Mapeamento do risco da esquistossomose em Minas Gerais usando k-NN e árvore de decisão

Flávia de Toledo Martins[1]
Luciano Vieira Dutra[1]
Eliana Pantaleão[2]
Sandra Sandri[1]
Corina da Costa Freitas[1]
Ricardo de Souza Paula Guimarães[3]

[1]Instituto Nacional de Pesquisas Espaciais – INPE
Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil
{flavinha, dutra, corina}@dpi.inpe.br, sandri.at.lac.inpe.br@gmail.com

[2]Universidade Federal de Uberlândia – UFU
Campus Avançado de Patos de Minas
epantaleao@ufu.br

[3]Instituto Evandro Chagas – IEC
Rodovia BR-316 km 7 Levilândia, CEP 67030-000 - Ananindeua - PA, Brasil
ricardojpsg@gmail.com

Abstract. Of all the parasitic diseases that affect humans, schistosomiasis is one of the most widespread. Considered a serious public health problem, the disease affects thousands of people in Brazil. Since the implementation of schistosomiasis control program in the state of Minas Gerais, stock control and surveillance have been conducted. To contribute to the control and mapping of endemic areas, the aim of this study is to obtain thematic maps showing the risk factor for schistosomiasis mansoni in Minas Gerais. Schistosomiasis is a disease caused by a worm that uses a snail as intermediary host. The worm uses the water to go from the snail to humans. Several variables can contribute for a high risk of a population contracting the disease. In this study, this risk is evaluated from climate, socioeconomic and remote sensing variables, which include MODIS and SRTM data. In this work, two pattern recognition techniques were used to generate two risk maps, with several parameter configurations. The first one is decision trees, for which a total of 19 classifications were generated. The second one technique is the nearest neighbour classification. For this method, only the number of neighbours varied, and 11 classifications were generated. Results showed a better result for the decistion trees in most part of the tests.

Palavras-chave: classification, image processing, k-NN, decision tree, classificação, processamento de imagens, esquistossomose, VMP, árvore de decisão.

1. Intrudução

A esquistossomose *mansoni* é um dos graves problemas de saúde publica que afetam milhares de pessoas em todo mundo (WHO, 1985). No Brasil, a esquistossomose é causada pelo agente etiológico *Schistosoma mansoni*, que tem como hospedeiro intermediário caramujos do gênero *Biomphalaria* (AMARAL et al., 2006). O parasita utiliza a água como meio para infectar o homem (hospedeiro definitivo), que através de suas fezes infectadas contamina a água, possibilitando a infecção do caramujo e dando origem a um novo ciclo. Assim, para estudar a transmissão dessa doença, além de combinar fatores ambientais e sociais, relacionados ao caramujo e ao homem, é importante relacionar esses fatores a aspectos espaciais, visto que locais próximos a áreas endêmicas são locais de potencial risco de contaminação.

A distribuição da esquistossomose nos Estado de Minas Gerais é irregular, intercalando-se em áreas de transmissão baixa ou nula com áreas de maior prevalência da doença. Na Figura 1, é possível verificar que a doença é endêmica nas regiões norte (compreendendo as zonas do Médio São Francisco e Itacambira), oriental e central (zonas do Alto Jequitinhonha, Metalúrgica, Oeste e Alto São Francisco) e que os maiores índices de infecção são encontrados nas regiões nordeste e leste do Estado, que compreendem as zonas do Mucuri, Rio Doce e da Mata (CARVALHO et al., 1987, 2005).

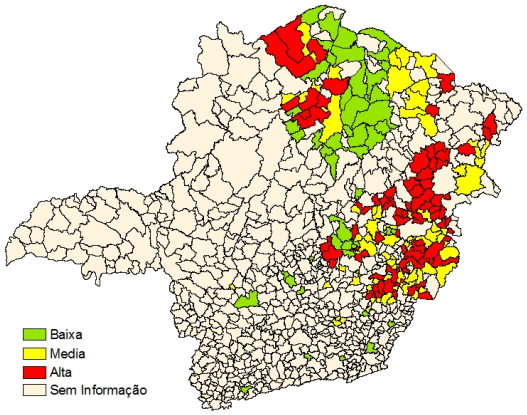


Figura 1. Distribuição da esquistossome em municípios do estado de Minas Gerais.

Neste trabalho foram usados dados da prevalência da esquistossomose em 197 municípios do estado de Minas Gerais para mapear o risco da prevalência da doença em todo o estado. Este mapeamento foi gerado usando dois classificadores: o k vizinhos mais próximos (k-NN, do inglês *k Nearest Neighbors*) e a árvore de decisão (DT, do inglês *Decision Tree*). A diferença básica entre esses classificadores é método usado para verificar o quanto um caso é semelhante a outro.

O classificador k-NN, introduzido por Fix e Hodges (1951), é uma das técnicas mais populares de reconhecimento de padrões. Essa técnica consiste em atribuir uma classe a um elemento com rótulo desconhecido usando, a classe da maioria de seus vizinhos mais próximos. No k-NN tradicional, os k vizinhos mais próximos são determinados segundo a distância Euclidiana no espaço de atributos (WEBB, 2002).

O classificador DT, é uma técnica de reconhecimento de padrões e um modelo prático de inferência indutiva. A DT é construída de acordo com um conjunto de casos previamente classificados. Posteriormente, outros casos são classificadas de acordo com essa mesma árvore. A estratégia dos algoritmos baseados em árvores de decisão é particionar sucessivamente o espaço de busca em subespaços de menores dimensões. As partições são feitas até que cada um dos subespaços contemple apenas uma classe ou até que uma das classes demonstre uma clara maioria, não justificando posteriores divisões. Como é evidente, a classificação consiste apenas em seguir o caminho ditado pelos sucessivos testes colocados ao longo da árvore até que seja encontrada uma folha que conterá a classificação correspondente (FONSECA, 1994).

Estudos anteriores apresentam alguns modelos elaborados para o mapeamento do risco da esquistossomose *mansoni* no estado de Minas Gerais. Tais como: Fonseca et al. (2014), Guimarães et al. (2012, 2010, 2008) e Martins-Bedê et al. (2010, 2008). Nesses estudos foram usadas ferramentas de geoprocessamento, mapas com características ambientais e dados de prevalência da doença. Esses dados foram utilizados para identificar fatores ambientais, que influenciam a distribuição da esquistossomose em Minas Gerais.

O objetivo deste estudo é gerar mapas temáticos que apontem os municípios que, de acordo com o estudo, possuam fatores de risco favoráveis para a doença. Esses mapas podem, a critério da Secretaria de Saúde do Estado de Minas Gerais, ser usados como subsídio para o mapeamento e controle da esquistossomose nos municípios do Estado.

Este artigo está organizado da seguinte maneira: As Seções 2 e 3 apresentam os materiais e métodos usados neste estudo, respectivamente. A Seção 4 apresenta os resultados e as conclusões.

2. Materiais

O Estado de Minas Gerais localiza-se na região Sudeste do Brasil e é dividido politicamente entre 853 municípios e possui área de aproximadamente 590.000 km^2. A população é de aproximadamente 20 milhões de habitantes e o clima é tropical (IBGE, 2013).

Os dados sobre a doença, usados neste trabalho, foram disponibilizados pela Secretaria de Saúde do estado de Minas Gerais em escala municipal. Dos 853 municípios apresentados na Figura 1, 197 possuem informação positiva de prevalência da esquistossomose. As faixas de prevalência abaixo de 5%, entre 5% e 15% e acima de 15%, apresentados na Figura 1, são definidas como baixa, média e alta, respectivamente, de acordo com classificação da Secretaria de Saúde do Estado de Minas Gerais.

Como conjunto de atributos, são utilizados variáveis extraídas de sensoriamento remoto (SR), obtidas pelos sensores *Moderate Resolution Imaging Spectroradiometer* (MODIS) e *Shuttle Radar Topography Mission* (SRTM), variáveis meteorológicas (chamadas neste trabalhos de climáticas) e variáveis socioeconômicas. Em Martins (2009) pode-se obter uma descrição mais detalhada do conjunto de atributos a nível municipal. A descrição dos atributos a nível local é retratada em Guimarães (2010).

As variáveis de SR e climáticas em escala municipal foram obtidas em 2003 em um projeto financiado pela FAPEMIG (processo: 1775/03), são compostas pela média por município. Das variáveis de SR derivadas do MODIS, são usadas: banda azul (Blue), vermelho (Red), infravermelho próximo (NIR), e infravermelho médio (MIR), os índices de vegetação melhorado (EVI), o índice de vegetação da diferença normalizada (NDVI), e as imagens-fração derivadas do MLME, vegetação (Veg), solo (Solo) e sombra (Sombra). Das variáveis obtidas através do SRTM, são usados o modelo digital de elevação (DEM) e a declividade (Dec), derivada do DEM. Das variáveis climáticas (Cli), são usadas a precipitação acumulada (Prec), a temperatura mínima (Tmin) e a temperatura máxima (Tmax).

As variáveis socioeconômicas foram divididas em dois tipos: IDHs e situação por domicílio (Sit). Essas variáveis foram obtidas do Sistema Nacional de Indicadores Urbanos (SNIU) e foram também usadas no projeto da FAPEMIG em 2003. Além do IDH são usados o IDH de educação (IDHE), de longevidade (IDHL) e de renda (IDHR) dos anos de 1991 e 2000. Os dados referentes à situação por domicílio são do ano 2000 e representam: (i) renda do responsável pelo domicílio; (ii) anos de estudo do responsável pelo domicílio; (iii) tipo de saneamento (esgoto e água).

Os atributos foram inicialmente selecionados usando a correlação entre as variáveis observadas e os dados sobre a doença. Dos 62 atributos disponibilizados, 29 foram selecionados por possuírem correlação com a prevalência da esquistossomose superiores a 30%. Dos 29 atributos selecionados, 11 são atributos da Sit, 8 são de IDHs, 6 de SR e 4 de Cli.

3. Métodos

Neste trabalho, foram geradas classificações usando o k-NN e DT. Para cada classificador, foi selecionada uma classificação representante, em um ambiente de simulação. As classificações de cada método foram comparadas de forma pareada, com o uso do índice de desempenho, termo proposto por Martins-Bedê (2014). Esse índice é usado para identificar, em um ambiente de simulação, quantas vezes a acurácia de um classificador é superior a de outro.

As amostras foram estratificadas e foram separados aleatoriamente $2/3$ destas como conjunto de treinamento e os $1/3$ restante como conjunto de teste. A classificação do conjunto de teste e conjunto de dados não rotulados (i.e. municípios sem informação) foi realizada baseada no conjunto de treinamento. Foram realizadas várias classificações, variando os parâmetros para o método DT e k-NN. Para o método DT foram realizados testes com 2 a 20 amostras mínimas por folha, gerando 19 classificações. Para o k-NN foram testados de 1 a 11 vizinhos,

gerando 11 classificações. Segundo Bishop (2006), o limite de vizinhos (11) é dado por uma aproximação da raiz quadrada do número de amostras de treinamento.

Para cada método, foi selecionada uma classificação representante pela média das acurácias. A seleção das classificações representantes foram feitas em três etapas, baseadas em um estudo Monte Carlo. Primeiramente, as amostras do conjunto de teste foram estratificadas e foram selecionados aleatoriamente 1.000 conjuntos com reposição. Esses conjuntos foram usados em todas as comparações. Em seguida, para cada configuração do classificador, foram contabilizadas as acurácias para cada um dos 1.000 conjuntos e foi calculada a média dessas acurácias. A configuração que apresentou a maior média foi selecionada para ser comparada com as outras configurações. Por fim, as 1.000 acurácias da configuração selecionada foram comparadas, de forma pareada, com as acurácias de cada uma das outras configurações, visando verificar a diferença estatística pelo teste T (GUIMARÃES, 2008). Em casos em que as acurácias foram estatisticamente iguais, a classificação selecionada foi a que apresentou a configuração mais simples (i.e. no caso do k-NN, o menor número vizinhos e para a árvore de decisão, o menor número de objetos por folha).

Como resultado final foram gerados 2 mapas com a melhores classificações de cada método. Esses mapas foram gerados usando a classificação que deu origem aos conjuntos para o estudo Monte Carlo.

4. Resultados e Conclusões

Com o uso da metodologia descrita na Seção 2 foi selecionada uma classificação para cada método. A classificação selecionada para o k-NN foi a com 6 vizinhos mais próximos. Para o método DT foi selecionada a classificação com no mínimo 2 objetos por folha.

Para se ter uma ideia geral dos resultados das 1.000 acurácias, foram gerados gráficos do tipo boxplot desses resultados. Esse gráfico é apresentado na Figura 2.

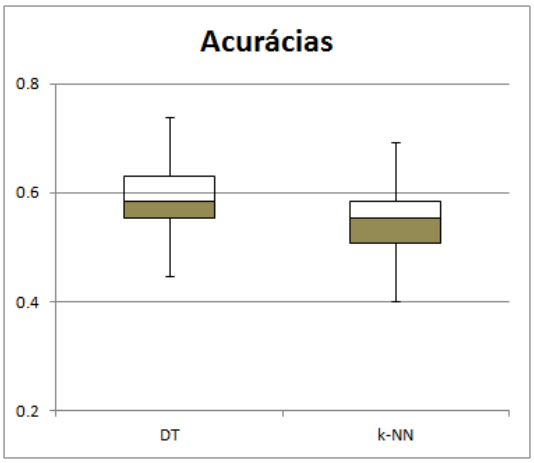


Figura 2. Acurácias da melhor configuração de cada método.

A partir da Figura 2 pode-se ter a impressão que as acurácias das classificações são similares. No entanto, é importante ressaltar que no teste pareado considera-se a ordem das acurácias das classificações e neste tipo de gráfico perde-se esta ordem. Contudo, de acordo com o intervalo de credibilidade, a árvore de decisão obteve acurácias mais altas em 70% dos casos.

As matrizes de confusão para os classificadores DT e k-NN são apresentadas nas Tabelas 1 e 2, respectivamente.

Neste estudo de caso, as classes são ordenadas, por isso alguns erros são considerados inaceitáveis. Um exemplo desse tipo de erro é classificar um município que originalmente é da classe Baixa como Alta, ou vice-versa. Neste trabalho esse tipo de erro é considerado ruim.

Tabela 1. Matriz de confusão do método árvore de decisão usando 2 objetos mínimos por folha.

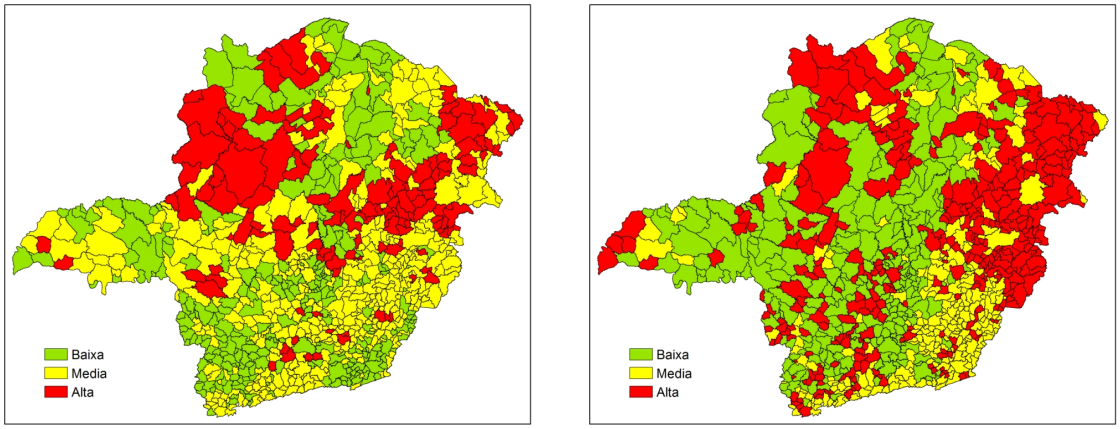
Referência	Baixa	Média	Alta
Baixa (15)	10	4	1
Media (24)	6	17	1
Alta (26)	3	12	11

Tabela 2. Matriz de confusão do método k-NN usando 6 vizinhos.

Referência	Baixa	Média	Alta
Baixa (15)	9	3	3
Media (24)	3	7	14
Alta (26)	1	5	20

A concepção desse trabalho é que o mapa com a classificação possa ser usado como subsídio para alocações de recursos para os municípios que mais precisam. Com esses erros, municípios poderiam não receber recursos necessários ou os recursos poderiam ser alocados indevidamente. Nas matrizes de confusão apresentadas, são poucos os erros considerados ruins, apenas 4 para os dois classificadores.

A Figura 3 apresenta os mapas com as classificações da DT e do k-NN. Visualmente, as classificações resultantes são bastante condizente com os dados de prevalência originais (ver Figura 1). A maior parte dos municípios classificadas como de prevalência Alta estão na região noroeste e leste do Estado. Entretanto, alguns municípios foram classificados como de prevalência Alta embora estejam na área indene (i.e. onde não tem doença).



(a) Árvore de decisão. (b) k-NN.

Figura 3. Classificações selecionadas.

De modo geral, a DT é superior ao k-NN. É conveniente destacar também, que a árvore de decisão confere uma melhor interpretação dos resultados. Os atributos selecionados neste estudo são condizentes com as caracteristicas ambientais e sociais necessárias para a transmissão da esquistossome. Trabalhos semelhantes a este apresentaram acurácias parecidas em Martins-Bedê et al. (2008, 2010). Como trabalhos futuros, pretende-se utilizar outros classificadores como, redes neurais e máquinas de vetores de suporte (SVM), visando encontrar uma melhor classificação.

Referências

AMARAL, R. S.; TAUTIL, P. L.; LIMA, D. D.; ENGELS, D. An analysis of the impact of the schistosomiasis control programme in Brazil. **Mem Inst Oswaldo Cruz**, v. 101, p. 79–85, 2006.

BISHOP, C. M. **Pattern recognition and machine learning**. New York, NY: Springer, 2006. (Information Science and Statistics). ISBN 0387310732.

CARVALHO, O. S.; DUTRA, L. V.; MOURA, A. C. M.; FREITAS, C. d. C.; AMARAL, R. S.; DRUMMOND, S. C.; FREITAS, C. R.; SCHOLTE, R. G. C.; GUIMARÃES, R. J. d. P. Souza e; MELO, G. d. R.; CORREIA, V. R. d. M.; GUERRA, M. Desenvolvimento de um sistema de informações para o estudo, planejamento e controle da esquistossomose no estado de minas gerais. In: **Simpósio Brasileiro de Sensoriamento Remoto, 12. (SBSR)**. São José dos Campos: INPE, 2005. p. 2083–2086. ISBN 85-17-00018-8.

CARVALHO, O. S.; ROCHA, R. S.; MASSARA, C. L.; KATZ, N. Expansão da esquistossomose mansoni em Minas Gerais. **Memórias do Instituto Oswaldo Cruz**, scielo, v. 82, p. 295 – 298, 1987. ISSN 0074-0276.

FIX, E.; HODGES, J. **Discriminatory analysis, nonparametric discrimination: Consistency properties**. Randolph Field, Texas: , 1951.

FONSECA, F. R.; FREITAS, C.; DUTRA, L. V.; GUIMARÃES, R.; CARVALHO, O. Spatial modeling of the schistosomiasis mansoni in minas gerais state, brazil using spatial regression. **Acta Tropica**, v. 133, n. 1, p. 56–63, 2014. ISSN 0001-706X and 1873-6254.

FONSECA, J. M. M. R. D. **Indução de árvores de decisão**. 151 p. Dissertação (Mestrado) — Universidade Nova de Lisboa, Lisboa, 1994.

GUIMARÃES, P. R. B. **Métodos quantitativos estatísticos**. Paris: IESDE Brasil S.A, 2008. 245 p.

GUIMARÃES, R. J. d. P. S.; FREITAS, C. d. C.; DUTRA, L. V.; DUTRA, L. V.; SCHOLTE, R. G. C.; MARTINS-BEDÉ, F. T.; FONSECA, F. R.; AMARAL, R. S.; DRUMMOND, S. C.; FELGUEIRAS, C. A.; OLIVEIRA, G. C. A geoprocessing approach for studying and controlling schistosomiasis in the state of minas gerais, brazil. **Memórias do Instituto Oswaldo Cruz**, v. 105, n. 4, p. 524–531, 2010. ISSN 0074-0276.

GUIMARÃES, R. J. P. S. **Ferramentas de geoprocessamento para o estudo e controle da esquistossomose no Estado de Minas Gerais**. 172 p. Tese (Doutorado em Biomedicina) — Santa Casa de Belo Horizonte, São José dos Campos, 2010.

GUIMARÃES, R. J. P. S.; ALVES, L. G.; FREITAS, C. C.; DUTRA, L. V.; MOURAD, A. C. M.; AMARAL, R. S.; DRUMMONDF, S. C.; SCHOLTE, R. G. C.; CARVALHO, O. S. Schistosomiasis risk estimation in minas gerais state, brazil, using environmental data and GIS techniques. **Acta Tropica**, v. 108, n. 2-3, p. 234–241, 2008. ISSN 0001-706X and 1873-6254.

GUIMARÃES, R. J. P. S.; FREITAS, C. d. C.; DUTRA, L. V.; FELGUEIRAS, C. A.; DRUMMOND, S. C.; TIBIRIÇÁ, S. H. C.; OLIVEIRA, G.; CARVALHO, O. S. Use of indicator kriging to investigate schistosomiasis in minas gerais state, brazil. **Journal of Tropical Medicine**, v. 2012, n. Article number837428, p. 1–10, Jan. 2012. ISSN 1687-9686. Setores de Atividade: Administração pública, defesa e seguridade social, Saúde humana e serviços sociais, Pesquisa e desenvolvimento científico.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Canais**: Estados. 2013. Disponível em: <http://www.ibge.gov.br>. Acesso em: 07 dez. 2013.

MARTINS-BEDÊ, F. d. T.; FREITAS, C. d. C.; DUTRA, L. V.; SANDRI, S. A.; FONSECA, F. R.; DRUMMOND, I. N.; GUIMARÃES, R. J. d. P. S. e.; AMARAL, R. S. d.; CARVALHO, O. d. S. Risk mapping of the schistosomiasis in minas gerais, brasil, using modis and socioeconomic spatial data. **IEEE Transactions on Geoscience and Remote Sensing**, v. 1, p. 1–10, July 2008. ISSN 0196-2892. Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International.

MARTINS-BEDÊ, F. T. **Uma extensão do classificador k-NN para múltiplos espaços**. 152 p. Tese (Doutorado) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2014-02-26 2014.

MARTINS-BEDÊ, F. T.; DUTRA, L. V.; DUTRA, L. V.; FREITAS, C. d. C.; GUIMARÃES, R. J. P. S.; AMARAL, R. S.; DRUMMOND, S. C. Schistosomiasis risk mapping in the state of minas gerais, brazil, using a decision tree approach, remote sensing data and sociological indicators. **Memórias do Instituto Oswaldo Cruz**, v. 105, n. 4, p. 541–548, jul. 2010. ISSN 0074-0276.

MARTINS, F. T. **Mapeamento do risco da esquistossomose no Estado de Minas Gerais, usando dados ambientais e sociais**. 144 p. Dissertação (Mestrado) — Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2008-02-26 2009.

WEBB, A. R. **Statistical pattern recognition**. 1. ed. Chichester: John Wiley & Sons, 2002.

WORLD HEALTH ORGANIZATION (WHO). **The control of schistosomiasis**. 1985. 113 p.