Revista Energia na Agricultura

ISSN 1808-8759

COMPARANDO PREDIÇÕES POR MODELOS GEOESTATÍSTICO E ADITIVO GENERALIZADO PARA RECONSTITUIÇÃO DE SUPERFÍCIES CONTÍNUAS GAUSSIANAS[1]

WAGNER HUGO BONAT[2]; PAULO JUSTINIANO RIBEIRO JUNIOR[3] & WALMES MARQUES ZEVIANI[4]

RESUMO: Os desempenhos dos modelos geoestatístico e aditivo generalizado com *thin plate splines* para a reconstituição de superfícies contínuas gaussianas são comparados através de um estudo de simulação. O procedimento proposto para a comparação leva em consideração fatores sobre o processo gerador da superfície, variabilidade e suavidade, bem como, para o processo de estimação, o tamanho da amostra. Os resultados mostraram que o modelo geoestatístico apresentou resultados superiores aos do *thin plate splines*, nas medidas de adequação de predição utilizadas, erro quadrático médio de predição e nível de cobertura. Os resultados indicaram que a predição por *thin plate splines* foi razoável apenas como análise inicial, ou quando não se está interessado na medida de incerteza das predições espaciais.

Palavras-chave: Geoestatística, modelos aditivos generalizados, predição espacial.

[1] Trabalho aceito para apresentação oral no II Simpósio de Geoestatística Aplicada em Ciências Agrárias

[2] Professor do Departamento de Estatística, LEG/DEST/UFPR - Curitiba/PR – Brasil. wagner@leg.ufpr.br

[3] Professor do Departamento de Estatística, LEG/DEST/UFPR - Curitiba/PR – Brasil. paulojus@leg.ufpr.br

[4] Professor do Departamento de Estatística, LEG/DEST/UFPR - Curitiba/PR – Brasil. walmes@leg.ufpr.br

COMPARING PREDICTIONS BY GEOSTATISTICAL AND GENERALIZED ADDITIVE MODELS FOR RECONSTRUCTING GAUSSIAN SPATIALLY CONTINUOUS SURFACES

SUMMARY: *The performance of geostatistic and generalized additive models with thin plate splines in the reconstruction of spatially continuous surfaces is assessed by a simulation study. The comparison accounts for different factors affecting the generating process, variability and smoothness as well as the sample size for the estimation procedure. The geostatistic model has clearly a better performance compared to the thin plate splines for the assessment measures, mean square error and coverage level. The results indicates that the generalized additive model with thin plate splines is suitable for an initial exploratory analysis or when there is no interest in the uncertainty about the spatial predictions.*

Keywords: *Geostatistics, generalized additive models, spatial prediction.*

1 INTRODUÇÃO

Em diversas situações que requerem análise espacial o interesse é recuperar uma superfície originalmente contínua, a partir de amostras obtidas em um conjunto discreto de localizações dentro de uma área de estudo. Tais situações ocorrem naturalmente, por exemplo, na geologia para estimação de depósitos minerais, na agronomia, para fins de zoneamento agrícola, na epidemiologia para o estudo da distribuição de doenças, na entomologia para a estimação da densidade vetorial de mosquitos transmissores de doenças, na construção civil para o monitoramento de grandes obras como barragens, entre outras.

De forma geral, estes problemas têm em comum o fato de que as amostras têm uma localização no espaço e o fenômeno varia continuamente em uma determinada área ou região de estudo. A presença da componente espacial torna as análises estatísticas convencionais inadequadas, principalmente devido à suposição de independência. A ênfase da análise espacial é mensurar propriedades e relacionamentos, levando em consideração a localização espacial do fenômeno em estudo de forma explícita.

Duas possíveis abordagens são: modelar a resposta como uma função das coordenadas ou modelar a estrutura de covariância, como é tipicamente feito nos procedimentos de geoestatística.

O termo geoestatística refere-se a modelos e métodos para dados seguindo as seguintes características: primeiro, os valores $Y_i : i = 1,...n$ são observados em um conjunto discreto de localizações amostrais, x_i, em alguma região espacial A; assume-se cada valor Y como uma versão ruidosa de um fenômeno espacial contínuo não observável, $S(x)$, nas correspondentes localizações amostrais

(DIGGLE; RIBEIRO Jr, 2007). O objetivo mais comum neste tipo de análise é recuperar o processo $S(x)$ em qualquer conjunto arbitrário de localizações x e tipicamente definindo um mapa de valores da variável na área.

A metodologia comumente utilizada é o modelo geoestatístico, porém algumas outras abordagens também são possíveis. Neste artigo, será considerada uma abordagem completamente distinta em concepção, porém com objetivo idêntico, recuperar o processo $S(x)$. A metodologia conhecida como modelo aditivo generalizado (GAM) (HASTIE; TIBISHIRANI, 1990), pode ser descrita como uma extensão do modelo linear generalizado (McCULLAGH; NELDER, 1989), porém com um ou mais preditores lineares envolvendo a soma de funções suaves (*smooth functions*), no caso espacial de coordenadas geográficas. Neste artigo, esta função suave será representada por um *thin plate splines* (WOOD, 2003), como será descrito com mais detalhes na próxima seção.

O objetivo deste artigo foi comparar através de um processo de simulação o desempenho destas duas abordagens, para recuperar superfícies contínuas gaussianas, ou seja, a distribuição de probabilidade assumida para a variável de interesse (resposta) é a gaussiana, e esta escolha deve-se ao fato desta ser a distribuição de maior uso na comunidade científica em geral.

Nesta primeira seção apresentou-se uma visão geral da aplicação dos modelos estatísticos considerados na análise e os objetivos do trabalho. Na segunda apresentam-se de forma ampla os modelos envolvidos, bem como, o procedimento de simulação utilizado na comparação. A seção três apresenta os principais resultados e a quarta as principais conclusões e recomendações para trabalhos futuros.

2 MATERIAL E MÉTODOS

Um modelo para análise de dados em espaço contínuo é o modelo geoestatístico, como apresentado em Diggle e Ribeiro Jr (2007). O método para estimação de parâmetros adotado neste trabalho é o da máxima verossimilhança, usando a implementação computacional do pacote geoR (RIBEIRO Jr; DIGGLE, 2001).

Como forma alternativa de análise, considerou-se um modelo aditivo generalizado com função suave bidimensional do tipo *thin plate splines* (WOOD, 2003). Este tipo de função suave é preferível, por exemplo, a *tensor product splines* (WOOD, 2006) para suavizar superfícies, uma vez que uma das 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 a rotações no sistema de coordenadas das covariáveis preditoras. Isso torna este um suavizador adequado para representar interações entre covariáveis medidas na mesma unidade, como coordenadas geográficas, quando isotropia é assumida como adequada.

Como o interesse do trabalho foi comparar duas abordagens para recuperar uma superfície contínua, em um primeiro momento é necessário, ter algum mecanismo capaz de gerar diferentes tipos de superfícies controlando alguns fatos relevantes sobre o processo, por exemplo, força da dependência espacial, variabilidade e suavidade do processo. Estes fatos importantes sobre o processo podem ser adequadamente expressos através da sua estrutura de variância-covariância. Diversas funções para descrever esta estrutura são usadas na literatura (DIGGLE; RIBEIRO Jr., 2007). Neste artigo, adotou-se a função de correlação de Matérn (MATERN, 1986), que tem a expressão,

$\rho(u) = \left(2^{\kappa-1}\Gamma(\kappa)\right)^{-1}(u/\varphi)^{\kappa}K_{\kappa}(u/\varphi)$, onde a correlação entre pares de pontos, assumida isotrópica, é função da distância u entre os pontos, $K_{\kappa}(.)$ denota a função de Bessel modificada de ordem κ, $\varphi > 0$ é um parâmetro de escala que associa à extensão da distância até a qual a correlação espacial é relevante, e $\kappa > 0$, é um parâmetro de corpo que determina a suavidade do processo. Adotando uma distribuição Normal Multivariada para descrever o processo gerador de dados, têm-se mais três parâmetros relevantes. A variância do processo espacial $S(x)$ que será denotada por $\sigma^2 > 0$, a variância do ruído, denotado por $\tau^2 > 0$ e por fim a média geral do processo sobre a área denotada por β. Controlando estes cinco parâmetros, é possível gerar diversos tipos de processos com características bem definidas.

Para gerar amostras deste processo, o método padrão é simular amostras independentes $Z = (Z_{1,}...Z_n)$ provenientes de uma distribuição Normal padrão, e aplicar uma transformação linear, $S = AZ$, onde A é uma matriz tal que $A'A = \Sigma$, a matriz de variância-covariância do modelo, o qual é assumido como gerador do processo. Os valores de Y são então obtidos adicionando-se desvios normais de variância τ^2. A função de Matérn foi usada neste ponto para definir os elementos da matriz Σ. Após este passo, foi aplicada a decomposição de Choleski para obter A e consequentemente as amostras do processo espacial. As superfícies geradas foram sempre em um quadrado unitário em um gride de 100 x 100 pontos. Para comparação dos modelos foi utilizado o seguinte algoritmo.

1. Gerar uma realização do processo, usando a função de correlação de Matérn com os seguintes conjuntos de parâmetros:

- Conjunto 1 - $(\beta = 50, \sigma^2 = 0{,}9, \tau^2 = 0{,}1, \varphi = 0{,}25, \kappa = 0{,}5/1{,}5/2{,}5)$,
- Conjunto 2 - $(\beta = 50, \sigma^2 = 0{,}75, \tau^2 = 0{,}25, \varphi = 0{,}25, \kappa = 0{,}5/1{,}5/2{,}5)$,
- Conjunto 3 - $(\beta = 50, \sigma^2 = 0{,}5, \tau^2 = 0{,}5, \varphi = 0{,}25, \kappa = 0{,}5/1{,}5/2{,}5)$,
- Conjunto 4 - $(\beta = 50, \sigma^2 = 0{,}25, \tau^2 = 0{,}75, \varphi = 0{,}25, \kappa = 0{,}5/1{,}5/2{,}5)$;

2. Retirar amostras aleatórias de tamanho $n = 150$, $n = 250$ e $n = 500$;
3. Para cada tamanho de amostra ajustar os dois modelos;
4. Usar os modelos ajustados, para recuperar a superfície completa;
5. Calcular o Erro Quadrático Médio de predição, e o nível de cobertura;
6. Repetir o procedimento 1000 vezes.

Todos os procedimentos descritos foram implementados em linguagem R (R Development Core Team, 2010) com a implementação de modelos GAM disponível no pacote mgcv (WOOD, 2008) e modelos geoestatísticos no pacote geoR (RIBEIRO; DIGGLE, 2001). Os códigos de análise estão disponíveis em www.leg.ufpr.br/papercompanions.

É importante notar que as parametrizações escolhidas contemplam diversas características importantes sobre o processo. A primeira é o tamanho da variância proveniente do sinal (σ^2) e a variância do ruído (τ^2). No primeiro esta razão é de 90% do sinal, indicando forte dependência espacial, até apenas 25% no quarto, o que representa uma superfície com baixa dependência espacial, consequentemente um alto nível de ruído, o que deve prejudicar as predições. Além disso, o parâmetro de suavidade do processo κ foi tomado de 0,5 até 2,5, o que indica um processo, não diferenciável, uma e duas vezes diferenciável. Outro aspecto importante considerado é o tamanho da amostra, que deve ter impacto direto nas medidas avaliadas e também no processo de estimação dos modelos.

3 RESULTADOS E DISCUSSÃO

A apresentação dos resultados será feita usando técnicas gráficas. A primeira medida avaliada é o Erro Quadrático Médio de predição. A Figura 1 apresenta os resultados comparando o modelo geoestatístico com o modelo aditivo generalizado, de acordo com o tamanho da amostra e conjunto de parâmetros geradores do processo.

De forma geral, o modelo geoestatístico (GEO) apresenta EQM menores em todas as configurações consideradas. De acordo com o procedimento de avaliação proposto, têm-se três causas conhecidas de variação que afetam o EQM, são elas: tamanho da amostra, proporção entre a variância do sinal e do ruído e suavidade do processo.

Um primeiro resultado evidente nos gráficos e esperado, é a queda do EQM com o aumento da amostra. Este comportamento é consistente para todas as parametrizações e níveis de suavidade considerados. Com relação às diferentes combinações entre variância do sinal e do ruído, os resultados mostram uma clara tendência de aumento do EQM quando se aumenta a quantidade de ruído no processo, por exemplo, o EQM médio com 90% de sinal para n = 150 pelo modelo geoestatístico é de 0,297, passando para 0,843 na parametrização 4 com o mesmo tamanho de amostra, um aumento de 183%.

Para os diferentes níveis de suavidade do processo, verificou-se que quanto mais suave foi o processo melhor foi a reconstituição da superfície, por exemplo, para $\kappa = 0,5$ o EQM médio na parametrização 1 com n = 250 pelo GAM é de 0,334, sob as mesmas condições porém $\kappa = 2,5$ o EQM é de 0,107, uma queda de 212%.

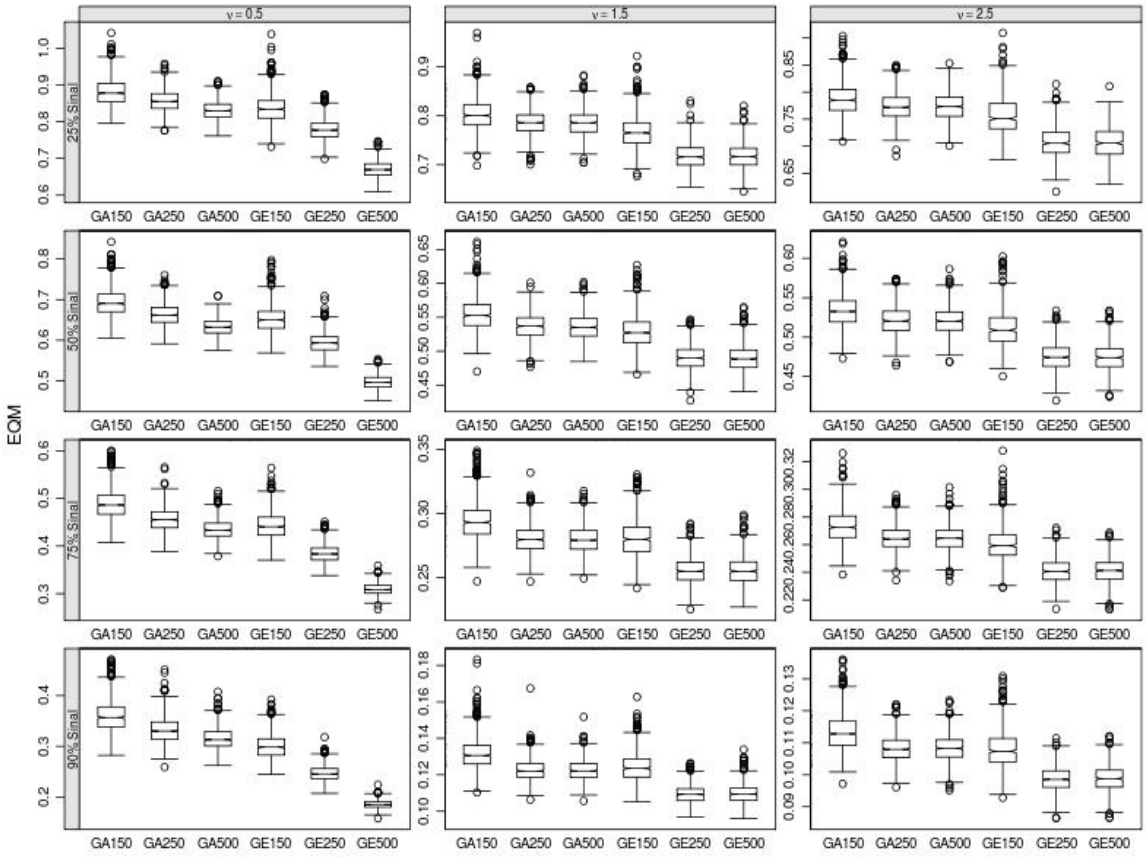

Figura 1 - Comparação do erro quadrático médio por tamanho de amostra, parametrização e modelos.

É interessante observar que o efeito do tamanho da amostra, da proporção entre a variância do sinal e do ruído e da suavidade do processo, afeta os dois modelos de forma muito próxima. Uma outra característica é que os modelos são mais sensíveis ao tamanho da amostra quando o processo é menos suave ($\kappa = 0,5$).

Comparando o EQM entre os dois modelos, o GEO apresentou os menores valores em todas as configurações avaliadas. Sendo que, a diferença foi mais acentuada quando o tamanho da amostra aumenta e o processo foi pouco suave. Por exemplo, com n = 500, $\kappa = 0,5$ na parametrização 1, o EQM médio do GEO foi de 0,186 enquanto que, na mesma condição para o EQM do GAM foi de 0,314, com aumento de 68,82% no erro médio cometido. Sob as mesmas condições, porém com $\kappa = 2,5$, essa

diferença caiu para 9,18%.

O EQM é uma medida geral do ajuste que leva em consideração apenas as predições médias, desprezando a incerteza associada. 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 conteve os valores reais. A figura 2, compara o nível de cobertura pelas duas abordagens consideradas.

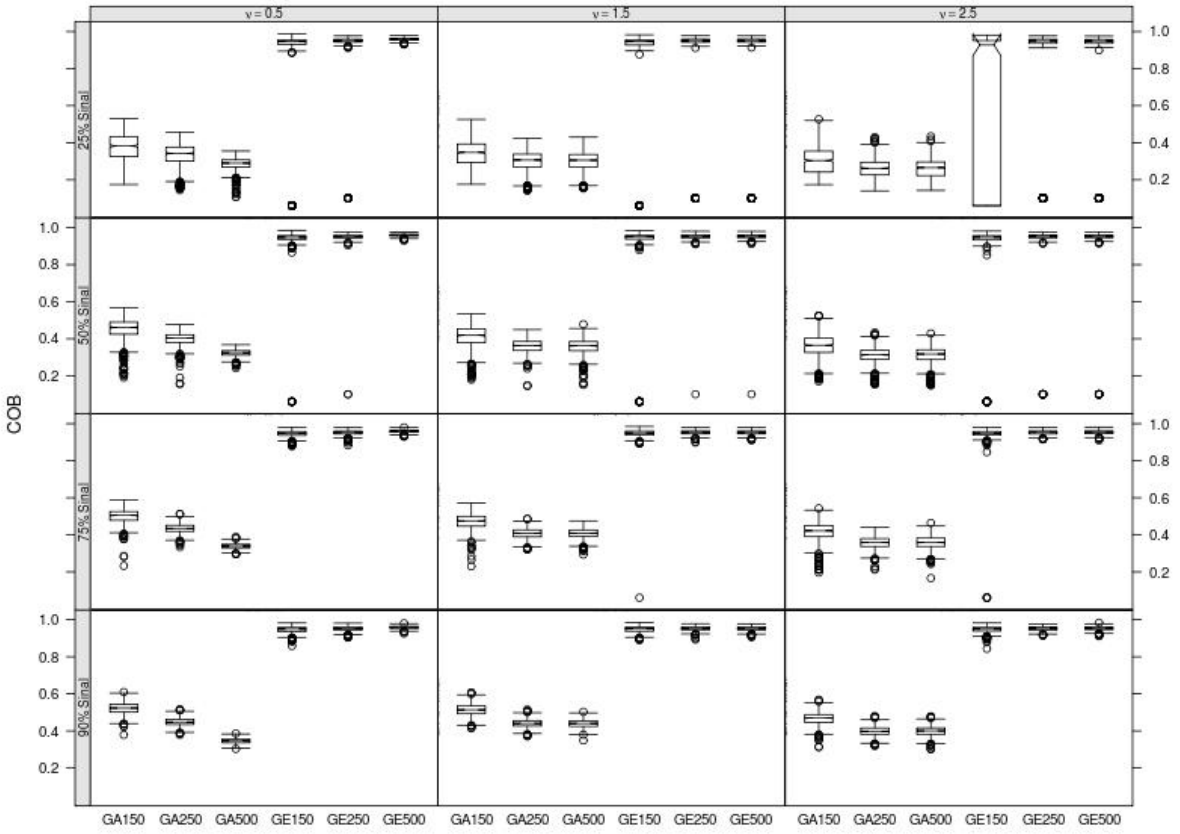

Figura 2 - Comparação do nível de cobertura por tamanho de amostra, parametrização e modelos.

Pelos gráficos apresentados na Figura 2, é claro o mau desempenho do GAM quando avaliado pelo nível de cobertura. Em nenhuma, das condições avaliadas este modelo atingiu o nível nominal de 95%, estando sempre abaixo com coberturas oscilando em torno de 30% a 60%. O modelo geoestatístico apresentou níveis de cobertura muito próximos ao nominal em todas as condições consideradas.

Outro resultado interessante foi o impacto do aumento do tamanho da amostra no nível de cobertura pela abordagem GAM, que tendeu a diminuir com o aumento da amostra. Este comportamento foi consistente em todas as parametrizações utilizadas. Isto não era um resultado esperado, pois o nível de

cobertura não deve depender do tamanho da amostra, uma vez que tamanhos de amostras menores devem refletir em intervalos de maior amplitude, porém com o mesmo nível de cobertura.

É interessante notar que alguns valores discrepantes apareceram pela abordagem geoestatística, em geral quando a proporção sinal ruído foi alta e o tamanho da amostra pequeno (n = 150), isto mostra a dificuldade de identificação deste modelo, sob estas condições. O caso mais patológico ocorreu na parametrização 4, n = 150 e $\kappa = 2,5$, onde o mínimo e o primeiro quartil coincidem com um nível de cobertura de apenas 6%. Este resultado mostra claramente a dificuldade de identificabilidade do modelo geoestatístico nesta situação, sendo necessário um tamanho de amostra maior para estimar adequadamente os parâmetros envolvidos no modelo.

Um fato interessante nesta parametrização é o número de vezes em que o algoritmo falhou (419 vezes em 1000). Novamente mostrando a dificuldade do processo de estimação do modelo geoestatístico nesta situação.

Estes problemas numéricos indicaram que também é interessante comparar o número de vezes que o algoritmo de estimação falhou durante o processo de estimação. A tabela 1 apresenta o número de falhas do algoritmo de estimação do modelo geoestatístico nas primeiras 1000 tentativas de ajuste. Não é apresentado o número de falhas para o modelo GAM, uma vez que este não falhou nenhuma vez durante todo o processo de simulação.

Tabela 1 - Número de falhas do algoritmo de estimação do modelo geoestatístico 1000 tentativas

Parâmetros	$\kappa=0,5$			$\kappa=1,5$			$\kappa=2,5$		
Amostra	150	250	500	150	250	500	150	250	500
90 % sinal	15	31	0	17	55	46	248	199	196
75% sinal	35	11	2	152	102	112	359	303	284
50% sinal	100	52	17	264	225	225	454	420	380
25% sinal	161	147	80	323	317	311	419	462	461

Como mostram os resultados da Tabela 1, o nível de suavidade do processo afeta o sucesso do algoritmo de estimação, sendo mais difícil estimar os parâmetros em processos mais suaves. A razão entre a variância do sinal e do ruído também afeta o ajuste do modelo geoestatistico. Quanto mais ruído no processo mais difícil a estimação. A última fonte de variação é o tamanho da amostra, que como era esperado tem um papel muito importante no processo de estimação.

De forma geral, na situação analisada amostras maiores tenderam a estabilizar o algoritmo numérico, uma vez que, com mais informação a curvatura da função de verossimilhança tendeu a ser melhor identificada pelo algoritmo numérico. Porém, em situações mais gerais uma amostra muito grande,

acima de 2000 observações pode tornar o problema não tratável numericamente, impedindo assim o uso do modelo geoestatístico, na solução usada neste trabalho.

É claro que as falhas encontradas nas primeiras 1000 simulações são facilmente contornadas mudando valores iniciais do algoritmo de otimização, ou algumas vezes retirando observações atípicas. Porém, estes números servem para mostrar que o modelo geoestatístico apesar de intensamente utilizado, ainda apresenta problemas com relação a procedimentos numéricos de maximização da função de verossimilhança, e requer conhecimentos do analista, mesmo que iniciais, de otimização numérica, ao passo que o modelo aditivo generalizado é um procedimento totalmente automático e robusto.

4 CONCLUSÕES

As comparações de erro quadrático médio de predição e nível de cobertura permitem concluir para as condições deste trabalho que o modelo geoestatístico é superior ao *thin plate splines* para a reconstituição de superficies gaussianas, levando em consideração diferente aspectos que afetam a variabilidade dos erros de predição, como tamanho da amostra, razão entre a variância do sinal e do ruído, além do nível de suavidade do processo. Embora a diferença de desempenho de EQM seja pequena em termos absolutos, e no geral as predições fornecidas pelo modelo com *thin plate splines* acompanhem as do modelo geoestatístico, a modelagem por GAM mostrou níveis de cobertura, que combinam predições e incertezas a elas associadas, muito abaixo do nível nominal.

Desta forma, sugeriu-se o modelo *thin plate splines,* como uma opção razoável ao modelo geoestatístico, apenas em estágios iniciais e exploratórios de análise, ou quando não se está interessado na medida de incerteza das predições espaciais ou que necessitem de modelagem mais simples evitando a especificação e estimação de funções de correlação e de computação mais simplificada. Como exemplo pode-se citar sistemas de vigilância, tais como o de vigilância entomológica descrito em Regis et al (2008) e Bonat, et. al. (2009), nos quais são necessários procedimentos totalmente automatizados e robustos, aplicados em grandes conjuntos de dados e com atualizações constantes das análises e resultados. Nestes casos o uso de modelos aditivos generalizados parece ser uma abordagem promissora adequada as implementações totalmente automatizadas. Sugere-se a adoção do modelo geoestatístico para uma análise mais refinada em que houver a necessidade e possibilidade de maior intervenção do investigador na modelagem.

Deixa-se como futuras agendas de pesquisa a investigação das causas do mal desempenho da *thin plate splines,* na medida de nível de cobertura, bem como, propor novas formas de construção de intervalos de confiança para funções suaves bidimensionais, que sejam capazes de descrever melhor a incerteza associada as predições.

5 REFERÊNCIAS

BONAT, W. et. al. RDengue um ambiente para o monitoramento de ovos do mosquito *Aedes aegypti*. In: X SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA: Rio de Janeiro p. 163-170, 2008.

DIGGLE, P. ; RIBEIRO Jr. P. J. **Model-based geostatistics**. New York: Springer, 2007.

HASTIE, T. J. ; TIBSHIRANI, R.J. **Generalized Additive Models.** London: Chapman & Hall, 1990.

MATÉRN, B. **Spatial Variation**. Berlin: Springer, 1986.

McCULLAGH, P. ; NELDER, J. A. **Generalized Linear Models.** London: Chapman & Hall, 1989.
R Development Core Team. *R: A language and Environment for Statistical Computing.* Vienna, Austria, 2009. Disponível em: <http://www.r-project.or>. Acesso em: 29 dez. 2010.

REGIS, L. ; MONTEIRO, A.M. ; MELO-SANTOS, M. ; SILVEIRA, J. ;FURTADO, A. F. ; ACIOLI, R. ; SANTOS, G. ; NAKAZAWA, M. ; CARVALHO, M. S. ; RIBEIRO Jr, P. J. ; SOUZA, W. Developing new approaches for detecting and preventing Aedes aegypti population outbreaks: basis for surveillance, alert and control system. **Memórias do Instituto Oswaldo Cruz**, v. 103, p. 50-59, 2008.

RIBEIRO Jr., P. J. ; DIGGLE, P. geoR: A package for geostatistical analysis. **R-NEWS**, p.15-18, 2001.
WOOD, S. N. **Generalized Additive Models: An Introduction with R.** Boca Raton: Chapman & Hall, 2006.

WOOD, S. N. Thin-plate regression splines. **Journal of the Royal Statistical Society (B)**, p. 95-114, 2003.

WOOD, S. N. **Gams with gcv smoothness estimation and gamms by reml/pql. R package version,** 2008**.** Disponível em: <http://cran.r-project.org/web/packages/mgcv>. Acesso em: 29 dez. 2010.