COMPONENTES PRINCIPAIS E GEOESTATÍSTICA PARA CARACTERIZAR A VARIABILIDADE ESPACIAL DA PRODUTIVIDADE DE SOJA NO ESTADO DE MINAS GERAIS EM ESCALA MUNICIPAL

João Marcos Louzada [1], Marcelo Silva de Oliveira [2], Marcelo de Carvalho Alves [3]

1 Introdução

A produção de soja no Brasil expandiu-se rapidamente no início dos anos 70 com uma produção tipicamente agroindustrial. Atingiu um pico em 1989, com 24 milhões de toneladas, caindo no início da década de 90 (inferior a 20 milhões ton/ano), mas vem se recuperando progressivamente. Atualmente, o Brasil é o segundo maior produtor mundial de soja, sendo o Estados Unidos o atual lider de produção, com uma safra de $71.448t$ em 2007/08. Já a produção brasileira atingiu a marca de $61.000t$ (atualização em agosto de 2007) nesse mesmo período. Devido à grande importância dessa cultura para economia brasileira, visto que o Brasil é o maior exportador da mesma, utilizou-se os métodos da Geoestatística para se detectar o padrão de dependência espacial, e, posteriormente, mapear as áreas de produção por meio do interpolador linear de krigagem. Assim, pretende-se com esse estudo gerar informações de suma importância para o manejo da soja, bem como, a verificação espacial da cultura para investimentos futuros.

2 Material e Métodos

2.1 Dados e recursos computacionais

Os dados sobre produtividade de soja (ton/ano) no Brasil foram referentes ao período de 1990 a 2005. A produção foi obtida pela rede de coleta do IBGE (Instituto Brasileiro de Geografia e Estatística 2007), mediante consulta a entidades públicas e privadas, a produtores, técnicos e órgãos ligados direta ou indiretamente aos setores da produção, comercialização,industrialização e fiscalização de produtos agrícolas. A unidade de investigação no inquérito estatístico foi em nível de município, cuja as sedes desses munícipios foram georeferenciadas, com o intuito de se realizar as análises Geoestatística.

As análises foram realizadas no programa **R** (R Development Core Team 2007) , da seguinte modo: Para a análise de Geoestatística utilizou-se o pacote "geoR" (jr. e Diggle 2001) e a ACP's foi desenvolvida pacote "stats" (R Development Core Team 2007).

2.2 Modelos da Geoestatística

A variável alvo, pode ser compreedida como uma realização de um processo estocático, sendo investigada sob condição de *variável regionalizada* que são processos típicos de mecanismos gerados por realizações de funções de variáveis aleatórias (Isaaks e Srivastava 1989).

[1]Lic. Matemática, Doutorando, Depto. de Ciências Exatas, Campus dea UFLA, jmarlo@bol.com.br

[2]Engenheiro Agrícola, Prof. Doutor, Depto. de Ciências Exatas , Campus da UFLA, marcelo.oliveira@ufla.br

[3]Dr. Bolsista de pós-doutorado Júnio/CNPq, Depto. de Engenharia, Campus da UFLA, marcelocarvalhoalves@gmail.com

Esse modelo de processo é visto como uma quantidade que varia continuamente por toda a região $\mathcal{R}$ do inquérito. Pode ser simbolizada por $\mathbf{z}(\mathbf{x}) = \{z(x_1), z(x_2), \cdots, z(x_n)\}$, $\forall\ \mathbf{x} \in \mathcal{R}^2$, onde n representa o total de observações tomadas nos locais amostrados representados por $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$.

O estudo de modelos de autocorrelação entre dados distribuídos espacialmente é denominado de análise estrutural ou *modelagem do semivariograma*, a ferramenta central da Geoestatística. O estimador clássico do semivariograma proposto por (Matheron 1963) é dado pela equação:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum (Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h}))^2,$$

em que, $Z(\mathbf{x})$ e $Z(\mathbf{x}+\mathbf{h})$ são variáveis regionalizadas, $N(\mathbf{h})$ determina o número de pares de valores medidos, das variáveis em estudo, separados por um vetor $\mathbf{h}$. O gráfico plotado em relação aos correspondente valores de $\mathbf{h}$ é denominado variograma ou semivariograma.

Desse modo, primeiramente, efetuou-se o cálculo do semivariograma experimental com base nos dados de soja, devidamente arranjados, para se detectar a presença de autocorrelação espacial. É importante salientar que para a realização desse cálculo, considerou-se as coordenadas geográficas de cada município, sendo a sede municipal escolhida como ponto de referência dos mesmos para se tomar as coordenadas espaciais. De fato, a aplicação do semivariograma como medida de continuidade espacial requer apenas que os dados satisfaçam a hipótese intrínseca para uma variável regionalizada (Jounel e Huijbregts 1991). Isso implica que a componente determinística, $m_{\mathbf{x}}$, deve ser constante (sem tendência na região), que se formaliza: $E[Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})] = 0$ ou $E[Z(\mathbf{x}+\mathbf{h})] = E[Z(\mathbf{x})] = m_{\mathbf{x}} = m$, e, que a variância das diferenças depende somente do vetor distância $\mathbf{h}$, conforme descrito abaixo:

$$Var[Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})] = E[(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2] = 2\gamma(\mathbf{h}).$$

Já a hipótese de estacionaridade de segunda ordem é mais restritiva, e, exige que a covariância existente entre quaisquer dois pares, $Z(\mathbf{x})$ e $Z(\mathbf{x}+\mathbf{h})$, separados por um vetor distância $\mathbf{h}$, seja somente dependente desse vetor distância $\mathbf{h}$. Então:

$$C(\mathbf{h}) = Cov[Z(\mathbf{x}), Z(\mathbf{x}+\mathbf{h})] = E[Z(\mathbf{x}) \cdot Z(\mathbf{x}+\mathbf{h})] - m^2,\ \forall\ \mathbf{x}.$$

A estacionariedade da covariância também implica na estacionariedade do semivariograma, definido por:

$$2\gamma(\mathbf{h}) = E[(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2].$$

Após o cálculo de semivariogram experimental, ajustou-se os modelos teóricos conforme trata a literatura, ver (Isaaks e Srivastava 1989; Cressie 1993; Chilès e Delfiner 1999). A princípio, testou-se o modelo *cardinal-seno* (modelo wave) - pertencente a família dos modelos "hole effect", utilizando-se dois métodos de ajuste: mínimos quadrados ordinários (MQO) e mínimos quadrados ponderados (MQP). Posteriormente, verificou-se a qualidade desses ajustes comparando as somas de quadrados dos modelos ajustados; isto é: escolheu-se o modelo com a menor soma de quadrados minimizada. Então, conhecidas as estimativas dos parâmetros (pepita, patamar e alcance) do semivariograma teórico, construiu-se o modelo de covariância espacial que foi determinante no processo de predição.

O modelo wave é dado pela seguinte expresão:

$$C(|\mathbf{h}|) = \left(\frac{a}{|\mathbf{h}|}\right) \sin\ \left(\frac{|\mathbf{h}|}{a}\right),\ \text{com } \mathbf{h} \in \mathcal{R}^3.$$

A notação $|\mathbf{h}|$ indica que o padrão de dependência espacial é o mesmo em toda direção (fenômeno isotrópico), para detalhes consulte (Isaaks e Srivastava 1989). $C(|\mathbf{h}|)$ é a covariância em função da distância $\mathbf{h}$, e a é o alcance teórico. Para o momento, um semivariograma é dito exibir um "hole effect" quando o seu crescimento não é monótono, e, esse efeito pode aparecer em modelos com ou sem patamar. Para mais informações consulte (Chilès e Delfiner 1999)

Um dos principais objetivos dos métodos geoestatísticos é o de predizer valores não amostrados, em um dado local espacial (pontual ou área), a partir de observações amostradas em outros locais. As técnicas de predição(krigagem) são versões refinadas das técnicas de médias móveis ponderadas usadas por Krige (Henley 1981).

Construído o modelo de covariância espacial, conforme descrito acima, realizou-se o procedimento de krigagem com o obejtivo de mapaer toda as possíveis regiões de cultivo da soja no estado de Minas Gerais. As várias técnicas de krigagem são todas baseadas em um simples modelo linear, definido como segue:

$$\hat{Z}_{KO}(\mathbf{x}_0) = \sum_{\alpha=1}^{k} w_\alpha Z(\mathbf{x}_\alpha),$$

que é uma média ponderada dos dados, sendo k o número de pontos amostrais vizinhos utilizados na predição, em um ponto $\mathbf{x}_0$, e w_α são os pesos atribuídos a cada realização $Z(\mathbf{x}_\alpha)$ do processo estocático. Para que o estimador de krigagem seja ótimo (sem tendência e de variância mínima), necessariamente, deve-se assegurar condição de universalidade, em que $\sum_\alpha^m = 1$ e de otimalidade (minimiza a variância do erro), isto é, $\sigma^2_{\mathbf{x}_0} = E[\hat{Z}(\mathbf{x}_0) - Z(\mathbf{x}_0)] = 0$, onde $\sigma^2_{\mathbf{x}_0}$ é a variância da estimativa, e $E[\hat{Z}(\mathbf{x}_0) - Z(\mathbf{x}_0)]$ é a esperança matemática do erro da estimativa (Burrough e Mcdonnell 1998).

2.3 Análise de componente principal

A análise de componentes principais (ACP's) pode ser compreendida como uma técnica de transformação linear que transforma um dado conjunto de variáveis correlacionadas em fatores não correlacionados. Em geral, essa técnica pode ser usada com as seguintes finalidades: redução dos dados, detecção de "outlier" multivariado, decifrar a matriz de correlação, identificar fatores subjacentes e detectar correlação intrínseca (Wackernagel 2003). Nesse trabalho, buscou-se, por meio da ACP's, explicar a estrutura de variância-covariância por meio de combinações lineares das variáveis originais. Dessa forma, os objetivos da análise foram, principalmente: reduzir a dimensão original dos dados, facilitar a interpretação espacial das análises e determinar a existência de correlação entre os anos de produtividade avaliados.

Calculou-se os componentes principais (CP's) relativos aos dezesseis anos de produção de soja, e, utilizou-se a combinação linear com máxima variância do componente principal com a maior porcentagem de explicação da produtividade de soja ao longo do período considerado. O critério de escolha dos CP's foi por inspeção gráfica: observando o gráfico de cotovelos (ou "screeplot"), omitido nesse artigo.

(Y_i)	$\text{Var}(Y_i)$	% explicada	% acumulada	–
Y_1	$12,86$	$91,920$	$91,92$	–
Y_2	$0,66$	$4,720$	$96,65$	–
Y_3	$0,12$	$0,008$	$97,47$	–
Y_4	$0,96$	$0,007$	$98,57$	–
Método	Pepita	Patamar	Alcance	SQM
MQO	$1.877.192,88$	$187.855.786,41$	$800.429,36$	$1,66 \times 10^{18}$
MQP	$0,00$	$1.881.128.578,44$	$810.831,60$	$6,44 \times 10^{21}$

TABELA 1: Resultados da análie de componentes principais (Y_i); estimativas dos parâmetros do modelo teórico wave: efeito pepita, patamar, alcance prático (unidade de medida é o metro linear) e a soma de quadrados minimizada: MQO significa mínimos quadrados ordinários, e MQP signfica mínimos quadrados ponderados.

Algebricamente, os CP's são combinações lineares das p variáveis aleatórias (VA's) $X_1, X_2, \cdots, X_P$, e dependentes somente da matriz de covariância amostral $\mathbf{S}$ (ou da matriz de correlação ρ multivariada. Dado o vetor de variáveis aleatórias $X' = [X_1, X_2, \cdots, X_P]$ e a sua matriz de covariância $\mathbf{S}$ com os autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Considerando-se a combinação linear $Y_i = \mathbf{e_i'X}$, com $\{i = 1, \cdots, p\}$, os CP's foram as combinações lineares não correlacionadas $Y_1, Y_2, \cdots, Y_p$ com as maiores variâncias possíveis. Seu desenvolvimento não requer suposição de normalidade multivariada.

3 Resultados e Discussão

A Tabela 1 mostra os resultados obtidos da ACP's, bem como, os parâmetros estimados do modelo de covariância, a partir do ajuste do semivariograma teórico (modelo wave). O primeiro CP explicou, aproximadamente, 92% da variação total dos dados, conforme mostra a Tabela 1. Assim, a técnica de componentes principais pôde ser utilizada para substituir as variáveis originais (16 anos de produção soja), sem muita perdas de informações. Embora, a análise de CP seja uma técnica para dados multivariados, nesse caso apenas gerou-se os "escores" para o cálculo do semivariograma com o intuito de inferir conjuntamente sobre o fenômeno.

A Figura 1 mostra que foi possível descrever o padrão de dependência espacial para os dados de soja, por meio da análise variográfica. Selecionou-se o modelo wave ajustado pelo método de MQO, visto que esse método obteve a menor soma de quadrados (veja Tabela 1). O alcance encontrado foi em torno de $800km$, indicando que há uma grande concentração de produção em minicípios próximos uns do outro, formando um grande cinturão ao longo da região noroeste de Minas Gerais e na região do triângulo mineiro - De fato, há um forte indício de que existe uma grande extensão com o cultivo da soja contribuindo para o potêncial econômico dessas regiões. Esse fato pode ser constatado por meio do mapa de krigagem ordinária, exibido na Figura 2, onde a variabilidade espacial da produtividade de soja estão em destaque.

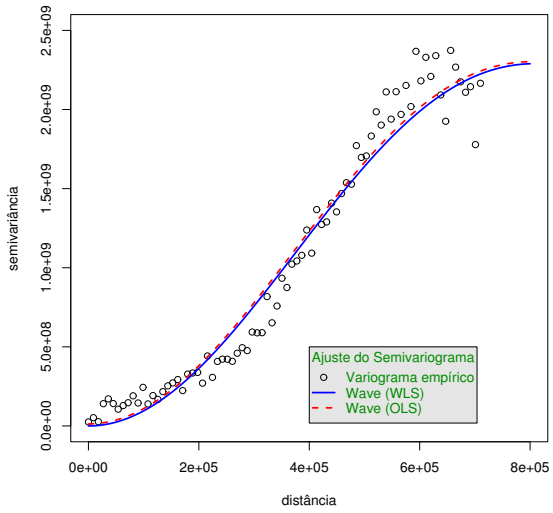

FIGURA 1: Semivariogramas teóricos: modelo cardinal-seno (wave) ajustados pelos métodos MQO e MQP, com patamar bem definido.

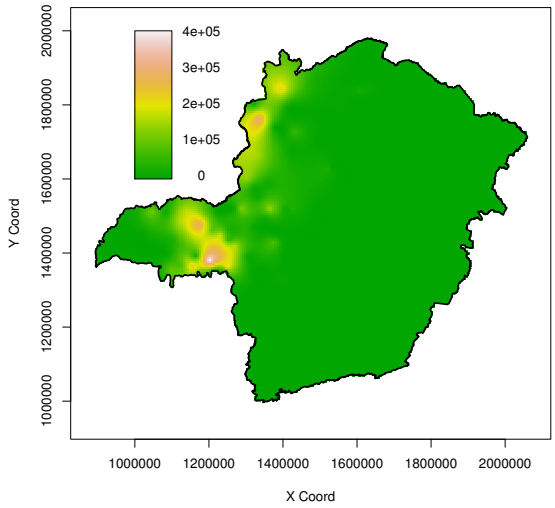

FIGURA 2: Mapa de krigagem ordinária, mostrando a variabilidade espacial da produtividade de soja no Estado de Minas Gerais.

4 Conclusões

Os dados de produtividade de soja do estado de Minas Gerais, avaliados no período de 1990 a 2005, apresentaram dependência espacial, com base na análise descritiva do semivariograma e por meio do ajuste do medelo "hole effect" (wave).

A Geoestatística a aliada à técnica de análise de componentes principais, se torna uma poderosa ferramenta de análise quando se estuda uma massa de dados com grande número de realizações no tempo - nesse caso, a técnica de componente principal facilitou bastante a análise Geoestatística.

Com a técnica da krigagem ordinária foi possível mapear as regiões produtoras de soja, do estado de Minas Gerais, satisfatoriamente, no período de 1990 a 2005.

REFERÊNCIAS BIBLIOGRÁFICAS

[Burrough e Mcdonnell 1998]BURROUGH, P. A.; MCDONNELL, R. A. *Principles of geografhical information systems: Spatial Information Systems and Geostatistics*. 2. ed. Oxford: Oxford University Press, 1998. 333 p.

[Chilès e Delfiner 1999]CHILÈS, J.-P.; DELFINER, P. *Geostatistics: modeling spatial uncertainty*. United States of America: John Wiley and Sons, 1999.

[Cressie 1993]CRESSIE, N. A. *Statistics For Spatial Data*. Revised edition. Iowa State Univesity, New York: A Wiley Interscience Publication, 1993.

[Henley 1981]HENLEY, S. *Nonparamentric Geostatistics*. London-U: Elsevier Applied Science Publishers Ltd, 1981. (1 ed.). 145p.

[Instituto Brasileiro de Geografia e Estatística. 2007]INSTITUTO Brasileiro de Geografia e Estatística.: Produção agrícola municipal: Culturas temporárias e permanentes. Brasil, 2007. Disponível em: <http://www.sidra.ibge.gov.br>.

[Isaaks e Srivastava 1989]ISAAKS, E. H.; SRIVASTAVA, R. M. *Applied Geostatistics*. New York: Oxford University Press, 1989.

[Jounel e Huijbregts 1991]JOUNEL, A. G.; HUIJBREGTS, C. J. *Mining Geostatisitcs*. 5. ed. Bury St Edumunds, Suffolk: St Edmundsbury Press Limited, 1991. 600 p.

[Jr e Diggle 2001]JR, P. J. R.; DIGGLE, P. J. geoR: a package for geostatistical analysis. *R-NEWS*, v. 1, n. 2, p. 14–18, June 2001. ISSN 1609-3631. Disponível em: <http://CRAN.R-project.org/doc/Rnews/>.

[Matheron 1963]MATHERON, G. Principles of geostatistics. n. 58, p. 1246–1266, 1963.

[R Development Core Team 2007]R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2007. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

[Wackernagel 2003]WACKERNAGEL, H. *Multivariate Geostatistics: an introduction with applications*. Berlin: Springer, 2003. (3ed.(rev.)). 387p.