GEOESTATÍSTICA NO R: UM ESTUDO COMPARATIVO ENTRE DOIS SCRIPTS

Batista, A. P. B.[1], Santos, G. R.[2], Mello, J. M.[3], Silva, K. E.[4], Kaleita, A.[5]

[1]Doutorando em Engenharia Florestal, UFLA/LEMAF/Departamento de Ciência Florestal/ CEP: 37200-000, Lavras, MG, anderson_pedro22@yahoo.com.br;
[2]Professor Doutor, UFV/ Departamento de Estatística / CEP: 36570-000,Viçosa-MG, gerson.santos@ufv.br;
[3]Professor Doutor, UFLA/LEMAF/Departamento de Ciência Florestal/ CEP: 37200-000, Lavras, MG, josemarcio@dcf.ufla.br;
[4]Pesquisadora da Embrapa Amazônia Ocidental / CEP: 69010-970 Manaus - AM, katia.emidio@embrapa.br
[5]Associate Professor, Iowa State University/Agricultural and Biosystems Engineering / Zip: 50010, Ames, Iowa, EUA, kaleita@iastate.edu.

Resumo - A geoestatística se destaca como uma metodologia da estatística espacial, que utiliza inclusive a posição geográfica dos dados amostrais para caracterizar uma ou mais variáveis em estudo, além de interpolar de maneira ótima, estatisticamente, em locais não amostrados. Com a popularização dessa metodologia, várias formas diferentes de análises têm surgido, inclusive utilizando o mesmo programa. Dentro desse contexto, o objetivo deste trabalho foi verificar se existem diferenças consideráveis nos resultados de uma análise geoestatística entre duas formas distintas de análise, através de dois *scripts* diferentes, utilizando o pacote geoR, para o mesmo conjunto de dados, no programa R. Foram avaliados 35 indicadores, que foram divididos em tabelas, de acordo com a etapa do processo analítico. Em termos gerais, o *script 1* apresentou pequena superioridade em relação ao *script 2*.

Palavras-chave: Geoestatística; pacote geoR; Programa R.

Geostatistics in R: a comparative study between two scripts

Abstract - Geostatistics stands as a methodology of Spatial Statistics which also uses the geographical position of the sample data to characterize one or more variables under study, and interpolate optimally, statistically, in unsampled locations. With the popularity of this method, several different forms of analysis have appeared, including using the same software. In this context, the aim of this study was to determine whether there are considerable differences in the results of a geostatistical analysis of two different forms of analysis, through two different scripts, using the geoR package for the same data set, in the program R. 35 indicators were chosen to this evaluation, divided into tables, according to the analytical process step. Overall the script 1 showed a slight superiority of the script 2.

Key words: Geostatistics, geoR package, program R.

Introdução

Entre as metodologias da estatística espacial, a geoestatística destaca-se por utilizar toda a informação disponível, inclusive a posição geográfica dos dados amostrais, para caracterizar a variável em estudo, levando em consideração a estrutura de dependência espacial dos dados, e interpolá-la através de um BLUP – *Best Linear Unbiased Predictor* (melhor preditor linear não viesado) de variância mínima (VIEIRA, 2000; SANTOS et al., 2011; YAMAMOTO e LANDIM, 2013).

O programa R é um dos melhores para análise estatística existentes na atualidade. Todo o conteúdo da janela do console pode ser salvo, marcado e eliminado utilizando os recursos do Windows e da barra de ferramentas. Uma maneira amplamente disseminada de utilizar o R é digitar os programas e macros que serão executados em uma janela denominada *script* (FERREIRA, 2013). Como principais vantagens do R é o fato de ser um programa gratuito, e os pacotes (*libraries/packages*) são escritos e desenvolvidos por pesquisadores das mais diferentes áreas do conhecimento. Dentre os pacotes, destaca-se o geoR utilizado para análises geoestatísticas (RIBEIRO JÚNIOR; DIGGLE, 2001).

Atualmente, estão disponíveis alguns programas na área de geoprocessamento que realizam análises geoestatísticas, porém com procedimento de ajuste automático, que podem comprometer o resultado final. Segundo Ferreira et al. (2013) o procedimento de ajuste não deve ser direto e/ou automático, mas sim interativo, pois neste processo o analista deve interagir o todo tempo com os resultados parciais obtidos.

Apesar de se tratar de uma metodologia já consagrada no meio científico e profissional, muitos trabalhos comparativos têm surgido nessa área, em busca da melhoria continua do processo analítico e representativo das regiões de estudo, independente da variável. Sendo assim, o objetivo deste trabalho foi verificar se existem diferenças

consideráveis nos resultados específicos e gerais de uma análise geoestatística entre duas formas distintas de análise, através de dois *scripts* diferentes do pacote geoR, para o mesmo conjunto de dados, no programa R.

Material e Métodos

Os dados utilizados para a realização desse estudo comparativo são da variável teor de cobre, coletados no campo experimental da Embrapa Amazônia Ocidental, localizado no município de Rio Preto da Eva – AM, com coordenadas 59º59'42.6" W e 2º32'49.7" S, conforme Figura 1.

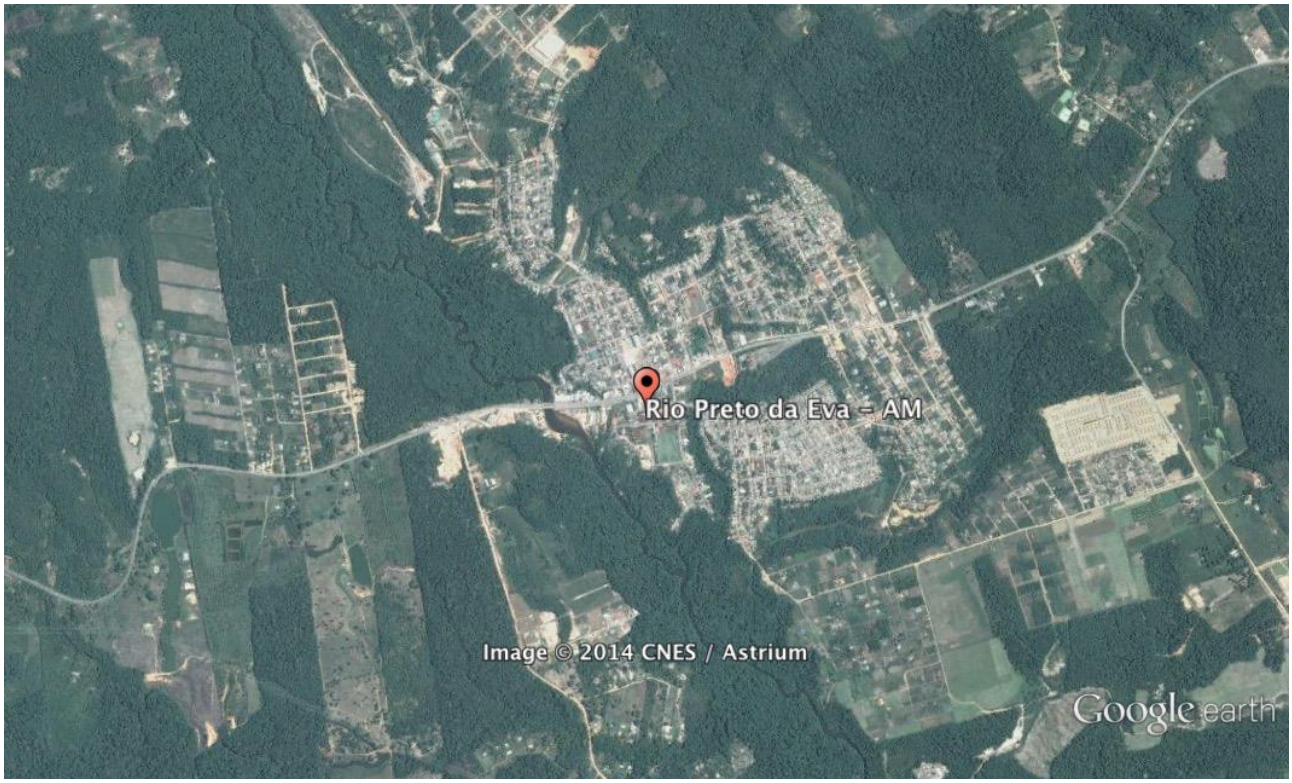


Figura 1. Imagem de satélite de Rio Preto da Eva – AM, localização da área de amostragem.
Fonte: Google Earth

Foram coletados 199 amostras de solo, cujo grid amostral foi de espaçamento irregular, com distâncias entre pontos variando de 3 a 425 metros. As amostras foram obtidas com trado a uma profundidade de 0 a 20 cm. Após a coleta e cuidados iniciais, as amostras foram enviadas ao Laboratório de Solos e Plantas da Embrapa Amazônia Ocidental para análise química e física.

Entre todas as variáveis envolvidas na amostragem, o teor de cobre foi escolhido devido às características apresentadas, como é o caso de normalidade dos dados, ausência de *outliers*, isotropia, ausência de tendência e atendimento aos pressupostos da geoestatística, conforme recomendam Vieira (2000), Santos et al. (2011) e Yamamoto e Landim (2013).

Os critérios de comparação foram adotados conforme a presença do critério em um dos dois *scripts*. Alguns critérios elencados foram apontados como presentes ou ausentes, sem comentários exclusivos, como é o caso do critério "Teste de independência", mas outros critérios considerados de maior importância para o estudo foram quantificados e receberam comentários exclusivos, como é o caso do critério "Média dos resíduos padronizados".

As análises foram realizadas com auxílio do pacote geoR (RIBEIRO JÚNIOR; DIGLLE, 2001) do software R (R DEVELOPMENT CORE TEAM, 2014).

Resultados e Discussão

Foram avaliados 35 indicadores existentes pelo menos em um dos *scripts* avaliados. Os indicadores foram divididos em Tabelas 1, 2, 3 e 4, para facilitar a compreensão dos resultados do estudo. Em termos gerais, o *script 1* apresentou pequena superioridade em relação ao *script 2*, mas nada muito significativo.

Tabela 1. Indicadores iniciais de uma análise geoestatística realizada no pacote geoR do programa R considerando dois *scripts* distintos.

Indicadores	***Script 1***	***Script 2***
Leitura multivariada dos dados	Sim	Não
Análise exploratória clássica	Sim	Sim
Teste de independência	Sim	Não
Teste de normalidade (SW)	Sim	Sim
Análise espacial da malha-visual	Sim	Sim
Análise de tendência - visual	Sim	Sim
Detecção de outlier	Sim	Sim

Conforme Tabela 1, não se constata no *script 2* apenas a leitura multivariada dos dados, que contribui na agilidade da análise de dados, e um teste de independência dos dados, importante na constatação de um dos mais importantes pressupostos teóricos para a aplicação da estatística. Contudo, nessa fase da análise, tal teste não é essencial, pois através do comportamento do variograma, tem-se um parecer preciso do comportamento da variável em estudo. Dessa forma, ambos os *scripts* apresentam características similares.

A fase seguinte da análise compreende toda a análise variográfica, ou seja, estimação do variograma, estudo de anisotropia e modelagem do variograma empírico, conforme apresentado na Tabela 2.

Tabela 2. Indicadores da fase de análise de variograma de uma análise geoestatística realizada no pacote geoR do programa R considerando dois *scripts* distintos.

Indicadores	***Script 1***	***Script 2***
Estimação do "melhor" variog	Sim	Sim
Determinação do nº de pares	Não	Sim
Estimação dos variogramas direcionais	Sim (Isotropia)	Sim
Sistema eyefit de ajuste do variograma (A sentimento)	Sim	Sim
Métodos de mínimos quadrados de ajuste	Sim	Sim
Métodos de máxima verossimilhança de ajuste	Sim	Sim (apenas ML)

Conforme Tabela 2, as características dos dois *scripts* apresentam características semelhantes nessa fase, exceto a determinação do número de pares de pontos da estimação do variograma empírico. Para este conjunto de dados, esta característica não se mostrou inadequada, mas certamente em outros conjuntos é importante estabelecer um número mínimo de pares de pontos, uma vez que o variograma é a etapa mais importante de uma análise geoestatística (VIEIRA, 2000; SANTOS et al., 2011; YAMAMOTO e LANDIM, 2013).

Para a fase de escolha de modelos do variograma, é usual adotar a autovalidação *leave-one-out*, adotando os critérios estabelecidos por Vieira (2000), conforme apresentado na Tabela 3.

Através dos resultados da autovalidação, é possível determinar o modelo mais adequado para o variograma empírico estimado, segundo Yamamoto e Landim (2013). Sendo assim, pode-se perceber uma pequena vantagem numérica do primeiro *script*, uma vez, que em média os resíduos deveriam ser nulos, e o desvio-padrão dos resíduos padronizados deveria ser unitário. Além disso, Vieira (2000) recomenda a realização de uma regressão linear simples entre os valores observados e os valores preditos na autovalidação, e subsequente análise dos parâmetros. Teoricamente, o parâmetro beta 0 deveria ser nulo, e o parâmetro beta 1 ser unitário. Dessa forma, pode-se perceber que o ajuste do modelo Gaussiano ao variograma empírico foi razoável. Apresenta-se ainda outras informações importantes, como é o caso do alcance, efeito pepita, contribuição e estimação da variância dos dados através do patamar (contribuição + efeito pepita).

Santos et al. (2011) mostram em média, e utilizando a variância de krigagem de parâmetro de decisão, que a krigagem indicativa é mais precisa do que outras krigagens lineares. Dessa forma, pode-se perceber que, através do *script 1*, que provavelmente devido esse motivo a variância de krigagem foi de 3 a 4 vezes menor. Além disso, o *script 1* apresenta ainda um mecanismo de delimitação da área de estudo que utiliza o alcance estimado na modelagem do variograma como gerador do buffer de extrapolação.

Tabela 3. Indicadores da fase de autovalidação de uma análise geoestatística realizada no pacote geoR do programa R considerando dois *scripts* distintos.

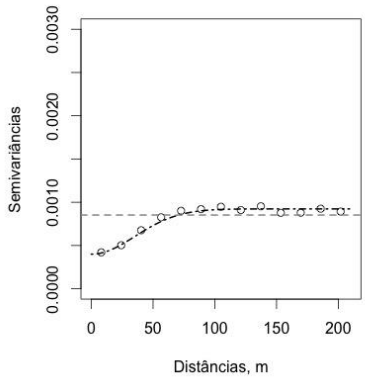
Indicadores	***Script 1***	***Script 2***
Autovalidação	Sim	Sim
Média dos resíduos padronizados	- 0.0001869	0.0003235
Desvio-padrão dos resíduos padronizados	1.0247	1.0537
RLS – Regressão Linear Simples	Sim	Não
RLS – Beta 0 estimado	0.0038	Não
RLS – Beta 1 estimado	0.9617	Não
"Melhor" método de ajuste	OLS	OLS
Modelo	Gaussiano	Gaussiano
Alcance prático	86.54 m	103.84 m
Efeito pepita	0.0004	0.0004 (fixo)
Contribuição	0.000523	0.000501
Variância dos dados	0.00085	0.000852

Finalmente, apresenta-se na Tabela 4, a interpolação da geoestatística, krigagem, com os respectivos indicadores de qualidade.

Tabela 4. Indicadores da fase de interpolação, krigagem, de uma análise geoestatística realizada no pacote geoR do programa R considerando dois *scripts* distintos

Indicadores	***Script 1***	***Script 2***
Krigagem	Simples	Ordinária
Delimitação da área krigada conforme alcance	Sim	Não
Média da variância de krigagem	$1.3x10^{-4}$	$4.8x10^{-4}$
Variância da var. de krig.	$4.1x10^{-9}$	$2.8x10^{-9}$
Desvio-padrão da var. de krig.	$6.4x10^{-5}$	$4.8x10^{-5}$
Média dos valores preditos	0.1001	0.1004564
Variância dos valores preditos	0.0005115	0.0003052
GDE	56.67	55.55
AIC	Não se aplica em OLS	454.16
Número de pares mínimos (pairs.min)	Automático (108)	Sim (15)

Apresenta-se na Figura 2 o variograma empírico, já modelado via o método de mínimos quadrados ordinários, usando o modelo Gaussiano, e a krigagem da variável teor de cobre (imagens de um dos *scripts* analisados).



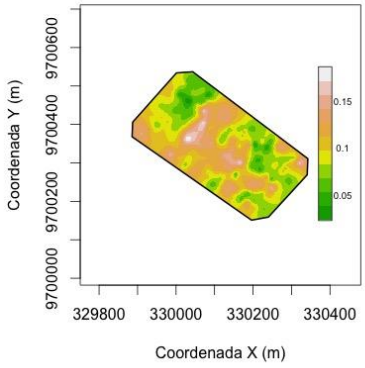


Figura 2 – Apresentação do variograma empírico (pontos do gráfico à esquerda), e respectiva modelagem (pontilhado mais escuro) e mapa de interpolação via krigagem (imagem à direita) do teor de cobre.

Conclusão

Ao comparar-se duas formas distintas de uma análise geoestatística, utilizando o pacote geoR e o programa R, conclui-se que, apesar das diferenças de formulação dos roteiros, denominados no trabalho como *scripts*, os resultados foram muito similares.

No geral, o *script 1* apresenta uma pequena vantagem em relação ao *script 2*, devido uma preocupação maior em se verificar o atendimento dos pressupostos teóricos que o uso da metodologia exige.

A maior diferença percebida foi na verificação da qualidade da análise realizada, através da variância de krigagem, em que o *script 1* foi de 3 a 4 vezes mais preciso do que o *script 2*, resultado justificado provavelmente pela escolha do interpolador.

Assim, este trabalho mostra a robustez de uma análise geoestatística feita no programa R, independente das diferenças apresentadas pelos dois roteiros escolhidos para este estudo.

Referências

FERREIRA, I. O.; SANTOS, G. R. RODRIGUES, D. D. Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas. **Revista Brasileira de Cartografia**, Rio de Janeiro, v. 65, n.5, p. 831-842, 2013.

FERREIRA, D. F. **Estatística computacional utilizando R**. Universidade Federal de Lavras, Departamento de Ciências Exatas, 2013, 125 p.

YAMAMOTO, J. K., LANDIM, P. M. B. **Geoestatística: conceitos e aplicações**. 1ed. São Paulo: Oficina de Textos, 2013, 215p.

RIBEIRO JÚNIOR, P. J.; DIGGLE, P. J. geoR: a package for geostatistical analysis. **R-NEWS**, Pelotas, v. 1, n. 2, p. 15-18, 2001.

R DEVELOPMENT CORE TEAM. R: **A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna. Disponível em:< http://www.r-project.org >. Acesso em: 08 jul. de 2014.

SANTOS, G. R.; OLIVEIRA, M. S.; LOUZADA, J. R.; SANTOS, A. M. R. T. KS versus KU: qual o preditor mais preciso? In: SIMPÓSIO DE GEOESTATÍSTICA EM CIÊNCIAS AGRÁRIAS, 2., 2011, Botucatu, **Anais...** Faculdade de Ciências Agronômicas-UNESP.

VIEIRA, S. R. **Geoestatística em estudos de variabilidade espacial do solo**. In: NOVAIS, R.F.; ALVAREZ, V., V.H. & SCHAEFER, G.R., eds. Tópicos em ciência do solo. Viçosa, MG, Sociedade Brasileira de Ciência do Solo, 2000. v. 1. p.1-54.