UNIVERSIDADE ESTUDUAL DO OESTE DO PARANÁ *CAMPUS* CASCAVEL

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA:

ENGENHARIA DE SISTEMAS AGROINDUSTRIAIS E TECNOLOGIA DA PRODUÇÃO AGRÍCOLA

ANÁLISE DA PRODUTIVIDADE DA SOJA ASSOCIADA A FATORES AGROMETEOROLÓGICOS, POR MEIO DE ESTATÍSTICA ESPACIAL DE ÁREA, NA REGIÃO OESTE DO ESTADO DO PARANÁ

EVERTON COIMBRA DE ARAÚJO

Cascavel – Paraná – Brasil

Dezembro– 2012

EVERTON COIMBRA DE ARAÚJO

ANÁLISE DA PRODUTIVIDADE DA SOJA ASSOCIADA A FATORES AGROMETEOROLÓGICOS, POR MEIO DE ESTATÍSTICA ESPACIAL DE ÁREA, NA REGIÃO OESTE DO ESTADO DO PARANÁ

Tese apresentada ao Programa de Pós-Graduação em Engenharia Agrícola em cumprimento parcial aos requisitos para obtenção do título de Doutor em Engenharia Agrícola, área de concentração em Engenharia de Sistemas Agroindustriais e Tecnologia da Produção Agrícola, da Universidade Estadual do Oeste do Paraná, *campus* Cascavel.

Orientador: Prof. Dr. Miguel Angel Uribe Opazo

Coorientador: Prof. Dr. Jerry Adriani Johann

Cascavel – Paraná – Brasil

Dezembro– 2012

EVERTON COIMBRA DE ARAÚJO

Dados Internacionais de Catalogação-na-Publicação (CIP)
Biblioteca Central do Campus de Cascavel – Unioeste
Ficha catalográfica elaborada por Jeanine da Silva Barros CRB-9/1362

C633a Coimbra, Everton
Análise da produtividade da soja associada a fatores agrometeorológicos, por meio de estatística espacial de área na Região Oeste do Estado do Paraná. /— Cascavel, PR: UNIOESTE, 2012.
116 f. ; 30 cm.

Orientador: Prof. Dr. Miguel Angel Uribe Opazo
Co-orientador: Prof. Dr. Jerry Adriani Johann
Tese (Doutorado) – Universidade Estadual do Oeste do Paraná.
Programa de Pós-Graduação Stricto Sensu em Engenharia Agrícola, Centro de Ciências Exatas e Tecnológicas.
Bibliografia.

1. Autocorrelação espacial. 2. Similaridade espacial. 3. Regressão espacial. I. Universidade Estadual do Oeste do Paraná. II. Título.

CDD 21. ed. 631.86

Revisão de português, inglês e normas realizadas em 10 de janeiro de 2013 por Dhandara Soares de Lima.

EVERTON COIMBRA DE ARAÚJO

Análise da produtividade da soja associada a fatores agrometeorológicos, por meio de estatística espacial de área, na região oeste do estado do Paraná

Tese apresentada ao Programa de Pós-Graduação "*stricto-sensu*" em Engenharia Agrícola em cumprimento parcial aos requisitos para obtenção do título de Doutor em Engenharia Agrícola, área de concentração em Engenharia de Sistemas Agroindustriais e Tecnologia da Produção Agrícola, da Universidade Estadual do Oeste do Paraná, *campus* Cascavel, apresentada à seguinte banca examinadora:

Orientador: Prof. Dr. Miguel Angel Uribe Opazo

Centro de Ciências Exatas e Tecnológicas, UNIOESTE

Coorientador: Prof. Dr. Jerry Adriani Johann

Centro de Ciências Exatas e Tecnológicas, UNIOESTE

Banca 1: Prof. Dr. Rubens Augusto Camargo Lamparelli

UNICAMP, Núcleo Interdisciplinar de Planejamento Energético

Banca 2: Profa. Dra. Luciana Pagliosa Carvalho Guedes

Centro de Ciências Exatas e Tecnológicas, UNIOESTE

Banca 3: Prof. Dr. Adair Santa Catarina

Centro de Ciências Exatas e Tecnológicas, UNIOESTE

Banca 4: Profa. Dra. Rosangela Aparecida Botinha Assumpção

Universidade Tecnológica Federal do Paraná – Campus Dois Vizinhos

Cascavel – Paraná – Brasil

Dezembro– 2012

BIOGRAFIA

Everton Coimbra de Araújo, graduado em Tecnologia em Processamento de Dados pelo CESUFOZ em 2000; especialista em Projeto e Desenvolvimento de Sistemas Baseados em Objetos para Internet pela UTFPR em 2006; e Mestre em Ciência da Computação pela UFSC em 2002. Ingressou no ano de 2009 no Programa de Pós-Graduação em Engenharia Agrícola – Engenharia de Sistemas Agroindustriais/Tecnologia da Produção Agrícola, sob a orientação do Prof. Dr. Miguel Angel Uribe-Opazo, estudando o tema: Estatística Espacial de Área na Produtividade da Soja e Fatores Agrometeorológicos na Região Oeste do Estado do Paraná.

Agradecimentos

Agradeço a todos que acompanharam mais esta etapa de minha vida pessoal e profissional e, de alguma maneira, me auxiliaram.

Agradeço com um carinho especial meu Orientador, Prof. Dr. Miguel Angel Uribe-Opazo, que deu a mim a oportunidade e a confiança para ingressar no programa de Pós-Graduação em Engenharia Agrícola (PGEAGRI) da UNIOESTE, *campus* de Cascavel, e que me propiciou grandes aprendizados, não só na vida acadêmica, mas também na pessoal, estando sempre presente nos momentos em que precisei de apoio.

Agradeço ao Prof. Dr. Jerry Adriani Johann, meu coorientador, pela importante orientação durante o desenvolvimento da tese.

Obrigado.

ANÁLISE DA PRODUTIVIDADE DA SOJA ASSOCIADA A FATORES AGROMETEOROLÓGICOS, POR MEIO DE ESTATÍSTICA ESPACIAL DE ÁREA, NA REGIÃO OESTE DO ESTADO DO PARANÁ

RESUMO

Este trabalho apresenta métodos para serem aplicados na estatística espacial de área na produtividade da soja e fatores agrometeorológicos na região oeste do estado do Paraná. Os dados utilizados estão relacionados aos anos-safra de 2000/2001 a 2007/2008, sendo as variáveis: produtividade da soja (t ha^{-1}) e agrometeorológicas, tais como precipitação pluvial (mm), temperatura média (^oC) e radiação solar global média (W m^{-2}). Em uma primeira fase foram utilizados índices de autocorrelação espacial (Moran Global e Local) e apresentados modelos de regressão espacial múltipla, com avaliações de desempenho. A estimativa dos parâmetros dos modelos ajustados se deu pelo uso do método de Máxima Verossimilhança e a avaliação do desempenho dos modelos foi realizada com base no coeficiente de determinação (R^2), no máximo valor do logaritmo da função do máximo valor do logaritmo da função verossimilhança e no critério de informação bayesiano de Schwarz. Em uma segunda etapa foram realizadas análises de agrupamento espacial por meio da estatística multivariada, buscando identificar associações no mesmo conjunto de variáveis, porém com um número maior de anos-safra. Finalmente, os dados de um ano-safra foram aplicados em uma abordagem baseada em agrupamento difuso, por meio do algoritmo *Fuzzy c-Means*, tendo a similaridade medida pela definição de um índice com este objetivo. O estudo da primeira fase permitiu verificar a correlação e a autocorrelação espacial entre a produtividade da soja e os elementos agrometeorológicos, por meio da análise espacial de área, usando técnicas como o índice *I* de Moran Global e Local uni e bivariado e os testes de significância. Foi possível demonstrar que, por meio dos indicadores de desempenho utilizados, os modelos SAR e CAR ofereceram melhores resultados em relação ao modelo de regressão múltipla clássica. Na segunda fase, foi possível apresentar a formação de grupos de municípios utilizando as similaridades das variáveis em análise. A análise de agrupamento foi um instrumento útil para uma melhor gestão das atividades de produção da agricultura, em função de que, com o agrupamento, foi possível se estabelecer similaridades que proporcionem parâmetros para uma melhor gestão dos processos de produção que traga, quantitativa e qualitativamente, resultados almejados pelo agricultor. Na etapa final, por meio do algoritmo *Fuzzy c-Means,* foi possível a formação de grupos de municípios similares à produtividade de soja, utilizando o Método de Decisão pelo Maior Grau de Pertinência (*MDMGP*) e o Método de Decisão pelo Limiar *β* (*MDL β*). Posteriormente, a identificação do número adequado de agrupamentos foi obtida utilizando a Entropia de Partição Modificada. Para mensurar o nível de similaridade de cada agrupamento, foi criado e utilizado um Índice de Similaridade de *Clusters* (ISC_l), que considera o grau de pertinência de cada município dentro do agrupamento a que pertence. Dentro das perspectivas deste estudo, o método empregado se mostrou adequado, permitindo identificar agrupamentos de municípios com graus de similaridades da ordem de 60 a 78%.

Palavras-chave: Autocorrelação espacial, Similaridade espacial, Regressão espacial

ANÁLISE DA PRODUTIVIDADE DA SOJA ASSOCIADA A FATORES AGROMETEOROLÓGICOS, POR MEIO DE ESTATÍSTICA ESPACIAL DE ÁREA, NA REGIÃO OESTE DO ESTADO DO PARANÁ

ABSTRACT

This paper aimed to present methods to be applied in the area of spatial statistics on soybean yield and agrometeorological factors in Western Paraná state. The data used, related to crop years from 2000/2001 to 2007/2008, are the following variables: soybean yield (t ha^{-1}) and agrometeorological factors, such as rainfall (mm), average temperature (oC) and solar global radiation average (W m^{-2}). In the first phase,it was used indices of spatial autocorrelation (Moran Global and Local) and presented multiple spatial regression models, with performance evaluations. The estimation of parameters occurred when using the Maximum Likelihood method and the performance evaluation of the models was based on the coefficient of determination (R^2), the maximum value of the function of the logarithm of the maximum value of the likelihood function logarithm and the Bayesian information criterion of Schwarz. In a second step, cluster analysis was performed using spatial statistical multivariate associations, seeking to identify the same set of variables, but with a larger number of crop years. Finally, the data from one crop year were utilized in an approach based on fuzzy clustering, through the Fuzzy C-Means algorithm and the similarity measure by defining an index for this purpose. The first phase of the study showed the correlation between spatial autocorrelation and soybean yield and agrometeorological elements, through the analysis of spatial area, using techniques such as index Global Moran's I and Local univariate and bivariate and significance tests. It was possible to demonstrate, through the performance indicators used, that the SAR and CAR models offered better results than the classical multiple regression model. In the second phase, it was possible to present the formation of groups of cities using the similarities of the variables under analysis. Cluster analysis is a useful tool for better management of production activities in agriculture, since, with the grouping, it was possible to establish similarities parameters that provide better management of production processes that bring quantitative and qualitatively better, results sought by the farmer. In the final step, through the use of Fuzzy C-Means algorithm, it was possible to form groups of cities of similar soybean yield using the method of decision by the Higher Degree of Relevance (MDMGP) and Method of Decision Threshold by β (β CDM). Subsequently, identification of the adequate number of clusters was obtained using modified partition entropy. To measure the degree of similarity of each cluster, a Cluster Similarity Index (ISC_l) was designed and used, which considers the degree of relevance of each city within the group to which it belongs. Within the perspective of this study, the method used was adequate, allowing to identify clusters of cities with degrees of similarities in the order of 60 to 78%.

KEY WORDS: Spatial autocorrelation, spatial similarity, similarity index.

SUMÁRIO

Lista de Tabelas

Lista de Figuras

Lista de Siglas

AEDE	Análise Exploratória de Dados Espaciais
CAR	*Conditional Auto Regressive Model*
CS	*Compactness and Separation*
EEA	Estatística Espacial de Área
FCM	*Fuzzy c-Means*
FPI	*Fuzziness Performance Index*
ICC	*Inter Class Contrast*
LISA	*Local Indicator of Spatial Association*
MPE	*Modified Partition Entropy*
MQO	Mínimos Quadrados Ordinários
PIB	Produto Interno Bruto
RMSSTD	*Root Mean Square Standard Deviation*
RS	*R-Square*
SAR	*Spatial Auto Regressive Model*
SIG	Sistemas de Informação Geográfica

INTRODUÇÃO

De acordo com Guimarães e Alvarez (2011), nas transformações técnico-produtivas da agricultura brasileira, iniciadas na década de 1960, a soja tem se destacado como o principal produto do agronegócio, trazendo ao país, desde 1976, a posição de segundo maior produtor mundial, sendo superado apenas pelos Estados Unidos. Em 2010, o Brasil respondeu por 26,2% da produção mundial de soja (CONAB, 2010), que correspondeu a 67,5 milhões de toneladas de soja, cultivada em uma área de 24,2 milhões de hectares (área equivalente ao território do Reino Unido) (CONAB, 2010). Em termos comerciais, a soja foi responsável por cerca de 9% das exportações brasileiras, perfazendo R$ 17,5 bilhões. Em relação ao PIB do agronegócio desse mesmo ano, a *commodity* respondeu por 5,6% de um total de R$ 821,8 bilhões, que correspondeu a uma participação de 1,25% do PIB nacional (BRASIL, 2012; CEPEA, 2012).

Em relação aos estados produtores, Guimarães e Alvarez (2011) destacam que o Paraná, de 1960 até o final da década de 1990, foi o principal estado produtor do país, tanto em área cultivada quanto em volume produzido, sendo esse estado responsável, ainda em 2010, por 21% da soja colhida no Brasil. Os autores ressaltam, entretanto, que em decorrência da expansão agrícola em direção ao Cerrado, na década de 1980, o Paraná perdeu a liderança produtiva para Mato Grosso, que responde atualmente por cerca de 27% da produção brasileira (CONAB, 2010).

A geração de informações relacionadas aos cultivos agrícolas, como área cultivada, produção e rendimento de grãos, é um dos objetivos das estimativas de safras. Aliado a essas estimativas, o conhecimento de sua distribuição no espaço geográfico se torna uma informação importante para o planejamento, a logística e a segurança alimentar, além da extrema relevância para a formação de preços (FIGUEIREDO, 2005; ASSAD et al., 2007).

A precipitação pluvial, a radiação solar global e a temperatura média são elementos agrometeorológicos limitantes e o conhecimento das ocorrências deles dentro do ciclo das culturas permite entender a importância deles em estimativas de safra (CARGNELUTTI FILHO et al., 2009). 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.

Modelos que empregam variáveis agrometeorológicas geralmente integram o acúmulo (ou a perda) de biomassa das culturas ao longo do tempo, utilizando informações de dados pontuais de estações meteorológicas de superfície. Então, faz-se necessário interpolar os dados para obtenção dos resultados, por exemplo, em escalas regionais, estaduais e outras. Posteriormente os resultados podem ser classificados ou agrupados por unidades de área e apresentados na forma de mapas (ROMANI et al., 2003).

Os Sistemas de Informação Geográfica (SIG) facilitam e contribuem no processo de análise dos dados, pois fornecem recursos para visualização, manipulação, armazenamento e processamentos de variáveis georreferenciadas. Quando utilizadas em conjunto com SIG, técnicas estatísticas para análise de dados espaciais de áreas podem ser desenvolvidas, permitindo e subsidiando a Análise Espacial de Área (ZIBORDI et al., 2006).

Uma vez que Estatística Espacial de Área (EEA) faz uso das coordenadas espaciais no processo de coleta, descrição e análise dos dados, esta técnica concentra seu interesse nos processos que ocorrem no espaço e buscam, por meio do emprego de seus métodos, descrever e analisar o comportamento desses processos. As áreas (com contagens) utilizadas na EEA representam dados agregados (polígonos), como os municípios deste estudo. A apresentação usual desses dados se faz pelo uso de mapas temáticos, com cores destacando o padrão espacial do fenômeno em estudo. A análise espacial de área busca por um modelo inferencial que incorpore explicitamente as relações espaciais constituintes deste fenômeno, objetivando identificar padrões de dependência espacial das variáveis em estudo.

Desta maneira, este método busca descrever a distribuição espacial, os padrões de associação espacial (*spatial clusters*), verificar a existência de diferentes regimes espaciais ou outras formas de instabilidade espacial, além de identificar observações atípicas. Esta situação pode também ser subsidiada por meio da análise de agrupamento (*cluster analysis*), uma técnica oferecida pela análise multivariada, que identifica grupos, com propriedades homogêneas entre os elementos amostrais, em objetos de dados multivariados.

Outra técnica para análise de agrupamentos é a teoria de conjuntos nebulosos, conhecida como teoria dos conjuntos *fuzzy*, que se mostra boa para modelar a relação entre a produtividade da soja e as variáveis agrometeorológicas, pois tem sido utilizada por se basear na caracterização de classes que não possuem limites rígidos entre si.

O objetivo geral deste trabalho foi estudar técnicas de Estatística Espacial de Área nas formas Univariada e Multivariada no estudo da produtividade da soja (t ha^{-1}) na região oeste do estado do Paraná, da safra 2000/2001 até a safra 2007/2008, associadas aos fatores agrometeorológicos: precipitação pluvial (mm), temperatura média do ar (^{o}C) e radiação solar global (W m^{-2}).

Esta tese está organizada em quatro capítulos. O primeiro capítulo apresenta um levantamento bibliográfico sobre as metodologias adotadas. No segundo, objetivou-se analisar espacialmente, para os anos-safras 2005/2006 a 2007/2008, a produtividade da soja e as variáveis agrometeorológicas, por meio dos índices de correlação e autocorrelação espacial (índices de Moran Global e Moran Local (*LISA*) uni e bivariado) e seus testes de significância e gerar modelos de regressão espacial múltipla autorregressivos (SAR) e modelos de erro espacial (CAR) entre as variáveis estudadas. O terceiro capítulo buscou

realizar uma análise de agrupamento da variabilidade espacial da produtividade da soja e de variáveis agrometeorológicas e do índice de Moran Local univariado para a produtividade da soja (LISA). Finalizando, no quarto capítulo foram classificadas, por meio da técnica *fuzzy* para agrupamentos, áreas associadas à produtividade da soja (t ha^{-1}) na região oeste do estado do Paraná, considerando as variáveis agrometeorológicas.

1 ANÁLISE ESPACIAL DA PRODUTIVIDADE DA SOJA E DOS DADOS AGROMETEOROLÓGICOS

1.1 A cultura da soja

De acordo com Klosowski (1997), devido à farta aplicabilidade de seus produtos e da facilidade de seu cultivo, a soja é extremamente importante para a humanidade, o que vem motivando sua expansão no Brasil. É cultivada há mais de cinco mil anos no Oriente, especialmente na China, região caracterizada por clima temperado. No Ocidente, tornou-se conhecida no século XX, por meio de sua exploração comercial nos Estados Unidos. Na Europa, segundo Costa (1996) e Embrapa (2006), a soja foi introduzida na metade do século XVIII pelos holandeses; entretanto, só depois de 1914 é que começou a despertar interesse nos meios agronômicos.

Na América do Sul, de acordo com Freire e Vernetti (1999), a soja foi introduzida inicialmente na Argentina (final do século XIX). No Brasil, foi cultivada por imigrantes japoneses, primeiramente no estado de São Paulo e em seguida nos estados de Minas Gerais, Paraná, Santa Catarina e Rio Grande do Sul (também pelos japoneses). Em termos de produção espacial, a cultura da soja vem ampliando sua área, com destaque para a Região Sul e a Centro-Oeste, embora as fronteiras agrícolas tenham avançado muito nas últimas décadas (YOKOO; SILVEIRA, 2006).

Yokoo e Silveira (2006) ressaltam que a demanda por informações concretas e eficientes sobre o desenvolvimento das culturas agrícolas ao longo de seus ciclos vem aumentando constantemente. Esse aumento se deve tanto por razões que implicam o aumento da produtividade como por questões de ordem econômica e ambiental, pois, ressaltamos autores, a soja tem estado entre os cultivos mais representativos na pauta do mercado externo.

No cenário paranaense, Freire e Vernetti (1999) comentam que a cultura da soja obteve destaque em meados da década de 1950, pois até então sua pequena produção era destinada ao consumo doméstico e à alimentação de suínos. Na Região Sul do estado e em pequenas áreas, afirmam os autores, era utilizada como alternativa ao lado do arroz sequeiro. A cultura da soja, até os anos 1950, não figurava como cultivo comercial para as regiões norte, noroeste, oeste e sudoeste do Paraná. O que a impulsionou foi a grande geada de 1953, que destruiu parte dos cafezais das regiões norte e noroeste do estado, o que forçou a maioria dos agricultores ao cultivo de cereais, intercalando com a cultura do café.

De acordo com a Embrapa (2005), foi na região sul do estado do Paraná, particularmente nos Campos Gerais, que houve um maior desenvolvimento da cultura da

soja. Na região sudoeste do estado, inicialmente a soja era plantada em pequenas propriedades. O oeste do Paraná, em razão da fertilidade das terras, do seu baixo preço e das condições de clima propícias, foi alvo de interesse de agricultores sulistas no final da década de 1960 (YOKOO; SILVEIRA, 2006).

Yokoo e Silveira (2006) ressaltam que a partir de 1965 foi constatado grande aumento da área cultivada com a soja no Paraná, dada a facilidade de comercialização. Até esse período, a tecnologia de cultivo vinha principalmente do Rio Grande do Sul e de São Paulo, pois no Paraná a pesquisa com a soja se restringia a alguns experimentos de variedades e de épocas de semeadura realizados pela Secretaria da Agricultura e pelo instituto de pesquisas IRI. Foi a partir da década de 1970 que essa cultura adquiriu maior expressividade, acentuando-se a partir de 1975 (KASTER et al., 1989; ALMEIDA et al., 1999).

Na região oeste do estado do Paraná, acultura da soja tem importância social e econômica pela elevada produtividade e pela extensão da área cultivada, constituindo-se em uma das principais regiões produtoras do estado (DERAL/SEAB, 2000; ROESE et al., 2001). Ayoade (1986) ressalta que o cultivo da soja ocorre nas estações de primavera/verão.

1.2 Dados agrometeorológicos

Uma observação meteorológica de superfície consiste na medição ou determinação de todos os elementos que, em seu conjunto, representem as condições meteorológicas em um dado momento e em um determinado lugar, utilizando-se de instrumental adequado e valendo-se do sentido da visão. As observações realizadas de maneira sistemática, uniforme, ininterrupta e em horas estabelecidas permitem conhecer as características e variações dos elementos atmosféricos. Esses elementos constituem os dados básicos para informar o tempo que está ocorrendo nas diferentes estações meteorológicas (INMET, 1999).

De acordo com Vianello e Alves (2001), os dados agrometeorológicos, como temperatura média do ar, precipitação pluvial e radiação solar, podem ser obtidos mediante leituras ou registros contínuos, diretamente dos instrumentos. As observações meteorológicas são realizadas em estações meteorológicas, que são locais tecnicamente escolhidos e preparados para tais fins. Em relação aos elementos do clima, os autores os definem como grandezas meteorológicas que comunicam ao meio atmosférico suas propriedades e características peculiares. Os principais elementos, pertinentes a este estudo, são: temperatura do ar, precipitação pluvial e radiação solar.

Segundo Carmo Neto *et al.* (2011), a agricultura é a atividade econômica que apresenta maior dependência das condições climáticas, as quais são consideradas como

um dos principais fatores responsáveis pelas oscilações nas produções das culturas. As variáveis climáticas influenciam todo o ciclo fenológico das plantas e também determinam a produção e a produtividade das culturas.

Yokoo e Silveira (2006) ressaltam que se torna cada vez mais fácil identificar, dentro do ano e das regiões, por meio das previsões agrometeorológicas e com o apoio de análises de séries históricas de dados agrometeorológicos, quais épocas são mais adequadas para o cultivo de cada cultivar. Conforme Mota (2002), a previsão agrometeorológica trata da avaliação do estado presente e futuro das culturas, inclusive das datas do desenvolvimento e da produtividade da colheita (quantidade e qualidade), assim como outros fatores que afetam a produção, como a densidade da semeadura e a escolha das áreas a serem plantadas. O autor ainda ressalta que esse processo é diferente das previsões meteorológicas para a agricultura, pois esta trata das previsões dos elementos meteorológicos que afetam as atividades agrícolas, como, por exemplo, previsões para fumigação e para estimar a probabilidade de ocorrência de condições potencialmente perigosas (geada, incêndio, granizo e chuva forte).

1. 2. 1 Temperatura do ar

Silva (2008) afirma que, como a maioria dos gases, o ar não é um bom condutor de calor e tarda muito a alcançar o equilíbrio térmico com os demais corpos com os quais se acha em contato. A autora ressalta que nas camadas de ar adjacentes ao solo é que se verificam as variações mais rápidas dos valores de temperatura do ar e a partir de determinada altitude (correspondente à superfície de 850 hPa) é que se verifica um decréscimo mais ou menos regular. É lembrado ainda pela autora que a distribuição de temperatura no planeta é influenciada por diversos fatores, tais como a latitude, a distribuição dos continentes e mares, as correntes marítimas, os ventos predominantes e pela ação das massas de ar.

De acordo com Silva (2008), a temperatura do ar é controlada principalmente pela radiação solar e sua distribuição depende da latitude. A autora ressalta que cidades na mesma latitude estão à mesma distância do equador e tendem a ter a mesma temperatura, mas a temperatura depende de outros fatores também, como a altitude. A oscilação diurna da temperatura, conforme afirmação da mesma autora, varia notavelmente de amplitude segundo as condições locais e a época do ano, de tal maneira que se considera a referida amplitude como um dos índices climatológicos mais significativos.

Segundo Caramori (2003), entre os elementos meteorológicos que mais afetam a produtividade agrícola no mundo destacam-se a temperatura e a precipitação. A temperatura, de acordo com o autor, é de tal maneira limitante aos cultivos que a distribuição geográfica das espécies vegetais no globo está confinada aos limites térmicos

tolerados por cada espécie ou variedade. Por outro lado, afirma o autor, a disponibilidade hídrica é o fator que mais causa frustrações de safra em todo o mundo.

1. 2. 2 Precipitação pluvial

A precipitação é definida como conjunto de partículas líquidas ou sólidas que caem das nuvens (chuva, chuvisco, granizo e neve), conjunto de partículas em suspensão na atmosfera (nevoeiro e bruma) e como partículas que se depositam (geada e orvalho) (INMET, 1999). Para as condições climáticas do Brasil, a chuva é a precipitação mais significativa em termos de volume. É o elemento alimentador da fase terrestre do ciclo hidrológico e constitui, portanto, fator importante para os processos de escoamento superficial direto, infiltração, evaporação, evapotranspiração, recarga de aquíferos e vazão dos rios. A precipitação sempre equilibra a evaporação em termos globais, a fim de manter harmônico o equilíbrio hidrológico (SILVA, 2008).

A chuva, por sua grande variabilidade em termos espacial e temporal, constitui-se em um dos elementos climáticos de maior importância para a agricultura, por sua grande influência em todos os estágios do desenvolvimento das plantas (VILHENA et al., 2009). O excesso ou a deficiência hídrica em determinados subperíodos de desenvolvimento dos cultivos agrícolas pode acarretar prejuízos, em termos de produtividade e de economia, sendo, portanto, de grande importância os estudos voltados para a avaliação da influência dos regimes pluviométricos na produção agrícola (SILVA et al., 2011).

1. 2. 3 Radiação Solar Global

A radiação solar incidente no topo da atmosfera terrestre varia basicamente com a latitude e o tempo, a qual, ao atravessar a atmosfera, interage com seus constituintes. Parte dessa radiação que é espalhada em outras direções é específica da radiação solar difusa; a outra parte, que chega diretamente à superfície do solo, é denominada de radiação solar direta. Somando a radiação difusa com a direta obtém-se a radiação solar global (SILVA et al., 2009).

Ayoade (1986) considera que a radiação solar é a energia que aciona o sistema agrícola, determinando as características térmicas do ambiente, especialmente as temperaturas do ar e do solo. Determina também a duração do dia, ou seja, o fotoperiodismo (resposta dos vegetais à luminosidade). O autor ainda complementa que, se não houver radiação suficiente, o sistema radicular da planta não se desenvolve completamente.

1.3 Geoprocessamento

A obtenção de informações sobre a distribuição geográfica de fenômenos e objetos é parte importante das atividades de organização da sociedade. Antes contidas em mapas e documentos em papel impresso, o desenvolvimento da Informática, na segunda metade do século XX, possibilitou armazenar e representar tais informações em ambiente computacional, culminando no advento da prática do Geoprocessamento, tido como: "[...] um ramo do processamento de dados que opera transformações nos dados contidos em uma base de dados referenciada territorialmente (geocodificada), usando recursos analíticos, gráficos e lógicos, para a obtenção e apresentação das transformações desejadas" (SILVA, 1992).

Reúnem-se *hardware*, *software*, base de dados, metodologias e operador, que analogicamente correspondem às ferramentas materiais e virtuais de trabalho, à matéria-prima, às técnicas do ofício e ao trabalhador. Com os componentes técnicos de suporte material (*hardware*) e os programas de manipulação de dados no suporte lógico (*software*), trabalhar com Geoprocessamento significa utilizar computadores como instrumentos de manuseio de dados para representação digital do espaço geográfico (DOMINGUES; FRANÇOSO, 2008). O conjunto de dados cujo significado contém associações ou relações de natureza espacial formam uma informação geográfica (GONÇALVES, 2008), dispostas em planilhas alfanuméricas, matrizes e representações gráficas vetoriais. Para que essas informações sejam submetidas ao processamento computacional, a cada tipo de informação é associado um valor numa escala de medida ou referência, o que insere a representação dos fenômenos geográficos na lógica dos sistemas de informação (MATIAS, 2002).

Várias são as Ciências que se beneficiam de seus resultados, como a Agronomia e o Urbanismo. Transpondo limites científicos disciplinares por meio dos trabalhos de localização dos fenômenos e equacionamento e esclarecimento das condições espaciais, o Geoprocessamento é:

> [...] uma tecnologia transdisciplinar que, através da axiomática da localização e do processamento de dados geográficos, integra várias disciplinas, equipamentos, programas, processos, entidades, dados, metodologias e pessoas para coleta, tratamento, análise e apresentação de informações associadas a mapas digitais georreferenciados. (ROCHA, 2002, p.210)

1. 3. 1 Sistemas de Informações Geográficas (SIG)

O termo Sistemas de Informação Geográfica (SIG) é aplicado em sistemas que realizam o tratamento computacional de dados geográficos e armazenam a geometria e os atributos dos dados que estão georreferenciados, isto é, localizados na superfície terrestre e representados numa projeção cartográfica (DRUCK et al., 2004). A principal diferença de um

SIG para um sistema de informação convencional é sua capacidade de armazenar tanto os atributos descritivos como as geometrias dos diferentes tipos de dados geográficos (NALON et al., 2011).

Desde sua concepção inicial, mais simplista e voltada para o projeto e para a construção de mapas, os SIG têm incorporado uma crescente variedade de funções. Em especial, apresentam mecanismos sofisticados para manipulação e análise espacial de dados, permitindo uma visualização bem mais intuitiva dos dados do que a obtida por meio de relatórios e gráficos convencionais (SILVA et al., 2009).

Pinheiro e Silva (2009) dividem a evolução dos SIG em três fases: manipulação e visualização de banco de dados (primeira fase), operações analíticas de dados não-gráficos e estrutura organizacionais (segunda fase) e análise espacial (terceira fase).

Iniciada na década de 1950, a primeira fase é marcada pela necessidade de armazenar, organizar, processar e visualizar dados, originando os SIG baseados na manipulação e visualização de dados. Na segunda fase, o aumento da capacidade de processamento e de memória dos computadores possibilitou novas concepções e a popularização dos SIG, conforme Teixeira *et al.* (1995). Nesta fase, as operações analíticas são enfatizadas por meio de modelos matemáticos. A terceira fase, década de 1980, foi marcada pela redução de recursos para a pesquisa científica enquanto havia um crescimento do setor industrial e comercial dos SIG. Nesta fase, o potencial dos SIG foi mais explorado, combinando atributos não-geográficos com as relações topológicas dos objetos geográficos para efetuar análises espaciais sobre dados georreferenciados (PINHEIRO;SILVA, 2009). Os SIG também podem ser considerados como um tipo de Sistema de Informação, que envolve de maneira sistêmica e interativa um Banco de Dados, Tecnologia e Pessoal, sendo capaz de realizar Análises Espaciais, armazenar, manipular, visualizar e operar dados georreferenciados para a obtenção de novas informações.

Conforme Miranda (2005), a abordagem mais adequada para a definição de SIG é a que enfatiza a importância da análise espacial e da modelagem que pode ser realizada, na qual o SIG é visto mais como uma ciência de informação espacial do que uma tecnologia. As definições de SIG refletem, cada uma à sua maneira, a multiplicidade de usos e visões possíveis dessa tecnologia e apontam para uma perspectiva interdisciplinar de sua utilização. A partir desses conceitos, é possível indicar duas importantes características de SIG. Primeiro, tais sistemas possibilitam a integração, em uma única base de dados, de informações geográficas provenientes de fontes diversas, como dados cartográficos, dados de censo e cadastro urbano e rural, imagens de satélite e modelos numéricos de terreno. Segundo, SIG oferecem mecanismos para recuperar, manipular e visualizar esses dados, por meio de algoritmos de manipulação e análise (CAMARA et al., 1996).

Desta maneira, o SIG pode ser entendido como ferramenta computacional para o geoprocessamento, que permite realizar análises complexas, ao integrar dados de diversas

fontes (fotos aéreas, imagens de satélite, cartas topográficas, imagens vetoriais e dados cadastrais das regiões observadas) e criar bancos de dados georreferenciados (CÂMARA et al., 2001).

1. 3. 1. 1 Áreas de Aplicação do SIG

Oliveira (1997) apresenta uma relação das diversas áreas de aplicação de SIG, divididas em cinco grupos principais:

- ocupação humana;
- uso da terra;
- uso de recursos naturais;
- meio ambiente; e
- atividades econômicas.

Segundo Oliveira (1997), a noção de análise espacial num SIG baseia-se na ideia da integração de dados espaciais e de atributos alfanuméricos, traduzindo-se numa série de funções relacionadas com a seleção, a pesquisa e a modelagem de dados.

1.4 Análise espacial

A compreensão da distribuição espacial de dados oriundos de fenômenos ocorridos no espaço para a elucidação de questões centrais em diversas áreas do conhecimento seja em ambiente, em saúde, em geologia, em agronomia, entre outras, constitui um grande desafio. A ideia central é incorporar o espaço à análise que se deseja fazer (SANTOS et al., 2004).

Bailey (1994) define a análise espacial como uma ferramenta que possibilita manipular dados espaciais de diferentes formas e extrair conhecimento adicional como resposta, incluindo funções básicas como consulta de informações espaciais dentro de áreas de interesse definidas, manipulação de mapas e produção de alguns breves sumários estatísticos dessa informação, incorporando também funções como a investigação de padrões e relacionamentos dos dados na região de interesse, buscando, assim, um melhor entendimento do fenômeno e a possibilidade de se fazer predições.

Um exemplo pioneiro do uso da análise espacial de área, ao qual intuitivamente se incorporou a categoria espaço às análises dos eventos, foi realizado no século XIX por John Snow. Em 1854, ocorria em Londres uma das várias epidemias de cólera trazidas das Índias. Pouco se sabia então sobre os mecanismos causais da doença. Uns achavam que estava relacionado aos gases e odores, concentrados nas regiões baixas e pantanosas da cidade, e outros à ingestão de água insalubre. Um mapa localizava a residência dos óbitos ocasionados pela doença (representado por pontos) e as bombas de água que abasteciam

a cidade (representado por cruzes), permitindo visualizar claramente o epicentro da epidemia (Figura 1). Estudos posteriores confirmaram essa hipótese, corroborada por outras informações, tais como a localização do ponto de captação de água desta bomba à jusante da cidade em local onde a concentração de dejetos, inclusive de pacientes coléricos, era máxima. Essa é uma situação típica em que a relação espacial entre os dados contribuiu significativamente para o avanço na compreensão do fenômeno, sendo um dos primeiros exemplos da análise espacial (CÂMARA et al., 2002).

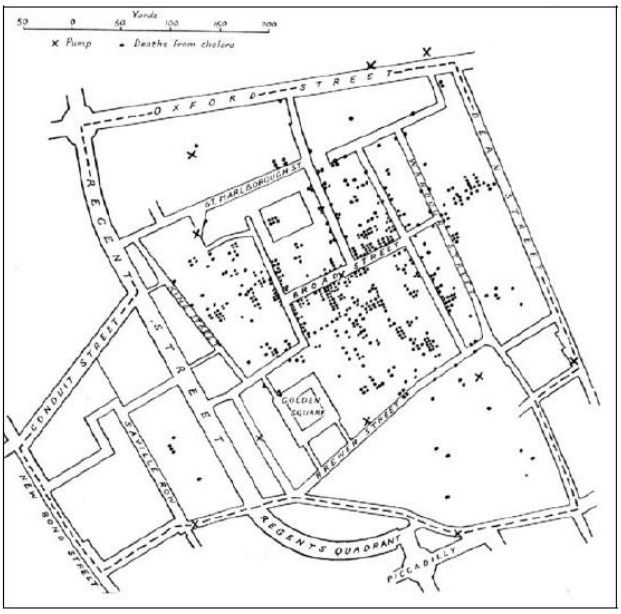


Figura 1 Mapa de John Snow (1855) mostrando os locais de ocorrência de epidemia de cólera em Londres em 1854 (CÂMARA; MONTEIRO, 2004).

Para Meneses (2003), a análise espacial apresenta duas vertentes principais: estatística espacial e geocomputação. A primeira gera modelos matemáticos de distribuição e correlação, os quais incorporam propriedades de significância e incerteza, resultantes da dimensão espacial. Já a geocomputação usa técnicas de redes neurais, busca heurística e autômatos celulares para explorar grandes bases de dados e gerar resultados empíricos (não-exatos) melhores que as técnicas convencionais, mas com ampla aplicabilidade prática. Esses instrumentos de análise espacial proporcionam maior confiabilidade aos resultados de investigações sobre a realidade modelada (CÂMARA, 2001).

Segundo Câmara *et al.* (2002), as técnicas subsidiadas pela estatística espacial de área permitem descrever a distribuição das variáveis de estudo, identificar observações atípicas não só em relação ao tipo de distribuição, mas também em relação aos vizinhos e buscar a existência de padrões na distribuição espacial. Por meio desses procedimentos, é possível estabelecer hipóteses sobre as observações, de maneira a selecionar o modelo inferencial melhor suportado pelos dados.

A localização espacial dos dados pode ser representada de forma regular ou irregular e seus índices podem ser definidos a partir de uma área no espaço, ora fixados

contáveis (LI, 2007). Os dados podem ser classificados seguindo outra denominação: dados de processos pontuais e dados de áreas. Esses dados guardam, respectivamente, forte relação com os dados ambientais e socioeconômicos. Para Cressie (1993), os dados de superfície contínua são ainda denominados, respectivamente, de *Geostatistical data* (dados contínuos no espaço) e os dados de área *Lattice data* (dados agrupados em áreas). O primeiro grupo se refere a dados contínuos, como uma amostra de uma distribuição contínua. O segundo grupo consiste em uma coleção fixa de localizações espaciais discretas (pontos ou polígonos).

1. 4. 1 Análise Exploratória de Dados Espaciais (AEDE)

A análise de dados espaciais pode ser empreendida sempre que as informações estiverem espacialmente localizadas e quando for preciso levar em conta, explicitamente, a importância do arranjo espacial dos fenômenos na análise ou na interpretação de resultados desejados (BAILEY; GATTREL, 1995).

O objetivo da análise espacial é aprofundar a compreensão do processo, avaliar evidências de hipóteses a ele relacionadas ou, ainda, tentar prever valores em áreas onde as observações não estão disponíveis (BAILEY; GATTREL, 1995). Como salientaram os autores, basicamente pode-se distinguir entre os vários métodos aqueles que:

- são essencialmente voltados à visualização dos dados espaciais;
- são exploratórios, investigando e resumindo relações e padrões mapeados; e
- contam com a especificação de um modelo estatístico e a estimação de parâmetros.

A visualização gráfica é uma etapa fundamental da análise espacial. Por meio dela é possível identificar padrões espaciais nos dados, gerando hipóteses testáveis, bem como avaliar o ajuste de modelos propostos ou, ainda, a validade das previsões resultantes. De fato, há uma série de questões que poderão justificar uma visualização criteriosa dos dados, tais como (CARDOSO et al., 2011):

- Há variáveis com valores extremos (muito altos ou muito baixos)?
- As observações se dividem em grupos distintos?
- Existem associações entre as variáveis?

As questões anteriores podem ser respondidas com o auxílio de métodos gráficos ou estatísticas descritivas. Essas técnicas são conhecidas como Análise Exploratória de Dados Espaciais, podendo ser classificadas em Univariadas ou Multivariadas, dependendo do número de variáveis envolvidas. Dentre as técnicas univariadas, destacam-se os histogramas, os mapas, as estimativas de densidade e boxplots, enquanto entre as técnicas

multivariadas poderão ser empregadas matrizes de dispersão, gráficos ligados aos mapas (*linked plots*) e gráficos de coordenadas paralelas, por exemplo (LIMA, 2010).

A análise exploratória é uma ferramenta utilizada na caracterização do arranjo espacial dos eventos. Os indicadores da análise exploratória buscam avaliar não apenas a posição absoluta dos eventos, mas também identificam a sua distribuição relativa, de maneira a buscar padrões de associações espaciais (*clusters* espaciais), regimes espaciais ou outras formas de instabilidade espacial (não-estacionaridade). A análise consiste na observação de algum tipo de padrão sistemático ou se estão distribuídos aleatoriamente no espaço (MONTENEGRO, 2008). De acordo com Messner *et al.* (1999), a AEDE é um conjunto de técnicas de análise estatística de informação geográfica.

A AEDE contribui para indicar uma apropriada modelagem econométrica espacial, ao permitir a identificação de localidades atípicas (*outliers* espaciais). A partir desse método, é possível extrair medidas de autocorrelação espacial e local, investigando a influência dos efeitos espaciais por intermédio de instrumentos quantitativos – e não somente pela inspeção visual de mapas (MONTENEGRO, 2008).

Os métodos convencionais, como regressões múltiplas, não são formas apropriadas de lidar com dados georreferenciados, visto que não são confiáveis para detectar agrupamentos e padrões espaciais significativos. Dessa maneira, a AEDE deve ser a primeira etapa para revelar padrões espaciais, que deverão anteceder quaisquer modelos espaciais (ANSELIN; BERA, 1988).

Anselin e Bera (1998) assumem que é difícil diferenciar autocorrelação espacial de heterogeneidade espacial. Os autores argumentam que em uma *cross-section* os dois problemas podem ser equivalentes do ponto de vista da observação, gerando dificuldades em determinar se o problema é ocasionado pela heterocedasticidade ou pela autocorrelação espacial. *Cross-sectional data* ou *cross section* (de uma população de estudo) em estatística e econometria é um tipo de dado unidimensional definido. *Cross-sectional data* refere-se aos dados coletados através da observação de muitos assuntos (como indivíduos, empresas ou países/regiões) no mesmo ponto do tempo ou sem levar em conta as diferenças de tempo. A análise dos dados *cross-sectional* geralmente consiste em comparar as diferenças entre os sujeitos.

1. 4. 2 Matriz de proximidade espacial

Um procedimento necessário para a análise de dados de área é a construção de uma matriz de vizinhança, também conhecida como matriz de distância, matriz de conectividade ou matriz de proximidade. Essa matriz de conectividade indica a relação espacial de cada área com as demais, podendo ser composta, por exemplo, apenas pela

lista de vizinhos de cada município; ou a distância entre municípios ligados por estradas; ou a conectividade ponderada pelo comprimento da fronteira comum (BRASIL, 2006).

A Figura 2 apresenta os municípios do estado de Roraima, a partir dos quais foi construída uma matriz baseada na vizinhança, por meio da atribuição de valores: 1 para os municípios que possuem fronteiras em comum, e 0 para os municípios que não compartilham fronteiras (BRASIL, 2006). A matriz de vizinhança referente ao mapa da Figura 2 pode ser vista na Figura 3. Observe, por exemplo, que a capital Boa Vista tem sete vizinhos e que o município de Uiramutã, no extremo norte, tem apenas dois vizinhos.

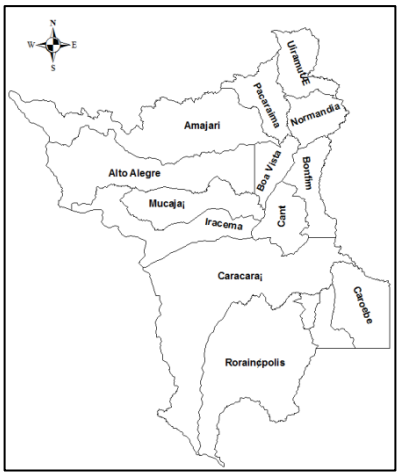


Figura 2 Mapa do estado de Roraima com divisão por municípios

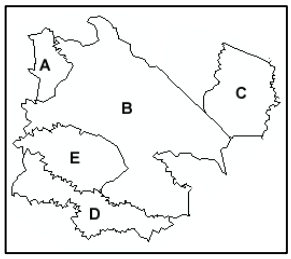
	Amajari	Alto Alegre	Boa Vista	Bonfim	Cantá	Caracaraí	Caroebe	Iracema	Mucajaí	Normandia	Pacaraima	Rorainópolis	São João da Baliza	São Luiz	Uiramutã
Amajari		1	1	0	0	0	0	0	0	0	1	0	0	0	0
Alto Alegre	1		1	0	0	0	0	1	1	0	0	0	0	0	0
Boa Vista	1	1		1	1	0	0	0	1	1	1	0	0	0	0
Bonfim	0	0	1		1	1	0	0	0	1	0	0	0	0	0
Cantá	0	0	1	1		1	0	1	1	0	0	0	0	0	0
Caracaraí	0	0	0	1	1		1	1	0	0	0	1	1	1	0
Caroebe	0	0	0	0	0	1		0	0	0	0	0	1	0	0
Iracema	0	1	0	0	1	1	0		1	0	0	0	0	0	0
Mucajaí	0	1	1	0	1	0	0	1		0	0	0	0	0	0
Normandia	0	0	1	1	0	0	0	0	0		1	0	0	0	1
Pacaraima	1	0	1	0	0	0	0	0	0	1		0	0	0	1
Rorainópolis	0	0	0	0	0	1	0	0	0	0	0		1	1	0
São João da Baliza	0	0	0	0	0	1	1	0	0	0	0	1		1	0
São Luiz	0	0	0	0	0	1	0	0	0	0	0	1	1		0
Uiramutã	0	0	0	0	0	0	0	0	0	1	1	0	0	0	

Figura 3 Matriz de vizinhança para os municípios do estado de Roraima

De acordo com Câmara *et al.* (2002), a matriz de vizinhança, também chamada de proximidade espacial *W*, é uma ferramenta básica para estimar a variabilidade espacial de dados de áreas. Como demonstrado graficamente pelas Figuras 4 e 5, dado um conjunto de *n* áreas $\{A_1,..,A_n\}$, constrói-se a matriz *W*, *n x n*, onde cada um dos elementos w_{ij} representa uma medida de proximidade entre A_i e A_j. Essa medida de proximidade pode ser calculada a partir de um dos seguintes critérios:

- w_{ij} = 1, se o centroide de A_i está a uma determinada distância de A_j; caso contrário w_{ij} = 0; para $i \neq j = 1, 2,..., n$;
- w_{ij} = 1, se A_i compartilha um lado comum com A_j, caso contrário w_{ij} = 0; para $i \neq j = 1, 2,..., n$;
- $w_{ij} = l_{ij}/l_{i,}$ sendo l_{ij}é o comprimento da fronteira entre A_i e A_j e l_i é o perímetro de $A_{i;}$ para $i \neq j = 1, 2,..., n$.

Como a matriz de proximidade é utilizada em cálculos de indicadores na fase de análise exploratória, é muito útil normalizar suas linhas, para que a soma dos pesos de cada linha seja igual a 1. Isso simplifica muito vários cálculos de índices de autocorrelação espacial. A Figura 4 ilustra um exemplo simples de matriz de proximidade espacial normalizada, em que os valores dos elementos da matriz refletem o critério de adjacência.



	A	B	C	D	E
A	0	1	0	0	0
B	0.25	0	0.25	0.25	0.25
C	0	1	0	0	0
D	0	0.5	0	0	0.5
E	0	0.5	0	0.5	0

Figura 4 Matriz de proximidade espacial de primeira ordem, normalizada pelas linhas

É importante convencionar as formas de vizinhança quando se utiliza matrizes de proximidade espacial que considerem a contiguidade. Desta maneira, os critérios baseiam-se em movimentos de peças presentes no jogo de xadrez, tais como a rainha (*Queen*), a torre (*Rook*) e o bispo (*Bishop*) (LESAGE, 1999; RODRIGUES et al., 2009). A Figura 5 apresenta esses critérios para identificar os vizinhos da área J.

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

(a)

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

(b)

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

(c)

Figura 5 Representação dos tipos de contiguidade entre áreas. (a) Contiguidade Queen (rainha), (b) Contiguidade Rook (torre) e (c) Contiguidade Bishop (bispo)

A Figura 5a demostra a adoção do critério *Queen*, em que todas as áreas que tiverem intersecção não-nula com a área *J* serão vizinhas de *J*. Na Figura 5b apresenta-se o critério *Rook*, que tem como vizinhos apenas os que tiverem um lado em comum da área *J*. Já a Figura 5c representa o critério *Bishop,* que relaciona como vizinhos da área *J* apenas as áreas que se localizam nas diagonais.

A ideia da matriz de proximidade espacial pode ser generalizada para vizinhos dos vizinhos, e isso caracteriza a ordem da matriz. Considerando o critério de vizinhança *Rook* e uma matriz de ordem 2 na Figura 5b, os vizinhos de *F*, *I*, *N* e *K* também seriam considerados vizinhos de *J*.

1. 4. 3 Vetor dos desvios e vetor de médias ponderadas

Sabe-se que a cada área *i* está associado um número real ($\boldsymbol{x}_i$), que representa o valor do atributo na área *i*. Para o cálculo do vetor de desvios *Z*, é calculada, primeiramente, a média (μ) dos valores dos atributos, considerando as *n* áreas. Cada elemento *i* de *Z*, denominado z_i, é obtido subtraindo-se o valor da média, do valor do atributo correspondente, ou seja, $z_i = x_i - \mu$, para $i = 1,..., n$. Em caso de não ter a média populacional μconsidera-se a média amostral $\bar{x}$ (ANSELIN, 1996).

O vetor de médias ponderadas (*Wz*) é obtido pela multiplicação do vetor transposto dos desvios, pela matriz de proximidade espacial com linhas normalizadas, onde cada elemento de uma linha *i* qualquer, originariamente com valor *1*, é dividido pelo número de elementos não-nulos da mesma linha. Desta maneira, como resultado, cada elemento Wz_i, contém um valor correspondente à média dos desvios dos vizinhos da área *i*, caracterizando uma média móvel espacial.

Com a definição do vetor de médias ponderadas, pode-se estabelecer outro mapa em uma análise exploratória, simplesmente calculando a média móvel espacial dos atributos estudados. Segundo Druck *et al.* (2004), o cálculo de uma média móvel espacial é uma maneira de explorar a variação da tendência espacial dos dados, pois a operação tende a produzir uma superfície com menor flutuação que os dados originais.

1. 4. 4 Dependência Espacial

Um conceito-chave na compreensão e análise dos fenômenos espaciais é a dependência espacial. Essa noção parte da primeira lei da geografia: “Todas as coisas são parecidas, mas coisas mais próximas se parecem mais que coisas mais distantes” (TOBLER, 1979). Pode-se afirmar que a maior parte das ocorrências, sejam estas naturais ou sociais, apresentam entre si uma relação que depende da distância. Esse princípio quer dizer que, se encontrarmos poluição num trecho de um lago, é provável que locais próximos a esta amostra também estejam poluídos. Ou que, se a presença de uma árvore adulta inibe o desenvolvimento de outras, esta inibição diminui com a distância e, após determinado raio, outras árvores grandes serão encontradas (CÂMARA et al., 2004).

A dependência espacial, ou autocorrelação espacial, refere-se à correlação entre o mesmo atributo em dois locais ou em dois períodos de tempo. Na ausência de dependência espacial, a proximidade das duas localidades não influencia o comportamento conjunto de atributos observados. Quando há dependência espacial (autocorrelação espacial positiva ou negativa), então as observações mais próximas são mais semelhantes do que as observações distantes (LI, 2007).

Após a obtenção de dados espaciais, a primeira questão que surge é se existe algum padrão espacial, ou seja, esses locais ou pequenas áreas que são próximas umas das outras tendem a se comportar da mesma forma que aqueles mais distantes uns dos outros? A questão pode ser colocada como um teste, onde, se houver dependência espacial nos dados, então se deseja medi-la e estima-la (CÂMARA et al., 2004).

De acordo com Anselin (1988), a dependência espacial se manifesta pela falta de independência que geralmente está presente entre as observações *cross-section*. Segundo Chasco (2003), à primeira vista, a dependência espacial pode parecer similar à mais conhecida dependência presente nos testes econométricos de correlação de séries, nos

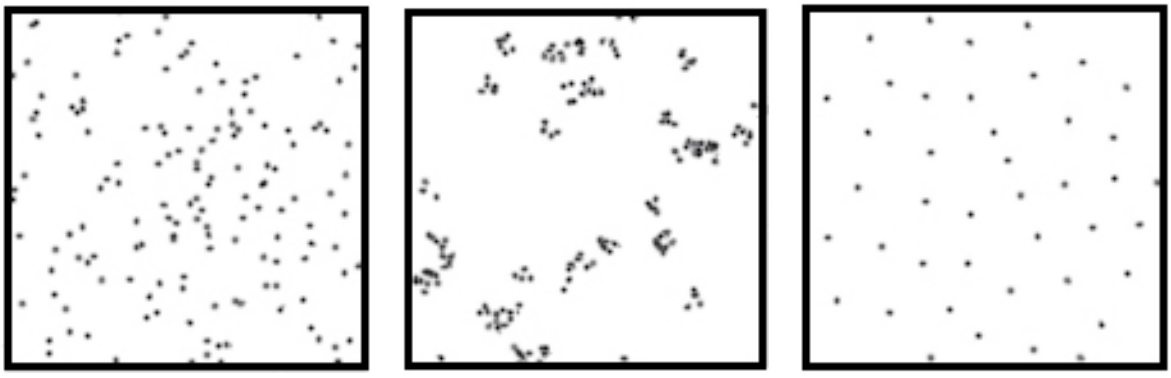
modelos de distribuição de atrasos e em outras análises de séries temporais. No entanto, essa semelhança apenas é real em parte, devido à natureza multidirecional da dependência no espaço que, frente à clara situação unidirecional do tempo, faz necessário o uso de uma estrutura metodológica diferente.

Segundo Lesage (1998), a presença de dependência espacial significa que uma observação está associada a uma localização *i* depende de observações nas localizações *j*, sendo que *i* ≠ *j*. $f = (x_j) i = 1, \ldots, n,\ i \neq j$. Anselin (1988) baseou a dependência espacial por meio da noção de contiguidade binária entre as unidades espaciais, ou seja, a estrutura dos vizinhos era expressa a partir dos valores binários, 0 e 1. Dessa maneira, se duas unidades espaciais têm uma fronteira comum, então tais unidades são consideradas contíguas e recebem o valor um. Contrariamente, as unidades não-vizinhas recebem o valor zero para classificá-las. Isso implica afirmar que, conforme Lesage (1998), observações que estão mais próximas uma das outras devem refletir um maior grau de dependência espacial do que as mais distantes. Consequentemente, o poder da dependência espacial entre as observações deve declinar com a distância entre elas. “Quanto à dependência espacial, as unidades vizinhas devem apresentar um maior grau de dependência espacial do que as unidades localizadas distantes”. (LESAGE, 1998).

Dados espaciais não formam um conjunto de amostras independentes. Uma importante diferença em relação a dados sem essa peculiaridade é que cada observação não traz uma informação independente, e o conjunto de todas as observações é utilizado integralmente para descrever o padrão do fenômeno estudado (BRASIL, 2006).

Duas questões estatísticas devem levar em consideração a presença de dependência espacial: a **identificação de padrões espaciais** e a **análise do efeito** de algum fator de risco sobre um desfecho para a localização geográfica. Para a identificação de padrões espaciais, é preciso estimar a presença, forma e intensidade da dependência espacial (BRASIL, 2006).

Espacialmente aleatória é qualquer ocorrência cuja distribuição espacial não apresente qualquer padrão espacial detectável. Na Figura 6a os pontos estão distribuídos aleatoriamente, na Figura 6b estão aglomerados (clusterizados) e na Figura 6c distribuídos de forma regular, ou seja, não-aleatória (BRASIL, 2006).



(a) (b) (c)

Figura 6 Padrões de distribuição espacial de pontos

1. 4. 5 Estatística Espacial de Áreas

A necessidade de quantificação da dependência espacial presente em um conjunto de geodados levou ao desenvolvimento da chamada estatística espacial. Segundo Anselin (1992), "[...] a característica que distingue a análise estatística dos dados espaciais é que seu foco principal está em inquirir padrões espaciais de lugares e valores, a associação espacial entre eles e a variação sistemática do fenômeno por localização".

As técnicas de estatística espacial distinguem-se das demais técnicas empregadas em análise estatística por considerar explicitamente as coordenadas dos dados no processo de coleta, descrição ou análise dos dados. Utiliza-se o termo autocorrelação espacial para diferenciar da correlação da estatística convencional, tendo em vista que nessa a correlação é obtida a partir de duas variáveis diferentes, sem referência a sua posição no espaço; no caso da autocorrelação, empregam-se no cálculo os valores de uma mesma variável em duas posições diferentes (ROCHA, 2004).

A análise de dados espaciais de área está associada a métodos utilizados quando a localização está associada a áreas delimitadas por polígonos, o que ocorre com muita frequência quando se lida com eventos agregados por municípios, bairros ou setores censitários, em que não se dispõe da localização exata dos eventos, mas de um único valor por área (DRUCK et al., 2004). Esses métodos podem ser divididos entre: métodos que estão relacionados à visualização dos dados, métodos chamados exploratórios e aqueles centralizados na especificação do modelo estatístico e na estimativa de parâmetros (AVELAR, 2008).

Para uma análise da distribuição espacial, levando em conta a localização das amostras, é necessário aplicar técnicas da estatística espacial para analisar dados que podem ser classificados em eventos de padrões espaciais, superfícies contínuas ou áreas com contagens (CRESSIE, 1993). Uma das técnicas para a área de contagens é o índice global de Moran (GETIS; ORD, 1992; BAILEY;GATRELL, 1995).

Com base na coleta sistemática de informações quantitativas, os objetivos da estatística espacial são: descrição cuidadosa e precisa de eventos no espaço geográfico (incluindo a descrição de padrões); exploração sistemática do padrão de eventos e de sua associação no espaço, com o objetivo de ganhar o melhor entendimento dos processos que podem ser responsáveis pela distribuição observada, e melhora da habilidade de predizer e controlar eventos que possam ocorrer nos espaços geográficos (AVELAR, 2008). Assunção *et al.* (2001a) afirmam que a característica fundamental da estatística espacial, que se diferencia da estatística clássica, é o uso explícito da referência geográfica no modelo, isto é, o uso explícito das coordenadas espaciais no processo de coleta, descrição e análise dos dados. Assim, o interesse está centrado nos processos que ocorrem no espaço e os métodos empregados buscam descrever e analisar o comportamento desses processos.

Essa característica faz com que estudos sobre o assunto exibam comportamento complexo, para serem analisados por métodos tradicionais de estatística (ASSUNÇÃO, 2001).

Uma vez que a estatística espacial de área faz uso da referência geográfica no modelo, isto é, das coordenadas espaciais no processo de coleta, descrição e análise dos dados, seu interesse está centrado nos processos que ocorrem no espaço e os métodos empregados buscam descrever e analisar o comportamento desses processos (ASSUNÇÃO, 2001b).

A forma usual de apresentação de dados agregados por áreas é o uso de mapas de diferentes tonalidades de cores com o padrão espacial do fenômeno (CÂMARA et al., 2002). A Figura 7 exibe a distribuição espacial da produtividade da soja para 48 municípios da região oeste do estado do Paraná, para os dados da safra 2001/2002. Verifica-se que 91,66% dos 48 municípios estudados, para os dados da safra de 2001/2002 estão acima da média estadual (2.766 kg ha^{-1}), segundo os dados da Conab (2010). Para a média nacional (2.407 kg ha^{-1}), verifica-se que 100% dos municípios estudados tiveram uma produtividade maior.

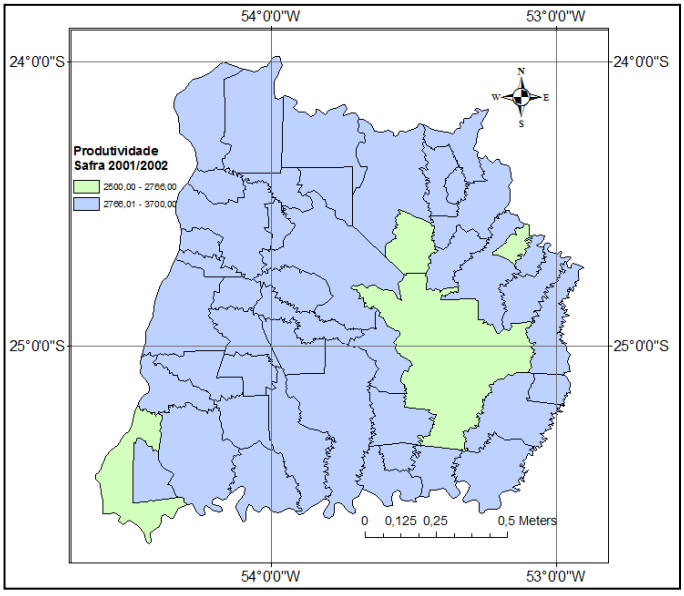


Figura 7 Valores de Produtividade para os dados da safra de 2001/2002, em 48 municípios da região oeste do estado do Paraná agrupados pela média estadual.

1. 4. 6 Análise de variáveis espaciais de áreas

Uma das técnicas mais utilizadas no estudo de fenômenos de áreas é a Análise de Autocorrelação Espacial (CARVALHO, 1997). Essa técnica permite identificar a estrutura de correlação espacial que melhor descreve o padrão de distribuição dos dados. A ideia básica é estimar a magnitude da Autocorrelação Espacial entre as áreas, evidenciando como os valores estão correlacionados no espaço (ANSELIN, 2002).

Neste caso, as técnicas são utilizadas para estimar quanto do valor observado de um atributo numa região é dependente dos valores dessa mesma variável nas localizações

vizinhas. Enquadram-se nessa categoria o Índice Global de Moran (TEXEIRA; BERTELLA, 2010).

Os indicadores globais de autocorrelação espacial, como o Índice de Moran, fornecem um único valor como medida da associação espacial para todo o conjunto de dados, o que é útil na caracterização da região de estudo como um todo. No entanto, quando se lida com um grande número de áreas, é muito provável que ocorram diferentes regimes de associação espacial e que apareçam locais em que a dependência espacial é ainda mais pronunciada (CÂMARA et al., 2002).

Por meio das “análises locais” ou “modelagens locais”, busca-se testar a presença de diferenças espaciais ao invés de assumir que estas não existem. Essas análises desagregam as estatísticas globais segundo seus constituintes locais, concentrando-se mais nas exceções locais do que na busca por regularidades globais (FOTHERINGHAM et al., 2000).

Entre as técnicas univariadas aplicadas à análise local existem as abordagens gráficas e aquelas voltadas para o desenvolvimento formal de estatísticas univariadas locais. Entre as abordagens gráficas busca-se, prioritariamente, identificar exceções locais às tendências gerais na distribuição dos dados e nas relações entre variáveis. Trabalha-se, neste sentido, com o auxílio de histogramas, gráficos de dispersão e gráficos em três dimensões (MELO;HEPP, 2008).

As técnicas gráficas mais complexas para demonstrar relações locais em bancos de dados univariados incluem o *Spatial Lagged Scatterplot*, o *Variogram Cloud Plot* e o *Moran Scatterplot*. Destaca-se, nesta lista, o *Moran Scatterplot*, que, além de permitir a identificação de grupos de valores, também permite a identificação de valores extremos na distribuição e apresenta uma visualização do nível de autocorrelação espacial existente (SALAME, 2008).

1.5 Autocorrelação espacial

A estrutura de dependência entre os valores observados nas várias áreas do fenômeno em estudo é analisada pela função de autocorrelação espacial. Autocorrelação, como o próprio nome indica, mede a correlação da própria variável, e, sendo espacial, no espaço. A correlação de uma variável com ela mesma, medida no mesmo local, será sempre 1 (UM). Entretanto, a correlação de uma variável com ela mesma, porém medida nas áreas vizinhas, terá um valor que varia entre -1 e 1 (como qualquer medida de correlação). Quanto mais próximo de 1 (UM), maior a semelhança entre vizinhos. O valor 0 (zero) indica inexistência de correlação, e valores negativos indicam dessemelhança (CÂMARA et al., 2002).

1. 5. 1 Autocorrelação espacial global univariada

A autocorrelação espacial global pode ser calculada por meio da estatística *I* de Moran (BAILEY; GATRELL, 1995), a qual permite analisar se os dados são aleatoriamente distribuídos no espaço, ou seja, se a variável sob análise está autocorrelacionada espacialmente.

Segundo Cressie (1993), formalmente, a estatística *I* de Moran pode ser expressa pela equação (1).

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j w_{ij}}{S_0 \sum_{i=1}^{n} z_i^2} \qquad \text{Eq.(1)}$$

em que *n* é o tamanho da amostra; $z_i = (x_i - \bar{x})$ e $z_j = (x_j - \bar{x})$ são as variáveis das populações *i* e *j* centradas na média; w_{ij} é o elemento da matriz quadrada e simétrica *W*, *n x n*, a qual expressa a relação espacial entre as *n* populações, e S_0 é o somatório dos elementos w_{ij} da matriz simétrica de pesos espaciais *W*.

Tendo como exemplo a produtividade de soja, a indicação de autocorrelação espacial positiva revela que há similaridade entre os municípios, ou seja, municípios com alta produtividade tendem a estarem rodeados por municípios vizinhos que também apresentam alta produtividade ou municípios com baixa produtividade rodeados por vizinhos que possuem baixa produtividade. Por outro lado, a autocorrelação espacial negativa indica que existe uma dissimilaridade entre os valores do atributo estudado e da localização espacial deste atributo. Assim, nesse exemplo, municípios com baixa produtividade estão rodeados por municípios que apresentam alta produtividade ou municípios com alta produtividade rodeados por vizinhos que apresentam baixos valores desta variável de interesse.

De acordo com Jing e Cai (2009), um aspecto interessante na estatística *I* de Moran é que é possível a visualização como sendo uma inclinação em um gráfico de dispersão da variável espacialmente defasada (*Wx*) sobre a variável original (x), ou o chamado *Moran Scatter plot*. Isso fornece uma maneira fácil de categorizar a natureza da autocorrelação espacial em quatro tipos, correspondentes aos *clusters* espaciais e *outliers* espaciais.

A estimativa da significância do índice de Moran, de acordo com Kampel *et al.* (2000) e Câmara *et al.* (2002), pode ser abordada de duas maneiras: a primeira associa o índice a uma distribuição estatística, onde geralmente considera a variável como sendo uma distribuição normal padrão, cuja significância é obtida por comparação direta do valor de Z com o valor da probabilidade tabelada; a segunda abordagem é um teste de pseudossignificância que gera diferentes permutações dos valores de atributos associados às zonas, onde cada permutação produz um novo arranjo espacial dos valores

redistribuídos entre as áreas, sendo a sua significância obtida a partir de uma distribuição empírica *I* de Moran. Se o valor do índice *I* de Moran medido corresponder a um “extremo” da distribuição simulada, então se trata de um evento com significância estatística.

1. 5. 2 Autocorrelação espacial global multivariada

A autocorrelação espacial global verifica a existência de um padrão de associação espacial entre duas variáveis. O objetivo é revelar se os valores da variável observada em uma dada região guardam uma relação com os valores de outra variável observada em regiões vizinhas. Isso significa que a estatística I de Moran (Equação 2) pode ser calculada para duas variáveis em estudo.

$$I^{yx} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})w_{ij}(y_j - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{Eq. (2)}$$

A interpretação para o *I* de Moran multivariado pode ser descrita da mesma maneira que a estatística *I* de Moran univariada, caso o valor do I^{yx} de Moran multivariado for positivo, municípios que apresentam valores elevados (*y*) estão rodeados por municípios vizinhos que apresentam nível (*x*) alto. De outra forma, municípios com baixos valores de (*y*) são vizinhos de outros com baixo nível de (*x*) (FERRARIO et al., 2009).

Para uma análise multivariada, com mais de duas variáveis em estudo, faz-se uso da Matriz de Diagramas de Dispersão de Moran (*Moran’s Scatterplot Matrix*). Nesta matriz, os eixos inferiores são as variáveis em estudo (todas normalizadas), nos eixos verticais estão as variáveis espacialmente defasadas (com os *lags* espaciais aplicados às variáveis normalizadas). Essa ferramenta permite uma visão do padrão espacial de cada variável com ela própria, bem como a defasagem espacial com as outras variáveis (ANSELIN et al., 2004). A Figura 8 apresenta uma matriz de diagramas de dispersão de Moran.

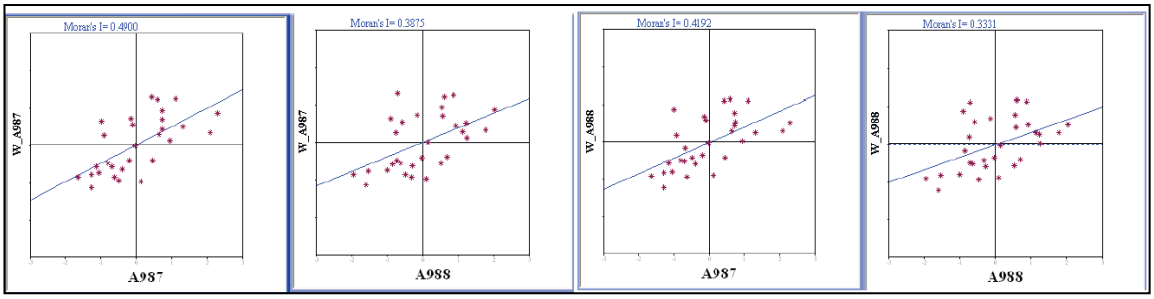


Figura 8 Matriz de Diagramas de Dispersão de Moran apresentado por Anselin *et al.* (2004)

1. 5. 3 Autocorrelação espacial local

O objetivo da autocorrelação espacial é captar padrões de associação local (*clusters* ou *outliers* espaciais). Embora seja capaz de apontar a tendência geral de agrupamento dos dados, o *I* de Moran é uma medida global e por isso não revela padrões locais de associação espacial, quer dizer, são geralmente ocultados pelas estatísticas de autocorrelação global.

O*I* de Moran pode não identificar *clusters* locais importantes em uma região global, quer sejam *clusters* positivos ou *clusters* negativos. A autocorrelação local pode ser calculada pela estatística *I* de Moran local, também conhecido como *Local Indicator of Spatial Association* (LISA) (ANSELIN, 1995).

1. 5. 4 Indicadores Locais de Associação Espacial (LISA) Univariado

Segundo Anselin (1995), um *Local Indicator of Spatial Association* (LISA) será qualquer estatística que satisfaça a dois critérios:

a) um indicador LISA deve possuir para cada observação, uma indicação de *clusters* espaciais significantes de valores similares em torno da observação (região, por exemplo);
b) o somatório dos LISAs para todas as regiões é proporcional ao indicador de autocorrelação espacial global.

Segundo Le Gallo e Erthur (2003), a estatística LISA, baseada no *I* de Moran local para a variável *x*, no período *t*, $X_t=(x_1,..x_n)^t$, pode ser especificada da seguinte forma:

$$I_{i,t} = \frac{x_{i,t} - \mu_t}{\sigma_0^2} \sum_{j=1}^{n} w_{ij}(x_{j,t} - \mu_t) \qquad \text{Eq. (3)}$$

sendo σ_0^2 a variância dos dados populacionais, com

$$\sigma_0^2 = \frac{\sum_{i=1}^{n} (x_{i,t} - \mu_t)^2}{n} \qquad \text{Eq. (4)}$$

Na qual $x_{i,t}$ é a observação de uma variável de interesse na região *i* para o período (ano por exemplo) *t* (ou espaço *t*), μ é a média das observações entre as regiões no período *t* para a qual o somatório em relação a *j* é tal que somente os valores vizinhos diretos de *j* são incluídos no cálculo da estatística.

A estatística pode ser interpretada da seguinte maneira:

- valores positivos de $I_{i,t}$ 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.

De acordo com Anselin (1995), a estatística LISA é utilizada também para medir a hipótese de ausência de associação espacial local. É importante salientar que, assim como a distribuição para as estatísticas globais, a distribuição genérica para a estatística LISA também é de difícil apuração. Portanto, para solucionar tal problema, deve-se trabalhar com resultados assintóticos. Logo, a alternativa é a utilização de uma aleatorização que permita auferir pseudoníveis de significância.

Rusche (2009) ressalta que a estatística local de Moran pode ser utilizada para uma avaliação inicial da estrutura local dos regimes espaciais, uma vez que o *I* de Moran pode ser calculado com a média aritmética dos valores de Moran local para todas as observações.

1. 5. 5 Indicadores Locais de Associação Espacial (LISA) Multivariado

A autocorrelação espacial local multivariada (LISA M) é representada pela seguinte fórmula:

$$I_i = \frac{(x_i - \bar{x})\sum_{j=1}^{n} w_{ij}(y_{ij} - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 / n} \qquad \text{Eq. (5)}$$

Em que x_i e y_j são variáveis em estudo cujo somatório sobre j é tal que somente os valores dos vizinhos $j \in J_i$ são incluídos. Comparando-se à fórmula de cálculo do I_i de Moran, o conjunto J_i abrange os vizinhos do município *i*, definidos conforme a matriz de pesos espaciais escolhida.

A interpretação dessa estatística (LISA M), segundo Anselin *et al.* (2004), representa uma indicação do grau de associação linear (positiva ou negativa) entre o valor de uma determinada variável em um dado local (município por exemplo) *i* e a média de uma outra variável nos locais vizinhos.

1. 5. 6 Análise Gráfica da Autocorrelação Espacial

Segundo Almeida *et al.* (2005), o diagrama de dispersão de Moran é uma representação do coeficiente de regressão linear por Mínimos Quadrados Ordinários (MQO), mediante um gráfico de duas variáveis *z* e *Wz*, na qual o coeficiente da inclinação da curva de regressão é dado pela estatística *I* de Moran. A inclinação da curva é obtida pela regressão de *Wz* contra *z*, e essa inclinação fornece o grau de ajustamento.

O diagrama de dispersão de Moran (Figura 9), que é a forma de visualizar o indicador global de autocorrelação espacial, mostra a defasagem espacial da variável de

interesse (ou seja, a média do atributo nos vizinhos) no eixo vertical e o valor da variável de interesse no eixo horizontal. Além da medida global de associação linear espacial, esse diagrama está dividido em quadrantes: Alto-Alto (AA), Baixo-Baixo (BB), Alto- Baixo (AB) e Baixo-Alto (BA) (ANSELIN, 1996).

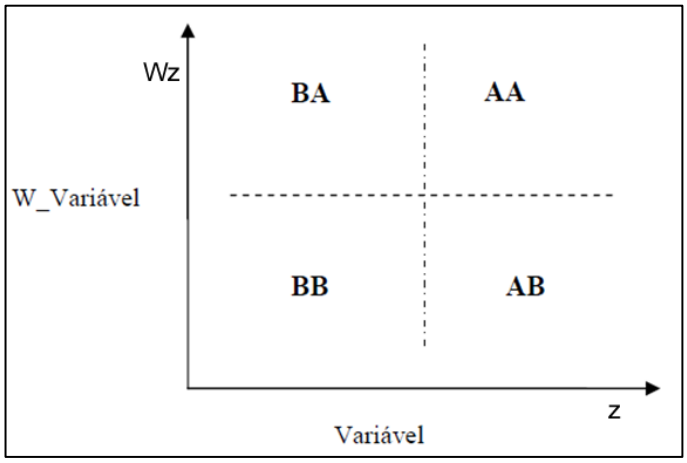


Figura 9 Estrutura do diagrama de dispersão de Moran onde W_Variável caracteriza a variável de interesse defasada espacialmente

As áreas em estudo (como municípios) localizadas nos quadrantes Alto-Alto (AA) e Baixo-Baixo (BB) significam localidades com valores altos (acima da média) e/ou baixos da variável de interesse, rodeadas por áreas que apresentam valores também altos e/ou baixos. Já as áreas situadas no quadrante Baixo-Alto (BA) e Alto-Baixo (AB) representam um grupo que está circundado por regiões com alto valor e/ou baixos valores da variável de interesse (como pluviosidade, por exemplo).

Segundo Anselin *et al.* (2004), o Diagrama de Dispersão de Moran é um *scatterplot* especializado, com a transformação espacialmente defasada de uma variável sobre o eixo y e a variável original no eixo x, após a padronização da variável de tal forma que a média é zero e a variância, um.

Para melhor compreensão do diagrama de espalhamento de Moran, é conveniente apresentá-lo associado com um mapa temático bidimensional, no qual cada polígono da região estudada é apresentado segundo seu quadrante no diagrama de espalhamento de Moran. Esse mapa é conhecido como *Box Map*, e a estrutura de sua legenda é indicada na Tabela 1 e um exemplo na Figura 10a.

Tabela 1 Mapa Box Map.

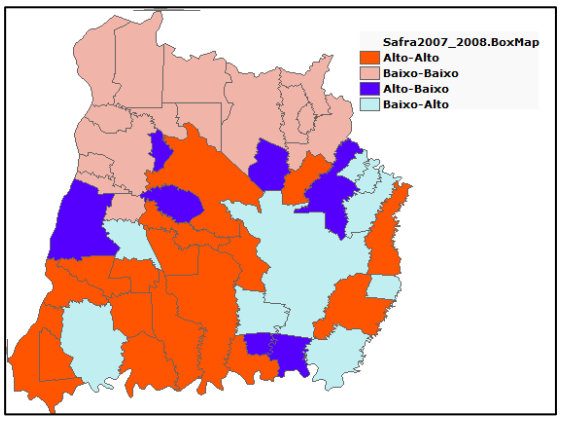
Cores	Quadrantes	Valor da área/Valor da média local
Cor 1 – escura	AA	Alto- Alto
Cor 1 – clara	BB	Baixo-Baixo
Cor 2 – escura	AB	Alto-Baixo
Cor 2 – clara	BA	Baixo-Alto

A autocorrelação do índice local de Moran é calculada a partir do produto dos desvios em relação à média, como uma medida de covariância. Dessa maneira, valores significativamente altos indicam altas probabilidades de que haja locais de associação

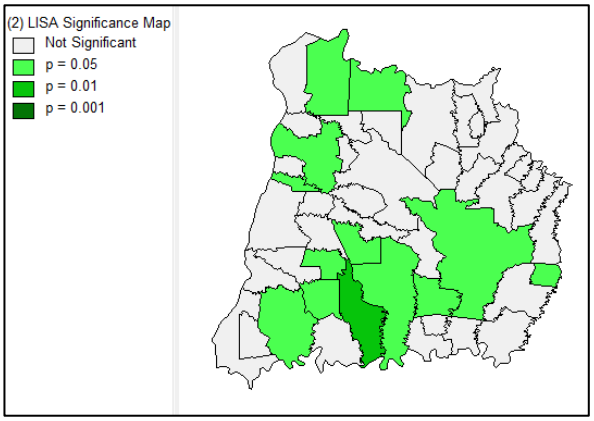
espacial, tanto de polígonos com altos valores associados, como com baixos valores associados. Por outro lado, baixos valores apontam para um padrão que pode ser entendido como locais de comportamento mais errático da variável observada entre um polígono e seus vizinhos (QUEIROZ, 2003).

Uma vez determinada a significância estatística desse índice, é útil elaborar um mapa indicando as regiões que apresentam correlação local significativamente diferente do resto dos dados. Essas regiões podem ser vistas como bolsões de homogeneidade, no caso regiões de concentração de valores elevados dos atributos e das regiões com valores reduzidos dos atributos, separadas por uma região de transição que não indica uma coisa nem outra. Essas áreas possuem dinâmica espacial própria e merecem análise detalhada. Esse mapa é chamado de Lisa Map (Figura 10b) e, na sua geração, os valores do índice local de Moran são classificados em quatro grupos: não-significantes, com significância de 95, 99 e 99,9%.

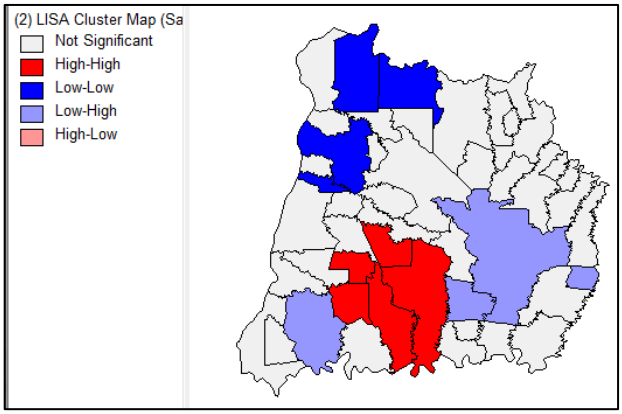
Outro tipo de mapa que também pode ser utilizado para identificação de padrões de associação espacial é o *Moran Map* (Figura 10c). Esse mapa classifica somente os objetos para os quais os valores do índice local de Moran foram considerados significantes, sendo destacados conforme sua localização no quadrante do gráfico do Diagrama de Espalhamento de Moran, ficando os demais objetos classificados como “sem significância estatística” (QUEIROZ, 2003).



(a)Box Map



(b)Lisa Map



(c)Moran Map

Figura 10 Mapas para uma análise gráfica da autocorrelação espacial

Por meio da análise gráfica de autocorrelação espacial, é possível realizar outra análise, a dos *outliers* globais e locais. *Outliers*, ou pontos discrepantes, podem ser definidos como observações que não seguem o mesmo padrão que a maioria dos dados e podem ser classificados de duas formas: *outliers* globais e *outliers* espaciais ou locais. Os *outliers* globais são observações que se distanciam muito do restante das outras observações tanto para cima (superior) quanto para baixo (inferior). A identificação dos *outliers* globais pode ser através dos instrumentos: *box-plot* e *box-map*[1], caracterizada como uma ferramenta para detectar *outliers* globais superiores (ANSELIN, 1993).

Os *outliers* espaciais ou locais podem ser definidos como observações que não seguem o mesmo processo de dependência espacial que o padrão da maioria dos dados. Anselin (1996) afirma que os *outliers* espaciais podem ser sinais de má especificação da matriz de pesos espaciais (*W*) ou de inadequada escala espacial dos dados.

É importante salientar a diferença entre *outlier* espacial e pontos de alavancagem no espaço. Pontos de alavancagem são observações que, embora seguindo a mesma associação espacial dos restantes dos dados, exercem uma influência grande na determinação do grau de associação espacial.

Tanto os pontos de alavancagem quanto os *outliers* podem ser identificados através do diagrama de dispersão de Moran (ANSELIN, 1996). A indicação de autocorrelação espacial positiva, ou seja, quando a inclinação da reta da regressão é positiva para uma variável, significa que a maioria das observações está localizada nos quadrantes AA e BB. Ao contrário, uma autocorrelação espacial negativa indica que as observações situam-se nos quadrantes AB e BA (ANSELIN, 1996).

Nesse sentido, pode-se identificar os *outliers* espaciais como observações localizadas nos quadrantes AB e BA, enquanto observações situadas nas associações AA e BB representam pontos de alavancagem.

1.6 Áreas de influência

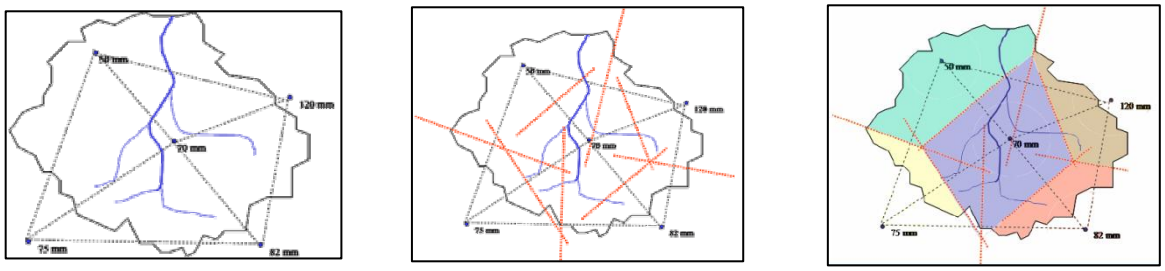
A precipitação varia temporal e espacialmente e o conhecimento dessa distribuição e variação da precipitação é imprescindível para estudos hidrológicos. Para calcular a precipitação média de uma superfície qualquer, é necessário utilizar as observações dos postos dentro dessa superfície e nas suas vizinhanças. Existem três métodos para o cálculo da chuva média: método da média aritmética, método de Thiessen e método das Isoietas (PEDRAZZI, 1999).

[1]*Box-map* é uma ferramenta que realiza o mapeamento dos valores dos quartis registrados nos respectivos municípios. Normalmente associado a cores, o *box map* assinala os *outliers* globais superiores e inferiores, que são identificados através do *box plot*, facilitando a observação e análise exploratória espacial.

O método dos Polígonos de Thiessen é indicado quando não há distribuição uniforme dos postos pluviométricos dentro da bacia hidrográfica. Consiste em atribuir um fator de peso aos totais precipitados medidos em cada posto pluviométrico, sendo esses pesos proporcionais à área de influência de cada posto. São considerados os postos inseridos na bacia e postos localizados na região de entorno e que exercem influência na bacia (CECÍLIO, 2006).

As análises Thiessen, também conhecidas como análises do diagrama de Voronoi ou Tesselação de Delaunay, podem ser aplicadas na gestão de diversos temas, tais como: meio ambiente, *marketing*, segurança e saúde, entre outros. Os diagramas podem ser obtidos com o uso de um Sistema de Informações Geográficas (SIG). Cada polígono do tema Thiessen contém os atributos do ponto dentro dele.

Unwin e Unwin (1998) ressaltam que a região de influência de cada estação meteorológica pode ser obtida pela aplicação do método dos polígonos de Thiessen. O primeiro passo é traçar linhas que unem os postos pluviométricos mais próximos. A seguir, é determinado o ponto médio em cada uma dessas linhas e, a partir desse ponto, é traçada uma linha perpendicular. A interceptação das linhas médias entre si e com os limites da bacia definirão a área de influência de cada um dos postos. Essa sequência é demonstrada pela Figura 11.



Traçar linhas que unem os postos pluviométricos mais próximos entre si.

Traçar linhas médias perpendiculares às linhas que unem os postos pluviométricos.

Definir a região de influência de cada posto pluviométrico e medir a sua área.

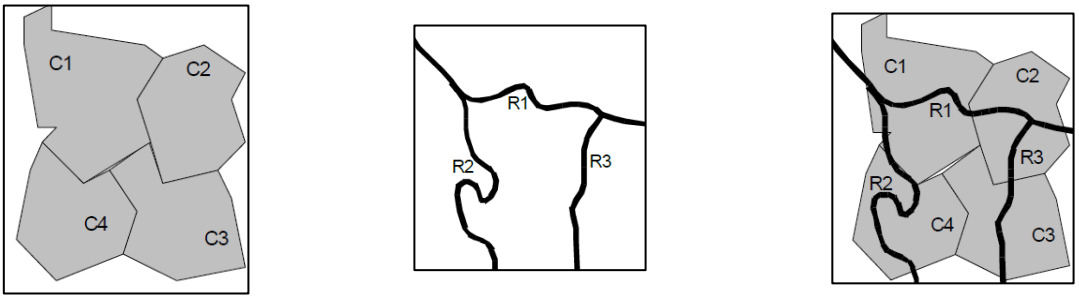
Figura 11 Determinação de áreas de influência pelo método de Thiessen

Com esse processo realizado, cada polígono, com sua área calculada representa um valor específico da variável medida, no exemplo, a precipitação. Em relação aos municípios, eles podem fazer parte de um ou mais polígonos e o valor da variável precisa ser computado para cada município, levando em consideração sempre a área do município que faz parte do polígono.

A operação de *Spatial Join* permite que o valor da variável em estudo amostrada em cada polígono de Thiessen seja atribuído a cada município contido pelos polígonos. Quando uma área (como um município) se encontra em mais de um polígono, o índice obtido para cada polígono é utilizado para que seus valores possam ser utilizados como amostra de influência para as áreas. Veenhof *et al.* (1995) definem *Spatial Join* como uma

operação entre dois conjuntos em um espaço multidimensional que seleciona pares de objetos que satisfaçam uma relação entre eles, que envolvam os valores de seus atributos, como uma intersecção.

A junção espacial combina duas relações de geometrias georreferenciadas de acordo com algum predicado espacial, como intersecção e distância entre objetos (FORNARI, 2006). A Figura 12 ilustra a situação de junção entre um mapa de municípios e um mapa de rios.



(a) Municípios (b) Rios (c) Sobreposição dos conjuntos

Figura 12 Exemplo de junção espacial

Na Figura 12a estão polígonos representando a área de cada município. Na Figura 12b encontram-se polilinhas representando os rios da região. Finalmente, na Figura 12c é apresentado o resultado da sobreposição de ambas as camadas de informação espacial. A execução de uma Junção Spacial (*Spatial Join*) seria, portanto, formada pelo conjunto de pares RS = {(C1, R1), (C2, R1), (C1, R2), (C4, R2), (C2, R3), (C3, R3)}.

1.7 Modelagem espacial

A modelagem espacial é uma das abordagens possíveis em análise espacial que está direcionada à estruturação, ao funcionamento e à dinâmica dos sistemas, incluindo um espectro abrangente de modelos, que vão desde os de planejamento, de interação espacial e de economia regional até os de localização-alocação e de escolha espacial, como o modelo de alocação por múltiplos critérios.

Christofoletti (1999) define um modelo como sendo “[...] uma estruturação simplificada da realidade que supostamente apresenta, de forma generalizada, características ou relações importantes.” Pode-se considerar que os modelos são aproximações subjetivas, por não incluírem todas as observações ou medidas associadas, mas se constituem em importantes instrumentos para análise por obscurecerem detalhes acidentais e por permitirem o aparecimento dos aspectos fundamentais da realidade. Em geral, os modelos não representam a realidade em si, mas sim a visão do modelador para definir a forma como ele percebe e compreende essa realidade.

Uma das funções dos modelos é servir como instrumento para o planejamento, a partir da simulação de cenários possíveis em função de mudanças ambientais. Atualmente, o potencial desses instrumentos tem sido bastante explorado em meios científicos e acadêmicos e sua incorporação ao planejamento vem sendo cada vez mais rápida, pelo desenvolvimento e pela aplicação de metodologias de suporte à tomada de decisão (SANTOS et al., 2004). Isso porque a modelagem pode auxiliar o planejamento na realização de previsões, considerando as implicações de planos alternativos, sem os custos de esperar ou de colocá-los em prática. A partir dessas previsões, pode-se tomar decisões e fazer escolhas entre os cenários simulados pela modelagem com mais segurança e possibilidade de sucesso.

Os modelos de suporte à decisão podem ser definidos como "[...] um sistema interativo que proporciona ao usuário acesso fácil a modelos decisórios e dados a fim de dar apoio a atividades de tomadas de decisões semiestruturadas ou não estruturadas" (CHRISTOFOLETTI, 1999). Esses modelos podem ter objetivos genéricos ou específicos. Os que adotam objetivos genéricos organizam uma arquitetura com ponto de partida para a solução de diversos problemas, mas possuindo sempre uma trajetória similar para as soluções pretendidas, são procedimentos metodológicos. Já os modelos específicos baseiam-se nos dados disponíveis, no problema concreto que deve ser solucionado e nos instrumentos que podem ser utilizados, tendo uma aplicação direcionada.

A principal vantagem da aplicação de modelos para o planejamento está na possibilidade do estudo de vários cenários diferentes e de forma rápida, muitos deles ainda não explorados em experimentos reais. Outra importante vantagem da utilização de simulação de cenários está associada ao seu baixo custo. Na maioria das aplicações, o custo de executar um programa computacional é bem menor do que o correspondente custo relativo à investigação experimental. A maior limitação ao uso de modelos é a dificuldade em trabalhar grande quantidade de dados que descrevem a heterogeneidade dos sistemas naturais. Por essas razões, Sistemas de Informações Geográficas (SIGs) são empregados na criação do banco de dados desses modelos.

A construção de modelos ou modelagem envolve a formulação, o ajuste e o diagnóstico do modelo de uma maneira iterativa e interativa (CHATFIELD, 1995). A formulação envolve considerações do problema em estudo, hipóteses, teorias. Isso indicará as possíveis variáveis que entrarão no modelo e, também, indicará restrições nos parâmetros e nas variáveis. O diagnóstico do modelo é uma etapa fundamental da modelagem. Nesta etapa, verifica-se o ajuste do modelo e se as suposições acerca do modelo são satisfeitas. Técnicas gráficas são as indicadas ou preferenciais. Se necessário, o modelo é modificado e um novo modelo é ajustado, isso indica que o processo é iterativo. Como existe a participação ativa do analista, o processo também é interativo.

1. 7. 1 Modelos de regressão espacial

Dados espaciais agregados são caracterizados pela dependência (autocorrelação espacial) e pela heterogeneidade ou estrutura espacial (ANSELIN, 1988). Esses efeitos espaciais são importantes, pois em alguns casos são os principais responsáveis pela realização dos eventos. Entretanto, invalidam os resultados dos modelos tradicionais de regressão, por violarem alguns pressupostos como a independência e a homocedasticidade. Assim, pela necessidade de se incorporar tais fenômenos à estrutura de um modelo é que foram desenvolvidos os modelos de regressão espacial como são conhecidos na literatura (SILVA, 2006).

Tipicamente, quando se faz uma análise de regressão, procura-se encontrar um bom ajuste entre os valores preditos pelo modelo e os valores observados da variável dependente. Além disso, procura-se descobrir quais das variáveis explicativas contribuem de forma significativa para o relacionamento linear. A hipótese padrão é que as observações não são correlacionadas e, portanto, os resíduos do modelo são independentes e não correlacionados com a variável dependente, além de apresentar Distribuição Normal com média zero e variância constante. No caso de dados, onde está presente a dependência espacial, é bem pouco provável que a hipótese padrão de observações não correlacionadas seja verdadeira. No caso mais comum, os resíduos continuam apresentando a autocorrelação espacial presente nos dados, que pode se manifestar por diferenças regionais sistemáticas ou, ainda, por uma tendência espacial contínua (LOPES et al., 2006).

Desta maneira, a investigação dos resíduos da regressão, em busca de sinais da estrutura espacial, pode fornecer um indicativo da necessidade da utilização de um modelo de regressão espacial. As ferramentas usuais de análise gráfica e o mapeamento dos resíduos podem fornecer a primeira indicação de que os valores observados estão mais correlacionados do que seria esperado sob uma condição de independência (FOTHERINGHAM et al., 2000). Somado à análise gráfica, pode-se fazer uso de testes estatísticos para verificação de autocorrelação espacial nos resíduos da regressão, como a análise do índice *I* de Moran.

A análise de regressão em dados espaciais incorpora, na modelagem, a dependência espacial entre os dados, melhorando o poder preditivo do modelo. Primeiramente, faz-se a análise exploratória com o intuito de identificar a estrutura de dependência nos dados, visando a definição da forma de incorporação dessa dependência ao modelo de regressão. Existem dois tipos básicos de modelagem que permitem incorporar o efeito espacial: de forma Global e Local (ANSELIN, 2002; CÂMARA et al., 2002; FOTHERINGHAM et al., 2000).

Os modelos de forma Global capturam a estrutura espacial por meio de um único parâmetro que é adicionado ao modelo de regressão tradicional. Os modelos mais simples

são: modelo espacial autorregressivo misto (*Spatial Auto Regressive Model*, SAR ou *Spatial Lag Model*) e modelo do erro espacial (*Conditional Auto Regressive Model,* CAR ou *Spatial Error Model*) (LOPES et al., 2006).

1. 7. 1. 1 Regressão linear espacial

Modelos de regressão linear espacial podem ser vistos como uma generalização do modelo padrão de regressão linear, de tal forma que a autocorrelação espacial é possível e contabilizada explicitamente por modelos espaciais. Os parâmetros do modelo incluem coeficientes usuais de regressão das variáveis explicativas (β) e da variância do termo de erro (σ^2). Além disso, os mais utilizados modelos de regressão espacial têm um coeficiente espacial autorregressivo (ρ), que mede a força de autocorrelação espacial. Uma matriz de pesos espaciais (*W*) correspondente a uma estrutura de vizinhança e uma variância da matriz de peso (*D*) são pré-especificadas (CHI;ZHU, 2007).

1. 7. 1. 2 SAR (Spatial Auto Regressive Model) ou Spatial Lag Model

No modelo SAR (ou LAG) (6), a autocorrelação espacial ignorada é atribuída à variável dependente Y. Considera-se a dependência espacial através da adição, ao modelo de regressão, de um novo termo na forma de uma relação espacial para a variável dependente (ANSELIN, 2002).

$$Y = \rho WY + X\beta + \varepsilon \qquad \text{Eq. (6)}$$

em que ***Y*** – vetor da variável dependente, *n* x 1; ***X*** – matriz de variáveis independentes, *n* x *p*; β - vetor de coeficientes de regressão, *n* x 1; ε - vetor de erros aleatórios, *n* x 1 com vetor de médias zero e covariância σ^2 I_n; ***W*** - matriz de vizinhança espacial ou matriz de ponderação espacial, *n* x *n*; ρ - coeficiente espacial autorregressivo, I_n matriz-identidade *nxn*..

A hipótese nula para a não-existência de autocorrelação é que H_0: $\rho = 0$ *versus* a hipótese alternativa H_1: $\rho \neq 0$ a um nível de α% de significância A ideia básica é incorporar a autocorrelação espacial como componente do modelo.

1. 7. 1. 3 CAR (Conditional Auto Regressive Model) ou Spatial Error Model

O segundo tipo de modelo de regressão espacial com parâmetros globais, também referido como *SpatialError Model*, considera que os efeitos espaciais é ruído ou perturbação, ou seja, precisa ser removido. Neste caso, os efeitos da autocorrelação espacial são

associados ao termo de erro $\mathcal{E}$ e o modelo pode ser expresso pela Equação 7 (LOPES et al., 2006).

$$Y = X\beta + \varepsilon,\ \varepsilon = \lambda W_{\varepsilon} + \xi \qquad \text{Eq. (7)}$$

Onde:

$\boldsymbol{W_{\varepsilon}}$ – vetor de erros com efeito espacial, *n* x 1; ε - vetor erros aleatórios, *n* x *n*, com vetor de média zero e variância $\sigma^2 I_n$; $\boldsymbol{\lambda}$ - coeficiente autorregressivo, I_n matriz-identidade *nxn*e ξ é o vetor n x 1 do componente do erro com média zero, variância constante e não correlacionada (ruído).

A hipótese nula para a não-existência de autocorrelação é que H_0:$\lambda=0$, ou seja, o termo de erro não é espacialmente correlacionado *versus* a hipótese alternativa H_1: $\lambda\neq0$.

Câmara *et al.* (2002) salientam que, na prática, a distinção entre os dois tipos de modelos de regressão espacial com parâmetros globais é difícil, pois, apesar da diferença nas suas motivações, eles são muito próximos em termos formais.

1.8 Estatística multivariada

Para qualquer tomada de decisão, é preciso levar em consideração diversos fatores. É certo que cada um desses fatores precisa ser visto e analisado de maneira ou peso diferenciados para cada decisão a ser tomada. O uso da intuição nessas tomadas de decisão não permite a identificação sistemática desses fatores (ou variáveis), ou seja, não se identifica de maneira precisa, qual (ou quais) variável(is) afetaram a decisão tomada.

Segundo Brasil (2006), quando se analisa o mundo que nos cerca, identifica-se que todos os acontecimentos, sejam eles culturais ou naturais, envolvem um grande número de variáveis. As diversas ciências têm a pretensão de conhecer a realidade e de interpretar os acontecimentos e os fenômenos, baseados no conhecimento das variáveis intervenientes, consideradas importantes nesses eventos.

É papel da ciência o estabelecimento de relações, a busca ou proposição de leis que expliquem algum fato. Porém, para que isso seja possível, torna-se necessário o controle, a manipulação e a medição de variáveis que sejam relevantes à compreensão do objeto ou fenômeno de análise. Existem diversas dificuldades para se obter uma tradução das informações para a forma de conhecimento, principalmente quando se busca uma avaliação estatística dessas informações.

Os métodos estatísticos, para analisar variáveis, estão dispostos em dois grupos: um que trata da estatística, que olha as variáveis de maneira isolada – a estatística univariada, e outro que olha as variáveis de forma conjunta – a estatística multivariada (BRASIL, 2006).

A análise de variáveis, antes da existência de computadores, era realizada de maneira isolada e então inferências eram realizadas sobre a realidade. Quando existe a

dependência de diversas variáveis, existem possibilidades de falha nessa análise, pois o conhecimento das informações estatísticas de maneira isolada é insuficiente, uma vez que o necessário é conhecer essas informações de maneira total, assim como as relações entre as variáveis desse conjunto. Quando não são percebidas as relações existentes entre as variáveis, ocorrem dificuldades na interpretação do fenômeno.

O desenvolvimento tecnológico, oriundo das descobertas científicas, tem apoiado o próprio desenvolvimento científico, ampliando, em várias ordens de grandeza, a capacidade de obter informações de acontecimentos e fenômenos que estão sendo analisados. Uma grande massa de informação deve ser processada antes de ser transformada em conhecimento. Portanto, cada vez mais se necessita de ferramentas estatísticas que apresentem uma visão mais global do fenômeno, que aquela possível numa abordagem univariada. A denominação “Análise Multivariada” corresponde a um grande número de métodos e técnicas que utilizam, simultaneamente, todas as variáveis na interpretação teórica do conjunto de dados obtidos (DOURADO NETO, 2004; BRASIL, 2006).

Os pesquisadores devem ter cautela ao trabalhar com as técnicas de análise multivariada, pois a arte do seu uso está na escolha das opções mais apropriadas para detectar os padrões esperados nos seus dados, e as opções mais apropriadas podem não estar no *software* escolhido. Leva-se algum tempo até escolher as melhores opções em análises multivariadas, e recomenda-se o exercício, com cautela, durante o tempo necessário para apreender as limitações dessas análises, antes de tentar explorar suas grandes potencialidades (MAGNUSSON, 2003).

Os métodos multivariados são escolhidos de acordo com os objetivos da pesquisa, pois se sabe que a análise multivariada é uma análise exploratória de dados, prestando-se a gerar hipóteses, e não tecer confirmações a respeito deles, o que seria uma técnica confirmatória, como nos testes de hipótese, nos quais se tem uma afirmação a respeito da amostra em estudo, embora, às vezes, possa ser utilizada para confirmação dos eventos (BRASIL, 2006). Desta maneira, a estatística multivariada e seus diferentes métodos diferem de um conjunto de elementos que possuam a mesma função, pois a fundamentação teórica e sua aplicabilidade são particulares a cada método. Dois métodos são destacados quando se busca verificar o relacionamento entre as amostras: a análise de agrupamento hierárquico e a análise fatorial com análise de componentes principais.

Brasil (2006) ressalta que ao realizar um estudo estatístico, quer seja univariado ou multivariado, sempre existirá a perda de informação, pois no momento que se está reduzindo um conjunto de dados para ser representado pela sua média, no caso univariado, perde-se informação. O mesmo ocorre quando se aplica uma técnica multivariada, pois ao reduzir a dimensionalidade de um problema também se perde informação. O *trade-off* do pesquisador então reside em obter a informação e saber que tem um erro que foi quantificado ou não.

1. 8. 1 Análise de Agrupamentos (AA)

De acordo com Brasil (2006), a AA, em sua aplicação, engloba uma variedade de técnicas e algoritmos, sendo que o objetivo é encontrar e separar objetos em grupos similares. Essa técnica pode ser observada, por exemplo, se se tiver várias máquinas agrícolas em uma determinada área de uma revendedora e distribuir essas máquinas, na revendedora, segundo suas características, de uma mesma finalidade ou o mesmo valor para venda, por exemplo. Aí se está a praticar AA. Agora, se essas máquinas estiverem espalhadas por toda a revendedora, significa que se terá mais de uma característica e, para que se possa uni-los por características comuns, será muito trabalhoso, exigindo conceitos mais sofisticados de semelhança e procedimentos mais científicos para juntá-los. É em relação a esse procedimento multidimensional que se trabalhará.

O conhecimento de algumas características sobre um dado grupo de um conjunto de elementos amostrais torna-se necessário em alguns estudos, principalmente se esse grupo resulta de uma ou mais variáveis. Quando a mensuração é obtida de diferente natureza, pode-se identificar a existência de similaridades nesse conjunto de dados.

A análise de agrupamentos estuda todo um conjunto de relações interdependentes, não distinguindo variáveis dependentes e independentes como na regressão.

Conforme Brasil (2006), a AA pretende resolver o seguinte problema: “[...] dada uma amostra de *n* objetos (ou indivíduos), cada um deles medidos segundo *p* variáveis, procurar um esquema de classificação que agrupe os objetos em *g* grupos. Deve ser determinado, também, o número de variáveis desses grupos”. Portanto, a finalidade dessa técnica é reunir os objetos (indivíduos, elementos) verificados nos grupos em que exista homogeneidade dentro do grupo e heterogeneidade entre os grupos, objetivando propor classificações. Os objetos em um grupo são relativamente semelhantes, em termos dessas variáveis, e diferentes de objetos de outros grupos. Quando utilizada dessa forma, a AA é o inverso da análise de fatores, pelo fato de reduzir o número de objetos, e não o número de variáveis, concentrando-os em um número muito menor de grupos.

A AA constitui uma metodologia numérica multivariada, com o objetivo de propor uma estrutura classificatória ou de reconhecimento da existência de grupos, objetivando, mais especificamente, dividir o conjunto de observações em um número de grupos homogêneos, segundo algum critério de homogeneidade (BRASIL, 2006). Muitas vezes, nessa técnica, são feitas afirmativas empíricas, que nem sempre têm respaldo teórico. Muitas técnicas são propostas, mas não há ainda uma teoria generalizada e amplamente aceita. Devido a isso, deve-se utilizar vários métodos e comparar os resultados, para que a análise dos dados seja realizada pela técnica mais adequada.

Brasil (2006) afirma que a AA é um método simples, calcada nos cálculos de distância. No entanto, não requerem conhecimento estatístico para a sua aplicação, como é

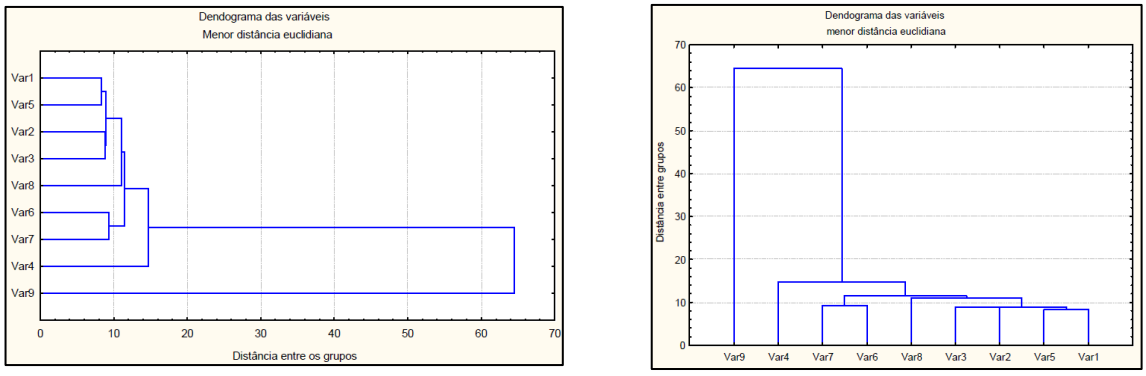
o caso quando se aplica análise de variância, de regressão ou fatorial. O primeiro caso, AA não requer o uso de um modelo; os demais casos necessitam. Para a aplicação da AA, as estatísticas e os conceitos, a seguir, serão utilizados:

- **Esquema de aglomeração**: informa sobre objetos ou casos a serem combinados em cada estágio de um processo hierárquico de aglomeração.
- **Centroide do agrupamento**: representam os valores médios das variáveis para todos os casos ou objetos em um agrupamento particular.
- **Centros de agrupamentos**: são os pontos iniciais em um agrupamento não-hierárquico. Os agrupamentos são construídos em torno desses centros.
- **Composição de um Agrupamento**: indica o agrupamento ao qual pertence cada objeto ou caso (MALHOTRA, 2001).

1. 8. 2 Dendograma

Um dendograma é uma representação gráfica em forma de árvore que resume o processo de agrupamento em uma análise de *clusters*. Os objetos similares se conectam mediante ligações cuja posição no diagrama está determinada pelo nível de similaridade/dissimilaridade entre os objetos (VILLARDÓN, 2011).

No dendograma (horizontal ou vertical), um dos eixos representa os grupos unidos por ordem decrescente de semelhança e o segundo indica as distâncias entre os grupos que foram formados. O dendograma é lido de cima para baixo, quando for feito na forma horizontal ou da direita para a esquerda na forma vertical (Figura 13).



(a) Horizontal (b) Vertical

Figura 13 Dendogramas

Verifica-se, na Figura 13, que as variáveis *Var1* e *Var5* são as que possuem a maior semelhança, no dendograma, por possuírem a menor distância euclidiana, sendo essas a formarem o primeiro grupo. Logo em seguida, vêm as variáveis Var2, Var3, Var8 e assim sucessivamente as variáveis serão agrupadas, por ordem decrescente de semelhança, ou seja, a Var9 formou o último grupo do dendograma, o qual manteve-se

distinto dos demais grupos formados, pelo fato de essa variável possuir pouca semelhança em relação às outras.

De forma geral, os dendogramas apresentam estruturas de agrupamentos de objetos homogêneos. Entretanto, a falta de critérios objetivos para se determinar o ponto de corte no dendograma (número ótimo de grupos) ainda é um problema em estudos que utilizam a análise de agrupamentos. Um método considerado como “objetivo”, entre os poucos existentes, é o Método de Mojema. Esse Método é um procedimento baseado no tamanho relativo dos níveis de fusões (distâncias) no dendograma (FARIA, 2009).

No presente trabalho, propõe-se ainda outro critério, de fácil entendimento, para determinação do número ótimo de grupos, baseado nas trajetórias das curvas dos índices RMSSTD e RS.

1. 8. 3 Índice RMSSTD e RS

O índice RMSSTD (*Root Mean Square Standard Deviation*), cuja tradução pode ser raiz quadrada do desvio padrão médio, é usado para calcular a homogeneidade dos agrupamentos (SHARMA, 1996). Em outras palavras, quanto mais compactos forem os grupos formados, situação esta verificada na presença de um grande número de grupos, menores os valores para essa estatística. Assim, é possível visualizar um gráfico (Figura 14a) que mostra o decréscimo do RMSSTD em função do aumento do número de *clusters*, todavia, essa trajetória não é linear e o seu ponto de máxima curvatura indica um limiar entre uma fase de decréscimo e uma fase de estabilização. Após esse ponto, denominado de ótimo, mesmo aumentando o número de *clusters* não se verifica grandes declínios nos valores do RMSSTD.

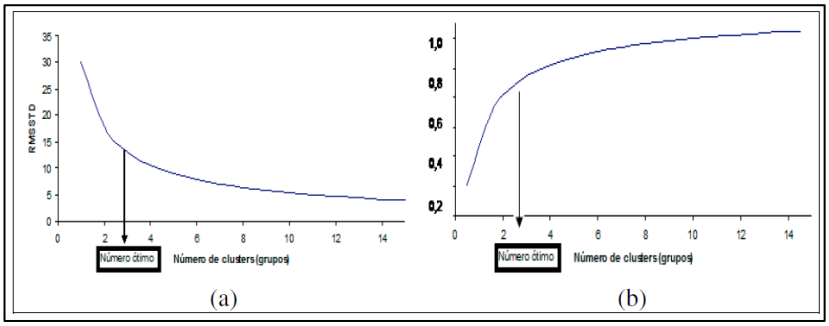


Figura 14 Trajetória dos índices RMSSTD (a) e RS (b) em função do aumento do número de *clusters* (grupos)

O índice R-square (RS), ou coeficiente de determinação, é usado para calcular a dissimilaridade entre agrupamentos. Um alto valor de RS indica dissimilaridade mais alta entre grupos (SHARMA, 1996) e tal situação é representada na presença de um alto número de grupos. Graficamente, o aumento no número de *clusters* proporciona um aumento nos

valores do RS (Figura 14b), e esta trajetória também é não-linear, o que realça a importância de se calcular um ponto de máxima curvatura.

Na literatura estatística, pontos de máxima curvatura geralmente são calculados em estudos de determinação de tamanho ótimo de parcelas experimentais; portanto, os métodos empregados nessa ocasião podem ser usados para estimar o número ótimo de *clusters* em se tratando das trajetórias mostradas na Figura 14.

1.9 Conjuntos *fuzzy* como modeladores de incerteza

A teoria dos conjuntos *fuzzy* vem se desenvolvendo, ganhando espaço e está sendo usada como instrumento para a criação de modelos em diferentes áreas de conhecimento. A teoria foi introduzida em 1965 pelo matemático Lotfi Asker Zadeh, com a intenção de dar um tratamento matemático a certos termos linguísticos subjetivos como: “aproximadamente”, “em torno de”, entre outros. Pode-se dizer que a Teoria dos Conjuntos Fuzzy representa um primeiro passo no sentido de se programar e armazenar conceitos vagos em computadores, tornando possível a produção de cálculos com informações imprecisas, a exemplo do que faz o ser humano.

Para descrever certos fenômenos (relacionados ao mundo sensível), utilizam-se graus que representam qualidades ou verdades parciais ou ainda padrões do melhor. Esse é o caso, por exemplo, dos conceitos de “úmido”, “quente”, “maduro” e “produtivo”.

Um exemplo a ser citado poderia ser a fixação de um conjunto de plantas adultas. Uma proposta para formalizar matematicamente tal conjunto poderia ter pelo menos duas abordagens. A primeira (clássica), distinguindo a partir de que valor da altura uma planta é considerada madura. Nesse caso, o conjunto está bem definido. A segunda, menos convencional, é dada de maneira que as plantas sejam consideradas maduras com mais ou menos intensidade, ou seja, existem elementos que pertenceriam mais à classe dos maduros que outros. Isso significa que quanto menor for a medida da altura da planta, menor será seu grau de pertinência a essa classe. Desse modo, pode-se dizer que as plantas pertencem à classe das plantas maduras, com mais ou menos intensidade. É com base em desafios como esse, no qual a propriedade que define o conjunto é incerta, que surgiu a teoria dos conjuntos *fuzzy*, que tem crescido consideravelmente, tanto do ponto de vista teórico como nas aplicações em diversas áreas de estudo.

A palavra “*fuzzy*”, de origem inglesa, significa incerto, vago, impreciso, subjetivo, nebuloso, difuso. Embora a teoria dos conjuntos *fuzzy* estude casos de incertezas, vale lembrar que tal teoria é muito bem definida. O que é incerto é a propriedade que define o conjunto em questão. Para obter a formalização matemática de um conjunto *fuzzy*, o matemático Zadeh baseou-se no fato de que qualquer conjunto clássico pode ser caracterizado por uma função característica, cuja definição é dada a seguir.

Definição 1. Seja *U* um universo de discurso e *A* um subconjunto de *U*. A função característica de *A* é dada por

$$X_A{}^{(x)} = \begin{cases} 1, sex \in A \\ 0, sex \notin A \end{cases} \quad \text{Eq. (8)}$$

Assim, *A* é uma função cujo domínio é *U* e a imagem está contida no conjunto {0, 1}. Aqui, $X_A{}^{(x)} = 1$ indica que o elemento *x* está em *A*, enquanto $X_A{}^{(x)} = 0$ indica que *x* não é elemento de A. Dessa maneira, a função característica descreve completamente o conjunto *A*, já que tal função indica quais elementos do conjunto universo *U* são também elementos de *A*. No entanto, existem casos em que a pertinência entre elementos e conjuntos não é precisa, ou seja, não é possível dizer se um elemento pertence efetivamente a um conjunto ou não.

O que é admissível é dizer qual elemento do conjunto universo se enquadra "melhor" ao termo que caracteriza o subconjunto. Por exemplo, considerando o subconjunto dos números reais "próximos de 2".

$$A = \{x \in \mathbb{R}: x\acute{e}pr\acute{o}ximode\ 2\} \quad \text{Eq. (9)}$$

Poderia ser perguntado se o número 7 e o número 2,001 pertencem a *A*. A resposta é incerta, pois não é sabido até que ponto pode-se dizer objetivamente quando um número está próximo de 2. O que pode afirmar é que 2,001 está mais próximo de 2 do que 7. Desta maneira, apresentam-se na sequência as formalizações dos conceitos da teoria dos conjuntos *fuzzy* com a noção de subconjunto *fuzzy*.

Definição 2. Seja *U* de um universo de discurso. Um subconjunto *fuzzy F* de *U* é caracterizado por uma função pré-fixada, chamada função de pertinência do subconjunto *fuzzy F*.

$$\varphi = \{U \to [0,1]\}, \quad \text{Eq. (10)}$$

A classe de todos os subconjuntos *fuzzy* de *U* é denominada por *F(U)*. O valor $\varphi_F{}^{(x)} = 0$F(x) indica o grau com que o elemento *x* de *U* está no conjunto *fuzzy F*. Em particular, $\varphi_F{}^{(x)} = 0\ e \varphi_F{}^{(x)} = 1$ indicam, respectivamente, a não-pertinência e a pertinência completa de *x* ao conjunto *fuzzy F*.

Do ponto de vista formal, a definição de subconjunto *fuzzy* foi obtida simplesmente ampliando-se o contradomínio da função característica, que é o conjunto {0,1}, para o intervalo [0,1]. Nesse sentido, pode-se dizer que um conjunto clássico é um caso particular de conjunto *fuzzy*, cuja função de pertinência φ_F é sua função característica X_F. Na linguagem *fuzzy*, um subconjunto clássico costuma ser denominado subconjunto *crisp*.

1. 9. 1 Conceito de *Fuzzy C-means*

Fuzzy C-means, ou FCM, é um algoritmo para separar em grupos de similaridades, ou *clusters*, um conjunto de pontos ou dados, minimizando uma função objetiva *fuzzy* escolhida previamente. Os *clusters* obtidos por esse algoritmo são representados por vetores protótipos. A similaridade entre esses vetores protótipos e os vetores de dados é representada por um grau de pertinência entre 0 e 1. Deste modo, um modo mais simples de se definir o FCM é como sendo um algoritmo para obtenção de funções de pertinência (FUJIMOTO, 2005).

1. 9. 1. 1 Similaridade

Um conceito muito importante na lógica *fuzzy* e utilizada no conceito do *Fuzzy C-Means* é a similaridade, por ser uma das medidas amplamente utilizadas e discutidas para obter as funções de pertinência. A similaridade entre dois elementos podem ser definida de duas formas:

- **similaridade estrutural**: semelhança de forma, evolução e característica de uma determinada função que descreve o elemento;
- **similaridade pontual**: pela medida de distância no espaço de característica que descreve o elemento.

A diferença entre essas duas definições é mais bem observada quando se faz uma comparação entre duas funções no tempo ou em qualquer outro domínio (por exemplo, sinais de vibrações e espectros de frequência). Uma função pode ser tratada de duas maneiras, por meio dos pontos de sua trajetória (uma vez que na prática normalmente se trabalha com funções discretas) ou por características numéricas que descrevem o perfil desta função.

No primeiro, ao se comparar duas funções pelos pontos de sua trajetória, pode-se considerar cada um desses pontos como sendo uma característica distinta. Com isso, alcança-se um espaço de característica de *N* dimensões, sendo *N* o número de pontos da função. A similaridade pontual considera esse espaço de característica para definir o grau de similaridade. Esse grau de similaridade é obtido por meio de uma medida de distância.

Já a similaridade estrutural está mais ligada às características numéricas que descrevem o perfil de uma função. Entre essas características, pode-se citar o RMS, mínimo e máximo global, características estatísticas (*Skewness* e *Kurtosis*), entre outros. O grau de similaridade é obtido comparando um conjunto dessas características, e, portanto, pode-se dizer que essa similaridade está relacionada à escolha dessas características. Essa similaridade pode ser obtida também por meio da medida de distância.

Por meio da Figura 15, pode-se exemplificar essa diferença de similaridade. Considerando a similaridade pontual, as funções A e B são mais similares entre sique em relação a C e D. Porém, ao se considerar a similaridade estrutural (neste exemplo, apenas a ondulação da função), as funções A e B perdem essa similaridade. Na Figura 15, a função estruturalmente mais similar à A é a função C, e a função mais similar à B é a função D.

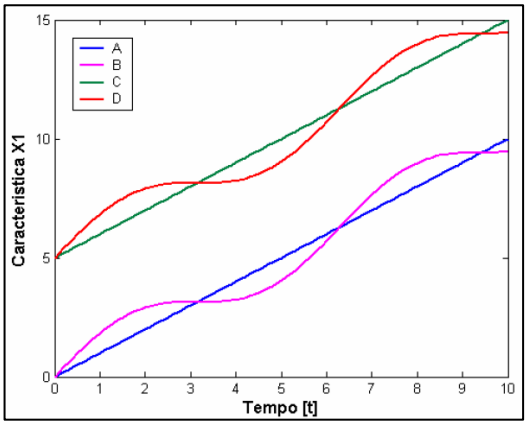


Figura 15 Exemplo de similaridades

Fica claro, desta maneira, que existe um ponto de sobreposição entre essas duas definições de similaridade. Essa sobreposição se refere ao fato de se considerar os pontos da trajetória da função como características que a descrevem o seu perfil, e posteriormente tratá-los de modo semelhante à similaridade pontual (por meio de medida de distância). Outra diferença entre similaridades é que a pontual trabalha com características de mesma natureza e a estrutural trabalha com características de diferentes naturezas.

1. 9. 2 Medindo a Validade do Agrupamento

Milagre (2008) afirma que o objetivo de agrupar é descobrir, automaticamente, o intrínseco agrupamento de um conjunto de dados (LAW; JAIN, 2003). Assim, a principal preocupação em um processo de agrupamento é revelar a organização dos padrões dentro de grupos sensíveis, os quais permitem descobrir similaridades e diferenças e derivar inferências úteis sobre eles (HALKIDI et al., 2002).

Diferentes algoritmos de agrupamento já foram propostos para diferentes aplicações e tamanhos de bases de dados (JAIN et al., 1999; LAW;JAIN, 2003; HALKIDI et al., 2002). A aplicação de um algoritmo a uma base de dados tem como objetivo – assumindo-se que a base de dados oferece uma tendência ao agrupamento – descobrir suas partições naturais. Entretanto, o processo de agrupamento é considerado um processo não-supervisionado, tendo em vista que não são predefinidas classes nem exemplos que mostrem que tipo de relações desejáveis devem ser válidas entre os dados. Desta maneira, os vários algoritmos de agrupamento baseiam-se em algumas suposições para definir o particionamento de uma base de dados, o que pode levar a resultados diferentes,

dependendo das características do conjunto de dados, como, por exemplo, geometria e densidade de distribuição dos grupos e dos valores dos parâmetros de entrada (HALKIDI et al., 2002).

Desta maneira, adicionadas as diferenças entre os resultados dos algoritmos, deve-se considerar que os algoritmos de agrupamentos produzem um modelo de particionamento para uma base de dados, quer existam ou não, tornando-se necessária a validação de tais agrupamentos, seja verificando se o modelo de agrupamento obtido é o que mais se adequa ao conjunto de dados ou avaliando-se a qualidade do agrupamento.

O procedimento de avaliação que pode fornecer uma resposta quantitativa para essas questões é denominado validade do agrupamento (*cluster validity*), o qual é o objetivo de muitos esforços de pesquisadores (LAW;JAIN, 2003; PAL;BEZDEK, 1995; HALKIDI et al., 2002).

Em termos gerais, existem três métodos para investigar a validade de um agrupamento, que são os critérios de avaliação externos, internos e relativos. Nos critérios de avaliação externos, a avaliação dos resultados do algoritmo de agrupamento baseia-se na comparação entre o particionamento obtido pelo algoritmo com uma estrutura de grupos pré-especificada que é imposta ao conjunto de dados e reflete nossa intuição sobre a estrutura que ela deve possuir. Nos critérios internos, o objetivo é avaliar o agrupamento resultante, utilizando-se somente características intrínsecas da base de dados, ou seja, baseiam-se em medidas de avaliação internas. No método de critérios relativos, a ideia básica é avaliar a estrutura do agrupamento por meio da comparação dele com outros esquemas de agrupamentos resultantes do mesmo algoritmo, mas com diferentes valores para os parâmetros (HALKID*et al.*, 2002; LAW;JAIN, 2003).

1. 9. 2. 1 Fuzziness Performance Index (FPI)

De Franco (2002) informa que o Índice de Performance da Nebulosidade – FPI (*Fuzziness Performance Index*) é uma medida de validação do número de categorias ideal de um conjunto amostral (ROUBENS, 1982). Ele estima o grau de nebulosidade de um sistema gerado por um método de categorização. O número ótimo de categorias é obtido pelo valor mínimo de *FPI*. A medida de validação *FPI* é estritamente nebulosa e sua fórmula é dada por:

$$FPI = 1 - \frac{c}{c-1}\left[1 - \sum_{k=1}^{n}\sum_{i=1}^{c} {(u_{ik})^2}/{n}\right] \qquad \text{Eq. (11)}$$

em que: *c* representa a quantidade de agrupamentos; *n* corresponde ao número de polígonos; u_{lk} corresponde ao grau de pertinência do polígono *k* do agrupamento *l*. Os

valores de *FPI* pertencem ao intervalo 0 ≤ *FPI* ≤ 1. Se a matriz *U* é rígida, o valor de *FPI* é 0, enquanto que para *FPI* = 1, o sistema atinge seu maior grau de nebulosidade.

1. 9. 2. 2 Modified Partition Entropy (MPE)

De Franco (2002) apresenta a Entropia de Partição Modificada – MPE (*Modified Partition Entropy*), ressaltando que é uma medida de validação do número ideal de categorias (ROUBENS, 1982). A medida MPE também é estritamente nebulosa e ela calcula o grau de desorganização gerado por cada número de categorias em que um espaço amostral foi dividido, sendo calculada pela fórmula:

$$MPE = \frac{-\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}\, log(u_{ik})/n}{log(c)} \qquad \text{Eq.(12)}$$

Os valores de MPE estão no intervalo 0 ≤ *MPE* ≤ 1. Quando *MPE*=0, a disposição das categorias do sistema se aproxima mais da forma rígida, enquanto *MPE*=1 indica o maior grau de nebulosidade possível. Quanto menor o valor de *MPE*, mais organizado é o sistema analisado e assim pode ser escolhido o melhor número de categorias deste.

1. 9. 2. 3 Compactness and Separation (CS)

A Compacidade e Separação – CS (*Compactness and Separation*) é uma medida de validação mais completa, pois avalia tanto a compacidade das categorias geradas como a qualidade da separação entre estas. Quanto menor o valor de CS, melhor a disposição das categorias. Minimizar CS corresponde a minimizar a função objetivo *Jm* (Equação 13), que é a finalidade do método FCM. Portanto, a execução que obteve o menor valor de *CS* é a execução do método de categorização que foi mais bem-sucedida na minimização da função objetivo, possuindo assim a melhor compacidade e separação (FRANCO, 2002).

$$J_m(U, V; X) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m \|x_k - v_i\|^2 \qquad \text{Eq. (13)}$$

sendo $V = \{v_1,...,v_c\}$ um conjunto de vetores que representa os *c* centroides dos agrupamentos, também chamados de protótipos (PAL;BEZDEK, 1995; PEDRYCZ;VUKOVICH, 2004), sendo $v_i = (v_{i1}; \ldots, v_{ip})^{\mathrm{T}} \in \boldsymbol{R}^p$ para *i*=1,...,*c*. Cada dado x_k será avaliado segundo sua proximidade a cada centroide v_i. Essa comparação é efetuada segundo a norma Euclidiana entre x_k e v_i, isto é, $\|x_k - v_i\| = \left(\sum_{l=1}^{p} x_{kl} - v_{il}\right)^{1/2}$. Os centroides dos agrupamentos v_i, e os graus de pertinência u_{ik} são obtidos segundo as Equações 14 e 15 para m≠1, respectivamente.

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \qquad \text{Eq. (7)}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|}\right)^{\frac{2}{m-1}}} \qquad \text{Eq.(14)}$$

Para a constante nebulosa *m*, CS é escrita como:

$$CS = \frac{J_m}{\mathrm{n}(d_{min})^2} \quad \text{Eq.(15)}$$

A fórmula de CS pode ser alterada para validar categorias geradas por métodos de categorização com diferentes funções objetivo. O fator d_{min} é a distância Euclidiana mínima entre dois centros de catego0rias. É ele quem mede a separação entre as categorias e é dado pela fórmula:

$$d_{min} = \min_{i \neq j} |c_i - c_j| \quad \text{Eq. (16)}$$

Devido à sua implementação, a CS só é uma medida inócua quando o número de categorias tende a *n*. Mas isso não chega a ser um problema, pois *c* sempre é bem menor que *n* nos problemas de categorização. A medida CS é a mais completa e precisa medida de validação dentre as apresentadas, pois, além de validar o número de categorias, ela também analisa o grau de separação entre elas, não perdendo exatidão mesmo quando a sobreposição das categorias é alta. Além de avaliar partições nebulosas, a medida CS também pode ser usada para avaliar partições rígidas, desde que os graus de inclusão nebulosos sejam substituídos por graus de inclusão rígidos.

1. 9. 2. 4 Inter Class Contrast (ICC*)*

A ICC foi criada para ser capaz de avaliar um espaço particionado por uma ferramenta de categorização nebulosa ou rígida, levando em conta a separação entre as categorias geradas e a compacidade destas. Ela também foi moldada para detectar centros alocados muito próximos, o que compromete uma boa categorização (FRANCO, 2002). Quanto maior o seu valor, melhor o particionamento do conjunto de dados. Sua fórmula é dada por:

$$ICC = \frac{|S_{Be}|}{\mathrm{n}} D_{min}\sqrt{c}, \quad \text{Eq. (17)}$$

S_{Be}é uma métrica que estima a qualidade da alocação dos centros e é dada pela Equação 18.

$$S_{Be} = \sum_{i=1}^{c}\sum_{k=1}^{n} \mu_{ik}(m_{ei} - md)^2 \quad \text{Eq. (18)}$$

sendo *md* o centroide do conjunto de todas as amostras e é quantificado pela equação 19.

$$md = \frac{1}{n}\sum_{k=1}^{n} x_k \quad \text{Eq. (19)}$$

m_{ei} é o centroide da categoria *i*, *i*=1,...,*c* definido na pela equação 20.

$$m_{ei} = \frac{\sum_{k=1}^{n} \mu_{ik} x_k}{\mu_i} \quad \text{Eq. (20)}$$

e μ_i é o somatório dos graus de pertinência de todas as amostras na categoria *i*,*i*=1,...,*c*, determinado pela equação 21.

$$\mu_i = \sum_{j=1}^{n} \mu_{ij} \quad \text{Eq. (21)}$$

Para evitar o falso crescimento da ICC, quando o número de categorias é maior que o número de classes, devido ao comportamento do termo s_{Be}, o termo D_{min} foi acrescentado à sua fórmula, dado que quando duas ou mais categorias são associadas a uma mesma classe, a distância mínima entre os centros D_{min} decresce abruptamente. Assim, D_{min} evita que o máximo valor de ICC seja atingido para um valor de *c* maior que o valor ideal. D_{min} é a distância Euclidiana mínima entre os centros das categorias e é dado pela fórmula:

$$D_{min=}^{min}{}_{1\leq i\leq c} \left[\min_{i+1\leq j\leq c\leq 1} \left\| m_{ei} - m_{ej} \right\| \right] \quad \text{Eq. (22)}$$

Quando uma ou mais categorias englobam mais de uma classe, ou seja, quando o número de categorias é menor que o número de classes, a distância mínima D_{min} entre os centros aumenta, aumentando o valor da ICC. Para evitar que a ICC atinja seu valor máximo para um valor de *c* menor que o ótimo, a raiz quadrada do número de categorias foi introduzida em sua fórmula. Essa condição pode ocorrer quando uma ou mais categorias representam mais de uma classe do problema e os centros dessas categorias estão longe um dos outros, gerando valores altos de D_{min} e, consequentemente, valores altos da medida ICC.

Assim, o termo $\sqrt{c}$ garante que a medida ICC cresça juntamente com o número de categorias, alcançando seus valores máximos próximos do valor ótimo de *c*, enquanto D_{min} evita que o valor máximo de ICC seja atingido para um valor de *c* maior que o valor ótimo.

O fator 1/n é um fator de escala, usado para compensar a influência do número de pontos no termo S_{Be}.

1.10 REFERÊNCIAS

ALMEIDA, L. A.; KIIHL, R. A. S.; MIRANDA, M. A. C.; CAMPELO, G. J. A. Melhoramento da soja para regiões de baixas latitudes. In: Embrapa Semi-Árido; Embrapa Recursos Genéticos/Biotecnologia. (Org**.). Recursos Genéticos e Melhoramento de Plantas para o Nordeste brasileiro.** 10. ed. Petrolina-PE; Brasília-DF: Embrapa, v. 1, 1999.

ALMEIDA, E. S.; HADDAD, E. A.; HEWINGS, G. J. D. The spatial pattern of crime in Minas Gerais: an exploratory analysis. **Economia Aplicada**, v. 9, n. 1, p.1-27.2005.

ANSELIN, L. **Spatial econometrics**: Methods and models. Dordrecht, Netherlands). Kluwer Academic Publishers, 1988.193p.

ANSELIN, L. **SpaceStat Tutorial**. 1992. Disponível em: <http://www.spacestat.com>. Acesso em: 10 set. 2012.

ANSELIN L. **Exploratory Spatial Data Analysis and Geographic Information Systems**. DOSES/Eurostat Workshop on New Tools for Spatial Analysis, Lisboa, Portugal, (West Virginia University, Regional Research Institute, Research Paper 9329), 1993.

ANSELIN, L. **Local indicators of spatial association - LISA**. Geographical Analysis, Ohio/USA, 27:91-115, 1995.

ANSELIN, L. The Moran scatterplot as ESDA tool to assess localinstability in spatial association. In: M. Fisher, H. J. Scholten and D.Unwin (ed). **Spatial Analytical Perspectives on GIS.London**, Taylor &Francis, p111-126, 1996.

ANSELIN, L.; BERA, A. **Spatial dependence in linear regression models with an introduction to spatial econometrics**.In: Handbook of applied economic statistics, edited by Amman Ullah and David E.A. Giles.New York: Marcel Dekker, 1998, 640p.

ANSELIN, L. **Properties of Tests for Spatial Error Components**.Regional Science and Urban Economics.v.33, n.5, p.595–618,2002.

ANSELIN L.; SYABRI, I.; SMIRNOV, O. **Visualizing multivariate spatial correlation with dynamically linked windows**. Urbana-Champaign: Spatial Analysis Laboratory, Department of Agricultural and Consumer Economics, University of Illinois, 2004. 13 p. Disponível em: <http://www.real.illinois.edu/d-paper/01/01-t-10.pdf>. Acesso em: 02 set. 2010.

ASSAD, E. D.; MARIN, F. R.; MEDEIROS, S. R. E.; PILAU, F. G.; FARIAS, J. R. R.; PINTO, H. S.; ZULLO JR, J. Sistema de previsão de safra de soja para o Brasil. **Pesquisa Agropecuária Brasileira**, Brasília/DF,v. 42, n.5, p. 615-625, 2007.

ASSUNÇÃO, R. M.; FILHO, C. C. B.; SILVA, B. F. A.; MARINHO, F. C.; REIS, I. A.; ALMEIDA, M. C. M. Conglomerados de homicídios e o tráfico de drogas em Belo Horizonte, Minas Gerais, Brasil, de 1995 a 1999. **Caderno de Saúde pública**, Rio de Janeiro, v. 17 n. 5, p. 1163-1171, 2001.

ASSUNÇÃO, R., M. **Estatística Espacial com aplicações em Epidemilogia, Economia e Sociologia**. 7ª Escola de Modelos de Regressão. São Carlos-SP. 2001b, 131p.

AVELAR, M. B. L. **Análise da agregação espacial do bicho-mineiro do cafeeiro (leucoptera coffeella (Guérin-Mèneville & Perrotter, 1842) (Lepidoptera:Lyonetiidae)) em lavoura cafeeira (coffea arabica L.) orgânica em formação**. Dissertação (Mestrado)-Universidade Federal de Lavras, Lavras, MG, 2008.

AYOADE, J.O. **Introdução a climatologia para os trópicos**. São Paulo: Difel, 1986, 332p.

BAILEY, T.C. **A review of statistical spatial analysis in geographical information systems**. Em: Fotheringham, S., Rogerson, P. Spatial Analysis and GIS. London: Taylor and Francis.p.13-44.1994.

BAILEY, T. C.; GATRELL, A. C. **Interactive spatial data analysis**.Essex: Longman Scientific, 1995.413p.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Fundação Oswaldo Cruz. SANTOS, S.M.; SOUZA, W. (Orgs.). **Introdução à estatística espacial para a saúde pública**. Brasília, 2006. 124p.

BRASIL. Ministério de Desenvolvimento, Indústria e Comércio. **Comércio exterior: balança comercial brasileira.** Disponível em:

<http://www.desenvolvimento.gov.br/sitio/interna/interna.php?area=5&menu=1161>. Acesso em: 21 jun. 2012.

CÂMARA, G.; CASANOVA, M.A.; HEMERLY, A.; MEDEIROS, C. M. B.; MAGALHÃES, G. C. **Anatomia de Sistemas de Informações Geográficas**. UNICAMP, Campinas. IX Escola de Computação. 1996, 26p.

CÂMARA, G.; DAVIS, C.; MONTEIRO, A. M. V. **Introdução à Ciência da Geoinformação**. São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2001, 345p.

CÂMARA, G.; CARVALHO, M. S.; CRUZ, O. G.; CORREA, V. Análise Espacial de Área. In: FUCKS, S. D.; CARVALHO, M. S.; CÂMARA, G.; MONTEIRO, A. M. V. **Análise Espacial de Dados Geográficos**. São José dos Campos: Instituto Nacional de Pesquisas Espaciais – Divisão de Processamento de Imagens, 2002.

CÂMARA, G.; MONTEIRO, A. M. V. **Conceitos básicos em ciência da geoinformação**. São José dos Campos: INPE, 2004, 36p.

CARAMORI, P. H. O clima e a agricultura. In: **Semana de geografia**, 13. Maringá. Anais. Maringá: Universidade Estadual de Maringá, p. 13-22.2003.

CARDOSO, W. S.; DE SÁ, M., E. R.;CRUZ, S. H. R. Indicadores socioespaciais urbanos nos assentamentos precários em Belém/PA. **XIVencontro nacional da ANPUR**. Rio de Janeiro. 2011.

CARGNELUTTI FILHO, A.; MATZENAUER, R; MALUF, J. R. T.; RADIN, B. Variabilidade temporal e espacial da precisão das estimativas de elementos meteorológicos no Rio Grande do Sul. **Ciência Rural**,Santa Maria/RS, v.39, n.4. p.962-970, 2009.

CARMO NETO, O. V.; LUI, J. J.:,PIRES, L. P. M.; CANCELLIER, L. L.; PELUZIO, J. M. Desempenho de genótipos de cana-de-açúcar em três cortes na região sul do estado do tocantins. Revista verde de agroecologia e desenvolvimento sustentável, v. 6, n.4, p.19-27, 2011.

CARVALHO, M. S. **Aplicação de métodos de análise espacial na caracterização de áreas de risco a saúde**. 149f.Tese. Universidade Federal do Rio de Janeiro. Programa de Pós-Graduação em Ciências em Engenharia Biomética. 1997.

CECÍLIO, R. A. **Precipitação**, 2006. Disponível em: <www.nedtec.ufes.br/prof/Roberto/disciplinas/manejo/03%20-%20Precipitação.pdf>. Acesso em 05 nov. 2012.

CEPEA. **Centro de Estudos Avançados em Economia Aplicada**. PIB do agronegócio. Disponível em: <http://www.cepea.esalq.usp.br/pib/>. Acesso em: 21 jun. 2012.

CHASCO, C. **Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales**. Madri (España): Consejería de Economía e Innovación Tecnológica, 430p. 2003.

CHATFIELD, C. Model uncertainty, data mining and statistical inference (with discussion). **Journal of the Royal Statistical Society Series A**, Londres/Inglaterra,158, p. 419–66, 1995.

CHI, G.; ZHU, J. Spatial regression models for demographic analysis.**Population Research andPolicy Review**,Berlin/Alemanha., 027, p. 17 -42, 2007.

CHRISTOFOLETTI, J. C. Considerações sobre a derivanas pulverizações agrícolas e seu controle. São Paulo:Teejet South América, 15p, 1999.

CONAB. Companhia Nacional de Abastecimento. **Soja-Brasil: série histórica de produção: safras 1976/77 a 2009/10.** 2010. Disponível em: <http://www.conab.gov.br/conteudos.php?a=1252&t=2&Pagina_objcmsconteudos=2#A_objcmsconteudos>. Acesso em:m21 jun. 2012.

COSTA, J. A. **Cultura da soja**. Porto Alegre: I. Manica, J. A. Costa, 1996, 233p.

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

DA SILVA, A. R. **Avaliação de modelos de regressão espacial para análise de cenários do transporte rodoviário de carga**. Dissertação de mestrado. Universidade de Brasília. Faculdade de tecnologia. 138p. 2006.

DERAL/SEAB - **Soja**. Curitiba: Departamento de Economia Rural/Secretaria do Estado da Agricultura e do Abastecimento do Paraná, 2000.

DOMINGUES, C. V.; FRANÇOSO, M. T. Aplicação de Geoprocessamento no Processo de Modernização da Gestão Municipal. RBC. **Revista Brasileira de Cartografia**,Rio de Janeiro/RJ, v. 60, n.1, p. 71-78, 2008.

DOS SANTOS, C. AL.; MACEDO, M. R. A.; MAIA, B. S. C.; DINIZ e SILVA, R. W.; NASCIMENTO, Y. K. O. Análise Exploratória de Dados Espaciais para Vítimas de Atentado Violento ao Pudor Contra Crianças e Adolescentes no Município de Belém no ano de 2009. **Anais XV Simpósio Brasileiro de Sensoriamento Remoto - SBSR**, Curitiba, PR, Brasil, 30 de abril a 05 de maio de 2011, INPE 3867p.

DOURADO NETO, D.; SPAROVEK, G.; FIGUEREDO JÚNIOR, L.G.M. de; FANCELLI A.L.; MANFRON, P.A.; MEDEIROS, S.L.P. Modelo para estimação da produtividade de grãos de milho deplecionada com base no balanço hídrico no solo. **Revista Brasileira de Agrometeorologia,** Santa Maria, v.12, n.2, p-359-367, 2004.

DRUCK, S.; CARVALHO, M.S.; CÂMARA, G.; MONTEIRO, A.V.M. (Ed). **Análise Espacial de Dados Geográficos**. Brasília: EMBRAPA, 2004, 209p.

EMBRAPA. Centro Nacional de Pesquisa do Soja. **Tecnologias de produção de soja – Paraná,** 2005. Londrina.

EMBRAPA, Empresa brasileira de pesquisa agropecuária**. Tecnologias de produção de soja - Paraná.** Londrina: Embrapa Soja, 2006. Disponível em: <http://www.cnpso.embrapa.br/download/tpsoja_2007_pr.pdf>. Acesso em: 24 jun. 2012.

FARIA, P. N. **Avaliação de métodos para determinação do número ótimo de *clusters* em estudo de divergência genética entre acessos de pimenta**. Dissertação de Mestrado. Universidade Federal de Viçosa. 67p. 2009.

FERRARIO, M. N.; SANTOS, A. A. L.; PARRÉ, J. L.; LOPES,R. Uma análise espacial do crescimento econômico do estado do paraná para os anos 2000 e 2004. **Revista Brasileira de Estudos Regionais e Urbanos**, Recife, v. 3, n. 1, p.154-177,2009.

FIGUEIREDO, D.C. Projeto GeoSafras - aperfeiçoamento do sistema de previsão de safras da CONAB. **Revista de Política Agrícola**, Brasília/DF,v.14, n.2, p.110-120, 2005.

FORNARI, M. R. **Análise e Desenvolvimento de um Novo Algoritmo de Junção Espacial para SGBD Geográficos**. 131f.Tese de Doutorado. Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre. 2006.

FOTHERINGHAM, A.S.; BRUNSDON, C.; CHARLTON, M. **Quantitative Geography**; Perspectives of Spatial Data Analysis. London Sage Publications. Ch4,.2000.288p.

FUJIMOTO, R. Y. **Diagnóstico Automático de Defeitos em Rolamentos Baseado em Lógica Fuzzy**. Dissertação. Escola Politécnica da Universidade de São Paulo. 2005. 181p.

FRANCO, C. R. **Novos Métodos de Classificação Nebulosa e de Validação de Categorias e suas Aplicações a Problemas de Reconhecimento de Padrões**. Dissertação. Universidade Federal do Rio de Janeiro. 2002.133p.

FREIRE, J. R. J.; VERNETTI, F. J.. A Pesquisa com soja, a seleção de rhizóbio e a produção de inoculantes no Brasil. **Pesquisa Agropecuária Gaúcha**, Porto Alegre/RS,v. 5, n.1, p. 117-126, 1999.

GETIS, A.; ORD, J.K.The analisys of spatial association by use of distance statistics.**Geographical Analisys**, Londres/Inglaterra,v.24, n.3, p.189-206, 1992.

GONÇALVES, R. P. Modelagem Conceitual de Bancos de Dados Geográficos para o Cadastro Técnico Multifinalitário em Municípios de Pequeno e Médio Porte. IP. **Informática Pública**,Belo Horizonte/MG, v. 61, n.3, p. 117-118, 2008.

GUIMARÃES, T. A.; ALVAREZ, V. M. P. Análise do Processo de Difusão Tecnológica para Cultivares de Soja da Embrapa no Paraná. **Revista de Política Agrícola**,Brasília/DF, v. 20, n.3, p. 19-34, 2011.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. **Cluster validity checking methods:** Part II. ACM SIGMOD, ACM Press, New York, NY, USA, v.31, n.3, p.19-27, September 2002.

INMET. **Manual de Observações Meteorológicas**. 3a ed., Brasília: Ministério da Agricultura e do Abastecimento, 1999, 84p.

JAIN, A. K.; MURTY, M.N.; FLYNN, P.J. **Data clustering**: a review. ACM Computing Surveys (CSUR), ACM Press, New York, NY, USA, v.31, n.3, p.264-323, September 1999.

JING, N.; CAI, W. **Analysis on the spatial distribution of logistics industry in the developed East Coast Area in China**.The Annals of Regional Science, [s.l.], v. 45, n. 2, p. 331-350, 2009.Disponível em: <http://www.springerlink.com/content/x2102235q143331u/>. Acesso em: 12 nov. 2010. DOI: 10.1007/s00168-009-0307-6

KAMPELL, S. A.; CÂMARA, G.; QUINTANILHA, J. A. Análise exploratória das relações espaciais do desflorestamento da Amazônia Legal. In: Gis Brasil 2000 - Congresso e Feira, 2000, Salvador. **Anais do GIS Brasil 2000**, 2000.

KASTER, M.; PALUDZYSZYN FILHO E.; KIIHL, R.A.S.; KRZYZANOWSKI, F.C., CARBONELL, S.A.M. Mejoramiento de la calidad fisiologica de la semilla de soja y, metodologia de evaluacion. In: **Conferencia Mundial de Investigacion em Soja**, 4. Buenos Aires. Argentina. Actas. A.J. Pascale (ed.). p.1106-1111, 1989.

KLOSOWSKI, E. S. Estimativa da produtividade de cultivares de soja por meio do modelo Soygro para Londrina, Estado do Paraná. **Revista Unimar**, Maringá, v. 19, n.3, p. 751-765, 1997.

LAW, M.H.; JAIN, A.K. **Cluster Validity by Bootstrapping Partitions**.East Lansing, Michigam, USA, 2003.

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**,Londres/Inglaterra, v. 82, n. 2, p. 175-201, 2003.

LESAGE, J. P. **Spatial econometrics**.[S.l.: s.n.], 1998, 284p. Disponível em: <http://www.spatial-econometrics.com/html/wbook.pdf>.Acesso em: 10 set. 2012

LESAGE, J. P. **The Theory and Practice of Spatial Econometrics**.University of Toledo, 1999, 309p.

LI, H. **Approximate Profile Likelihood Estimation for Spatial-Dependence Parameters**. Tese (Doutorado em Filosofia)- The Ohio State University Graduate School, Columbus OH,.2007. 156p.

LIMA, M. T.. **Uma proposta de modelagem para o risco de sofrer acidente de trabalho em Piracicaba/SP**. 86f. Dissertação. Escola Superior de Agricultura “Luiz de Queiroz”. 2010.

LOPES, S. B.; BRONDINO, N.C.M.; SILVA, A.N.R. Análise do desempenho de modelos de regressão espacial na previsão de demanda por transportes. In: CONGRESO PANAMERICANO DE INGENIERÍA DE TRÁNSITO Y TRANSPORTE, 14., 2006. PANAM 14., 2006, Las Palmas de Gran Canaria. **Anais...** Las Palmas de Gran Canaria: [s.n.], 2006.

MAGNUSSON, W. E.; MOURÃO, G. **Estatística sem matemática**: a ligação entre as questões e a análise. Curitiba: 2003.136p.

MALHOTRA, N. **Pesquisa de marketing**: uma orientação aplicada. Trad.Laura Bocco. 4 ed. Porto Alegre: Bookman, 2006.720p.

MATIAS, L.F. Sistemas para Informação. **Espaço & Geografia**,Brasília/DF, v.5, n.1. p.101-118. 2002.

MELO, A. S.; HEPP, L. U. **Ferramentas estatísticas para análises de dados provenientes de biomonitoramento**. Oecologia Brasiliensis 12(3):463-486, 2008.

MILAGRE, S., T. **Análise do Número de Grupos em Bases de Dados Incompletas Utilizando Agrupamentos Nebulosos e Reamostragem Bootstrap**. 157f. Tese de doutorado. Escola de Engenharia de São Carlos, da Universidade de São Paulo. 2008

MIRANDA, J. I. **Fundamentos de Sistemas de Informações Geográficas**. Brasília, DF, Embrapa Informação Tecnológica. Editora Perfil. 2005, 425p.

MENESES, H. B. **Interface Lógica em Ambiente SIG para Bases de Dados de Sistemas Centralizados de Controle do Tráfego Urbano em Tempo Real**.204f. Dissertação de Mestrado, Centro de Tecnologia, Universidade Federal do Ceará, Fortaleza,2003.

MESSNER, S. F.; ANSELIN, L.; BALLER, R.D.; HAWKINS, D. F.; DEANE, G.; TOLNAY, S. E. The spatial patterning of country homicide rates: an application of exploratory spatial data analysis. **Journal of Quantitative Criminology**, v. 15, n. 4, p. 423-450, 1999. Disponível em: <http://dx.doi.org/10.1023/A:1007544208712>. Acesso em: 12 nov. 2010. DOI 10.1023/A:1007544208712

MONTENEGRO, R. L. G. **Diversificações e Especializações Tecnológicas:** uma análise da atividade inovativa paulista. 138f. Dissertação (Mestrado em Economia Aplicada)- Universidade Federal de Juiz de Fora, Juiz de Fora, 2008.

MOTA, F. S. **Agrometeorologia**: uma seleção de temas e casos. 4. ed. Pelotas: Edição do autor, 2002, 340p.

NALON, F. R.; LISBOA FILHO, J.; BRAGA, J. L.; BORGES, Karla A. V.; ANDRADE, M. V. A.Applying the model driven architecture approach for geographic database design using a UML profile and ISO standards. **Journal of Information and Data Management**, Belo Horizonte/MG,v. 2, n. 2, p. 171-180, 2011.

OLIVEIRA, M. P. G. **Sistema Espacial de Apoio à Decisão:** modelos para análise do adensamento de atividades econômicas no espaço urbano. Dissertação de Mestrado, Escola de Governo de Minas Gerais da Fundação João Pinheiro, 1997.

PAL, N.R.; BEZDEK, J. C.**On cluster validity for the fuzzy c-means model**. IEEE Transactions on Fuzzy Systems, IEEE Computer Society, Washington, DC, USA, v.3, n.3, p.370-379, August 1995.

PEDRAZZI, J.A. **FACENS – Hidrologia Aplicada**, 1999. Disponível em: <http://www.facens.br/site/alunos/download/hidrologia>. Acesso em 05 nov. 2012

PINHEIRO, D. M.; SILVA, N. C**Desenvolvimento do programa wingis para análise de informações geográficas**. Espaço e Geografia (UnB), v. 12, n.2, p. 205-221, 2009.

QUEIROZ M. P. **Análise espacial dos acidentes de trânsito do município de Fortaleza** [on line]. 2003. Disponível: <http://www.det.ufc.br/petran/teses/tese27.pdf> Acesso em 05 nov. 2012.

ROCHA, C. H. B. **Geoprocessamento**. Tecnologia transdisciplinar. 2. ed. rev., atual e amp. Juiz de Fora, MG: Ed. do Autor, 2002. 220p.

ROCHA, M. M. **Modelagem da Dispersão de Vetores Biológicos com emprego da Estatística Espacial**.93f. Dissertação de Mestrado, Instituto Militar de Engenharia-IME, Rio de Janeiro, 2004.

RODRIGUES, M. A.; MONTEIRO, W. F.; CAMPOS, A. C; PARRE, J. L. Identificação e análise espacial das aglomerações produtivas do setor de confecções na região sul. In: **XXXVII Encontro Nacional de Economia**, 2009, Foz do Iguaçu - PR. Encontro Nacional de Economia ANPEC, 2009.

ROESE, A. D.; ROMANI, R. D.; FURLANETTO, C.; TANGARLIN, J. R.; PORTZ, R. L. Levantamento de doenças na cultura da soja, Glycine max (L.) Merrill, em municípios da região oeste do estado do Paraná. **Acta Scientiarum**,Maringá/MG, v. 23, n. 5, p. 1293-1297, 2001

ROMANI, L. A. S.; SANTOS, E. H. dos; EVANGELISTA, S. R. M.; ASSAD, E.D.; PINTO, H. S. Utilização de estações vizinhas para estimativa de temperatura e precipitação usando o inverso do quadrado da distância. In:**CONGRESSO BRASILEIRO DE AGROMETEOROLOGIA**, 13, 2003, Santa Maria.Situação atual e perspectivas da agrometeorologia: anais. Santa Maria: Unifra: SBA: UFSM, 2003. p. 717-718.

ROUBENS, M. Fuzzy clustering algorithms and their cluster validity.**European Journal of Operational Research,**Amsterdam/Holanda, 10 (3), p. 294–301, 1982.

RUSCHE, K. Quality of life in the regions: an exploratory spatial data analysis for West German labor markets. **Jahrbuch für Regional wissenschaft**, [s.l.], v. 30, n. 1, p. 1-22, 2009. Disponível em: < http://www.springerlink.com/content/x228008u0216176u/>. Acesso em: 12 nov. 2010. DOI 10.1007/s10037-009-0042-6

SALAME, C. W. **Análise espaço-temporal da ocorrência de queimadas e desmatamento no estado do Pará no período de 1999 a 2004**. 67f. Dissertação. Programa de Pós-Graduação em Matemática e Estatística (PPGME) da Universidade Federal do Pará. 2008.

SANTOS, C.B.; HINO P.; CUNHA, T.N.L VILLA, T.C.; MUNIZ, J.N. **Utilização de um Sistema de Informação Geográfica para descrição dos casos de tuberculose**. Bol Pneumol Sanitária 12(1):7-12, 2004.

SHARMA, S. **Applied multivariate techniques**. New York: John Wiley, 1996.493p.

SILVA, J. X. Geoprocessamento e análise ambiental. **Revista Brasileira de Geografia**, Rio de Janeiro: SBG, n.54, p. 47-61, jul/set 1992.

SILVA, A. R**. Avaliação de Modelos de Regressão Espacial para Análise de Cenários do Transporte Rodoviário de Carga**. Dissertação de Mestrado, Departamento de Engenharia Civil e Ambiental, Faculdade de Tecnologia, Universidade de Brasília. Brasília DF, 2006.

SILVA, A. T. A. da.**Aspectos meteorológicos e balanço hídrico em um aterro de resíduos sólidos urbanos.** 141f. Dissertação de mestrado. COPPE/UFRJ, 2008.

SILVA, T. O. da; CARVALHO, C. A. B.; CALIJURI, M. L.; LIMA, D. C. Sistemas de Informações Geográficas como Suporte à Gerência de Manutenção de Rodovias Vicinais não Pavimentadas. **Revista Brasileira de Cartografia** (Impresso),Rio de Janeiro/RJ, v. 61, n. 3, p. 301-309, 2009.

SILVA, V. P. R.; PEREIRA, E. R. R.; AZEVEDO, P. V.; SOUSA, F. A. S.; SOUSA, I. F. Análise da pluviometria e dias chuvosos na região nordeste do Brasil. **Revista Brasileira de Engenharia Agrícola e Ambiental (Online),**Campina Grande/PB, v. 15, n. 2, p. 131-138, 2011.

TEIXEIRA, A.; MATIAS, L.; NOAL, R.; MORETTI, E. Qual a melhor definição de SIG. **Fator GIS** – A revista do geoprocessamento. v.3, n.11, p.20-22. 1995.

TEXEIRA, R.. A. P.; BERTELLA, M. A. **Curva de Kuznets Ambiental para o Estado de Mato Grosso**: Modelagem Espacial. XIII Encontro Regional de Economia – ANPEC Sul. Porto Alegre – RS, 2010.

TOBLER, W. Cellular geography. In: S. Gale and O. G. (ed). **Philosophy in Geography.Dordrecht**, Reidel, 1979, p.379-386.

UNWIN A.; UNWIN D. **Spatial Data Analysis with Local Statistics**. Journal of the Royal Statistical Society: Series D (The Statistician) v. 47, n. 3, p 415–421, 1998.

VEENHOF, H. M.; APERS, P. M. G.; HOUTSMA, M. A. W. Optimization of spatial joins using filters.Advances in databases. **Lecture Notes in Computer Science**,Berlim/Alemanha,v. 940, p136-154, 1995.

VIANELLO, R. L.; ALVES, A. R. **Meteorologia básica e aplicações**. 3. ed. Viçosa: Ed. da Universidade Federal de Viçosa, 2001, 449p.

VILHENA, J. E. S.; CERQUEIRA, H. D. V.; BISPO, C. J. C.; GOMES, N. V.; ROCHA, E. J. P. **Análise da Precipitação no Período Chuvoso Sobre Cultivo de Soja para Paragominas-PA**. Centro regional de bibliografia agrometeorologica de la AR-III, Lima - Peru, p. 06 - 23, 01 dez. 2009.

VILLARDÓN, J. L. V. **Valoración y análisis de la diversidad funcional y su relación con los servicios ecosistémicos**. Informe Técnico No. 384. Centro Agronómico Tropical de Investigación y Enseñanza, CATIE Turrialba, Costa Rica, 2011.

YOKOO, S. C.; SILVEIRA, L. M. As relações entre a precipitação pluvial e a produção e a produtividade da soja no município de Campo Mourão - PR. In: **IV Congresso Brasileiro de Biometeorologia,** Ribeirão Preto. Sociedade Brasileira de Biometeorologia. Ribeirão Preto: Biomet. v. I. p. 124-128, 2006.

ZIBORDI, M. S.; CARDOSO, J. L.; VILELA FILHO, L. R.. Análise de aspectos socioeconômicos e tecnológicos da agropecuária na Bacia Hidrográfica do Rio Mogi Guaçu.**EngenhariaAgrícola**. [online],Jaboticabal/SP, v. 26, n.2, p. 644-653, 2006.

2 MODELO DE REGRESSÃO ESPACIAL PARA ESTIMATIVA DA PRODUTIVIDADE DA SOJA ASSOCIADA A VARIÁVEIS AGROMETEOROLÓGICAS NA REGIÃO OESTE DO ESTADO DO PARANÁ

RESUMO: Este trabalho apresenta o Modelo de Regressão Espacial Autoregressivo Misto (SAR) e Modelo do Erro Espacial (CAR) no intuito de investigar a associação entre a produtividade da soja e as variáveis agrometeorológicas relacionadas à precipitação pluvial, temperatura média e radiação solar global. O estudo foi realizado com os dados das safras dos anos agrícolas 2005/2006 a 2007/2008 da região oeste do estado do Paraná. Como os dados agrometeorológicos estão disponíveis apenas para oito municípios da região em estudo, as estimativas foram obtidas por meio do uso de Polígonos de Thiessen. A estimativa de parâmetros dos modelos ajustados foi obtida utilizando o método de Máxima Verossimilhança. A avaliação do desempenho dos modelos foi realizada com base no coeficiente de determinação (R^2), no máximo valor do logaritmo da função verossimilhança e no critério de informação bayesiano de Schwarz (*BIC*). Este estudo também permitiu verificar a correlação e a autocorrelação espacial entre a produtividade da soja e os elementos agrometeorológicos, por meio da análise espacial de área, usando de técnicas como o índice *I* de Moran Global e Local uni e bivariado e os testes de significância. O estudo pôde demonstrar que, por meio dos indicadores de desempenho utilizados, os modelos SAR e CAR ofereceram melhores resultados em relação ao modelo de regressão múltipla clássica.

PALAVRAS-CHAVE: Autocorrelação espacial; Estatística espacial de área, Modelos espaciais SAR e CAR.

SPATIAL REGRESSION MODEL FOR THE SOYBEAN CROP IN THE WESTERN REGIONOF THE STATE OF PARANA

ABSTRACT: This study presents the Spatial Lag Model (SAR) and Conditional Auto Regressive Model (CAR) in order to investigate the association between soybean yield and agrometeorological variables related to medium temperature and global solar radiation. The study was carried out with data from the agricultural years from 2005/2006 to 2007/2008 crops of West Region of the state of Parana. As agrometereological data was available only for eight cities of the region under study, the estimates were obtained through the use of Thiessen polygons. The estimation of parameters of the adjusted models was obtained using the method of maximum likelihood. The evaluation of the performance of models was held based on the coefficient of determination (R^2), maximum value of the logarithm of the likelihood function and Bayesian Information Criterion of Schwarz (*BIC*). This work also allowed to verify the correlation and the spatial autocorrelation between soybean yield and the agrometeorological factors by analyzing spatial area, through use of Global and Local uni- and bivariate and significance tests. Using such performance indicators, the study demonstrated that the SAR and CAR models offered better results than the classical multiple regression model.

KEY WORDS: Spatial autocorrelation; Moran's I index; Spatial statistics area; Spatial SAR and CAR models.

2.1 INTRODUÇÃO

A Estatística Espacial de Área (EEA) é um método estatístico que faz uso da referência geográfica no modelo, isto é, das coordenadas espaciais no processo de coleta, descrição e análise dos dados. Assim sendo, o interesse está centrado nos processos que ocorrem no espaço e os métodos empregados buscam descrever e analisar o comportamento desses processos.

A identificação e a quantificação das relações entre a produtividade das culturas agrícolas e os elementos agrometeorológicos têm sido tema de muitos estudos (FONTANA et al., 2001; DOURADO NETO et al., 2004). Segundo Berlato *et al.* (1992), os elementos críticos associados à produtividade agrícola são a radiação solar, a temperatura do ar e a precipitação. Após a identificação desses elementos e o período dentro do ciclo das culturas em que elas são limitantes, é possível a derivação de modelos de previsão de produtividade com boa acurácia.

Entretanto, é necessário, para facilitar esta análise, o uso de Sistemas de Informação Geográfica (SIG), já que contribuem fornecendo meios para visualização, manipulação, armazenamento e processamentos destas variáveis georreferenciadas. Aliado a isso, há o desenvolvimento de técnicas estatísticas para análise de dados espaciais de áreas, que, combinadas a um SIG, permitem a Análise Espacial de Área.

A análise espacial de área compõe um conjunto de procedimentos cujo objetivo é encontrar um modelo inferencial que incorpore explicitamente as relações espaciais constituintes de um fenômeno. Normalmente, a modelagem é iniciada pela análise exploratória de dados espaciais associada à visualização dos dados por meio de gráficos e mapas e, posteriormente, identificam-se padrões de dependência espacial das variáveis em estudo. Almeida *et al.* (2008) ressaltam que a Análise Exploratória de Dados Espaciais trata diretamente de efeitos decorrentes da dependência espacial e da heterogeneidade espacial.

Estendendo o estudo para análise espacial multivariada com dados de área, é possível utilizar modelos de regressão espacial linear por meio da regressão espacial múltipla, que permite constatar a relação entre uma variável dependente e diversas variáveis independentes envolvidas considerando a localização onde foram coletados os dados. Se constatada tal relação, busca-se ajustar um modelo estatístico que permita descrever uma determinada variável em relação às demais, considerando a localização dos dados (LOURENÇO; LANDIM, 2004).

Dentre as culturas de grande valor econômico, a soja se destaca como um dos principais produtos da agricultura brasileira, assumindo grande importância econômica nas exportações. Na última década, a região Sul foi, em média, a segunda maior região produtora de soja, respondendo por 39,8% da área plantada e 35,8% da produção brasileira. O estado do Paraná foi responsável por 19,2% da área plantada brasileira e 48,6% da

região Sul, produzindo 20,2% destes grãos do Brasil e 57,2% da região Sul. Enquanto que a produtividade média brasileira de soja foi de 2,60 t ha^{-1},o Paraná apresentou produtividade média de 2,73 t ha^{-1} no período (IBGE, 2011).

Nesse sentido, o objetivo deste trabalho foi analisar, espacialmente, para os anos-safras 2005/2006 a 2007/2008, na região oeste do estado do Paraná, a produtividade da soja (t ha^{-1}) e as variáveis agrometeorológicas, por meio dos índices de correlação e autocorrelação espacial. Além disto, foram gerados modelos de regressão espacial múltipla entre as variáveis estudadas.

2.2 MATERIAIS E MÉTODOS

A área de estudo deste trabalho (Figura 16) compreende quarenta e oito (n = 48) municípios (população) da região oeste do estado do Paraná. O período utilizado foi das safras 2005/2006 a 2007/2008 e as variáveis utilizadas foram: Produtividade da soja (t ha^{-1}) [Prod], Precipitação (mm) [Prec], Temperatura Média (ºC) [TMed] e Radiação Solar Global Média (W m^{-2}) [Rs], sendo os dados independentes. O período das safras utilizado para obtenção dos dados agrometorológicos diários foi de 1º de outubro do ano inicial da safra até 28 de fevereiro de seu ano final. A Precipitação utilizada foi obtida por meio da soma dos dados do período de cada safra e a Temperatura Média e Radiação Solar Global Média pela média aritmética. Os dados referentes à produtividade foram fornecidos pela SEAB (2010) e os dados agrometeorológicos (precipitação, temperatura média e radiação solar global média), pelo SIMEPAR (2010). Como os dados agrometeorológicos estão disponíveis apenas para oito municípios da região em estudo, sua estimação foi obtida por meio do uso de Polígonos de Thiessen (ANDRADE et al., 2008) e Junção Espacial (JACOX;SAMET, 2007).

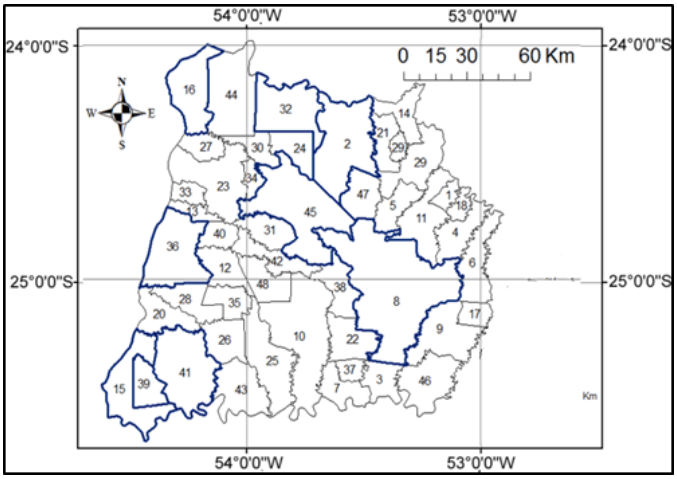


Figura 16 Região Oeste do Paraná, com destaque para os municípios com estações meteorológicas: (2) Assis Chateaubriand, (8) Cascavel, (15) Foz do Iguaçu, (16) Guaíra, (32) Palotina, (36) Santa Helena, (41) São Miguel do Iguaçu, e (45) Toledo.

A seleção dos anos-safra utilizados para este estudo foi baseada na identificação da média da produtividade de todos os municípios entre 2000/2001 e 2007/2008. Assim, foram selecionados os anos-safra com a menor média de produtividade (2005/2006), aquela com maior média de produtividade (2007/2008) e uma com a média mais próxima de todo o período (2006/2007).

Em relação aos dados agrometeorológicos, mais precisamente para os dados diários de precipitação, em alguns dias dos anos-safra escolhidos e para alguns municípios, não houve medição. Para estes dias e municípios, as mesmas técnicas de Junção Espacial e Polígonos de Thiessen foram utilizadas para a estimativa destes dados.

Para desenvolver a análise espacial de área foram utilizados os softwares ArcMap 9.3 (ESRI, 2011) e OpenGeoda 0.9.9.6 (OPENGEODA, 2011).

Para o desenvolvimento da modelagem estatística espacial, utilizou-se o índice de Moran global (*I*) e local (*LISA*), com a finalidade de estimar o nível de autocorrelação espacial entre as áreas (municípios). O *I* de Moran, que calcula a autocorrelação espacial global de cada variável, permitiu analisar se os dados estavam autocorrelacionados espacialmente, sendo determinado pela Equação 23.

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j w_{ij}}{S_0 \sum_{i=1}^{n} z_i^2} \qquad \text{Eq. (23)}$$

em que *n* é o número de populações (*n* polígonos); $z_i = (x_i - \bar{x})$ e $z_j = (x_j - \bar{x})$ para $i \neq j = 1,...,n$ são os valores observados das populações *i* e *j* centradas na média da variável *x* em estudo; w_{ij} é o elemento da matriz de proximidade *W*, *n x n*, a qual expressa a relação espacial entre as *n* populações e S_0 é definida pela Equação 24:

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \qquad \text{Eq. (24)}$$

De acordo com Câmara e Monteiro (2004) e Anselin *et al.*(2007), cada elemento w_{ij} da matriz de proximidade espacial *W*, representa uma medida de proximidade entre as populações (polígonos) A_i e A_j, a qual pode ser calculada a partir de um dos seguintes critérios:

- Critério da Distância entre Centroides

  w_{ij} = 1, se o centroide de A_i está a uma determinada distância de A_j; caso contrário w_{ij} = 0; para i ≠ j = 1,..., n.

- Critério de contiguidade (torre, rainha e bispo)

w_{ij} = 1, se A_i compartilha um lado comum com A_j, caso contrário w_{ij} = 0; para i ≠ j = 1,..., n.

- Critério de número de vizinhos mais próximos

  $w_{ij} = A_{ij}/A_i$, onde A_{ij} é o comprimento da fronteira entre A_i e A_j e A_i é o perímetro de A_i; para i ≠ j = 1,..., n.

Com a estatística *I* de Moran como medida de dependência espacial é possível realizar o teste de hipótese, se existe ou não autocorrelação espacial. A estatística *I* de Moran tem valor esperado *E(I) = –[1/(n-1)]* sob a hipótese de não-existência de autocorrelação (H_0). Dessa maneira, os valores de *I* que excederem *–[1/(n-1)]* indicam autocorrelação espacial positiva. Consequentemente, valores de *I* abaixo do valor esperado sinalizam uma autocorrelação negativa (DRUCK et al.,2004).

A autocorrelação espacial positiva, no contexto deste estudo, pode ser representada pela similaridade entre as populações (polígonos), ou seja, os municípios que possuem uma alta/baixa produtividade de soja (t ha^{-1}) tendem a ser vizinhos de populações (polígonos) que também possuam uma alta/baixa produtividade de soja. Em contrapartida, a autocorrelação espacial negativa indica que existe uma dissimilaridade entre os valores de produtividade da soja e de sua localização espacial. Assim, os municípios (polígonos) com alta/baixa produtividade de soja são vizinhos de municípios (polígonos) que apresentam um baixo/alto valor para a mesma variável.

A autocorrelação espacial local busca captar padrões de associação, pois embora seja capaz de apontar a tendência geral de agrupamento dos dados, o *I* de Moran é uma medida global e por isso não revela padrões locais de associação espacial. A autocorrelação local pode ser calculada pela estatística *I* de Moran local, também conhecido como Indicador Local de Associação Espacial (*LISA*). Esta estatística deve satisfazer aos seguintes critérios: um indicador *LISA* deve possuir para cada município, uma indicação de agrupamentos espaciais significantes de valores similares em torno do município; o somatório dos *LISA*s para todas os municípios é proporcional ao *I* de Moran Global (ANSELIN, 1995).

Segundo Celebioglu e Dall'erba (2009), a estatística *LISA*, ou índice *I* de Moran local, pode ser especificado pela Equação 25:

$$I_i = \frac{x_i - \mu}{\sigma_0^2} \sum_{j=1}^{n} w_{ij}(x_j - \mu), \; i = 1,...,n \qquad \text{Eq. (25)}$$

em que, σ_0^2 a variância populacional da variável em estudo dos *n* municípios; x_i é a observação de uma variável de interesse no município *i* para $i = 1,...,n$; μ é a média dos *n* municípios (populações).

A estatística *LISA* I_i, para i =1,..., n, pode ser interpretada da seguinte forma: valores positivos de I_i significam que existem agrupamentos espaciais com valores similares (alto ou baixo); valores negativos significam que existem agrupamentos espaciais com valores diferentes entre as regiões e seus vizinhos.

A significância do índice de Moran Global e Local, segundo Nicolau *et al.* (2009), pode ser abordada por um teste de pseudo-significância que gera diferentes permutações dos valores de atributos associados às zonas, onde cada permutação produz um novo arranjo espacial dos valores redistribuídos entre as áreas, sendo a sua significância obtida a partir de uma distribuição empírica da estatística *I* de Moran. Se o valor do índice *I* de Moran medido corresponder a um "extremo" da distribuição simulada, então se trata de um evento com significância estatística.

Para o estudo de duas variáveis espacialmente georreferenciadas, o índice de Moran bivariado denotado como I_{xy} é um índice de correlação espacial entre duas variáveis (X e *Y*), cada uma sendo obtidas nos *n* municípios. O índice Moran bivariado I_{xy} é obtido da forma mostrada na Equação 26:

$$I_{xy} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} u_i z_j w_{ij}}{S_0\sqrt{S_u^2 S_z^2}} \qquad \text{Eq. (26)}$$

em que *n* é o número de municípios(populações); $z_j = (x_j - \bar{x})$ e $u_i = (y_i - \bar{y})$ são os valores observados centrados nas médias das variáveis *X* e *Y* em estudo, respectivamente; w_{ij} é o elemento da matriz proximidade *W*, *n*x*n*; S_0 é definido na Equação 19; $S_u^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$ e $S_z^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ as respectivas variâncias de *Y* e *X* (ANSELIN et al., 2003).

Um modelo de regressão baseia-se na relação entre duas ou mais variáveis, de forma que uma delas (variável dependente) possa ser explicada em função de outra ou outras variáveis (variáveis independentes). No caso de dados espaciais, havendo correlação espacial, o modelo gerado deve incorporar a estrutura espacial, já que a dependência entre as observações afeta a capacidade de explicação do modelo (CÂMARA et al., 2002).

Para introduzir explicitamente efeitos espaciais em modelos de regressão espacial, há diferentes formas, sendo a mais simples denominada modelo com efeitos espaciais globais, que busca capturar a estrutura de correlação espacial em apenas um parâmetro e adicioná-lo no modelo de regressão. Dessa maneira, têm-se duas alternativas. A primeira delas é pelo modelo SAR, que atribui à variável resposta *Y* a autocorrelação espacial ignorada. Formalmente, ela é definida pela Equação 27:

$$Y = X\beta + \rho WY + \varepsilon \qquad \text{Eq. (27)}$$

em que: *Y* é o vetor *n* x 1 das respostas dos *n* municípios, *X* é a matriz *n* x *p*, de $X_1,...,X_p$ variáveis explicativas ou covariáveis, β é um vetor de *p* x 1 coeficientes de regressão desconhecidos a serem estimados, *W* é a matriz de proximidade espacial *n* x *n*; o produto *WY* expressa a dependência espacial em *Y*; ρ é o coeficiente espacial autorregressivo, ε é o vetor de erros aleatórios, *n* x 1 com média zero e variância constante não correlacionada (BAILEY;GATRELL, 1995).

A segunda alternativa, como consta em Druck*et al.* (2004), é pelo CAR, que considera os efeitos espaciais como um ruído, isto é, como um fator a ser removido, e é descrito pela Equação 28:

$$\begin{aligned} Y &= X\beta + \varepsilon, \\ \varepsilon &= \lambda W_{\varepsilon} + \xi \end{aligned} \qquad \text{Eq. (28)}$$

em que: W_ε é a componente do erro com efeitos espaciais, *n* x 1, λ é o coeficiente autorregressivo, onde a hipótese nula para a não-existência de autocorrelação é H_0:λ=0; e ξ é o vetor n x 1 do componente do erro com média zero, variância constante e não-correlacionada (ruído).

As estimativas dos parâmetros das Equações 27 e 28 são obtidas pelo método de estimação da Máxima Verossimilhança (MV).

A avaliação do desempenho dos modelos foi realizada com base no coeficiente de determinação - R^2, no máximo valor do logaritmo da função verossimilhança – *MVLFV* (MCBRATNEY; WEBSTER, 1986) e no critério de informação bayesiano – *BIC* (SCHWARZ, 1978),que é definido na Equação 29. Segundo Kuha (2004), o critério *BIC* apresenta um melhor desempenho que o critério de interferência Akaike.

$$BIC = -2\ln(L(\theta)) + (p+1)\ln(n) \qquad \text{Eq. (29)}$$

em que ln($L(\theta)$) é o logaritmo da função verossimilhança $L(\theta)$, *p* é a dimensão do vetor de parâmetros, *n* é o número de população e θ o vetor de parâmetros desconhecidos para cada modelo em estudo. O objetivo de utilizar mais de uma estatística para avaliação do desempenho dos modelos baseou-se em se ter mais critérios de avaliação da performance dos modelos, uma vez que apenas o R^2 nem sempre é suficiente para avaliar a qualidade dos ajustes (BROWN et al., 1999).

2.3 RESULTADOS E DISCUSSÃO

Na Tabela 2 são apresentados os índices de autocorrelação espacial, *I* de Moran Global de cada variável em estudo e seus respectivos níveis de significância (entre parênteses), segundo os critérios de contingência torre, distância entre centroides e vizinho mais próximo para a matriz de proximidade *W*. Pode-se observar que os níveis descritivos (p-valor) são menores que 0,05 (nível de significância), concluindo que todas as variáveis têm autocorrelação espacial significativa a 5% de probabilidade.

Tabela 2 Índice *I* de Moran Global de autocorrelação espacial para as variáveis em estudo.

	Índice Global de Moran								
Variáveis	**Contiguidade (Torre)**			**Distância entre centroides**			**Vizinhos mais próximos**		
	2005/ 2006	**2006/ 2007**	**2007/ 2008**	**2005/ 2006**	**2006/ 2007**	**2007/ 2008**	**2005/ 2006**	**2006/ 2007**	**2007/ 2008**
Prod	0,5632 (0,001)	0,4900 (0,001)	0,3112 (0,001)	0,6187 (0,001)	0,5094 (0,001)	0,2968 (0,004)	0,5539 (0,001)	0,5193 (0,001)	0,3052 (0,001)
Prec	0,2834 (0,002)	0,2703 (0,001)	0,2203 (0,009)	0,3237 (0,007)	0,2904 (0,005)	0,2224 (0,031)	0,2087 (0,011)	0,1658 (0,032)	0,0975 (0,11)
TMed	0,8359 (0,001)	0,8632 (0,001)	0,8233 (0,001)	0,8521 (0,001)	0,8924 (0,001)	0,8952 (0,001)	0,7875 (0,001)	0,8313 (0,001)	0,8401 (0,001)
Rs	0,7925 (0,001)	0,7641 (0,001)	0,7916 (0,001)	0,8883 (0,001)	0,8116 (0,001)	0,8257 (0,001)	0,8285 (0,001)	0,7482 (0,001)	0,7364 (0,001)

Prod: produtividade de soja (t ha^{-1}); Prec: precipitação (mm); TMed: temperatura média do ar (ºC); Rs: radiação solar global média (W m^{-2}). Entre parênteses tem-se o nível descritivo p-valor.

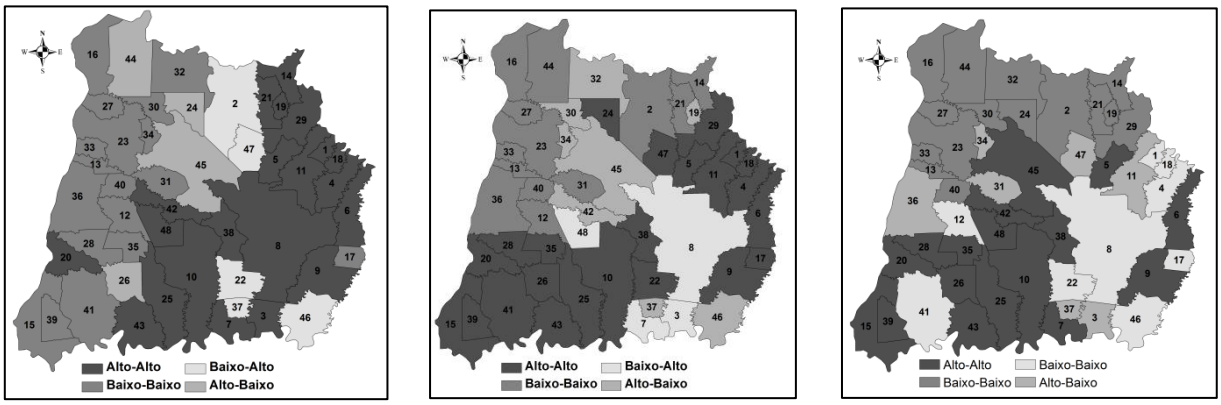
Com exceção da TMed, as maiores autocorrelações espaciais foram encontradas para o ano-safra 2005/2006, independente da matriz de proximidade utilizada. Em contrapartida, dentre as variáveis estudadas, independente do ano-safra, a TMed foi a que apresentou os maiores valores de autocorrelação espacial (*I*> 0,79), seguido da Rs (*I*> 0,73), o que já era esperado, já que, dentre as variáveis agrometeorológicas, estas são as que apresentam a menor variabilidade espacial. A Prec apresentou os menores valores de autocorrelação espacial (*I*< 0,33), o que também se justifica, já que ela é a que apresenta a maior variabilidade espacial, devido, principalmente, à ocorrência de precipitações convectivas, ou “chuvas de verão”, que produzem grandes volumes de água em pequenas áreas (DEPPE et al., 2006 e 2007). Já a Prod apresentou uma autocorrelação espacial moderada (*I*), variando entre 0,30 e 0,62.

Dado o fato de que todas as variáveis em estudo, para os três anos-safra, tiveram autocorrelação espacial positiva significativa (Tabela 2), prosseguiu-se em busca de modelos que incorporassem essa informação.

A Figura 17 apresenta o mapa de espalhamento de Moran Global para a Prod, usando o critério de contiguidade (torre) para a matriz de proximidade (*W*). A escolha por este critério se deu pelo fato de ele ser mais simples, pois considera apenas as regiões que possuam uma fronteira comum (PIMENTEL; HADDAD, 2004). Nota-se que a maior parte dos municípios da região oeste do estado do Paraná indica associação espacial positiva.

Estes municípios são identificados na legenda como Alto-Alto (valores positivos, médias positivas) e Baixo-Baixo (valores negativos, médias negativas). Os demais municípios, indicados pelas tonalidades claras do cinza, podem ser vistos como municípios que não seguem o mesmo processo de dependência espacial dos demais municípios.

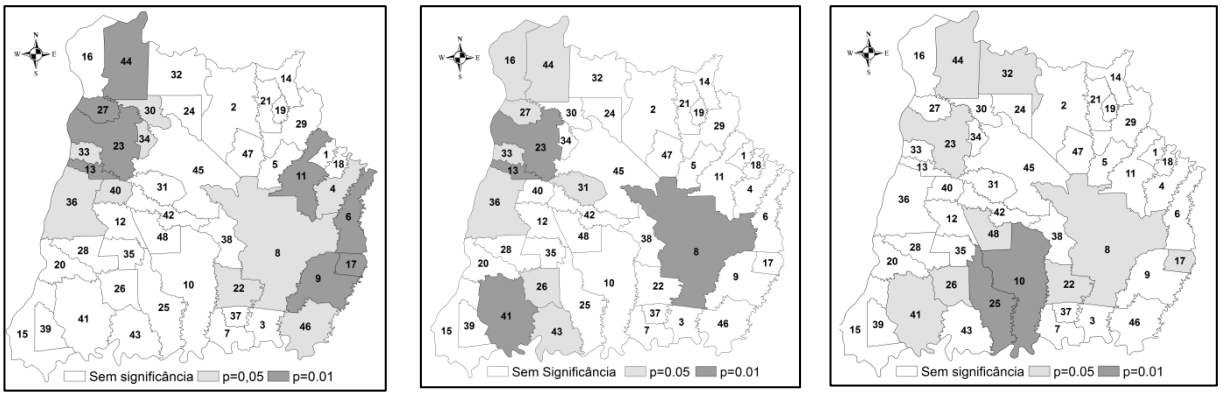
Para estimar a significância dos índices globais de Moran da Tabela 2, foi realizado o teste de pseudo-significância, com 999 permutações. Como o valor do índice de Moran medido (I = 0,5632 (2005/2006); I = 0,4900 (2006/2007); e I = 0,3112 (2007/2008)) corresponde a um extremo (à direita) da distribuição simulada, então se trata de um valor com significância estatística. Além disso, verificou-se que a esperança do índice global de Moran é de -0,0213 e o p-valor = 0,001 (nível descritivo), para as três safras, rejeitando a hipótese de ausência de autocorrelação espacial a 5% de significância.



2005/2006 (a) 2006/2007 (b) 2007/2008 (c)

Figura 17 Mapa de espalhamento de Moran Global para a variável Produtividade da Soja.

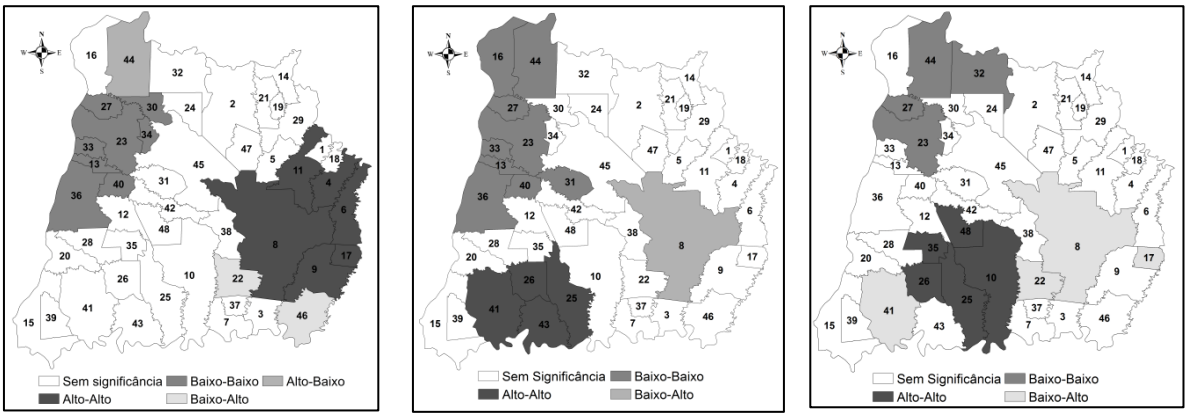
Conforme Druck*et al.* (2004), para um grande número de áreas, como neste caso com 48 municípios, é importante utilizar indicadores de associação espacial local que permitam identificar melhor os agrupamentos. Calculando-se o Índice de Moran Local (*LISA*) foi possível classificar os municípios em função do nível de significância dos valores de seus índices locais. Na Figura 18, pode-se identificar os municípios não-significativos ao nível de 5% de probabilidade (sem cor). Já os municípios que apresentaram *LISA* significativo são ilustrados em diferentes tons de cinza, tanto a 5% significância (tonalidades claras do cinza) como a 1% de significância (tonalidades escuras do cinza) e são aqueles com características próprias, que merecem uma análise detalhada.



2005/2006 (a) 2006/2007 (b) 2007/2008 (c)

Figura 18 Indicador local de autocorrelação espacial (*LISA*) para a variável Produtividade de Soja.

A Figura 19 mostra que ao sul da região em estudo, para os anos-safra de 2006/2007 e 2007/2008, existe um agrupamento com alta produtividade da soja (t ha^{-1}) significativa (1 e 5%). Para o ano-safra 2005/2006 (Figura 19a), ano de menor produtividade média de soja entre 2000/2001 e 2007/2008, a alta produtividade da soja (t ha^{-1}) foi significativa (1 e 5%) a leste da região estudada. Os agrupamentos com baixa produtividade da soja (t ha^{-1}) significativa, para as três safras, se encontram a oeste e nordeste. Além disto, para o ano-safra 2005/2006 (Figura 19a), diferentemente do que ocorreu para as demais safras estudadas, houve um município (44: Terra Roxa) que teve alta produtividade de soja, porém, municípios vizinhos apresentaram baixa produtividade (Alto-Baixo). Os municípios de Lindoeste (22) e Três Barras do Paraná (46) foram categorizados como de Baixo-Alto e, portanto, tiveram baixa produtividade média de soja, e fazem fronteira com municípios que apresentaram alta produtividade. Constata-se ainda que, nas três safras, municípios que se encontram na área central (38:Santa Tereza do Oeste, 42:São Pedro do Iguaçu, 45:Toledo e 47:Tupãssi) no norte e nordeste (2:Assis Chateaubriand, 5:Cafelândia, 14:Formosa do Oeste, 19:Iracema do Oeste, 21:Jesuítas e 29:Nova Aurora) e na região sul (3:Boa Vista da Aparecida, 7:Capitão Leônidas Marques e 17:Ibema) da área estudada, encontram-se na categoria de Sem Significância estatística; todavia, estes municípios fazem vizinhança com outros que possuem autocorrelação, seja ela positiva ou negativa. Com esta análise, ressalta-se a importância de um estudo que possa aproximar os municípios sem significância estatística em suas autocorrelações com os que possuem autocorrelação positiva e verificar se os que fazem vizinhança com os que possuem autocorrelação espacial negativa influenciam a falta de significância de seus vizinhos.



2005/2006 (a) 2006/2007 (b) 2007/2008 (c)

Figura 19 Mapa de espalhamento de Moran local para a variável Produtividade da soja.

A correlação espacial entre as variáveis agrometeorológicas pode ser verificada pelo coeficiente *I* de Moran bivariado dado na matriz da Tabela 3.

Tabela 3 Índice *I* de Moran Bivariado e nível descritivo (p-valor).

Variáveis	Prod			Prec			TMed		
	2005/ 2006	**2006/ 2007**	**2007/ 2008**	**2005/ 2006**	**2006/ 2007**	**2007/ 2008**	**2005/ 2006**	**2006/ 2007**	**2007/ 2008**
Prec	**-0,2395** (0,004)	-0,1081 (0,167)	0,0706 (0,154)						
TMed	**-0,3866** (0,001)	**-0,1820** (0,022)	-0,0516 (0,389)	**0,1952** (0,021)	**0,2146** (0,013)	0,1069 (0,091)			
Rs	**-0,4204**(0,001)	-0,1537 (0,061)	**-0,2197** (0,011)	**0,2205** (0,012)	0,1177 (0,078)	0,0879 (0,107)	**0,4197** (0,001)	**0,4181** (0,001)	0,0333 (0,273)

Prod: produtividade de soja (t ha^{-1}); Prec: precipitação (mm); TMed: temperatura média do ar (ºC); Rs: radiação solar global média (W m^{-2}). Entre parênteses tem-se o nível descritivo p-valor. Os valores em negrito representam correlações significativas a 5% de probabilidade.

Pela análise da Tabela 3, verificou-se que houve correlação significativa (5%) entre todas as variáveis estudadas para o ano-safra 2005/2006, o que não ocorreu para os demais anos estudados. Houve uma correlação espacial positiva e significativa (diretamente proporcional) entre as variáveis: Prec e TMed para os anos-safra de 2005/2006 (0,1952) e 2006/2007 (0,2146); Prec e Rs para o ano-safra 2005/2006 (0,2205) e para TMed e Rs para os anos-safra de 2005/2006 (0,4197) e 2006/2007 (0,4181), onde foram encontrados os maiores valores, o que se justifica, já que a TMed tende a ser maior em dias com maior incidência solar. De acordo aos dados obtidos, esta análise indica a rejeição da hipótese nula, que refere-se à não-existência de correlação espacial bivariada.

Foi encontrada correlação espacial negativa e significativa (inversamente proporcional) entre a Prod e todos os elementos agrometeorológicos para o ano-safra 2005/2006 (Prec = 0,2395, TMed = -0,3866, Rs = -0,4204). Para 2006/2007, somente para Prod e TMed (-0,1820) e em 2007/2008 entre Prod e Rs (-0,2197). Isto comprova que os

elementos agrometeorológicos escolhidos realmente têm impacto na produtividade da soja, como descreve a literatura.

Considerando-se a existência de autocorrelação espacial, o modelo completo SAR foi estimado por Prod = $\hat{\beta}_0$+ $\hat{\beta}_1$Prec + $\hat{\beta}_2$TMed + $\hat{\beta}_3$Rs + $\widehat{\rho}$*W*Prod, cujos parâmetros foram estimados por MV.

Identificou-se que para os anos-safra de 2005/2006 e 2006/2007 o modelo que melhor explicou a produtividade da soja foi o modelo sem a Prec. Para a safra de 2007/2008, o melhor modelo foi aquele que não fez uso da TMed; porém, verificou-se que, por meio do coeficiente de determinação (R^2), o melhor modelo explica muito pouco esta produtividade. Os modelos identificados como melhores, de acordo com os critérios de avaliação de desempenho utilizados, são apresentados na Tabela 4.

Tabela 4 Resumo de modelos ajustados e da análise com os parâmetros obtidos para o Modelo SAR.

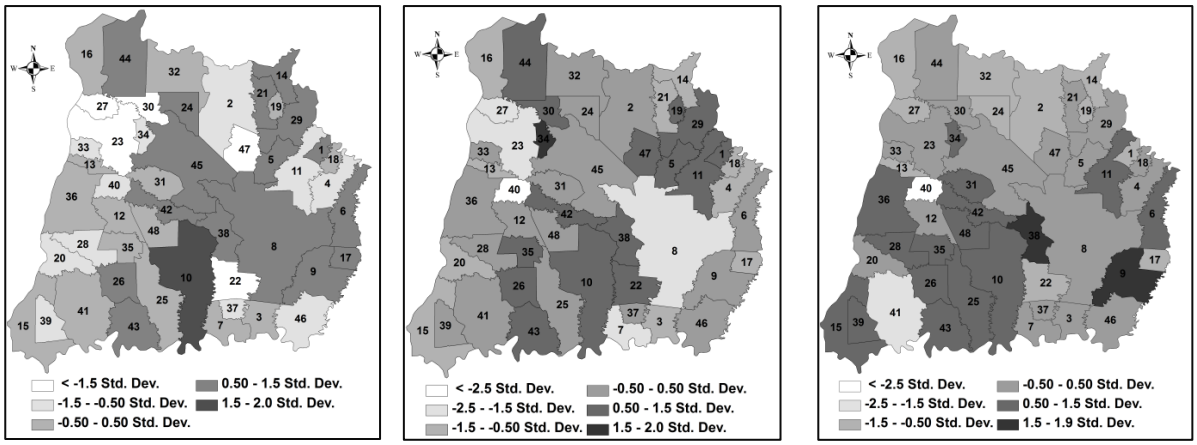
Anos-Safra	Variável	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\rho}$	R^2	MVLFV	*BIC*
2005/ 2006	Prod	2,27187		-0,034983	-0,002141	0,64151	50,58%	-20,13	55,74
2006/ 2007	Prod	1,57486		-0,020036	-0,000576	0,70879	47,49%	1,30	12,89
2007/ 2008	Prod	1,81152	0,000046		-0,000508	0,49698	24,46%	9,35	-3,22

$\hat{\beta}_0$: estimativa do coeficiente linear; $\hat{\beta}_1$: estimativa do parâmetro associado à precipitação pluvial (mm); $\hat{\beta}_2$: estimativa do parâmetro associado à temperatura média do ar (ºC); $\hat{\beta}_3$: estimativa do parâmetro associado à radiação solar global média (W m^{-2}); $\hat{\rho}$: estimativa do coeficiente exponencial autoregressivo; R^2: coeficiente de determinação; MVLFV: máximo valor do logaritmo da função verossimilhança; *BIC*: critério de informação bayesiano.

Constata-se assim que, para todos os anos-safra estudados, a Rs foi utilizada nos modelos. A Prec foi utilizada apenas no modelo SAR de 2007/2008 e os baixos valores de R^2 se justificam porque tanto a Prec como a Rs tiveram parâmetros com valores muito baixos. Para os dois primeiros anos-safra avaliados, a TMed foi a que teve menor peso nas estimativas da produtividade (menores valores dos parâmetros). O efeito da temperatura é oposto à produtividade, ou seja, quanto maior a temperatura média durante o ciclo de desenvolvimento da cultura, menor foi a produtividade média alcançada pelos municípios. O coeficiente espacial autorregressivo estimado, $\widehat{\rho}$, foi significante a 5% de probabilidade para os modelos identificados para os três anos-safra estudados, tendo seu maior valor encontrado para o modelo do ano-safra de 2006/2007, corroborando a existência de autocorrelação.

Os resíduos do modelo SAR estão espalhados aleatoriamente em torno da sua média (Figura 20) e, de acordo com o teste de Anderson-Darling, a 5% de significância, têm

distribuição normal, com p-valor = 0,214 (2005/2006); p-valor = 0,022 (2006/2007) e p-valor = 0,298 (2007/2008). O índice global de Moran para esses resíduos foi de 0,0001 (com E(*I*) = -0,0213 e p-valor = 0,391 para 2005/2006); de 0,0319 (com E(*I*) = -0,0213 e p-valor = 0,283 para 2006/2007); e de -0,0132 (com E(*I*) = -0,0213 e p-valor = 0,552), podendo serem considerados iguais a zero ao nível de significância de 5%. Isso indica que a inclusão da componente *WY* nos modelos praticamente eliminou a autocorrelação espacial, fazendo com que a inclinação da reta, que representa o índice de Moran, no diagrama de espalhamento, fosse muito pequena. Portanto, o modelo SAR, permitiu gerar resíduos distribuídos aleatoriamente pela área de estudo, como pode ser observado na Figura 20, que representa o mapa dos resíduos padronizados, gerado pelo método do desvio padrão resultante da aplicação do modelo SAR.



2005/2006 (a) 2006/2007 (b) 2007/2008 (c)

Figura 20 Mapa de espalhamento de Moran local para os resíduos padronizados do modelo SAR.

O modelo completo CAR foi estimado por Prod = $\hat{\beta}_0$ + $\hat{\beta}_1$Prec + $\hat{\beta}_2$TMed + $\hat{\beta}_3$Rs + $\hat{\lambda}W_\varepsilon$. Verificou-se que para as safras de 2006/2007 e 2007/2008, o modelo que melhor explicou a produtividade da soja foi o modelo sem a TMed. Para a safra de 2005/2006 o melhor modelo identificado foi o que não fez uso da Rs. Os modelos identificados como melhores, de acordo com os critérios de performance dos modelos utilizados, são apresentados na Tabela 5.

Tabela 5 Resumo de modelos ajustados e da análise com os parâmetros obtidos para o Modelo CAR.

Anos-Safra	Variável	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\hat{\lambda}$	R^2	MVLFV	*BIC*
2005/2006	Prod	3,00721	-0,000008	-0,03666		0,70677	50,61%	0,51	53,39
2006/2007	Prod	3,60973	-0,000011		-0,001586	0,72680	47,99%	1,28	9,05
2007/2008	Prod	3,36910	0,0000306		-0,000299	0,49650	22,94%	8,88	-6,14

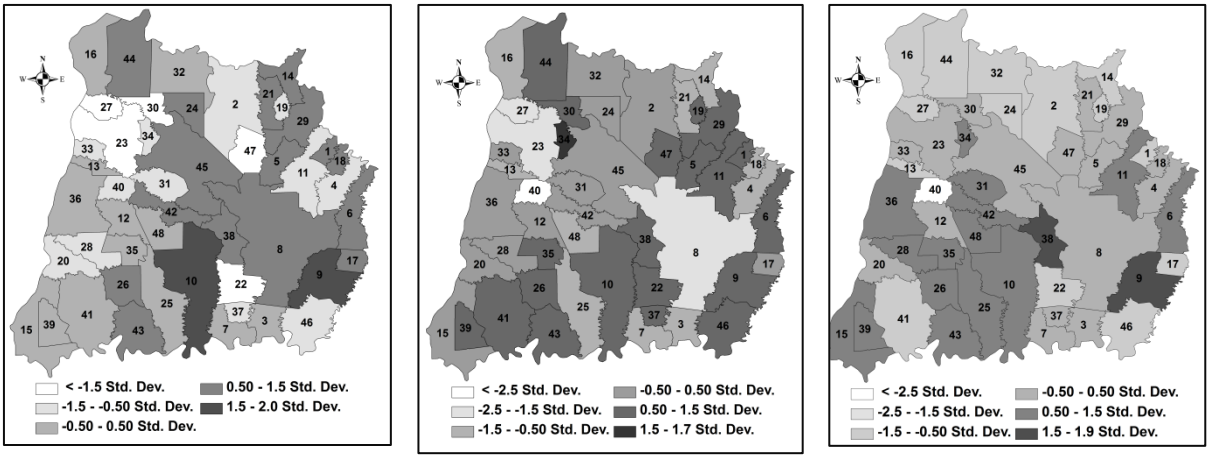
$\hat{\beta}_0$: estimativa do coeficiente linear;$\hat{\beta}_1$: estimativa do parâmetro associado à precipitação pluvial (mm); $\hat{\beta}_2$: estimativa do parâmetro associado à temperatura média do ar (ºC); $\hat{\beta}_3$: estimativa do parâmetro associado à radiação solar global média (W m^{-2});$\hat{\lambda}$: estimativa do coeficiente exponencial autoregressivo; R^2: coeficiente de determinação; MVLFV: máximo valor do logaritmo da função verossimilhança;*BIC*: critério de informação bayesiano.

Embora diferentes elementos agrometeorológicos tenham sido utilizados para os modelos SAR e CAR, para os anos-safra 2005/2006 e 2006/2007, diferentemente do que ocorreu para 2007/2008, em que os mesmos elementos foram selecionados, para o melhor modelo escolhido, verificou-se que em termos de R^2, praticamente não houve diferenças, valendo, portanto, as análises realizadas para o modelo SAR. O coeficiente espacial autorregressivo estimado, $\hat{\lambda}$, foi significante a 5% de probabilidade para os modelos identificados para os três anos-safra estudados, tendo seu maior valor encontrado para o modelo do ano-safra de 2006/2007.

Em relação aos baixos valores identificados pelo coeficiente de determinação, tanto para o SAR, como para o CAR, Brown *et al.* (1999) apontaram limitações em relação a esta estatística, atribuindo deficiências do R^2 à ocorrência de heterocedasticidade. Desta forma, utilizou-se *BIC* e *MVLFV*. Para este estudo os critérios *MVLFV* para o modelo SAR e o *BIC* para o modelo CAR foram os que melhor avaliaram os modelos identificados.

Como no estudo dos modelos SAR, os resíduos dos modelos CAR estão espalhados aleatoriamente em torno da média zero e tiveram distribuição normal de probabilidades a 5% de significância, pelo teste de Anderson-Darling, (p-valor = 0,275 para 2005/2006; p-valor = 0,039 para 2006/2007; p-valor = 0,190 para o ano-safra 2007/2008).

O índice global de Moran para esses resíduos foi de 0,0049 (com E (*I*) = -0.0213 e p-valor = 0,369 para 2005/2006), 0,0492 (com E(*I*) = -0,0213 e p-valor = 0,222 para 2006/2007) e 0,0029 (com E(*I*) = -0,0213 e p-valor = 0,388 para 2007/2008), podendo ser considerado igual a zero, ao nível de 5% de significância, ou seja, os resíduos dos modelos estimados não são autocorrelacionados espacialmente. Assim, a inclusão das componentes W_ε e ξ também eliminaram a autocorrelação espacial, ou seja, permitindo a geração de resíduos distribuídos aleatoriamente pela área de estudo, como ilustra a Figura 21.



2005/2006 (a) 2006/2007 (b) 2007/2008 (c)

Figura 21 Mapa dos resíduos padronizados da regressão espacial gerada pelo modelo Spatial Error, considerando o método do desvio-padrão.

Na Tabela 6 é apresentado o resumo de modelos ajustados por um modelo de regressão múltipla clássica. Observa-se que, quando comparados aos modelos SAR e CAR,

apresentaram pior desempenho, em função de não considerar a dependência espacial dos dados.

Tabela 6 Resumo de modelos ajustados e da análise com os parâmetros obtidos para o Modelo de Regressão Múltipla Clássica.

Anos-Safra	Variável	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	R^2	MVLFV	*BIC*
2005/2006	Prod	7,97095	-0.000070	-0.23572		24,15%	-27.64	70,76
2006/2007	Prod	4.77207	0.0000043	-0.06044	-0.000674	5,02%	-9,35	34,19
2007/2008	Prod	4.22996	0.0000712	-0.02404	-0.001052	6,95%	5,87	3,75

$\hat{\beta}_0$: estimativa do coeficiente linear; $\hat{\beta}_1$: estimativa do parâmetro associado à precipitação pluvial (mm); $\hat{\beta}_2$: estimativa do parâmetro associado à temperatura média do ar (ºC); $\hat{\beta}_3$: estimativa do parâmetro associado à radiação solar global média (W m^{-2}); R^2: coeficiente de determinação; MVLFV: máximo valor do logaritmo da função verossimilhança; *BIC*: critério de informação bayesiano.

2.4 CONCLUSÕES

Verificou-se autocorrelação espacial da produtividade da soja com os elementos agrometeorológicos nas safras de 2005/2006, 2006/2007 e 2007/2008por meio da análise exploratória espacial por áreas, usando técnicas estatísticas como o índice *I* de Moran univariado. No caso global, houve autocorrelação espacial entre as regiões, detectando áreas de agrupamentos e áreas de transição. No caso local, por meio do índice *LISA*, áreas com características individuais correlacionadas também foram possíveis de identificação.

Existe correlação espacial bivariada direta entre precipitação e temperatura, precipitação e radiação solar, temperatura e radiação solar e inversa entre produtividade e as três variáveis agrometeorológicas, embora estas correlações não ocorram em todos os anos-safra.

Aplicaram-se dois modelos de regressão espacial (SAR e CAR) com efeitos globais que incorporam a dependência espacial. Estes modelos apresentaram melhores resultados quando comparados ao modelo de regressão múltipla clássica, indicando que a inclusão da dependência espacial nos modelos melhora a estimativa da produtividade de soja da região oeste do Paraná.

Portanto, de forma geral, os métodos estatísticos espaciais aplicados neste trabalho apresentaram-se eficientes na identificação de padrões de área, na quantificação da autocorrelação espacial, da correlação espacial e na aplicação das regressões espaciais.

2.5 AGRADECIMENTOS

À CNPq, à CAPES e à Fundação Araucária pelo apoio financeiro e ao SIMEPAR e ao SEAB pelo encaminhamento dos dados.

2.6 REFERÊNCIAS

ALMEIDA, E. S. de; PEROBELLI, F. S.; FERREIRA, P. G. C. Existe convergência espacial da produtividade agrícola no Brasil?.**Revista de Economia e Sociologia Rural**, Brasília, v. 46, n. 1, pp. 31-52. 2008.

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/SP,v.27, n.2, p.237-248. 2008.

ANSELIN, L. *Local indicators of spatial association - LISA*.**Geographical Analysis**, Ohio/USA,v.27, (2), p. 93-115. 1995.

ANSELIN, L.; SYABRI, I.; SMIRNOV, O. Visualizing multivariate spatial correlation with dynamically linked windows. In ANSELIN, L. and REY, S., editors, **New Tools for Spatial Data Analysis:** Proceedings of the Specialist Meeting. Center for Spatially Integrated Social Science (CSISS),University of California, Santa Barbara. 20p. 2003.

ANSELIN L.; SRIDHARAN S.; GHOLSTON S. Using exploratory spatial data analysis to leverage social indicator databases: the discovery of interesting patterns. **Social IndicatorsResearch**, [s. l.], v. 82, n. 2, p.287-309, 2007.

BAILEY, T. C.; GATRELL, A. C. **Interactive spatial data analysis**. Essex: Longman Scientific, 1995.

BERLATO, M.A.; FONTANA, D.C.; GONÇALVES, H.M. Relação entre rendimento de grãos de soja e variáveis meteorológicas. **Pesquisa Agropecuária Brasileira**. Brasília/DF, v.27, n.5, p.695-702, maio, 1992.

BROWN, S.; LO, K.; LYS, T. Use of R-squared in Accounting Research: Measuring changes in value relevance over the last four decades. **Journalof Accounting & Economics**, Amsterdam, v. 28, n. 2, p. 83-115, Dez. 1999.

CÂMARA, G; CAMARGO, E. C. G.; FUCKS, S. D. Análise Espacial de Superfícies.In: **Análise espacial de dados geográficos, eds**. FUKS, S.D.; CARVALHO, M. S.; CÂMARA, G. A. M. V. – Divisão de Processamentos de Imagens – Instituto Nacional de Pesquisas Espaciais – São José dos Campos – Brasil. 2002. Disponível em: <http://www.dpi.inpe.br/gilberto/livro/analise/cap1- introducao.pdf>Acesso em: 04 abr. 2010.

CÂMARA, G.; MONTEIRO, A. M. V*.* **Conceitos básicos em ciência da geoinformação**. São José dos Campos: INPE, 2004. 346p.

CELEBIOGLU, F.; DALL'ERBA, S. Spatial Disparities across the regions of Turkey: an exploratory spatial data analysis, **The Annals of Regional Sciences**. v.45, n.2, p.379-400, 2009.

DEPPE, F.; MARTINI, L.; LONHMANN, M.; ADAMI, M. Validation studies of ECMWF precipitation data with observed SIMEPAR ground data (meteorological stations). p.83-92. In: 2° INTERNATIONAL WORKSHOP ON CROP MONITORING AND FORECASTING IN SOUTH AMERICA, 2006. **Proceedings...** Montevideo: South America Scientific Network on Crop Monitoring and Forecasting, 2006.

DEPPE, F.; MARTINI, L.; LONHMANN, M.; CALVETTI, L.; ADAMI, M. Comparação de estimativas de precipitação com dados observados (estações meteorológicas). In: XII SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO (SBSR), 2007. Florianópolis/SC. **Anais…** São José dos Campos: INPE, p.3319-3326, 2007.

DOURADO NETO, D.; SPAROVEK, G.; FIGUEREDO JÚNIOR, L.G.M. de; FANCELLI A.L.; MANFRON, P.A.; MEDEIROS, S.L.P. Modelo para estimação da produtividade de grãos de milho deplecionada com base no balanço hídrico no solo. **Revista Brasileira de Agrometeorologia,** Santa Maria, v.12, n.2, p-359-367, 2004.

DRUCK, S.; CARVALHO, M.S.; CÂMARA, G.; MONTEIRO, A.V.M. (ed). *Análise* **Espacial de Dados Geográficos**. Brasília: EMBRAPA, 2004. 209p.

ESRI. ArcGIS Spatial Analyst. 2011. Disponível em: <http://www.esri.com/software/arcgis/extensions/spatialanalyst/surface.html>Acesso em: 10 set. 2012.

FONTANA, D.C.; BERLATO, M.A.; LAUSCHNER, M.H.; MELLO, R.W. Modelo de estimativa de rendimento de soja no Estado do Rio Grande do Sul. **Pesquisa Agropecuária Brasileira**,Brasília/DF, v.36, n.3, p.399-403, março, 2001.

GUIMARÃES, R. M. L.; GONÇALVES, A. C. A.; TORMENA, C. A.; FOLEGATTI, M. V.; BLAINSKI, E. Variabilidade espacial de propriedades físico-hídricas de um nitossolo sob a cultura do feijoeiro irrigado. **EngenhariaAgrícola**.,Jaboticabal/SP, v.30, n.4, p.657-669, 2010.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Banco de Dados Agregados - Sistema IBGE de Recuperação Automática – SIDRA**. 2011. Disponível em: <http://www.sidra.ibge.gov.br>. Acesso em 30 nov. 2011.

JACOX, E. H.; SAMET H. Spatial join techniques, **ACM Transactions on Database Systems (TODS),**Arizona/USA, v.32 n.1, p.7-es, March 2007KUHA, J. AIC and BIC: Comparisons of assumptions and performance. **SociologicalMethods&Research**, Massachusetts/USA, v.33, n.3, p. 417-417, 2004.

LOURENÇO, R.W.; LANDIM, P.M.B. Análise de regressão múltipla espacial. IGCE/DGA/Lab. **Geomatemática**, Rio Claro, 34 p. 2004.

McBRATNEY, A.; WEBSTER, R. Choosing functions for semi-variograms of soil properties and fitting them to sample estimates. European **Journal of Soil Science**, Londres/Inglaterra,37:617-639, 1986.

NICOLAU, R.; MACHADO, A.; NUNES, B. Análise da variação concelhia da mortalidade anual média por neoplasias malignas dos órgãos do aparelho respiratório e intra-torácicos em Portugal Continental. **Revista Portuguesa de Saúde Pública**,Lisboa/Portugal, v.27, n.2, p.7-16. ISSN 0870-9025.jul. 2009.

OPENGEODA.**GeoDa Center for Geospatial Analysis and Computation**.2011. Disponível em: <http://geodacenter.asu.edu/about>.Acesso em: 8 set. 2012

PIMENTEL, E.A.; HADDAD, E.A. **Desigualdades regionais em Minas Gerais**: análises espaciais do fenômeno, 1991- 2000. Trabalho apresentado ao 3º Encontro da Associação Brasileira de Estudos Regionais, Belo Horizonte, 2004.

SCHWARZ, G. Estimating the dimensional of a model. **Annals of Statistics**, Hayward, v.6, n.2, p.461-464, 1978.

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

SIMEPAR. **Sistema Meteorológico do Paraná**. 2010.

VALCU, M.; KEMPENAERS, B. **Spatial autocorrelation**: an overlooked concept in behavioral ecology. Behavioral Ecology, [s.l.], v. 21, n. 5, p. 902-905, 2010.

ZIBORDI, M. S.; CARDOSO, J. L.; VILELA FILHO, L. R.. Análise de aspectos socioeconômicos e tecnológicos da agropecuária na Bacia Hidrográfica do Rio Mogi Guaçu. **EngenhariaAgrícola**.Jaboticabal/SP, v.26, n.2, p. 644-653, 2006.

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

RESUMO: O presente trabalho realizou uma análise de agrupamentos espaciais por meio da estatística multivariada, no intuito de investigar a associação entre a produtividade da soja e as variáveis agrometeorológicas: precipitação pluvial, temperatura média do ar, radiação solar global e o índice local de Moran (*LISA*) da produtividade. O estudo foi realizado com os dados das safras dos anos agrícolas 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 uma melhor gestão das atividades de produção da agricultura, em função de que, com o agrupamento, foi possível estabelecer similaridades que proporcionem parâmetros para uma 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.

ANALYSIS GROUPING OF SPATIAL VARIABILITY OF SOYBEANS PRODUCTIVITY AND AGROMETEREOLOGICAL FACTORS IN WEST PARANÁ

ABSTRACT: This work carried out an analysis of spatial cluster using multivariate statistics, in order to investigate the association between soybean productivity and meteorological variables precipitation, average air temperature and solar radiation and Moran index local (*LISA*). The study was conducted with data from the harvest of the crop years 2000/2001 to 2007/2008 in the West region of Paraná state. 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 cities 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 grouping is possible to establish similarities that provide parameters for better management of production processes that bring, both quantitatively and qualitatively,the satisfactory results sought by the farmer.

KEYWORDS: Spatial statistics area; Spatial similarity; Multivariate analysis.

3.1 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, 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 e 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 são as 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, Kunz *et al.* (2008) e Kóvals*et al.* (2005) 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 (t ha^{-1}) e de variáveis agrometeorológicas: precipitação pluvial (mm), temperatura média do ar (ºC), radiação solar global média (W m^{-2}) e índice de Moran Local univariado para a produtividade da soja (LISA) para região oeste do Estado do Paraná.

3.2 MATERIAL E MÉTODOS

A área de estudo deste trabalho é apresentada nas Figuras 22e 23e compreende 48 municípios da região oeste do Estado do Paraná. Foram utilizados dados dos anos-safra 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 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 a temperatura média do ar e radiação solar global média pela média aritmética. Os dados referentes à produtividade da soja em (t ha^{-1}) foram fornecidos pela SEAB (2010) e os dados agrometeorológicos pelo SIMEPAR (2010). Os dados agrometeorológicos: temperatura média do ar (ºC), radiação solar global média (W m^{-2}) e precipitação pluvial

(mm) 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 (mm) 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).

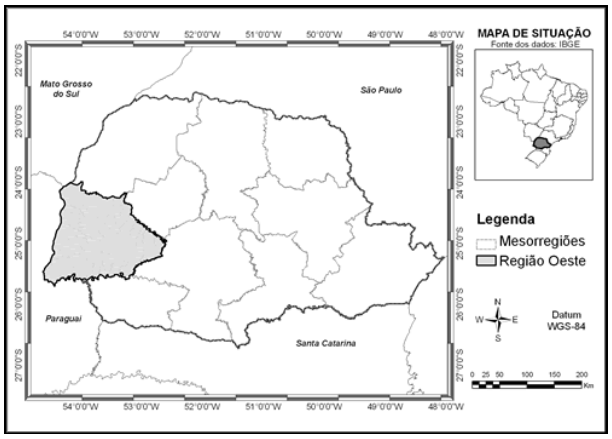


Figura 22 Mapa de localização da região oeste do estado do Paraná.

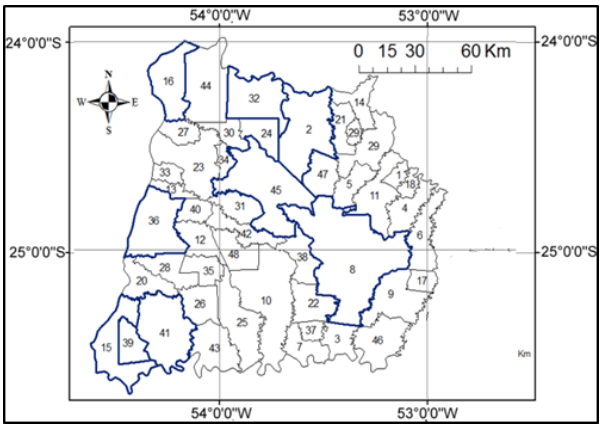


Figura 23 Região oeste do Paraná, com destaque para os municípios com estações meteorológicas.

(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 é da forma representada na Equação 30, 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.

$$X_i=(X_{i1}, X_{i2},..., X_{ip}), \text{ para } i=1,..., n, \qquad \text{Eq.(30)}$$

Seja X uma matriz de observações de n x p de n elementos amostrais em p variáveis, escrita da forma especificada na Equação 31:

$$X = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2j} & \dots & X_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{ip} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ X_{n1} & X_{n2} & \dots & X_{nj} & \dots & X_{np} \end{bmatrix} = [X_1 X_2 \dots X_n]^T \quad \text{Eq.(31)}$$

A medida mais utilizada na indicação da proximidade entre dois objetos *i* e *k* é a distância euclidiana, representada por Johnson e Wichern (1992) pela Equação32.

$$d_{ik} = d(i,k) = [\sum_{j=1}^{p} (X_{ij} - X_{kj})^2]^{1/2} \quad \text{Eq.(32)}$$

em que $i \neq k=1,\dots, n$ (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 $X_{i1},\dots, X_{ip}$ do vetor X_i, usando a transformação descrita na Equação 33.

$$z_{ij} = \frac{X_{ij} - \bar{X}_{.j}}{s_{.j}} \quad \text{Eq.(33)}$$

em que $i=1,\dots, n$; $j=1,\dots, p$; e $\bar{X}_{.j}$ e $s_{.j}$ 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 34, que é a soma dos desvios padronizados.

$$d_{ik} = d(i,k) = [\sum_{j=1}^{p} (z_{ij} - z_{kj})^2]^{1/2} \quad \text{Eq.(34)}$$

De acordo a Boschi *et al.* (2011), a técnica de agrupamento apresenta grande eficiência, e Bussab (1990) sugere 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), que é definida pela Equação 35:

$$d_{(kl)j} = \frac{(d_{kl} + d_{lj})}{2} \quad \text{Eq.(35)}$$

em que $d_{(kl)j}$ é a distância entre o agrupamento (*kl*) e o agrupamento *j*; d_{kl} e d_{lj} 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 36:

$$s(ij) = 100\left(1 - \left(\frac{d_{ij}}{d_{max}}\right)\right) \qquad \text{Eq.(36)}$$

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

A autocorrelação espacial local univariada (*LISA*) busca captar padrões de associação local. A autocorrelação local pode ser calculada pela estatística *I* de Moran local, também conhecido como *Local Indicator of Spatial Association* (*LISA*) (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 (região, por exemplo).

Segundo Le Gallo e Erthur (2003), o índice *LISA* univariado, baseada no *I* de Moran local, pode ser especificado para uma determinada variável X_j, *j*=1,...,*p*, da forma descrita na Equação 37:

$$I_{ij} = \frac{X_{ij} - \overline{X}_{.j}}{\sigma_j^2} \sum_{k=1}^{n} w_{ik}^{(j)} (X_{kj} - \overline{X_{.j}}),\ i = 1,...,n \text{ para cada j=1,...,}p, \qquad \text{Eq.(37)}$$

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, 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 ótimo de grupos 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 38.

$$RMSSTD = \left[\frac{\sum_{k=1}^{c} \sum_{i=1}^{n_k} \sum_{j=1}^{p} (X_{ij}^{(k)} - \overline{X_{.j}^{(k)}})^2}{p \sum_{k=1}^{c} (n_k - 1)} \right]^{1/2} \quad \text{Eq. (38)}$$

O *RS* pode ser considerado uma medida da dissemelhança 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 homogêneo será cada grupo e vice-versa. O *RS* é definido pela Equação 39.

$$RS = 1 - \frac{\sum_{k=1}^{c} \sum_{i=1}^{n_k} \sum_{j=1}^{p} (X_{ij}^{(k)} - \overline{X_{.j}^{(k)}})^2}{\sum_{k=1}^{c} \sum_{i=1}^{n_k} \sum_{j=1}^{p} (X_{ij}^{(k)} - \overline{X}_{..})^2} \quad \text{Eq.(39)}$$

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 alocado 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_ke *j=1,...,p.*

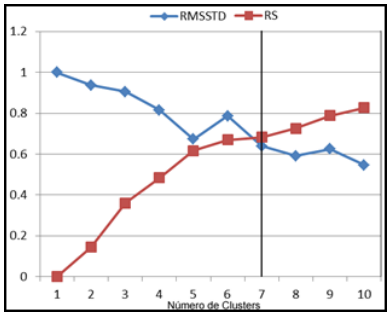
Para cada ano-safra (2000/2001 a 2007/2008), tendo como dados todas as variáveis do estudo (produtividade da soja (t ha^{-1}) [Prod], precipitação pluvial (mm) [Prec], temperatura média do ar (ºC) [TMed], radiação solar global média (W m^{-2}) [Rs] e índice de Moran Local da produtividade da soja [*LISA*], foram geradas as estatísticas *RMSSTD* e *RS* para até 10 grupos de *clusters*, com o objetivo de identificação do melhor número de *clusters* para cada ano-safra. Em um segundo estudo, também com até 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).

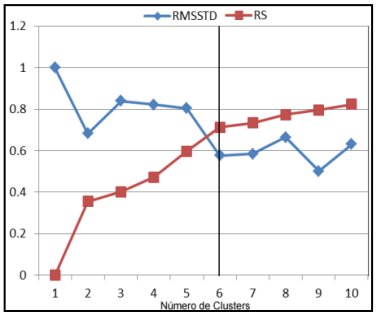
3.3 RESULTADOS E DISCUSSÃO

Na busca por um número ótimo para a quantidade de *clusters*, determinaram-se as estatísticas *RMSSTD* e *RS* para cada ano-safra estudado (Figura 24). 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).

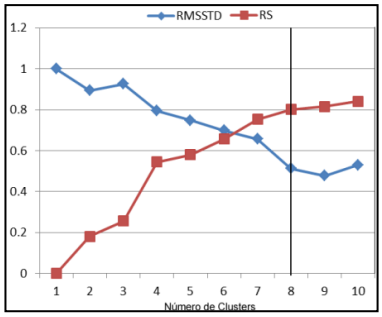
Na execução das estatísticas (Figura 24), em algumas safras, o número de *clusters* identificado como ótimo causou a existência de municípios isolados, sem pertencer a nenhum *cluster*. Os resultados dos agrupamentos por nível de similaridade são apresentados na Tabela 7 e podem ser visualizados na Figura 25.



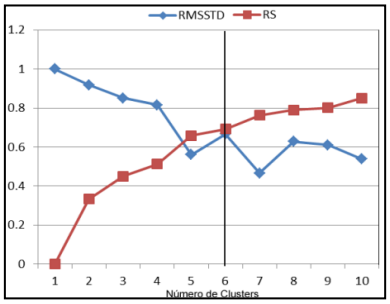
(a) Safra 2000/2001



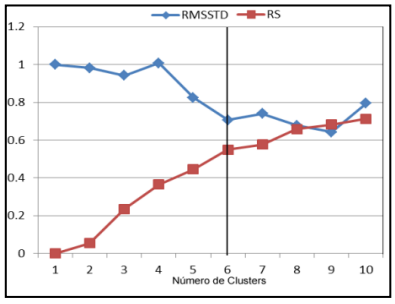
(b) Safra 2001/2002



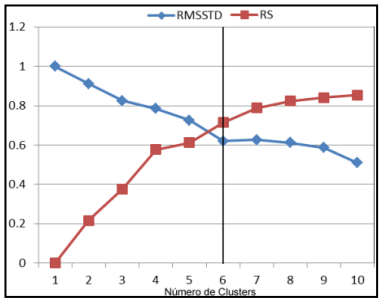
(c) Safra 2002/2003



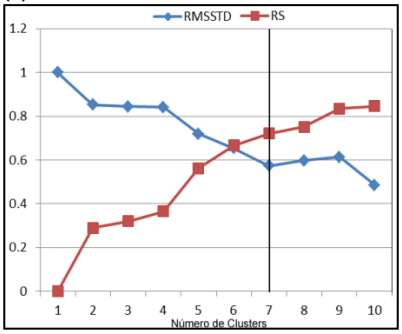
(d) Safra 2003/2004



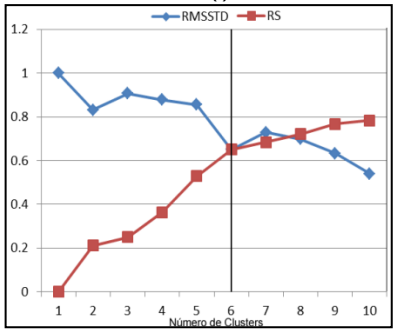
(e) Safra 2004/2005



(f) Safra 2005/2006



(g) Safra 2006/2007



(h) Safra 2007/2008

Figura 24 Gráfico de estimação do número ótimo de *clusters* para os anos-safra em estudo por meio das estatísticas RMSSTD e RS

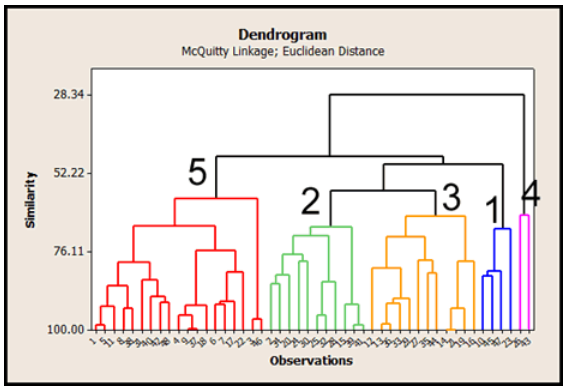
Tabela 7 Processo de agrupamento por similaridade e distância euclidiana dos municípios da área em estudo, considerando as variáveis Prod, Prec, TMed, Rs, *LISA*.

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**

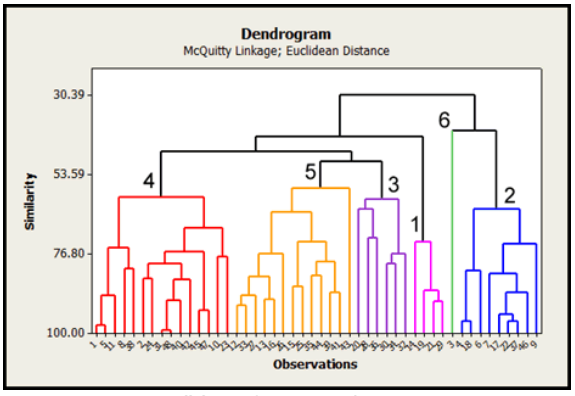
Em negrito estão os municípios que ficaram isolados, sem associação a *clusters*.

Para os anos-safra 2001/2002, 2003/2004, 2004/2005 e 2007/2008, um total de seis *clusters* foram identificados como o número ótimo (Figura 24). Para cada um desses anos-safra, houve um município isolado dos demais *clusters* (3: Boa Vista da Aparecida, 1: Anahy 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. Os anos-safra 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

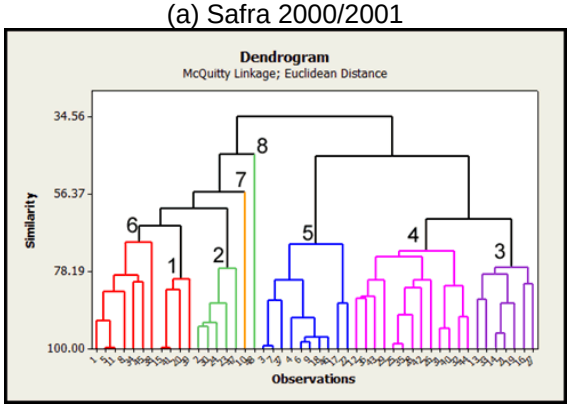
número ótimo de *clusters* para o ano-safra de 2002/2003 foi de oito, porém dois municípios (10: Céu Azul e 48: Vera Cruz do Oeste) ficaram isolados.



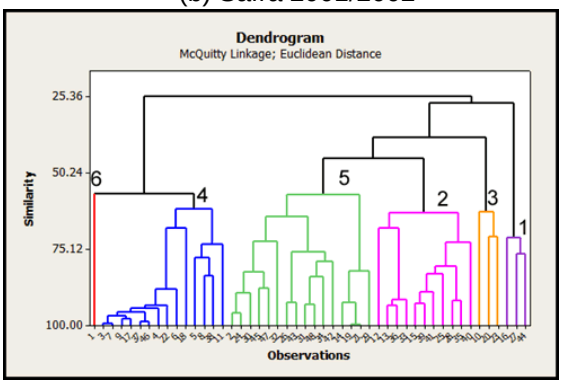
(a) Safra 2000/2001



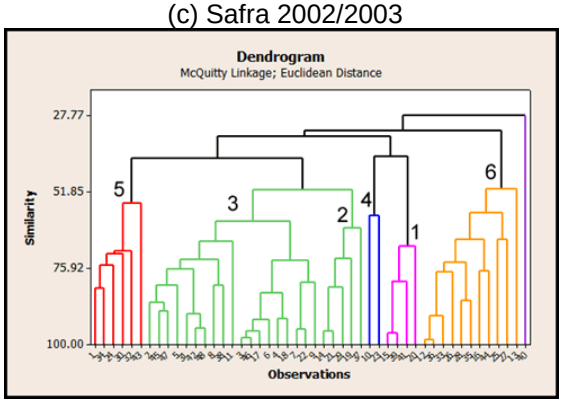
(b) Safra 2001/2002



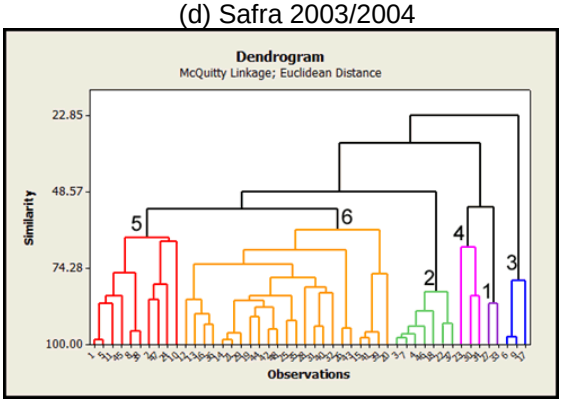
(c) Safra 2002/2003



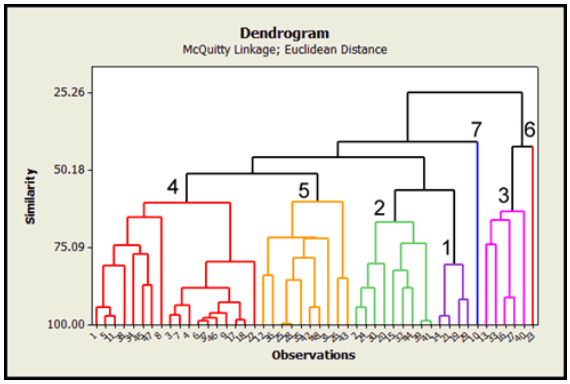
(d) Safra 2003/2004



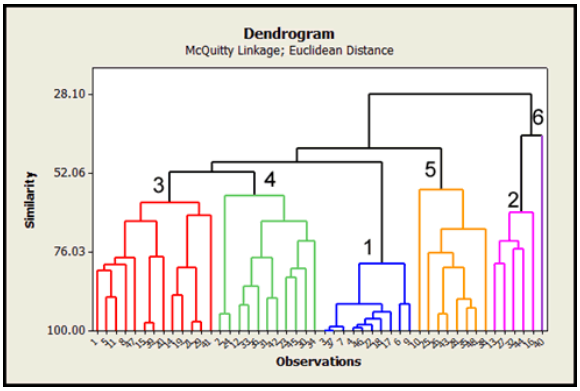
(e) Safra 2004/2005



(f) Safra 2005/2006



(g) Safra 2006/2007



(h) Safra 2007/2008

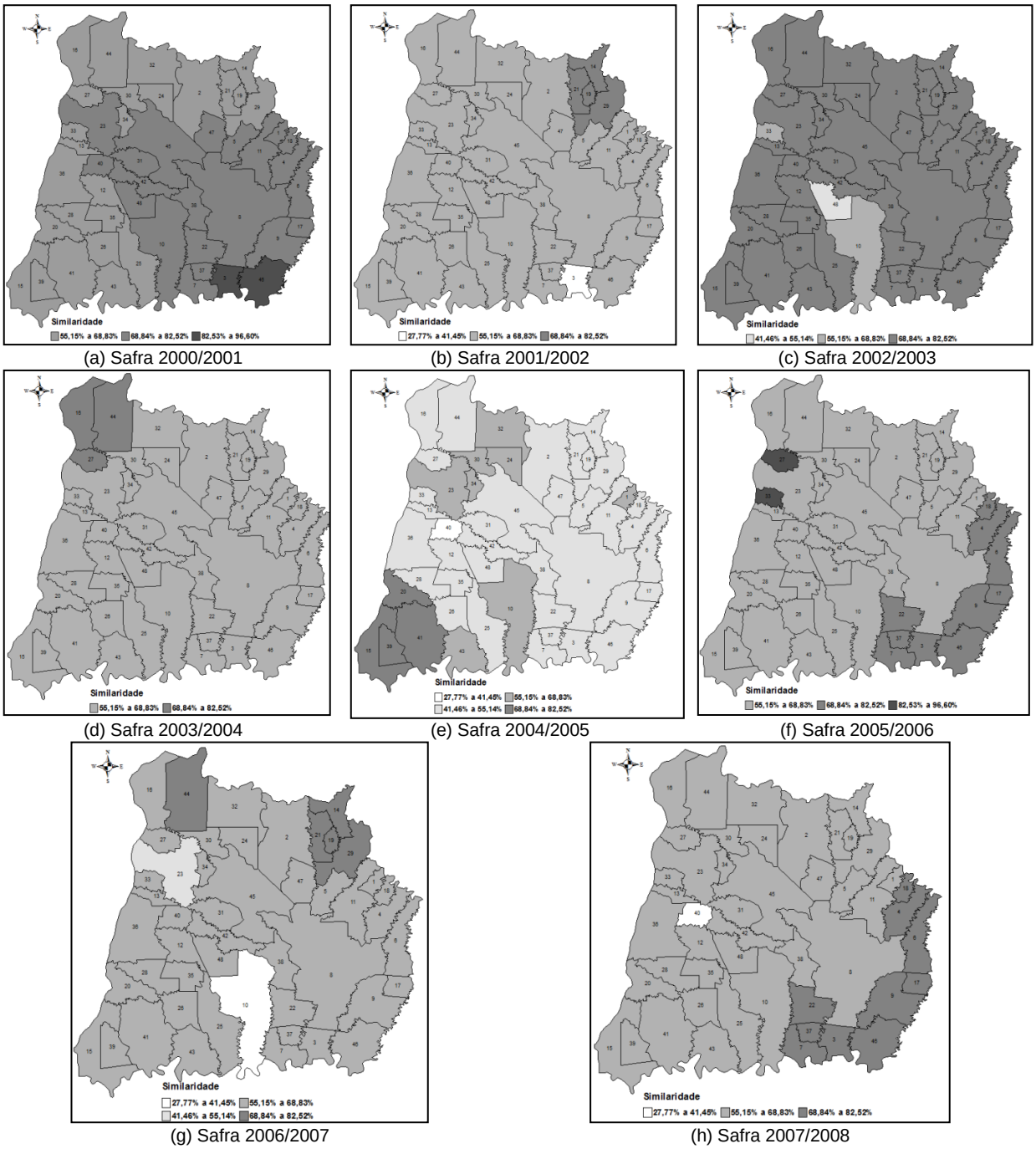
Figura 25 Dendogramas gerados com as variáveis produtividade da soja (t ha^{-1}), precipitação pluvial (mm), temperatura média do ar (ºC), radiação solar global média (W m-2) e índice *LISA* para os 48 municípios da área de estudo em oito anos.

Na análise de similaridade dos municípios apresentada na Tabela 7 e Figura 25, identificou-se a quantidade de *clusters* indicada pelas estatísticas *RMSSTD* e *RS* para cada um dos anos-safra estudados. A similaridade variou entre 27,77% (*cluster* 6 do ano-safra 2004/2005) e 96,60% (*cluster* 1 do ano-safra 2005/2006) com média de 63,89% para os *clusters* e anos-safra 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 (1998), 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 ao que foi apresentado na Tabela 7, 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 25 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 26, 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*.



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

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

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

Figura 26 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.

Verifica-se na Figura 26 que dois anos-safra (2000/2001 e 2005/2006) possuem seus municípios agrupados entre 55,15 e 96,60% de similaridade, e no ano-safra 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-safra 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 se repetem nestes agrupamentos. Como exemplos, nas safras de 2000/2001 (29a) e 2002/2003 (29c) 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, na safra de 2001/2002 (29b), 2003/2004 (29d), 2005/2006 (29f) e 2007/2008 (29h) estes mesmos municípios se encontram no *cluster* de faixa entre 55,15 a 68,83%. Na safra de 2004/2005 (29e) destes municípios apenas 45: Toledo e 47: Tupãssi se encontram no mesmo *cluster*, de 41,46 a 55,14%, enquanto o município 23: Marechal Cândido Rondon encontra-se no intervalo de 55,15 a 68,83%, mantendo a similaridade das safras de 2001/2002. Na safra de 2006/2007 (29g), novamente apenas 45: Toledo e 47: Tupãssi, se encontram 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 27 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 8 apresenta a análise de similaridade dos municípios para os sete *clusters* e o resultado em forma de mapa temático pode ser visualizado na Figura 27, juntamente com o dendograma.

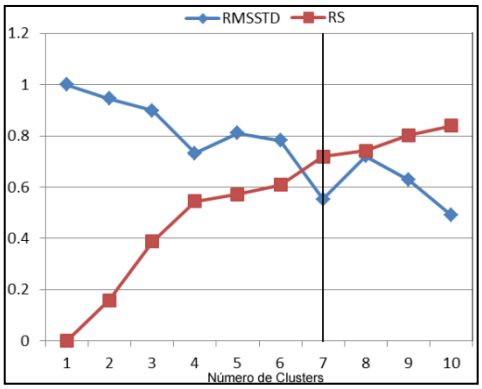


Figura 27 Gráfico de estimação do número ótimo de *clusters* para todas as safras em estudo, como um único conjunto, por meio das estatísticas RMSSTDe RS.

Tabela 8 Processo de agrupamento por similaridade e distância dos municípios da área em estudo para os anos-safra de 2000/2001 a 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

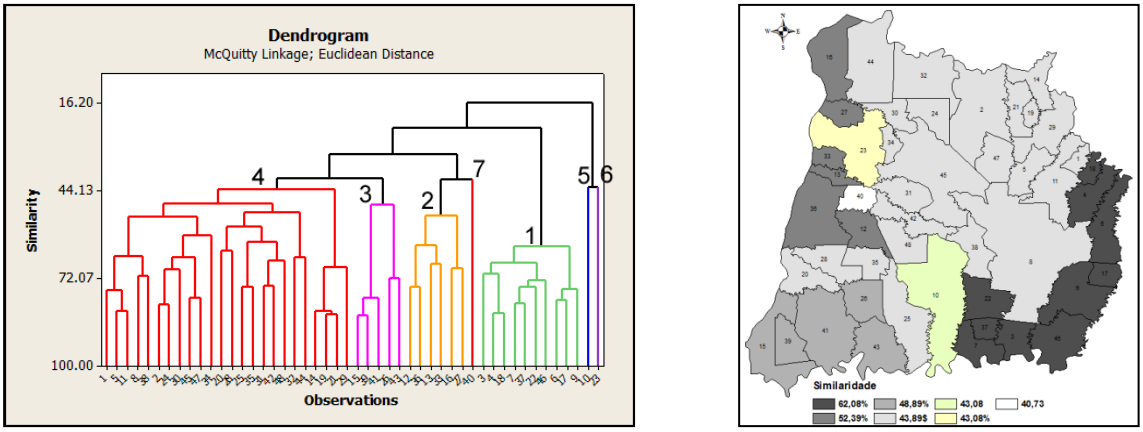


Figura 28 Mapa temático e dendograma de análise dos agrupamentos dos municípios da pesquisa com base no índice de similaridade considerando as variáveis na produtividade da soja (t ha^{-1}), precipitação pluvial (mm), temperatura média do ar (ºC), radiação solar global média (W m^{-2}) e índice *LISA* univariado, para todos os anos-safra em estudo.

Pela Tabela 8 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 tem como característica um relevo mais acidentado, quando comparado 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, 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 8 foram exatamente os mesmos na análise de agrupamentos por safra nos anos-safra de 2002/2003 e 2007/2008, 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 os 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 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 1487,2 mm, temperatura média do ar 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 dos municípios se localizarem distribuídos dentro da área da pesquisa, e em alguns casos eles encontrarem-se 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 26, verifica-se 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 existe 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 nas 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 28foi 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 (onde o município 7: Capitão Leônidas Marques fez parte de outro *cluster*)e 2001/2002 onde 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.

3.4 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 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 útil para uma melhor gestão das atividades de produção da agricultura, em função de permitir estabelecer 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 se estabelecer um padrão para os processos produtivos. Isso fica evidente nas diferenças encontradas nos 48 municípios analisados na pesquisa.

3.5 AGRADECIMENTOS

Ao CNPq, à CAPES e à Fundação Araucária pelo apoio financeiro e ao SIMEPAR e ao SEAB pelo encaminhamento dos dados.

3.6 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/SP, v. 27, n.2, p. 237-248. 2008.

ANSELIN, L. **Local indicators of spatial association – LISA**. Geographical Analysis, Ohio/USA, 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/SP, v. 31, n. 6, p. 1189-1201,2011.

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**, Reino Unido,v. 11, n.1, p.4-5, 1996.

CORRAR, L. J.; PAULO, E.; FILHO, J. M. D. **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. 511-521, 2008.

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.

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[online].Brasília/DF, v. 45, n.10, p.1095-1101, 2010.

HARDLE, W.; SIMAR, L. **Applied Multivariate Statistical Analysis**. 2. ed. Berlin: Springer, 2007. 486p.

JACOX, E. H.; SAMET H. **Spatial join techniques**.ACM Transactions on Database Systems (TODS),Arizona/USA, v. 32, n. 1, p. 7-75, march 2007.

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

KOVÁCS, F.; LEGÁNY, C.; BABOS, A. Cluster validity measurement techniques**.Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence**, Budapest, p. 18-19, 2005.

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/SP, v. 28, n. 1. p. 125-35, 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**,Londres/Inglaterra, v. 82, n. 2, p. 175-201, 2003.

LUDEWIG, D. R.; URIBE-OPAZO, M. A.; GIMENES, R. M. T.; SOUZA, E.G. – O Processo de Gestão de Custos e Planejamento de Resultados utilizando técnicas de análise estatística de agrupamentos. **Revista Scientiarum Techonology**, Maringá (PR), v.1 n. 2, p. 215-220. 2009.

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,**Tóquio/Japão,v. 8, p. 88-93, 2009.

4 CLASSIFICAÇÃO DE ÁREAS ASSOCIADAS À PRODUTIVIDADE DA SOJA E VARIÁVEIS AGROMETEOROLÓGICAS POR MEIO DE AGRUPAMENTO FUZZY

RESUMO

Este trabalho teve como objetivo aplicar uma abordagem baseada em agrupamento difuso para classificação de áreas associadas à produtividade da soja, conjuntamente comas variáveis agrometeorológicas: precipitação pluvial, temperatura média do ar e radiação solar global média. O estudo foi realizado envolvendo 48 municípios da região oeste do estado do Paraná, Brasil, com dados da safra do ano-agrícola 2007/2008. Por meio do algoritmo *Fuzzy c-Means,* foi possível formar grupos de municípios similares à produtividade de soja, utilizando o Método de Decisão pelo Maior Grau de Pertinência (*MDMGP*) e o Método de Decisão pelo Limiar *β* (*MDL β*). Posteriormente, a identificação do número adequado de agrupamentos foi obtida utilizando a Entropia de Partição Modificada. Para mensurar o nível de similaridade de cada agrupamento, definiu-se e empregou-se o Índice de Similaridade de *Clusters* (ISC_I). Dentro das perspectivas deste estudo, o método empregado se mostrou adequado, permitindo identificar agrupamentos de municípios com graus de similaridades da ordem de 60 a 78%.

Palavras-chave: *Fuzzy c-Means*, Métodos de decisão, Índice de similaridade.

CLASSIFICATION OF AREAS ASSOCIATED WITH SOYBEANS PRODUCTIVITY AND AGROMETEREOLOGICAL FACTORS THROUGH FUZZY CLUSTERING

ABSTRACT: This study aimed to apply an approach based on fuzzy clustering for classification of areas associated with soybean yield, together with meteorological factors: rainfall, average air temperature and average global solar radiation. The study was conducted involving 48 cities in the western region of Paraná State, Brazil, with crop data from the crop-year of 2007/2008. Through Fuzzy C-Means algorithm, it was possible to form groups of counties similar to soybean yield using the Decision Method by the Higher Degree of Relevance (*DMHDR*) and Method of Decision Threshold by β (*MDT gnβ*). Subsequently, identification of the adequate number of clusters was obtained using the Modified Partition Entropy. To measure the degree of similarity of each cluster, it was designed and used the Index of Similarity Clustering (ISC_I). In the perspective of this study, the method used was adequate, making it possible to identify clusters of cities with degrees of similarities in the order of 60 to 78%.

KEY WORDS: Fuzzy c-Means, Decision Methods, Index of similarity.

4.1 INTRODUÇÃO

A soja tem sido objeto de estudos que buscam compreender as relações que as variáveis agrometeorológicas têm na formação da produção da cultura (DALACORT et al., 2006; TOLEDO et al., 2010; CARMELLO, 2011). Na região oeste do estado do Paraná, essa produção tem sido responsável por contribuir com a balança comercial do estado e consequentemente do Brasil. A soja é um dos principais produtos agrícolas brasileiros

sendo que, somente em 2011, gerou uma receita da ordem de US$ 24,9 bilhões (CONAB, 2008).

Os procedimentos para geração de estimativas de safra agrícola, assim como o conhecimento da sua distribuição espacial, constituem uma importante informação para o setor agricultura. Grande parte desses procedimentos envolvem técnicas de previsão baseadas na agrometeorologia, fundamentando-se na relação estatística entre a variável dependente produtividade da cultura de soja e as variáveis independentes, como: precipitação pluvial, temperatura média e radiação solar global (BERLATO et al., 1992;DALACORT et al., 2006; ASSAD et al., 2007). Com relação aos fatores agrometeorológicos, como seca, excesso de chuvas, temperaturas muito altas ou muito baixas e a baixa luminosidade, ressalta-se que estes podem ocasionar reduções significativas na produtividade da soja, restringindo inclusive as áreas onde espécies comercialmente importantes podem ser cultivadas (FARIAS, 2011).

Para o desenvolvimento da soja, a disponibilidade de água é de extrema importância, principalmente em dois períodos: germinação-emergência e floração-enchimento de grãos. Durante o primeiro período, tanto o excesso como a falta de água é prejudicial para obtenção de uma boa uniformidade na população de plantas, sendo o excesso hídrico mais limitante do que o déficit (EMBRAPA, 2008). Outro fator agrometeorológico causador de impactos nesse processo é a temperatura. A soja se adapta melhor às regiões onde a temperatura oscila entre 20 e 30°C, sendo que a semeadura não deve ser realizada quando a temperatura do solo estiver abaixo dos 20°C, pois a germinação e a emergência da planta ficam comprometidas. Outro componente ambiental relevante ao desenvolvimento da soja é a radiação solar, pois, além de fornecer energia luminosa para a fotossíntese, fornece sinais ambientais para uma gama de processos fisiológicos para essa cultura. Assim, além da intensidade da radiação, a duração e a qualidade do espectro luminoso são determinantes de respostas morfológicas e fenotípicas marcantes, tais como estatura da planta, indução ao florescimento e ontogenia (THOMAS,1994).

Com vistas a investigar a relação entre a produtividade da soja e as variáveis agrometeorológicas, a teoria de conjuntos nebulosos conhecida como teoria dos conjuntos *fuzzy*, foi adotada. Essa abordagem baseia-se na caracterização de classes que não possuem limites rígidos entre si (BURROUGH;MCDONNEL, 1998; RODRIGUES JÚNIOR et al., 2011), sendo indicada quando se busca trabalhar com informações em um ambiente de incerteza, imprecisão, ambiguidades, abstrações e ambivalências em modelos matemáticos complexos, em que os limites difusos comuns em processos que ocorrem no espaço são representados (TAYLOR et al., 2007; YAN et al., 2007).

Rodrigues Junior *et al.* (2011) fizeram uso do *Fuzzy c-Means* para definirem zonas de manejo para cafeicultura. Os autores tiveram como base determinações realizadas com

sensor de clorofila e por análise foliar. Odeh *et al.* (1992) identificaram classes de solo com amostras obtidas de dois perfis, utilizando também o classificador *Fuzzy-c-means*.

Neste contexto, o objetivo deste trabalho foi classificar áreas associadas à produtividade da soja (tha^{-1}) na região oeste do estado do Paraná, considerando as seguintes variáveis agrometeorológicas: precipitação pluvial (mm), temperatura média do ar (^oC) e radiação solar média (Wm^{-2}).

4.2 MATERIAIS E MÉTODOS

Área de Estudo: compreende 48 municípios da região oeste do estado do Paraná, Brasil, localizados entre as Longitudes W 52^o 54' e W 54^o 36' e Latitudes S 23^o 58' e S 25^o 40', conforme ilustra a Figura 29, com destaque para os municípios que possuem estações meteorológicas.

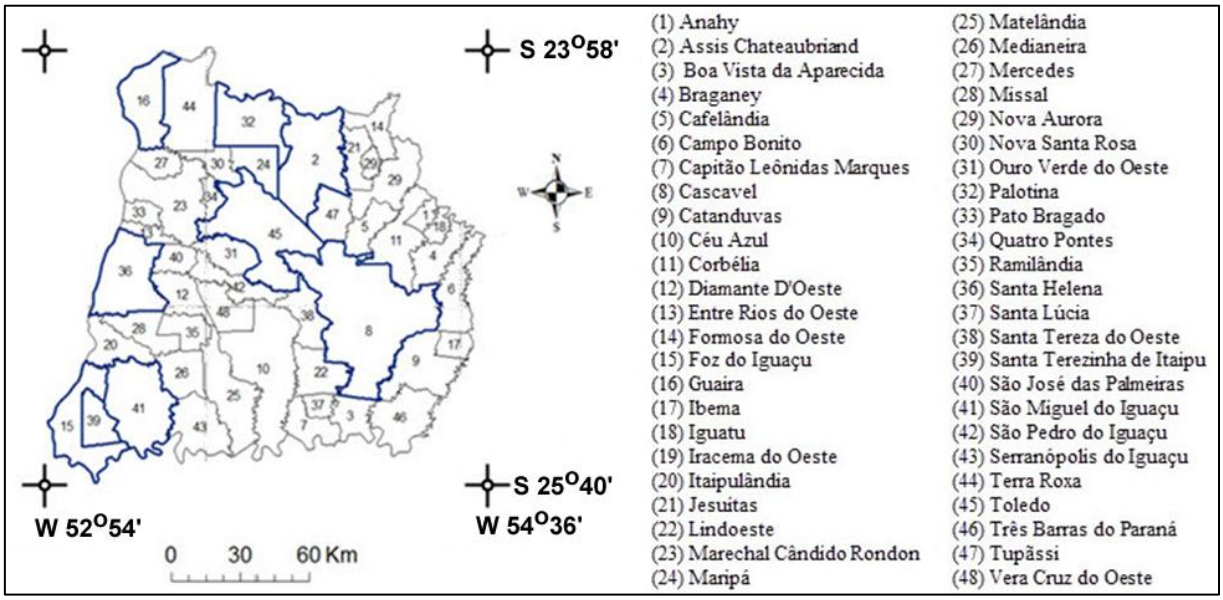


Figura 29 Região oeste do Paraná com destaque para os municípios com estações meteorológicas

Organização dos Dados: no domínio de Sistemas de Informação Geográfica (SIG), as considerações para organização do Banco de Dados Geográfico (BDG) são aplicáveis de acordo com os diferentes tipos e natureza dos dados, que necessitam ser organizados e armazenados. Em geral, e também para fins deste estudo, os dados para o BDG são classificados como: dado espacial e dado não-espacial.

Dado espacial: está associado a elementos geográficos ou espaciais, como o conjunto de polígonos que representam o mapa de municípios que cobrem a região de estudo (vide Figura 29). Atualmente, muitos desses mapas são fornecidos no formato digital.

Para este estudo, o mapa de municípios da região oeste do Paraná foi obtido do IBGE (2012).

Dado não-espacial: se refere a um conjunto de atributos que complementam o dado espacial, descrevendo o que está associado em um ponto, ao longo de uma linha ou em um polígono. Neste trabalho, os atributos foram organizados em uma tabela de 48 linhas (número de polígonos) por 5 colunas (número de atributos). Posteriormente, esta tabela foi incorporada ao BDG e ligada ao mapa de municípios. A definição dos atributos foi baseada na identificação da maior média da produtividade de soja de todos os municípios entre os anos-safra de 2000/2001 e 2007/2008, sendo selecionado o do ano-safra 2007/2008. Assim, para cada um dos 48 municípios foram estabelecidos os seguintes atributos:

- identificador do polígono (PID);
- produtividade da soja (PROD), medida em (tha^{-1}), fornecida pela SEAB (2010);
- três atributos agrometeorológicos advindos do SIMEPAR (2010): precipitação pluvial (PREC), medida em mm, temperatura média do ar (TMED), em ºC, e radiação solar global média (RSGM), em Wm^{-2}. Inicialmente, essas variáveis estavam disponíveis somente para oito municípios da região de estudo, conforme destacados anteriormente na Figura 29. Para os demais municípios sem medições, foram obtidas estimativas por meio do uso de Polígonos de Thiessen (ANDRADE et al., 2008) e *Spatial Join* (JACOX; SAMET, 2007);

  O período das safras utilizado para obtenção dos dados agrometorológicos diários foi de 1^{o} de outubro de 2007 até 28 de fevereiro de 2008. A precipitação pluvial utilizada foi obtida por meio da soma dos dados do período e a temperatura média e radiação solar global média pela média aritmética.

Método: o método empregado neste estudo envolve um procedimento de 5 etapas:

i. Padronização dos atributos (PROD, PREC, TMED, RSGM), conforme segue:

$$Vatr_{pad} = \frac{(Vatr - Vatr_{\min})}{(Vatr_{\max} - Vatr_{\min})} \qquad \text{Eq.(37)}$$

em que: $Vatr_{pad}$ é o valor do atributo padronizado, variando de 0 a 1*;* *Vatr* é o atributo observado; $Vatr_{min}$ e $Vatr_{max}$ referem-se aos valores mínimos e máximos do atributo observado, respectivamente. Essa forma de padronização é necessária para o emprego do *FCM*, garantindo que todos os atributos tenham a mesma ordem de grandeza, isto é, variando de 0 a 1 (GOMES et al., 2011);

ii. Aplicaçãodo *Fuzzy c-Means* (*FCM*) para se obter agrupamentos ou grupos similares em um conjunto de dados. O *FCM* é um algoritmo iterativo e, a cada iteração, novos centros de agrupamentos e graus de pertinência são calculados, buscando sempre minimizar a métrica euclidiana entre cada dado e o centro do

agrupamento. A responsabilidade de verificar essa convergência cabe à função objetivo. Uma descrição mais detalhada do *FCM* pode ser vista em Dunn (1973), Bezdek (1981) e Bezdek e Pal (1992);

iii. Alocação dos dados nos agrupamentos estabelecidos pelo *FCM* foi realizada por meio de métodos de decisão. Para esse estudo, foram utilizados dois métodos: Método de Decisão pelo Maior Grau de Pertinência (*MDMGP*) e Método de Decisão pelo Limiar *β* (*MDLβ*) (GUIERA et al., 2005; FERREIRA et al., 2008; NG et al., 2008; WANG, 2009). No *MDMGP,* a determinação a qual agrupamento o dado pertencerá é dada pelo maior grau de pertinência. Isto garante que todos os dados sejam alocados. Já o *MDL β,* a determinação se dá por meio de um limiar *β*, isto é, um dado pertencerá a um agrupamento desde que seu grau de pertinência seja maior ou igual a *β*. Diferente do *MDMGP*, o *MDL β* pode gerar um conjunto de dados que não estejam alocados a nenhum agrupamento, justamente por estarem abaixo do limiar *β* estabelecido (GUIERA et al., 2005). A partir da alocação dos dados nos agrupamentos, torna-se importante uma avaliação da qualidade do resultado obtido. Neste estudo, empregou-se a Entropia de Partição Modificada (*MPE*) para a indicação do melhor número de agrupamentos (BUDAYN et al., 2009; SUN et al., 2012):

$$MPE = \frac{-\sum_{k=1}^{n}\sum_{i=1}^{c} u_{lk}\log(u_{lk})/n}{\log(c)} \qquad \text{Eq.(38)}$$

em que: *c* representa a quantidade de agrupamentos; *n* corresponde ao número de polígonos; u_{lk} corresponde ao grau de pertinência do polígono *k* do agrupamento *l*.

O *MPE* mede o grau de desorganização entre os agrupamentos de um conjunto de dados. Seu valor varia de 0 a 1. Valores próximos de zero indicam agrupamentos distintos, apresentando pequeno grau de compartilhamento entre os dados. Valores próximos de um indicam não haver agrupamentos distintos, apresentando elevado grau de compartilhamento dos dados (MCBRATNEY;MOORE, 1985; FRIDGEN et al., 2004). Segundo Boydell e Mcbratney (2002), o melhor número de agrupamento de um conjunto de dados é estabelecido com base no valor mínimo de *MPE*;

iv. Mensuração do nível de similaridade de cada agrupamento. Para tal, definiu-se e empregou-se o Índice de Similaridade de *Clusters* (ISC_l), conforme segue:

$$ISC_l = \left[\frac{\sum_{k=1}^{n_l} u_{lk}}{n_l}\right] 100 \qquad \text{Eq.(39)}$$

em que n_l corresponde ao número (n) de polígonos do agrupamento *l* e u_{lk} segue a mesma definição dada na Equação 38.

v. Resultados do *FCM* foram transportados para BDG e conectados ao mapa de polígonos para posteriores análises.

Este trabalho foi realizado com auxílio dos seguintes programas: ArcMap 9.3 (ESRI, 2011) e Matlab R2010a (MATLAB, 2011).

4.3 RESULTADOS E DISCUSSÃO

Inicialmente foi realizada uma análise preliminar dos dados. A Tabela 9 sintetiza as principais estatísticas descritivas das variáveis observadas, bem como de seus valores padronizados.

Tabela 9 Estatísticas descritivas das variáveis e de seus respectivos valores padronizados no ano-safra de 2007/2008.

Atributo	Média	Desv. Padrão	Coef. Var.	Mín.	Máx.	Mediana
PROD (t ha^{-1})	3,27	0,22	6,86	2,50	3.70	3,29
PROD padronizada	0,64	0,19	29,03	0	1	0,65
PREC (mm)	3959	2022	51,07	826	10962	3630
PREC padronizada	0,31	0,20	64,54	0	1	0,28
TMED (ºC)	24,33	0,86	3,53	22,90	25,30	24,60
TMED padronizada	0,60	0,36	60,02	0	1	0,71
RSGM (W m^{-2})	442,51	42,74	9,66	366,05	536,05	446,35
RSGM padronizada	0,45	0,25	55,90	0	1	0,47

Pela Tabela 9, verifica-se que a produtividade média obtida no ano-safra de 2007/2008 foi de 3,27 tha^{-1}, produtividade média considerada alta em comparação com a média de produção nacional 2,82 t ha^{-1} (CONAB, 2008),demonstrando o potencial de produção da região oeste do Paraná. A precipitação pluviométrica média diária do período de estudo foi de 26 mm, com coeficiente de variação da ordem de 51%, favorecendo o desenvolvimento da cultura (EMBRAPA SOJA, 2007). A temperatura média de 24,33°C da região está dentro dos padrões em que a soja se adapta melhor (20 a 30°C).

A fim de investigar os agrupamentos de municípios com respeito à produtividade da soja (PROD) e variáveis agrometeorológicas (PREC, TMED, RSGM) aplicou-se o *FCM* com apoio do programa Matlab R2010a (MATLAB, 2011). Das dez tentativas realizadas para 4 agrupamentos, o algoritmo *Fuzzy c-Means* atingiu sua melhor condição de parada em 16 iterações. Para a primeira iteração, a função objetivo forneceu o valor de 4,108424 e, para a última iteração, o valor de 2,101852. Esses valores corroboram a convergência dos dados, minimizando a distância de qualquer dado de um agrupamento em relação ao seu centro (CHEN; WANG, 2009; ZHU et al*.,* 2012).

Após a execução do *FCM*, a alocação dos dados nos agrupamentos foi submetida ao índice *MPE* para qualificar a separação dos agrupamentos. Depois de 10 execuções do

MPE, foram identificados quatro agrupamentos. De acordo com as características do *MPE*, buscou-se validar a indicação de um número ótimo para a quantidade de agrupamentos. Isto foi realizado por meio do Método de Decisão pelo Maior Grau de Pertinência (*MDMGP*),que quantifica o grau de inclusão (sobreposição) entre agrupamentos. O resultado é apresentado na Tabela 10.

Tabela 10 Graus de inclusão entre os agrupamentos estabelecidos pelo método MDMGP

(A,B)	*S*(A,B)	(A,B)	*S*(A,B)	(A,B)	*S*(A,B)	(A,B)	*S*(A,B)
1,2	0,4274	2,1	**0,5374***	3,1	0,4576	4,1	0,2595
1,3	0,3383	2,3	0,3644	3,2	0,3922	4,2	0,2488
1,4	0,2306	2,4	0,2780	3,4	0,3087	4,3	0,2567

***** representam os agrupamentos com graus de sobreposição considerados altos (> 0,5). *S*(A, B) refere-se ao grau de sobreposição entre os agrupamentos A e B.

Em relação aos valores de inclusão entre agrupamentos, quanto menor o valor identificado, mais definido é o agrupamento, pois não se aproxima de outros. Em contrapartida, valores acima de 0,5 identificam uma nebulosidade entre os agrupamentos, pois são próximos. Desta maneira, a qualidade dos quatro agrupamentos, apontada pelo *MPE* foi considerada satisfatória para esse estudo, uma vez que apenas uma pertinência foi identificada acima de 0,5, a *S*(2,1) = 0,5374.

Estabelecido o número de agrupamentos e seus respectivos graus de inclusão, buscou-se quantificar a distribuição dos municípios segundo seus agrupamentos. Para tal, foram empregados os métodos de decisão *MDMGP* e *MDL β*. Os resultados são apresentados na Tabela 11, na qual a coluna (M) indica o número de municípios alocados para cada agrupamento e a coluna (%) indica o percentual de municípios alocados em relação ao total (48).

Tabela 11 Distribuição dos municípios nos agrupamentos de acordo com os métodos de pertinência *MDMGP* e *MDLβ*

	Agrup. 1		**Agrup. 2**		**Agrup. 3**		**Agrup. 4**		**Não Alocado**	
Método	(M)	(%)	(M)	(%)	(M)	(%)	(M)	(%)	(M)	(%)
MDMGP	13	27,08	13	27,08	9	18,76	13	27,08	0	0
MDLβ, com β=0,8	4	8,33	3	6,25	3	6,25	9	18,76	29	60,16
MDLβ, com β=0,65	10	20,83	7	14,58	8	16,67	9	18,75	14	29,17
MDLβ, com β=0,50	11	22,92	7	14,58	8	16,67	10	20,83	12	25

Para o método *MDMGP*, 100% dos municípios foram alocados nos agrupamentos estabelecidos. Isso se dá pelo fato de a alocação dos dados ser determinada pelo valor do maior grau de pertinência encontrado em cada agrupamento. Já a distribuição dos municípios nos agrupamentos pelo método *MDL β* depende do valor do limiar *β*. Quando *β*=0,8 (0,8 ≤ grau de pertinência ≤ 1), aproximadamente 60% dos municípios foram alocados e 40% não foram alocados (29 municípios). Para *β*=0,65 (0,65 ≤ grau de pertinência ≤ 1) há em torno de 70% de municípios alocados e 30% não-alocados (14 municípios). Para *β*=0,5

(0, 5 ≤ grau de pertinência ≤ 1) 75% dos municípios foram alocados e apenas 25% (12 municípios) não foram alocados.

A distribuição dos municípios impostas pelos métodos *MDMGP* e *MDL β* pode ser espacialmente visualizada na forma de mapa. Neste trabalho, os mapas foram gerados com auxílio do programa ArcMap 9.3 (ESRI, 2011), conforme ilustra a Figura 30, em que as tonalidades das cores, da mais clara para a mais escura, denota o nível da pertinência do município para o agrupamento em que está alocado. As tonalidades foram divididas em 3 classes: mais clara >0 e ≤0,5; intermediária >0,5 e ≤ 0,75; e a mais escura >0,75 e ≤1,0. Os valores apresentados para cada cor representam os níveis de similaridades obtidos para cada agrupamento, sendo o agrupamento4 o mais similar, da ordem de 78%. Estes níveis de similaridade foram obtidos por meio do ISC_i.

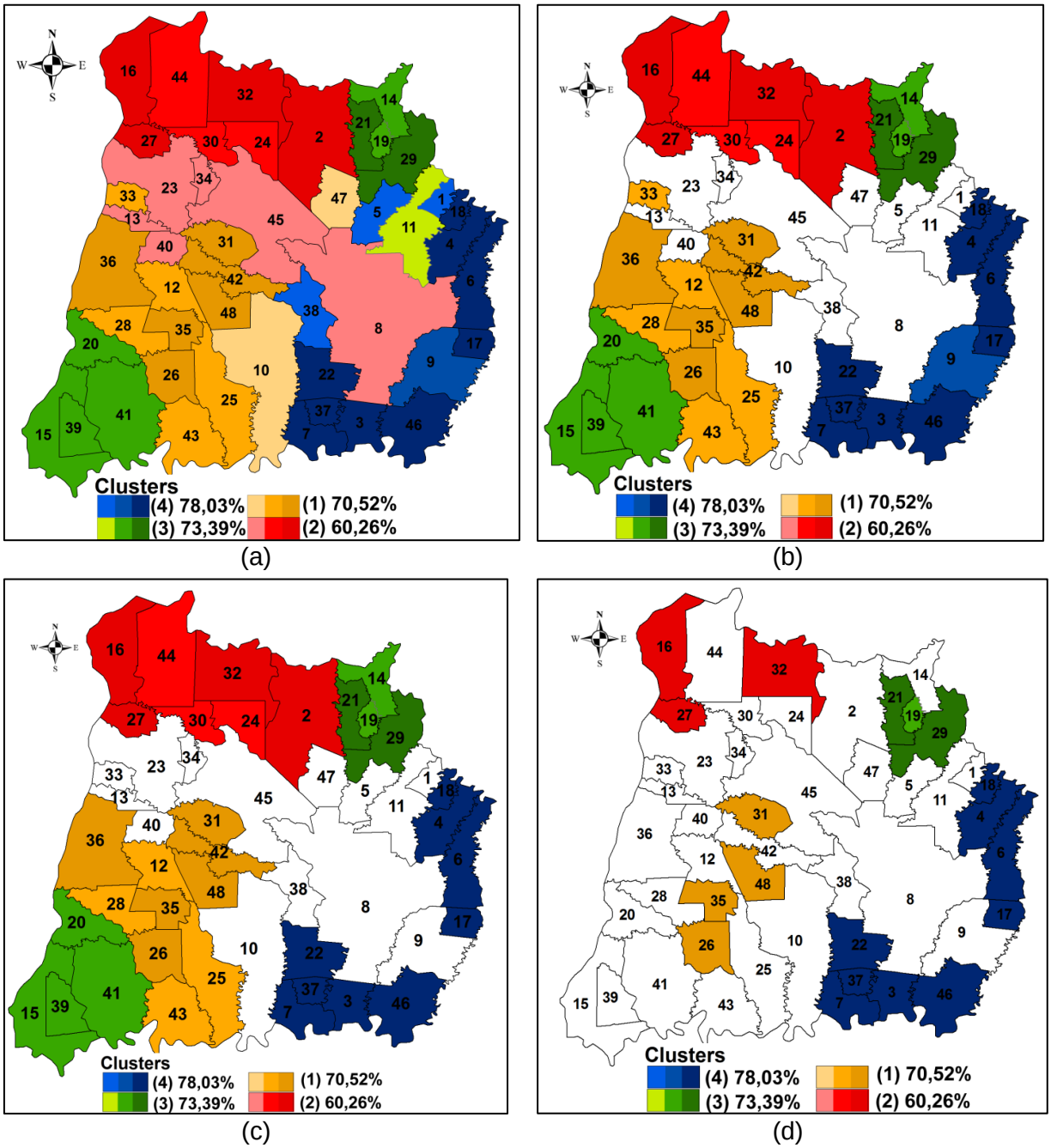


Figura 30 Distribuição dos agrupamentos impostos pelo FCM decorrente dos métodos: MDMGP(a), MDLβ 0,5 (b), MDLβ 0,65 (c) e MDLβ 0,8 (d)

Dentre os quatro mapas apresentados na Figura 30, optou-se pelo mapa da Figura 30(a), que representa a classificação gerada pelo método *MDMGP*, uma vez que esse método, por sua característica, classificou todos os municípios da região de estudo. Os demais mapas, decorrentes do método *MDL β*, permitiram verificar os níveis de pertencimento dos municípios aos agrupamentos identificados. Essa situação admite supor que, por exemplo, quando o limiar for 0,8, os municípios agrupados são mais similares, ou menos nebulosos.

Por meio da Tabela 12, que sintetiza as principais estatísticas descritivas dentro dos agrupamentos encontrados pelo método *MDMGP*, buscou-se analisar a produtividade da soja e as variáveis agrometeorológicas. Identificou-se que a maior produtividade média está localizada no agrupamento 1, onde também foram obtidos, em relação também à produtividade, menor desvio padrão e menor coeficiente de variação dentre todos os agrupamentos. O agrupamento número 1 também tem nele identificado a maior temperatura média, com os menores valores para desvio padrão e coeficiente de variação. Em relação à precipitação, o maior volume de chuva (total e média) ocorreu no agrupamento de número 2, que teve a menor produtividade média dentre os agrupamentos, identificando ainda, para a precipitação, o maior desvio padrão e coeficiente de variação em relação aos demais grupos. A radiação solar teve, também no agrupamento 2, o seu maior valor médio identificado.

Tabela 12 Estatísticas para as variáveis do estudo em cada agrupamento da região de estudo

Cluster	Variável	Mínimo	Máximo	Intervalo	Média	Mediana	Desvio Padrão	C.V. (%)
1	Prod	3,20	3,60	0,40	3,42	3,47	0,1298	3,80
	Prec	2322	8097	5775	4262	3968	1586	37,21
	TMedia	24,130	25,200	1,070	24,964	25,100	0,327	1,31
	RS	425,03	468,20	43,17	455,61	453,70	13,17	2,89
2	Prod	2,50	3,50	1,00	3,07	3,00	0,2475	8,06
	Prec	1917	10962	9045	4718	3485	2768	58,67
	TMedia	23,750	25,200	1,450	24,480	24,450	0,368	1,51
	RS	450,60	536,05	85,45	489,78	490,33	25,39	5,18
3	Prod	3,00	3,47	0,47	3,28	3,27	0,1793	5,47
	Prec	2500	6492	3992	4080	3803	1362	33,39
	TMedia	23,900	25,300	1,400	24,963	24,900	0,445	1,78
	RS	366,05	402,00	35,95	377,21	373,90	13,75	3,65
4	Prod	3,15	3,70	0,55	3,33	3,30	0,1661	4,99
	Prec	826	6448	5622	2812	2518	1560	55,48
	TMedia	22,900	23,900	1,000	23,119	22,900	0,418	1,81
	RS	402,00	450,60	48,60	427,35	430,10	12,59	2,95

C.V.: Coeficiente de Variação

Com o objetivo de validar os agrupamentos identificados pelo *FCM* (Figura 30a) e as estatísticas apresentadas na Tabela 12, o mapa apresentado na Figura 31 traz a produtividade da soja em destaque para cada município. Entre os principais municípios produtores de cereais, grãos e oleaginosas no Paraná, o IBGE destaca Assis Chateaubriand, Toledo, Cascavel e Terra Roxa. A dispersão do agrupamento 2 pode ser decorrente de prejuízos advindos de adversidades climáticas, com chuvas bastante irregulares dentro de uma mesma região para a safra 2007/2008. Sobre a situação de plantio e colheita das principais lavouras em todo o estado, depois das chuvas, o levantamento do Deral revela que por causa da escassez de chuvas, o plantio tardio comprometeu a produção da soja.

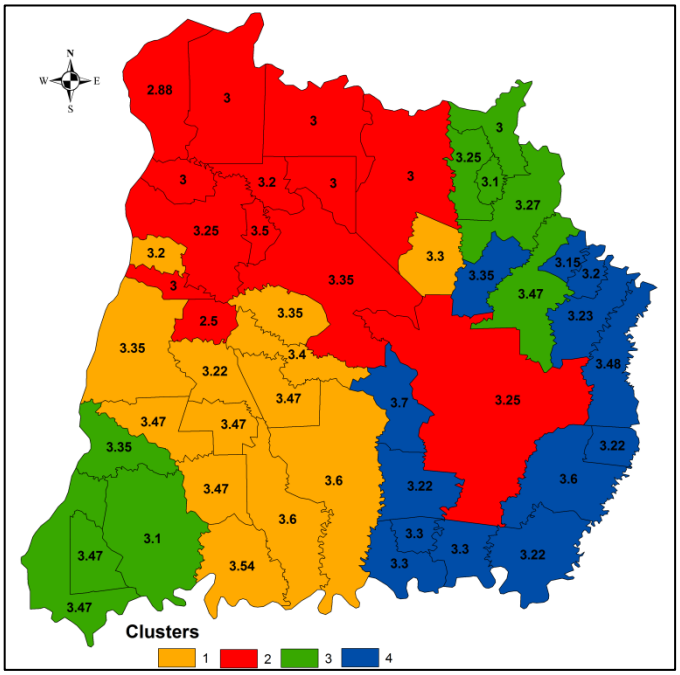


Figura 31 Mapa temático da produtividade da soja

4.4 CONCLUSÃO

Por meio da aplicação do algoritmo *Fuzzy C-Means*, obteve-se uma classificação dos municípios, com graus de similaridades da ordem de 60 a 78%. Dos dois métodos aplicados para classificação dos municípios, o Método de Decisão pelo Maior Grau de Pertinência (*MDMGP*), por sua característica de garantir que todos os dados pertençam a um grupo, apresentou melhores resultados.

Com as classificações obtidas foi possível identificar diferentes similaridades, tanto nos municípios que compuseram cada agrupamento, como entre os agrupamentos obtidos. A mensuração da similaridade entre os municípios de cada agrupamento foi possível por meio do Índice de Similaridade de *Clusters* (*ISC*$_l$). Em relação à similaridade entre agrupamentos, ela foi subsidiada pelo indicador que mede o grau de inclusão entre os agrupamentos.

Com estes resultados é possível subsidiar futuros estudos com metodologias que possam, por exemplo, considerar a correlação espacial entre as unidades de áreas (municípios).

4.5 **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/SP,v.27, n.2, p.237-248. 2008.

ASSAD, E. D.; MARIN, F. R.; MEDEIROS, S. R. E.; PILAU, F. G.; FARIAS, J. R. R.; PINTO, H. S.; ZULLO JR, J. Sistema de previsão de safra de soja para o Brasil. **Pesquisa Agropecuária Brasileira**, Brasília/DF, v. 42, n.5, p. 615-625, 2007.

BERLATO, M.A.; FONTANA, D.C.; GONÇALVES, H.M. Relação entre rendimento de grãos de soja e variáveis meteorológicas. **Pesquisa Agropecuária Brasileira,** Brasília/DF, v.27,n.5, p.695-702, maio, 1992.

BEZDEK, J. C. **Pattern recognition with fuzzy objective function algorithms**. New York: Plenum, 1981.256p.

BEZDEK, J.C.; PAL, S.K. **Fuzzy Models for Pattern Recognition**, IEEE Press, New York (1992)

BOYDELL, B.; MCBRATNEY, A. B. **Identifying potential within:** reflectance spectra: Algorithm development for remote sensing of field management zones from cotton-yield estimates. Precis.Agric. chlorophyll. J. Plant Physiol. 148:494–500. 3:9–23, 2002.

BUDAYAN, C.; DIKMEN, I.; BIRGONUL, M. T. Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping.**Expert Systems with Applications**, Reino Unido,v.36, n.9, p.11772–11781, 2009.

BURROUGH P.A., MCDONNELL R.A. **Principles of GIS**. Oxford University Press, Oxford, UK.1998.

CARMELLO, V. Vulnerabilidade agrícola da produção de soja na região metropolitana de Londrina – PR: análise da safra de 2005/06. **Revista Geográfica de América Central**.Costa Rica,v.2, No 47E, p. 1-16. (2011).

CHEN, W.; WANG, M.A fuzzy c-means clustering-based fragile watermarking scheme for image authentication.**Expert Systems with Applications**,Reino Undo,v. 36, n. 2, Part 1, p 1300–1307. 2009,

CONAB.**Acompanhamento da Safra Brasileira**, 2008.Disponível em:<http://www.conab.gov.br/conabweb/download/safra/12_levantamento_set2008.pdf>.Acesso em: 23 out. 2008.

DALLACORT, R.;FREITAS, P. S. L.;FARIA,R. T. de F.; GONÇALVES,.A. C. A.;REZENDE, R.; BERTONHA, A.Utilização do modelo Cropgro-soybean na determinação de melhores épocas de semeadura da cultura da soja, na região de Palotina, estado do Paraná. **Acta Scientiarum.Agronomy**, Maringá - PR, v. 28, n. 04, p. 583-589, 2006.

DUNN, J. C**.** A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. **Journal Cybernetics and Systems**,França, v. 3, n.3, 1973.

EMBRAPA, Empresa brasileira de pesquisa agropecuária**. Tecnologias de produção de soja - Paraná - 2007**. Londrina: Embrapa Soja, 2006. Disponível em: <http://www.cnpso.embrapa.br/download/tpsoja_2007_pr.pdf>. Acesso em: 24 jun. 2012.

EMBRAPA. Centro Nacional de Pesquisa de Soja. **Tecnologias de produção de soja** – região central do Brasil – 2009 e 2010. Londrina: EMBRAPA-CNPSo: EMBRAPA-CPAC: EMBRAPA-CPAO, 2008. 261p.

ESRI. **ArcGIS Spatial Analyst**. 2011. Disponível em: <http://www.esri.com/software/arcgis/extensions/spatialanalyst/surface.html>. Acesso em: 9 set. 2012

FARIAS, J. R. B. Limitações climáticas à obtenção de rendimentos máximos de soja.**Mercosoja 2011.** Quinto congreso de la Soja en el Mercosur. 2011.

FERREIRA, G. C. N.; GAMA, R. A. T. S. da; CAVALCANTI, M. C.; MOURA, A. M. de C. Organização automática de páginas Web para exibição em portais semânticos. WebMedia '08 Companion. Proceedings **of the XIV Brazilian Symposium on Multimedia and the Web** Pages 161-163. ACM, New York, NY, USA, 2008.

FRIDGEN, J. J.; KITCHEN, N. R.; SUDDUTH, K. A.; DRUMMOND, S. T.; WIEBOLD, W. J.; FRAISSE, C. W. **Management Zone Analyst (MZA)**: software for sub-field management zone delineation. Agron. J. 96, 100–108, 2004.

GOMES, A. da S.; PIRES, M. de M.; ALMEIDA, V. M.de; ROSADO, P. L.; SANTOS, P. R. P. SÃO JOSÉ, A. R. Análise dos territórios da região sudoeste da Bahia na perspectiva do desenvolvimento rural. **Revista Desenbahia**, Salvador/BA,v. 1, p. 59-82, 2011.

GUIERA, A. J. A.; Centeno, T. M.; DELGADO, M. R.; MULLER, M. Segmentação por Agrupamentos Fuzzy C-means em Imagens LiDAR Aplicados na Identificação de Linhas de Transmissão de Energia Elétrica. **Espaço Energia**,Paraná, v. 3, p. 24-31, 2005.

IBGE - Instituto Brasileiro de Geografia e Estatística. **Mapas Digitais**. Disponível em: <http://www.ibge.gov.br/home/download/geociencias.shtm>. Acesso em: 20 set. 2012

JACOX, E. H.; SAMET, H. ACM Transactions on Database Systems (TODS).**Spatial join techniques**, Arizona/USA, [s.l.] v. 32 n. 1, p.7-es, 2007.

MALHOTRA, N. **Pesquisa de marketing**: uma orientação aplicada. Trad.Laura Bocco. 4 ed. Porto Alegre: Bookman, 2006. 720p.

MAGNUSSON, W. E.; MOURÃO, G. **Estatística sem matemática**: a ligação entreas questões e a análise. Curitiba: 2003.136p.

MATLABR2010a. Disponível em: <http://www.mathworks.com/help/techdoc/?s_cid=ML2012_bb_doc>. Acesso em: 23 jun. 2011.

MCBRATNEY, A. B.; MOORE, A. W. **Application of fuzzy sets to climatic classification**.Agr. Forest Meteorol., 35, p. 165–185, 1985.

NG, Hsiao Piau; ONG, Sim Heng; Weng, KELVIN; FOONG, Chiong; GOH, Poh Sun; Wieslaw; Nowinski L. Fuzzy c-means algorithm with local thresholding for gray-scale images. **International Journal on Artificial Intelligence Tools**.Reino Unido, v. 17, n. 4, p.765–775, 2008.

ODEH, I.O.; McBRATNEY, A.B.; CHITTLEBOROUGH,D.J. Soil pattern recognition with fuzzy-c-mean: applicationto classification and soil-landform interrelationships**. Soil Science Society American Journal**, USA,v. 56, p. 505-516, 1992.

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

SEAB, 2010. **Agropecuária - Estatísticas, Produção agropecuária, Produção Agrícola Paranaense por Município - últimas 5 safras.** Disponível em: <http://www.agricultura.pr.gov.br>.Acesso em: 11 set. 2012

SIMEPAR. Sistema Meteorológico do Paraná. 2010.

SUN, X.; ZHAO, Y.; WANG, H.; YANG, L., QIN, C.; ZHU, A.; ZHANG, G.; PEID, T.; LI, B. **Sensitivity of digital soil maps based on FCM to the fuzzy exponent and the number of clusters**. Geoderma.V. 171-172, pages 24-34, 2012.

TAYLOR, J. A.; MCBRATNEY, A. B.; WHELAN, B. M. Establishing management classes for broadacre agricultural production.**Agronomy Journal**,Madison/USA, v.99, p.1366-1376, 2007.

THOMAS, J.F. Ontogenetic and morphological plasticity in crop plants. In: BOOTE, K.J. *et al.*(Comp.). **Physiology and determinations of crop yield.** Madison: ASA/CSSA/SSSA, Cap. 7B, p. 181-185, 1994.

TOLEDO, N. T.; MULLER, A. G.; BERTO, J.L.; MALLMANN, C. E. S. Ajuste do modelo fototérmico de estimativa do desenvolvimento e do índice de área foliar de soja. **Revista brasIleira de engenharia agrícola ambiental** [online], Campina Grande/PB. 2010, v.14, n.3, pp. 288-295. ISSN 1807-1929.

UNWIN A.; UNWIN D. Spatial Data Analysis with Local Statistics. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Londres/Inglaterra,v.47, n.3, p.415–421, 1998.

VEENHOF, H. M.; APERS, P. M. G.; HOUTSMA, M. A. W. Optimization of spatial joins using filters.Advances in databases. Lecture Notes in Computer Science, Volume 940/1995, 136-154, 1995.

WANG, H; FEI, B.**A modified fuzzy C-means classification method using a multiscale diffusion filtering scheme**.Med. Image Anal.13(2), 193-202 (2009).

YAN, L.; ZHOU, S.; FENG, L.; HONG-YI, L. **Delineation of site specific management zones using fuzzy clustering analysis in a coastal saline land**.Computers and Electronics in Agriculture, p.174-186, 2007.

ZHU, W.; JIANG, J.; SONG, C.; BAO, L. Clustering Algorithm Based on Fuzzy C-means and Artificial Fish Swarm**.Procedia Engineering**, Reino Unido,v. 29, p. 3307–3311, 2012

CONSIDERAÇÕES FINAIS

Por meio das metodologias apresentadas e aplicadas com as variáveis disponíveis, foi possível desenvolver estudos relacionados à estatística espacial de área na região oeste do estado do Paraná, com vistas à análise da produtividade da soja nos municípios desta região. Nestes estudos, por meio de indicadores de autocorrelação espacial, a identificação de significância estatística possibilitou a classificação de municípios em forma de agrupamentos, onde similaridades podem ser estudadas e trabalhadas. Em conjunto com esta técnica, a geração de modelos de regressão múltipla espacial permite subsidiar estudos explicativos e preditivos em relação à produtividade da soja, tendo sempre as variáveis agrometeorológicas associadas aos modelos obtidos e validados.

Este trabalho buscou apresentar métodos para serem aplicados na estatística espacial de área na produtividade da soja e fatores agrometeorológicos na região oeste do estado do Paraná. Os dados utilizados estão relacionados aos anos-safra de 2000/2001 a 2007/2008, sendo as variáveis: produtividade da soja (t ha^{-1}) e agrometeorológicas, tais como precipitação pluvial (mm), temperatura média (^{o}C) e radiação solar global média (W m^{-2}). Em uma primeira fase foram utilizados índices de autocorrelação espacial (Moran Global e Local) e apresentados modelos de regressão espacial múltipla, com avaliações de desempenho. Em uma segunda etapa foram realizadas análises de agrupamento espacial por meio da estatística multivariada, buscando identificar associações no mesmo conjunto de variáveis, porém com um número maior de anos-safra. Finalmente, os dados de um ano-safra, foram aplicados em uma abordagem baseada em agrupamento difuso, por meio do algoritmo *Fuzzy c-Means*, tendo a similaridade medida pela definição de um índice com este objetivo.

Com o intuito de identificar similaridades entre os municípios, fez-se uso também da estatística multivariada, com índices para validação dos agrupamentos identificados e apresentados por meio de dendogramas. Neste estudo, além dos atributos disponíveis foi também utilizado o índice de autocorrelação espacial local univariado (*LISA*), responsável pela indicação do nível de autocorrelação para a variável produtividade da soja. Concluindo os objetivos propostos para este estudo, foi feito uso da técnica nebulosa (*fuzzy*), por meio do algoritmo *Fuzzy c-Means* para uma nova classificação dos municípios, porém, nesta etapa, a informação espacial trabalhada foram os centroides dos municípios. Para a validação da similaridade identificada nos agrupamentos, foi definido um índice para esta mensuração, o ISC_i.

Com os resultados obtidos com a aplicação das técnicas propostas, verifica-se que estas ferramentas são extremamente úteis no estudo de classificação, explicação e estimativa de produtividade. Entretanto, recomenda-se para estudos futuros um

aprofundamento nas técnicas relacionadas aos agrupamentos difusos em um conjunto de dados maior, unindo-se, inclusive à mineração de dados.