Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura e do Abastecimento

Comunicado Técnico da Embrapa Informática Agropecuária

Nº. 14, março/2001 ISSN 1516-5620

Ministério da Agricultura e do Abastecimento

ANÁLISE DE CORRESPONDÊNCIA - UMA FERRAMENTA ÚTIL NA COMPARAÇÃO DE MAPAS DE PRODUTIVIDADE

José Ruy Porto de Carvalho[1], Sidney Rosa Vieira[2], Regina Célia Carvalho Pinto Moran[3]

Termos de indexação: Mapa de produtividade; Variabilidade espacial; Variabilidade temporal; Análise de correspondência simples; Análise de correspondência múltipla; Krigagem.
Index Terms: Yield mapping; Spatial variability; Temporal variability; Simple correspondence analysis; Multiple correspondence analysis; Kriging.

1. Introdução

A agricultura tradicional, influenciada pelo processo de globalização da economia, vem sendo desafiada na obtenção de níveis de produtividade que sejam compatíveis internacionalmente. Entretanto, as medidas que vêm sendo adotadas sugerem a utilização de áreas de grande extensões as quais são consideradas homogêneas. Neste sentido, utiliza-se o conceito da necessidade média para aplicação de insumos (Capelli, 1999), ou seja, os mesmos níveis de fertilizantes, defensivos, água, etc. - são aplicados em toda área sem considerar as suas necessidades específicas e localizadas, resultando em uma produtividade não uniforme para a lavoura, além do fato de acarretar problemas ambientais e ecológicos.

As tecnologias e práticas emergentes da Agricultura de Precisão vêm combater o conceito da necessidade média. O manejo da variabilidade é o principal foco de sua atenção. Três formas de variabilidade são imediatamente detectadas: espacial, temporal e preditiva. A variação espacial é a variação vista no campo onde, por exemplo, as bordas do campo apresentam sempre menor produção do que seu interior, facilmente identificadas nos mapas de produtividade ou de solos. A variação temporal é a variação que ocorre quando se compara mapas de produtividade de um ano para outro. A variação preditiva é vista quando são feitas predições sobre qualquer variável no futuro, mas estas predições não são realizadas (Blackmore & Larscheid, 1998).

Com o uso de GPS (Global Positioning Systems) na área agrícola, combinado com dados de produção obtidos de uma colheita, é possível a produção de mapas de produtividade. Estes mapas são centrais para a agricultura de precisão, pois promovem o uso mais eficiente das informações, promovendo desta maneira a administração da variabilidade no campo.

São freqüentes comentários como: "Olha como estes mapas são similares". Mas o que determina similaridade? Seriam algumas poucas áreas que aparecem moldadas semelhantemente no mapa? Seriam fatores como seleção de cor, número de classes e pontos de ruptura temáticos? Seriam a presença, ausência ou controle da variabilidade fatores de influência em similaridade? É importante lembrar que os padrões formados pelas características dos mapas são subjetivos. Sugestão pode influenciar na interpretação de mapas. Como então proceder para comparar mapas de produtividade?

2. Identificação do problema

O método estatístico usualmente utilizado para avaliar diferenças de médias de dois grupos de dados é o teste t de Student.

[1]Ph.D. em Estatística, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (jruy@cnptia.embrapa.br)
[2]Ph.D. Em Conservação de Solos, Pesquisador do Instituto Agronômico de Campinas, Caixa Post
[3]Ph.D. em Estatística, Professora do IMEC-UNICAMP, Caixa Postal 6065, Barão Geraldo - 130

Tiragem: 120 exemplares

Entretanto, as condições para que este teste seja apropriadamente aplicado são de que os dados devam ser independentes e e normalmente distribuídos com variância homogênea. O problema é que estas condições raramente se aplicam em mapas de produtividade devido as dependências espacial e temporal das observações. A associação entre dois mapas pode também ser medida através do coeficiente de correlação entre todos os dados ou grupos de dados que compõem estes mapas. Correlação é uma medida empírica da intensidade da associação entre duas variáveis. Para que esta técnica estatística tenha validade, assume-se que existe na população uma dependência linear entre as variáveis. Esta é uma condição aceitável quando a amostra é obtida de uma população com distribuição normal bivariada (Draper & Smith, 1981). Infelizmente esta pressuposição é difícil de acontecer nas condições atuais.

Uma outra forma de efetuar a comparação de mapas é através do Índice KAPPA de concordância, também conhecido como KHAT ou KAPPA coeficiente de concordância. Esta medida estatística foi introduzida por psicólogos e adaptada como medida de concordância para mapas por Congalton & Mead (1983). Este índice testa a associação entre mapas. Ele nos ajuda a entender se os mapas diferem devido à alguma variação casual ou se há uma real concordância. Ele também permite a comparação de dados com estruturas definidas Rosenfield & Fitzpatrick-Lins (1986). O índice KAPPA varia de 0 à 1 onde 0 indica que os resultados acontecem totalmente ao acaso e 1 indica concordância perfeita.

Em resumo, existem diversas maneiras de se tratar o problema. Tem-se também os métodos de classificação cruzada, coeficiente de similaridade de Qui-Quadrado e Cramer, etc. Entretanto todos dependem de certas condições pré-estabelecidas que são difíceis de acontecer para dados temporal e espacialmente dependentes.

A Análise de Correspondência é uma técnica multivariada para análise exploratória de dados categorizados. A preocupação com a análise destes tipos de dados já vem desde o começo do século. Mas, somente na década de 60, por intermédio de Jean-Paul Benzécri, que definiu um método mostrando suas propriedades algébricas e geométricas denominado "Analyse Factorialle des Correspondences", que esta técnica se difundiu (Benzécri, 1992). Ela converte uma matriz de dados não negativos em um particular tipo de gráfico que exibe as linhas e colunas da matriz como pontos de um espaço vetorial de dimensão menor que a original, de maneira que as relações entre as linhas, entre as colunas e entre linhas e colunas possam ser interpretadas. É exatamente por estar interessada em estudar a correspondências entre variáveis, que esta técnica recebeu o nome de Análise de Correspondência. Sua geometria e álgebra fazem com que pertença a uma família de técnicas de disposição gráfica que são baseadas em aproximação de uma matriz por outra de posto menor, por meio da decomposição em valores singulares. Ou seja, o objetivo desta análise é achar um subespaço que melhor ajuste a nuvem de pontos no espaço euclidiano. Este ajuste é feito pelo método de quadrado mínimo ponderado onde a distância euclidiana generalizada (ponderada) é utilizada em um sistema de massas pontuais (Greenacre & Hastie, 1987).

3. Objetivo

O objetivo deste trabalho é apresentar a análise de correspondência como um método estatístico multivariado que pode ajudar na comparação de mapas de produtividade.

4. Hipótese

O uso de mapas de produtividade está estritamente relacionado a variabilidade de seus componentes, tornado-se desta maneira, preocupante seu uso indiscriminado sem se considerar as variações espaciais e temporais inerentes.

5. Metodologia

Na área experimental do Instituto Agronômico de Campinas - IAC, localizada em Votuporanga, São Paulo, em latossolo vermelho arenoso, foram coletadas amostras de componentes de produção para arroz em 1994 - A4, milho em 1995 - M5, algodão em 1996 - A6 e milho em 1998 - M8, em uma malha com 110 pontos, espaçados de 10 x 10 m. Em cada ponto, foram coletados as partes aéreas de 5 m^2, de onde também foi medida a produtividade.

Os mapas de produtividade ou mapa de isolinhas para as culturas dos quatro anos foram obtidos pela interpolação dos valores em qualquer posição no campo de estudo, sem tendência e com variância mínima, através do método de interpolação chamado krigagem. Segundo Vieira et al. (1983), a precisão da localização das isolinhas entre dois pontos é extremamente dependente da densidade de pontos por área e, conseqüentemente, da distância entre os pontos. A maneira mais comum de localizar uma isolinha entre dois pontos é pela interpolação linear.

A estimativa Z* na posição x_0, pode ser obtida através de:

$$Z^*(x_0) = \sum_{i=1}^{N} \lambda_i Z(x_i) \qquad (1)$$

onde N é o número de valores medidos, Z(x_i) é o valor medido na posição x_i e λ_i é o peso associado ao valor medido na posição x_i. Segundo Vieira (1997), impondo as condições de estimativa sem vício e com variância mínima na equação 1, chega-se ao sistema de equações da krigagem:

$$\sum_{j=1}^{N} \lambda_j \gamma(x_i, x_j) + \mu = \gamma(x_i, x_0), \quad i = 1 \text{ a } N$$

$$\sum_{j=1}^{N} \lambda_j = 1 \qquad (2)$$

onde:

$\gamma(X_i, X_j)$ é a semivariância estimada usando o modelo ajustado ao semivariograma, correspondente à distância entre os pontos localizados nas posições x_i e x_j; $\gamma(x_i, x_o)$ é a semivariância correspondente à distância entre os pontos localizados nas posições x_i e x_o. A solução do sistema de equações de krigagem (2) gera N valores de pesos λ e um valor do multiplicador de Lagrange, μ. Os valores dos pesos são substituídos na equação (1), estimando-se, dessa maneira, todos os valores para a posição x_o no espaço amostrado. Desta maneira os mapas de isolinhas representando as produtividades das três culturas nos quatro anos podem ser construídos para exame e interpretação da variabilidade.

A Análise de Correspondência Simples parte de uma matriz de dados representado por uma tabela de contingência. O desenvolvimento do algoritmo e sua geometria fornecem as regras básicas para a interpretação. O algoritmo adaptado de Greenacre (1984) e Barioni Júnior (1995) segue os seguintes passos.

Seja N uma tabela de contingência com I categorias de linhas e J categorias de colunas, com elementos não negativos, de modo que a soma de cada linha ou coluna seja não nula. Seja n_{ij} o número de freqüências observadas pela interseção da i-ésima categoria da variável A com a j-ésima categoria da variável B, logo $N = [n_{ij}]_{I \times J}$ é a matriz de freqüências absolutas. A matriz N pode transformar-se em uma matriz de freqüências relativas expressa por: $P = (1/n)\,N$, onde P denomina-se matriz de Correspondência.

Em função da matriz de correspondência P, define-se o i-ésimo perfil linha como o vetor: $a_i = (p_{i1}/p_{i.}, p_{i2}/p_{i.}, \ldots, p_{ij}/p_{i.})^T$ onde cada vetor a_i $(i = 1, \ldots, I)$ representa uma distribuição muitinomial, condicionada à i-ésima categoria da variável A. Conseqüentemente, a matriz de perfis linha é definida por: $R = D_r^{-1}P$ e a matriz de perfis coluna por: $C = D_c^{-1}P'$. Cada perfil linha ou coluna é afetado por um peso $p_{i.}$ para $i = 1,\ldots,I$ ou $p_{.i}$ para $j = 1,\ldots,J$, proporcional aos respectivos totais de linha ou de coluna dos dados originais. Estes pesos são denominados de massas onde $r_i = p_{i.} = n_{i.}/n$ para $i = 1,\ldots,I$ é a massa da i-ésima linha e $c_j = p_{.i} = n_{.i}/n$ para $j = 1,\ldots,J$ é a massa da j-ésima coluna.

O centróide da linha ou coluna de uma tabela de contingência indica geometricamente a posição média dos perfis linha ou coluna, como se fosse o centro de gravidade ou ponto de equilíbrio da matriz de dados. Como cada perfil linha ou coluna está associado a uma massa, define-se como centróide a média ponderada dos perfis linha ou coluna. Ou seja, o centróide dos perfis linha $Ce_{(r)} = \sum_i r_i a_i = \mathbf{R'r} = \mathbf{c}$ e centróide dos perfis coluna $Ce_{(c)} = \sum_j c_j b_j = \mathbf{C'c} = \mathbf{r}$.

Cada perfil linha a_i define um ponto em R^J, logo, neste espaço acomoda-se uma nuvem de pontos representando os I perfis linha. Como os elementos de cada perfil linha somam 1, implica que a nuvem esta restrita a uma região, no subespaço de dimensão J-1 do espaço das colunas, conhecida como Simplexo. A interpretação do espaço dos perfis coluna é feita de forma análoga. O formato da nuvem ou o aparecimento de pontos isolados pode indicar a existência ou não de associações entre pontos. Essas associações são medidas através das distâncias entre os pontos chamada distância Qui-Quadrado (χ^2) que é uma distância euclidiana ponderada entre os perfis linha e o centróide Ce, sendo definida por: $D^2\{a_i, Ce\} = (R\text{-}Ce'1)^T D_{Ce}^{-1} (R\text{-}Ce'1)$ onde D_{Ce}^{-1} é a matriz de ponderação com diagonal representada pelo inverso das coordenadas do centróide. Se a métrica $RD_{Ce}^{-1/2}$ com suas respectivas massas for aplicada, têm-se a Distância Euclidiana clássica.

De acordo com a natureza dos dados a analisar, a nuvem terá maior ou menor dispersão em relação ao centróide. A dispersão pode ser calculada como a média dos desvios dos vários pontos ao centróide. O quantificador normalmente usado da dispersão dos pontos dessa nuvem no espaço R^J é chamado de Inércia total dos perfis linha.

Na comparação de mapas de produtividade, além de se estar interessado na similaridade dos mapas, é interessante o estudo das similaridades de classes ou áreas dentro dos mapas. Essas diferenças entre classes ou áreas são observáveis através da investigação das associações existentes entre as categorias dentro de cada classe e/ou entre as categorias de diferentes classes. Ela segue o mesmo objetivo da Análise Discriminante Canônica de Fisher, ou seja, o de maximizar a variabilidade entre

Classes, representado na Análise de Correspondência pela inércia entre classes.

O tratamento matemático para a execução da Análise de Correspondência em classes segue o mesmo princípio para a Análise e de Correspondência Múltipla. Enquanto na Análise e de Correspondência Simples a tabela de contingência é usada, na Análise de Correspondência Múltipla usa-se a tabela de Incidência ou tabela de Burt. Nesta, todas as associações entre pares de variáveis são analisadas, bem como a associação entre uma variável e ela mesma. A inclusão das associações de uma variável com ela mesma, o que não acontece na Análise de Correspondência Simples, é que permite a Análise de Correspondência Múltipla de ser eficiente em sua parte computacional. Entretanto, sua presença aumenta a inércia total dos pontos e dificulta sua interpretação geométrica (Greenacre, 1983; Pamplona, 1998).

6. Comentários

Com o uso da técnica estatística da Análise de Correspondência, pretende-se testar a hipótese de que a utilização de mapas de produtividade para delimitar regiões homogêneas no campo, depende da estabilidade espacial e temporal da variável produção, tornando-se desta maneira, preocupante seu uso indiscriminado como ferramenta de informação preditiva.

7. Referências bibliográficas

BARIONI JÚNIOR, W. **Análise de correspondência na identificação dos fatores de risco associados à diarréia e à performance de leitões na fase de lactação**. 1995. 97 f. Tese (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba.

BENZéCRI, J.P. **Correspondence analysis handbook**. New York: Marcel Decker, 1992. 665 p. (Statistics: Textbooks and Monographs, 125).

BLACKMORE, B. S.; LARSCHEID, G. **Strategies for managing variability**. Silsoe: Cranfield University - The Centre for Precision Farming, [1998]. Disponível em: <http://www.silsoe.cranfield.ac.uk/cpf/papers/strategies/strategies.pdf>. Acesso em: 28 jul. 2000.

CAPELLI, N. L. **Agricultura de precisão novas tecnologias para o processo produtivo**. [S.l.: s.n., 1999]. Disponível em: <http://www.cria.org.br/gip/gipaf/capelli.htm>. Acesso em: 28 jul. 2000.

CONGALTON, R. G.; MEAD, R. A. A. Quantitative method to test for consistency and correctness in photointerpretation. **Photogrammetric Engineering and Sensing**, v. 49, p. 69-74, 1983.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 2.ed. New York: John Wiley, 1981. 709 p. (Wiley Series in Probability and Mathematical Statistics).

GREENACRE, M. J. **Correspondence analysis in practice**. London: Academic Press, 1983. 195 p.

GREENACRE, M. J. **Theory and applications of correspondence analysis**. London: Academic Press, 1984. 364 p.

GREENACRE, M. J.; HASTIE, T. The geometric interpretation of correspondence analysis. **Journal of the American Statistical Association**, v. 82, n. 398, p. 437-447, 1987.

PAMPLONA, A. S. **Análise de correspondência para dados com estrutura de grupo**. 1998. 163 f. Tese (Mestrado em) Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas.

ROSENFIELD, G.H.; FITZPATRICK-LINS, K. A coefficient of agreement as a measure of thematic classification accuracy. **Photogrammetric Engineering and Remote Sensing**, v.52, n.2, p.223-227, 1986.

VIEIRA, S.R. Variabilidade espacial de argila, silte e atributos químicos em uma parcela experimental de um latossolo roxo de Campinas (SP). **Bragantia**, Campinas, v.56, n.1, p.1-17, 1997.

VIEIRA, S.R.; HATFIELD, J.L.; NIELSEN, D.R; BIGGAR, J.W. Geostatistical theory and application to variability of some agronomical properties. **Hilgardia**, Berkeley, v.51, n.3, p. 1-75, 1983.

IMPRESSO

Empresa Brasileira de Pesquisa Agropecuária
Centro Nacional de Pesquisa Tecnológica em Informática para a Agricultura
Ministério da Agricultura e do Abastecimento
Rua Dr. André Tosello, s/nº Caixa Postal 6041 - Barão Geraldo
13083-970 - Campinas, SP
Fone (19) 3789-5700 Fax (19) 3289-9594
E-mail: sac@cnptia.embrapa.br
http://www.cnptia.embrapa.br

MINISTÉRIO DA AGRICULTURA E DO ABASTECIMENTO

GOVERNO FEDERAL
Trabalhando em todo o Brasil