Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 21-26 abril 2007, INPE, p. 2955-2962.

Simulação conjunta de variáveis correlacionadas para aplicação em modelagem espacial

Jussara de Oliveira Ortiz [1]
Carlos Alberto Felgueiras [1]
Camilo Daleles Rennó [1]

[1] Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12245-970 - São José dos Campos - SP, Brasil
{jussara, carlos, camilo}@dpi.inpe.br

Abstract. This work presents a study case about joint simulation of correlated variables. The procedure is interesting when in spatial modeling is necessary to evaluate uncertainty propagation. The simulation of several correlated variables have been replaced to the simulation of several independent factors. These factors have been gotten by principal components. The standart deviation from estimated values, by sequential indicator simulation, is used as a metric of uncertainty. The results show that the uncertainty propagation is lower with the purposed case and then the estimated values from spatial modeling are more realistic.

Palavras-chave: indicator simulation, uncertainty, correlated variables, principal components.

1. Introdução

Estudos exploratórios e qualitativos, que envolvem as atividades humanas e os processos físicos e químicos que ocorrem na natureza, são importantes para entender os diversos problemas ambientais que vêm ocorrendo nas últimas décadas. No entanto, segundo Myers (1997), há que se considerar, também, que as ações para detecção, monitoramento, avaliação e conseqüente tomada de decisão em um determinado problema, estão baseadas na quantificação da informação disponível e nas incertezas envolvidas. Esse fato gera a necessidade de desenvolver procedimentos que proporcionem avaliações confiáveis, incluindo a quantificação das incertezas inerentes ao próprio processo de decisão.

O procedimento de simulação geoestatística não linear possibilita a construção de cenários associados a modelos que utilizam dados espaciais. Em modelagem espacial, quando se deseja avaliar a propagação de incertezas do modelo final, uma solução apropriada, segundo Heuvelink (1999), é utilizar a simulação conjunta de variáveis. As realizações geradas por simulação serão utilizadas como entrada no modelo e consequentemente as incertezas de cada variável serão propagadas para o resultado.

O objetivo deste trabalho é explorar um procedimento para simulação geoestatística de variáveis correlacionadas. A abordagem utiliza a simulação sequencial por indicação sobre fatores independentes, resultantes da decorrelação de duas variáveis, através do método de principais componentes. Um estudo de caso com duas propriedades químicas do solo, CTC (Capacidade de Troca Catiônica) e V% (Soma de Bases).

2. Aspectos Conceituais

2.1. Simulação

A simulação é uma abordagem adotada, em geral, para resolver problemas nos quais não se dispõe de solução analítica. Em um contexto mais amplo, a simulação refere-se à construção de modelos de qualquer natureza (físicos, matemáticos, sistemas produtivos e de distribuição) e simulação de cenários para o apoio à tomada de decisão. Em termos mais práticos, a simulação consiste na construção de um modelo de um sistema real (ou

ainda por existir) e, através do uso do computador, torna possível a realização de experimentos com vários cenários deste modelo (Saliby, 2001).

No entanto, para se construir e empregar modelos para prever e explicar fenômenos com precisão é necessária uma etapa de seleção cuidadosa das variáveis mais significativas que vão descrever o comportamento do sistema. A razão é que embora seja necessário um grande número de variáveis e ou parâmetros para prever um fenômeno com exatidão, um pequeno número de variáveis, em geral, explica grande parte dele. Deste modo, o ponto inicial de um processo de modelagem consiste em identificar as variáveis certas e as relações entre elas (Gavira, 2003).

O modelo é, portanto, gerado a partir de uma abstração do problema, onde são envolvidas as variáveis consideradas. O nível de complexidade do modelo é diretamente proporcional à abstração efetuada, ou seja, ao número e a quais características serão utilizadas para a abstração (Lobato, 2000).

A simulação estocástica ou probabilística de um sistema oferece meios para a geração de inúmeras seqüências independentes do fenômeno. Cada sorteio gera uma nova série, diferente, porém, com as mesmas propriedades estatísticas e igualmente prováveis. Com as séries distintas entre si, são obtidos diversos resultados provenientes das simulações, ao invés de um único resultado, como ocorre no processo de estimativa. Assim, torna-se possível que o planejador, por exemplo, tome sua decisão baseado não em um evento isolado, mas na análise probabilística do fenômeno estudado (Fonseca, 2004). A característica principal da simulação é a capacidade de reproduzir a variação dos dados de entrada, tanto no sentido univariado (via histograma) quanto espacialmente (através do variograma ou outro modelo de covariância) (Vann et al.; 2002).

2.1.1. Simulação Geoestatística

Uma variante particular da simulação de Monte Carlo, denominada simulação geoestatística, definida em Deutsch e Journel (1998), também tem sido usada para avaliação de incerteza (Vann et al., 2002; Krivoruchko e Crawford, 2003; Pebesma e Heuvelink, 1999). Seguindo esta abordagem, a entrada do modelo é composta por um conjunto de dados espaciais, por exemplo, medidas de propriedades físicas do solo de uma região de interesse, e a saída, por exemplo, fertilidade do solo, requer representações espaciais detalhadas dos parâmetros medidos, ou amostrados. Tais representações podem ser geradas por estimadores geoestatísticos.

2.1.1.1. Simulação Seqüencial Condicionada

O procedimento geoestatístico de simulação estocástica possibilita a criação de campos aleatórios semelhantes, segundo critérios probabilísticos. Estes campos, realizações de funções aleatórias, são usados para caracterizar distribuições de probabilidades, que modelam as incertezas associadas aos valores de variáveis ou campos aleatórios.

Considerando uma área A e uma variável aleatória (VA) Z na posição geográfica u, o procedimento de simulação seqüencial condicionada é apresentado em Deutsch e Journel (1998) e usa a função de distribuição acumulada condicionada (fdac) para obter valores $z(u)$, da VA Z, em cada posição $u \hat{I} A$. As fdac podem ser estimadas através de algorítmos de krigeagem.

Na distribuição de um conjunto de dados amostrais, um determinado número de cortes K e seus valores de corte z_k, $k=1,...,K$, são definidos. A codificação por indicação, se processa para cada valor de corte z_k, e gera um conjunto amostral por indicação $i(u, z_k)$, do tipo (Druck et al. 2002): $i(u; z_k) = 1$, se $z(u) \leq z_k$ e $i(u; z_k) = 0$, se $z(u) \geq z_k$.

A codificação por indicação é aplicada sobre todo o conjunto amostral criando, para cada valor de corte, um conjunto cujos valores são 0 ou 1. Os K valores de corte são definidos em função do número de amostras e devem ser escolhidos de forma que os K+1 cortes contenham aproximadamente as mesmas freqüências.

Ao utilizar um estimador de krigeagem por indicação, por exemplo, a krigeagem por indicação simples, o procedimento de krigeagem linear simples é aplicado ao conjunto amostral codificado por indicação em $z = z_k$, fornecendo para cada para cada valor de corte, z_k, uma estimativa que é também a melhor estimativa mínima quadrática da esperança condicional da VA $I(u; z_k)$. Utilizando esta propriedade pode-se calcular estimativas dos valores da fdac de Z(u) para vários valores de z_k, pertencentes ao domínio de $Z(u)$. O conjunto dos valores estimados para das fdacs de $Z(u)$, nos valores de corte, é considerado uma aproximação discretizada da fdac real de $Z(u)$. Quanto maior a quantidade de valores de corte melhor é a aproximação (Felgueiras, 1999).

O condicionamento considera os dados amostrais originais e também os valores pré-simulados dentro da vizinhança de u. Esta característica da simulação condicionada faz com que esse procedimento seja mais completo que o da krigeagem no que se refere ao modelo de covariância. As fdac estimadas por krigeagem estão condicionadas apenas às amostras, caracterizando a incerteza estimada a partir dessas, como local. As fdac estimadas por simulação consideram também os valores pré-simulados, permitindo informação sobre a incerteza conjunta das variáveis.

A simulação seqüencial indicadora apresenta os seguintes passos (Ortiz et al. 2004):

(a) Definição de um caminho aleatório para cada nó da grade em cada ponto (u);
(b) Determinação da probabilidade condicional da ocorrência da classe z_k, $\hat{f}\left(u, z_k \middle| (n)\right)$, utilizando a krigeagem ordinária,
(c) Correção dos desvios de ordem (que podem ocorrer);
(d) Inferência das distribuições condicionais de probabilidade acumulada através de $\hat{F}\left(\mathrm{u}, z_k \middle| (n)\right) = \sum_{k=1}^{K} \hat{f}\left(\mathrm{u}, z_k \middle| (n)\right) \quad k = 1,\ldots,K$;
(e) Simulação do valor $z_k^l(u)$ a partir da distribuição acumulada inferida anteriormente;
(f) Adição do novo valor simulado $z_k^l(u)$ ao conjunto amostral de dados utilizados durante a simulação.

O conjunto de realizações pode ser usado na determinação de parâmetros estatísticos da fdac local de uma VA ou da fdac conjunta de uma função aleatória (FA). A partir da fdac torna-se possível definir vários intervalos de probabilidade que podem ser usados para medidas associadas a incertezas (Felgueiras et al., 1999).

2.1.2. Simulação Conjunta de Múltiplas Variáveis

Muitas aplicações em geoprocessamento requerem uma simulação conjunta de atributos interdependentes. Caso contrário corre-se o risco de simular valores em condições irreais. Em modelagem espacial de dados, se ocorre correlação positiva entre as variáveis, por exemplo, altos valores de um atributo associados a altos valores de um segundo atributo, ou o contrário, a amostragem a partir das fdac independentes pode vir a combinar altos valores de um atributo com baixos valores do outro. Se essa combinação não representa a realidade os resultados da modelagem estrão comprometidos.

Quando a dependência espacial entre as variáveis é considerada importante no modelo deve-se considerar a correlação espacial cruzada entre elas e realizar uma

simulação condicional de mais de uma variável ou uma simulação conjunta. De acordo com os autores Deutsch e Journel (1998) e Goovaerts (1997) o problema, neste caso, reside na inferência e modelagem da matriz de covariância cruzada. Por causa dos problemas de implementação, a co-simulação é raramente implementada, porém, algumas aproximações são possíveis:

(a) Substituir a simulação de M variáveis dependentes por M fatores independentes, a partir dos quais as variáveis originais possam ser reconstituídas;

(b) Utilizar a variável mais importante, chamada variável primária, para ser simulada primeiro; então, todas as outras variáveis covariadas (variáveis secundárias) são simuladas por realizações obtidas de distribuições condicionais específicas. As autocorrelações das variáveis secundárias são indiretamente reproduzidas a partir daquela da variável primária. Portanto, esta aproximação da co-simulação não é recomendada quando é importante reproduzir acuradamente as covariâncias de variáveis secundárias que diferem acentuadamente daquela da variável primária.

A maioria dos algoritmos de simulação pode ser generalizado, ao menos em teoria, para simulação conjunta de várias variáveis, ao considerar uma função aleatória vetorial do tipo $Z_M(u) = \{Z_1(u), Z_2(u), ..., Z_M(u)\}$. O problema está, segundo Deutsch e Journel (1998), na inferência e modelagem da matriz de covariância cruzada, $\sum = \left|C_{M,M'}\right| = [Cov\{Z_M(u), Z_{M'}(u)\}]$.

Ao substitur a simulação das M variáveis dependentes, $Z_M(u)$,pela simulação de M fatores independentes, $Y_M(u)$,as variáveis Z originais deverão ser reconstituídas. Assim, se $Y = j(Z)$, então, $Z = j^{-1}(Y)$ e $z_M^l(u) = j^{-1}(y_M^l(u))$. Segue também que $y_M^l(u) = \left[y_1^l(u), ..., y_M^l(u)\right]$ é o conjunto de fatores independentes simulados e $z_M^l(u) = \left[z_1^l(u), ..., z_M^l(u)\right]$ é o conjunto de valores simulados resultante onde a interdependência é garantida pela transformada inversa comum j^{-1}.

Exceto para casos específicos onde fatores Y sejam baseados em relações não lineares evidentes, a maior parte das aplicações pode considerar fatores lineares baseados em uma decomposição ortogonal da matriz de covariância $\sum$ das M variáveis originais $Z_M(u)$. A decomposição da matriz de covariância, segundo Deutsch e Journel (1998) é dada por $\sum_{(h_0)} = \left[Cov\{Z_M(u), Z_{M'}(u+h_0)\}\right]$ e corresponde a um simples vetor de separação h_0. As correlações cruzadas de Z são reproduzidas somente para este vetor específico de separação. A distância zero, h_0=0, é frequentemente escolhida por conveniência, sendo o argumento tal que a correlação cruzada é máxima para a distância h_0=0.

3. Metodologia

3.1. Área de Estudo

A área de estudo pertence à Fazenda Santo André situada no município de Carambeí, Estado do Paraná, Brasil. Trata-se de uma das fazendas monitoradas pela empresa de consultoria e assessoria agrícola IMPAR (http://www.imparag.com.br). A fazenda está adotando o sistema de agricultura de precisão, sendo que a principal atividade agrícola está relacionada ao plantio das culturas de Soja e Milho. A fazenda localiza-se entre as

coordenadas geográficas: 24º 45’ 33” e 24º 48’ 56” de latitude Sul; 50º 12’ 05” e 50º 16’ 41” de longitude Oeste.

Foi realizado um levantamento de solos, pela empresa de consultoria em agricultura de precisão, IMPARag do Praná (IMPAR, 2005), no qual as propriedades químicas do solo da fazenda foram medidas. Para este trabalho foram selecionadas a capacidade de troca catiônica (CTC) e a soma de bases (V%), as quais serão, posteriormente, utilizadas como variáveis de entrada em um modelo para predição das necessidades de calcário da área.

A metodologia proposta é exemplificada espacializando-se as variáveis selecionadas através da simulação seqüencial por indicação (ou não linear). O procedimento considerou que as duas variáveis deveriam ser simuladas em conjunto para atingir a proposta do cálculo das necessidades de calcário. Estas variáveis foram amostradas nas mesmas posições do espaço geográfico e a configuração das amostras é apresentada na Figura 1.

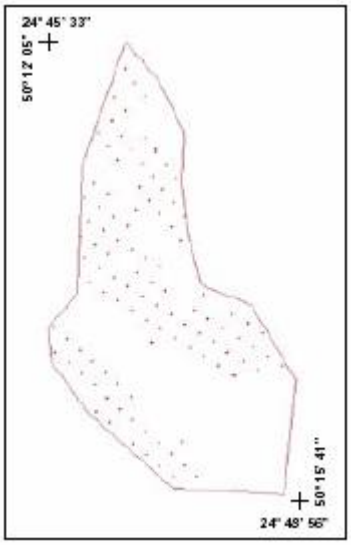


Figura 1- Disposição das amostras de CTC e V% na área de estudo

Ao dispor de algorítmos apenas para a realização de simulação com variáveis independentes, este estudo considerou a técnica de principais componentes.

A simulação foi, portanto, efetuada sobre os fatores independentes ou componentes, resultantes da transformação das variáveis originais usando principais componentes seguindo a teoria apresentada no item 2.1.2.

3.2. Resultados e Discussão

Foi efetuada uma análise para identificação de correlação, no caso positiva, entre as variáveis CTC e V% (ρ=0,57), indicando que o procedimento para decorrelacionar as variáveis poderia ser aplicado.

A técnica de principais componentes foi aplicada sobre os dois conjuntos de dados, conforme metodologia apresentada em Richards (1986), e o resultado gerado corresponde a dois novos conjuntos de dados, denominados também fatores, os quais são totalmente decorrelacionados.

No espaço das principais componentes, o conjunto de pontos de cada fator foi espacializado utilizando a simulação seqüencial por indicação. Novos conjuntos de dados foram definidos nessa etapa, pois os valores foram transformados em binários, conforme apresentado em Deutsch e Journel (1998). Neste caso, três valores de corte

foram definidos em quantis (25%, 50% e 75%) e, portanto, três conjuntos binários, para cada fator, foram gerados. Sobre esses conjuntos binários foi efetuada a análise de variografia.

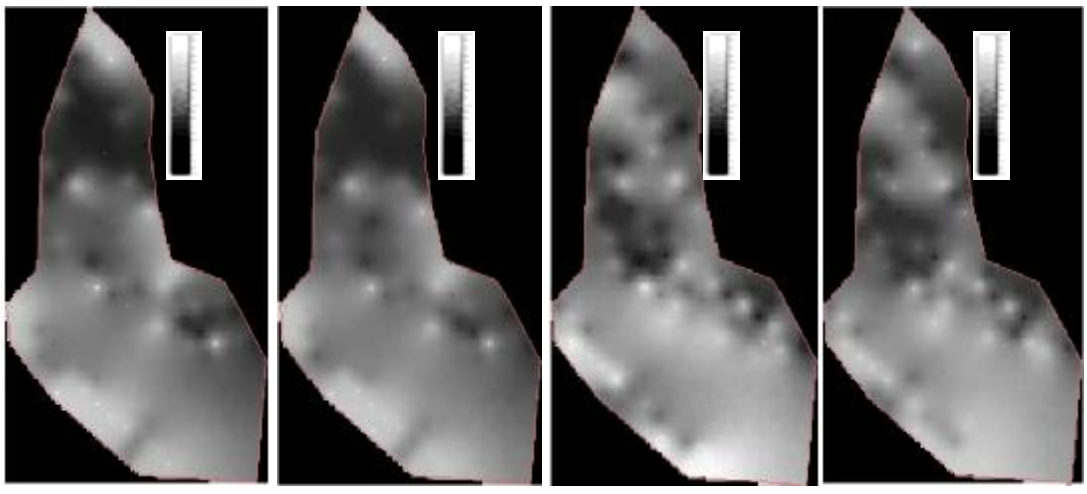
Nessa análise de variografia, foram definidas as estruturas de correlação que representam a variação espacial dos fatores. No caso da abordagem por indicação, o padrão de continuidade espacial pode ser definido para cada corte, o que torna a técnica mais potente que outras que utilizam uma única estrutura para descrever a continuidade espacial do fenômeno. Os modelos experimentais utilizados nesse estudo foram o exponencial e o esférico.

Segundo a abordagem descrita no item 2.1.2, este trabalho adotou a premissa de que em h_0=0, a correlação cruzada é considerada máxima. Cabe ressaltar, que nesse experimento, os dados foram amostrados na mesma posição espacial.

A simulação seqüencial indicadora, conforme apresentada no item 2.1.1, utilizou-se das fdacs para realização de valores no campo das principais componentes, totalizando 400 realizações para cada fator.

A transformação inversa foi aplicada sobre os fatores simulados, garantindo que a correlação necessária entre as variáveis estivesse presente nos procedimentos de modelagem que exijam tal condição.

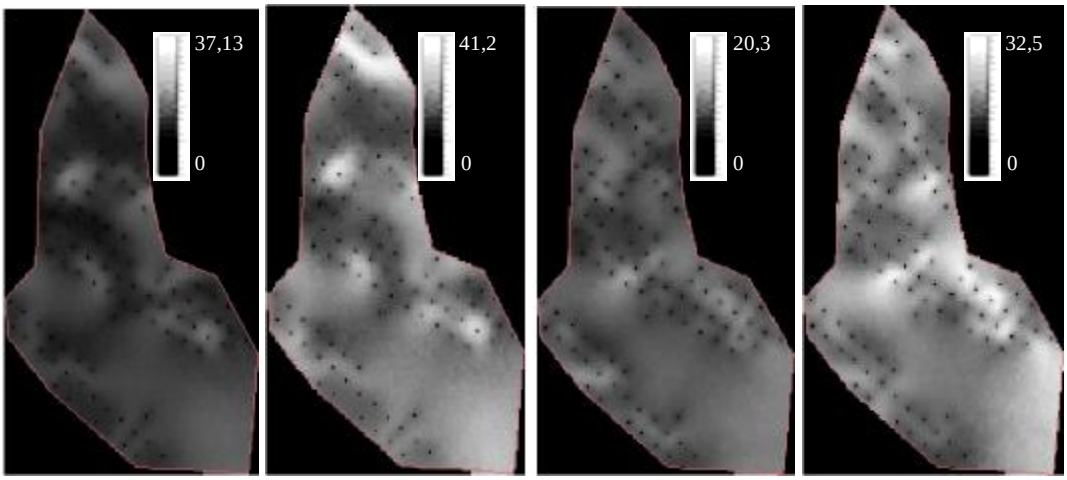
As estimativas para as variáveis CTC e V% foram obtidas utilizando o operador de média. A abordagem por indicação também gera os valores de incerteza (desvio padrão) associados às estimativas. Para efeito de comparação, o mesmo procedimento de simulação por indicação foi realizado sem decorrelacionar as variáveis. Observa-se nas Figuras 2a e 2b que não há diferenças visualmente significativas entre as imagens, as quais correspondentem à média dos valores simulados para a variável CTC, decorrelacionando os dados e não aplicando a decorrelação, respectivamente. O mesmo ocorreu para a variável V%, cujas imagens apresentadas na Figura 2c e 2d também não apresentaram, visualmente, mudanças significativas. Nessas imagens, os níveis de cinza mais claros correspondem aos valores de média mais altos e os níveis de cinza mais escuros, aos valores de média mais baixos.



(a) **(b)** **(c)** **(d)**

Figura 2 – Média resultante da simulação: a) CTC considerando decorrelação; b) CTC sem considerar decorrelação; c)V% considerando decorrelação; d) V% sem considerar decorrelação.

Observa-se, no entanto, nas Figuras 3a, 3b, 3c e 3d, que as incertezas sobre os dados simulados sem decorrelação prévia mostram-se maiores, tanto para a variável CTC quanto para variável V%. Os valores de incerteza estão sendo representados pelo desvio padrão e, analogamente às imagens de médias anteriores, os níveis de cinza mais escuros são associados aos valores mais baixos; portanto, com incertezas menores sobre as estimativas e vice-versa.



(a) **(b)** **(c)** **(d)**

Figura 3 – Médias dos desvios padrão: a) CTC considerando decorrelação; b)CTC sem considerar decorrelação; c) V% considerando decorrelação; d) V% sem considerar decorrelação.

Esse fato pode ser explicado devido ao próprio processo de simulação, nesse caso, não ter considerado a correlação entre as variáveis. A simulação deveria ter sido realizada de forma condicionada, de modo que, a estimativa de uma das variáveis seja obtida levando-se em consideração que a outra variável entra no processo de estimativa como auxiliar. A correlação existente entre as variáveis favorece tal procedimento.

Ao decorrelacionar as variáveis, no entanto, a simulação pode ser realizada de forma independente. No caso desse estudo, a relação entre as variáveis é considerada linear e propicia utilização da técnica de principais componentes. Posteriormente, a transformada linear inversa garantiu que a correlação entre as variáveis fosse restabelecida.

4. Conclusões

A proposta metodológica desse trabalho apresenta uma forma simples de efetuar a simulação de variáveis correlacionadas que tenham sido amostradas na mesma posição espacial, quando não se dispõe de ferramenta para realização de simulação conjunta.

A aplicação do procedimento proposto mostrou que é importante considerar adequadamente a correlação entre variáveis quando o objetivo é a modelagem espacial em geoprocessamento. Ao constatar que os desvios padrão das estimativas tornam-se menores, evita-se que incertezas maiores sejam propagadas nos valores finais do modelo, gerando superestimativas ou subestimativas a partir do modelo.

Ao dispor de estimativas com menor desvio padrão estas tornam-se mais próximas da realidade, capacitando o usuário de geoprocessamento a tomar uma melhor decisão.

5. Referências Bibliográficas

Deutsch, C.V.; Journel, A.G. GSLIB: geostatistical software library and user's guide. New York: Oxford University Press, 1998. 369p, 1 CD.

Druck, S.; Carvalho, M.S.; Camara, G.; Monteiro, A.M.V. Análise Espacial de Dados Geográficos. Versão revisada em julho, 2002. Disponível em: (http://www.dpi.inpe.br/gilberto/livro/analise).

Felgueiras, C.A. **Modelagem ambiental com tratamento de incertezas em sistemas de informação geográfica: o paradigma geoestatístico por indicação**. 1999. 165p. Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Publicado em <http://www.dpi.inpe.br/teses/carlos/>, 1999.

Fonseca, W. S. **Contribuição da simulação de monte carlo na projeção de cenários para gestão de custos na área de laticínios**. Dissertação de mestrado em Engenharia de Produção. 2005. 137p. Universidade Federal de Itajubá. Itajubá, MG. Dezembro. 2004

Gavira, M.O. **Simulação computacional como uma ferramenta de aquisição de conhecimento**. Dissertação de mestrado em Engenharia de Produção. São Carlos, USP, SP. Março. 2003.

Goovaerts, P. Geostatistics for natural resources evaluation. New York : Oxford University Press, 1997. 483p.

Heuwelink, G. B. M. **Error propagation in environmental modeling with GIS**. Bristol: Taylor and Francis Inc, 1998. 127p.

Impar - Consultoria e Assessoria Agrícola. Acesso: (http://www.imparag.com.br). 2005

Lobato, D.C**. Proposta de um Ambiente de Simulação e Aprendizado Inteligente para RAID**. 2000. 167p. Dissertação de mestrado-USP, São carlos, SP. Abril, 2000.

Ortiz, J.O.; Felgueiras, C.A.; Druck, S.; Monteiro, A.M.V. Modelagem de fertilidade do solo por simulação estocástica com tratamento de incertezas. Pesquisa Agropecuária Brasileira-**PAB**, Brasília, v.39. n.4, p.379-389, 2004.

Richards, J.A. **Remote sensing digital image analysis: an introduction**. Springer Verlag, New York. 1986. 281p.

Saliby, E.; Araújo, Marcos M. S. Cálculo do valor em risco através da simulação de Monte Carlo: uma avaliação de uso de métodos amostrais mais eficientes em portfólios com opções. In: Simposio Brasileiro de Pesquisa operacional, 23., 2001, Campos do Jordão. **Anais**. Rio de Janeiro: Sociedade Brasileira de Pesquisa Operacional, 2001. Disponível em http://www.sobrapo.org.br/simposios/xxxiii/artigos/080-ST280.pdf.

Vann, J.; Bertoli, O.; Jackson, S. An Overview of Geostatistical Simulation for Quantifying Risk. In: Association of Australasia symposium "Quantifying Risk and Error" March, 2002.