Spatial modeling using mixed models: an ecologic study of visceral leishmaniasis in Teresina, Piauí State, Brazil

Modelagem espacial utilizando modelos mistos: um estudo ecológico sobre leishmaniose visceral em Teresina, Piauí, Brasil

Guilherme L. Werneck [1,2]
James H. Maguire [3]

[1] *Núcleo de Estudos de Saúde Coletiva, Universidade Federal do Rio de Janeiro. Centro de Ciências da Saúde, Bloco K, Cidade Universitária, C. P. 68037, Rio de Janeiro, RJ 21941-590, Brasil. gwerneck@nesc.ufrj.br*
[2] *Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro. Rua São Francisco Xavier 524, Rio de Janeiro, RJ 20559-900, Brasil.*
[3] *Department of Immunology and Infectious Diseases, Harvard School of Public Health. 665 Huntington Ave., SPH1, Room 711, Boston, MA 02115, U. S. A.*

Abstract *Most ecologic studies use geographical areas as units of observation. Because data from areas close to one another tend to be more alike than those from distant areas, estimation of effect size and confidence intervals should consider spatial autocorrelation of measurements. In this report we demonstrate a method for modeling spatial autocorrelation within a mixed model framework, using data on environmental and socioeconomic determinants of the incidence of visceral leishmaniasis (VL) in the city of Teresina, Piauí, Brazil. A model with a spherical covariance structure indicated significant spatial autocorrelation in the data and yielded a better fit than one assuming independent observations. While both models showed a positive association between VL incidence and residence in a favela (slum) or in areas with green vegetation, values for the fixed effects and standard errors differed substantially between the models. Exploration of the data's spatial correlation structure through the semivariogram should precede the use of these models. Our findings support the hypothesis of spatial dependence of VL rates and indicate that it might be useful to model spatial correlation in order to obtain more accurate point and standard error estimates.*

Key words *Visceral Leishmaniasis; Spatial Analysis; Ecologic Studies; Epidemiology*

Resumo *A maioria dos estudos ecológicos utiliza áreas geográficas como unidades de observação. Uma vez que as áreas geograficamente próximas tendem a ser mais semelhantes do que as distantes, as estimativas da magnitude do efeito e dos intervalos de confiança devem levar em conta a auto-correlação espacial das medidas. Neste estudo demonstramos um método para modelar a auto-correlação espacial dentro de um referencial de modelo misto, utilizando dados sobre determinantes ambientais e sócio-econômicos da incidência de leishmaniose visceral (LV) na cidade de Teresina, Estado do Piauí. Um modelo com uma estrutura de covariância esférica indicou uma auto-correlação espacial significativa nos dados e produziu melhor ajuste quando comparado com outro modelo que pressupunha observações independentes. Embora ambos modelos tenham demonstrado associações positivas entre incidência de LV e residência em favelas ou áreas com vegetação verde, os valores para os efeitos fixos e erros-padrão diferiram substancialmente entre os modelos. A estrutura da correlação espacial dos dados deve ser explorada através do semivariograma, antes da utilização destes modelos. Nossos achados favorecem a hipótese da dependência espacial dos coeficientes de incidência de LV e sugerem que a modelagem da correção espacial poderia ser útil para obter estimativas pontuais e de erros-padrão mais acuradas.*

Palavras-chave *Leishmaniose Visceral; Análise Espacial; Estudos Ecológicos; Epidemiologia*

Introduction

In ecologic studies, geographical areas are the usual units of observation, data on outcome are expressed as incidence rates, and data on explanatory variables may include aggregate, environmental or global measures (Morgenstern, 1998). Data taken within specific regions (areal data) typically possess spatial structure, in the sense that observations closer together tend to be more alike than observations farther apart (Cressie, 1991). Accordingly, areal data can be considered a two-dimensional counterpart of time series data, in which observations are correlated in the single dimension of time. As in the analysis of time series, it is important to model the spatial correlation structure among observations in order to obtain valid estimates of effect size, confidence intervals, and significance levels (Cressie, 1991).

In this paper we describe a strategy for modeling spatial covariance structure in ecologic studies within a mixed model framework. The observed data in mixed models consist of fixed effects, which define the expected values of the observations, and random effects, which define the variance and covariance of the observations (Littell et al., 2000). Since errors in mixed models for spatial data are correlated, spatial covariance is modeled through the error term. As an illustration of these methods, we present data of an ecologic study of environmental and socioeconomic determinants of the incidence of visceral leishmaniasis (VL) in Teresina, Piauí, Brazil.

Methods

Study area

Teresina, capital of the State of Piauí, was the site of Brazil's first urban epidemic of VL in 1980-1985 (Costa et al., 1990). In a second epidemic, from 1992 to 1996, more than 1,200 cases were reported among a population of 650,000. Factors that favor transmission of *Leishmania chagasi* by the sand fly vector *Lutzomyia longipalpis* include the city's tropical climate and vegetation. Grass, shrubs, and sparse mango and palm trees are found throughout the city, and tropical forest and farmland surround the urban periphery.

Data on visceral leishmaniasis

The age, date of diagnosis, and geographic location of the residence of 1,061 persons with VL in Teresina between 1993 and 1996 were obtained from the National Health Foundation (Fundação Nacional de Saúde – FUNASA) and confirmed from clinical and laboratory records from all hospitals in Teresina. This figure represents approximately 95% of the total number of cases reported to FUNASA during this period. It is likely that few cases of VL were overlooked, since there is no alternative center for treating the disease close to Teresina, and, by law, all suspected and confirmed cases are reported to FUNASA, which is the sole distributor of anti-leishmanial drugs in Brazil.

VL incidence rates were calculated for each of the city's 494 census tracts using data from the 1991 and 1996 national censuses. For the analysis, the original census tracts were consolidated into 430 areas ("consolidated census tracts") so that at least one case of VL would be expected in each tract had cases been distributed uniformly throughout the city. Each of the 64 census tracts with less than one expected case of VL was joined to an adjoining census tract with a similar socioeconomic profile. Census tracts were considered neighbors if they shared a common boundary. Similarity of socioeconomic profiles was based on a score derived by principal components analysis (SAS – SAS Institute, 1996) of census data on household characteristics such as running water, indoor plumbing, garbage collection, level of schooling, and family income. Geographical coordinates indicating the longitude and latitude of its centroid identify each census tract.

Explanatory variables

From the 1991 census data, each consolidated census tract was characterized as consisting of a slum (*favela*) or non-slum as a proxy for its socioeconomic status (SES).

Landscape features were identified by remote sensing (RS) using a Landsat 5 Thematic Mapper (TM) scene (6 bands, 30m resolution) of Teresina from October 1995. Digital maps of the consolidated census tracts were produced using CartaLinx (The Clark Labs, 1998). IDRISI software (The Clark Labs, 1997) was used to overlay the digital map on the RS image to extract the land cover information for each consolidated census tract. Environmental features were also characterized by the Normalized Difference Vegetation Index (NDVI). The NDVI is a widely used vegetation index in remote sensing and is defined as (Hay et al., 1996):

$$NDVI = (Ch2 - Ch1) / (Ch2 + Ch1)$$

where Ch1 is the reflectance from each pixel in the red wavelength band (Landsat band 3) and Ch2 is the reflectance in the near-infrared wavelength band (Landsat band 4). NDVI varies from -1.0 to +1.0 with positive values generally indicating green vegetation, and negative values indicating lack of green vegetation. NDVI correlates positively with rainfall and humidity, factors that are related to sand fly abundance (Hay et al., 1996; Thomson et al, 1997). In this study we determined the mean NDVI over the pixels in each consolidated census tract.

Statistical analysis

- ### Modeling the spatial covariance structure

A general model for our data can be conceptualized as follows:

$$LINC_i = \beta_0 + \beta_1 NDVI_i + \beta_2 SES_i + e_i \quad (1)$$

The natural logarithm of the VL incidence rates for the i^{th} consolidated census tract ($LINC_i$) is the continuous outcome variable, the explanatory variables are $NDVI_i$ and SES_i for each i^{th} consolidated census tract, and e_i is the random error. Unlike inference from the ordinary least squares regression, inference from this model cannot assume an independent error structure because of spatial autocorrelation. Accordingly, we employ a mixed linear model in which spatial autocorrelation is modeled through the error term, and the data are allowed to exhibit correlation and heteroscedasticity, thereby generalizing the standard linear model. The mixed model framework permits modeling of not only the fixed-effects parameters β, but their variances (*Var*) and covariances (*Cov*) as well.

In general, spatial correlation models can be defined by letting (Littell et al., 1996):

$$Var(e_i) = \sigma_i^2 \text{ and } Cov(e_i,e_j) = \sigma_{ij} \quad (2)$$

Let the spatial location of $LINC_i$ be expressed by s_i, which is specified by the two coordinates, latitude and longitude. The covariance is assumed to be a function of the distance (d_{ij}) between locations s_i and s_j, and has the general form (Littell et al., 1996):

$$\mathrm{Cov}(e_i,e_j) = \sigma^2[f(d_{ij})] \quad (3)$$

We have chosen to model $f(d_{ij})$ by using the spherical function, available in the procedure MIXED of SAS software:

$$f(d_{ij}) = [1\text{-}1.5(d_{ij}/\rho)+0.5(d_{ij}/\rho)^3]1(d_{ij}<\rho) \quad (4)$$

This model indicates that the degree of correlation decreases as the distance between two observations increases, but it may not adequately account for abrupt changes over relatively small distances (Littell et al., 1996). It is possible to model these changes by adding an additional parameter σ_l^2, the *nugget*. The resulting covariance models with a nugget effect are:

$$Var(e_i) = \sigma^2 + \sigma_l^2 \text{ and } Cov(e_i,e_j) = \sigma^2[f(d_{ij})] \quad (5)$$

Conveniently, some parameters involved in models (4) and (5) correspond to parameters described by the semivariogram, the standard statistical measure of spatial variability as a function of the distance between observations (Cressie, 1991). In this model the parameters σ_l^2, $\sigma^2 + \sigma_l^2$, and ρ correspond to the geostatistics parameters nugget, sill, and range, respectively (Cressie, 1991). The *nugget* represents micro-scale variation or measurement error. The *sill* corresponds to the variance of the random field. The *range* is defined as the distance at which the semivariogram reaches the sill. For distances less than the range, observations are spatially correlated. For distances greater than or equal to the range, spatial correlation is effectively zero.

SAS PROC MIXED does not compute the semivariogram *per se*, but estimates of the range, sill, and nugget from other software packages can facilitate working with SAS PROC MIXED, as described below (Littell et al., 1996).

Estimates of the variance and covariance of these models are obtained through a Restricted Maximum Likelihood (REML) approach, and estimates of βs are obtained through solutions to mixed model equations (Littell et al., 1996).

- ### Modeling strategy

The first strategy in the fitting process was to explore the empirical semivariogram of the residuals of model described in (1), and to fit a model for it using S+SPATIALSTATS (Figure 1). The semivariogram detects spatial dependence in the residuals and provides estimates for the sill, nugget, and range, for use within SAS PROC MIXED. Because most models for spatial covariance structure require that the assumption of isotropy (spatial dependence is the same in all directions) (Cressie, 1991), we next estimated empirical semivariograms for 4 different directions (Figure 2). Correlation structures were fairly similar in all 4 directions, suggesting that the assumption of isotropy holds for the VL data.

Figure 1 depicts the best model for residuals of equation (1). The spherical model fit the

Figure 1

Fitted spherical semivariogram for the visceral leishmaniasis data.

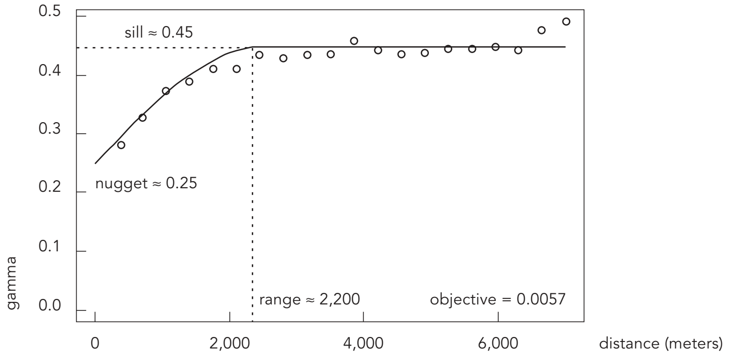
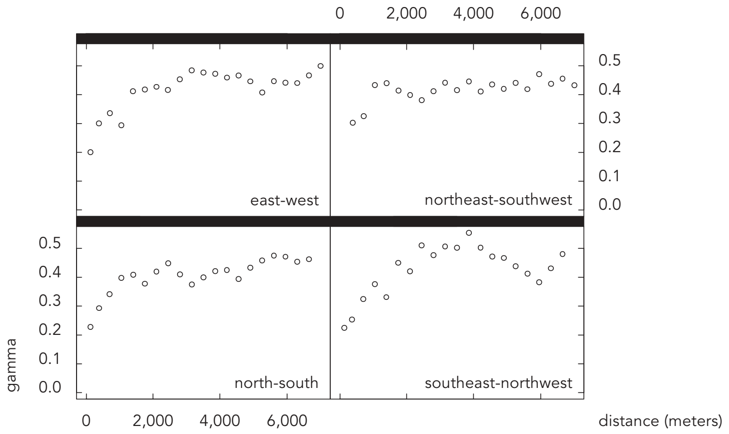


Figure 2

Empirical semivariograms in four directions for the visceral leishmaniasis data.



data well, and suggests a range of about 2,200 meters, a nugget effect of about 0.25, and a sill of approximately 0.45. Based on these estimates, the spatial covariance for the residuals of equation (1) was modeled using the spherical function in SAS PROC MIXED. This model was then compared to the independent model (that ignored the presence of spatial correlation), using the Likelihood Ratio Test (based on -2REML Log Likelihoods).

Results

The results obtained with the spherical model for the covariance structure indicate that there is significant spatial correlation in the VL data. According to this model, the estimate for the range is 2101.3, indicating spatial correlation in VL rates of consolidated census tracts that are within 2km of each other. The estimates for the nugget and the sill were 0.28 and 0.51, respectively.

The model with a spherical covariance structure fit the data better than the independence model ($p < 0.0001$, Likelihood Ratio Test) (Table 1). Both models provided the same qualitative result, that is, living in a *favela* and/or in areas covered by green vegetation was positively associated with the incidence of VL. However, the values for the fixed effects and standard errors of mean NDVI changed substantially when moving from the independence model to the spherical model, with a relative increase of about 60% and 17%, respectively. No major changes were detected in the fixed effects or standard errors associated with SES.

Discussion

Several studies have used different approaches to explore spatial autocorrelation when modeling areal data (Clayton et al., 1993; Cook & Pocock, 1983; Cressie & Chan, 1989; Leyland et al., 2000; Richardson et al., 1995). Following previous studies (Penello et al., 1999; Pickle, 2000; Pickle et al., 1999), in this report we use a mixed model framework for modeling areal data. Compared to models that do not take spatial autocorrelation of measurements into account, this approach provides more valid estimates for fixed effects and standard errors, and provides a description of the spatial structure underlying the data. Modeling spatial dependence in this way in essence requires only a single step using a restricted maximum likelihood estimation process.

We identified three important issues for the researcher who intends to use these models. The first is that most models for spatial covariance structure require that the assumption of isotropy hold, which can be tested by comparing empirical semivariograms for different directions (Cressie, 1991). When the isotropic assumption is violated, strategies for removing large-scale variation (spatial gradient), such as median polishing techniques should be tried (Cressie & Read, 1989).

The second issue pertains to the choice of a model for spatial covariance. A preliminary exploration of the features of the semivariogram is highly recommended to obtain a sense of the spatial correlation structure. This step will facilitate the selection of the best model from the various available models such as the linear, linear-log, Gaussian, and others, and avoid the extensive effort required to sort these out as well as problems with convergence.

The third issue refers to the fact that Generalized Linear Mixed Models with the Poisson link function are the most appropriate way of dealing with count data. Unfortunately, procedures for running models in programs such as SAS PROC GEE at present do not allow the specification of spatial covariance structures. Therefore, it is necessary to run SAS PROC MIXED using transformed data, with the attendant loss of direct interpretability of the regression coefficients and covariance parameters.

Our application of this approach to the data from Teresina supports the hypothesis of spatial correlation of VL rates. Modeling spatial correlation using mixed models yielded more accurate point and standard error estimates. The experience reported here argues for inclusion of spatial modeling in the analysis of ecologic studies.

Table 1

Fixed effects estimates and log-likelihood (REML) for the Spherical and Independence models.

Parameter	Estimate	Standard Error	Z	Pr > lZl	-2REMl_LL
Spherical model					841.21
Non-slum	-0.386	0.103	-3.75	0.0002	
Mean NDVI	1.934	0.503	3.85	0.0001	
Independence model					895.09
Non-slum	-0.433	0.102	-4.24	0.0001	
Mean NDVI	1.191	0.429	2.78	0.0057	

NDVI = Normalized Difference Vegetation Index.

References

CLAYTON, D. G.; BERNARDINELLI, L. & MONTOMOLI, C., 1993. Spatial correlation in ecologic analysis. *International Journal of Epidemiology*, 22:1193-1202.

COOK, D. G. & POCOCK, S. J., 1983. Multiple regression in geographic mortality studies with allowance for spatially correlated errors. *Biometrics*, 39:361-371.

COSTA, C. H. N.; PEREIRA, H. F. & ARAÚJO, M. V., 1990. Epidemia de leishmaniose visceral no estado do Piauí, Brasil, 1980-1986. *Revista de Saúde Pública,* 24:361-372.

CRESSIE, N., 1991. *Statistics for Spatial Data*. New York: John Wiley & Sons.

CRESSIE, N. & CHAN, N. H., 1989. Spatial modeling of regional variables. *Journal of the American Statistical Association*, 31:699-719.

CRESSIE, N. & READ, T. R. C., 1989. Spatial data analysis of regional counts. *Biometrical Journal*, 31: 699-719.

HAY, S. I.; TUCKER, C. J.; ROGERS, D. J. & PACKER, M. J., 1996. Remotely sensed surrogates of meteorological data for the study of the distribution and abundance of arthropod vectors of disease. *Annals of Tropical Medicine & Parasitology,* 90:1-19.

LEYLAND, A. H.; LANGFORD, I. H.; RASBASH, J. & GOLDSTEIN, H., 2000. Multivariate spatial models for event data. *Statistics in Medicine*, 19:2469-2478.

LITTELL, R. C.; MILLIKEN, G. A.; STROUP, W. W. & WOLFINGER, R. D., 1996. *SAS System for Mixed Models*. Cary: SAS Institute.

LITTELL, R. C.; PENDERGAST, J. & NATARAJAN, R., 2000. Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, 19:1793-1819.

MORGENSTERN, H., 1998. Ecologic studies. In: *Modern Epidemiology* (K. J. Rothman & S. Greenland, ed.), pp. 459-480, Philadelphia: Lippincott-Raven Publishers.

PENNELLO, G. A.; DEVESA, S. S. & GAIL, M. H., 1999. Using a mixed effects model to estimate geographic variation in cancer rates. *Biometrics*, 55:774-781.

PICKLE, L. W., 2000. Exploring spatio-temporal patterns of mortality using mixed effects models. *Statistics in Medicine*, 19:2251-2263.

PICKLE, L. W.; MUNGIOLE, M.; JONES, G. K. & WHITE, A. A., 1999. Exploring spatial patterns of mortality: The new atlas of United States mortality. *Statistics in Medicine*, 18:3211-3220.

RICHARDSON, S.; MONFORT, C.; GREEN, M.; DRAPER, G. & MUIRHEAD, C., 1995. Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. *Statistics in Medicine*, 14:2487-2501.

THOMSON, M. C.; CONNOR, S. J.; MILLIGAN, P. & FLASSE, S. P., 1997. Mapping malaria risk in Africa: What can satellite data contribute? *Parasitology Today*, 13:313-318.

Submitted on 23 January 2001
Final version resubmitted on 9 July 2001
Approved on 8 October 2001