Schistosomiasis risk mapping in the state of Minas Gerais, Brazil, using a decision tree approach, remote sensing data and sociological indicators

Flávia T Martins-Bedê[1], Luciano V Dutra[1]/+, Corina C Freitas[1], Ricardo JPS Guimarães[2,3], Ronaldo S Amaral[4], Sandra C Drummond[5], Omar S Carvalho[6]

[1]Instituto Nacional de Pesquisas Espaciais, CP 515, 12201-970 São José dos Campos, SP, Brasil [2]Instituto de Pesquisa René Rachou-Fiocruz, Belo Horizonte, MG, Brasil [3]Programa de Pós-Graduação em Clínica Médica e Biomedicina, Santa Casa de Misericórdia, Belo Horizonte, MG, Brasil [4]Secretaria de Vigilância em Saúde, Ministério da Saúde, DF, Brasil [5]Secretaria de Estado de Saúde, Belo Horizonte, MG, Brasil

Schistosomiasis mansoni is not just a physical disease, but is related to social and behavioural factors as well. Snails of the Biomphalaria *genus are an intermediate host for* Schistosoma mansoni *and infect humans through water. The objective of this study is to classify the risk of schistosomiasis in the state of Minas Gerais (MG). We focus on socio-economic and demographic features, basic sanitation features, the presence of accumulated water bodies, dense vegetation in the summer and winter seasons and related terrain characteristics. We draw on the decision tree approach to infection risk modelling and mapping. The model robustness was properly verified. The main variables that were selected by the procedure included the terrain's water accumulation capacity, temperature extremes and the Human Development Index. In addition, the model was used to generate two maps, one that included risk classification for the entire of MG and another that included classification errors. The resulting map was 62.9% accurate.*

Key words: schistosomiasis - risk mapping - public health - decision tree

Schistosomiasis is caused by the trematoda *Schistosoma mansoni*, whose main intermediate hosts in Brazil are snails of the genus *Biomphalaria* (*Biomphalaria glabrata*, *Biomphalaria tenagophila* and *Biomphalaria straminea*). The disease is related to social and behavioural factors, particularly inadequate public and environmental sanitation and a low level of education about health in the populations involved (Doumenge et al. 1987).

Once schistosomiasis risk is identified by both environmental and social factors, new computerised analytical tools, known as Geographic Information Systems and Remote Sensing Data Analysis, have been used to map epidemiological data or analyse satellite images (Bavia et al. 2001, Freitas et al. 2006).

In Brazil, geo-processing tools have been used in the study of schistosomiasis in the states of Bahia (BA) and Minas Gerais (MG). These studies have provided risk maps for schistosomiasis infection on a municipality basis by using multiple regression analyses that include environmental features, *a priori* disease prevalence data and other spatial data (Bavia et al. 1999, 2001, Freitas et al. 2006, Guimarães et al. 2006, 2008, 2009, Martins-Bedê et al. 2009).

In the present paper, a standard data-mining technique, the decision tree, was used to identify the severity of disease prevalence. This technique is based on a recursive partitioning of predictor variables in which knowledge of the problem is represented by a decision rules structure.

Predictive models are used to classify different samples whose values or labels are not known. In this paper, a decision tree model is used to classify the schistosomiasis prevalence risk for the whole state. Remote sensing data and spatial sociological indicators are used to map potential risk areas which are not covered by the schistosomiasis control program. This map of potential risk areas can be used by a decision maker to evaluate municipalities outside of the program that have similar environmental and social conditions to the municipalities covered by the program. Moreover, the map can serve as a guide for future disease-control efforts in the state.

MATERIALS AND METHODS

Study area - The study area was MG, Brazil. MG is 590,000 square kilometres in size and is politically divided into 853 cities; the area has a tropical climate and includes approximately 18 million inhabitants (IBGE 2008).

The distribution of schistosomiasis in MG is irregular. The state has areas with a high prevalence and areas where transmission is low or non-existent. The disease is endemic in the northern, eastern and central regions of the state, but is not endemic in the Triângulo Mineiro region or the northwestern and southern parts of the state. The greatest rates of infection are found in the northeastern and eastern regions of the state in the Mucuri, Rio Doce and Mata zones (Pellon & Teixeira 1950, Katz et al. 1978, Carvalho et al. 1987, 2005).

Included variables - Disease prevalence data have been provided by the Health Secretary of MG. The prevalence is known for 197 out of 853 municipalities in the

Financial support: INPE, CAPES, CNPq, NIH-Fogarty
+ Corresponding author: dutra@dpi.inpe.br
Received 7 April 2009
Accepted 16 October 2009

state (Fig. 1). These data have been used to construct a decision tree to classify schistosomiasis risk in all of the municipalities in MG.

Sixty-two variables were used. Of these, 22 were derived from both MODIS (Moderate Resolution Imaging Spectroradiometer) and SRTM (Shuttle Radar Topography Mission) information, six were related to climate conditions and 34 emerged from socioeconomic data.

The data were pre-processed so the variables would fit the tool's format. In addition, the classification algorithm required that a nominal variable type be predicted. Disease prevalence data were therefore classified into three categories according to the standards of the Secretary of Health: low (0-5%), average (> 5-15%) and high (> 15%).

The 62 variables were allocated into six groups relevant to the study of schistosomiasis: socioeconomic and demographic data, basic sanitation, presence of water bodies accumulation and dense vegetation during the summer, presence of water accumulation and dense vegetation during the winter, climate and variables related to the terrain. Supplementary data presents a description of the variables included in each group.

Data on socioeconomic, demographic and sanitation variables (Groups 1 and 2) were provided by the Sistema Nacional de Indicadores Urbanos (SNIU 2005); some of the human development indices were previously used by Guimarães et al. (2006). These variables range from 0-1 and are sometimes expressed as percentages.

Other variable groups (such as the presence of water and dense vegetation during the summer and winter) were determined by Guimarães et al. (2008) using four images (h14v10, h14v11, h13v10, h13v11). These images were provided by MODIS sensor and reflect two dates: one during the summer and another during the winter (17 January 2002 and 28 July 2002). The eight images were reprojected using the MODIS Reprojection Tool. Further processing was performed using Spatial Planning for Regions in Growing Economies, Environment for Visualizing Images and ArcGis. Further information about this data is available in Guimarães et al. (2008). The blue band in the summer, blue band in the winter, RedS (red band in the summer), RedW (red band in the winter), near infrared band in the summer, near infrared band in the winter, middle infrared band in the summer and middle infrared band in the winter variables are each expressed as reflectance (%). The variables generated using the linear mixture model [RedS, RedW, soil in the summer, soil in the winter, shade in the summer and ShadeW (shade in the winter)] are also expressed as percentages and the EVIS (enhanced vegetation index in the summer), EVIW (enhanced vegetation index in the winter), normalized difference vegetation index in the summer and normalized difference vegetation index in the winter variables vary from -0.2-1.

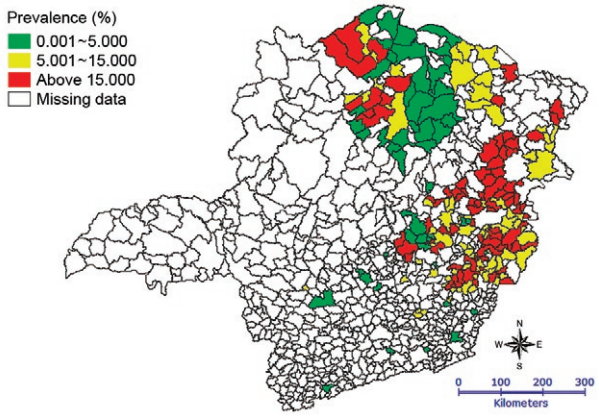

Fig. 1: schistosomiasis prevalence. Source: Health Secretary of the state of Minas Gerais.

The climate group variables were collected by Guimarães et al. (2006) with data collection platforms obtained from Centro de Previsão de Tempo e Estudos Climáticos-Instituto Nacional de Pesquisas Espaciais for both the summer (from 17 January - 17 February 2002) and winter (from 28 July - 12 August 2002) seasons. The temperature variables [average of daily maximum temperature in the winter, TmaxS (average of daily maximum temperature in the summer), TminW (average of daily minimum temperature in the winter) and average of daily minimum temperature in the summer] are expressed in ^{o}C and the precipitation variables are expressed in mm.

The terrain group variables [Dem (digital elevation model of terrain) and Dec (slope declivity)] were obtained from SRTM Dem obtained during an 11-day mission in February 2000. Topographic data of the earth's surface were gathered using orbital Synthetic Aperture Radar Interferometry according to the method described by Guimarães et al. (2008). The Dem variable is given in metres and the Dec variable is given in degrees. The other two variables of this group, average of accumulated water and median of accumulated water, were obtained using the water accumulation area map generated from the SRTM Dem. This map measures the number of possible pathways water can run to reach a particular site in a hydrographic basin and the resulting values vary from 0-80,000.

Variable selection - The variables under study are highly correlated and were preselected; this process reduced the total of 62 variables to 12 (2 variables for each of the 6 groups). First, three variables from each group that are most correlated with the disease were selected. Next, the two least-correlated variables were determined; this process ensured, as much as possible, the use of the most informative variables. For each variable, the ratio between its correlation with prevalence and its correlation with the remaining 15 variables was calculated; the two variables in each group with the highest ratios were ultimately selected.

Software - Waikato Environment for Knowledge Analysis (Weka), a public domain software program from the University of Waikato, New Zealand, was used to perform the pattern recognition analyses. From the available resources, the well known decision tree algorithm C4.5, developed by Quinlan (1993) and implemented as the J4.8 procedure inside the Weka software, was used to estimate the decision trees (Witten & Frank 2005).

Decision trees - The decision tree is a pattern recognition technique and a practical model used in inductive inferences. These trees are constructed according to a previously classified sample set and, afterwards, other unlabelled samples are classified according to this same tree. Decision trees are frequently used in applied fields like finance, marketing, engineering, health and remote sensing (Mitchell 1997, Witten & Frank 2005). The algorithms ID3, ASSISTANT and C4.5 (Quinlan 1993) are examples of algorithms used for building these trees. These algorithms generally rely on the expected distribution of values for each variable or on the independent relationships among variables; the C4.5 algorithm is based on the entropy concept.

Decision trees consist of a hierarchy of internal nodes and leaves that are connected by edges (branches). Interior nodes correspond to one of the input variables. Each leaf represents a value of the target variable given ranges of values of the input variables represented by the path from the root to the leaf.

In these structures, the leaves represent classifications and are associated with a label or a value. Decision makers can employ decision trees to identify the best strategy to reach an objective (Mitchell 1997, Witten & Frank 2005, Theodoridis & Koutroumbas 2006).

This method performs both classification and predictive functions simultaneously. Decision trees use a sequence of questions and rules to classify an object or an incident within predetermined classes based on attribute values. A decision is reached by seeking a path in the tree based on a comparison of the value of each input variable with the value of its corresponding node; the chosen direction depends on whether the variable's value is bigger than, smaller than or equal to the node value.

The rationale behind any decision tree-based algorithm is to break down a problem into sub-problems through a successive division of the feature space; ultimately, a solution for each smaller problem can be found. Under this principle, the classifiers based on decision trees seek to find ways to successively split the universe into several subsets by creating nodes that include the respective tests. This process continues until each of the nodes becomes a unique class or until one of the classes demonstrates a clear dominance that precludes further divisions (generating, in this case, 1 leaf that contains the majority class). Classification entails following a path determined by the successive tests placed along the tree until a classification leaf is found (Mitchell 1997, Witten & Frank 2005, Theodoridis & Koutroumbas 2006).

During the tree-building process, called the training phase (Mitchell 1997, Witten & Frank 2005, Theodoridis & Koutroumbas 2006), it is possible to establish a minimum number of training samples to be considered in each leaf. The decision tree can be analysed by an expert and, if necessary, it can be modified according to rules underlying the system's rationale.

Results assessment - To evaluate the accuracy and robustness of the decision rules, the 197 samples were randomly allocated into three sets of 132 samples each; one third of the samples were always set apart to test each estimated decision tree. Decision trees were built for each set, compared, and a final decision tree was constructed with all 197 samples. All known samples used for building the tree or for testing were classified; the results were summarised in a classification matrix, in which the reference samples are placed on the rows (Table I). From this matrix, the global accuracy (AC) and user accuracy were calculated; the AC is the ratio of the number of correctly classified samples to the total number of samples and the user accuracy is the relative number of correctly classified samples in a particular class compared to the total number of samples in that class. The samples that were not initially included in the training set but were used as a test set were used to evaluate the generalisation potential of the classification method.

TABLE I

Confusion matrix for the tree with 197 samples

Global accuracy (62.9%)			
Class (n)	Classification		
	Low n (%)	Median n (%)	High n (%)
Low (46)	31 (67.4)	8 (17.4)	7 (15.2)
Median (73)	8 (10.9)	29 (39.7)	36 (49.3)
High (78)	2 (2.6)	12 (15.4)	64 (82.1)
User accuracy (%)	75.6	59.2	59.8

RESULTS

The 12 selected variables included: HDI-91 (Human Development Index in year 1991) and HDIE-91 (Education Human Development Index in year 1991) from the socioeconomic and demographic group; % of residence with access to water supply by means of wells and % of residence with sewage connected to rudimentary cesspit from the sanitation group; EVIS and vegetation in the summer from the group concerning the presence of water and dense vegetation during the summer; EVIW and ShadeW from the group concerning the presence of water and dense vegetation during the winter; TmaxS and TminW from the climate group; and AWater2 (median of the accumulated water area) and Dec from the terrain group. These variables, together with the prevalence data, were used to build the decision tree based on all 197 samples.

A minimum of 12 samples per leaf was set as a parameter for the estimation process. This parameter limits the depth of the estimated tree and results, in general, in a three-level tree; this depth is a good compromise between accuracy over the training samples and generalisation power (i.e., good accuracy over test samples). The decision rules generated by the J4.8 Weka algorithm, with 62.9% AC, are shown in Fig. 2.

The total number of classified samples, as well as those incorrectly classified, is displayed in parentheses for each leaf in Fig. 2 according to each classification rule.

The variable showing the highest amount of information AWater2 is placed at the root of the decision tree.

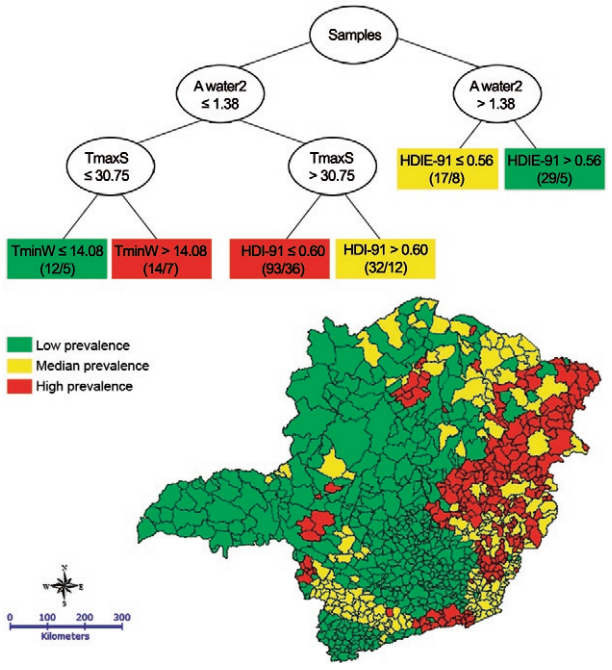

Fig. 2: graphical representation of the decision tree. Temperature in ºC. Human Development Index (HDI) is defined between 0-1. The median of the accumulated water area (AWater2) is a dimensionless counting. HDI-91: Human Development Index in year 1991; HDIE-91: Education Human Development Index in year 1991; TmaxS: average of daily maximum temperature in the summer; TminW: average of daily minimum temperature in the winter.

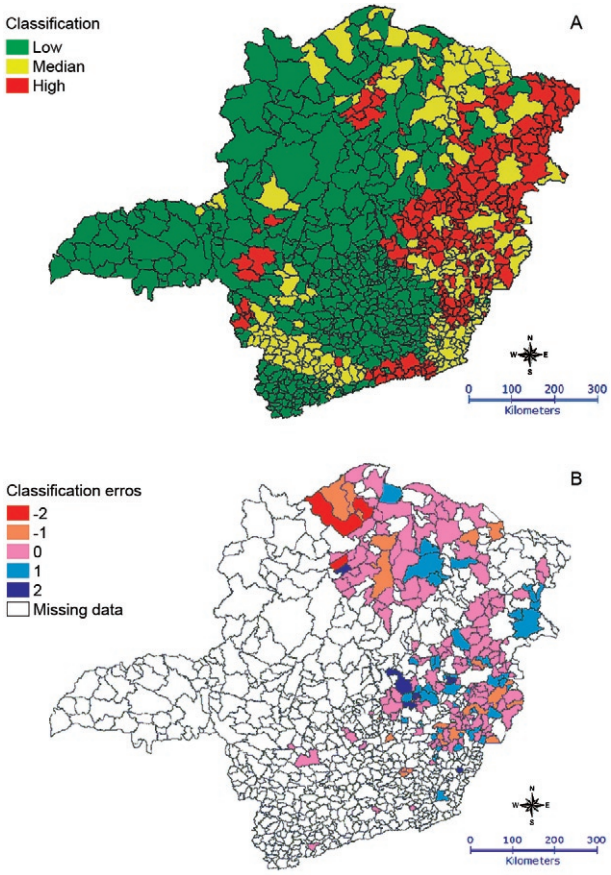

Fig. 3: A: schistosomiasis prevalence classified by the decision tree; B: classification errors.

Increasingly homogeneous sets are formed in each tree node. When values of this variable are greater than 1.386, the samples are classified into median or low prevalence depending on the HDIE-91 value: HDIE-91 values less than or equal to 0.561 reflect median prevalence and other samples are classified as low prevalence. The classification of samples with AWater2 values less than or equal to 1.38 will depend on both temperature (TmaxS and TminW) and HDI variables.

When TmaxS values are less than or equal to 30.75ºC, the samples can be classified as either low prevalence (if TminW values are less than or equal to 14.08ºC) or high prevalence (if TminW values are greater than 14.08ºC). This classification can be considered coherent since studies performed in BA by Bavia et al. (2001) and in Ethiopia by Malone et al. (2001) showed that the temperature was a good predictor of the disease.

When TmaxS values are greater than 30.75ºC and HDI-91 values are less than or equal to 0.607 the samples are classified as high prevalence; when HDI-91 values are greater than 0.607 the samples are classified as median prevalence. These variables were also selected by Guimarães et al. (2006) as predictor variables.

The selected variables are related to the ideal habitat for snails and to people's life conditions, both of which are important factors for disease occurrence. Fig. 3 depicts the thematic map obtained by applying the classification rules of the decision tree for the whole MG (Fig. 3A) and the classification errors for the 197 samples (Fig. 3B).

In Fig. 3B, the light and dark orange areas are municipalities where classification is underestimated by either two prevalence classes or one prevalence class, respectively. That is, the two cities in orange were classified as low prevalence class, but they actually belong to the high prevalence class. The light and dark blue colours highlight municipalities where the classification is overestimated by either two or one prevalence class, respectively. Pink-coloured municipalities were correctly classified.

Table I displays the confusion matrix generated by the decision tree. Low prevalence classes presented a user accuracy of 75.6%; this means that 75.6% of the samples that were classified as low actually belong to the low prevalence class. Similarly, the user classification accuracy was about 60% for both high and median prevalence classes. These results indicate that there was a significant improvement in predictive power when using the classification procedure (52.3%, 22.1% and 20.2% for low, median and high prevalence classes, respectively), when compared with the *a priori* probability of class occurrences (calculated as the ratio of the number of *a priori* samples of each class to the total number of samples). The improvements are particularly important for median and high prevalence classes (from approximately 20-60%).

Decision tree robustness evaluation - Using the three sets with 132 training samples, three decision trees (A, B

and C) were generated; these trees used the same input parameter as the previously built tree, which included at least 12 samples per leaf. Fig. 4 displays the graphical representation of decision trees A, B and C. These trees reached AC values of 65.2%, 62.1% and 65.9%, respectively. These values are similar to the value reached (62.9%) by the general tree built with all 197 samples. In addition, with the exception of tree A (Fig. 4A), the estimated trees presented approximately the same topology as the global tree. These findings allow us to infer a relative robustness of these procedures; the change in initial conditions did not result in major instabilities to the tree generation.

For a better visualisation, Table II shows the variables and groups selected by the general decision tree and by the three other trees (A, B and C).

Note that tree A is quite different from the other two trees. While all of the trees used different variables, several of the variables belonged to the same groups: variables EVIW and ShadeW, for example, were from "the presence of water and dense vegetation during the winter group", and variables Dec and AWater2 were from the "terrain" group.

There are additional similarities among these three trees: variables EVIW and AWater2, selected for the first division of the trees, do not belong to the same variable group, but are highly (0.47) inversely correlated; variables TmaxS and TminW also do not belong to the same groups as variables HDI-91 and HDIE-91, but are highly (0.53 and 0.91) inversely correlated.

Of the six groups of variables, only the variables from the "sanitation" and "presence of water and dense vegetation during the summer" groups were not selected in any of the four trees (Table II). Most likely, these variables were not excluded because they are unimportant, but because they are highly correlated with variables from other groups that better explain the disease.

Comparisons among the three sub-trees (Fig. 4) and the general tree (Fig. 2) show that tree B (Fig. 4B) and tree C (Fig. 4C) are very similar to the general tree. It

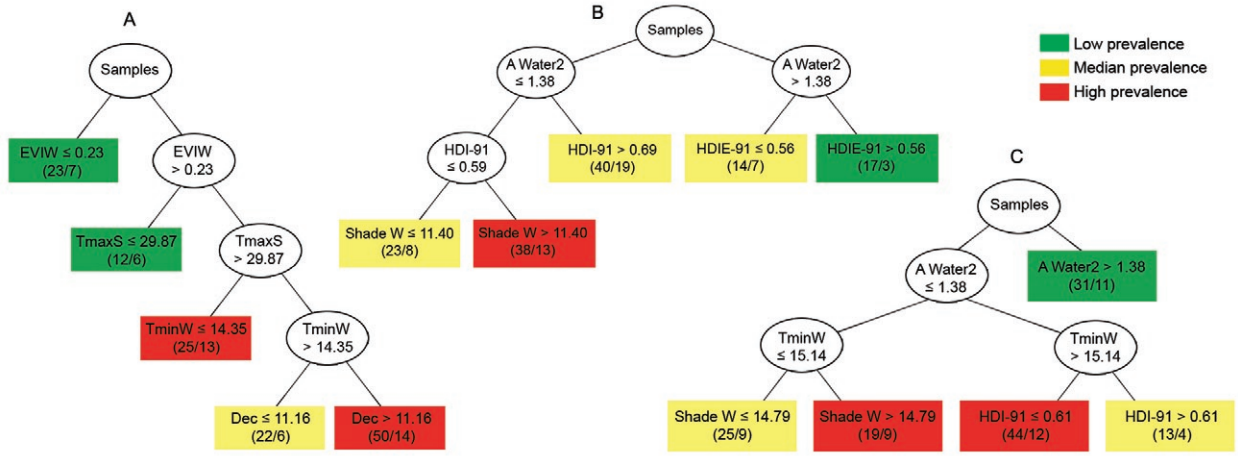

Fig. 4: graphical representation of decision trees A (A), B (B) and C (C). Temperature in ºC. The median of the accumulated water area (AWater2) is a dimensionless counting. Slope declivity (Dec) is given in degrees. The shade in the winter (ShadeW) is given in percentage (%) between 0-100. Human Development Index (HDI) is defined between 0-1. EVIW: enhanced vegetation index in the winter; HDI-91: Human Development Index in year 1991; HDIE-91: Education Human Development Index in year 1991; TmaxS: average of daily maximum temperature in the summer; TminW: average of daily minimum temperature in the winter.

TABLE II

Selected variables

Trees	Selected variables	Groups
General	AWater2, TmaxS, TminW, HDI-91, HDIE-91	Terrain, climate, socioeconomic
A	EVIW, TmaxS, TminW, Dec	Presence of water and dense vegetation during winter, climate, terrain
B	AWater2, HDI-91, ShadeW, HDIE-91	Terrain, socioeconomic, presence of water and dense vegetation during winter
C	AWater2, TminW, ShadeW, HDI-91	Terrain, climate, socioeconomic, presence of water and dense vegetation during winter

AWater2: median of accumulated water; Dec: slope declivity; EVIW: enhanced vegetation index in the winter; HDI-91: Human Development Index in year 1991; HDIE-91: Education Human Development Index in year 1991; ShadeW: shade in the winter; TmaxS: average of daily maximum temperature in the summer; TminW: average of daily minimum temperature in the winter.

is significant that the right branches of both the general tree and tree B are identical and that the central branches of the general tree and tree C differ only because of the presence of TmaxS in the former and TminW in the latter. Such comparisons allow us to demonstrate that, although there are clear differences among the trees, strong similarities among variables can also be observed in each of them. These similarities are either because the variables belong to the same group or because variables from different groups are highly correlated. Table III shows the association of the confusion matrix to the classifications of the three sub-trees.

Similarities can be also observed when comparing the general confusion matrix (Table I) with the A, B and C confusion matrices (Table III). In particular, it is difficult to separate median and high classes.

Table III shows the user accuracy values for the low prevalence classification; tree B has the highest value while trees A and C have lower, but similar, values. The highest user accuracy value for the median prevalence classification is found in tree A. On the other hand, the samples classified as high prevalence showed approximately the same accuracy values in all trees; the highest value was found in tree C.

To evaluate the generalisation potential of trees A, B and C, the samples that were not initially included were classified using the decision rules of each tree. Table IV displays the confusion matrices of the test samples for trees A, B and C and the total confusion matrix that is the sum of the confusion matrices of all trees. As expected, these values are lower than those found in the training sets.

Table IV shows that the highest user accuracy value for the low prevalence classification was found in test matrix B; this finding is similar to that for training data (sub-tree B, Table III). The highest user accuracy value for the high prevalence classification was also found in test matrix B; this finding is different from the matrix of sub-tree B (Table III). Finally, the highest user accuracy value for the median prevalence classification was found in test matrix A (Table IV) and also occurs in the matrix of sub-tree A (Table III).

TABLE III
Confusion matrix of the sub-trees A, B and C

Tree A - Global accuracy (65.2%)			
Class (n)	Classification		
	Low n (%)	Median n (%)	High n (%)
Low (28)	22 (78.57)	0 (0)	6 (21.43)
Median (48)	11 (22.92)	16 (33.33)	21 (43.75)
High (56)	2 (3.57)	6 (10.71)	48 (85.71)
User accuracy (%)	62.86	72.73	64
Tree B - Global accuracy (62.1%)			
Class (n)	Classification		
	Low n (%)	Median n (%)	High n (%)
Low (31)	14 (45.16)	12 (38.71)	5 (16.13)
Median (54)	3 (5.56)	43 (79.63)	8 (14.81)
High (47)	0 (0)	22 (46.81)	25 (53.19)
User accuracy (%)	82.35	55.84	65.79
Tree C - Global accuracy (65.9%)			
Class (n)	Classification		
	Low n (%)	Median n (%)	High n (%)
Low (33)	20 (60.61)	5 (15.15)	8 (24.24)
Median (45)	7 (15.56)	25 (55.56)	13 (28.89)
High (54)	4 (7.41)	8 (14.81)	42 (77.78)
User accuracy (%)	64.52	65.79	66.67

TABLE IV
Confusion matrices of the test set of the sub-trees A, B and C

Test sub-tree A - Global accuracy (38.5%)			
Class (n)	Classification		
	Low n (%)	Median n (%)	High n (%)
Low (18)	6 (33.33)	2 (11.11)	10 (55.56)
Median (25)	2 (8)	5 (20)	18 (72)
High (22)	3 (13.64)	5 (22.73)	14 (63.64)
User accuracy (%)	54.55	41.67	33.33
Test sub-tree B - Global accuracy (50.8%)			
Class (n)	Classification		
	Low n (%)	Median n (%)	High n (%)
Low (15)	10 (66.67)	4 (26.67)	1 (6.67)
Median (19)	1 (5.26)	14 (73.68)	4 (21.05)
High (31)	3 (9.68)	19 (61.29)	9 (29.03)
User accuracy (%)	71.43	37.84	64.29
Test sub-tree C - Global accuracy (40%)			
Class (n)	Classification		
	Low n (%)	Median n (%)	High n (%)
Low (13)	9 (69.23)	1 (7.69)	3 (23.08)
Median (28)	5 (17.86)	4 (14.29)	19 (67.86)
High (24)	1 (4.17)	10 (41.67)	13 (54.17)
User accuracy (%)	60	26.67	37.14
Test A, B and C - Global accuracy (43.1%)			
Class (n)	Classification		
	Low n (%)	Median n (%)	High n (%)
Low (46)	25 (54.35)	7 (15.22)	14 (30.43)
Median (72)	8 (11.11)	23 (31.94)	41 (56.94)
High (77)	7 (9.09)	34 (44.16)	36 (46.75)
User accuracy (%)	62.50	35.94	39.56

The user accuracy for the sum of test matrices A, B and C for the median and high prevalence classifications are approximately 20% different from the user accuracy shown in the main confusion matrix. On the other hand, the user accuracy of the low prevalence classification was 13.11%. This result was expected because of the difficulty of separating the median and high classes in all four trees.

The variables selected by this methodology are those that are normally accepted as important factors for the presence of schistosomiasis (i.e., the ideal habitat for the transmission host and human population life conditions). Some of the variables, such as the HDI, temperature and vegetation indices, were previously used in other investigations focused on using multiple regression analysis to predict the disease. This technique was fairly reliable and is consistent with other explanations of schistosomiasis prevalence. Furthermore, this classification method, expressed in terms of simple decision rules, is usually easy to understand.

DISCUSSION

The results demonstrate the difficulty of separating median and high classes of prevalence. Nevertheless, the improvement in user accuracy was significant, moving from *a priori* estimates of roughly 20% to *a posteriori* estimates of roughly 60% for these two classes. Approximately 63% of the samples were correctly classified with trees built from either all of the samples or two thirds of them. This indicates that, in spite of the rather low AC, the procedure is robust enough that the results do not change considerably when the training samples are varied. This paper and others show that the generalisation of environmental variables at the municipal level has reached its limits. To improve the quality of predictive maps it will be necessary to improve the geo-localisation of training data. Public resources for combating the disease are scarce, but the use of decision rules, which are easy for public policy experts to understand, will allow limited resources to be put to more efficient use.

REFERENCES

Bavia ME, Hale LF, Malone JB, Braud DH, Shane SM 1999. Geographic information systems and the enviromental risk of schistosomiasis in Bahia, Brazil. *Am J Trop Med Hyg 60*: 566-572.

Bavia ME, Malone JB, Hale L, Dantas A, Marroni L, Reis R 2001. Use of thermal and vegetation index data from earth observing satellites to evaluate the risk of schistosomiasis in Bahia, Brazil. *Acta Trop 79*: 79-85.

Carvalho OS, Dutra LV, Moura AC, Freitas CC, Amaral RS, Drummond SC, Freitas CR, Scholte RG, Souza e Guimarães RJ, Melo GR, Ragoni V, Guerra M 2005. *Desenvolvimento de um sistema de informações para o estudo, planejamento e controle da esquistossomose no Estado de Minas Gerais*, Simpósio Brasileiro de Sensoriamento Remoto INPE, Goiânia, 2083-2086.

Carvalho OS, Rocha RS, Massara CL, Katz N 1987. Expansão da esquistossomose mansoni em Minas Gerais. *Mem Inst Oswaldo Cruz 82* (Suppl. IV): 295-298.

Doumenge JP, Mott KE, Cheung C, Villenave D, Chapuis O, Perrin MF, Reaud-Thomas G 1987. *Atlas of the global distribution of schistosomiasis*, Universitaires de Bordeaux Press, Bordeaux, 399 pp.

Freitas CC, Guimarães RJ, Dutra LV, Martins FT, Gouvêa EJ, Santos RA, Moura AC, Drummond SC, Amaral RS, Carvalho OS 2006. *Remote sensing and geographic information systems for the study of schistosomiasis in the state of Minas Gerais, Brazil*, IEEE International Geoscience and Remote Sensing Symposium International Geoscience And Remote Sensing Symposium, Denver, p. 2436-2439.

Guimarães RJ, Freitas CC, Dutra LV, Felgueiras CA, Moura AC, Amaral RS, Drummond SC, Scholte RG, Oliveira G, Carvalho OS 2009. Spatial distribution of *Biomphalaria mollusks* at São Francisco River Basin, Minas Gerais, Brazil, using geostatistical procedures. *Acta Trop 109*: 181-186.

Guimarães RJ, Freitas CC, Dutra LV, Moura AC, Amaral RS, Drummond SC, Scholte RG, Carvalho OS 2008. Schistosomiasis risk estimation in Minas Gerais state, Brazil, using environmental data and GIS techniques. *Acta Trop 108*: 234-241.

Guimarães RJ, Freitas CC, Dutra LV, Moura AC, Amaral RS, Drummond SC, Guerra M, Scholte RG, Freitas CR, Carvalho OS 2006. Analysis and estimative of schistosomiasis prevalence for the state of Minas Gerais, Brazil, using multiple regression with social and environmental spatial data. *Mem Inst Oswaldo Cruz 101* (Suppl. I): 91-96.

IBGE - Instituto Brasileiro de Geografia e Estatística 2008. [homepage on the Internet]. Available from: http://www.ibge.gov.br/home/.

Katz N, Motta E, Oliveira VB, Carvalho EE 1978. *Prevalência da esquistossomose em escolares no estado de Minas Gerais*, XIV Congresso da Sociedade Brasileira de Medicina Tropical, João Pessoa, 102 pp.

Malone JB, Bergquist NR, Huh OK, Bavia ME, Bernardi M, El Bahy MM, Fuentes MV, Kristensen TK, McCarroll JC, Yilma JM, Zhou XN 2001. A global network for the control of snail-borne disease using satellite surveillance and geographic information systems. *Acta Trop 79*: 7-12.

Martins-Bedê FT, Freitas CC, Dutra LV, Sandri SA, Fonseca FR, Drummond IN, Guimarães RJ, Amaral RS, Carvallho OS 2009. Risk mapping of schistosomiasis in Minas Gerais, Brazil, using MODIS and socioeconomic spatial data. *IEEE Transactions on Geoscience and Remote Sensing 47*: 3899-3908.

Mitchell T 1997. *Machine learning,* McGraw Hill, p. 352.

Pellon AB, Teixeira I 1950. *Distribuição da esquistossomose mansônica no Brasil*, Congresso Brasileiro de Higiene, Recife, p. 117.

Quinlan JR 1993. *C4.5*: *Programs for machine learning*, Morgan Kaufmann Publishers, San Francisco, p. 302.

SNIU - Sistema Nacional de Indicadores Urbanos 2005. [database on the Internet]. Available from: http://www.cidades.gov.br/index.php?option=content&task=section&id=49.

Theodoridis S, Koutroumbas K 2006. *Pattern recognition*, Academic Press, p. 885.

Witten IH, Frank E 2005. *Data mining*: *practical machine learning tools and techniques,* 2nd. Morgan Kaufmann, San Francisco, p. 525.