COMPARAÇÃO ENTRE O MODELO GEOESTATÍSTICO E MODELO ADITIVO GENERALIZADO GAUSSIANOS PARA A RECONSTITUIÇÃO DE SUPERFÍCIES CONTÍNUAS.

Wagner Hugo BONAT[1]

■ RESUMO: O presente artigo tem o objetivo de comparar o modelo geoestatístico, com o modelo aditivo generalizado, para a reconstituição de uma superfície contínua, através de simulações. O processo de simulação, consistiu em gerar uma malha de 2500 pontos, desta malha retirar amostras de tamanho $n = 50$,$n = 150$ e $n = 250$ e ajustar os dois modelos. Com os modelos ajustados, reconstituir a malha e avaliar as algumas medidas de desempenho preditivo como, erro quadrático médio esperado, coeficiente de correlação de Pearson esperado e nível de cobertura esperado. Os resultados mostram que o modelo geoestatístico tem melhor desempenho em todas as medidas consideradas. Destaca-se o fraco desempenho do modelo aditivo generalizado, quando avaliado pelo intervalo de cobertura, e a má estimação dos parâmetros de variância do modelo geoestatístico, na presença de alta proporção de ruído no sinal, quando a amostra é pequena.

■ PALAVRAS-CHAVE: Modelo geoestatístico; Modelo aditivo generalizado; simulação.

1 Introdução

Em diversas situações práticas, se deseja recuperar uma superfície originalmente contínua, através de algumas amostras obtidas em um conjunto discreto de localizações, dentro de uma área de estudo. Tais situações, ocorrem naturalmente por exemplo na geologia, onde se deseja estimar a extensão de um depósito mineral em uma região a partir de amostras. Na agronomia, o objetivo pode ser analisar uma região para fins de zoneamento agrícola, na entomologia estimar a ocorrância de determinado mosquito a partir de amostras coletadas em armadilhas, dentro de um município. Esses e outros inúmeros exemplos, existem na realidade cotidiana de muitos pesquisadores.

[1]LEG - Laboratório de Estatística e Geoinformação, Universidade Federal do Paraná, Caixa Postal 19.081 , CEP: 81531-990, Curitiba, Paraná, Brasil, E-mail: *wbonat@gmail.com*

Todos estes problemas tem em comum, o fato de que as amostras tem uma localização no espaço, e o fenômeno em estudo varia continuamente em uma determinada área ou região de estudo. A presença de um componente espacial na coleta dos dados, torna as análises estatísticas convencionais inadequadas, principalmente devido a suposição de amostras independentes. A ênfase da análise espacial é mensurar propriedades e relacionamentos, levando em conta a localização do fenômeno em estudo de forma explícita. Ou seja, a idéia geral é incorporar o espaço na análise que se deseja fazer.

O termo geoestatística, refere-se a modelos e métodos para dados seguindo as seguintes características. Primeiro, os valores $Y_i : i = 1, \ldots, n$ são observados em um conjunto discreto de localizações amostrais, x_i, em alguma região espacial A. Segundo, cada valor observado Y_i é uma versão ruidosa de um fenômeno espacial contínuo não observável, $S(x)$, nas correspondentes localizações amostrais x_i, (Diggle e Ribeiro Jr, 2007). Importante notar que o fenômeno de interesse, $S(x)$, varia continuamente sobre toda a região, porém só se é capaz de medir uma versão ruidosa deste, e em algumas localizações espaciais. O objetivo mais comum deste tipo de análise, é recuperar o processo $S(x)$. Os modelos estatísticos comumente utilizados para analisar este tipo de dados, são os modelos geoestatísticos.

Uma abordagem diferente para recuperar o processo, $S(x)$, é conhecida como Modelo Aditivo Generalizado (Hastie e Tibishirani, 1990), que pode ser descrito como uma extensão do Modelo Linear Generalizado (McCullagh e Nelder, 1983), porém com um ou mais preditores lineares envolvendo a soma de funções suaves (*smooth functions*), no caso espacial de coordenadas geográficas.

O objetivo deste artigo é comparar o desempenho do modelo geoestatístico, com o modelo aditivo generalizado, para recuperar uma superfície contínua, atribuindo para a variável resposta a distribuição Gaussiana, por esta ser a distribuição de maior uso na comunidade científica em geral. Esta comparação será realizada através de simulações, para cada classe de modelo serão calculadas as seguintes medidas de desempenho preditivo: Erro quadrático médio esperado, coeficiente de correlação de Pearson esperado e nível médio de cobertura esperado.

O presente artigo encontra-se dividido em 4 seções, esta primeira busca dar uma visão geral da aplicação dos modelos estatísticos considerados na análise e apresenta os objetivos do trabalho. Na segunda seção apresenta-se os aspectos inferenciais sobre os modelos geoestatístico e aditivo generalizado, e o procedimento de simulação usado para fazer a comparação. A seção três apresenta os principais resultados, obtidos a partir da análise das simulações. A quarta e última seção traz as principais conclusões e recomendações da utilizaçõ dos diferentes modelos.

2 Metodologia

Esta seção, traz um resumo sobre o processo de estimação dos modelos geoestatístico e aditivo generalizado, além do procedimento de simulação utilizado na análise.

2.1 Modelo geoestatístico

Esta seção, vai explorar o processo de estimação do modelo geoestatístico, baseado apenas pelo estudo da função de verossimilhança, para uma visão das várias outras formas de estimação do modelo geoestatístico ver (Diggle e Ribeiro, 2007).

O método de Máxima Verossimilhança tem propriedades ótimas para grandes amostras. Sob certas condições de regularidade (Cox e Hinkley, 1974), o estimador de máxima verossimilhança é assintóticamente normalmente distribuído, não viesado e eficiente. O processo de estimação apresentado nesta seção segue a contribuição de Diggle e Ribeiro, 2007.

Por simplicidade, considera-se um modelo geoestatístico gaussiano com uma tendência linear, denotado por $\mu(x)$. Isto segue da mesma forma para a inclusão de tendências polinômiais, ou mais geralmente, covariáveis referenciadas espacialmente. Assim para $\mu(x) = D\beta$,

$$Y \sim N(D\beta, \sigma^2 R(\phi) + \tau^2 I)$$

onde D é uma matriz de covariáveis $n \quad x \quad p$, β é o correspondente vetor de parâmetros de regressão,R é uma função de correlação válida. Para este trabalho, definiu-se esta função como sendo a função de correlação exponencial, que é escrita da seguinte forma $\exp(-u/phi)$ que depende de um parâmetro, ϕ, e de uma matriz de distâncias, entre cada par de pontos u. O parâmetro σ^2 pode ser considerado como a amplitude do sinal $S(x)$ e o parâmetro τ^2 como o erro de medida cometido (ruído no sinal). Note que a soma dos dois efeitos é a variância total, sendo uma parte atribuída ao sinal $S(x)$ e outra parte a erros de medida, as vezes também referida como efeito de pequena escala ou efeito pepita.

Sendo assim, a função de log-verossimilhança é dada por:

$$\begin{aligned} L(\beta, \tau^2, \sigma^2, \phi) = -0.5nlog(2\pi) + log|(\sigma^2 R(\phi) + \tau^2 I)| + \\ (y - D\beta)^T (\sigma^2 R(\phi) + \tau^2 I)^{-1} (y - D\beta) \end{aligned}$$

a maximização desta função com relação aos parâmetros desconhecidos, traz as estimativas de máxima verossimilhança, para os parâmetros envolvidos no modelo. A proposta de algoritmo para a maximização da função log-verossimilhança é como segue.

Primeiro, procede-se uma reparemetrização fazendo $\upsilon^2 = \tau^2/\sigma^2$ e escreve-se $V = R(\phi) + \upsilon^2 I$. Considerando V conhecido, a log-verossimilhança é maximizada com,

$$\widehat{\beta}(V) = (D^T V^{-1} D)^{-1} D^T V^{-1} y,$$

e

$$\widehat{\sigma^2}(V) = n^{-1} y - D\widehat{\beta}(V)^T V^{-1} y - D\widehat{\beta}(V).$$

Substituindo as expressões de $\widehat{\beta}(V)$ e $\widehat{\sigma^2}(V)$ na função de log-verossimilhança, obtem-se a log-verossimilhança concentrada.

$$L_c(\upsilon^2, \phi) = -0.5nlog(2\pi) + nlog\widehat{\sigma^2}(V) + log|V| + n$$

Esta função pode ser otimizada numericamente com respeito a ϕ e υ, e por substituição se obtem $\widehat{\sigma^2}$ e $\widehat{\beta}$.

Uma vez ajustado o modelo, pode-se obter a predição espacial. Em termos gerais, o problema de predição pode ser resumido da seguinte forma. Deixe Y denotar o vetor das realizações da variável aleatória observada, e deixe T denotar uma outra variável aleatória a qual deseja-se predizer com base nos valores de Y. Um preditor pontual para T é uma função de Y, que pode-se denotar por $\widehat{T} = t(Y)$.

O erro quadrático médio de predição de $\widehat{T}$ é

$$EQM(\widehat{T}) = E[(\widehat{T} - T)^2]$$

onde a esperança é com respeito a distribuição conjunta de T e $\widehat{T}$ ou, equivalentemente, a distribuição de T e Y. Supondo inicialmente que o objetivo é predizer o valor do sinal $S(x)$ em uma localização arbitraria, ou seja, o objetivo é predizer $T = S(x)$. Então, (T, Y) é uma Normal Multivariada, e pode-se obter o menor erro quadrático médio de predição, $\widehat{T}$ usando o seguinte resultado padrão da distribuição Normal Multivariada.

Seja $X = (X_1, X_2)$ sendo a distribuição conjunta Normal Multivariada, com um vetor de média $\mu = (\mu_1, \mu_2)$ e matriz de covariância:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Assim, $X \sim NMV(\mu, \Sigma)$. Então, a distribuição condicional de X_1 dado X_2 é uma Normal Multivariada, $X_1|X_2 \sim NMV(\mu_{1|2}, \Sigma_{1|2})$, onde

$$\mu_{1|2} = \mu_1 + \Sigma_{1|2}\Sigma_{2|2}^{-1}(X_2 - \mu_2)$$

e

$$\Sigma_{1|2} = \Sigma_{1|1} - \Sigma_{1|2}\Sigma_{2|2}^{-1}\Sigma_{2|1}$$

Aplicando o resultado acima para o problema de predição, note que (T, Y) é uma Normal Multivariada com vetor de média μ_1 e matriz de variância

$$\begin{bmatrix} \sigma^2 & \sigma^2 r' \\ \sigma^2 r & \sigma^2 V \end{bmatrix}$$

onde r é o vetor com os elementos $r_i = R(||x - x_i||) : i = 1, \ldots n..$ Seguindo, o resultado fazendo $X_1 = T$ e $X_2 = Y$ tem-se que o preditor com o menor erro quadrático médio para $T = S(x)$ é

$$\widehat{T} = \mu + r'V^{-1}(Y - \mu_1)$$

com variância de predição

$$Var(T|Y) = \sigma^2(1 - r'V^{-1}r)$$

Note que neste caso especial da distribuição Normal Multivariada, a variância condicional não depende de Y, e o erro quadrático médio é igual a variância de predição.

2.2 Modelo Aditivo Generalizado

Um modelo aditivo generalizado, (Hastie e Tibishirani, 1990) pode ser descrito como uma extensão do modelo linear generalizado (McCullagh e Nelder, 1983), porém com um ou mais preditores lineares envolvendo a soma de funções suaves *(smooth functions)* das covariáveis. O modelo torna-se semi-paramétrico e pode ser escrito da seguinte forma:

$$g(\mu_i) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots \quad (1)$$

onde

$$\mu_i \equiv E(Y_i) \quad \text{e} \quad Y_i \sim \quad \text{família exponencial.}$$

Y_i é a variável resposta com distribuição de probabilidade na família exponencial, X_i^* é uma linha da matriz do modelo, para a parte estritamente paramétrica, θ é o vetor de parâmetros correspondentes e as f_j são funções suaves das covariáveis x_k.

As funções suaves, podem ter mais de uma covariável como argumento, conforme ilustrado por f_3 nessa expressão. O modelo definido desta forma, proporciona flexibilidade na especificação da forma da relação entre a variável resposta e as covariáveis. Casos particulares, como a especificação do modelo apenas em termos de funções suaves também são possíveis, evitando desta forma, qualquer suposição de relacionamento linear entre as variáveis. Entretanto, a flexibilidade de tais modelos vem acompanhada de dois novos problemas teóricos: como representar as funções suaves e como estimar os parâmetros envolvidos neste modelo.

O problema de como representar a função suave univariada, pode ser resolvido usando uma *spline*. Uma *spline* cúbica, é uma curva composta por seções de polinômiais cúbicas, juntas de modo que componham uma função contínua que permita primeira e segunda derivada. Os pontos onde as seções se juntam são conhecidos como *knots* do *spline*. Para um *spline* comum os nós (*knots*) ocorrem onde quer que haja uma curva de referência. Para regressão *spline* que é o interesse aqui, a localização dos nós deve ser especificada.

Tipicamente os *knots* são espalhados de maneira uniforme em toda a extensão dos valores observados de x, ou nos quantis da distribuição de x. Seja qual for o método para encontrar os *knots* eles serão denotados por $x_i^* : i = 1, \ldots, q-2$.

Dada a localização dos *knots*, tem-se muitas formas alternativas e equivalentes de escrever a base para a *spline* cúbica. Uma base simples para ser utilizada, pode ser encontrada nos livros de (Wahba,2000) e (Gu, 2002). Embora a expressão da

função de base dada a seguir seja ligeiramente intimidadora e a definição dada seja um pouco vaga, seu uso na prática é fácil.

$$R(x,z) = \frac{[(z-\frac{1}{2})^2 - \frac{1}{12}][(x-\frac{1}{2})^2 - \frac{1}{12}]}{4} - \frac{[(|x-z|-\frac{1}{2})^4 - \frac{1}{2}(|x-z|-\frac{1}{2})^2 + \frac{7}{240}]}{24}$$

Para mais detalhes sobre esta função de base (Gu, 2002, p.37). Usar esta base de *spline* cúbica para f significa que o modelo em (1) torna-se um modelo linear da forma $y = X\beta + \epsilon$, onde a $i - esima$ linha da matriz do modelo é:

$$X_i = [1, x_i, R(x_i, x_1^*), R(x_i, x_2^*), \ldots, R(x_i, x_{q-2}^*)]$$

Desta forma o modelo pode ser estimado por mínimos quadrados ordinários, que é simples computacionalmente. Apesar deste modelo ser satisfatório, a escolha do grau de suavidade do modelo é essencialmente arbitrária, controlada pela dimensão q da base escolhida.

Uma outra opção de função *spline* para o caso de mais de uma covariável é a *thin plate*. Segundo Wood, 2006 a *thin plate* é uma solução elegante e geral para o problema de estimar uma função suave de variáveis preditoras múltiplas. Wahba, 2000 mostra que *thin plate splines* são uma generalização natural da *spline* polinomial univariada, para duas ou mais dimensões.

A dificuldade com *thin plate splines* é o custo computacional, dado que estes suavizadores têm tantos parâmetros desconhecidos quanto dados (estritamente, número de combinações únicas do preditor). Exceto no caso do preditor simples, o custo computacional da estimação do modelo é proporcional ao cubo do número de parâmetros.

Apesar disto, neste trabalho ele foi preferido e amplamente usado. Uma de suas principais características é a isotropia da penalidade das ondulações, onde tais ondulações são em todas as direções igualmente tratadas, com o ajuste inteiramente invariante para a rotação do sistema de coordenadas das covariáveis preditoras. Segundo Wood, 2006, pg.228 a *thin plate splines* é adequada para suavizar interações entre variáveis medidas na mesma unidade, como coordenadas geográficas, onde a isotropia é assumida como adequada.

Seja qual for a função *spline* escolhida, a escolha da base q que determina o grau de suavidade do modelo é arbitrária. Uma forma simples, para escolher o grau de suavização, é tentar fazer uso de testes de hipóteses, para selecionar q por uma seleção da forma *backward*. Porém, tal opção é um tanto problemática, dado que um modelo com $k-1$ *knots* espalhados de forma uniforme, geralmente não é aninhado a um modelo com k *knots* espalhados de forma uniforme. Outra opção, seria começar com uma grade de valores para os *knots* e dar saltos sequenciais, como parte da seleção *backward*. Entretanto, o resultado de espaçamentos diferentes para os *knots* conduzem a modelos com baixo desempenho, além de os resultados ficarem muito dependentes da posição escolhida para os *knots*.

Uma alternativa para controlar a suavização sem alterar a dimensão da base, é manter a dimensão da base fixa, em um tamanho um pouco maior do que se acha necessário, porém o controle da suavização do modelo é feito adicionando

uma penalização "rugosidade"(*wigglines*) na função objetivo do ajuste por mínimos quadrados. Por exemplo, o ajuste usual feito via mínimos quadrados, visa minimizar a seguinte função objetivo:

$$||y - X\beta||^2.$$

enquanto que, seguindo a idéia de penalização, minimiza-se a seguinte função objetivo:

$$||y - X\beta||^2 + \lambda \int_0^1 [f''(x)]^2 dx$$

em que a integral do quadrado da segunda derivada, do modelo penalizado é o termo *wiggly*, de penalização a falta de suavidade. O relacionamento entre o modelo ajustado e o modelo suavizado é controlado pelo parâmetro de suavização λ. Tem-se desta forma que $\lambda \to \infty$ vai alisar muito os dados. Quando $\lambda = 0$ resulta em não penalização e uma simples regressão é ajustada.

Como f é uma função linear nos parâmetros β_i, a penalização pode sempre ser escrita como uma forma quadrática em relação a β.

$$\int_0^1 [f''(x)]^2 dx = \beta^T S \beta$$

onde S é uma matriz de coeficientes conhecidos. É neste momento que o uso de uma base *spline* mostra suas vantagens, já que,$S_{i+2,j+2} = R(x_i^*, x_j^*)$ para $i, j = 1, \ldots, q-2$ quando a primeira e segunda linhas e colunas são iguais a zero. Desta forma, o problema a ser resolvido pela regressão por *splines* é simplesmente minimizar:

$$||y - X\beta||^2 + \lambda \beta^T S \beta$$

em relação a β. O problema de estimar o grau de suavização para o modelo é agora um problema de estimação do parâmetro λ.

Lembrando que se λ for muito pequeno o modelo não alisará os dados adequadamente, e se o λ for muito grande o modelo alisará muito os dados. Em ambos os casos isto significa que a *spline* estimada $\widehat{f}$ não é uma boa aproximação para a função real f. A melhor escolha para λ é aquela que mais aproxima $\widehat{f}$ da real função f. Um critério adequado para escolher λ seria escolher o λ que minimiza a seguinte expressão:

$$M = \frac{1}{n} \sum_{i=1}^{n} (\widehat{f}_i - f_i)^2,$$

onde a notação $\widehat{f}_i \equiv \widehat{f}(x_i)$ e $f_i \equiv f(x_i)$ foi adotada por conveniência.

Note que f é desconhecida, então M não pode ser usada diretamente. Entretanto é possível derivar uma estimativa da $E(M) + \sigma^2$ que é o erro quadrático médio esperado ao prever uma nova variável. Define-se $\widehat{f}^{[-i]}$ como o modelo ajustado com todos os dados exceto y_i, e define-se o escore ordinário da validação cruzada como:

$$\nu_0 = \frac{1}{n} \sum_{i=1}^{n} (f_i^{[-i]} - y_i)^2.$$

Este escore é o resultado de ajustar o modelo com uma parte dos dados, usar o modelo para predizer os valores retirados, calcular a diferença entre os retirados e os preditos e, com estas diferenças, calcular uma diferença média, passando por todos os dados. Substituindo $y_i - f_i + \epsilon_i$,

$$\begin{aligned}\nu_0 &= \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_i^{[-i]} - f_i - \epsilon_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_i^{[-i]} - f_i)^2 - (\widehat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2.\end{aligned}$$

Se $E(\epsilon_i) = 0$, e ϵ_i e $\widehat{f}_i^{[-i]}$ são independentes, o segundo termo da soma desaparece se a esperança for dada por:

$$E(\nu_0) = \frac{1}{n}E\left(\sum_{i=1}^{n}(\widehat{f}_i^{[-i]} - f_i)^2\right) + \sigma^2.$$

Agora, $\widehat{f}^{[-i]} \equiv \widehat{f}$ com igualdade quando o tamanho da amostra tende a infinito. Assim $E(\nu_0) \equiv E(M) + \sigma^2$ com igualdade quando o tamanho da amostra tende a infinito. Sendo assim escolher λ que minimize ν_0 é uma aproximação razoável para minimizar M. Este processo de escolha de λ para minimizar ν_0 é conhecido como "validação cruzada ordinária". Este método é uma aproximação razoável, já que ele minimiza o erro quadrático de predição. Se os modelos forem julgados apenas por sua capacidade preditiva, modelos mais complicados serão sempre escolhidos, ao invés de modelos mais simples. Desta forma, escolher um modelo que tem máxima sua habilidade em predizer os dados fora dos dados observados não resolve o problema. Além, de ser computacionalmente ineficiente, calcular ν_0 para todos os dados e ajustar o modelo para cada um dos n componentes do conjunto de dados. Felizmente, isto pode ser contornado uma vez que pode ser mostrado que:

$$\nu_0 = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \widehat{f}_i)^2}{(1 - A_{ii})^2}$$

onde $\widehat{f}$ é a estimativa vinda do ajuste com todos os dados e A corresponde a matriz de influência. Na prática, os pesos $1 - A_{ii}$, são frequentemente substituídos pela média ponderada, $tr(I - A)/n$, chegando assim ao escore da validação cruzada generalizada,

$$\nu_g = \frac{n\sum_{i=1}^{n}(y_i - \widehat{f}_i)^2}{[tr(I - A)]^2}.$$

O escore da validação cruzada generalizada (GCV) tem vantagens computacionais sobre o escore de validação cruzada ordinária (OCV), e também tem vantagens em termos de invariância ver (Wahba, 2000, pg.53). É também possível mostrar que GCV minimiza $E(M)$ quando a amostra tende a infinito ou é relativamente grande.

Até o momento, foi discutido de forma rápida, como estimar de forma pontual os parâmetros $\beta's$, além dos parâmetros de suavização $\lambda's$, é necessário quantificar a incerteza associada com estes parâmetros. Em particular, é de interesse obter intervalos de confiança para os parâmetros β e λ. Existem duas abordagens para estimar a incerteza. Primeiro, escrevendo $S = H + \sum_i \lambda_i S_i$ e lembre que os estimadores dos parâmetros são da forma,

$$\widehat{\beta} = (X^T W X + S)^{-1} X^T W y$$

onde os dados, y, tem matriz de covariância $W^{-1}\phi$, tem-se que

$$V_e = (X^T W X + S)^{-1} X^T W X (X^T W X + S)^{-1} \phi$$

é a matriz de covariância para os estimadores $\widehat{\beta}$. Vindo da normalidade de y, ou para grandes amostras normalidade multivariada de $X^T W y$, e segue que aproximadamente,

$$\widehat{\beta} \sim N(E(\widehat{\beta}), V_e)$$

Geralmente $E(\widehat{(\beta)}) \neq \beta$, isto é um problema para usar este resultado para calcular intervalos de confiança. Entretando, se $\beta = 0$ então $E(\widehat{\beta}) = 0$, assim este resultado pode ser usado para testar termos do modelo para a igualdade a zero.

Uma alternativa é usar uma abordagem Bayesiana para quantificar a incerteza, com resultados vindos de uma matriz de covariância posteriori Bayesiana para os parâmetros,

$$V_\beta = (X^T W X + S)^{-1} \phi$$

e a correspondente distribuição a posteriori para estes parâmetros,

$$\beta \sim N(\widehat{\beta}, V_\beta)$$

Mesmo para dados não normais, a normalidade a posteriori para os parâmetros é uma aproximação justificada para grandes amostras. Este último resultado pode ser usado diretamente para calcular intervalos de credibilidade para os parâmetros, e foi a abordagem usada neste trabalho. Todos os procedimentos aqui descritos estão implementados no pacote *mgcv* (Wood, 2008).

2.3 Procedimento para a Simulação

Como se deseja estudar o comportamento dos modelos, para recuperar uma superfície contínua, em um primeiro moment,o se fará a simulação de uma grade irregular de $50x50$ pontos, dentro de uma quadrado unitário. Tentando de forma computacional, descrever uma superfície contínua que será considerada como os valores reais do fenômeno em toda a área.

Para gerar este conjunto de dados, o método padrão é simular amostras independentes $Z = (Z_1, \ldots, Z_n)$ provenientes de uma distribuição Normal padrão, e aplicar uma transformação linear,

$$S = AZ$$

onde A é uma matriz tal que $AA' = \Sigma$. Duas formas para construir a matriz A são comumente usadas, a decomposição de Cholesky e a decomposição em valores singulares.Neste artigo, se fará uso da decomposição de Cholesky. Para a avaliação dos modelos será seguido o seguinte algoritmo.

1. Gerar uma realização do processo, usando a função de correlação exponencial com os seguintes quatro conjuntos de parâmetros.
   - Conjunto 1 - $(\beta = 50, \sigma^2 = 1, \tau^2 = 0, \phi = 0.25)$.
   - Conjunto 2 - $(\beta = 50, \sigma^2 = 0.75, \tau^2 = 0.25, \phi = 0.25)$.
   - Conjunto 3 - $(\beta = 50, \sigma^2 = 0.5, \tau^2 = 0.5, \phi = 0.25)$.
   - Conjunto 4 - $(\beta = 50, \sigma^2 = 0.25, \tau^2 = 0.75, \phi = 0.25)$.
2. Da realização do processo retirar amostras de tamanho $n = 50, n = 150$ e $n = 250$.
3. Para cada tamanho de amostra, ajustar o modelo geoestatístico e o modelo aditivo generalizado.
4. Usar os modelos ajustados com as amostras, para recuperar a superfície completa.
5. Avaliar as estimativas de média e variância de predição para os dois modelos.
6. Calcular o EQM - Erro Quadrático Médio, coeficiente de correlação de Pearson entre os preditos e simulados e o nível de cobertura, definido pelo percentual de pontos, contidos no intervalo de predição para ambos os modelos.
7. Repetir o procedimento 200 vezes.
8. Obter o EQM - Erro Quadrático Médio esperado, coeficiente de correlação esperado e nível de cobertura esperado. Calculando os respectivos intervalos de confiança de 95% baseados na distribuição Normal.

Todos os procedimentos descritos foram implementados em R, as funções estão disponíveis em www.leg.ufpr.br/pessoais:wagner. Para gerar as simulações foi usada a função $grf()$ do pacote $geoR$ (Ribeiro Jr e Diggle, 2001). Para o ajuste do modelo geoestatístico foi usada a função $likfit()$ do mesmo pacote. Para o ajuste do modelo aditivo generalizado, foi usada a função gam do pacote $mgcv$.

3 Resultados

A apresentação dos resultados, será feita basicamente usando técnicas gráficas e estatísticas descritivas. A primeira medida avaliada é o erro quadrático médio. A figura 1 apresenta os resultados comparando o modelo geoestatístico com o modelo aditivo generalizado, com relação ao erro quadrático médio de predição, de acordo com o tamanho de amostra e conjunto de parâmetros geradores.

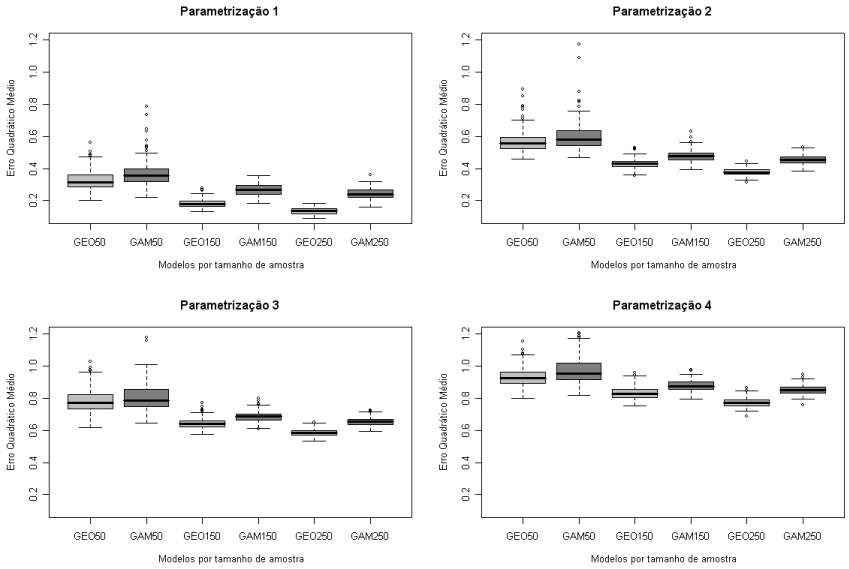

Figura - 1: Comparação do erro quadrático médio por tamanho de amostra e parametrização.

Pelos gráficos apresentados na figura 1, fica evidente o melhor desempenho do modelo geoestatístico, quando comparado ao modelo aditivo generalizado. Com o aumento do tamanho da amostra o EQM diminui de forma significativa, indicando que tamanhos de amostras maiores são mais confiáveis para fazer a predição, resultado esperado, pois mais amostras melhora a estimação dos parâmetros envolvidos nos modelos. Outro resultado bastante evidente, é que o EQM aumenta de acordo com a estrutura de parametrização utilizada, quando a proporção entre sinal σ^2 e ruído τ^2 é baixa o EQM é pequeno, ao aumentar a proporção de ruído mantendo a mesma variabilidade total, as predições ficam menos precisas, ou seja, aumenta o EQM. Para melhor explorar estes resultados, a tabela 3 traz a raiz quadrado do erro quadrático médio esperado, gerando portanto, o erro padrão médio esperado e intervalos baseados na distribuição Normal com 95% de confiança, para cada um dos modelos de acordo com o tamanho da amostra.

Pelos resultados apresentados na tabela 3, é possível dizer que o modelo

Tabela - 1: Erro padrão médio esperado e intervalos de confiança, de acordo com tamanho de amostra e parametrização para os modelos GEO e GAM.

Estatísticas	GEO50	GAM50	GEO150	GAM150	GEO250	GAM 250
Int. Máximo 1	0.581	0.615	0.430	0.521	0.372	0.498
EQM esperado 1	0.571	0.603	0.424	0.514	0.368	0.492
Int Mínimo 1	0.560	0.591	0.419	0.507	0.363	0.485
Int. Máximo 2	0.760	0.783	0.659	0.695	0.617	0.677
EQM esperado 2	0.751	0.772	0.654	0.690	0.614	0.673
Int. Mínimo 2	0.743	0.761	0.650	0.685	0.611	0.669
Int. Máximo 3	0.890	0.905	0.805	0.832	0.768	0.813
EQM esperado 3	0.882	0.896	0.801	0.828	0.765	0.810
Int. Mínimo 3	0.874	0.887	0.797	0.824	0.762	0.807
Int. Máximo 4	0.971	0.994	0.915	0.940	0.881	0.925
EQM esperado 4	0.965	0.986	0.911	0.936	0.878	0.922
Int. Mínimo 4	0.958	0.978	0.907	0.932	0.874	0.919

geoestatístico apresenta menor erro padrão médio, para praticamente todas as combinações entre tamanhos de amostra e parametrização, apenas para $n = 50$ na parametrização 3, os modelos são equivalentes. É interessante observar, que o erro padrão médio esperado para o modelo geoestatístico, é de $0.368(0.363 - 0.373)$ para $n = 250$ na parametrização 1, enquanto que para o modelo aditivo generalizado o erro padrão médio esperado, é de $0.492(0.485 - 0.498)$ para a mesma parametrização, ou seja, o modelo geoestatístico apresenta um erro padrão médio 33% menor que o modelo aditivo, nas condições consideradas.

Conforme aumenta a proporção de ruído, os modelos vão perdendo a capacidade preditiva, inflacionando o erro padrão médio. Com esse aumento, a diferença na performance entre os dois modelos também diminui, por exemplo, o erro padrão médio esperado do modelo geoestatísitco, para $n = 250$ na parametrização 4, é de $0.878(0.874 - 0.881)$ e para o modelo aditivo generalizado, é de $0.922(0.919 - 0.925)$ nas mesmas condições. A diferença da performance entre os modelos que era de 33%, na parametrização 1 passa a ser de apenas 5%, ainda mostrando o modelo geoestatístico como de melhor performance.

Também, se observa uma queda bastante rápida do erro padrão médio com o aumento da amostra, para os dois modelos. Enquanto, o erro padrão médio esperado, é de $0.571(0.560 - 0.581)$ para $n = 50$ quando a amostra aumenta para $n = 250$ o erro padrão médio esperado, passa a $0.368(0.363 - 0.372)$ uma queda de aproximadamente, 55% para o modelo geoestatístico, e de $0.603(0.591 - 0.615)$ para $0.492(0.485 - 0.498)$, uma queda de aproximadamente 24%, para o modelo aditivo generalizado, na primeira parametrização, que considera que os dados são medidos sem nenhum ruído $\tau^2 = 0$.

Quando aumenta-se o ruído a dimensão da queda do erro padrão médio diminui sensivelmente, por exemplo, para $n = 50$ o erro padrão médio é de $0.965(0.958 - 0.971)$ passando a $0.878(0.874 - 0.881)$ para $n = 250$ uma queda de

aproximadamente 8% para a parametrização 4 no modelo geoestatístico. No modelo aditivo generalizado, para $n = 50$ o erro padrão médio, é de $0.986(0.978 - 0.994)$ passando a $0.922(0.919 - 0.925)$, uma queda de aproximadamente 6% bem mais tímida que para a parametrização 1.

Avaliando o erro padrão médio para as diferentes parametrizações, observa-se um aumento bastante rápido aumentando a proporção sinal σ^2 e ruído τ^2. Na primeira parametrização ($\sigma^2 = 1 \quad \tau^2 = 0$) tem-se para $n = 250$ um erro padrão médio, de apenas $0.368(0.363 - 0.373)$ passando a $0.878(0.874 - 0.881)$ na quarta parametrização ($\sigma^2 = 0.25 \quad \tau^2 = 0.75$) um aumento de aproximadamente 138%, para o modelo geoestatístico, e de $0.492(0.485 - 0.498)$ para $0.922(0.919 - 0.925)$ um aumento de aproximadamente 87%, para o modelo aditivo generalizado, o mesmo se observa para os outros tamanhos amostrais.

Uma outra forma de avaliar a qualidade das predições obtidas pelos modelos ajustados, é calcular o coeficiente de correlação, entre os valores preditos e os valores simulados, que neste caso representam o valor real da localização que se deseja predizer. Espera-se que se o modelo tem uma boa capacidade de predição, esta correlação seja próxima da unidade. A figura 2 faz a comparação do modelo geoestatístico com o modelo aditivo generalizado, de acordo com a parametrização e tamanho de amostra, com relação ao coeficiente de correlação linear de Pearson.

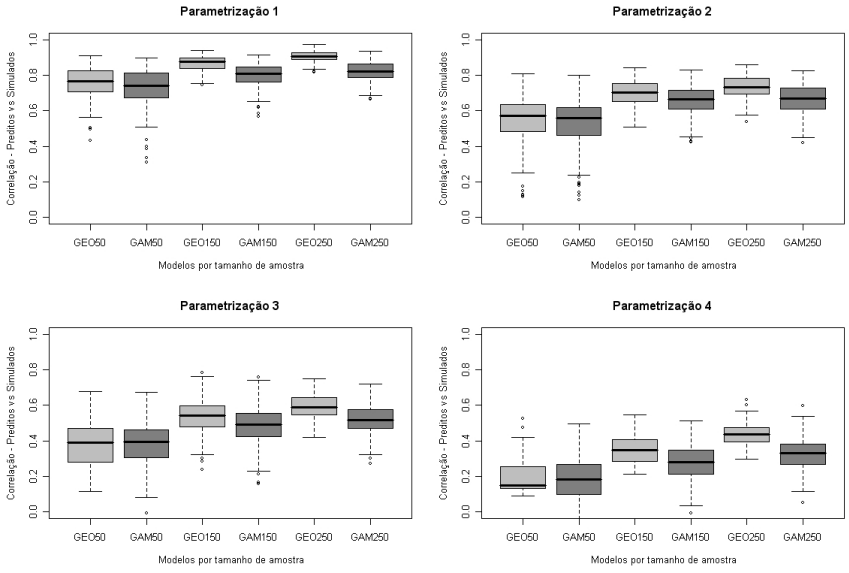

Figura - 2: Comparação do coeficiente de correlação por tamanho de amostra e parametrização.

Pelos gráficos apresentados na figura 2, é possível constatar o melhor desempenho do modelo geoestatístico, comparado com o modelo aditivo

generalizado, no que diz respeito, a correlação entre os preditos e simulados. Com o aumento do tamanho da amostra a correlação aumenta de forma lenta, mais chega próximo da unidade para o modelo geoestatístico, na parametrização 1. Outro resultado bastante evidente, é que a correlação diminui bastante de acordo com a estrutura de parametrização utilizada, quando a proporção entre sinal σ^2 e ruído τ^2 é baixa a correlação apresenta valores altos próximos da unidade, ao aumentar a proporção de ruído mantendo a mesma variabilidade total, as correlações diminuem bastante, ficando na faixa do 0.2 a 0.4 na parametrização 4. Para melhor explorar estes resultados a tabela 3, traz o coeficiente de correlação linear de Pearson, e seus respectivos intervalos baseados na distribuição Normal com 95% de confiança, para cada um dos modelos, de acordo com o tamanho da amostra e parametrização.

Tabela - 2: Coeficiente de correlação médio esperado e intervalos de confiança, de acordo com tamanho de amostra e parametrização para os modelos GEO e GAM.

Estatísticas	GEO50	GAM50	GEO150	GAM150	GEO250	GAM 250
Int. Máximo 1	0.778	0.75	0.877	0.813	0.911	0.834
COR esperado 1	0.761	0.731	0.869	0.800	0.905	0.823
Int Mínimo 1	0.743	0.710	0.861	0.787	0.899	0.812
Int. Máximo 2	0.578	0.561	0.713	0.673	0.748	0.685
COR esperado 2	0.552	0.534	0.699	0.657	0.735	0.668
Int. Míinimo 2	0.526	0.506	0.684	0.640	0.723	0.652
Int. Máximo 3	0.398	0.403	0.550	0.505	0.604	0.535
COR esperado 3	0.374	0.379	0.533	0.485	0.591	0.518
Int. Mínimo 3	0.344	0.354	0.515	0.465	0.577	0.501
Int. Máximo 4	0.213	0.203	0.365	0.296	0.447	0.339
COR esperado 4	0.195	0.179	0.349	0.278	0.435	0.322
Int. Mínimo 4	0.177	0.154	0.334	0.260	0.424	0.305

Pelos resultados apresentados na tabela 3, é possível dizer que o modelo geoestatístico apresenta maior coeficiente de correlação esperado, para praticamente todas as combinações, entre tamanhos de amostra e parametrização, apenas para $n = 50$ nas quatro parametrizações os modelos são equivalentes. O modelo geoestatístico, apresenta coeficientes elevados, por exemplo, na parametrização 1 com $n = 250$ o coeficiente, é de $0.905(0.899 - 0.911)$ e para o modelo aditivo generalizado, é de $0.823(0.812 - 0.834)$ uma diferença de aproximadamente 10%. Para a parametrização 4 e mesmo tamanho de amostra os resultados são, $0.435(0.424 - 0.447)$ para o geoestatístico, e de $0.322(0.305 - 0.339)$ para o aditivo generalizado, uma diferença de aproximadamente 35%, mostrando que por este critério ao aumentar a variabilidade dos dados, o modelo geoestatístico se mostra mais confiável para realizar as predições. Resultado diferente do encontrado para o erro padrão médio, que com o aumento da variabilidade os modelos apresentam resultados mais próximos.

É possível observar, um aumento bastante rápido do coeficiente de correlação com o aumento da amostra, para os dois modelos. Enquanto o coeficiente

de correlação esperado, é de $0.761(0.743 - 0.778)$ para $n = 50$, quando a amostra aumenta para $n = 250$ o coeficiente de correlação esperado passa a $0.905(0.899 - 0.911)$, um aumento de aproximadamente 18% para o modelo geoestatístico, e de $0.731(0.710 - 0.752)$ para $0.823(0.812 - 0.834)$ um aumento de aproximadamente 8%, para o modelo aditivo generalizado, para a primeira parametrização. Quando aumenta-se o ruído a dimensão da queda do coeficiente de correlação aumenta sensivelmente, por exemplo, para $n = 50$ a correlação, é de $0.195(0.177 - 0.213)$ passando a $0.435(0.424 - 0.447)$ para $n = 250$, um aumento de aproximadamente 123%, para a parametrização 4 no modelo geoestatístico. No modelo aditivo generalizado, para $n = 50$ a correlação, é de $0.179(0.154 - 0.203)$ passando a $0.322(0.305 - 0.339)$ um aumento de aproximadamente 79%, bem mais forte que para a parametrização 1, indicando que quando o nível de ruído é alto, um tamanho de amostra maior é necessário, para se ter confiança nas predições.

Avaliando o coeficiente de correlação esperado para as diferentes parametrizações, observa-se um decréscimo bastante rápido, aumentando a proporção sinal σ^2 ruído τ^2. Na primeira parametrização ($\sigma^2 = 1 \quad \tau^2 = 0$) tem-se para $n = 250$ uma correlação esperada de $0.905(0.899 - 0.911)$ passando a $0.435(0.424 - 0.447)$, na quarta parametrização ($\sigma^2 = 0.25 \quad \tau^2 = 0.75$) um decréscimo de aproximadamente 108%, para o modelo geoestatístico, e de $0.823(0.812 - 0.834)$ para $0.322(0.305 - 0.339)$ um decrescimo de aproximadamente 155%, para o modelo aditivo generalizado, o mesmo se observa para os outros tamanhos amostrais.

As duas medidas de desempenho anteriores, só levam em consideração as predições médias, desprezando a incerteza associada a cada predição. Para levar em consideração a incerteza, em cada simulação foi também calculado intervalos de confiança para cada predição feita. O nível de cobertura, é a proporção de vezes em que o intervalo de predição contêm os valores reais, no caso deste estudo dos valores simulados. Na situação ideal, como os intervalos construídos foram de 95% espera-se que a taxa de cobertura fique em torno deste valor, e o nível de cobertura esperado seja de aproximadamente 95% com igualdade quando o número de simulações tender a infinito. Também espera-se que o nível de cobertura não dependa de forma acentuada do tamanho da amostra, pois uma amostra menor deve refletir em intervalos maiores e não afetar a proporção de intervalos contendo o valor real da predição. A figura 3 traz os resultados do nï¿½vel de cobertura por tamanho de amostra e parametrizaï¿½ï¿½o utilizada na simulaï¿½ï¿½o.

O que salta aos olhos na figura 3 é o mal desempenho do modelo geoestatístico, na parametrização 4 com $n = 50$. Esse resultado pode ser explicado olhando a distribuição empirica dos estimadores de máxima verossimilhança deste modelo, nas condições consideradas, a figura 4 traz essas distribuições.

Como pode-se observar na figura 4 a distribuição do $\widehat{\beta}$ é simétrica e centrada no valor 50 como era esperado, mostrando que este parâmetro está sendo estimado de forma eficiente. Ao analisar a distribuição empirica do $\widehat{\tau^2}$ observa-se uma calda extremamente pesada a esquerda, tirando esta calda a distribuição mostra-se simétrica e centrada aproximadamente em 0.75, que seria o resultado esperado,

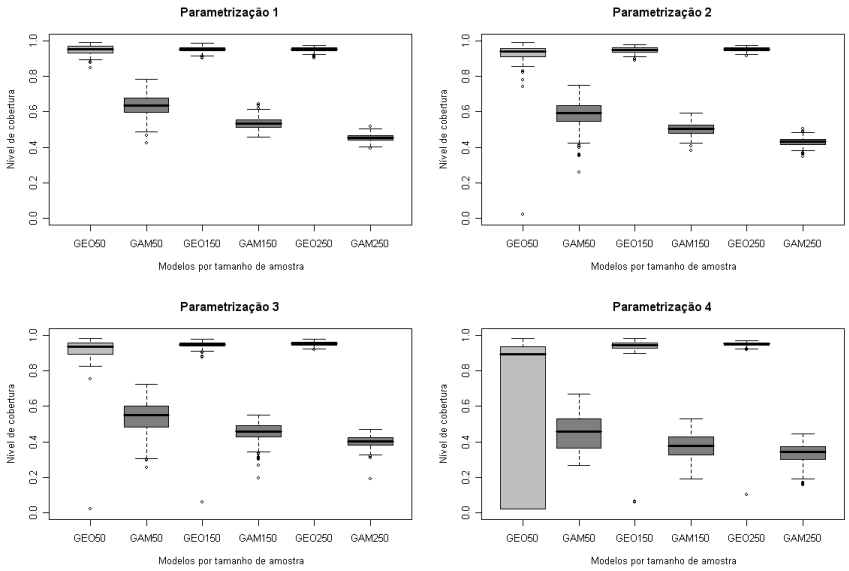

Figura - 3: Comparação do nível médio esperado de cobertura por tamanho de amostra e parametrização.

porém pela calda extremamente pesada, este parâmetro é mal estimado uma quantidade considerável de vezes. Avaliando a distribuição empirica do $\widehat{\sigma^2}$ verifica-se novamente uma quantidade excessiva de estimativas iguais a zero, para ser mais exato são 88 zeros em 200 estimativas, aproximadamente 44% mostrando que sob esta parametrização este parâmetro é mal estimado.

Isso é que explica o baixo desempenho do modelo geoestatístico, quando avaliado pelo intervalo de cobertura, apesar da sua estimativa de média ser boa, as suas estimativas de variância são muito ruins, chegando ao extremo de estimar a variância $\sigma^2 = 0$, o que faz com que as variâncias estimadas de predições sejam iguais a zero, e portanto, nenhum intervalo contêm os valores reais. Como o processo de predição usado pelo modelo geoestatístico interpola os valores exatos nos pontos de amostragem, é o que explica o boxplot começar em 0.02 que é exatamente 50/2500 sendo 50 pontos amostrais em uma malha de 2500 pontos.

Para terminar a análise, observa-se na distribuição empirica do $\widehat{\phi}$ novamente uma quantidade excessiva de zeros. Como era esperado, sempre que a estimativa do $\sigma^2 = 0$ a estimativa do ϕ também é igual a zero. Ou seja, o modelo é confundido pelo ruído forte nas observações e com esse tamanho pequeno de amostra não consegue separar o sinal do ruído e apresenta como sua melhor estimativa, uma simples média dos dados, como se os dados fossem independentes. Essa estimativa nesta situação não é tão ruim, como mostra o erro padrão médio e o coeficiente de correlação para o modelo geoestatístico, que mesmo nesta situação, apresentou resultados muito

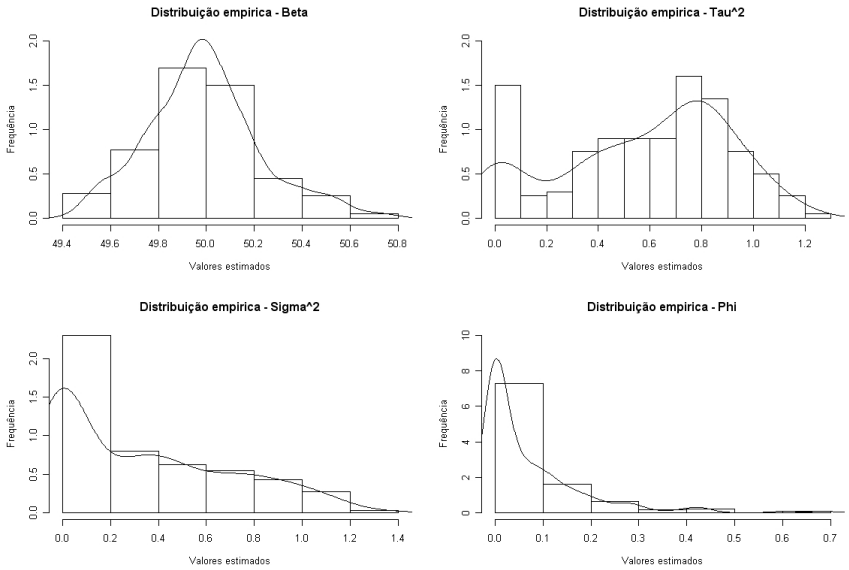

Figura - 4: Distribuição empirica dos estimadores de máxima Verossimilhança para os parâmetros do modelo geoestatístico, parametrização 4,$n = 50$.

próximos ao modelo aditivo generalizado, sendo melhor com relação ao erro padrão médio e pior na medida de correlação.

Porém, quando aumenta-se o tamanho da amostra o modelo geoestatístico estabiliza suas estimativas e apresenta resultados excelentes, com nível de cobertura muito próximo do valor esperado de 95%, enquanto o modelo aditivo generalizado apresenta resultados inferiores, tendo uma cobertura máxima de apenas 63%, mostrando que os intervalos construídos por este método são muito otimistas, sendo em geral, menores do que deveriam ser.

Para melhor ilustrar estes resultados na tabela 3, apresenta-se o nível de cobertura médio esperado e seus respectivos intervalos baseados na distribuiçã Normal com 95% de confiança.

Como mostra a tabela 3, o modelo geoestatístico é superior ao modelo aditivo generalizado, para todas as parametrizações e tamanhos de amostra considerados. O nível de cobertura não é afetado pelo tamanho de amostra no modelo geoestatístico, como pode-se observar para $n = 50$ o nível de cobertura, é de $0.949(0.944 - 0.954)$ e para $n = 250$, é de $0.953(0.950 - 0.955)$ o que mostra a equivalência, como era esperado. Já para o modelo aditivo generalizado, o nível de cobertua é fortemente influenciado pelo tamanho da amostra, e piora com amostras maiores, por exemplo, para $n = 50$ na parametrização 4 o nível de cobertura, é de $0.528(0.439 - 0.616)$ e para $n = 250$, é de $0.334(0.323 - 0.345)$ uma queda de aproximadamente 58%, resultado não esperado e que merece maior atenção para futuros trabalhos.

Tabela - 3: Nível de cobertura médio esperado e intervalos de confiança, de acordo com tamanho de amostra e parametrização para os modelos GEO e GAM.

Estatísticas	GEO50	GAM50	GEO150	GAM150	GEO250	GAM 250
Int. Máximo 1	0.954	0.645	0.956	0.543	0.955	0.459
Cobertura 1	0.949	0.633	0.953	0.537	0.953	0.454
Int Míinimo 1	0.944	0.621	0.950	0.531	0.950	0.450
Int. Máximo 2	0.939	0.597	0.952	0.510	0.954	0.434
Cobertura 2	0.913	0.582	0.948	0.503	0.952	0.430
Int. Mínimo 2	0.887	0.566	0.945	0.496	0.950	0.425
Int. Máximo 3	0.875	0.555	0.956	0.465	0.955	0.407
Cobertura 3	0.814	0.535	0.943	0.454	0.953	0.400
Int. Mínimo 3	0.753	0.515	0.931	0.443	0.951	0.393
Int. Máximo 4	0.616	0.472	0.856	0.385	0.953	0.345
Cobertura 4	0.528	0.453	0.789	0.370	0.922	0.334
Int. Mínimo 4	0.439	0.434	0.722	0.355	0.891	0.323

Comparando a qualidade dos dois modelos com relação ao nível de cobertura, é fácil observar que o modelo geoestatístico é muito melhor que o modelo aditivo generalizado, por exemplo, para $n = 250$ na parametrização 1, o nível de cobertura, é de $0.953(0.950 - 0.955)$ para o modelo geoestatístico, e de $0.457(0.450 - 0.459)$ nas mesmas condições para o modelo aditivo generalizado, uma queda de aproximadamente 108%, quantificando o quanto o nível de cobertura do geoestatístico é melhor que do aditivo generalizado. O nível de cobertura não é muito influenciado pelas diferentes parametrizações para amostras grandes > 150, porém para $n = 50$ a estimação do modelo geoestatístico é muito volátil, indicando que nestas situações o modelo aditivo pode ser preferido, porém com restrições as interpretações de seus intervalos de credibilidade.

4 Conclusões

A comparação de modelos estatísticos através de simulações, é uma atividade interessante e de muito valor para embasar futuras análises, utilizando os modelos considerados. Saber os pontos fortes e fracos de cada classe de modelos, pode ajudar de forma significativa na escolha de um deles para realizar uma análise com dados reais.

O presente artigo, buscou confrontar o modelo geoestatístico com o modelo aditivo generalizado, para a reconstituição de uma superfície contínua. É claro, que a forma de construção desta superfície baseada em função de correlação, e parâmetros originários do modelo geoestatístico, influenciaram os resultados em favor deste, mais este era um dos objetivos do estudo. Saber como se comporta o modelo aditivo generalizado, para tratar com dados da forma geoestatística.

Os resultados mostram, que nesta situação o modelo geoestatístico é muito superior ao modelo aditivo generalizado, em praticamente todos os critérios

utilizados para a comparação. O resultado mais intrigante é o mal desempenho dos intervalos de credibilidade calculados pelo modelo aditivo, que não chegou nem perto do nível esperado de 95%, e piorou sua cobertura com o aumento do tamanho da amostra. Esse com certeza é um resultado que merece mais dedicação e uma porta para novos estudos.

Deixa-se aqui como futuras extensões para este artigo, considerar outras parametrizações, tamanhos de amostras maiores na faixa de 1500 a 3000, para ver a estabilidade numerica nas implementações dos dois métodos. Também aumentar o número de simulações que aqui foram consideradas apenas 200 para uma malha de 2500 pontos, recomendaria aumentar para pelo menos 1000 modelos em uma malha de 10000 pontos.

Agradecimentos

Agradeço ao Professor Paulo Justiniano Ribeiro Junior pelas discussões durante a elaboração deste artigo. E ao LEG - Laboratório de Estatística e Geoinformação pelo uso dos recursos computacionais.

BONAT, W. H. Comparação entre o modelo geoestatístico e modelo aditivo generalizado Gaussianos para a reconstituição de superfícies contínuas . *Rev.Mat. Estat.,* São Paulo, v.xx, n.x, p.xx-xx, 2008. *Rev. Mat. Estat.* (São Paulo), v. 20, n.1, p. 1-10, 2000.

Referências

DIGGLE, P. J. ; RIBEIRO Jr, P. J. R. *Model-Based Geostatistics*. Hardcover: Springer, 2007.

HASTIE, T. J. ; TIBISHIRANI, R. J. *Generalized additive models*. London: Chapman and Hall, 1990.

McCULLACH, P.; NEDER, J. A. *Generalized linear models*. 2.ed. London: Chapman and Hall, 1989. 511p.

WAHBA, G. *Splines in nonparametric regression*. Madison, Department of Statistics: University of Wisconsin - Technical Report, 2000.

GU, C. *Smoothing splines anova models*. New York: Springer, 2002.

WOOD, S. N. *Generalized additive models: Introduction with R*. Boca Raton: Chapman and Hall, 2006.

WOOD, S.N. *GAMs with GCV smoothness estimation and GAMMs by REML/PQL*. R package version 1.3-31, 2008.

RIBEIRO Jr, P. J. R. ; DIGLLE, P. J. *geoR: a package for geostatistical analysis*. R package version 1.6-14, 2001.

Recebido em 01.01.2008.

Aprovado após revisão em 01.01.2008.