CLASSIFICAÇÃO DE ATRIBUTOS ESPACIAIS BASEADA EM INFORMAÇÃO DE INCERTEZAS. UMA METODOLOGIA DE APOIO A DECISÕES

CARLOS ALBERTO FELGUEIRAS[1]
SUZANA DRUCK FUKS[2]
ANTÔNIO MIGUEL VIEIRA MONTEIRO[1]

[1]Instituto Nacional de Pesquisas Espaciais - INPE/DPI
Av. dos Astronautas 1758 Jardim da Granja
CEP 12201-970 São José dos Campos SP Brasil
Fone: (0xx12) 345 6519 Fax: (0xx12) 345 6468
e-mails: carlos@dpi.inpe.br, miguel@dpi.inpe.br

[2]Empresa Brasileira de Agropecuária – EMBRAPA/CPAC
Br 020 Km 18 Rodovia Brasília Fortaleza
Planaltina Distrito Federal Brasil
Fone: (0xx61) 389 1121 Fax: (0xx61) 389 2953
e-mail: drucks@ensam.inra.fr

Abstract. This work explores the use of geostatistical procedures, known as geostatistical indicator approaches, in order to classify categorical information. This work shows that these approaches can be used to model the attribute behavior from a punctual sample set. The geostatistical attribute representation allows one to infer optimal estimates values and to model uncertainties associated to the stochastic model. These uncertainties can be used to qualify the inferences and can also be used to generate constraint classifications, that are important in decision makings evolved in planning activities. The concepts here presented are applied and tested in a case study developed for a sample set of soil texture.

Resumo. Este trabalho explora o uso de procedimentos geoestatísticos por indicação para classificação de atributos espaciais categóricos. Mostra-se, no trabalho, que esses procedimentos são usados para inferências, em localizações espaciais não observadas, a partir de um conjunto amostral puntual do atributo categórico de interesse. Pode-se inferir estimativas ótimas e também valores de incertezas associadas aos modelos que representam o atributo. Essas incertezas, além de qualificar as inferências, podem ser usadas para restringir as regiões classificadas. O trabalho mostra que as classificações com restrição são usadas para apoio a decisões em atividades de planejamento. Os conceitos aqui são ilustrados por um estudo de caso aplicado a um conjunto amostral de classes de textura de solo.

INTRODUÇÃO

Modelagem de aplicações geográficas, desenvolvidas em ambiente computacionais e usadas na simulação e estudo de fenômenos ambientais, requerem a modelagem dos atributos espaciais considerados relevantes para uma representação adequada do fenômeno considerado.

Interpoladores, determinísticos e estocásticos, para inferências de atributos espaciais de natureza contínua são comumente encontrados na literatura (Goovaerts, 1995, Camargo, 1997, De Oliveira, 1997, Heuvelink, 1998).

Porém muita das informações amostradas, sobre dados espaciais, são de natureza categórica. Existe uma quantidade enorme de amostras puntuais de atributos categóricos, tais como, amostras de tipos de solo, de classes de vegetação, de tipos de rochas, etc.., que são disponibilizadas, principalmente, a partir de trabalhos de levantamentos em campo. Apesar de importantes, metodologias de espacialização para esse conjunto de informações são raramente abordadas na literatura.

Os procedimentos geoestísticos por indicação, de krigeagem e de simulação estocástica, possibilitam a espacialização, condicionada ao conjunto amostral, de atributos categóricos para geração de mapas classificados. Ademais fornecem informação de incertezas das inferências, que podem ser usadas como restrições de qualidade no processo de classificação. Dessa forma, geram-se produtos, restritos a diferentes níveis de incerteza, que são mais apropriados a um determinado tipo de aplicação em estudo.

Nesse contexto, o objetivo principal deste trabalho é explorar esses procedimentos geoestatísticos por indicação como uma metodologia de modelagem de atributos espaciais categóricos, baseado nas incertezas das inferências, para apoio a decisões em atividades de planejamento, ambientais por exemplo.

O presente trabalho tem a seguinte organização: além desta introdução, a seção 2 apresenta o arcabouço conceitual importante para formalização da metodologia aqui explorada; a seção 3 mostra e analisa os resultados da aplicação dessa metodologia num estudo de caso, desenvolvido sobre dados categóricos, amostrados pontualmente, na fazenda experimental da Empresa Brasileira de Pesquisas Agropecuárias - EMBRAPA – de Canchim e; a seção 4 finaliza o trabalho com um conjunto de conclusões consideradas relevantes para o artigo.

ASPECTOS CONCEITUAIS

Representação de atributos categóricos pela geoestatística

A geoestatística modela os valores de um atributo, dentro de uma região A da superfície terrestre, como uma função aleatória (Isaaks, 1989, Deutsch, 1998). Para cada posição $\mathbf{u} \in \mathbf{A}$ o valor do atributo de um dado espacial é modelado como uma *variável aleatória* (VA) $Z(\mathbf{u})$. Isto significa que, na posição **u,** a VA $Z(\mathbf{u})$ pode assumir diferente valores desse atributo, cada valor com uma probabilidade de ocorrência associada. Nas n posições amostradas, $\mathbf{u}_\alpha$, $\alpha=1,2,...,n$, os valores $z(\mathbf{u}_\alpha)$ são considerados determinísticos, ou ainda, VA's cujo valor medido tem uma probabilidade de 100% de ocorrência. A função de distribuição, para atributos categóricos, de $Z(\mathbf{u})$ condicionada aos n dados amostrado, $F(\mathbf{u}; z|(n))$, *fdc* é definida por:

$$F(u; z \mid (n)) = \text{Prob}\{Z(u) = z \mid (n)\} \quad (1)$$

A $F(\mathbf{u}; z|(n))$ modela a incerteza sobre os valores de $z(\mathbf{u})$, em posições **u** não amostradas, considerando-se as **n** amostras. Esta seção formaliza os conceitos relacionados aos estimadores geoestatísticos por indicação, de krigeagem e de simulação estocástica. Também são apresentadas alternativas para o cálculo de incertezas associadas aos atributos inferidos pelos métodos por indicação.

Determinação da fdc de uma VA pela metodologia por indicação

A *fdc* univariada de uma VA categórica pode ser aproximada utilizando-se de uma metodologia geoestatística não paramétrica chamada *krigeagem por indicação*. Essa metodologia requer a transformação das VAs $Z(\mathbf{u})$ em VAs por indicação $I(u; z_k)$ pela seguinte relação:

$$I(\mathbf{u}; z_k) = \begin{cases} 1, & para \;\; Z(\mathbf{u}) = z_k \\ 0 & caso\, contrário \end{cases} \quad (2)$$

onde z_k é um valor de corte pertencente ao domínio do atributo. Neste caso os valores de corte são as classes, do domínio do atributo categórico, consideradas importantes para o contexto da modelagem.

O valor esperado da VA por indicação, $E\{ I(\mathbf{u}; z_k)|(n)$, fornece uma estimativa F^* da *fdc* de $Z(\mathbf{u})$, no valor de corte z_k e condicionado aos n dados amostrais, do atributo $z(\mathbf{u}_\alpha)$.

$$\begin{aligned} &E\{I(\mathbf{u}; z_k) \mid (n))\} = \\ &1 \cdot Prob\{I(\mathbf{u}; z_k) = 1 \mid (n)\} + 0 \cdot Prob\{I(\mathbf{u}; z_k) = 0 \mid (n)\} = \\ &1 \cdot Prob\{I(\mathbf{u}; z_k) = 1 \mid (n)\} = F^*(\mathbf{u}; z_k \mid (n)) \end{aligned} \tag{3}$$

Essa estimativa, quando realizada por uma krigeagem ordinária sobre o conjunto de valores por indicação, fornece uma inferência por regressão de mínimos quadrados para a *fdc* de $Z(\mathbf{u})$ no valor de corte z_k (Deutsch, 1998). Um conjunto de estimativas F^* em diferentes valores de corte leva a uma aproximação da *fdc* completa de $Z(\mathbf{u})$.

Inferências a partir da fdc de um atributo categórico

A *fdc* aproximada pode ser utilizada para inferências de estimativas e de incertezas relacionadas com o atributo categórico em questão. Tipicamente, a estimativa ótima $z^*(\mathbf{u})$, de um atributo categórico representado por $\mathbf{L}$ classes c_k k=1,...,$\mathbf{L}$, é determinada por:

$$z^*(\mathbf{u}) = c_j \quad sse \quad p_j(\mathbf{u}) > p_i(\mathbf{u}) \quad \forall i, j = 1,...,\mathbf{L} \tag{4}$$

onde $p_l(\mathbf{u}) = F^*(\mathbf{u}; z_k|(n))$, quando $z_k = c_l$, é o valor estimado da probabilidade da ocorrência da classe l na localização $\mathbf{u}$. Este classificador é conhecido como *classificador de moda*, ou estimador de moda, por ser o estimador que considera a maior probabilidade da $fdc(\mathbf{u})$.

Como inferência de incertezas, $Inc(\mathbf{u})$, para atributos categóricos costuma-se utilizar o complemento da probabilidade da moda da *fdc* que é definido por:

$$Inc(\mathbf{u}) = 1 - \underset{j=1}{\overset{\mathbf{L}}{Max}} \left[p_j(\mathbf{u}) \right] \tag{5}$$

O valor de incerteza em $\mathbf{u}$ pode, ainda, ser estimado pela entropia de Shannon (Shannon, 1949), que é definida como:

$$Inc(\mathbf{u}) = H = -\sum_{j=1}^{\mathrm{L}} p_j(\mathbf{u}) ln\left(p_j(\mathbf{u}) \right) \tag{6}$$

A entropia de Shannon tem a vantagem de considerar todas os valores inferidos da $fdc(\mathbf{u})$ e tem valores máximos onde a confusão entre as classes é maior, ou seja, quando a $fdc(\mathbf{u})$ tende para uma distribuição uniforme.

O estimador de moda puro, como formalizado na equação 4, tem a desvantagem de classificar todas a localizações u, dentro da região de interesse, ainda que a incerteza no ponto classificado seja alta, ou seja, as probabilidades da $fdc(\mathbf{u})$ sejam pequenas ou muito uniformes.

Uma metodologia alternativa consiste em se considerar a incerteza como restrição no processo de classificação. Nesse caso são classificadas apenas as localizações com um nível máximo de incerteza preestabelecido. Assim, consideram-se como classificadas apenas aquelas localizações $\mathbf{u}$ em que a incerteza é menor que um limiar *Imax*, ou seja:

$$z^*(\mathbf{u}) = c_j \quad sse \left(p_j(\mathbf{u}) > p_i(\mathbf{u}) \wedge Inc(\mathbf{u}) < Im\,ax \right) \ \forall i, j = 1,...,\mathbf{L}$$
$$z^*(\mathbf{u}) = f \quad caso\ contrário \tag{7}$$

onde $z^*(\mathbf{u}) = f$ significa valor não estimado ou localização não classificada.

Os produtos, mapas ou relatórios, gerados por esta metodologia discriminam regiões espaciais classificadas e não classificadas, segundo o modelo de incerteza e o valor de *Imax* adotado. Esses produtos se adequam melhor a procedimentos de tomadas de decisão uma vez que o risco que se pode assumir, na atividade de planejamento, está incorporado no classificador.

ESTUDO DE CASO

Caracterização do conjunto amostral categórico

A Figura 1 apresenta a distribuição de um conjunto amostral de classes de textura de solo utilizado para estudo de caso deste trabalho. As amostras foram coletadas numa fazenda experimental da EMBRAPA (Empresa Brasileira de Pesquisas Agropecuárias do Brasil) chamada Canchim. A fazenda Canchim situa-se no município de São Carlos, no estado de São Paulo, Brasil, envolvida pelas coordenadas s 21^o 55’ 00’’ a s 21^o 59’ 00’’ e o 47^o 48’ 00’’ a o 41^o 52’ 00’’.

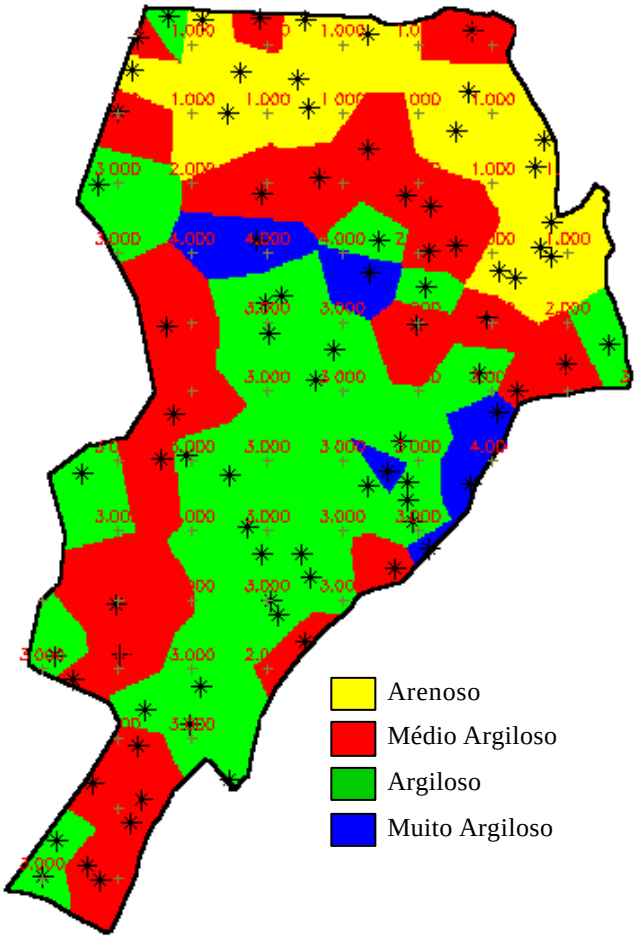

Figura 1: Distribuição das amostras de textura de solo na região de Canchim

Na Figura 1, as amostras estão superpostas a um mapa (grade regular) de textura de solo obtido por estimativas pelo valor da amostra vizinha mais próxima. Esse mapa mostra a região de influência de cada classe e serve de referência primária para conhecimento da distribuição espacial do atributo em estudo.

Modelagem e classificação da textura do solo

Modelos de variografia foram gerados para cada um dos conjuntos por indicação utilizando-se os procedimentos de análise exploratória, principalmente os procedimentos de geração de variograma de superfície e por indicação, do módulo de geoestatística do Sistema de Informação Geográfica SPRING (SPRING, 2000). Por esse módulo, foram definidos variogramas experimentais, e ajustados por modelos matemáticos, para cada uma das classes de textura de solo consideradas: Arenoso, Médio Argiloso, Argiloso e Muito Argiloso.

Sobre os dados amostrais, em conjunto com os variogramas ajustados para cada classe de textura, foi aplicado o procedimento de krigeagem por indicação. Obteve-se, assim, uma aproximação das *fdcs* das VAs com localizações determinadas por uma grade regular retangular, de 200 linhas por 200 colunas, envolvendo a região de Canchim. Essas *fdcs* foram, então, consideradas na geração dos mapas classificados e mapas de incerteza apresentados a seguir.

A Figura 2 abaixo mostra em (a) o mapa de textura classificado pelo estimador de moda definido segundo a equação 4. O mapa de incerteza mostrado na Figura 2 (b) foi gerado pela aplicação da metodologia formalizada na equação 5.

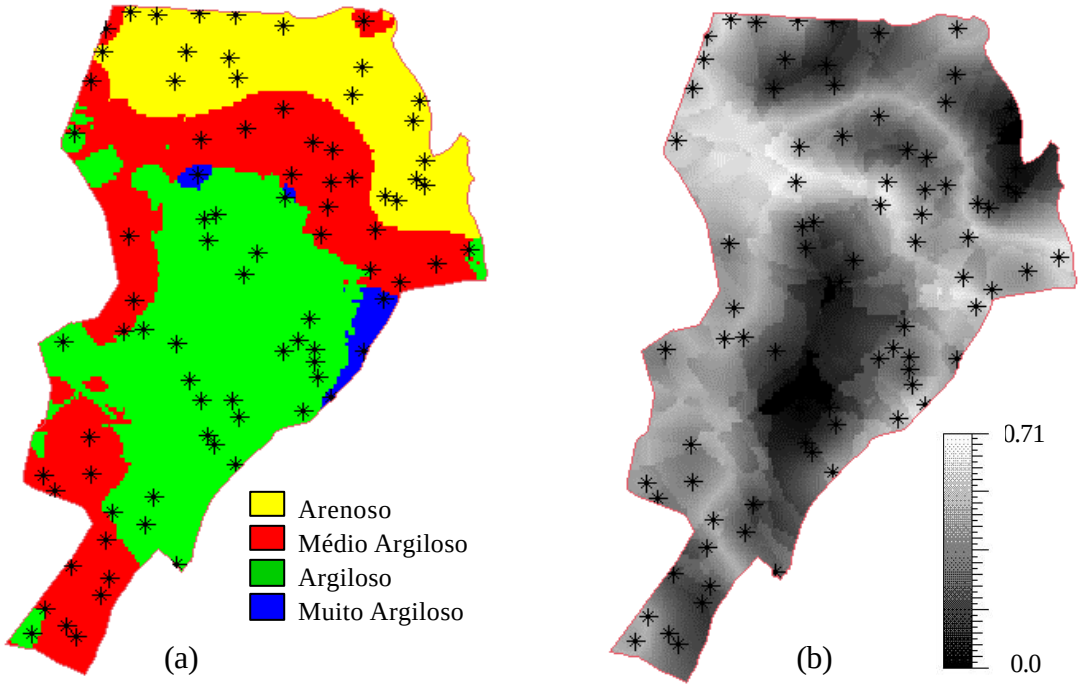

Figura 2: (a) Classificação pela moda e (b) Incerteza associada à classificação

A Figura 3 ilustra o uso do classificador que considera restrições quanto ao nível de incerteza admitido pelo planejador da aplicação. Na Figura 3(a) as classes de textura de solo foram determinadas a partir de um nível de incerteza máximo de 0.38. Por outro lado, o mapa da Figura 3(b) foi construído admitindo-se um nível máximo de incerteza igual a 0.50.

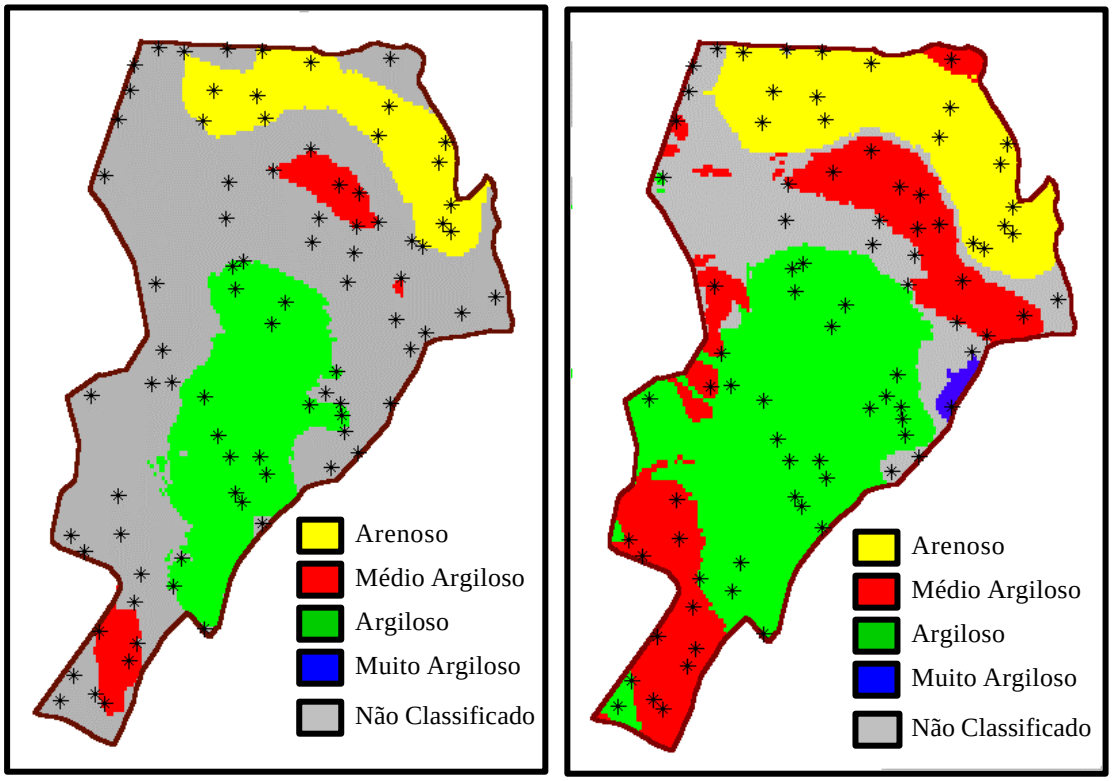

Figura 3: Classificações com nível máximo de incerteza igual a (a) 0.38 e (b) 0.50

Análise dos resultados

A partir de uma análise qualitativa, visual, dos mapas das Figura 1, 2 e 3 pode-se tecer as seguintes considerações:

1. Os mapas classificados apresentados nas Figuras 2 e 3 tem consistência com os valores e com a distribuição das amostras de textura, mostradas na Figura 1. Disso conclui-se que a krigeagem por indicação é um procedimento que pode ser aplicado a atributos categóricos com resultados coerentes.

2. O mapa de incertezas, apresentado na Figura2, tem uma conformação que está de acordo com a variabilidade do atributo em questão. Neste, as maiores incertezas aparecem nas regiões de maior variabilidade, e também nas faixas de transição, dos valores do atributo. Por outro lado, os mapas de incertezas apresentam valores menores em regiões mais homogêneas, onde as amostras de textura são todas de uma única classe.

3. Os mapas classificados da Figura 3 mostram que se pode incluir os dados de incerteza na classificação, como formalizado pela equação 7. Observa-se, nesses mapas, que ao aumentar-se o limiar de incerteza admitido na classificação, obtém-se maiores regiões classificadas. Por outro lado, as regiões classificadas diminuem para um valor menor de incerteza admitido no processo. Assim o planejador tem dois cenários possíveis, baseados em níveis distintos de incertezas, para tomada de decisões referentes a sua aplicação.

CONCLUSÕES

Apresentou-se uma metodologia de modelagem de atributos categóricos a partir de um conjunto amostral puntual. Essa metodologia está baseada em procedimentos geoestatísticos não lineares, por indicação, e apresentam com vantagens principais: partem da premissa que a variabilidade espacial do atributo em questão deve ser modelada a priori e utilizam essa informação nos processos de inferência e; possibilitam a inferência do modelo probabilístico do atributo em qualquer posição espacial de interesse. O modelo probabilístico é usado para:

- espacialização de atributos categóricos, materializados em campos, ou mapas classificados, o que não é possível ser realizado por procedimentos determinísticos ou geoestatísticos lineares;
- Permitem a inferência de informação de incertezas que qualificam as inferências e podem ser utilizadas para apoiar decisões em planejamento de aplicações em que o nível de incerteza da espacialização é considerado uma item importante.

A classificação pode ainda estar baseada em funções de custos, dependentes das incertezas das inferências. Isto esta sendo tema de estudos atuais pelos autores deste trabalho, que pretendem reportar seus resultados em artigos num futuro próximo.

REFERÊNCIAS BIBLIOGRÁFICAS

Burrough P. A. and McDonnell R. A. *Principles of Geographical Information Systems*, Oxford University Press, 1998.

Camargo E. C. G. *Desenvolvimento, implementação e teste de procedimentos geoestatísticos (krigeagem) no Sistema de Processamento de Informações Georeferenciadas (SPRING).* Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 1997.

De Oliveira J. L., Pires F. and Medeiros C. B., "An environment for modeling and design of geographic applications", *GeoInformatica*, 1, (1997), 29-58.

Deutsch C. V. and Journel A. G. *GSLIB Geostatistical Software Library and User's Guide*. Oxford University Press, 1998.

Felgueiras C. A. *Modelagem Ambiental com Tratamento de Incertezas em Sistemas de Informação Geográfica: O Paradigma Geoestatístico por Indicação.* Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Publicado em http://www.dpi.inpe.br/teses/carlos/, 1999.

Goovaerts, P. e Journel, A. G.. Integrating soil map information in modelling the spatial variation of continuous soil properties. **European Journal of Soil Science**, v. 46, n. 3, p. 397-414, 1995.

Goovaerts, P.. **Geostatistics for Natural Resources Evaluation**. New York, Oxford University Press, 1997. 481p.

Heuvelink G. B. M. *Error Propagation in Environmental Modeling with GIS*, Bristol, Taylor and Francis Inc, 1998.

Isaaks E. H. and Srivastava R. M. *An Introduction to Applied Geostatistics*, Oxford University Press, 1989.

Shannon, C. E. e Weaver, W.. **The Mathematical Theory of Communication**. Urbana, The University of Illinois Press, 1949. 117p.

SPRING V.3.4, (DPI/INPE) Sistema de Processamento de Informações Georeferenciadas – Divisão de Processamento de Imagens (DPI) do Instituto Nacional de Pesquisas Espaciais (INPE), http://www.dpi.inpe.br/spring/ , 2000.