Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 21-26 abril 2007, INPE, p. 2841-2848.

Uso de árvore de decisão para predição da prevalência de esquistossomose no Estado de Minas Gerais, Brasil.

Flávia de Toledo Martins[1]
Luciano Vieira Dutra[1]
Corina da Costa Freitas[1]
Fernanda Rodrigues Fonseca[1]
Ricardo José de Paula Souza e Guimarães[2,6]
Ana Clara Mourão Moura[5]
Ronaldo Guilherme Carvalho Scholte[2,6]
Ronaldo Santos Amaral[3]
Sandra Costa Drummond[4]
Charles R. Freitas[5]
Omar dos Santos Carvalho[2]

[1] Instituto Nacional de Pesquisas Espaciais- INPE
Caixa Postal 515 – 12202-970 – São José dos Campos-SP, Brasil
{flavinha,dutra,corina, ffonseca}@dpi.inpe.br

[2] Centro de Pesquisas René Rachou/FIOCRUZ-MG
{omar,ricardo,ronaldo}@cpqrr.fiocruz.br

[3] Secretaria de Vigilância em Saúde/MS
ronaldo.amaral@funasa.gov.br

[4] Secretaria de Estado de Saúde de Minas Gerais
sandra.drummond@saude.gov.br

[5] Laboratório de Geoprocessamento-IGC-UFMG
{anaclara,charlesrf}@ufmg.br

[6] Programa de Pós-Graduação da Santa Casa de Misericórdia de Belo Horizonte, MG, Brasil

Abstract
The aim of this paper is to determine the behavior pattern within data derived from remote sensing, social- economic variables and climatic variables, through the decision tree analysis by choosing the more appropriate ones to explain the prevalence of schistosomiasis in State of Minas Gerais, Brazil.

Palavras-chave: remote sensing, geographic information systems, schistosomiasis, data mining, sensoriamento remoto, sistemas de informação geográfico, esquistossomose, mineração de dados**.**

1. Introdução

A esquistossomose mansoni no Brasil tem como agente etiológico o trematódeo *Schistosoma mansoni*, cujos hospedeiros intermediários são moluscos límnicos do gênero *Biomphalaria* (*B. straminea*, *B. glabrata* ou *B. tenagophila*) (Doumenge et al., 1987). A doença tem caráter social e

comportamental, sobretudo em decorrência da insuficiência de saneamento domiciliar e ambiental e do baixo nível de educação em saúde da população que vive sob risco.

Visto que a doença é limitada no espaço e tempo por fatores ambientais, os sistemas de informação geográficos (SIG) e o sensoriamento remoto (SR) são úteis para identificar os fatores ambientais que permitem determinar e delimitar a área de risco, permitindo um alocamento efetivo de recursos para o controle da doença. A aplicação de produtos derivados de SR e o uso de SIG, tem sido muito utilizado no estudo da esquistossomose no Brasil e em outros países que também padecem da enfermidade (Bavia et al., 2001; Freitas et al 2006).

A mineração de dados fornece um grande potencial para buscar um padrão no comportamento das variáveis que determinam a ocorrência da esquistossomose, seja na compreensão do comportamento da doença, como no relacionamento dela com outras variáveis explicativas. Por meio de particionamento recursivo das variáveis preditoras, o conhecimento sobre o problema pode ser representado por meio de uma estrutura de árvore de decisão. A árvore de decisão é um método bastante utilizado para determinar o padrão de comportamento de conjuntos de dados em diferentes formas de representações. Entretanto, os classificadores baseados em árvores de decisão não têm sido muito utilizados pela comunidade de sensoriamento remoto (Marcelino, 2003; Araki, 2005).

O objetivo principal desse trabalho é aplicar técnica de mineração de dados (árvore de decisão) para estimar a prevalência da esquistossomose no Estado de Minas Gerais através de variáveis derivadas de SR, climáticas e sócio-econômicas utilizando ferramentas de SIG para dar subsídios às campanhas de controle da doença e conscientização da população pelos órgãos competentes.

2. Materiais e Métodos

A área de estudo é o Estado de Minas Gerais, com uma área de aproximadamente 590.000 km^2, dividido politicamente em 853 municípios, com aproximadamente 18 milhões de habitantes (IBGE, 2006).

2.1 Variáveis utilizadas

Foram utilizadas 197 amostras, de um total de 853 amostras (número de municípios de Minas Gerais), que possuíam dados de prevalência (porcentagem de casos positivos da doença em relação a no mínimo 80% da população do município). Esta amostra foi utilizada para a geração da árvore de decisão. Através da árvore selecionada foi possível extrapolar a estimativa da prevalência da esquistossomose para o restante das amostras.

Os dados de prevalência da doença foram disponibilizados pela Secretaria de Vigilância em Saúde e Secretaria de Estado de Saúde de Minas Gerais. Para explicar a prevalência foram utilizadas quarenta e quatro variáveis, das quais vinte e duas são derivadas de dados de sensoriamento remoto, seis são climáticas, e dezesseis são sócio-econômicas. Das variáveis derivadas de sensoriamento remoto, dezoito foram obtidas através do sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*) em duas épocas, uma no verão (17/01 a 01/02/2002) e outra no inverno (28/07 a 12/08/2002), compostas das bandas azul (BLUE), vermelho (RED), infravermelho próximo (NIR), infravermelho médio (MIR), índice de vegetação melhorada (EVI), índice de vegetação da diferença normalizada (NDVI) e os índices derivados do modelo linear de mistura espectral, vegetação (VEG), solo (SOLO) e sombra (SOMB). As outras quatro variáveis de sensoriamento remoto foram derivadas do SRTM (*Shuttle Radar Topography*

Mission), o modelo digital de elevação (DEM) e a declividade (DEC), derivada do DEM e outras duas com informações hidrográficas, uma obtida através do mapa de acumulação hídrica (que mede em cada ponto de uma bacia hidrográfica, os caminhos possíveis que a água pode transcorrer ao atingir esse determinado ponto) gerado a partir do DEM (Moura et al., 2005), a média de acumulação hídrica (AH1), e a mediana da acumulação hídrica (AH2). Das variáveis climáticas, três são da época de verão e três são de inverno: precipitação acumulada (Pa); média da temperatura mínima (Tmin); e média da temperatura máxima (Tmax), obtidas através da plataforma de coleta de dados do CPTEC/INPE. As variáveis sócio-econômicas foram obtidas do IBGE: os índices de desenvolvimento humano (IDH), educação (IDHE), longevidade (IDHL), renda (IDHR) dos anos de 1991 e 2000; e oito índices de conforto sanitário, porcentagem de domicílios com banheiro ligados a: rio ou lago (SAN1); vala (SAN2); fossa rudimentar (SAN3); fossa séptica (SAN4); rede geral (SAN5); outro tipo de esgotamento (SAN6); e porcentagem de domicílios com banheiro (SAN7) e sem banheiro (SAN8).

2.2 Mineração de dados

As árvores de decisão são um dos modelos mais práticos e mais usados em inferência indutiva. Este método representa funções como regra de decisão. Estas árvores são treinadas de acordo com um conjunto de amostras previamente classificadas e posteriormente, outras amostras são classificadas de acordo com essa mesma árvore. Para a construção destas árvores são usados algoritmos como o ID3, ASSISTANT e C4.5 (CFBioinfo, 2006). O C4.5 não depende de suposições sobre a distribuição dos valores das variáveis ou da independência entre si das variáveis. Isto é importante quando se utiliza dados de SIG juntamente com dados de imagem (Araki, 2005).

Para aplicar a técnica de mineração de dados foi utilizado o software de domínio público, denominado *Waikato Environment for Knowledge Analysis* (*Weka*), da Universidade de Waikato, Nova Zelândia. O pacote *Weka* consiste de uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Pode ser usado para aplicar método(s) de aprendizado a um conjunto de dados e analisar a saída para extrair informações a partir dos dados de entrada (Weka, 2006).

O *Weka* usa arquivos de dados de treinamento onde devem ser explicitadas quais variáveis são permitidas para uma relação específica, bem como o tipo de dado de cada variável (isto é, nominal ou valor numérico). O *Weka* pode detectar padrões em dados que podem ser explorados mediante regras. Dos recursos disponíveis, foi utilizado o sistema de aprendizado com o algoritmo de indução de árvore de decisão C4.5 desenvolvido por Quinlan e implementado em sua versão para linguagem Java (no *Weka*) com o nome J4.8, para gerar árvores de decisão (Weka, 2006).

A árvore de decisão pode ser analisada pelo especialista e, se necessário, pode ser modificada, para então ser convertida em regras que formam a base de conhecimento de um sistema. Cada caminho da raiz até a folha corresponde a uma regra de decisão ou classificação.

Para a utilização do conjunto de dados, foi necessário um pré-processamento nos dados a fim de torná-los compatíveis com o formato da ferramenta utilizada. Além disso, o algoritmo de classificação requer que a variável a ser explicada seja uma variável nominal; dessa forma, foi necessário discretizar os dados de prevalência para transformá-los em variável nominal. Foi utilizada uma regra simples no *Excel*, que classificou os dados em quatro categorias: baixa (0 a 5%); média (>5 a 15%); alta (>15 a 25%); e muito alta (>25%), como mostra a **Figura 1**.

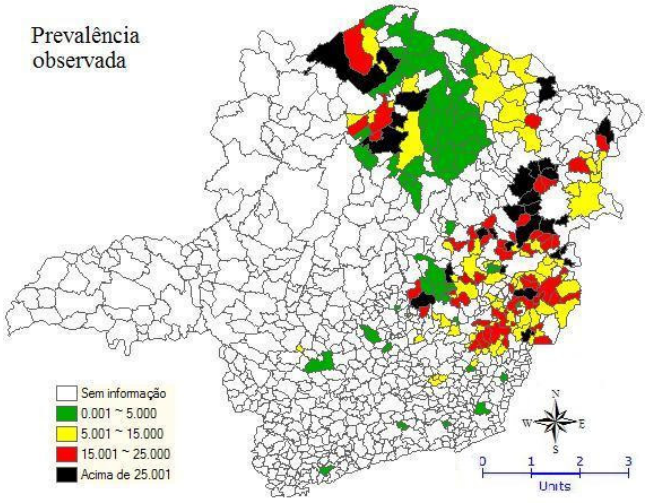

Figura 1 - Prevalência da esquistossomose em 197 municípios de Minas Gerais.

3. Resultados e Discussões

As classificações com pequeno número de amostras foram descartadas da árvore de decisão, visto que nos testes realizados com no mínimo 12 amostras em cada classe a árvore de decisão alcançou 65% de amostras classificadas corretamente. Esse valor baixo na classificação deve-se ao número reduzido de informação, e também por englobar casos de prevalência da esquistossomose que ocorreram desde 1984, além do fato de existirem poucas amostras com prevalência muito alta em relação às outras classes.

Com a classificação, buscou-se verificar as diferenças no padrão de comportamento das variáveis em relação à prevalência da doença. Para avaliação dessas classificações foi utilizada a estatística Kappa que, segundo Congalton e Green (1999), é definida como uma técnica de estatística multivariada discreta usada para a avaliação da precisão, determinada estatisticamente por uma matriz de erro. A classificação gerou um índice Kappa de 51%.

O algoritmo J4.8 fornece regras de decisão e uma matriz de erros (**Tabela 1**). Analisando a matriz de erros pode-se detectar possíveis problemas na classificação e também a separabilidade entre as classes. Na **Tabela 1** observa-se a maior confusão entre as classes de prevalência média e baixa, alta e média e entre as classes de prevalência muito alta e alta.

Tabela 1 – Matriz de erros

Classificada / Observada	Baixa	Média	Alta	Muito alta
Baixa (46)	42	03	01	0
Média (73)	17	47	09	0
Alta (51)	5	14	25	7
Muito alta (27)	2	04	08	13

Pode-se observar também na **Tabela 1** que das 197 amostras, 127 são classificadas corretamente, sendo que das 70 amostras classificadas incorretamente, 58 foram classificadas com um erro de classe, 10 com dois erros e apenas duas com três erros. Para as 46 amostras da classe de prevalência baixa, 91,3% das amostras foram classificadas corretamente e somente 1

amostra (2,2%) foi classificada como prevalência alta. Esse resultado pode ser considerado satisfatório, uma vez que os recursos para o combate a doença são escassos e com essa classificação esses recursos dificilmente serão alocados para áreas com menor prevalência. O mesmo acontece com as amostras com prevalência média, 64,4% delas são classificadas corretamente, e apenas 12% são classificadas como prevalência alta. Já para a classe de prevalência alta e prevalência muito alta foram classificadas corretamente 49% e 48% das amostras, respectivamente, sendo que 5 amostras (9,8%) de prevalência alta foram classificas como baixa e apenas 2 amostras (7,4%) de prevalência muito alta foram classificadas como baixa.

O algoritmo de aprendizagem de máquina determina a variável com maior quantidade de informação e a coloca na raiz da árvore de decisão. Em cada nó da árvore, foi feita a divisão em conjuntos cada vez mais homogêneos. A variável colocada na raiz da árvore foi VEG_I (VEG da época de inverno), correspondendo à divisão em dois grupos: para valores desta variável menores ou iguais a 27,262 as amostras já são classificadas como de baixa prevalência e acima deste valor tem-se outras regras para a classificação.

A **Figura 2** mostra a árvore de decisão (obtida a partir do algoritmo J4.8 do *Weka*) para a prevalência da doença em relação a algumas variáveis preditivas que foram selecionadas pelo *Weka* por conterem maior quantidade de informação. Foram ainda selecionadas outras variáveis tais como variáveis do inverno (Tmax_I), de verão (Tmin_V, BLUE_V, SOMB_V e EVI_V), sócio-econômicas (IDHL,SAN6) além da variável que mede a altitude do terreno (DEM). Algumas dessas variáveis foram selecionadas por Freitas et al (2006) e Guimarães et al (2006), imprimindo mais confiabilidade e demonstrando que a técnica pode ser considerada eficiente para explicar a prevalência da esquistossomose.

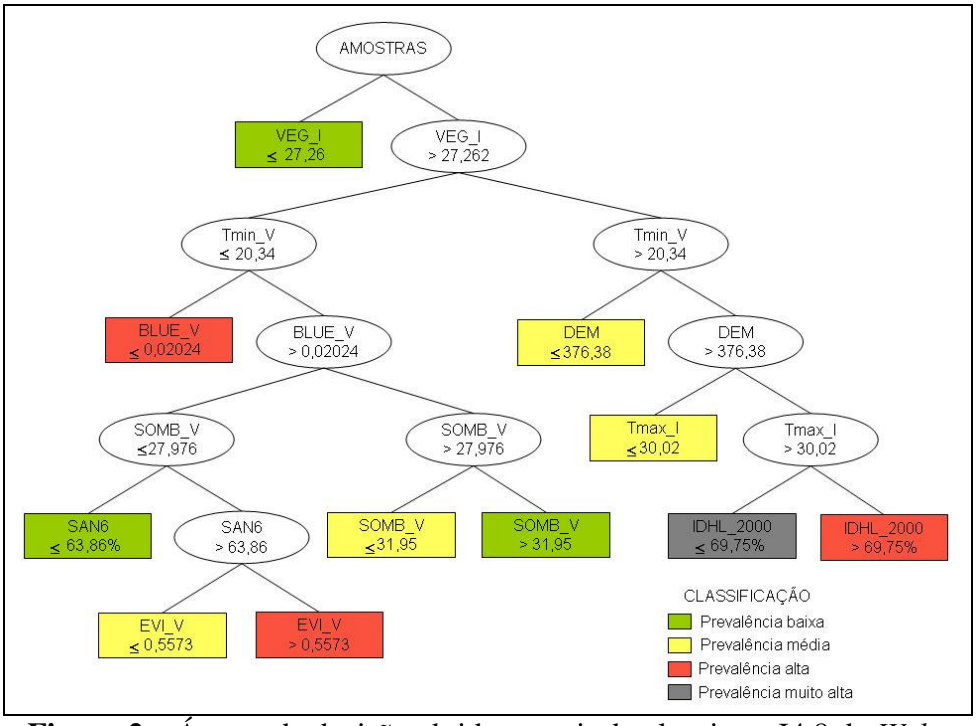

Figura 2 – Árvore de decisão obtida a partir do algoritmo J4.8 do *Weka*

A divisão seguinte da árvore é dada em função da variável Tmin_V. As amostras nesse ramo terão duas possibilidades: menor ou igual a 20,34º C e maior que este valor. Seguindo o ramo da menor temperatura, as amostras serão classificadas como prevalência baixa, média ou alta.

A próxima divisão das amostras é relativa à variável BLUE_V. Quando esta variável tiver valor menor ou igual a 0,02024 as amostras serão classificadas como prevalência alta. Caso contrário, quando o valor de BLUE_V for maior, haverá a divisão da variável SOMB_V, a qual está relacionada à topografia e também à existência de água:

- Quando o valor da variável SOMB_V for menor ou igual a 27,976, e a variável SAN6 apresentar valor menor ou igual a 63,857% as amostras serão classificadas como prevalência baixa. Porém se o valor for maior que este, e se a variável EVI_V for menor ou igual a 0,55728 as amostras serão classificadas como prevalência média, caso contrário, se EVI_V for maior, as amostras serão classificas como prevalência alta.
- Quando o valor da SOMB_V estiver entre 27,976 e 31,949, a prevalência será classificada como média. Se o valor da variável for maior que 31,949 a prevalência será classificada como baixa.

Da mesma forma, seguindo o ramo da maior temperatura as amostras poderão ser classificadas como prevalência média, alta ou muito alta.

Neste ramo, a divisão das amostras se inicia com a variável DEM. Quando o valor for menor ou igual a 376,377 as amostras serão classificadas como prevalência média. A mesma classificação ocorrerá se DEM for maior e se a variável Tmax_I for menor ou igual a 30º C.

Porém se a temperatura for maior que 30º C, as amostras serão classificadas como prevalência alta ou muito alta, se o IDHL_2000 (IDHL do ano de 2000) for maior que 69,75% ou menor ou igual a este valor, respectivamente.

O resultado da classificação pode ser considerado coerente em relação a realidade, o habitat ideal para o caramujo e condições de vida das pessoas são fatores importantes para a existência da doença. Porém em alguns ramos a classificação parece ser subestimada e em outros superestimada. Esse fato pode ser observado na **Figura 3.**

A **Figura 3 (a)** mostra os dados classificados através das regras da árvore de decisão espacializados num mapa temático utilizando o aplicativo *Terra View* e também os erros da classificação **(b)**.

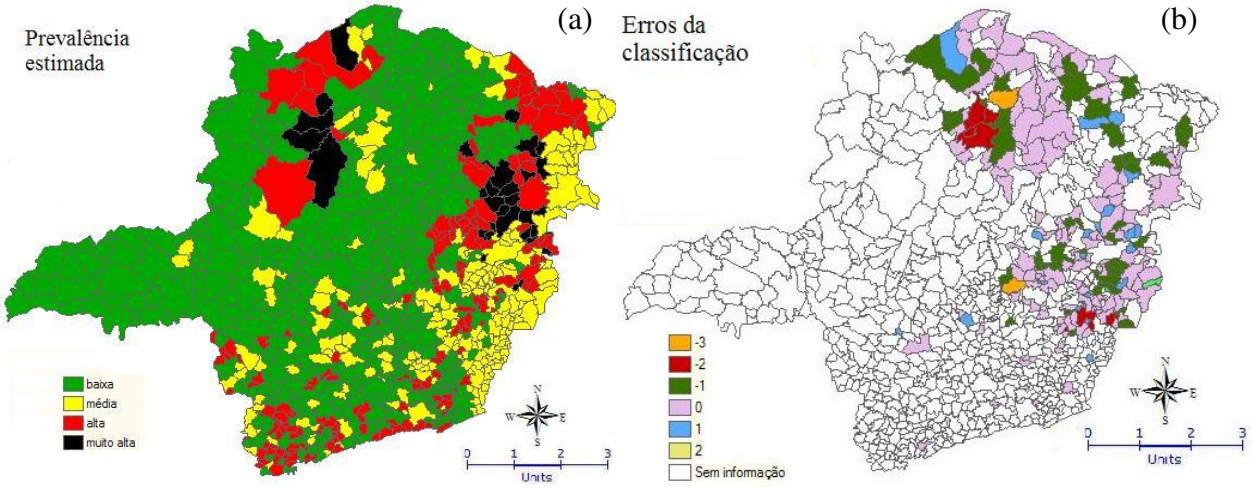

Figura 3 – **(a)**Prevalência da esquistossomose estimada através da árvore de decisão; **(b)** Erros de classificação.

Na **Figura 3 (b)** os municípios representados em laranja, vermelho e verde a classificação é subestimada em relação a 3, 2 e 1 classe observada, respectivamente (por exemplo, os dois municípios em laranja, foram classificados como prevalência baixa, porém pertenciam à classe de prevalência muito alta). Já nos municípios representados em azul e em um único município em amarelo, a classificação é superestimada em 2 e em uma classe respectivamente. Os municípios classificados corretamente estão representados em lilás.

As vantagens de árvores de decisão incluem a capacidade de lidar com dados que estão em diferentes escalas de medidas, não serem necessárias suposições sobre as distribuições de freqüência dos dados em cada uma das classes, a flexibilidade e a capacidade de lidar com relações não lineares entre variáveis e classes. Também, o analista pode interpretar uma árvore de decisão, uma vez que é gerado conhecimento simbólico, diferentemente de outras metodologias de aprendizagem (Pal, Mather, 2003).

4. Agradecimentos

Os autores reconhecem o suporte da CAPES; CNPq (processos 309922/2003-8; 305546/2003-1; 380203/2004-9; 384467/2006-7); Fapemig (processo EDP 1775/03; EDT 61775/03; CRA 0070/04).

5. Referencias Bibliográficas

Araki H. **Fusão de Informações Espectrais, Altimétricas e de dados auxiliares na classificação de Imagens de Alta Resolução Espacial**. 2005. 136 p. Tese (Doutorado em Ciências Geodésicas) Universidade Federal do Paraná, Curitiba, 2005.

Bavia, M. E.; Malone, J. B.; Hale, L.; Dantas, A.; Marroni, L.; Reis, R. Use of thermal and vegetation index data from earth observing satellites to evaluate the risk of schistosomiasis in Bahia, Brazil. **Acta Tropica**, v. 79, n. 1, p. 79-85, 2001

CFBioinfo. Genómica Funcional e Bioinformática. Disponível em: <http://bioinformatics.ath.cx/index.php?id=199>, acesso em outubro, 2006.

Congalton, R. G.; Green, K. **Assessing the accuracy of remotely sensed data**: principles and practices. New York: Lewis Publishers, 1999. 137 p.

Doumenge, J.P., Mott, K.E., Cheung, C., Villenave, D., Capui, O., Perrin, M.F. **Atlas of the Global istribution of Schistosomiasis**. 1987 Talence, Geneva.

Freitas, C. C.; Guimarães, R. J. d. P. S.; Dutra, L. V.; Martins, F. d. T.; Gouvêa, É. J. C.; Santos, R. A. T.; Moura, A. C. d. M.; Drummond, S. C.; Amaral, R. S.; Carvalho, O. d. S. Remote Sensing and Geographic Information Systems for the Study of Schistosomiasis in the State of Minas Gerais, Brazil, In: 2006 International Geoscience and Remote Sensing Symposium, Denver, USA, 31 July- 04 August, 2006 (in press). **Anais**... Denver, USA: IEEE, 2006.

Guimarães, R. J. d. P. S.; Freitas, C. C.; Dutra, L. V.; Moura, A. C. d. M.; Amaral, R. S.; Drummond, S. C.; Scholte, R. G. C.; Freitas, C. R.; Carvalho, O. d. S. Analysis and estimative of schistosomiasis prevalence for Minas Gerais state, Brazil, using multiple regression with social and environmental spatial data. **Mem Inst Oswaldo Cruz**, vol. 101 p. 91-96, 2006.

IBGE. Instituto Brasileiro de Geografia e Estatística. Disponível em: http://www.ibge.gov.br/estadosat/perfil.php?sigla=mg. Acesso em: abril, 2006.

Marcelino, I. P. O. **Análise de episódios de tornados em Santa Catarina: caracterização sinótica e mineração de dados.** 2003. 223p. (INPE-12145-TDI/969). Dissertação (Mestrado em Sensoriamento Remoto) Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 2003

Moura, A. C. M.; Freitas, C. R.; Dutra, L. V.; Melo, G. R.; Carvalho, O. S.; Freitas, C. C.; Amaral, R. S.; Scholte, R. G. C.; Drummond, S. C.; Guimarães, R. J. P. S. Atualização de mapa de drenagem como subsídio para montagem de SIG para a análise da distribuição da esquistossomose em Minas Gerais. Simpósio Brasileiro de Sensoriamento Remoto, 12. (SBSR), 16-21 abr. 2005, Goiânia. **Anais**... São José dos Campos: INPE, 2005. Artigos, p. 3551-3558.

Pal & Mather, 2003 M. Pal and P.M. Mather, An assessment of the effectiveness of decision tree methods for land cover classification, Remote Sensing of Environment 86 (2003), pp. 554–565.

Weka 3 - Data Mning Software in Java The University of Waikato,, disponível em: <http://www.cs.waikato.ac.nz/ml/weka/> ultimo acesso em: junho,2006.