Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brasil, 21-26 abril 2007, INPE, p. 2573-2580.

Desenvolvimento de um modelo de regressão linear para a predição da prevalência de esquistossomose no Estado de Minas Gerais

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

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

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

Abstract: The aim of this paper is to determinate a relationship between schistosomiasis prevalence and social–economic variables, spatial variables and data derived from remote sensing, in the state of Minas Gerais, Brazil, using multiple linear regression.

Palavras-chave: remote sensing, schistosomiasis, multiple linear regression, sensoriamento remoto, esquistossomose, modelo de regressão linear.

1. Introdução

A esquistossomose mansoni é uma doença endêmica, conhecida pelos brasileiros como barriga d'água, xistosa ou doença do caramujo. Seu agente etiológico é o *Schistosoma mansoni* e apresenta formas agudas ou crônicas, com sintomatologia variada, mas com predominância intestinal. A doença é caracterizada, na forma mais grave pelo aumento do fígado e do baço. O diagnóstico e o tratamento são simples, mas a erradicação da doença só é possível com medidas que interrompam o ciclo evolutivo do parasita, como a realização de saneamento básico e a mudança do comportamento das pessoas que vivem em áreas de risco (Katz & Almeida 2003).

Estudos realizados em diversas comunidades da zona rural brasileira demonstraram que a topografia, vegetação, temperatura, tipo de solo, diferentes níveis de saneamento básico, densidade populacional, número de moluscos, distribuição e índices de infecção dos hospedeiros intermediários e contato com coleções hídricas habitadas por moluscos infectados pelo *S. mansoni*, são determinantes para a prevalência da infecção humana (Marcal et al. 1991, Kloetzel & Vergetti 1988, Pieri & Thomas 1987, Bavia et al. 1999, 2001).

A esquistossomose, ainda hoje, pode ser considerada um caso de saúde pública devido a sua incidência em regiões pobres do país e, por isso, requer estudos que ajudem a correlacionar a prevalência da doença com variáveis espaciais, sócio-econômicas ou ambientais, determinando possíveis locais de transmissão da doença para que, efetivamente, possam ser feitas campanhas de combate e prevenção à doença pelos órgãos competentes.

O presente estudo teve como objetivo estimar a prevalência da esquistossomose no Estado de Minas Gerais, através de modelos estatísticos, utilizando dados provenientes de sensoriamento remoto, sócio-econômicos, espaciais e dados históricos de prevalência da doença. O trabalho foi dividido em quatro etapas, incluindo a coleta e preparação dos dados, redução do número de variáveis preditivas, refinamento e seleção do modelo, e validação do modelo.

1.1. Área de Estudo

Minas Gerais é o quarto maior estado do Brasil. Localiza-se no sudeste e limita-se a norte e nordeste com a Bahia, a leste com o Espírito Santo, a sudeste com o Rio de Janeiro, a sul e sudoeste com São Paulo, a oeste com o Mato Grosso do Sul e a noroeste com Goiás. É o segundo Estado mais populoso, com 18 milhões de habitantes que se distribuem politicamente em 853 municípios numa área aproximada de 590.000 km2 (IBGE, 2006).

Segundo dados da Secretaria de Saúde de Minas Gerais, a área em destaque na Figura 1 possui informações catalogadas sobre focos de esquistossomose. Dentre os 853 municípios do Estado de Minas Gerais, tem-se 197 municípios com dados históricos de prevalência da esquistossomose, dos quais foram usados 142 municípios para a construção do modelo e 55 para a validação do modelo de regressão linear.

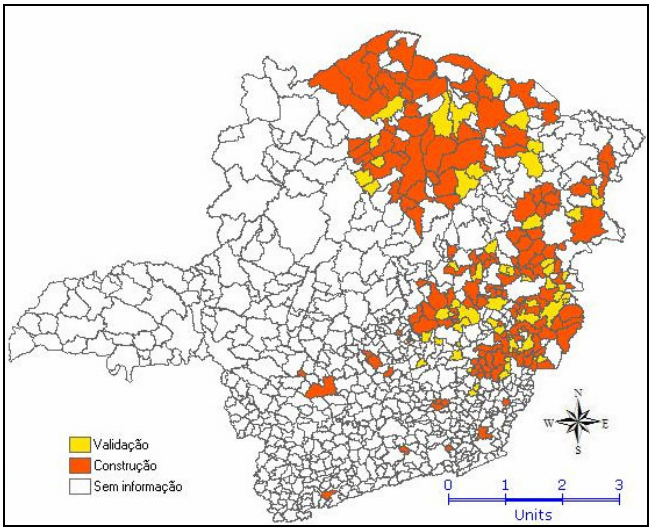


Figura 1: Municípios do Estado de Minas Gerais em estudo, destacando-se os que possuem informações de prevalência de esquistossomose.

2. Materiais

Para desenvolver um modelo de regressão linear para estimar a prevalência de esquistossomose no Estado de Minas Gerais, foram utilizadas quarenta e quatro variáveis, incluindo variáveis espaciais, de sensoriamento remoto, sócio-econômicas e a variável dependente com informações da prevalência da doença.

Os dados sobre a prevalência da esquistossomose (porcentagem dos casos positivos da doença em relação à população total do município) foram cedidos pela Secretaria de Estado de Saúde de Minas Gerais, a partir de dados históricos catalogados em 197 municípios do Estado. Segundo metodologia da Secretaria de Estado de Saúde de Minas Gerais, os dados de prevalência se tornam disponíveis quando são analisados no mínimo 80% da população. Desta forma, neste trabalho, considera-se como informação de prevalência estes dados disponibilizados pela Secretaria.

Os dados de sensoriamento remoto foram derivados do MODIS (*Moderate Resolution Imaging Spectroradiometer*) e SRTM (*Shuttle Radar Topography Mission*). Foram utilizadas nove variáveis do sensor MODIS, coletados em duas datas, uma no verão (17/01/2002 a 01/02/2002) e outra no inverno (28/07/2002 a 12/08/2002), e duas do SRTM. As variáveis do MODIS são 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 de mistura, vegetação (VEG), solo (SOLO) e sombra (SOMB). Os dados obtidos pelo sensor SRTM são o modelo digital de elevação (DEM) e a declividade (DEC), derivada do DEM.

As dezoito variáveis sócio-econômicas foram fornecidas pelo Instituto Brasileiro de Geografia e Estatística - IBGE, que incluíram dados de índice de desenvolvimento humano (IDH), de longevidade (IDHL), renda (IDHR), educação (IDHE), dos anos de 1991 e 2000; três variáveis com informações de qualidade de água do ano 2000, que se referem ao percentual de domicílios com acesso à rede geral de abastecimento de água (%Água), com acesso à água através de poço ou nascente (%Poço) e com outra forma de acesso à água (%Outros); e sete variáveis do ano de 2000 referentes à qualidade de esgoto dos municípios em estudo, com dados de percentual de domicílio com banheiro ou sanitário e outro tipo esgotamento (A), ligado a rio, lago ou mar (B), ligado a uma vala (C), fossa rudimentar (D), fossa séptica (E), rede geral (F) e percentual de domicílios sem banheiro ou sanitário (G).

As cinco variáveis espaciais (Esp1, Esp2, Esp3, Esp4, Esp5) são equivalentes aos valores de prevalência dos cinco municípios mais próximos ao município em questão, divididos pela raiz quadrada da distância a esses municípios. As distâncias dos municípios considerando seus centróides, foram obtidas pelas coordenadas UTM/Datum SAD 69 utilizando o fuso 22.

Durante o desenvolvimento do trabalho foram utilizados as ferramentas *Statistica* 6.0 e *Excel* para seleção de variáveis e análises estatísticas, e o software *Terra View* 3.1.3 para a geração de mapas e visualização de resultados com a aplicação do modelo de regressão linear.

3. Metodologia e Resultados

As quarenta e três variáveis independentes foram correlacionadas com a variável prevalência de esquistossomose (Y). A partir de uma análise de regressão, foi possível selecionar as variáveis que melhor explicam a prevalência da doença.

Após a realização da correlação entre as variáveis coletadas, obteve-se baixa correlação das variáveis independentes em relação à dependente. Para melhorar a correlação obtida, foram efetuadas transformações em Y. Desta maneira, foram testadas as transformações inversas (1/(Y+1)), logarítma (Ln(Y+1)), e raiz quadrada (Sqrt(Y)), com a finalidade de se normalizar a variável dependente e linearizar sua relação com as variáveis independentes.

Selecionou-se a transformação Ln(Y+1), a qual foi usada como variável dependente no restante do trabalho.

As variáveis independentes também foram submetidas a transformações, porém a maioria foi considerada em seu range original de valores, pois não apresentaram um aumento significativo na correlação com a variável dependente. Foi considerada no trabalho apenas as transformações Ln(Esp1+1) e 1/(%Poço+1) correspondentes às variáveis do primeiro município mais próximos com dados de prevalência dividido pela raiz quadrada da distância, e o percentual de domicílios com acesso à água através de poço ou nascente, respectivamente.

Através de uma análise detalhada da correlação das variáveis independentes, retirou-se do modelo as variáveis com baixa correlação com a variável dependente e as variáveis preditivas com correlação entre si acima de 80% . Feito isso, ainda com dezessete variáveis, foi efetuado o procedimento *best subset* (Neter et al., 1996), através dos critérios R2 (**Figura 2**), R2 ajustado e Mallows'Cp, definindo o melhor conjunto de variáveis explicativas.

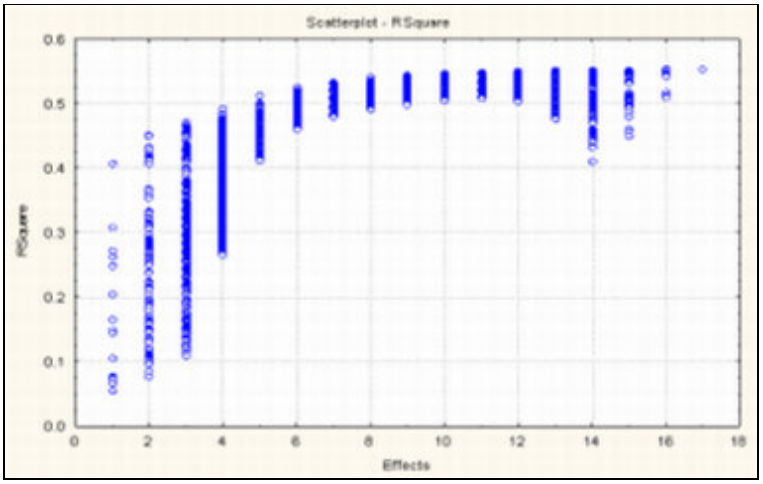


Figura 2 - R^2 versus Número de variáveis

Durante o procedimento de escolha das variáveis, verificou-se que os modelos com cinco ou quatro variáveis poderiam ser os mais adequados, devido à pequena diferença nos resultados dos testes e pela simplicidade do modelo em relação aos demais.

Dentre os modelos apresentados no procedimento *best subset*, foram selecionados dois modelos, um com cinco variáveis [Ln(Esp1+1), Esp2, DEM, NIR-I, SOMB-i], contendo duas variáveis com informações espaciais, e outro com quatro variáveis [DEM, NIR-I, SOMB-i, 1/(%Poço+1)] excluindo os dados espaciais, mas incluindo uma variável com correlação significativa com os dados espaciais. A escolha de dois modelos, um com informação da prevalência dos municípios mais próximos (informação espacial) e outro sem esta informação deveu-se ao fato de que o objetivo deste trabalho é estimar a prevalência para todo o Estado de Minas Gerais, e como a informação de prevalência existente se resume basicamente ao norte e leste do Estado, quando a estimação é efetuada para os municípios distantes (principalmente à oeste do estado) as informações espaciais perdem seu poder explicativo, devido à grande distância em relação aos municípios com informações de prevalência. Desse modo, adotou-se os dois modelos para que, a medida em que o modelo contendo variáveis espaciais perder sua capacidade explicativa, o modelo sem as variáveis espaciais se torna mais adequado para estimar a prevalência da doença.

Os dois modelos foram submetidos ao teste de Breusch-Pagan (Neter et al., 1996) para se testar a hipótese de homocedasticidade dos resíduos. Esta hipótese foi aceita ao nível de

confiança de 99% para o modelo com cinco variáveis, e ao nível de 95% para o modelo com quatro variáveis.

Além da análise de homocedasticidade, os modelos propostos foram analisados para verificar se havia a presença de *outliers*, e observou-se que não houve a influência de nenhum.

3.1. Validação dos modelos

Após a geração e escolha dos modelos de regressão, foi realizada a validação dos modelos selecionados com o intuito de observar se os resultados seriam equivalentes em toda a área de estudo. Dentre os vários métodos existentes, optou-se por aquele em que um novo conjunto de amostras é utilizado para checar o modelo e sua capacidade de predição (Neter et al., 1996). Para a validação dos modelos propostos foram utilizadas 55 amostras, previamente selecionadas.

Ao final da análise e construção dos modelos de regressão, ambos os modelos em estudo atenderam aos requisitos de validação. Desta forma, pode-se afirmar que os modelos são válidos para estimar a prevalência da esquistossomose no Estado de Minas Gerais.

Uma vez que os modelos foram validados, a equação final foi efetuada reunindo-se os dados de construção e os dados de validação, totalizando 197 municípios.

4. Conclusão e discussão dos resultados

O modelo final com cinco variáveis obteve o coeficiente de determinação de 43% e incluiu as variáveis: modelo digital de elevação (DEM), infravermelho próximo no inverno (NIR-i), sombra do modelo de mistura espectral no inverno (SOMB-i), a transformação da variável do primeiro município mais próximo dividido pela raiz quadrada da distância (Ln(Esp1+1)) e a prevalência do segundo município mais próximo dividido pela raiz quadrada de sua distância (Esp2).

A equação final, gerada pelo modelo de regressão com as cinco variáveis baseadas nos 197 municípios é:

$$Pv = e^{(-1{,}98+24{,}20*Ln(Esp_1+1)+38{,}74*Esp_2-0{,}00094*DEM+13{,}87*NIR_i+0{,}045*SOMB_i)} - 1 \quad \textbf{(1)}$$

Através da equação obtida, observa-se a correlação positiva em relação aos dois municípios mais próximos, chegando a uma conclusão de, quanto maior a prevalência dos municípios vizinhos à cidade em estudo, maior é a prevalência da esquistossomose. Esta relação pode ser justificada segundo a "lei de Tobler": "no mundo, todas as coisas se parecem; mas coisas mais próximas são mais parecidas que aquelas mais distantes" (Tobler, 1979). Ainda no mesmo modelo foi analisado que a prevalência é inversamente relacionada com o DEM, ou seja, quanto menor é a elevação do terreno maior a distribuição da doença. Foi observado também que quanto maior o NIR-i e SOMB-i maior a prevalência da doença. Isto coincide com as condições adequadas para o desenvolvimento dos moluscos, como a presença de água na vegetação definida pela variável NIR-i e a concentração de água ou topografia acidentada que pode estar associada á variável SOMB-i.

Já o modelo final com quatro variáveis obteve o coeficiente de determinação de 34% e incluiu as variáveis: modelo digital de elevação (DEM), infravermelho próximo no inverno (NIR-i), sombra do modelo de mistura espectral no inverno (SOMB-i), a transformação da variável de qualidade de água, que mostra o percentual de domicílios com acesso a água através de poço ou nascente [1/(%Poço+1)]. O modelo final, baseado nos 197 municípios, é:

$$Pv = e^{(-1{,}07309-0{,}00133*DEM+17{,}77706*NIRi+0{,}03478*SOMBi-8{,}51211*[1/\%Poço+1])} - 1 \quad \textbf{(2)}$$

A influência das variáveis DEM, NIR-i e SOMB-i, sobre a prevalência de esquistossomose é a mesma em relação ao modelo com cinco variáveis. O modelo mostra que quanto maior o percentual dos domicílios com acesso a água através de poço ou nascente maior a prevalência da doença.

Nota-se através dos coeficientes de determinação, que o melhor modelo para explicar a presença da doença, é o modelo que contém variáveis espaciais. Na análise dos dados, observou-se que 99% dos 197 municípios encontram-se a uma distância menor que 130 km do município mais próximo que possui informação de prevalência. Portanto, decidiu-se utilizar o modelo dado pela Equação (**1**) para os municípios distantes a menos de 130 km de municípios com informação de prevalência, e o modelo dado pela Equação (**2**) para o restante dos municípios. Assim, dos 853 municípios, observa-se que 673 municípios se adequam ao modelo com variáveis espaciais e 180 ao modelo sem as variáveis espaciais. A distribuição espacial destes municípios é apresentada na **Figura 3**.

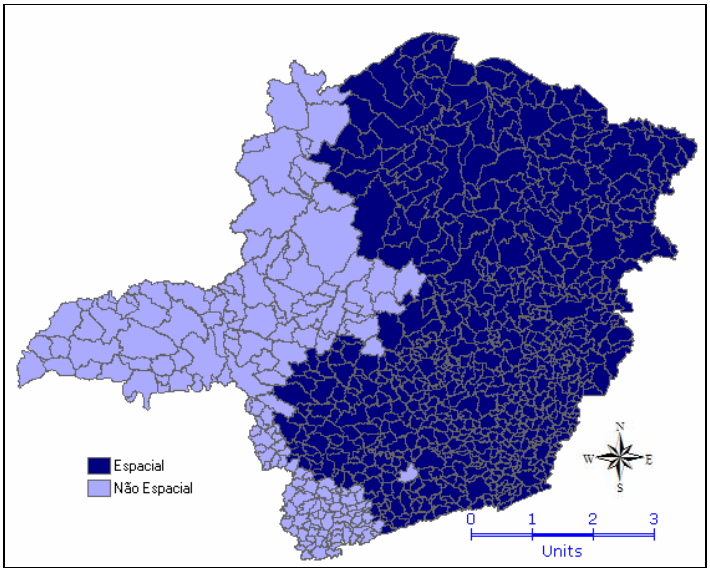


Figura 3 – Aplicação dos modelos nos municípios de Minas Gerais

A partir das equações propostas em cada um dos modelos selecionados, foi realizada a análise dos resultados através de mapas da estimativa da prevalência e mapas de resíduos. A seguir são apresentados os mapas representativos da prevalência observada (**Figura 4**), de prevalência estimada e dos resíduos dos modelos obtidos (**Figuras 5a e 5b**, respectivamente).

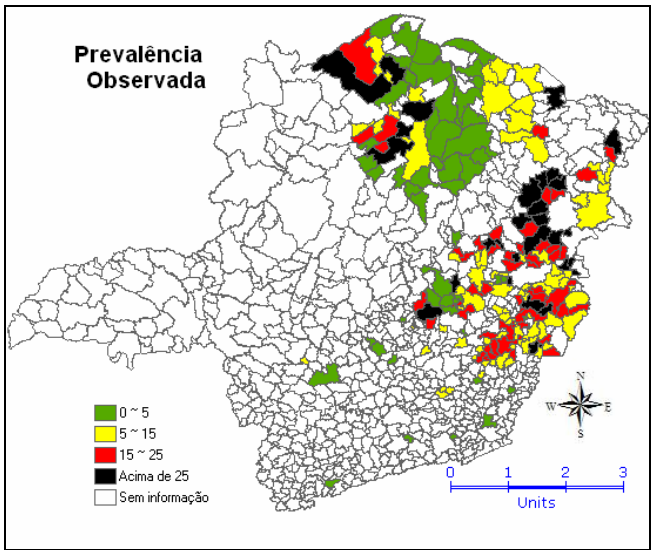
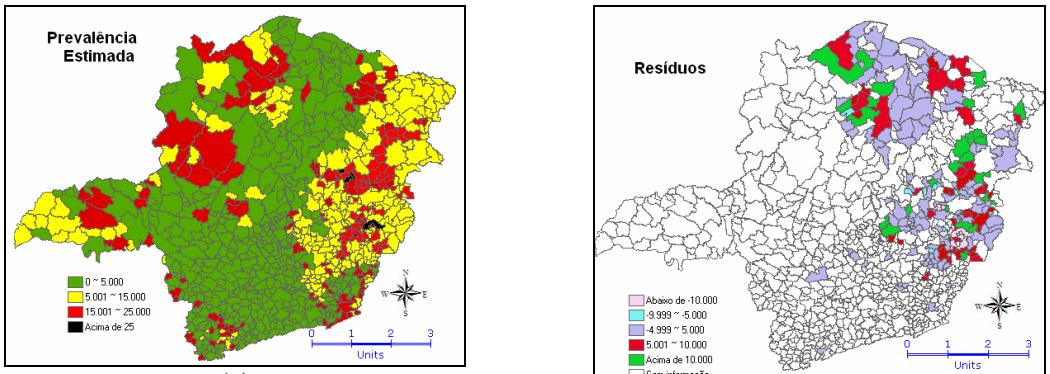


Figura 4 – Prevalência observada.



(a) (b)

Figura 5 – (a) Prevalência estimada e (b) resíduos dos modelos.

Na **Figura 5b**, observa-se a presença de resíduos resultantes da diferença entre o valor observado e o valor estimado da prevalência. As cores mais escuras (vermelho e verde) representam os dados subestimados e nas cores mais claras (azul e rosa) representam os dados superestimados e finalmente em lilás, faixa que varia entre -4,999 a 5,000, representam os dados que tiveram uma boa estimativa.

De acordo com os resultados encontrados no presente trabalho, verifica-se que os modelos podem ser considerados eficazes e confiáveis para predizer a prevalência da esquistossomose no Estado de Minas Gerais, possibilitando indicar uma melhor distribuição de recursos e direcionamento mais adequado para o controle do molusco. Sugere-se para trabalhos futuros o uso do desvio padrão, para prever o risco da doença para todos os municípios do Estado de Minas Gerais, uma vez que esta é uma das medidas mais comuns de dispersão estatística.

5. Agradecimentos

Os autores reconhecem o suporte do CNPq (processos 384467/2006-7; 305546/2003-1; 380203/2004-9; 304274/2005-4); Fapemig (processo EDP 1775/03; EDT 61775/03; CRA 0070/04).

6. Referências

Bavia ME, Hale LE, Malone JB, Braud DH, Shane SM. Geopraphic information systems and the environmental risk of Schistosomiasis in Bahia, Brazil. **Am J Trop Med Hyg** 60(4): 566-572, 1999.

Bavia ME, Malone JB, 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** 79: 79-85, 2001.

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

Kloetzel K, Vergetti AMA, 1988. Repeated mass treatment of Schistosoma mansoni: experience in hyperendemic areas of Brazil. II. Micro-level evaluation of results. **Ann Trop Med Parasitol** 28: 316-367.

Katz N, Almeida K. Esquistossomose, xistosa, barriga d'água. **Cienc. Cult.,** vol.55, no.1, p.38-43. ISSN 0009-6725, Jan./Mar 2003.

Marcal JR O, Patucci MJ, Dias LCS, 1991. Schistosomiasis mansoni in area of low transmission. I. Impact of control measures. **Rev Inst Med Trop** São Paulo 33: 83-90.

Neter, J.; Kutner, M. H.; Nachtssheim, C. J.; Wasserman, W. **Applied linear statistical models**. 4. Boston: WCB/McGraw-Hill, 1996.

Piere O S, Thomas J D .Snail host control in the eastern coastal areas of north-east Brazil**. Mem Inst Oswaldo Cruz** 82 (suppl IV): 197-201, 1987

Tobler, W. Cellular geography. In: S. Gale and O. G. (ed). **Philosophy in Geography**. Dordrecht, Reidel, 1979. v., p.379-386.