Acta Tropica 133 (2014) 56–63

Contents lists available at ScienceDirect

Acta Tropica

journal homepage: www.elsevier.com/locate/actatropica

Spatial modeling of the schistosomiasis mansoni in Minas Gerais State, Brazil using spatial regression

F. Fonseca[a,c,*], C. Freitas[a], L. Dutra[a], R. Guimarães[a,d], O. Carvalho[b]

[a] Instituto Nacional de Pesquisas Espaciais/INPE, Av. dos Astronautas, 1758 Jd. Granja, CEP 12227-010 São José dos Campos, SP, Brazil
[b] Centro de Pesquisas René Rachou/FIOCRUZ, Av. Augusto de Lima, 1715 Barro Preto, CEP 30190-002 Belo Horizonte, MG, Brazil
[c] Instituto Leônidas e Maria Deane/Fiocruz Amazônia, Rua Terezina, 476 Adrianópolis, CEP 69057-070 Manaus, AM, Brazil
[d] Instituto Evandro Chagas/IEC, Rodovia BR-316 km 7 Levilândia, CEP 67030-000 Ananindeua, PA, Brazil

ARTICLE INFO

Article history:
Received 28 June 2013
Received in revised form
19 December 2013
Accepted 30 January 2014
Available online 9 February 2014

Keywords:
Spatial relations
Generalized proximity matrices
Spatial analysis
Regression analysis
Schistosomiasis mansoni

ABSTRACT

Schistosomiasis is a transmissible parasitic disease caused by the etiologic agent *Schistosoma mansoni*, whose intermediate hosts are snails of the genus *Biomphalaria*. The main goal of this paper is to estimate the prevalence of schistosomiasis in Minas Gerais State in Brazil using spatial disease information derived from the state transportation network of roads and rivers. The spatial information was incorporated in two ways: by introducing new variables that carry spatial neighborhood information and by using spatial regression models. Climate, socioeconomic and environmental variables were also used as co-variables to build models and use them to estimate a risk map for the whole state of Minas Gerais. The results show that the models constructed from the spatial regression produced a better fit, providing smaller root mean square error (RMSE) values. When no spatial information was used, the RMSE for the whole state of Minas Gerais reached 9.5%; with spatial regression, the RMSE reaches 8.8% (when the new variables are added to the model) and 8.5% (with the use of spatial regression). Variables representing vegetation, temperature, precipitation, topography, sanitation and human development indexes were important in explaining the spread of disease and identified certain conditions that are favorable for disease development. The use of spatial regression for the network of roads and rivers produced meaningful results for health management procedures and directing activities, enabling better detection of disease risk areas.

1. Introduction

Schistosomiasis is a transmissible parasitic disease typical of areas with no sanitation or poor sanitation. In Brazil, schistosomiasis is caused by the etiologic agent *Schistosoma mansoni*, whose intermediate hosts are snails of the genus *Biomphalaria* (*B. straminea*, *B. glabrata* or *B. tenagophila*) (Katz and Almeida, 2003), and the prevalence of the disease can reach over 25% in some municipalities. The parasite life cycle is well defined, and the parasite uses water as a support to infect humans (definitive host). Human beings, by contaminating water with feces, infect the snail of genus *Biomphalaria*. Thus, the study of the transmission of schistosomiasis focuses on the combination of environmental characteristics related to human beings and the snail.

Human beings depend on roads to move geographically, and the snail depends on river systems. Thus, to study the spread of schistosomiasis, aspects related to the transport routes of the intermediate and definitive host must be taken into account. Previous works have studied the prevalence of this disease and other diseases that have a direct relationship with the environment by extracting environmental variables using remote sensing data. These studies provided the opportunity to better understand the disease distribution and thereby improve the knowledge about ecological influences (Freitas et al., 2006; Guimarães et al., 2008, 2009, 2010; Martins-Bedé et al., 2009; Scholte et al., 2012; Fonseca et al., 2012).

This paper takes another step in the epidemiologic modeling of schistosomiasis in the state of Minas Gerais, Brazil, using spatial dependence information about the disease obtained from the network of roads and rivers, in addition to variables related to climate, socioeconomic status and the environment.

Aiming to confirm the hypothesis that spatial information improves the estimation of the prevalence of schistosomiasis, models containing spatial information have been proposed and compared with traditional models that do not include this information. Spatial information was incorporated in these models in two ways:

- By proposing two variables that measure neighborhood influence (connectivity through roads and rivers). These variables are inserted into a multiple linear regression model to characterize the spatial dependence associated with schistosomiasis.

* Corresponding author. Tel.: +55 12 3208 6475.
E-mail address: ffonseca@dpi.inpe.br (F. Fonseca).

0001-706X/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.actatropica.2014.01.015

- By using spatial regression models that incorporate spatial dependency into the formula design. The roads and rivers network is also employed in this method.

2. Materials

Socioeconomic and environmental variables, in addition to spatial variables, have been collected to build a model for the spread of disease in the target area. A brief description of the variables used in this work is given in the following subsections.

2.1. Schistosomiasis prevalence data

The schistosomiasis prevalence data were provided by the Secretary of Health in the State of Minas Gerais from historical data collected from 1984 to 2005. The prevalence of schistosomiasis determined in this study was the average number of positive cases of the disease during this period in relation to the population investigated, and at least 80% of the municipality population had to be surveyed. Of the 853 municipalities of Minas Gerais, only 197 municipalities had information about the disease prevalence.

2.2. Environmental data

The environmental data were derived from the moderate resolution imaging spectroradiometer (MODIS) and surface models generated by the shuttle radar topography mission (SRTM). Nine variables were selected from the MODIS products, with the data collected on two different dates, one during the summer (rainy season, from January 17th to February 1st, 2002) and one during the winter (drought season, from July 28th to August 12th, 2002). Two variables came from the SRTM. The MODIS variables were blue, red, near-infrared and middle infrared bands; an enhanced vegetation index; a normalized difference vegetation index; and the vegetation, soil and shade indexes derived from mixture models (Shimabukuro and Smith, 1991). The variables derived from the SRTM were elevation and declivity. In addition, a water accumulation map (which measures the number of possible paths that water can run before reaching a particular point, at each point of a hydrographic basin) was generated from the digital elevation model (Moura et al., 2005). Based on this map, two hydrographic variables were derived: the mean and the median water accumulation inside each municipality. A water mobility index, based on declivity and the average water accumulation (Fonseca et al., 2007), was also calculated for both the rainy season and the drought season. The water mobility index provides an indication of the speed and abundance of water collection.

2.3. Climatic data

The climatic data, accumulated precipitation and maximum and minimum temperatures for the rainy season and drought season were obtained from the Brazilian Center for Weather Forecasting and Climate Studies (CPTEC, Portuguese acronym). The other climate input data were the daily differences between the maximum and minimum temperatures during summer and winter, as proposed by Malone et al. (1994).

2.4. Socioeconomic data

Eighteen socioeconomic variables provided by the Brazilian Institute of Geography and Statistics (IBGE) were used, including data from the human development index (longevity, income, education) for 1991 and 2000; three variables about the water quality in the year 2000, including the percentages of households with access to the main water supply network, access to water through wells or springs and other forms of access to water; and eight variables related to sanitation conditions in the year 2000, including the percentages of households with a toilet connected to a river or lake, a toilet connected to a ditch, rudimentary cesspools, septic tanks, a general sewage network and other types of sewage as well as the percentages with a toilet facility and without a toilet facility. Three other socioeconomic variables were also used in this study, obtained by the João Pinheiro Foundation (FJP, Portuguese acronym) in 2004: the health need index, an index of economic size and an allocation factor of investment resources for health care.

2.5. Spatial data

With the aim of using variables that characterize the mobility of the definitive host (human beings) and the intermediate host (snail), two variables were proposed:

- Connectivity through roads (*CTRoads*) is an indicator of the potential influence of having the disease in a particular municipality provided by the neighbor who has it. This variable refers to the prevalence in the municipality closest to the municipality being investigated divided by the square root of the distance with respect to the paved road network. Eq. (1) below formally describes the *CTRoads* variable, which was calculated for every municipality M_{i}:

$$CTRoads_i = \frac{Prev_e}{\sqrt{DistRoad_{ie}}} \tag{1}$$

where $Prev_e$ corresponds to the prevalence of schistosomiasis in the closest municipality (based on the paved road network) with prevalence information (M_e), and $DistRoad_{ie}$ represents the distance by road between the municipality investigated (M_i) and its nearest neighbor M_e.

- Connectivity through rivers (*CTRivers*) is an indicator of the potential influence of having the disease in a particular municipality, provided by the upriver neighbor who has it. The *CTRivers* variable is the prevalence in the closest upriver municipality to the municipality being investigated divided by the square root of the distance to this municipality, measured over the river network. Formally, the *CTRivers* variable was calculated using the following equation:

$$CTRivers_i = \frac{Prev_r}{\sqrt{DistRiver_{ir}}} \tag{2}$$

where $Prev_r$ corresponds to the prevalence of schistosomiasis in the nearest municipality (based on the river network) with prevalence information (M_r), and $DistRiver_{ir}$ represents the distance by river between the municipality investigated (M_i) and its nearest neighbor M_r. It should be highlighted that according to the definition of neighborhood through rivers (given in Section 3), municipality M_r is always located upriver in relation to the municipality M_i.

The distances (through paved roads and rivers) between municipalities were computed using generalized proximity matrix methodology, proposed by Aguiar et al. (2003) and described in detail in Section 3.1.

3. Methodology

To characterize the distribution of schistosomiasis mansoni and consequently prove that models containing spatial information improve the estimation of the prevalence of the disease in the state of Minas Gerais, models were developed using multiple linear regression (MLR) and spatial regression (SR) and compared with usual MLR models that do not have this information (UMLR). The models produced using MLR and SR incorporate spatial dependence by utilizing the prevalence of schistosomiasis in a nearby municipality, where the neighborhood was determined by the network of roads and rivers, using the generalized proximity matrix methodology (Section 3.1).

The difference in both (MLR and SR) regressions is related to the weighting of neighboring municipalities and the parameter estimation method. In MLR models, spatial dependence is associated with the prevalence in the first municipality divided by the square root of the distance via roads and/or rivers (*CTRoads* and *CTRivers* variables), using the variables described by Eqs. (1) and (2) as independent variables. On the other hand, in models generated with SR, spatial dependence is determined by averaging the disease prevalence in all neighboring municipalities through a network of roads and/or rivers (included parameter for the SR model). The UMLR model uses only socioeconomic and environmental variables to estimate the prevalence of disease.

Two types of networks (roads and rivers), which are exemplified in Fig. 1, were used to build the generalized proximity matrix used in the MLR and SR models. Details about how to build the generalized proximity matrix are given in Section 3.1. Section 3.2 presents the MLR methodology, which is basically estimating the parameters of Eq. (5), with the proposed variables that carry the spatial information. Section 3.3 describes the SR methodology, which is essentially described by Eq. (6).

3.1. *Generalized proximity matrix*

The generalized proximity matrix (GPM) described in Aguiar et al. (2003) is an extension of the spatial weight matrix used in many spatial analytical methods (Bailey and Gattrel, 1995), where the spatial relationships are calculated by taking into account the relationships not only in absolute space (such as Euclidean distance) but also in relative space (transport network). Using the GPM, two geographic objects (municipalities) can be considered close to each other if they are connected through either a transport or telecommunication network, even if they are thousands of kilometers apart.

To study the effects of human mobility and to measure the impact of the snail's presence in disease transmission, two types of transport networks were used to define the neighborhood: paved roads and river networks. The vector data for the network of roads and rivers used in this work, at a 1:1,500,000 scale, were acquired through the State Program for Integrated Use of Geoprocessing (GeoMinas). Fig. 1 illustrates the criteria for the neighborhood and for the construction of the generalized proximity matrix, where paved roads compose the connection network. The spatial objects (M_1, M_2, M_3, M_4, M_5 and M_6) are defined as municipalities, and their respective headquarters are represented by stars.

Six basic parameters are shown in Fig. 1(a–c):

- $d_{Localij}$ – the local distance between municipalities "*i*" and "*j*" (Fig. 1a), which represents the Euclidean distance between the investigated municipality "*i*" itself and its possible neighboring municipality "*j*";
- $d_{maxLocal}$ – the maximum local distance (Fig. 1a): the radius, measured in kilometers, that determines the area or circular region of influence of the investigated municipality;
- $d_{munNeti}$ – the distance from municipality "*i*" to the network (Fig. 1b), which is the Euclidean distance from the investigated municipality headquarter to the closest place via the road network;
- $d_{maxMunNet}$ – the maximum distance from the investigated municipality to the network (Fig. 1b): the distance, measured in kilometers, that determines the area or region of influence of the network;
- d_{Pathij} – the path between the investigated municipality "*i*" and its possible neighboring municipality "*j*" (Fig. 1c): the distance by road between the two municipalities. This distance is calculated as the sum of the Euclidean distance from the municipality "*i*" to the point (P_i) closest to the network ($d_{MunNeti}$), the Euclidean distance from the municipality "*j*" to the point (P_j) closest to the network ($d_{MunNetj}$) and the distance traveled through the network, in kilometers, from P_i to P_j;
- d_{maxNet} – the maximum network distance (Fig. 1c): a threshold value, measured in kilometers, used to identify the neighbors of municipality "*i*".

Eq. (3) defines the neighborhood criteria based on the road network. It is said that the municipality M_j is the neighbor of municipality M_i with respect to the road network if:

$$(((d_{munNeti} < d_{\max MunNet}) \cap (d_{munNetj} < d_{\max MunNet}) \cap (d_{Pathij} < d_{\max Net})) \cup (d_{Localij} < d_{\max Local})) \quad (3)$$

where "*i*" is the index of the investigated municipality and "*j*" is the index of the possible neighboring municipality ($i \neq j$).

As expressed in (3), a municipality is connected to its neighboring municipality by the paved road network (M_j will be neighboring M_i) if the Euclidean distance of the headquarters of the municipalities M_i (and M_j) to the closest road network is less than a specified threshold defined by $d_{maxMunNet}$ and if the distance by road between the two municipalities is less than the specified value indicated by d_{maxNet} or if the municipality M_j is located within the area defined by the maximum local distance from municipality M_i ($d_{maxLocal}$).

The GPM is defined for all municipalities "$\text{M}_{i''}$ that have information on the prevalence of schistosomiasis and that have at least one neighboring municipality. In this study, for the definition of the neighborhood based on the road network, the parameters were set to the following values: $d_{maxLocal}$ = 10 km, $d_{maxMunNet}$ = 30 km and d_{maxNet} = 80 km.

The neighborhood based on the river network was defined using equivalent distance criteria to those presented in (3). However, a new parameter was used to characterize the river flow, defined by the average altitude of the municipality. Based on the network of rivers, it is said that the municipality M_j is a neighbor of municipality M_i if:

$$(((d'_{munNeti} < d'_{\max MunNet}) \cap (d'_{munNetj} < d'_{\max MunNet}) \cap (d'_{Pathij} < d'_{\max Net}) \cap (Altitude_j \geq Altitude_i)) \cup (d'_{Localij} < d'_{\max Local})) \quad (4)$$

where d' is equivalent to d with respect to the river network, "*i*" is the index of the investigated municipality, "*j*" is the index of possible neighboring municipality and $Altitude_i$ indicates the average altitude of the municipality M_i.

As shown in (4), a municipality M_j is said to be the neighbor of M_i via the river network if the average altitude of municipality M_j is higher than or equal to the average altitude of the municipality M_i, if the Euclidean distance of the headquarters of the municipalities M_i and M_j to the closest river network is less than $d'_{\text{max}MunNet}$, if the distance by river between the two municipalities is less than $d'_{\text{max}Net}$, or if the municipality M_j is contained within the area defined by $d'_{\text{max}Local}$. For the definition of the neighborhood based on the river network, the parameters were set to the following values: $d'_{\text{max}Local}$ = 0 km, $d'_{\text{max}MunNet}$ = 10 km and $d'_{\text{max}Net}$ = 70 km.

To construct the GPM based on the network of roads and rivers, the values set for $d_{\text{max}Local}$, $d_{\text{max}MunNet}$ and $d_{\text{max}Net}$ were defined according to the scale applied and the level

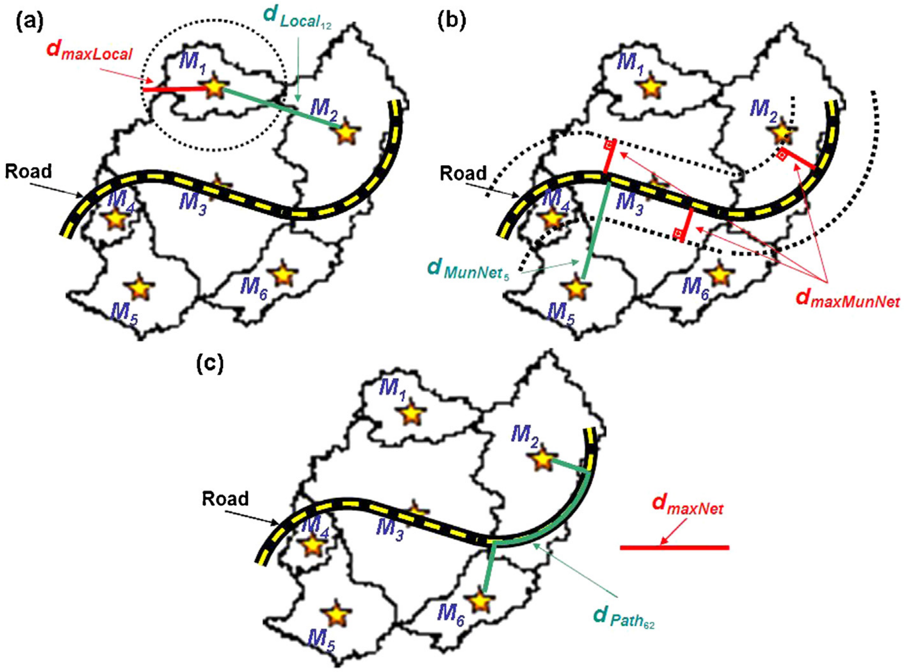

Fig. 1. Schematic of a road network. (a) Local distance between municipalities "i" and "j" ($d_{Localij}$) and maximum local distance ($d_{\max Local}$), (b) distance from municipality "i" to the network ($d_{munNeti}$) and maximum distance from the investigated municipality to the network ($d_{\max MunNet}$), (c) path between the investigated municipality "i" and its possible neighboring municipality "j" (d_{Pathij}) and maximum network distance ($d_{\max Net}$).

of neighborhood influence. The statistical analyses were performed using different values for $d_{\max Local}$, $d_{\max MunNet}$ and $d_{\max Net}$ after determining the ideal distance limits for the study area.

Once the neighbors of each municipality M_i have been defined, three generalized proximity matrices (GPM) can be defined: one using the paved road network neighborhood criterion (W_P), another using the river network neighborhood criterion (W_R), and the last using both network neighborhood criteria simultaneously (W_{PR}). The elements w_{ij} of each GPM are defined by

$$w_{ij} = \begin{cases} \dfrac{1}{n_i}, & \text{if } M_j \text{ is neighbor of } M_i \\ 0, & \text{otherwise} \end{cases}$$

where n_i is the number of neighbors of M_i based on each criterion.

3.2. Development of multiple linear regression models

Regression analysis is a statistical methodology that establishes a relationship between two or more variables such that one of them (dependent variable) can be predicted from the others (independent variables) (Neter et al., 1996). In multiple linear regression, it is assumed that there is a linear relationship between the dependent variable (Y) and K independent variables X_k (k = 1,. . .K). The model is given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5}$$

where $\boldsymbol{Y}$ is the vector representing the dependent variable; $\boldsymbol{X}$ is the matrix with the values of the independent variables; $\boldsymbol{\beta}$ is a vector of the regression parameters; and $\boldsymbol{\varepsilon}$ is a vector representing the random independent errors with zero mean and constant variance.

In the present work, the dependent variable is a function of the disease prevalence, and the independent variables are the socioeconomic, environmental and spatial variables described in Section 2. To distinguish between the models that use only socioeconomic and environmental variables and the one that also uses the spatial variables given in Eqs. (1) and (2), they are called the usual MLR (UMLR) and the MLR, respectively.

To establish a relationship between the dependent and independent variables, exploratory analyses of the data and the model performance were conducted by assessing dispersion, outliers, normality and the variance of the regression's residuals. In this work, the steps used to develop the predictive models from MLR were as follows:

- Data collection and preparation of the independent variables. In this step, all of the independent variables were correlated with the dependent variable. To improve the correlation, transformations were tested on the dependent and independent variables (e.g., quadratic, inverse, logarithmic, square root, etc.) to normalize the variables and linearize the relationship between the dependent variable and the independent ones. On this basis, the transformation $Ln\,(Prev + 1)$ was selected and used as the dependent variable in this work.
- Reduction of the number of independent variables. The correlations of the independent variables were analyzed in detail, and the variables with a low correlation with the dependent variables or with a high correlation with other independent variables were discarded.
- All possible regression models were generated, and R^2, adjusted R^2 and Mallow's Cp criteria were applied.
- Based on the results of the previous step, some likely regression models were selected.
- Refinement and selection of the best model. Among the possible models, the best was selected by performing a residual analysis of residues, which includes tests for normality, heteroscedasticity and autocorrelation as well as tests for the identification of outliers;
- Validation of the model. Finally, a validation of the selected model was carried out, using the RMSE (root mean square errors) based on a new set of samples (already separated initially).

3.3. Spatial regression models

Spatial regression models (SR) are regression models in which the spatial dependency is embedded in the model. Basically, the spatial effect can be incorporated in two ways: locally and globally (Fotheringham et al., 2000; Anselin, 2002). The local models capture the spatial structure through parameters that vary in space because of the existence of several spatial patterns. The global models capture the spatial structure through a single parameter that is added to the usual regression model. The global models can be called spatial error or spatial lag models, depending on whether spatial dependence is embedded in the term relative to the error or in the dependent variable.

The present study used *spatial lag*, which considers that the value of the dependent variable in a certain location is correlated with the observations in neighboring localities. This model is expressed by the equation:

$$\boldsymbol{Y} = \rho W Y + \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (6)$$

where $\boldsymbol{Y}$ is the vector representing the dependent variable; W is a matrix of spatial weights; WY, which expresses the spatial dependence in Y, is a new explanatory variable added to the regression model that represents the weighted average of the dependent variable in neighboring areas; ρ is the spatial autoregressive parameter, such that if $\rho = 0$, there is no autocorrelation and the model becomes equivalent to the usual linear regression model expressed in (5); $\boldsymbol{X}$ is the matrix with the values of the independent variables; $\boldsymbol{\beta}$ is a vector of the regression parameters, and $\boldsymbol{\varepsilon}$ is a vector representing the random independent errors with zero mean and constant variance.

Two models were developed using SR. The first model generated was based on the generalized proximity matrix through the paved road network (W_P), and W_PY is the vector of the means of the prevalence logarithm of the disease in all neighboring municipalities connected to the network through paved roads. The second model was developed from the generalized proximity matrix W_{PR} based on the junction of the paved road network and the river network, and $W_{PR}Y$ is the vector of the means of the logarithm of the prevalence of the disease in all neighboring municipalities connected to the network through paved roads and rivers.

The models were evaluated by determining whether the residuals demonstrated spatial autocorrelation. The best-fitting model was further verified using the Akaike information criterion (AIC). The AIC is a performance measure used to analyze spatially correlated data. The lower the AIC, the better the model fits.

4. Results and discussion

For the generation and estimation of the predictive models using multiple linear regression and spatial regression, the study area was divided into three groups based on the network of roads and rivers.

From the 853 municipalities in the state of Minas Gerais, Brazil, there are 388 municipalities that are connected to neighbors (with prevalence information) by roads (group 1), 113 municipalities that are connected to neighbors (with prevalence information) by both roads and rivers (group 2) and 352 municipalities that are not connected to neighbors (with prevalence information) by roads or by rivers (group 3). A regression model was generated for each group.

The division of the state of Minas Gerais, Brazil into three groups is presented below (Fig. 2).

The multiple linear regression model (MLR) and the spatial regression model (SR) were estimated for groups 1 and 2 and compared to assess which of the two methodologies with spatial information better estimates the prevalence of schistosomiasis mansoni. The UMLR was estimated considering all municipalities with prevalence information and applied to group 3, which has no spatial information. This UMLR model was also applied to the other groups for comparison purposes.

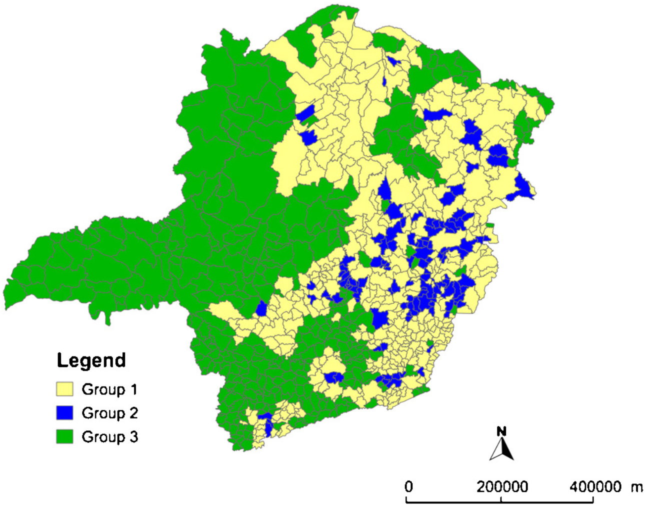

Fig. 2. The state of Minas Gerais divided into three groups: municipalities that have neighbors with prevalence information and are connected only by roads (group 1), municipalities that have neighbors with prevalence information and with connectivity by both roads and rivers (group 2); and municipalities that have neighbors with prevalence information that are not connected either by roads neither by rivers (group 3).

In group 1, the model derived using multiple linear regression obtained an R^2 value of 42% ($R = 65\%$) and included the following variables: the median water accumulation (*Ac*), winter precipitation (*Prec_w*), the minimum temperature during the summer (*Tmin_s*), and connectivity by roads (*CTRoads*). Eq. (7) shows the model generated by multiple linear regression for group 1.

$$\widehat{Prev} = \exp(-2.84 + 1.19(\sqrt{CTRoads}) - 7.07(\ln(Ac+1)) + 0.22(\ln(Prec\_w+1)) + 2.38(\sqrt{T\min\_s})) - 1 \qquad (7)$$

The model produced by spatial regression for group 1 obtained an AIC = 3.78 and a spatial dependence value of $\rho = 0.29$. Eq. (8) shows the model generated using spatial regression for group 1.

$$\widehat{Prev} = \exp(0.29W_PY - 3.39 - 5.63(\ln(Ac+1)) + 0.18(\ln(Prec\_w+1)) + 2.14(\sqrt{T\min\_s})) - 1 \qquad (8)$$

where W_PY is the weighted average of the natural logarithm of the prevalence values in the neighboring municipalities, using the paved roads network as the criterion for the construction of the proximity matrix.

Based on Eqs. (7) and (8), there is a negative relationship between the disease prevalence and the variable *Ac*. Thus, the smaller the area of water accumulation, the more concentrated the snails become and consequently, the higher the disease prevalence becomes. The variable *Prec_w* has a direct relationship to the disease, which means that greater accumulated precipitation in the dry season (winter) is associated with higher disease prevalence. The prevalence is also directly related to *Tmin_s*, that is, the higher the minimum temperature in the summer, the greater the disease distribution. This relationship can be explained by analyzing the life cycle of the disease: in general, the snails release the parasite larvae in the water during the hottest periods of the day, stimulated by intense sun light and the average temperature of 28 °C. Furthermore, it was observed that there was a direct correlation between the disease prevalence and the variable *CTRoads*, presented in (7), showing that higher prevalence rates in neighboring municipalities that are connected by the road network are associated with a greater prevalence of schistosomiasis in the municipality investigated. Analyzing the autoregressive parameter (ρ), present in (8), it is observed that there is spatial dependence at 0.29, which is related to the mobility of human beings among the municipalities, defined by $W_P Y$.

In group 2, the selected variables in the multiple linear regression model were the percentage of households with a toilet connected to a ditch (*TD*), the human development index of longevity in the year 2000 (*HDIL_00*), the water mobility index in the summer (*MI_s*), connectivity through roads (*CTRoads*), and connectivity through rivers (*CTRivers*). Eq. (9) shows the model generated by multiple linear regression for group 2, which reached a R^2 value of 48% ($R = 69\%$).

$$\hat{Prev} = \exp(3.99 + 1.53(\sqrt{CTRoads}) + 0.85(\sqrt{CTRivers}) - 0.004((TD)^2) - 3.49((HDIL\_00)^2) - 0.001((MI\_s)^2)) - 1 \qquad (9)$$

The model generated for group 2 using spatial regression showed an AIC = 3.59 and more pronounced spatial dependence than the model for group 1 ($\rho = 0.51$). This result is due to the increase in the possible interconnections between the municipalities, either by roads or by rivers. These results show that the higher the (lower) prevalence in the neighboring municipalities connected by roads and rivers, the greater (smaller) the prevalence in the observed municipality will be. Eq. (10) shows the model generated using spatial regression for group 2.

$$\hat{Prev} = \exp(0.51 W_{PR} Y + 3.17 - 0.004((TD)^2) - 3.08((HDIL\_00)^2) - 0.002((MI\_s)^2)) - 1 \qquad (10)$$

where $W_{PR} Y$ is the weighted average of the natural logarithm of the prevalence in the neighboring municipalities, using the network of roads and rivers as the criterion for constructing the proximity matrix.

As Eqs. (9) and (10) show, the variable *TD* has an inverse relationship to the disease, which means that lower percentages of households with a toilet connected to a ditch are associated with greater prevalence of the disease. In other words, the fewer households with basic sanitation systems, the greater the risk of contamination with the disease will be. We also observed an inverse relationship between *HDIL*_00 and the prevalence of schistosomiasis, indicating that the longevity of people living in places with a low prevalence of schistosomiasis is higher than in places with a high prevalence.

The variable *MI_s* also had an indirect relationship with the disease, coinciding with a study (Fonseca et al., 2007) that indicated that the dispersal and spread of the snails during the rainy season (summer), followed by periods of drought (winter) in stable habitats, low water flow, and easy human access to infected waters associated with a lively snail population favor the life cycle and transmission of *Schistosoma mansoni*. Based on Eq. (9), higher values of *CTRoads* and *CTRivers* are associated with higher disease prevalence. Thus, the higher the prevalence rates in the neighboring municipalities connected through the roads and rivers, the higher the disease prevalence will be in the municipality investigated.

For group 3, it was not possible to generate a spatial regression model because these municipalities do not have neighbors with prevalence information, and therefore, it is not possible to include spatial dependence, human mobility on the roads or snail movement along the rivers in the model. Thus, a multiple linear regression model with no spatial component was generated, which included the following variables: the percentage of households with a toilet connected to a river or lake (*TRL*), the median water accumulation (*Ac*), the fractional vegetation index in the winter (*Veg_w*), the minimum temperature in the summer (*Tmin_s*) and the interaction of the declivity of the terrain (*Dec*) with the fractional vegetation index in the winter (*Veg_w*) (Martins-Bedé et al., 2009). Eq. (11) shows the model generated using multiple linear regression for group 3.

$$\hat{Prev} = \exp(-3.30 + 0.01(TRL) - 2.54(Ac) + 0.07(Veg\_w) - 0.003(Dec * Veg\_w) + 0.37(T\min\_s)) - 1 \qquad (11)$$

With the intention of comparing the results from the models generated using spatial regression and those produced using multiple linear regression, with and without spatial information, Table 1 presents the root mean squared errors (RMSE) of the models obtained with spatial information (MLR and SR) and without this information (UMLR) and their respective numbers of samples/municipalities (*n*) with prevalence information. The table's second and third columns provide the values of the RMSE of the models produced by MLR and SR (with spatial information) applied to municipalities in groups 1 and 2. For the municipalities in group 3, which do not have neighbors connected by rivers and/or roads, the model without spatial information (Eq. (11)) was used. The RMSE using the UMLR for all groups is presented in the fourth column of Table 1.

Columns 5, 6 and 7 show, respectively, the percent improvement of the SR model in comparison to the MLR, of the MLR in comparison to the UMLR, and of the SR in comparison to the UMLR.

Comparing the results of the three models, the SR model provides better results with smaller RMSE values. The greatest improvement was observed for the municipalities of group 2 when spatial information was used, including the connectivity through both roads and rivers. The RMSE was 18% lower than that for the model without spatial information. For all 197 municipalities with prevalence information, the RMSE using the SR was almost 10% lower compared with the model that did not use spatial information and showed 2.5% improvement compared with the MLR model with spatial information. It was also observed that the MLR model with spatial information was 7% better compared with the MLR model without this information. These results confirm the importance of introducing spatial information (i.e., the connectivity of roads and rivers) and the superiority of the spatial regression model with respect to multiple linear regression models.

Table 1
Comparison of the models generated by SR, MLR, and UMLR without spatial information.

	N	RMSE			Improvement rate (%)		
		MLR (A)	SR (B)	UMLR – without spatial information (C)	$\frac{A-B}{A}$	$\frac{C-A}{C}$	$\frac{C-B}{C}$
Municipalities of group 1	109	9.35	8.98	9.92	3.99	5.72	9.49
Municipalities of group 2	60	6.73	6.72	8.19	0.28	17.76	17.99
Municipalities of group 3	28	10.19	10.19	10.19	0.00	0.00	0.00
Total of municipalities	197	8.77	8.55	9.47	2.54	7.33	9.69

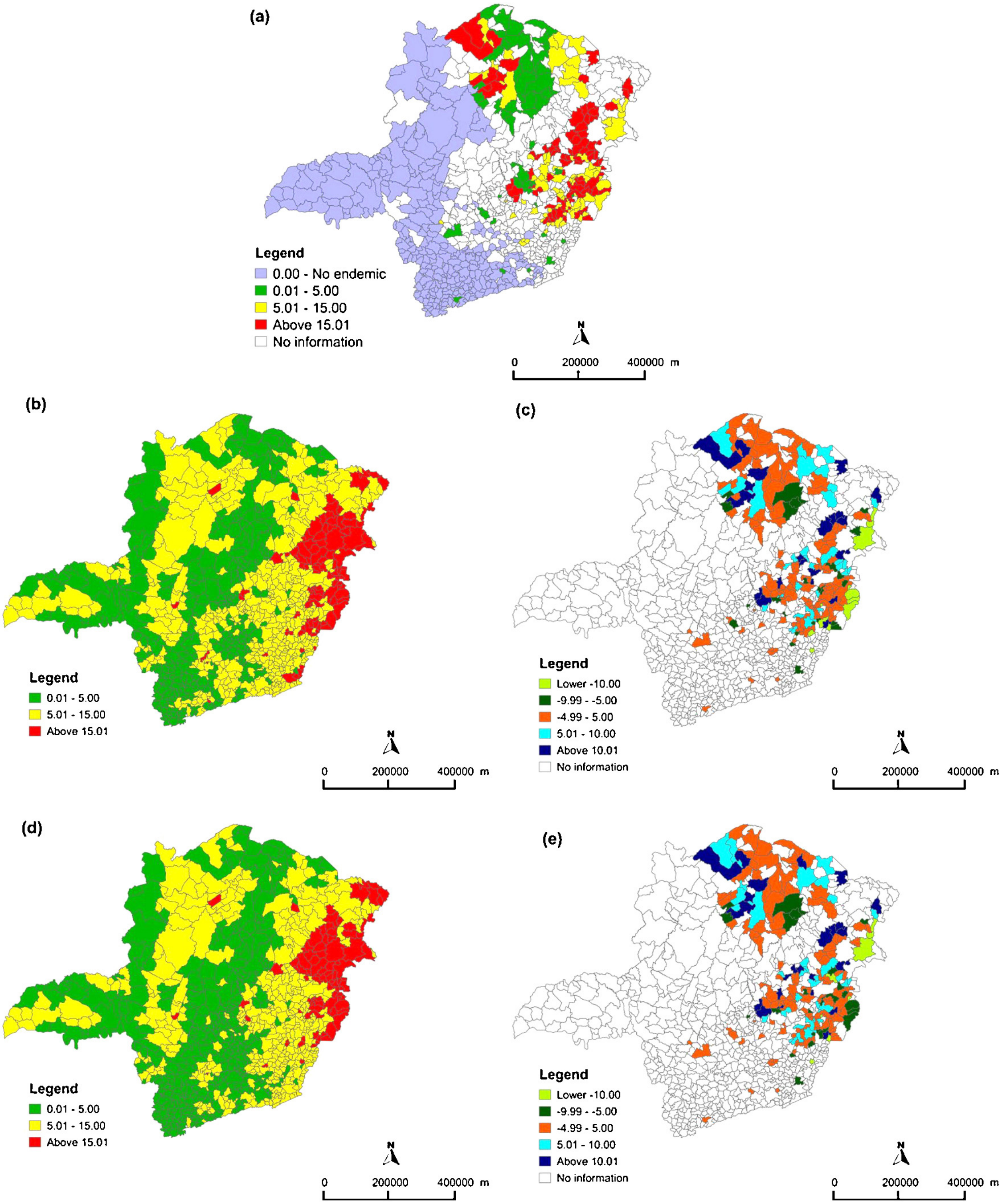

Fig. 3. Estimation of disease prevalence for the entire state of Minas Gerais by MLR and SR: (a) observed prevalence, (b) estimated prevalence by MLR, (c) residuals by MLR, (d) estimated prevalence by SR, (e) residuals by SR.

The accuracy of the models generated by MLR and SR was assessed by examining the distribution of the residuals and the estimates of the disease prevalence for the entire state of Minas Gerais. Fig. 3(a) shows the distribution of the observed prevalence, Fig. 3(b and c) shows the estimated prevalence according to the MLR and the corresponding residuals, and Fig. 3(d and e) shows the estimated prevalence and the corresponding residuals obtained with the SR model.

The accuracy of the models can be evaluated with Fig. 3(c and e), which shows where the models are more accurate (the municipalities in orange), where the predicted values are overestimated (the municipalities in light green and dark green) and where the predicted values are underestimated (the municipalities in light blue and dark blue).

The prevalence estimates in the western region (a non-endemic region, known as "Triângulo Mineiro") are estimated accurately, with most of the values smaller than 5%. The estimated prevalence did not reach high values in any of the municipalities considered non-endemic.

The matrix of neighbors connected by roads and rivers, used in the spatial regression models, and the variables *CTRoads* and *CTRivers*, used in the multiple linear regression models, reflected the movement of men and snails in neighboring municipalities along roads and rivers, respectively.

5. Conclusion

The results showed that the mobility of *S. mansoni* hosts is an important factor to consider in modeling and estimating the prevalence of schistosomiasis. Variables representing vegetation, temperature, precipitation, topography, sanitation and human development indexes were important in explaining the spread of disease and determining the favorable conditions for disease development. The use of multiple linear regression and spatial regression provided meaningful results for health management procedures and related activities, enabling better detection of disease risk areas. The methods can be expanded to other studies regarding disease vectors that display spatial dependence and whose hosts depend on some transportation network for dispersal.

Acknowledgments

This work was supported by CNPq, CAPES, FAPESP and FAPEMIG, Brazil. The authors also thank A.P.D. Aguiar for providing the first version of the software developed for the present analysis.

References

Aguiar, A.P.D., Câmara, G., Monteiro, A.M.V., Souza, R.C.M., 2003. Modeling spatial relations by generalized proximity matrices. In: Proceedings of the V Brazilian Symposium on Geoinformatics, Campos do Jordão, Brazil.

Anselin, L., 2002. Under the hood: issues in the specification and interpretation of spatial regression models. Agric. Econ. 27, 247–267.

Bailey, T., Gattrel, A., 1995. Spatial Data Analysis by Example. Longman, London.

Fonseca, F.R., Saraiva, T.S., Freitas, C.C., Dutra, L.V., Monteiro, A.M.V., Renno, C.D., Martins, F.T., Guimarães, R.J.P.S., Moura, A.C.M., Scholte, R.G.C., Amaral, R.S., Drummond, S.C., Carvalho, O.S., 2007. Desenvolvimento de um índice hidrológico para aplicação em estudos de distribuição da prevalência de esquistossomose em Minas Gerais. In: Proceedings of the XIII Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, Brazil, pp. 2589–2595.

Fonseca, F.R., Freitas, C.C., Dutra, L.V., 2012. Mapping schistosomiasis mansoni in the state of Minas Gerais, Brazil, using spatial regression. In: Proceedings of the International Geoscience and Remote Sensing Symposium, Munich, Germany, pp. 7228–7231.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2000. Quantitative Geography: Perspectives on Spatial Data Analysis. Sage, London.

Freitas, C.C., Guimarães, R.J.P.S., Dutra, L.V., Martins, F.T., Gouvêa, E.J.C., Santos, R.A.T., Moura, A.C., Drummond, S.C., Amaral, R.S., Carvalho, O.S., 2006. Remote sensing and geographic information systems for the study of schistosomiasis in the state of Minas Gerais, Brazil. In: Proceedings of the International Geoscience and Remote Sensing Symposium, Denver, USA, pp. 2436–2439.

Guimarães, R.J.P.S., Freitas, C.C., Dutra, L.V., Moura, A.C.M., Amaral, R.S., Drummond, S.C., Scholte, R.G.C., Carvalho, O.S., 2008. Schistosomiasis risk estimation in Minas Gerais State, Brazil, using environmental data and GIS techniques. Acta Trop. 108, 234–341.

Guimarães, R.J.P.S., Freitas, C.C., Dutra, L.V., Felgueiras, C.A., Moura, A.C.M., Amaral, R.S., Drummond, S.C., Scholte, R.G.C., Oliveira, G.C., Carvalho, O.S., 2009. Spatial distribution of *Biomphalaria* mollusks at São Francisco River Basin, Minas Gerais, Brazil, using geostatistical procedures. Acta Trop. 109, 181–186.

Guimarães, R.J.P.S., Freitas, C.C., Dutra, L.V., Scholte, R.G.C., Martins, F.T., Fonseca, F.R., Amaral, R.S., Drummond, S.C., Felgueiras, C.A., Oliveira, G.C., Carvalho, O.S., 2010. A geoprocessing approach for schistosomiasis studying and control in the state of Minas Gerais – Brazil. Mem. Inst. Oswaldo Cruz 105, 524–531.

Katz, N., Almeida, K., 2003. Esquistossomose, xistosa, barriga d'água. Ciência e Cultura 55, 38–55.

Malone, J.B., Huh, O.K., Fehler, D.P., Wilson, P.A., Wilensky, D.E., Holmes, R.A., Elmagdoub, A.I., 1994. Temperature data from satellite imagery and the distribution of schistosomiasis in Egypt. Am. J. Trop. Med. Hyg. 50, 714–722.

Martins-Bedé, F.T., Freitas, C.C., Dutra, L.V., Sandri, S., Drummond, I.N., Fonseca, F.R., Guimarães, R.J.P.S., Amaral, R.S., Carvalho, O.S., 2009. Risk mapping of schistosomiasis in Minas Gerais, Brazil, using MODIS and socioeconomic spatial data. IEEE Trans. Geosci. Remote Sens. 47, 3899–3908.

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., 2005. Atualização de mapa de drenagem como subsídio para montagem de SIG para a análise da distribuição da esquistossomose em Minas Gerais. In: Proceedings of the XII Simpósio Brasileiro de Sensoriamento Remoto, Goiânia, Brazil, pp. 3551–3558.

Neter, J., Kutner, M.H., Nachtssheim, C.J., Wasserman, W., 1996. Applied Linear Statistical Models. McGraw-Hill, Boston.

Scholte, R.G.C., Freitas, C.C., Dutra, L.V., Guimaraes, R.J.P.S., Drummond, S.C., Oliveira, G., Carvalho, O.S., 2012. Utilizing environmental, socioeconomic data and GIS techniques to estimate the risk for ascariasis and trichuriasis in Minas Gerais, Brazil. Acta Trop. 121, 112–117.

Shimabukuro, Y.E., Smith, J.A., 1991. The least-square mixing models to generate fraction images derived from remote sensing multispectral data. IEEE Trans. Geosci. Remote Sens. 29, 16–20.