UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA "LUIZ DE QUEIROZ"
DEPARTAMENTO DE CIÊNCIAS EXATAS
LCE 5801-5 Geoestatística – 2011/02
RESENHA 2

Nome: Elisângela Aparecida de Oliveira **Nº USP**: 7458358

Prof.: Paulo Justiniano Ribeiro Junior

GEOSTATISTICAL MODEL AVERAGING BASED ON CONDITIONAL INFORMATION CRITERIA

Chun-Shu Chen and Hsin-Cheng Huang

Springer, Environ Ecol Stat, April 2011.

Segundo os autores em muitos problemas de geoestatística, a variável de interesse é frequentemente observada juntamente com outras variáveis em algumas localidades. Tipicamente, um modelo de regressão geoestatístico é aplicado ao tratar a variável de interesse como resposta e algumas outras variáveis como variáveis explicativas. Como selecionar um subconjunto adequado de variáveis explicativas é crucial, porque a precisão da seleção afeta diretamente a estimação e predição. Seleção de modelos e modelos médios têm sido bem estudados para a regressão linear.

Neste artigo, os autores tiveram três **objetivos** e se concentraram na previsão espacial: Primeiro, introduziram uma classe de critérios de informação condicional indexada por um parâmetro de penalidade, o que acreditaram ser o mais adequado para seleção de modelos geoestatísticos como o AIC condicional (CAIC). Em segundo lugar, como os preditores espaciais obtidos a partir de um critério de informação condicional fixo são instáveis, propuseram estabilizar o preditor espacial pelo modelo de média local, utilizando perturbações nos dados, resultando em um preditor estabilizado que é diferenciável em relação às variáveis respostas. E em terceiro lugar, utilizando a propriedade diferenciável do preditor estabilizado, propuseram aplicar a estimativa de risco imparcial de Stein para selecionar entre um conjunto de critérios de informação condicional, levando a uma penalidade dos dados dependentes com uma característica adaptativa de tal forma que ele selecionasse com uma grande penalidade, e, portanto, um modelo pequeno (baseado em um pequeno número de variáveis explicativas), quando o modelo subjacente verdadeiro fosse pequeno, e vice-versa.

Alguns experimentos numéricos mostram a superioridade do método do modelo médio proposto sobre alguns métodos comumente utilizados nas de seleções de variáveis. Além disso, o método proposto é aplicado a um conjunto de dados de mercúrio dos lagos em Maine.

Os autores consideraram um processo espacial {S (s): s ∈ D} definido sobre uma região $D \subseteq IR^d$. O processo espacial pode ser decomposto em:

$$S(s) = \mu(s) + \eta(s); s \in D, \qquad (1)$$

em que: μ(·) é um processo determinístico da média, η(·) é a média-zero. Supondo que observamos variáveis resposta Z_i e um vetor p-dimensional das variáveis explicativas, $(x_1 (s_i), ...,x_p (s_i))'$, associado com Z_i no local $s_i \in D$; i = 1,. . . , *n*. O processo médio μ(s) é geralmente modelado como $\beta_0 + \sum_{j=1}^{p} \beta_j x_j(s)$, em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$ são parâmetros de regressão. Portanto, o modelo de regressão geoestatística pode ser escrito como:

$$Z_i = S(s_i) + \varepsilon(s_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j(s_i) + \eta(s_i)\,; i = 1, \cdots, n, \qquad (2)$$

em que $\varepsilon(s_1),. . . , \varepsilon(s_n) \sim N(0, \sigma^2_\varepsilon)$ são variáveis de ruído branco representando os erros de medição, e são independentes do processo espacial η(·). Aqui, o processo espacial dependente η(·) é geralmente considerado como estacionário com alguma classe de função de covariância paramétrica. Uma utilizada comumente é o modelo Matérn isotrópico.

Neste trabalho, os autores consideraram a seleção entre *p* variáveis explicativas. Cada modelo candidato corresponde a um subconjunto de *p* variáveis e é indexado por $\gamma \in \Gamma$, em que, $\Gamma \subset 2^{\{1,\ldots,p\}}$ é a classe de todos os modelos candidatos.

O objetivo é selecionar $\gamma \in \Gamma$ que minimize $E||\hat{S}_\gamma(\hat{\theta}_\gamma) - S||^2$. Alguns modelos de critérios de seleção comumente usados , tais como os critérios de AIC, BIC e AIC corrigido (AICC), são casos especiais do critério de informação generalizada. No entanto, estes critérios não visam minimizar $E||\hat{S}_\gamma(\hat{\theta}_\gamma) - S||^2$, e, portanto, podem não ser ideais para fins de previsão espacial. Com o objetivo de predição espacial em mente, foi introduzida uma classe de critérios de informação condicional, condicionado em η, o critério CAIC.

Em geral, o CAIC tende a selecionar um modelo super complexo, particularmente quando o modelo subjacente verdadeiro é pequeno. Um remédio natural é considerar a aplicação de uma penalidade maior em (6), levando ao seguinte critério de informação condicional generalizado ($CGIC_\lambda$) indexados por um parâmetro penalizado λ > 0. Mas na prática, o modelo subjacente verdadeiro é desconhecido, e, portanto, um critério $CGIC_\lambda$ pode funcionar bem apenas em algumas situações, o que não é adaptável.

Os autores propuseram uma característica adaptativa ao $CGIC_\lambda$ que se ajusta automaticamente ao selecionar um modelo apropriado $\hat{\gamma}(\hat{\lambda})$, em torno do qual um modelo preditor de média $E\left(\hat{S}^*_{\hat{\gamma}^*(\lambda)}\left(\boldsymbol{s}, \widehat{\boldsymbol{\theta}}^*_{\hat{\gamma}^*(\lambda)}\right)\middle| Z\right)$ de **S**(s) é obtido, independentemente de o modelo subjacente verdadeiro ser pequeno ou grande, e chamaram este método de "*geostatistical model averaging method* – GMA".

Nos estudos de simulação os autores concluíram que o método proposto GMA teve um desempenho melhor do que CAIC e CBIC para todos os casos.

Aplicação

Os autores aplicaram o método proposto GMA ao conjunto de dados de mercúrio previamente analisados por Hoeting e Olsen (1998) usando regressão linear múltipla para avaliar os níveis de mercúrio dos peixes em lagos de Maine. É sabido que o mercúrio é um metal tóxico, o que pode danificar o sistema nervoso humano se o nível de mercúrio no corpo humano ficar acima do limite de segurança. Por exemplo, o governo do estado de Maine sugere que o nível de mercúrio em partes por milhão (ppm) deve ser inferior a 0,43. Para avaliar se os peixes em lagos de Maine são seguros para comer, é importante estimar os níveis de mercúrio nos peixes especialmente em lagos onde nenhuma observação é tomada e identificar as variáveis explicativas importantes responsáveis por níveis elevados de mercúrio.

O conjunto de dados consiste em níveis de mercúrio (em ppm) como a variável resposta e 10 variáveis explicativas amostrados pela Agência de Proteção Ambiental dos EUA em 110 lagos de Maine. Considerarmos todas as combinações possíveis das variáveis explicativas tem-se 2^{10} modelos candidatos a serem selecionados.

Comparando os dois CBIC e CAIC, tem-se que o CBIC seleciona um modelo menor, com apenas uma variável. Em contraste, CAIC seleciona um modelo mais complexo tendo duas variáveis adicionais. O método GMA seleciona o modelo correspondente ao CBIC. Entre os 110 lagos, 75 deles, particularmente no sudeste de Maine, têm níveis de mercúrio acima do limite de segurança (ou seja, 0,43 ppm).

O artigo é bem interessante e é uma alternativa para a seleção de variáveis na regressão geoestatística, uma vez que teve bons desempenhos para a previsão espacial em alguns estudos de simulação.

Como os próprios autores levantaram a questão, uma Justificativa teórica seria de interesse, mas disseram que seria muito difícil, particularmente no âmbito do quadro de domínio fixo assintótico e, portanto, estava além do propósito do artigo, o que deixa uma proposta em aberto para futuros trabalhos nesta área.