APLICAÇÃO DE MÉTODOS GEOESTATÍSTICOS BAYESIANOS NA ANÁLISE DE DADOS COMPOSICIONAIS

Ana Patricia Bastos Peixoto [1]
Tiago Viana Flor de Santana [2]
Paulo Justiniano Ribeiro Junior [3]
Maria Cristina Neves de Oliveira [4]
José Ronaldo de Macedo [5]
Cláudio Lucas Capeche [6]

1 Introdução

A estatística clássica tem por base que as amostras podem ser repetidas indefinidamente ao acaso e que o resultado de uma amostra não influencia o resultado de outra, garantindo assim, a independência dos dados. Num estudo geoestatístico, estas variáveis têm em comum uma dupla característica, são aleatórias já que os valores numéricos observados podem variar consideravelmente de um ponto a outro no espaço; são espaciais porque apesar de muito variar dentro do espaço, os valores numéricos observados não são inteiramente independentes (Guerra, 1988). Variáveis com tal estrutura são chamadas de variáveis regionalizadas (VR) ou georreferenciadas, uma definição de Matheron (1963).

A adoção de métodos bayesianos, que incorporam a incerteza associada aos parâmetros nos procedimentos de predição permite, uma melhor definição e melhor caracterização da incerteza sobre zonas viáveis de manejo em experimentos agronômicos que envolvem a produção de mapas temáticos, sobretudo quando se dispõe de amostra pequena da variável de interesse. Permite estimar a média do processo gaussiano subjacente que reflete a produtividade total quando comparado a uma estimação por krigagem convencional.

A lógica da inferência bayesiana é realizar inferência associando incerteza aos parâmetros envolvidos no modelo, tratando-os também como variável aleatória. A proposta para sua aplicação com a geoestatstica é a de combinar estimação e predição a partir de um conjunto de observações associadas a um processo, em um alvo de predição. O objetivo desse estudo é identificar a dependência espacial nos dados utilizados, com o uso das ferramentas computacionais do software R, através da abordagem Bayesiana em dados composicionais.

2 Material e métodos

Os dados experimentais, que fazem parte deste trabalho, foram coletados do levantamento detalhado de solos da Estação Experimental de Campos, Fazenda Angra (Capeche et al., 1997), por pesquisadores da PESAGRO e Embrapa Solos. No estudo pedológico foram avaliadas as

[1]Doutoranda em Estatística e Experimentação Agronômica - LCE, ESALQ/USP: *apbpeixoto@usp.br*
[2]Doutorando em Estatística - IMECC UNICAMP: *tiagodesantana@yahoo.com.br*
[3]Professor Dr. do Laboratório de Estatística e Geoinformação - LEG UFPR: *paulojus@ufpr.br*
[4]D.Sc. Em Estatística e Experimentação Agronômica - EMBRAPA: *mmcno@cnpso.embrapa.br*
[5]Ph.D em Manejo e Conservação de Solos - EMBRAPA: *jrmacedo@cnps.embrapa.br*
[6]M.Sc. em Manejo e Conservação de Solos - EMBRAPA: *capeche@cnps.embrapa.br*

características morfológicas, físicas e químicas dos solos, e apresentadas também, informações referentes à distribuição geográfica. Para o estudo geoestatístico foi considerada a variável agronômica teor de Areia (%), Argila (%) e Silte (%), na camada de 0-20cm. A área compreendida pelo levantamento situa-se ao Norte do estado do Rio de Janeiro, à margem esquerda do Rio Paraíba do Sul, no município de Campos dos Goytacazes, entre os paralelos $21^{o}44'47''$ e $41^{o}18'24''$ WGr.

Em geral, o modelo geoestatístico é especificado por meio de dois sub-modelos: um sub-modelo para um processo espacial não observado $\left\{S(x) : x \in^2\right\}$, chamado de sinal e um sub-modelo para os dados $Y = \{Y_1, ..., Y_n\}$ condicionado ao $S(.)$. Usando θ como uma notação genérica para todos os parâmetros desconhecidos, uma notação formal para o modelo especificado é:

$$[Y, S|\theta] = [S|\theta]\,[Y|S, \theta]$$

em que, S denota o conjunto do processo de sinal $\left\{S(x) : x \in^2\right\}$.

A distribuição preditiva clássica S, é condicional a distribuição $[Z|Y, \theta]$ que, a princípio, pode ser obtida a partir da especificação de um modelo de aplicação do teorema de Bayes. Para qualquer objeto de predição G, como G é uma função deterministica de S a distribuição preditiva, pode ser ou não analiticamente calculada (DIGGLE & RIBEIRO Jr., 2007). Em qualquer caso, para gerar uma compreensão a partir da distribuição preditiva $[G|Y, \theta]$ precisa-se apenas gerar uma realização da distribuição preditiva $[Z|Y, \theta]$ e aplicar um calculo determinístico para converter S para G.

A informação vinda dos dados através da verossimilhança com a incorporação de conhecimentos a priori, ou seja, com as informações obtidas antes da coleta dos dados, obtém-se uma distribuição denominada distribuição a posteriori. A partir desta distribuição a posteriori é possível determinar as quantidades necessárias para o processo inferencial, tais como medidas de posição e dispersão.

Considere uma quantidade de interesse desconhecida θ. A informação de que dispomos sobre θ, resumida probabilisticamente através de $p(\theta)$ chamada distribuição a priori, pode ser aumentada observando-se uma quantidade aleatória X relacionada com θ .

A distribuição amostral $p(x|\theta)$ define esta relação. A idéia de que após observar $X = x$ a quantidade de informação sobre θ aumenta é bastante intuitiva e o teorema de Bayes é a regra de atualização utilizada para quantificar este aumento de informação,

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x|\theta)}{p(x)},$$

onde $p(\theta|x)$ é chamada densidade a posteriori, $p(x) = \int p(x|\theta)p(\theta)d\theta$ recebe o nome de densidade preditiva.

Para um valor fixo de x, temos que $L(\theta, x) = p(x|\theta)$ é a verossimilhança de cada um dos possíveis valores de θ . Como $p(\theta|x)$ é uma densidade para θ a observação x é apenas uma constante bem como $p(x)$; temos que a forma usual do teorema de Bayes é dada por,

$$p(\theta|D) \propto L(\theta)p(\theta),$$

onde D denota o conjunto de observações de x.

A comparação de modelos sob a teoria Bayesiana pode ser feita a partir de medidas de adequabilidade, como o Bayesian Information Criterion (BIC), o qual é uma aproximação do fator de Bayes. Carlin e Louis (2000) introduziram uma modificação do critério BIC dado por

$$BIC_i = -2E(\ln L(\theta_i|x, M_i) + p_i \ln(n)),$$

onde i indexa o modelo; n é o número da amostra e p_i é o número de parâmetros sob o modelo M_i. Menores valores do BIC indicam o melhor ajuste do modelo.

3 Resultados e discussão

Para inferência bayesiana sobre os parâmetros do modelo foi considerado um modelo isotrópico, sem tendência direcional ou efeito sistemático e função de correlaçõo de Matèrn com parâmetro de diferenciabilidade $k = 0.5$ e utilizamos a variável silte para apresentação dos resultados.

Tabela 1: Resumo dos parâmetros do modelo a posteriori e intervalo de credibilidade

Variável	Mínimo	1 Quartil	Mediana	Média	3 Quartil	Máximo	LI	LS
β_0	-33,2	622,7	746,2	750,7	877,1	1641,0	354,6	1155,5
β_1	-154,4	-78,7	-61,4	-61,6	-45,9	24,8	-109,9	-9,1
σ^2	29820,0	59170	77460	85230	106900	193000	40308,3	163477,5
ϕ	125,0	250,0	333,3	380,0	485,3	10000,0	0,42	0,13
τ^2_{rel}	0,042	0,042	0,042	0,056	0,083	0,291	0,042	0,125

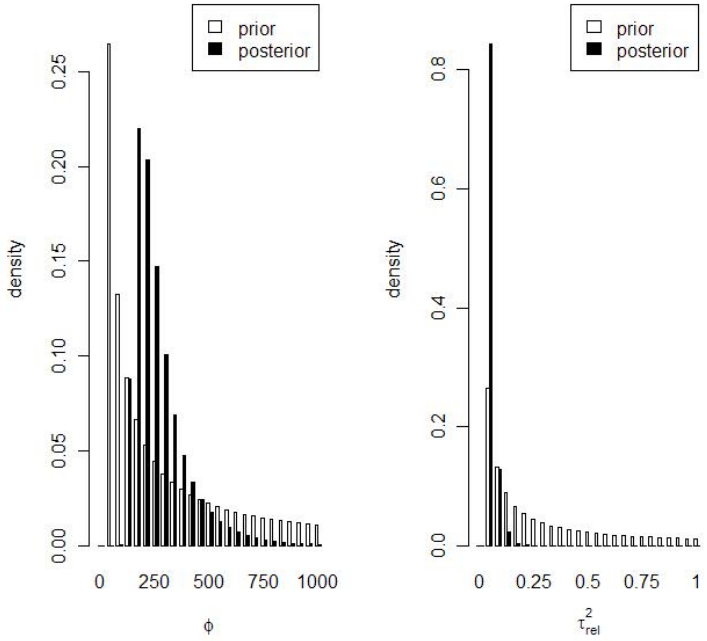

Figura 1: Distribuição priori e posteriori dos parâmetros do modelo ϕ no painel a esquerda e τ^2 no painel a direita.

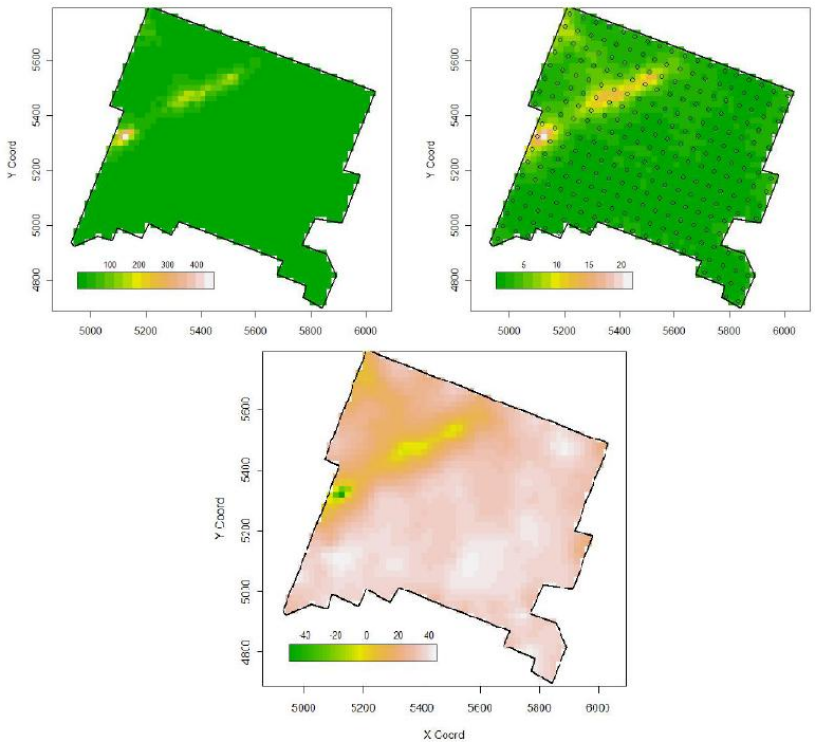

Figura 2: Krigagem Bayesiana a posteriori para o teor de silte.

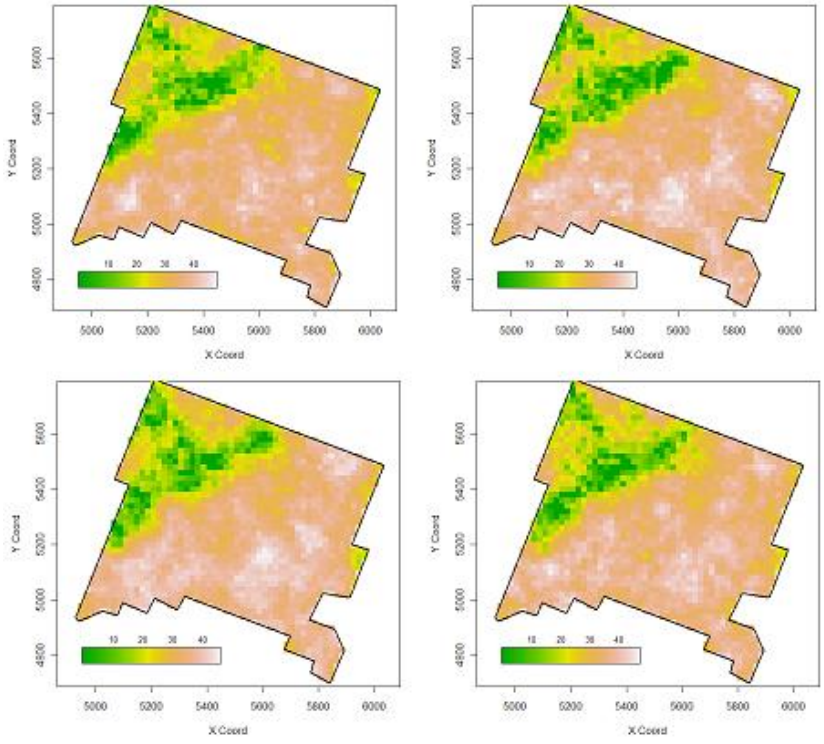

Figura 3: Gráfico das distribuições preditivas.

As distribuições para cada um dos parâmetros estimados apresentam diversos comportamentos. Tanto β_0 quanto β_1 apresentaram simetria nos sua distribuição, cujo comportamento é de uma distribuição normal, os demais paramétros apresentaram assimetria na sua estrutura. β_0

quanto $\beta_1$são parâmetro do efeito sistemático do modelo, σ^2 e ϕ são parâmetros da função de correlação, todos obtidos pela média das aproximações numéricas pelo método bayesiano. Neste processo, τ^2 é o parâmetro do erro, fixado em zero.

A Figura 2 mostra os valores preditos do teor de silte (%). Esses resultados indicaram a manutencção no erro relativo comparado com o resultado obtido pela predição por krigagem convencional. É posível observar a concentração do silte nas regiões mais claras. Nessas figuras nota-se que os mapas utilizaram informação de um padrão de variabilidade espacial.

4 Conclusão

A presença de indícies de teor de silte é mais fortes em torno de 40% nas áreas mais claras dos gráficos que foram observados, indicando que a prevalência da fração total dessa variável tem maior intesidade e interferência na produtividade do solo.

Referências

[1] CAPECHE, C.L.; MACEDO, J.R.; MANZATTO, H.R.H.; SILVA, E.F. Caracterização pedológica da fazenda Angra - PESAGRO/RIO - Estação experimental de Campos (RJ). (compact disc). In: CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO. 26., Informação, globalização, uso do solo; Rio de Janeiro, 1997. trabalhos. Rio de Janeiro: Embrapa/SBCS, 1997.

[2] CARLIN, B.P., Louis, T. Bayes and Empirical Bayes methods for data analysis, 2nded, London: Chapman and Hall, 2000.

[3] DIGGLE, P.;TAWN, J. A.; MOYEED, R. A. Model-based geostatistics. Applied Statistics, V.47, n.3, p.299-350, 1998. (citeseer.ist.psu.edu/diggle98modelbased.html)

[4] GUERRA, P.A.G. Geoestatística operacional. Brasília: Ministério das Minas e Energia, 1988. 145p.

[5] MATHERON, G. Principles of Geostatistics. Economic Geology, v.58, p.1246-1266. 1963.

[6] R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. Vienna, Austria, 2008. (http://www.R-project.org)