study, as a single set, by means of statistics RMSSTD and RS**.

TABELA 2. Processo de agrupamento por similaridade e distância dos municípios da área em estudo para os anos-safras de 2000/2001 a 2007/2008. **Process of grouping by similarity and distance of the municipalities in the study area for the harvest of 2000/2001 to 2007/2008**.

Cluster	Nível de Similaridade	Nível de Distância	Quantidade Municípios Agrupados	Municípios Agrupados
1	62,08	5,46	10	3, 4, 6, 7, 9, 17, 18, 22, 37 e 46
2	52,39	6,85	6	12, 13, 16, 27, 33 e 36
3	48,89	7,36	5	15, 26, 39, 41 e 43
4	43,89	8,08	24	1, 2, 5, 8, 11, 14, 19, 20, 21, 24, 25, 28, 29, 30, 31, 32, 34, 35, 38, 42, 44, 45, 47 e 48
5	43,08	8,20	1	10
6	43,08	8,20	1	23
7	40,73	8,53	1	40

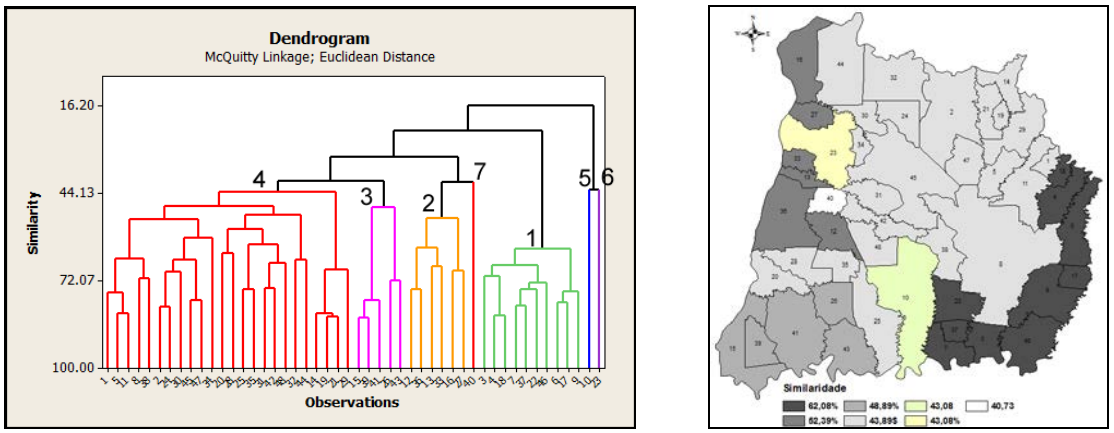


FIGURA 6. Mapa temático e dendograma de análise dos agrupamentos dos municípios da pesquisa com base no índice de similaridades considerando as variáveis na produtividade da soja (t ha^{-1}), precipitação pluvial (mm), temperatura média do ar (ºC), radiação solar global média (W m^{-2}) e índice *LISA* univariado, para todos os anos-safra em estudo. **Thematic map and dendogram analysis of groups of municipalities in the search based on the similarity index considering the variables on soybean yield (t ha^{-1}), rainfall (mm), average air temperature (° C), global solar radiation average (W m^{-2}) and *LISA* univariate index for all crops under study**.

Pela Tabela 2, observa-se que os municípios alocados ao *cluster* com maior nível de similaridade (62,08%) foram: 3: Boa Vista da Aparecida, 4: Braganey, 6: Campo Bonito, 7: Capitão Leônidas Marques, 9: Catanduvas, 17: Ibema, 18: Iguatu, 22: Lindoeste, 37: Santa Lúcia e 46: Três Barras do Paraná. Esses municípios estão localizados a leste da região em estudo e têm como característica um relevo mais acidentado, quando comparados aos demais municípios da região.

Os seis municípios (12: Diamante D'Oeste, 13: Entre Rios do Oeste, 16: Guaíra, 27: Mercedes, 33: Pato Bragado e 36: Santa Helena) agrupados com o segundo maior nível de similaridade (52,39%) estão concentrados na região centro-oeste e noroeste da região estudada, mostrando que apresentam características semelhantes, em termos climáticos e de produtividade, ao longo dos oito anos estudados.

Com 48,89% de similaridade (*cluster* 3), foram agrupados um total de 24 municípios, ou seja, a metade dos municípios estudados (1: Anahy, 2: Assis Chateaubriand, 5: Cafelândia, 8: Cascavel, 11: Corbélia, 14: Formosa do Oeste, 19: Iracema do Oeste, 20: Itaipulândia, 21: Jesuítas, 24: Maripá, 25: Matelândia, 28: Missal, 29: Nova Aurora, 30: Nova Santa Rosa, 31: Ouro Verde do Oeste, 32: Palotina, 34: Quatro Pontes, 35: Ramilândia, 38: Santa Tereza do Oeste, 42: São Pedro do Iguaçu, 44: Terra Roxa, 45: Toledo, 47: Tupãssi e 48: Vera Cruz do Oeste). De maneira geral, eles estão localizados onde se encontram os municípios de maior produção agrícola da região estudada, e onde, normalmente, são encontradas também as maiores produtividades de soja do Estado do Paraná.

Três municípios ficaram isolados, formando cada um deles um *cluster*. Esses municípios são: 10: Céu Azul (43,08%), 23: Marechal Cândido Rondon (43,08%) e 40: São José das Palmeiras (40,73%). Os mesmos municípios agrupados no *Cluster 1* e apresentados na Tabela 2 foram exatamente os mesmos na análise de agrupamentos por safra, nos anos-safras de 2002/2003 (*cluster* 2) e 2007/2008 (*cluster* 1), com maior produtividade, o que pode evidenciar uma necessidade de estudo na produtividade nessas duas safras e a relação das variáveis agrometeorológicas com ela.

A similaridade multivariada leva em consideração aspectos de coesão interna e isolamento externo entre os municípios, podendo-se avaliar, como diferentes, municípios com valores iguais nas variáveis em estudo. Como exemplo, pode-se verificar que os valores dos municípios (3: Boa Vista da Aparecida e 4: Braganey) da safra de 2000/2001 (produtividade da soja de 2,90 t ha^{-1} para os municípios 3: Boa Vista da Aparecida e 4: Braganey, precipitação pluvial de 1.487,2 mm, temperatura média do ar de 22,7 ºC, radiação solar global média de 498,7 W m^{-2}, índice *LISA* de 1,08 e -0,13), que apesar de sua semelhança numérica, analisados de maneira individual, são diferentes quando comparados na forma multivariada em função das variáveis em estudo que possam influenciar na produtividade. Esses aspectos levam a utilizar os critérios de análise de agrupamentos para definir estratégias de ações de planejamento para o cultivo da soja, aspectos esses que não devem ser negligenciados na atividade agrícola.

Apesar de os municípios se localizarem distribuídos dentro da área da pesquisa, e em alguns casos eles se encontrarem distantes uns dos outros, ainda assim se encontram níveis de similaridade satisfatórios nos valores das variáveis em estudo. Essa constatação leva a supor que esses valores possuem diferença numérica, mas na análise multivariada dos valores formam conjuntos similares.

Por meio dos mapas apresentados na Figura 4, pode-se verificar que há municípios com valores diferentes para as variáveis em estudo, que podem ser vistos como similares por meio da estatística multivariada, formando os *clusters*.

A localização dos municípios, em região contígua, pressupõe que exista uma forte relação entre os resultados obtidos em um município com os resultados do outro em função de estarem no mesmo espaço geográfico. Muitas suposições podem ser analisadas em função das similaridades encontradas. As cinco variáveis analisadas possuem diferenças numéricas, analisadas de maneira individual, tanto nos valores como em suas dimensões. Em termos de produtividade, o município 3: Boa Vista da Aparecida, obteve, na Safra de 2002/2003, a maior verificada em toda a área da pesquisa, 3,72 t ha^{-1}. Verificando os mapas da Figura 4, foi possível constatar que os municípios 7: Capitão Leônidas Marques, 37: Santa Lúcia e 46: Três Barras do Paraná estão sempre no mesmo

cluster, em praticamente todas as safras. A exceção está nas safras de 2000/2001 (em que o município 7: Capitão Leônidas Marques fez parte de outro *cluster*) e 2001/2002 em que o município 3:Boa Vista da Aparecida ficou isolado, tendo o menor valor de similaridade. É desejável que a similaridade com esses municípios seja tomada como meta, o que leva a concluir que os demais municípios da pesquisa associados aos mesmos *clusters* destes municípios, podem chegar ao mesmo valor de produtividade, baseado nas variáveis em estudo.

CONCLUSÕES

Pela técnica da análise multivariada de agrupamento, foi possível a formação de grupos de municípios utilizando as similaridades das variáveis em análise, mostrando-se assim como uma ferramenta valiosa no entendimento da distribuição espacial de dados agrometeorológicos e da produtividade da soja no oeste do Paraná.

A análise de agrupamento foi uma técnica útil para melhor gestão das atividades de produção da agricultura, em função de que, com o agrupamento, é possível estabelecer-se similaridades que proporcionem parâmetros para uma melhor gestão dos processos de produção que traga, quantitativa e qualitativamente, resultados almejados pelo agricultor.

As atividades agrícolas possuem características próprias, sendo difícil estabelecer-se um padrão para os processos produtivos. Isso fica evidente nas diferenças encontradas nos 48 municípios analisados na pesquisa.

AGRADECIMENTOS

Ao CNPq, à CAPES e à Fundação Araucária, pelo apoio financeiro.

REFERÊNCIAS

ANDRADE, N. L. R. de; XAVIER, F. V.; ALVES, E. C. R. de F.; SILVEIRA, A.; OLIVEIRA, C. U. R. Caracterização morfométrica e pluviométrica da bacia do Rio Manso – MT. *Revista Brasileira de Geociências*, São Paulo, v. 27, p. 237-248. 2008.

ANSELIN, L. Local indicators of spatial association – LISA. *Geographical Analysis,* Ohio, v. 27, n. 2, p. 93-115, 1995.

BOSCHI, R. S.; OLIVEIRA, S. R. de M.; ASSAD, E. D. Técnicas de mineração de dados para análise da precipitação pluvial decenal no Rio Grande do Sul. *Engenharia Agrícola,* Jaboticabal, v. 31, n. 6, p. 1.189-1.201, 2011.

BUSSAB, W. de O. *Introdução à análise de agrupamento*. São Paulo: IME-USP, 1990. 105p.

CONDIT, R. Defining and mapping vegetation types in mega-diverse tropical forests. *Trends in Ecology and Evolution,* v. 11, n.1, p .4-5, 1996.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. *Análise Multivariada:* para os cursos de Administração, Ciências Contábeis e Economia. São Paulo: Atlas, 2007. 344p.

ESRI. *ArcGIS Spatial Analyst*. 2011. Disponível em: <http://www.esri.com/software/arcgis/extensions/spatialanalyst/surface.html>. Acesso em: 17 jul. 2012.

FERREIRA, R. L. C.; MOTA, A. C.; SILVA, J. A. A.; MARANGON, L. C.; SANTOS, E. S. Comparação de duas metodologias multivariadas no estudo de similaridade entre fragmentos de Floresta Atlântica. *Revista Árvore.* Viçosa-MG, v. 32, n.3, p. 501-511, 2008.

FIORINI, C. V. A.; DA SILVA, D. J. H.; E SILVA, F. F.; MIZUBUTI, E. S. G.; ALVES, D. P.; CARDOSO, T. S. Agrupamento de curvas de progresso de requeima, em tomateiro originado de cruzamento interespecífico. *Pesquisa Agropecuária Brasileira*, Brasília, v. 45, n.10, p.1.095-1.101, 2010.

GIMENES, F. R.; GIMENES, R. M. T.; OPAZO, M. A. U. Os processos de integração econômica sob a ótica da análise estatística de agrupamento. *FAE,* Curitiba, v. 7, n. 2, p. 19-32, jul./dez. 2004.

HARDLE, W.; SIMAR, L. *Applied multivariate statistical analysis*. 2nd. ed. Berlin: Springer, 2007. 486p.

JACOX, E. H.; SAMET, H. Spatial join techniques. *ACM Transactions on Database Systems,* New York, v. 32, n. 1, p. 7-75, mar. 2007.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 3rd. ed. New Jersey: Prentice-Hall, 1992. 800p.

KOVÁCS, F.; LEGÁNY, C.; BABOS, A. *Cluster validity measurement techniques*. In: INTERNATIONAL SYMPOSIUM OF HUNGARIAN RESEARCHERS ON COMPUTATIONAL INTELLIGENCE, 6., 2005, Budapest. *Proceedings.*… p. 18-19.

KUNZ, V. L.; GABRIEL FILHO, A.; PRIMO, M. A.;GURGACZ, F.; FEY, E. Distribuição de palha por colhedoras autopropelidas na colheita da soja. *Engenharia Agrícola,* Jaboticabal, v. 28, n. 1. p.125-135, 2008.

LE GALLO, J.; ERTHUR, C. Exploratory spatial data analysis of the distribution of regional per capita GDP in Europe, 1980-1995. *Papers in Regional Science,* Urbana, v. 82, n. 2, p. 175-201, 2003.

MINITAB. *Minitab Statistical Software*. 2011. Disponível em: <http://www.minitab.com/en-US/Products/>. Acesso em: 17 jul. 2012.

OLIVEIRA, J. T. A.; BERGAMASCO, S. M. P. P. Impactos Ambientais de Sistemas de Produção Segundo as Lógicas Produtivas. *Revista Eletrônica do Mestrado em Educação Ambiental*. Rio Grande do Sul, v. 10, p. 51-61, 2003.

OPENGEODA. *GeoDa Center for Geospatial Analysis and Computation.* 2011. Disponível em: <http://geodacenter.asu.edu/about>. Acesso em: 17 jul. 2012.

SAS INSTITUTE. *SAS System:* SAS/STAT. Version 9.0 (software). Cary, 2011.

SEAB. *Secretaria da Agricultura e do Abastecimento do Paraná*. 2010.

SIMEPAR. *Sistema Meteorológico do Paraná*. 2010.

TEIXEIRA, R. F. A. P.; BERTELLA, M. A. A distribuição espacial da indústria do vestuário no Brasil. *Revista de Economia.* Curitiba, v. 36, p. 91-118, 2010.

WANG, K.; WANG, B.; PENG, L. CVAP: validation for cluster analyses. *Data Science Journal,* v. 8, p. 88-93, 2009.