IDRiM(2013)3(1) ISSN: 2185-8322
DOI10.5595/idrim.2013.0034

Journal of Integrated Disaster Risk Management

Original paper

Bayesian analysis of population vulnerability to rainfall events in Venezuela.

Jhan Rodriguez[1] and Lelys Isaura Guenni[2]*

Received: 29/09/2011 / Accepted: 26/04/2013 / Published online: 04/06/2013

Abstract We use a Bayesian hierarchical model to quantify, at the district scale, the vulnerability of population to rainfall-related events, such as floods, flash-floods, and landslides. As a measure of vulnerability quantification we use the Relative Risk (RR). The RR is defined for each district and a given time span, as the ratio of the (unknown) potential proportion of people affected in the district to a pre-fixed, data-based, expected proportion of people affected. Thus, the RR is a measure of deviation from the expected behaviour of damage to population in each district. It can be used as an indicator of anomalous damage behaviour, by identifying those districts having a RR (say) significantly different from one. The model employed for the RR analysis is a log-linear model which considers the number of affected people in each district as the realization of a Poisson variable, and allows the inclusion of district-specific covariates. The model also allows the inclusion of parameters that capture any structural spatial pattern on the underlying RR surface, namely the so-called Conditionally Auto-Regressive, or CAR, effects. An important result is the RR map of Venezuela, which summarizes the posterior distribution of the RR for each district, and indicates that the most vulnerable districts form clusters in Nord-central and Western Venezuela, in addition to other districts of high RR arranged in a less structured way.

Key words Vulnerability; Risk; Spatial hierarchical models; Bayesian modelling.

1. INTRODUCTION

Flash-floods, floods and landslides are frequent occurrences in most tropical countries around the world, and cause yearly considerable losses, both human and material (ISDR 2004). The losses, however, are not only the direct result of the occurrence of a natural phenomenon, but the combination of this with the existing situation or coping capacity of the population and goods in whose spatial extension the natural phenomenon occurs. For example, in (ISDR, 2004), 'Risk' is conceptualized as follows:

"Risk: The […] expected losses (deaths, injuries, property, livelihoods, economic activity disrupted or environment damaged) resulting from interactions between natural or human-induced hazards and vulnerable conditions".

[1] Department of Scientific Computing and Statistics, Universidad Simon Bolivar, Caracas, Venezuela.

[2] *Corresponding author: Department of Scientific Computing, Coordinator of the Environmental Risk Assessment Research Group, Universidad Simon Bolivar, Caracas, Venezuela. E-mail: lbravo@usb.ve

Immediately, the concepts ‘hazard’ and ‘vulnerability’ are clarified:

“Two elements are essential in the formulation of risk: a potential damaging event, phenomenon or human activity – hazard; and the degree of susceptibility of the elements exposed to that source – vulnerability.”

This discrimination between natural phenomenon and coping capacity or existing ‘situation’, has been acknowledged for some time by researchers, and has led to various model formulations (see, for example, Downing *et al.* 1999; Schulze 2001; Plate 1996) that consist of various interacting components, and of which the output is a measure of the impact of the natural event. Thus, depending on the actual state of these components in a given geographical region (whatever its definition may be), one can expect bigger or smaller damages. Of course, the state of the components is given or estimated, either deterministically (e.g. social indicators, amount of population present in the region), or stochastically (e.g. the occurrence or not of precipitation, and the level of its intensity).

Regardless of the specific model employed, it is always useful to produce maps of the estimated damage measure (e.g. risk), in order to effectively identify the regions to which attention should be paid in first place, and to gain better understanding of the overall state and distribution of risk in the area of study. Thus it is customary in risk research and reporting, the use of GIS (Geographic Information Systems), as they provide useful plotting and analysis enhancing capabilities.

In this study, vulnerability is considered a dimensionless quantity conceived as the degree of loss or damage, between 0 and 100%, of the Venezuelan population (number of fatalities, people affected or injured, hereafter summarized as the number of “people affected”) due to rainfall-related events, such as floods, flash-floods, and landslides. As a convenient implementation of this concept of vulnerability, the proportion of people affected to the existing population exposed constitutes the measure of vulnerability on which this work builds. Let p^* be the global proportion (for the whole country) of people affected within a given time-span; let p_i be the proportion of people affected during the same time-span at district i. Then it is possible to define a measure of relative vulnerability for each district, $\psi_i = \frac{p_i}{p^*}$. This measure conveys the relatively critical or satisfactory situation of each district concerning its vulnerability, as placed within the big picture of the whole country. The parameter ψ_i is customarily employed in disease mapping and receives the name of “Relative Risk” within that context (see: Lawson *et al.* 2003). It provides means of identifying “hotspots” in which the relative damages suffered by the population are significantly higher than the average, and thus can be used as support for, e.g. governmental decision making and intervention. This measure of (relative) vulnerability will be the object of this study, even though we keep the name “Relative Risk” due to the origins of the methodology and its use in disease mapping research. But we want to warn the reader that our main study variable is the population vulnerability which is quantified in relative terms. We shall deal in the following with the quantification and mapping of the relative risks of the Venezuelan districts. That is to say, districts constitute the spatial unit of analysis in this work.

A Bayesian approach is employed, so full probability distributions are obtained for the relative risk of each district, which enables an easy assessment of statistical significances and precision estimates for the relative risks. Specifically, we use a hierarchical model that can accommodate covariates, such as geographical variables. It also allows parameters for the identification and highlighting of relative risk regional clusters, the so called Conditional Auto-Regressive (CAR) random effects.

In section 2 we introduce the basics of the models to be used later. In section 3 the data and specific particulars of the study are given. In section 4 we present the results, such as relative risk map and its interpretation, the parameter’s posterior distribution summaries, and the identification of clusters and districts with high or low relative risks. Finally we provide conclusions and further possibilities for the method.

2. STATISTICAL MODELS

As stated before, in this study we will work at the district scale. We begin denoting by Y_i the number of people affected[3] in district *i*, in a given time span. The districts are here indexed as $i = 1,...,I$, where I is the total number of districts in the country.

We consider Y_i a random variable and assume a Poisson model on this variable, thus:

$Y_i \sim Poisson(\lambda_i)$, $i = 1,...,I$.

In the above model, the λ_i parameters could be related to one another (after all, they correspond to districts some of which are neighbours to other districts), or they could even be the same parameter: $\lambda_i = \lambda$, for $i = 1,...,I$.

Further, the given parameters are decomposed as follows:

$\lambda_i = E_i.\psi_i$, $i = 1,...,I$.

Where E_i is called the *expected risk* for district *i* ($i = 1,...,I$), or expected number of losses or of people affected, depending on the specific research context. (Desirable is that this expected number be as close to zero as possible, but experience holds, that this is not a realistic assignment). This expected risk is assigned in this work the value of $E_i = n_i.p^*$, where n_i stands for the total number of inhabitants (exposed people) at district *i*, and p^* denotes the global proportion of people affected for all the country.

The parameter ψ_i is called the *relative risk* for district *i* ($i = 1,...,I$), and thus is a parameter of much concern in this study. Now we see that ψ_i can be interpreted as a multiplying factor to the expected number of losses, given within the context of the Poisson model by $\lambda_i = E_i.\psi_i$. It becomes of interest to evaluate whether ψ_i is significantly different (greater or smaller) than one, or whether it can be accepted to be one, in which case the district is understood to behave as an average district of the country.

The model now focuses on the relative risks, providing prior probability distributions for these, as we are using the Bayesian approach. The prior distributions are updated to posterior distributions which are a compromise between the priors and the observed data. There are several possibilities for relative risks' prior distributions (see Lawson 2003, chapter 6; or Banerjee *et al.* 2004, section 5.4), we provide some below. On writing the prior distributions for the relative risks, we define different hierarchical models for the number of people affected, Y_i. Some detail is shown for the Gamma-Poisson model, but similar explanations apply to the other models.

Gamma-Poisson model:

In order to estimate the posterior distributions to apply the Bayesian paradigm, the data likelihood and the prior distributions for the data model parameters are defined as follows:

Likelihood level: $Y_i \sim Poisson(E_i.\psi_i)$, $i = 1,...,I$.

Prior for ψ_i: $\psi_i \sim Gamma(a,b)$, $i = 1,\ldots,I$.

The parameters *a* and *b* should be selected in a sensible fashion. Sensible fashion here means in such a

3 This concept will be clarified later in the text.

way that they portray valuable information acquired in other (former) studies, or else that the data itself determines most of the posterior distributions. For example (Banerjee *et al*. 2004), we could choose $a = 4$ and $b = 4$, which yield a mean of $\mu = a/b = 1$ (the "Null" value) and a standard deviation of $\sigma = \sqrt{(a/b^2)} = 0.5$ that is big for this scale. Then the posterior distribution for every ψ_i, in the light of available data y_i, becomes:

$$\psi_i \mid a,b,y_i \sim Gamma(a + y_i, b + E_i),\ i = 1,...,I .$$

And the posterior mean of each relative risk becomes:

$$E(\psi_i \mid y_i, a,b) = \frac{a + y_i}{b + E_i} = \omega_i SMR_i + (1 - \omega_i).\frac{a}{b}\,0$$

where $\omega_i = \dfrac{E_i}{b + E_i}$ and SMR_i is normally called in Epidemiology the *Standardized Mortality Ratio*, which is calculated as y_i / E_i. So the posterior mean is shown to be a weighted average of the *SMR* and the *a priori* mean a/b.

Customary is, however, to assign also priors to *a* and *b* rather than specific values, in order to allow data to dominate the posteriors, or else to smooth the influence of the parameters *a* and *b*. Thus the model is finally stated by adding these last priors, for example (Lawson 2003):

Priors for a and b:

$a \sim Exp(L_a)$
$b \sim Exp(L_b)$

Where $Exp(\theta)$ stands for the exponential distribution with parameter θ and mean $1/\theta$. The selection of L_a and L_b can be made through similar considerations as before, but now the posterior distributions for the ψ_i's are less sensible to this selection.

The sampling from the parameters' posterior distributions is done using a Markov Chain Monte Carlo (MCMC) simulation. To attain this, it is extremely useful to have the *posterior full conditional distribution*[4] of each parameter or vector of parameters, which is nothing but the distribution of each parameter in the model *given* all the other parameters and the data (Gilks *et al*. 1996, chapters 1 and 5).

In summary, the Gamma-Poisson model is written as:

Likelihood level: $Y_i \sim Poisson(E_i.\psi_i),\ i = 1,\ldots,I$.

Prior for ψ_i **:** $\psi_i \sim Gamma(a,b),\ i = 1,\ldots,I$.

Priors for a and b:

$a \sim Exp(L_a)$
$b \sim Exp(L_b)$

[4] Usually, researchers actually use the posterior full conditional *densities* in the MCMC sampling, but the 'distribution' terminology is mostly used in the literature.

Log-linear models

A weakness of the Poisson-Gamma model is that it does not allow for the inclusion of covariates (such as district-wise social or geographical variables), and does not allow for the explicit modelling of spatial correlation. To overcome these inadequacies, researchers have come up with models that are linear on the logarithm of the relative risk, and so belong to the realm of *log-linear* models. In this subsection, we consider two types of log-linear models, one of them allows for covariates, and the other for covariates *and* spatial correlation modelling.

Type 1 model:

The following decomposing of the logarithm of the relative risk is considered (Besag *et al.* 1991):

$\log(\psi_i) = \beta_0 + \vec{\beta}.\vec{x}_i + v_i$, $i = 1,\ldots,I$.

Where $\vec{x}_i$ stands for a vector of *p* covariates pertinent to district *i*; $\vec{\beta}$ is a vector of coefficients whose components relate the covariates to the log-relative risk; and β_0 is an overall mean of the log-relative risks. The component v_i is a random effect intended to capture additional unstructured variability, and is assumed to come from a Normal distribution with zero mean and variance $\sigma^2 = 1/\tau_h$ (to be estimated). For ease of further explanation, the overall mean β_0 and the vector of coefficients $\vec{\beta}$ are collapsed into a single vector $\beta := (\beta_0, \vec{\beta})$, and each covariates vector is added a 1 in its first coordinate: $x_i := (1, \vec{x}_i)$. Thus we have the equivalent expression $\log(\psi_i) = \beta.x_i + v_i$, for each log-relative risk.

In summary, this model allows for covariates adjustment, via the vector β , and also allows for further unstructured departure from the regression model, via the random effect v_i .

The model equations are:

Likelihood level: $Y_i \sim Poisson(E_i.\psi_i)$, $i = 1,...,I$.

where $\psi_i = \exp(\beta.x_i + v_i)$.

***Prior distributions*:**

$P(\beta) = 1$, for every β in $\Re^{p+1}$

$v_i \mid \tau_h \sim N(0, 1/\tau_h)$ (they are assumed independent among each other)

$\tau_h \sim Gamma(a_h, b_h)$

Commentaries:

1. The probability density assigned to β is not a proper density[5], and so does not provide a probability distribution for β . However, it is not strange in Bayesian statistics to make use of improper prior densities when the respective posterior densities of the parameters in question will result in proper densities. The reader interested in the necessary and sufficient conditions that ensure the propriety of the posterior distributions for the kinds of models presented in this article, is referred to: Sun *et al.* 2001, Ghosh *et al.* 1998, Song *et al.* 2006 and Eberly and Carlin 2000, and the references therein.

---

5 Namely, the integral of this function over $\Re^{p+1}$ is infinity.

2. The parameter τ_h, called the precision parameter of the distribution of v_i, is unknown and so it needs to be estimated. In the present model, we provide it with a prior distribution as well, and seek to estimate its posterior distribution. To this end, we assign a Gamma distribution with parameters a_h and b_h, and again wish the model to be insensible to the selection of these two parameters. One possible choice, and the one applied in this study, is to take $a_h = 0.5$, $b_h = 0.0005$ (Kelsal and Wakefield 1999), which results in a mean of $a_h / b_h = 1000$ and a variance of $a_h / b_h^2 = 2 \times 10^6$.

Type 2 model

After adjustment for covariates and beyond-covariates' heterogeneity effects, it is tenable to check whether geographical proximity remains a factor influencing the correlations among the relative risks. To this end it is adequate to add a component to the log-relative risks model that can adjust for correlation assignable to spatial proximity. In this study, we use the Conditionally Auto-Regressive (CAR) random effects model (Cressie and Chan 1989, Besag *et al.* 1991, Besag *et al.* 1995).

In this context, the model for the logarithm of the relative risk is extended to:

$$\log(\psi_i) = \beta_0 + \vec{\beta}.\vec{x}_i + v_i + c_i \ , \ i = 1,...,I \ .$$

Here, the random effect c_i is the Conditional Auto-Regressive effect for each district $i = 1,...,I$. Each of these random effects is influenced by those of the neighbouring districts. If a district *i* has a set ∂_i of m_i immediate neighbouring districts (i.e., $|\partial_i| = m_i$) then the *a priori* distribution for each of these effects, given all the others $c_{j \neq i}$ and a scale parameter τ_c, is defined to be:

$$c_i \mid \underline{c}_{-i}, \tau_c \ \sim \ N\left(\overline{c_i}, \frac{1}{\tau_c . m_i}\right) \ (*)$$

where $\overline{c_i} = \frac{1}{m_i} \sum_{j : j \in \partial_i} c_j$. That is, each CAR effect is a priori normally distributed with mean equal to the average of its neighbours, and a variance inversely proportional to the number of neighbours. The scale parameter τ_c is unknown and must be estimated, or its probability distribution estimated.

These conditionally defined distributions in (*), actually define a joint density (See Besag 1974, or Kaiser and Cressie 2000) for $\vec{c} = (c_1,\ldots,c_I)$ and it can be seen to be equivalent to the following formulation:

$$\Pr(\vec{c} \mid \tau_c) \propto \exp\left\{-\frac{\tau_c}{2} \vec{c}^{\,t} . Q . \vec{c}\right\}$$

where Q can be written in the form $Q = M^{-1}(1_I - B)$. M is a diagonal matrix of size *IxI* and components $1/\tau_c \, m_i$, 1_I is the identity matrix of size *I* and B is the adjacency matrix (with components b_{ij} 0's or 1's depending whether a district *j* is a neighbour or not of district *i*).

However, Q is singular and so the induced density is not proper. This is not strictly a problem, since the posterior for $\vec{c}$ ends up being proper, and this improper version of the CAR has been widely used in applications. Since sometimes it is better anyway to have an a priori proper distribution, one can try to keep as much as possible of the above-mentioned distribution and yet have it proper.

One remedy (Banerjee *et al.* 2004) is simply to include an additional parameter γ and "move it" until $Q_{(\gamma)} = M^{-1}(1_I - \gamma.B)$ becomes non-singular. With the matrix components defined as before, $Q_{(\gamma)} = M^{-1}(1_I - \gamma.B)$ happens to be non-singular for every $\gamma \in (0,1)$. A prior distribution can also be assigned to γ and its posterior distribution obtained; such an approach is followed in this study using a uniform distribution.

3. DATA AND STUDY AREA

3.1 Study Area

We perform the analysis on all of Venezuela, taking districts ("Municipios", in Venezuela) as spatial units. The boundaries of these districts have evolved during the time-span for which we collected data (1960 – 2000), so we use here the boundary configurations as of 2001. These were provided by the Instituto Nacional de Estadistica (INE). The data collected is described below.

3.2 Information about the adverse water-related events

By an "event" we mean any reported incident where at least 10 people were somehow affected; either on their property or in their physical integrity. We restrict here to the period January 1960 to December 2000. The collection of data has two sources: First we took the reports available at the database of the Centre for Research on the Epidemiology of Disasters (CRED). Second, a research was undertaken at the newspaper archive of the National Library of Venezuela in order to obtain more detailed information about those cases, and other cases not taken into account by CRED. We used, whenever available, the journals "El Nacional", "El Universal", and "Últimas Noticias", which have national coverage.

Out of these inquiries, we extracted the specific date and location of the events, sometimes at the district level, sometimes obtaining the specific name of the town. We also extracted the number of people affected on each event, subdivided according to: 1. Casualty, 2. Injured, and 3. Property lost or damaged. In figure 1, the accumulated number of people affected is presented on the 2001 district map.

3.3 Geographic data

We included in the analysis the following geographic variables derived from a Digital Elevation Model (DEM) having a resolution of 90 meters per pixel, provided by the Shuttle Radar Topographic Mission (SRTM). These variables were totalized for each district with the aid of the IDRISI Geographic Information System (GIS):

1. Elevation. From this variable the mean was obtained.
2. Slope. From this variable we considered the average, the percentage of the district having slopes greater than 30 degrees (see figure 2) and the percentage of the district having less than 5 degrees.

We used the available hydrographical data to build a map of distances from each 90 meter pixel to the closest river or lake. Unfortunately, this data was available at two different scales: North of the Orinoco river the scale is 1:100000, while in the South it is 1:500000. This issue is mirrored in the map produced, which is presented in figure 3.

In addition, we used the monthly "networked" precipitation anomaly maps at the 30"x30" scale of Fekete *et al.* 2001. In these maps, the downstream routing of the precipitation is taken into account, and

so the potential impact of an event downstream of a water body is accounted for. The anomaly for each pixel is obtained by dividing the estimated "networked" precipitation by a baseline average provided for each month of the year and each pixel. The data for the month of December 1999 is shown in figure 4 as an example.

The population density of each district as in the year 2001 was employed as a proxy for the densities during the study period. The data was kindly provided by the National Institute of Statistics (INE) and is shown in figure 5. The uneven distribution of the population and its higher concentration around the main urban centres is quite evident in this map.

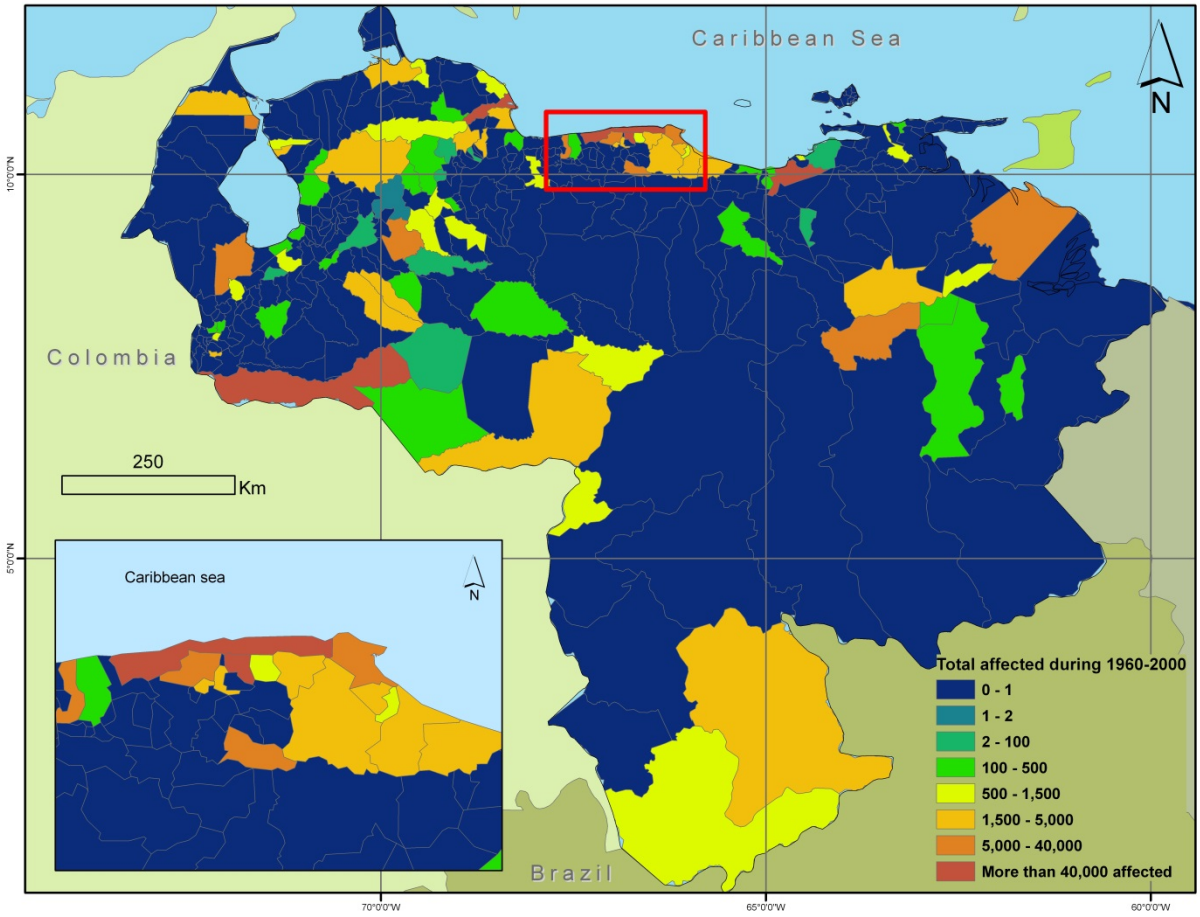

Figure 1: Total number of affected people per administrative unit during the period 1960-2000.

 ISSN: 2185-8322

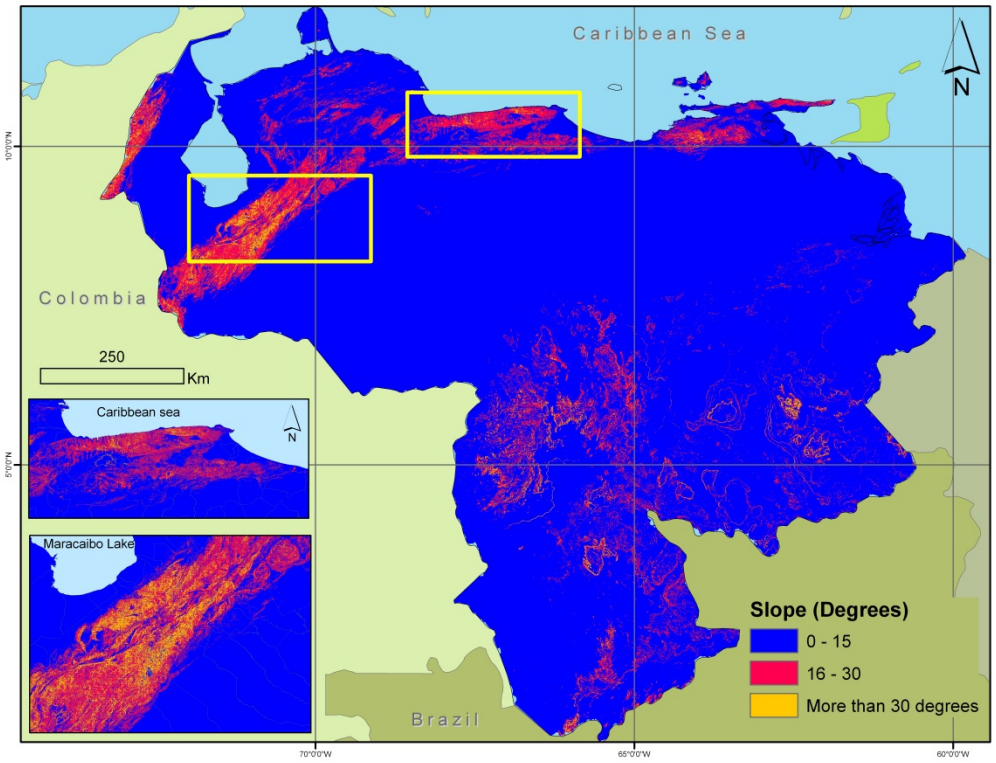

Figure 2: Locations having more than 30 degrees of slope

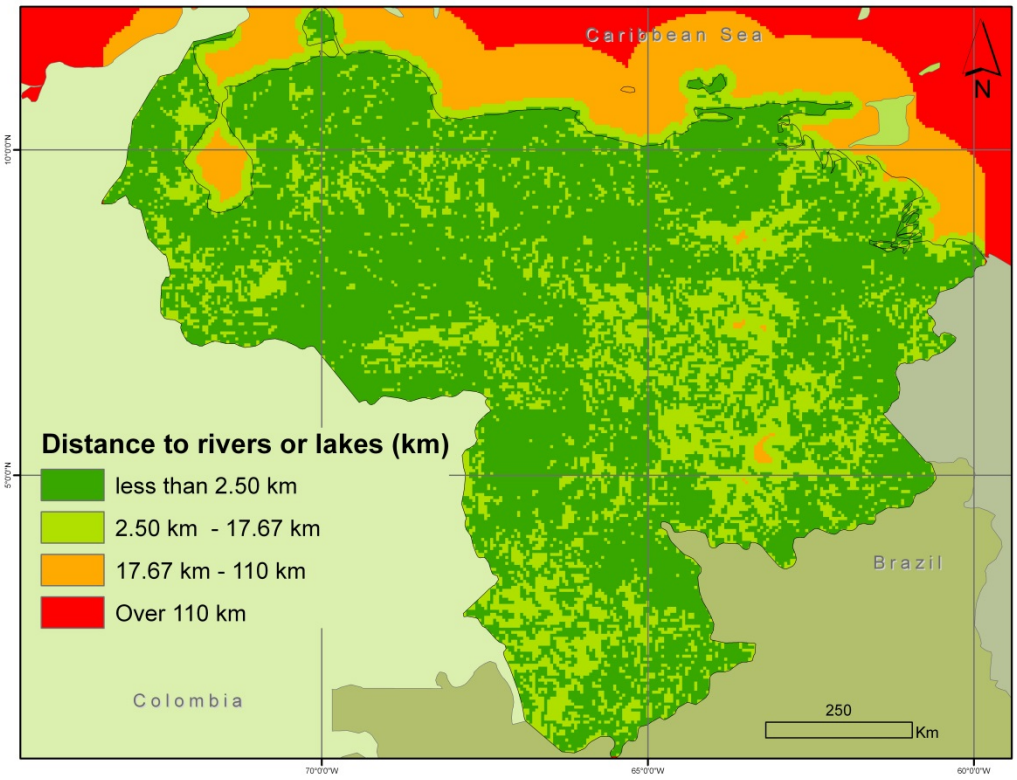

Figure 3: Map of distances (in kms) to the nearest river or lake.

 ISSN: 2185-8322

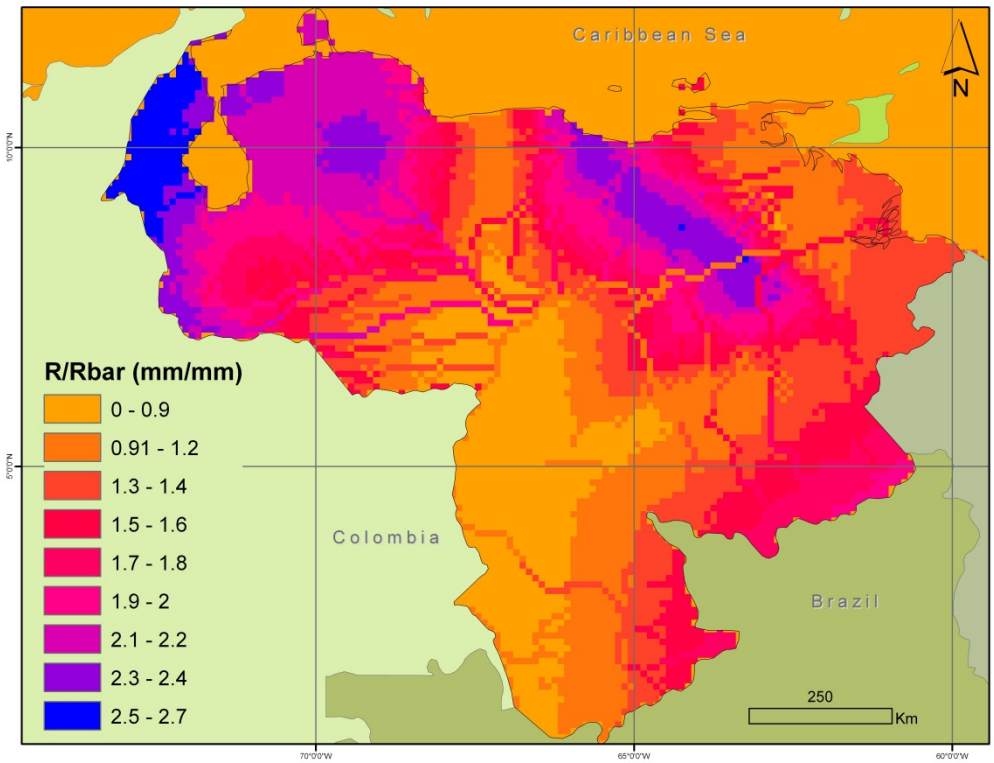

Figure 4: Networked rainfall anomalies for December 1999 with respect to the baseline rainfall climatology period 1960-1990.

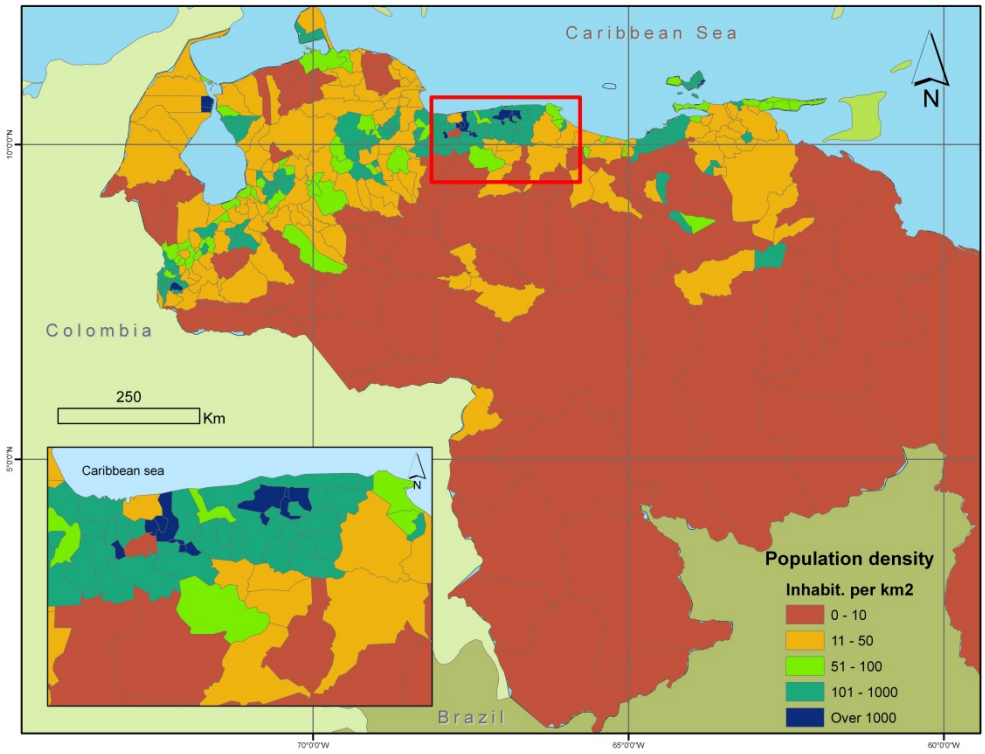

Figure 5: Population density in Venezuela, 2001.

4. RESULTS

As indicated above, we take districts as spatial units. Our objective is to estimate the probability distribution of the relative risk for each district and identify the way in which the geographical variables affect it. For each of these spatial units the variables presented in the previous section are obtained in the form of convenient statistics of the pixel-based data. Namely, the following variables were taken as explanatory variables for each spatial unit:

- X_1 : Proportion of the districts having a slope smaller than five degrees.
- X_2 : Proportion of the districts having a slope greater than thirty degrees.
- X_3 : Per district average distance to secondary water bodies[6], measured in kms.
- X_4 : District's average distance to primary water bodies, measured in kms.
- X_5 : District's average slope, in degrees.
- X_6 : Logarithm of the district's average elevation. The elevation was available in Metres Above Sea Level (MASL).
- X_7 : Logarithm of the district's 2001 population density, as portrayed by the National Institute for Statistics. The population density is expressed in inhabitants per squared kilometre.

For the sake of model fitting, al variables were standardized, as this transformation is prone to reduce MCMC chains' convergence problems. The Gamma-Poison model and the two types of log-linear models of section 2, using different subsets of variables X_1 through X_7 above as explanatory variables, were implemented in the WinBugs software (Spiegelhalter *et al.* 2003). For each model, two chains were produced and convergence of the MCMC chains was monitored by means of the Brooks and Gelman criterion implemented in WinBugs. Convergence in general was clearly attained for most parameters, although certainly not so for some CAR effects in the type two models. As model selection criterion, the criterion due to Gelfand and Gosh 1998 was used. The selected model is a type two model that includes X_3, X_4 and X_7 as explanatory variables, that is, the equation for the logarithm of the relative risk in each spatial unit *i*, is:

$$\log(\psi_i) = \beta_0 + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \beta_7 X_{i,7} + v_i + c_i \ , \ i = 1,...,I \ .$$

4.1 Effect of the selected explanatory variables

The MCMC based approximate densities for the coefficients of the included explanatory variables and that of the intercept are shown in Figure 6.

A decrease in variables X_3 and X_4 is associated with an increase in (log of) the relative risk, since the locations of these parameters' distributions are clearly to the left of zero. Let us remember that, in agreement with the definition of these variables, the smaller the values of X_3 and X_4, the higher the density of water bodies or rivers in the spatial unit. An important fraction of the lower income population in rural and urban areas has settled in the flood plains and river margins throughout the country; therefore the lower the values of X_3 and X_4 the larger the number of people potentially affected by rainfall related events.

[6] Those represented with lines at the working scale of 1:250,000. The primary water bodies are those represented with polygons.

A positive association is seen between population density and relative risk, as portrayed by the density of the coefficient of X_7. This is likely to mirror the fact that, in Venezuela, usually big cities are surrounded by or comprise inside of its poorer zones with very deficient infrastructure and services (the so-called marginal zones), built on inadequate land and which are prone to landslides and/or floods when strong precipitation occurs.

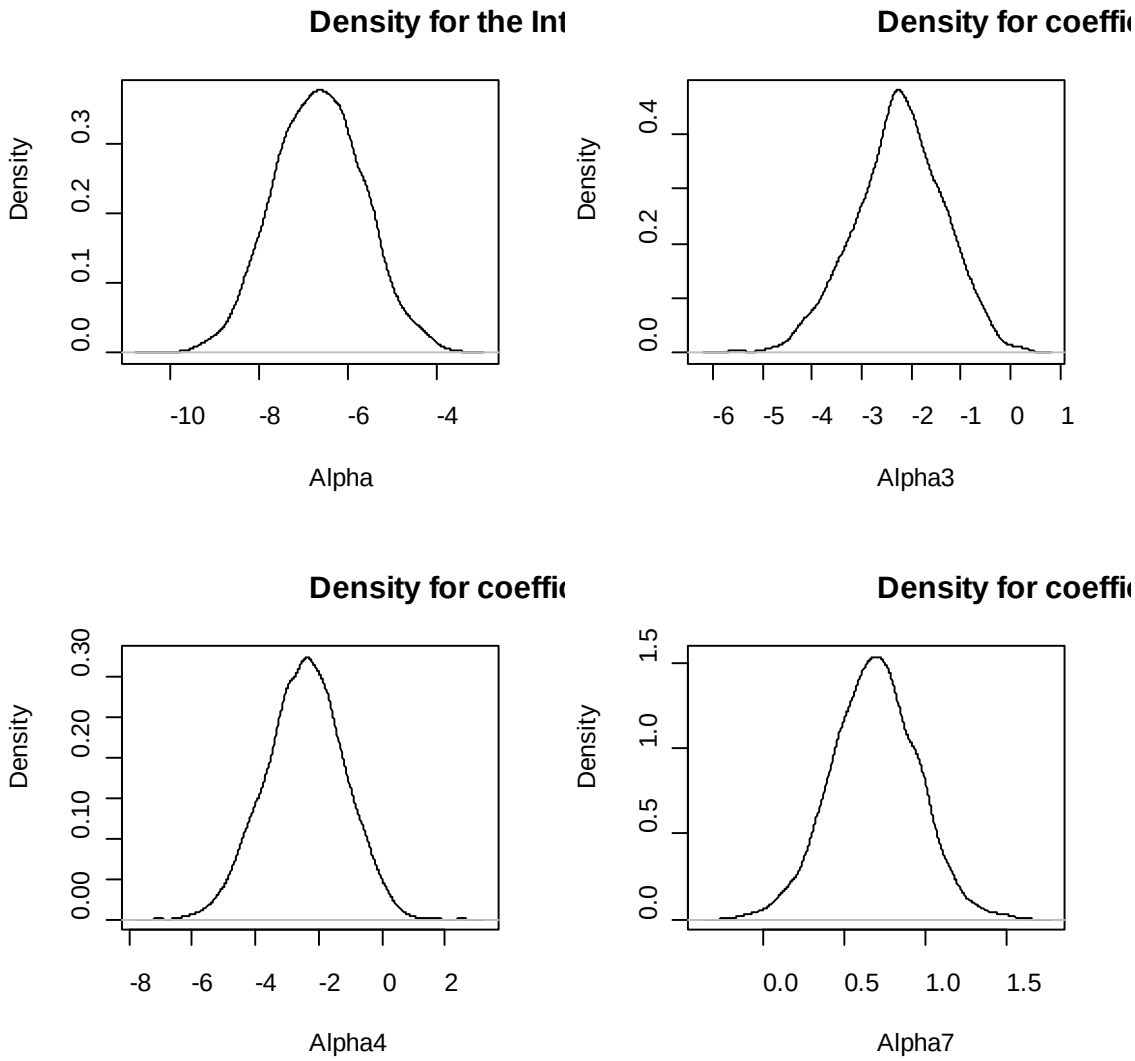

Figure 6: Kernel smoothing probability densities for the intercept and the three explanatory variables of the selected model. All densities are produced using samples from the MCMC simulated chains.

4.2 Relative Risks Map

The expected relative risks for all districts are presented in Figure 7. These are computed, for each district, as the average of the respective MCMC relative risk chain. The expected relative risks that are smaller than one are masked with lighter colours. A cluster of relative risks greater than one are visible in the North-Central region of the country. The two district clusters at the south of the country were induced by a single big event in June of 1996, which affected this region of rather poor infrastructure. The presence of risky districts in the west extreme of the country can be explained by the presence of the

 ISSN: 2185-8322

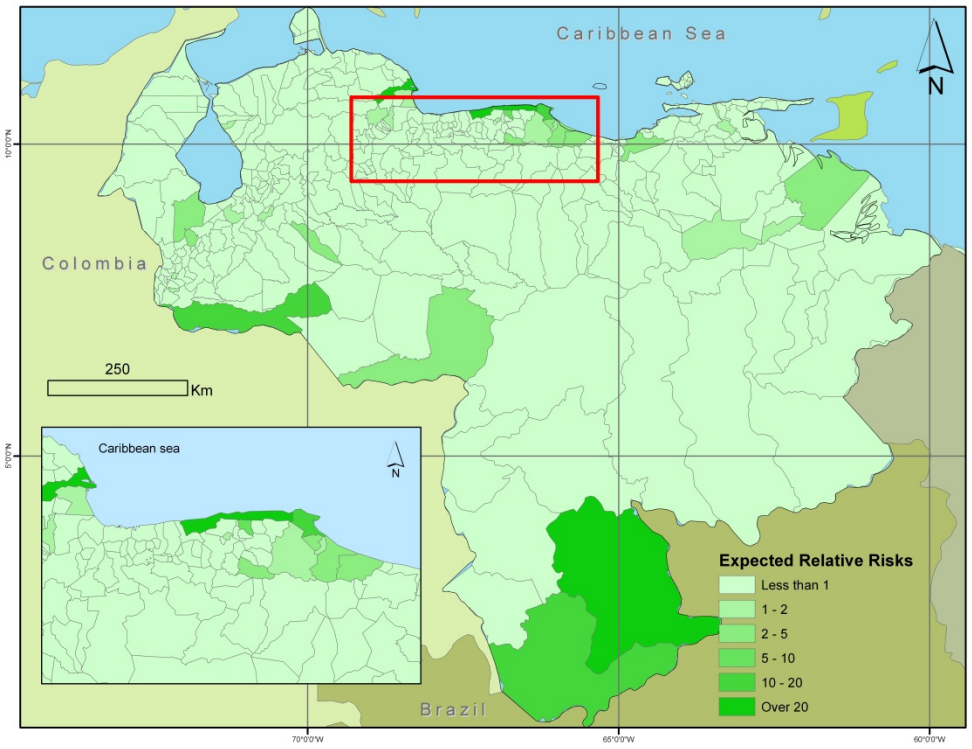

Figure 7: Expected Relative Risks

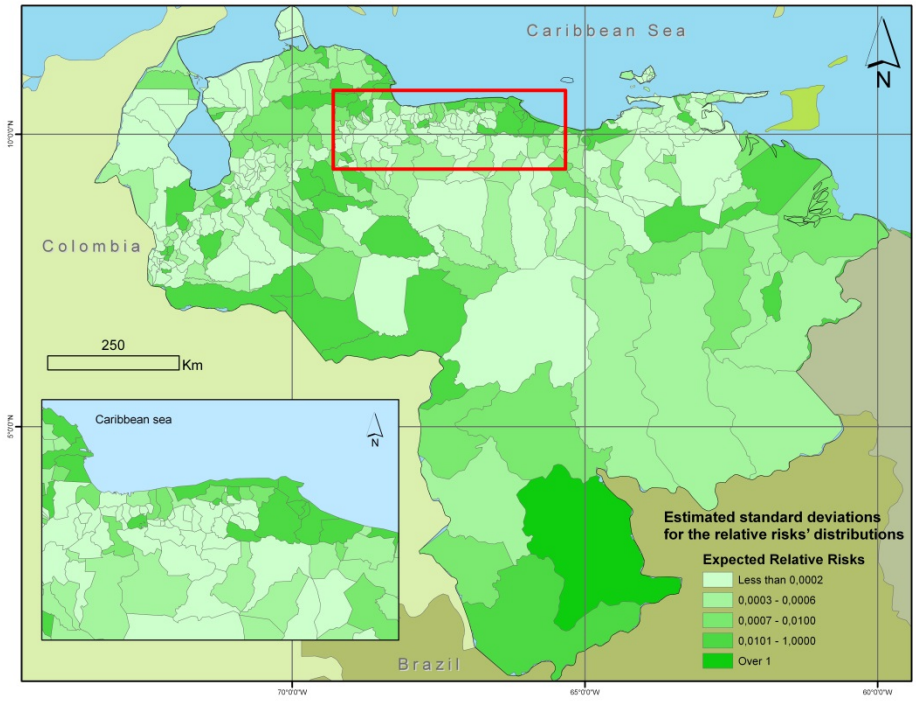

Figure 8: Estimated standard deviations for the relative risks' distributions

Andean ridge, which then abruptly descends towards the South-West of the country, into a very flat and low land. These characteristics can also explain the presence of the two risky districts in the South-West. As a complement to the expected relative risk map, the standard deviations of the relative risk distributions are shown in Figure 8 above.

4.3 Model Validation

In this section we report briefly on model validation. We employ a variant of cross-validation as means of identifying districts of which the observed risk is "too high" or "too low" under the assumed model, as measured through a suitable *discrepancy criterion* (*DC*) to be introduced shortly. Districts or areas presenting an extreme value of the *DC* will be named *divergent* districts or areas. The result of this validation analysis can be: 1. we find no divergent districts, which would imply a reasonable fitness of the model. 2. We find one or two divergent districts only, which would imply a reasonable fitness of the model but, at the same time, the presence of one or two "hot spots" or "interesting areas" where the influence of covariates or the spatial association pattern deviates considerably from that of the rest of the districts. 3. Relatively many (say, more that 1-2%) of the districts show a divergent response, which would imply the inadequacy of the fitted model.

The idea of checking the consistency of data with a model is standard in statistics and the general approach, of which the analysis here presented is a particular application, can be called the "predictive distribution approach". For an introduction to it, see chapter 9 of Gilks *et al.* 1996 and the references therein. The interested reader is referred to Stern and Cressie (2000) and Marshall and Spiegelhalter (2003) for further details on the application of the concept specifically to disease mapping.

Since we have fitted a model to the data presented in section 3, it is in principle possible, for any positive integer $y_{i,rep}$ representing a potential number of people affected at district *i* during a time-span of 40 years, to compute the probability of $y_{i,rep}$ according the fitted model. This probability would be, if we use a Poisson model such as those presented in section 2, given by $\Pr\left(Z_i = y_{i,rep} \middle| Data\right) = \frac{\hat{\lambda}^{y_{i,rep}} e^{-\hat{\lambda}}}{y_{i,rep}!}$, where Z_i stands for random variable "Number of people affected at district *i*", whereas $\hat{\lambda} = E_i . \exp\left(\hat{\beta}_0 + \hat{\beta}_3 X_{i,3} + \hat{\beta}_4 X_{i,4} + \hat{\beta}_7 X_{i,7} + \hat{v}_i + \hat{c}_i\right)$ is built from estimates of the model parameters such as, for example, the averages of the values with which plots in Figure 6 were created and their equivalents for effects c_i and v_i. One idea for testing the adequacy of the model could be to compute in a like manner the probabilities of the values actually observed at districts *i=1,…,I*; namely $\Pr\left(Z_i = y_{i,obs} \middle| Data\right) = \frac{\hat{\lambda}^{y_{i,obs}} e^{-\hat{\lambda}}}{y_{i,obs}!}$, and check whether these probabilities are "too low" in some sense. If the model declares unlikely even the data that was actually observed, then we have reasons for discarding such a model. This is basically the idea that we implement for our validation analysis, with three modifications:

Firstly, we shall use the probability of $\{Z_i \le y_{i,obs}\}$ instead of that of $\{Z_i = y_{i,obs}\}$ as a *discrepancy criterion*. A too high value of the criterion indicates that the fitted model predicts mostly values which are considerably under the observed ones (under-estimates losses), whereas a too low value of the criterion means that the model predicts mostly values above the observed ones (over-estimates losses). In this work, we define values above 0.9 as "too high" and values under 0.1 as "too low".

Secondly, we shall compute the probability constituting our discrepancy criterion for each district *i* on the basis of all observed data *except* for $y_{i,obs}$ itself, namely: $\Pr\left(Z_i \le y_{i,obs} \middle| Data - \left\{y_{i,obs}\right\}\right)$. Thus, the technique is a version of the cross-validation method. As shorthand for $Data - \left\{y_{i,obs}\right\}$ we shall write y_{-i} .

Thirdly, instead of computing plug-in estimates for the parameters and then computing the desired probability, we shall use the "posterior predictive probability" distribution (see, for example, chapter 1 of Gelman *et al*. 2004), which is customary in Bayesian statistics:

$$\Pr\left(Z_i = y_{i,obs} \middle| y_{-i}\right) = \int_{\Theta} \Pr\left(Z_i = y_{i,obs} \middle| \theta_i\right) \Pr\left(\theta_i \middle| y_{-i}\right) d\theta_i$$

which means that DC_i , the discrepancy criterion for district *i* is given by:

$$DC_i = \Pr\left(Z_i \le y_{i,obs} \middle| y_{-i}\right) = \sum_{n=0}^{y_{i,obs}} \left\{ \int_{\Theta} \Pr\left(Z_i = n \middle| \theta_i\right) \Pr\left(\theta_i \middle| y_{-i}\right) d\theta_i \right\} \qquad \text{(eq. 1)}$$

The integral in equation 1 is along all possible values of the parameters $\theta_i = \left\{\vec{\beta}, v_i, c_i\right\}$; this integral is normally found by simulation, on the basis of the MCMC chains of the model parameters. Thus, in principle, one should run the MCMC algorithm up to *I = 335* times, leaving one district at a time, in order to compute the discrepancy criterion for each district. This can be computationally very demanding. Ingenious techniques to approximate the criterion by running the MCMC algorithm only once are the object of the papers of Stern and Cressie 2000 and Marshall and Spiegelhalter 2003. For the current analysis, we employed the method proposed by Stern and Cressie.

Specific details are rather numerous and are not susceptible of being addressed within the small space of this section; the reader may see the bibliography given above for details. Hopefully we have here conveyed the basic idea behind this "posterior predictive" cross-validation variant.

In Figure 9, we present the result of applying cross-validation to the data and the model selected. Only 2 out of 335 districts presented an extreme value for the discrepancy criterion. As explained above, rather than a sign of model inadequacy, this is indication of a particular response of these districts, potentially due to one or more unconsidered factors present at each of them, which should be investigated. One of the anomalous districts (Zamora, Aragua state) exhibits a DC value greater than 0.9, which means that the model under-estimates the expected number of people affected. The other one (Iribarren, Lara state) presents a DC value smaller than 0.1, which indicates under-estimation.

Summarizing, the model is adequate, according to the cross-validation method employed. Also, two districts have been identified as providing a different response with respect to the covariates analysed and the spatial association with its neighbouring districts, as compared with all other districts studied.

4.4 Inclusion of the Precipitation Variable

The reader might have already noticed that precipitation data, in the form of monthly anomalies on a 30"x30" grid, was available for this study and yet is not reflected in the final fitted model. The inclusion of the precipitation variable as a time series of monthly rainfall anomalies implies a spatio-temporal model, in which the number of parameters to estimate increases considerably. In this first attempt to picture the overall situation of the relative risks in Venezuela, we considered it sounder not to include the precipitation variable. One of the problems with a spatio-temporal model for precipitation related events is the great number of no-case ("zero events") that is to be expected in many districts along the time-span considered. As an example, using the monthly precipitation anomalies of section 3 and a study time-span of 40 years yields 480 time-steps for each district, in most of which a zero will result, making inference very difficult and MCMC fitting (convergence) unfeasible. One possibility that may be worth trying in

the future is the use of the "zero inflated" models (Lambert 1992, Ghosh *et al.* 2006) where this kind of data becomes tractable.

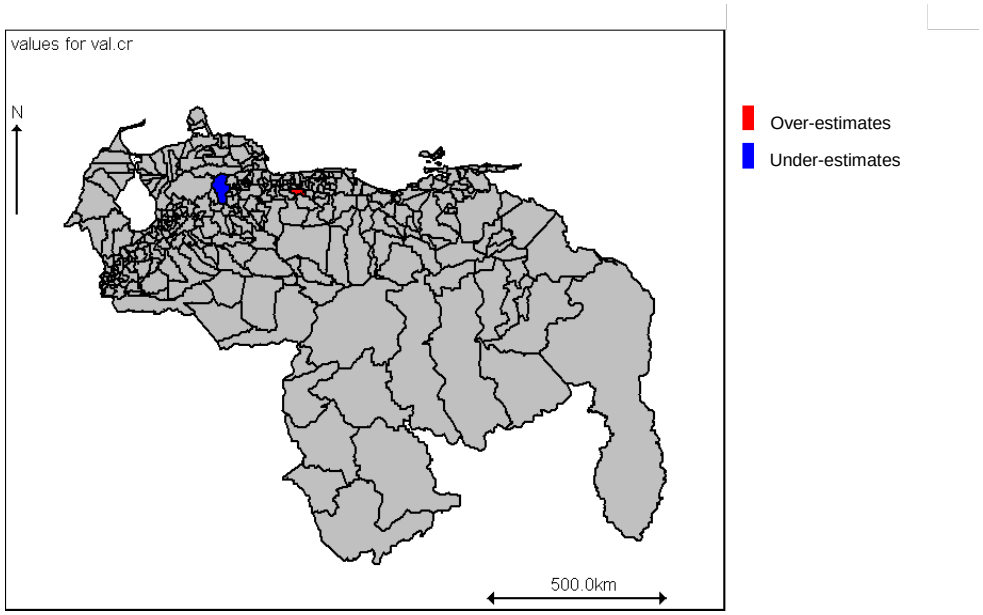

Figure 9: Divergent areas, showing a different response pattern as compared to the rest of the districts.

5. CONCLUSIONS

A methodology taken from epidemiology when dealing with *disease mapping* was successfully applied to the mapping of relative risks as a measure of vulnerability of people being affected by landslides and/or floods at the district scale in Venezuela. An overview of the situation for the whole country was then obtained, which can be a first step for future studies at local scales. Geographical variables, specifically, density of water bodies' network and population density were found to be positively associated with an increase in the relative risk at the district scale. The presence of the CAR random coefficients in the selected model, however, makes a case for a deeper understanding of the mechanism determining the spatial distribution of the relative risks.

ACKNOWLEDGEMENTS

The authors acknowledge Raul Ramirez Arbeláez for his tremendous assistance in the preparation of many of the figures of this paper; to the Instituto Nacional de Estadística (INE) for providing the population data sets; the Water System Analysis Group (WSAG) of the University of New Hampshire for providing the networked rainfall anomalies data, and the Fondo Nacional de Investigaciones Científicas y Tecnológicas (FONACIT), for partially funding this research under the project No. 2005-000184 .

REFERENCES

Banerjee, S., Carlin, B. and Gelfand, A. (2004) *Hierarchical Modeling and Analysis for Spatial Data*, Chapman&Hall/CRC.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems' (with discussions). *Journal of the Royal Statistical Society, Series B*, 36:192-236.

Besag, J., Green, P., Higdon, D., and Mengersen, K.L. (1995) Bayesian Computation and Stochastic Systems (With Discussion). *Statistical Science* 10(1): 3-41.

Besag, J., York, J. and Mollié, A. (1991) Bayesian Image Restoration with Two Applications in Spatial Statistics. *Annals of the Institute of Statistical Mathematics*, 43: 1-59.

Cressie, N. and Chan, N.H. (1989) Spatial modeling of regional variables. *Journal of the American Statistical Association*, 84:393-401.

Downing, T., Olsthoorn, A.J. and Tol, R.S.J. (eds) (1999) *Climate, Change and Risk*. Routledge, London, 407 pp.

Eberly, L.E. and Carlin, B.P. (2000) Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine*, 19(17-18):2279-2294.

Gilks, W.R., Richardson, S. and Spiegelhalter, D. (eds) (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, Boca Raton, Florida, 481 pp.

Gelfand, A.E. and Ghosh, S.K. (1998) Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85:1-11.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B (2004) *Bayesian data analysis*. Boca Raton: Chapman & Hall/CRC.

Ghosh, M., Natarajan, K., Stroud, T. and Carlin, B. (1998) Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93:273-282.

Ghosh, S.K., Mukhopadhyay, P. and Lu, J.C. (2006) Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*. 136(4):1360-1375.

ISDR (2004): Living with Risk: A global review of disaster reduction initiatives, 2004 version. http://www.unisdr.org/eng/about_isdr/bd-lwr-2004-eng.htm

Kaiser, M.S. and Cressie, N. (2000) The construction of multivariate distributions from Markov random fields. *Journal of Multivariate Analysis*, 73:199-220.

Kelsall, J. and Wakefield, J. (1999) Discussion of "Bayesian models for spatially correlated disease and exposure data. In *Bayesian Statistics 6*, Bernardo J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. (eds) Oxford University Press, Oxford, 151 pp.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to detects in manufacturing. *Technometrics*, 34: 1-14.

Lawson, A., Browne, W. and Vidal-Rodeiro, C. (2003) *Disease Mapping with WinBUGS and MLWiN*. Wiley.

Marshall, E. C., and Spiegelhalter, D.J. (2003) Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine, 22*(10), 1649–1660.

Plate, E.J. (1996) Risk Management for Hydraulic Systems under Hydrological Loads. Kovacs Symposium, UNESCO, Paris, France.

Schulze, R.E. (2001) *Risk, uncertainty and risk management in hydrology: A Conceptual Framework and*

a South African Perspective. Technical Report of Natal University, South Africa.

Song, J., Ghosh, M., Miaou, S. and Mallick, B. (2006) Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, 97(1): 246-273.

Spiegelhalter, D. J., Thomas, A., Best. N. G., and Lunn, D. (2003) *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge. http://www.mrc-bsu.cam.ac.uk/bugs.

Stern, H. S. and Cressie, N. (2000) Posterior predictive model checks for disease mapping models. *Statistics in medicine*, *19*(17-18), 2377–2397.

Sun D., Tsutakawa R., He, Z. (2001): 'Propriety of posteriors with improper priors in hierarchical linear mixed models'. Statistica Sinica 11(2001), pp. 77-95.