
# Non-Gaussian spatial modelling of Chagas disease in Argentina

Pablo Juan[1,*], Jorge Mateu[1] and Carlos Díaz-Avalos[2]

[1] *Department of Mathematics, University Jaume I, Castellón, Spain; mateu@mat.uji.es, juan@mat.uji.es*
[2] *Departamento de Probabilidad y Estadística, Universidad Nacional Autónoma de México; carlos@sigma.iimas.unam.mx*
[*] *Corresponding author*

***Abstract.***
*We model the spatial distribution of prevalence of the most important Chagas disease vectors to enable predictive mapping of univariate and multivariate prevalence of the species vectors and the disease risk. We analyse both the binary variable of presence-absence of Chagas disease and the species richness in Argentina, in combination with meteorological and topographical covariates associated to the grid. We use several statistical techniques to produce distribution maps of presence-absence, and species richness including a hierarchical Bayesian framework within the context of multivariate geostatistical modelling. Our results show that, as expected, the inclusion of covariates improves the quality of the fitted models, and that there is spatial interaction between neighboring cells/pixels.*

***Keywords.*** *Chagas disease; Climatic conditions; Population dynamics; Presence-Absence; Spatial distribution.*

## 1 Introduction

Chagas disease is caused by infection with the protozoa Trypanosoma cruzy. The disease has been considered an autoctonous disease of 22 countries in the continental Western Hemisphere (WHO, 2002). The estimated rate of prevalence for different countries ranges between 0.7 and 15.5%. For Argentina, the estimated prevalence rate of Chagas disease is 8.2% (Schmunis and Yadon, 2010).

Among the 17 most common species related to Chagas disease, Triatoma, Panstrongylus and Psammolestes are perhaps the most important and widespread vectors of Trypanosoma cruzy, the causative agent of Chagas disease. These vectors are widely distributed in Argentina and other South American countries, where they probably contribute to more than a half of the estimated 24 million cases of this disease. However, knowledge of the population characteristics of these vectors is limited to laboratory

studies and partial field observations. One of these field observational studies records data on presence-absence of the above mentioned disease vectors for a variety of species over a fine grid covering part of Argentina. In addition, climatic and topographical covariates were also measured.

We aim at describing the spatial distribution of prevalence of the disease vectors to enable predictive mapping of univariate and multivariate prevalence of the species vectors and the disease risk. We analyze the binary variable of presence-absence of Chagas disease in Argentina, with data obtained from a long term field survey on a grid covering the northern part of the country, in combination with meteorological and topographical covariates associated to the grid. We use several statistical techniques to produce distribution maps of presence-absence, and species richness including a hierarchical Bayesian framework within the context of multivariate geostatistical modeling. The fitted logistic regression models range from simple logistic regression to models with a spatial term. We also explore models to test the possibility of interaction between different vector species. Our results show that, as expected, the inclusion of covariates improves the quality of the models fitted, and that there is spatial interaction between neighboring pixels in the grid.

## 2 Statistical methodology

Given that the data are dichotomous, we fit logistic regression models of the form $\xi_i = logit(p_i) = \mathbf{x}_i'\beta$ to the presence-absence data of the five species considered as the most important. $\beta$ is a vector whose entries are the coefficients related to the covariates $\mathbf{x}$. Logistic regression assumes that the data are independent, but such assumption might not be adequate for spatial data. We thus alternatively fit the model $\xi_i = \mathbf{x}_i'\beta + \psi_i$, in which $\psi_i$ is a term that incorporates the possible spatial interaction between neighboring locations. The spatial term $\psi_i$ can be considered as a surrogate for unobserved variables that are correlated in space (Besag *et al.*, 1995). This corresponds to generalized linear mixed models (GLMM), a class of models useful in problems that involve the mapping of risks (Clayton and Kaldor, 1987).

Our approach to model fitting comes from the Bayesian context, that is, if $\theta$ is a vector of $k$ components containing all the parameters in the model, statistical inferences about $\theta$ are based on the posterior distribution $f(\theta|y) = \frac{L(\theta;y)\pi(\theta)}{\int_\Theta L(\theta;y)\pi(\theta)d\theta} \propto L(\theta;y)\pi(\theta)$

Our model formulation is as follows. We assumed a flat, non informative prior distribution (Box and Tiao, 1973) for the non spatial parameters in our model, which allows them to be assigned any arbitrary initial values. Following Besag *et al.* (1991), we assumed a Gaussian pairwise difference prior distribution for the spatio term, with precision $\lambda$, $\pi(\psi) \propto \lambda^{0.5N}|W|^{0.5}\exp\{-0.5\lambda\psi'W\psi\}$, where $\psi = (\psi_1,\ldots,\psi_N)$ is the vector of spatial components, $W$ is a matrix with $W_{ii} = \nu_i$, $W_{ij} = \frac{1}{2}1$ if pixels i and j are neighbors and $W_{ij} = 0$ otherwise. $|W|$ denotes the product of the nonzero eigenvalues of W. We are weighting each direction equally, because the geographic scale we are working with is not detailed enough to detect possible anisotropies. The prior density previously described belongs to the class of non stationary Gaussian intrinsic autoregressions, and may be considered as the stochastic equivalent of linear interpolation (Besag *et al.*, 1991).

Because both the columns and rows of $W$ add to zero, this prior is improper. However, the full conditional densities necessary to make statistical inferences on the $\psi_i, i = 1,\ldots,N$ are well defined (Besag, *et al.*, 1995). For the precision $\lambda$ we assumed a $G(1,1)$ prior density, which allows initially low values for $\lambda$ and, therefore, high variability in $\psi$. The model is completely specified by further assuming

independence between the components of $\theta = (\beta, \psi, \lambda)$ and by assuming that the observations $y_i$ are conditionally independent. The posterior density of the parameters is proportional to

$$\left( \prod_{i=1}^{N} \frac{\exp\{y_i x_i' \beta + \psi_i\}}{1 + \exp\{y_i x_i' \beta + \psi_i\}} \right) \times \lambda^{0.5N} |W|^{0.5} \exp\{-0.5\lambda \psi' W \psi\} \times \lambda^{a-1} \exp\{-b\lambda\}$$

And the full conditional distributions are given by

$$\pi(\beta|\cdot) \propto \left( \prod_{i=1}^{N} \frac{\exp\{y_i x_i' \beta + \psi_i\}}{1 + \exp\{y_i x_i' \beta + \psi_i\}} \right)$$

$$\pi(\psi|\cdot) \propto \left( \prod_{i=1}^{N} \frac{\exp\{y_i x_i' \beta + \psi_i\}}{1 + \exp\{y_i x_i' \beta + \psi_i\}} \right) \times \exp\{-0.5\lambda \psi' W \psi\}$$

$$\pi(\lambda|\cdot) \sim \Gamma(a + 0.5N, b + \sum_{i=1}^{N} \nu_i (\psi_i - \bar{\psi})^2)$$

where the notation $(u|\cdot)$ means the conditional distribution of one set of parameters given the rest of the components in the model. To avoid numerical problems during the MCMC computations, covariate values were normalized. We used the Hastings (1970) algorithm to update the nonconjugate distributions, generating the candidate values with a Gaussian density centered at the current value for each parameter. The dispersions of these candidate generators were tuned to get acceptance rates in the 25-60% range.

## 3 Data analysis

Chagas disease has been considered an autoctonous disease of 22 countries in the continental Western Hemisphere (WHO, 2002). Among the 17 most common species related to Chagas disease, Triatoma, Panstrongylus and Psammolestes are perhaps the most important and widespread vectors of Trypanosoma cruzy, the causative agent of Chagas disease.

These vectors are widely distributed in Argentina and other South American countries, where they probably contribute to more than a half of the estimated 24 million cases of this disease. One of these field observational studies records data on presence-absence of the above mentioned disease vectors for a variety of species over a fine grid covering part of Argentina. In addition, climatic and topographical covariates were also measured. Figure 1 represents the selected species we here analyze. And Figure 2 shows the other type of response variable that we analyze based on Poisson counts of different types of species (called species richness), and the NDVI as an example of covariate. We make use of logistic spatial regression to model the spatial distribution of prevalence of the disease vectors to enable predictive mapping of univariate and multivariate prevalence of the species vectors and the disease risk. We also model Poisson counts of species to obtain predictive richness of Chagas disease in Argentine.

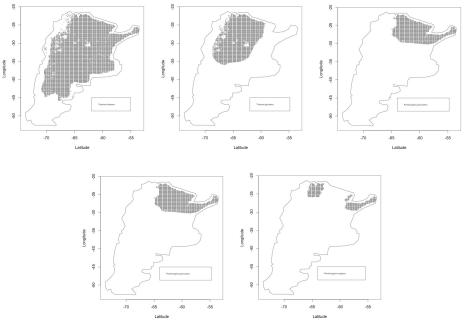

Figure 1: Species. *First row*: Triatoma.infestans, Triatoma.garciabes and Panstrongylus.guentheri. *Second row*: Panstrongylus.geniculatus and Panstrongylus.megistus.

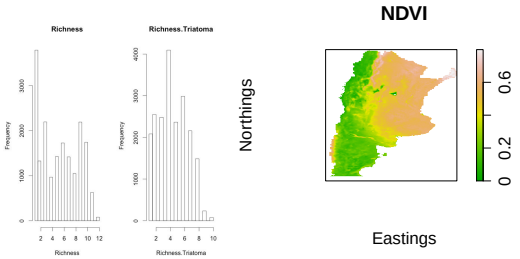


Figure 2: *Left*: Histograms of species richness and Triatoma richness. *Right*: NDVI covariate.

**Acknowledgments.** Work partially funded by grant MTM2010-14961 from the Spanish Ministry of Science and Education.

## References

[1] Besag, J., York, J. and Mollie,A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1-21.

[2] Besag, J., Green, P.J., Higdon, D. M. and Mengersen, K. L., (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3-66.

[3] Box, G. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis.* John Wiley and Sons.

[4] Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671-681.

[5] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

[6] Schmunis G.A. and and Yadon Z.E. (2010). Chagas disease: A Latin American health problem becoming a world health problema. *Acta Tropica*, **115**: 14–21.

[7] World Health Organization (2002). Control of Chagas disease. Second report of the WHO Expert Committee. W.H.O. Tech. Rep. Ser. 905, Geneva, pp. 1-109.