

# An analysis of edge effects in disease mapping

*Carmen L. Vidal Rodeiro<sup>1</sup>, Andrew B. Lawson<sup>1</sup>*

<sup>1</sup>Department of Epidemiology and Biostatistics  
Norman J. Arnold School of Public Health  
University of South Carolina, Columbia SC, U.S.A.  
Phone: +1 (803) 777 1232 Fax: +1 (803) 777 2524  
Email: cvidal@sc.edu

**Presented at GeoHealth 2002  
Victoria University of Wellington  
December 3-5<sup>th</sup> 2002**

## ABSTRACT

The construction and smoothing of disease maps has been the object of many methodological developments in recent years but the analysis of edge effects, that is, the effects upon the analysis brought by the proximity of a boundary, has been a neglected area in spatial epidemiology. This is regrettable since many analyses can be fundamentally altered by the inclusion of edge effects in different forms. The aim of this work is to find out how the estimation of the mortality relative risk from a particular disease at or near boundaries can be affected by the edge position.

**Keywords and phrases:** Bayesian models, data augmentation, disease mapping, edge effects, MCMC

## 1.0 INTRODUCTION

Spatial analyses are usually undertaken within a finite region. This property of finiteness means that a boundary is present and that any geographic distribution or spatial interaction occurring within the region may extend beyond its boundaries. For example, if we were to compute the variability of some relative risk estimates close to the external boundary of the study area, we will find that the value is elevated. The reason for this is that there is little information available around the edge regions to make the variability small. This is not important for the calculation of, say, the standardized mortality ratios (SMRs), but can become important when methods that "borrow" information from neighbouring regions are used. In such a case, the areas outside the boundary are missing and estimates close to the edge, which must be based on available data within the study area, are likely to be statistically biased.

In general, in mapping exercises where statistical data are to be represented, edge effects are present and may need to be accommodated in the analysis (Lawson (2001)). Traditionally, the most effective strategies that have been adopted to remove or compensate for edge effects (Griffith (1983)) are:

- Map a finite surface onto a torus. This correction is appropriate in the case of stationary Poisson point processes within a study region of regular geometry (Ripley (1988)). However, it is not usually appropriate in the analysis of spatial disease data because either non-stationary in mean or covariance is likely to be encountered and often the study region is highly irregular.
- Use of correction methods, such as weights, for boundary areas (Lawson *et al.* (1999)). The weight for an observation usually acts as a surrogate for the degree of missing information at a location. This correction is appropriate when only a small proportion of the study window is close to the boundaries and only general parameter estimation is concerned.
- Construct an internal buffer zone along the border of the region. The area is used in the estimation process but it is excluded from the reporting stage, as it will be prone to edge effects itself.
- Construct an external buffer zone. Each unit in this external area could be assigned the mean value of the observed spatial distribution, could be extrapolated using a trend surface model or treated as

a missing value. This last method has significant advantages if used within iterative simulation methods such as data augmentation or general *Markov Chain Monte Carlo* (MCMC) algorithms (Gilks *et al.* (1996), Tanner (1996)).

It is usually straightforward to adapt conventional estimation methods to accommodate edge-weighted data. If guard areas are selected and observations are available within the guard area, it is possible to proceed with inference by using the whole data. When external guard areas are available but no data are observed, missing data methods can be used. An intermediate situation arises when some external rates are available. In that case, missing counts are regarded as parameters in a hierarchical model and are sampled iteratively within Gibbs-Metropolis sampling.

In this work, and in order to highlight the importance of considering edge effects in a mapping exercise, we explore in detail how the estimation of the mortality relative risk from a particular disease at or near boundaries can be affected by the edge position. Also, we are interested in the propagation of the error due to the boundary effects as we go further into the study region. For this purpose, we use MCMC data augmentation methods (Tanner (1996)) to obtain the estimates of the relative risks. Four different basic models are considered to find out how the effects of the edges vary when different risk patterns (from single risk gradients to more complex risk structures, included spatial correlation). Data sets are examined and to make the results more rigorous, a simulation study was set up.

## 2.0 DATA AND METHODS

For basic notation, let  $\zeta_i$  denote the unknown relative risk for the  $i^{\text{th}}$  area,  $i=1, \dots, n$ . Also, let  $(O_1, \dots, O_n)$  and  $(E_1, \dots, E_n)$  denote the number of deaths and the expected number of deaths, respectively, from the disease during the study period.

In the study of the geographical variation of disease risk in count data there are basic models for the estimation of the relative risks that, at least as a starting point, are usually applied. In this work four different models were assumed. First, the basic model assumed for  $(O_1, \dots, O_n)$  is a Poisson likelihood with parameter  $E_i \zeta_i$ . This is called the classical model (no prior knowledge). The other models we examine here (Gamma-Poisson model, lognormal model and Besag, York and Mollié models) are extensions of this model when prior distributions for the relative risks are assumed. These models range from simple risk structures to more complex risk structures, including spatial correlation. In the following, they are described in more detail.

### 2.1 Classical model of relative risk

This approach is based on the assumption that, conditional on the  $E_i$  being known, the  $\zeta_i$ s are mutually independent and each  $O_i$  follows a Poisson distribution with mean  $E_i \zeta_i$ . Under these assumptions, the maximum likelihood estimator of  $\zeta_i$  coincides with the standardized mortality ratio

$$\hat{\zeta}_i = SMR_i = \frac{O_i}{E_i}$$

### 2.2 Gamma-Poisson model

When the disease is non-contagious and rare, the numbers of deaths in each area are assumed to be mutually independent and to follow Poisson distributions:

$$O_i \sim \text{Poisson}(E_i \zeta_i) \quad \forall i$$

A gamma prior for the relative risks combines conveniently with the Poisson likelihood to give a gamma posterior distribution. If the prior distribution for the relative risk is a *gamma*( $\nu, \tau$ ) then, the relative risk has the following posterior distribution

$$\zeta_i \sim \text{gamma}(\nu + O_i, \tau + E_i)$$

Prior distributions for  $\nu$  and  $\tau$  are specified; we considered exponential distributions with mean 0.1 for both parameters.

### 2.3 Lognormal model

A gamma prior for the relative risk is mathematically convenient, but may be restrictive because covariate adjustment is difficult and there is no possibility for allowing spatial correlation between risks in nearby areas. A lognormal model for the relative risks is more flexible:

$$\begin{aligned} O_i &\sim \text{Poisson}(E_i \xi_i) \\ \log \xi_i &= \theta_i \\ \theta_i &\sim \text{normal}(0, \tau) \end{aligned}$$

A prior distribution for  $\tau$  should be specified; an inverse gamma distribution was considered.

### 2.4 Besag, York and Mollié model

In this model for the relative risks, area-specific random effects are decomposed into a component that takes into account the effects that vary in a structured manner in space (correlated heterogeneity, CH) and a component that models the effects that vary in an unstructured way between areas (uncorrelated heterogeneity, UH). The model, introduced by Clayton and Kaldor (1987) and developed by Besag *et al.* (1991), is formulated as follows:

$$\begin{aligned} O_i &\sim \text{Poisson}(E_i \xi_i) \\ \log \xi_i &= \theta_i + \phi_i \end{aligned}$$

where  $\theta_i$  is the uncorrelated heterogeneity and  $\phi_i$  is the correlated heterogeneity. Prior distributions should be specified for these parameters. See Besag *et al.* (1991) for details.

The models described above have been applied to data on deaths from lung cancer among American white males over the period 1970-94. We selected a region formed by ten adjacent U.S. states by county: Colorado, Iowa, Kansas, Montana, Minnesota, Missouri, Nebraska, North Dakota, South Dakota and Wyoming. This area was chosen because it had a reasonable regular shape and it is big enough (760 counties) to examine what are the effects of the edges as we move away from them. Analyses were carried out with the whole data set and with other data sets based on the previous one where the edge counts were treated as missing values and estimated within an iterative sampling algorithm.

## 3.0 SIMULATION STUDY

To assess if the previous models are good at recovering the true spatial variation and the relative risks at or near boundaries, we decided to simulate data sets from a number of different possible relative risk models. To this end, we use the same geographical area as before to simulate relative risks within. In addition, we required using a set of fixed expected counts for the area. We also made use of the expected number of deaths from lung cancer among white males for the period 1970-94.

The models for the true relative risks defined in this simulation study were selected to represent some of the possible underlying risk that might be encountered:

- Constant risk model

$$\xi_i = \xi = 1 \quad \forall i$$

- Lognormal model

$$\begin{aligned} \xi_i &= \exp\{\theta_i\} \\ \theta_i &\sim \text{normal}(0, \sigma_\theta^2) \end{aligned}$$

where  $\sigma_\theta^2=1$ .

- Gamma model

$$\xi_i \sim \text{gamma}(\alpha, \beta)$$

with parameters  $\alpha=1, \beta=1$ .

- Besag, York and Mollié model (CH+UH)

$$\xi_i = \exp\{\phi_i + \theta_i\}$$

$$\phi_i \sim \text{car.normal}, \theta_i \sim \text{normal}(0, \sigma_\theta^2)$$

where  $\sigma_\theta \sim \text{inverse gamma}(0.5, 0.0005)$  and the car.normal distribution as defined in Spiegelhalter *et al.* (2000).

- Besag, York and Mollié model (CH)

$$\xi_i = \exp\{\phi_i\}$$

$$\phi_i \sim \text{car.normal}$$

where the car.normal distribution as defined in Spiegelhalter *et al.* (2000).

The model fitting was examined using the Pearson's chi-squared measure (*RSS*).

## 4.0 RESULTS

*Notation:* Set  $I, I=1, \dots, 11$ , consists of the counties that are in the  $i^{\text{th}}$  border of the region (Set 1 is compound by edge counties and Set 11 are the most internal ones).  $RR-J, J=1, \dots, 5$  represents the relative risk obtained when counts for sets  $I, I=1, \dots, J$  are supposed to be missing and, therefore, simulated. In step 0 all data are used to estimate the relative risks. In step  $I, I=1, \dots, 5$  the relative risks  $RR-J$  are estimated.

When edge counts are unknown and then simulated, the relative risks in the internal counties are quite similar to those obtained using the complete data set (Figure 1).

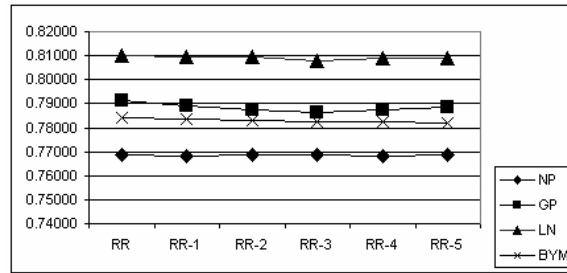


Figure 1: Mean of the relative risks in the internal counties.

In the edge and in those counties close to it, the values of the relative risks suffer a great variation (Figure 2) and they are highly dependent on the number of counties with simulated data in the edges, on the distance from such area to the edge of the region and on the model used to make the estimations (Figure 3).

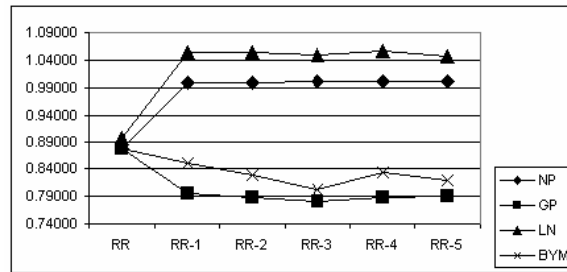


Figure 2: Mean of the relative risks in the edge counties.

In the situations with no prior knowledge and Gamma-Poisson models, the variation in the range of the relative risks in a particular set of counties does not depend on the number of simulated data sets; it only depends on if we know all information or we do not. However, in the lognormal and BYM situations there are differences in the range depending on the number of simulated sets. Also, the range of the relative risks is higher for the counties in Set 6 and it reduces as we go through Set 5 and Set 4 and then, it increases again. This is probably due to the fact

that the counties in Set 1 are already censored and so there may be less reduction for them while the inner sets are progressively censored.

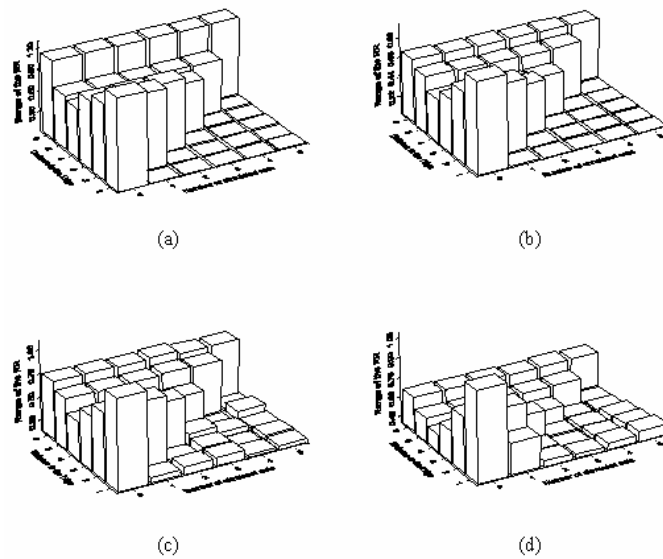


Figure 3: Range of the relative risk. (a) No prior knowledge, (b) Gamma-Poisson model, (c) lognormal model, (d) BYM model.

Geographical patterns of disease are also dependent on the number of counties with simulated data in the edges. Figure 4 shows the geographical distribution of the relative risks across the study region for the BYM model; similar patterns appear when one or two sets of counts are simulated (steps 1 and 2), and they are also similar to the pattern in step 0 although they are more smoothed. This can be due to the information of the remaining counties still influence the estimations in the edges. In the other cases, the pattern changes and it tends to make a very clear distinction between the northwest and southeast areas. Furthermore, the values of the relative risk change a lot in the northwest region depending on the fitting step.

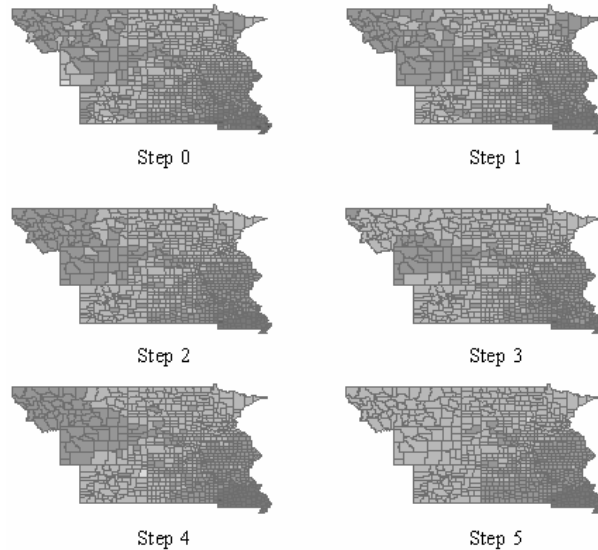


Figure 4: Geographical distribution of lung cancer mortality. BYM model, steps 0 to 5.

It is important to relate these previous results to the internal regions. This can be done by examining the relative risks in the internal regions and seeing how they change with changes in the edges.

Spatial smoothing methods, such as the BYM model, use data from different regions to estimate a value at a location. At or close to an edge, the values of the relative risks are estimated using information of the edge counties. Then, some distortion will result if counts are unknown and, therefore, estimated. If the counties are not very close to the edge (for example in regions that are 8, 9, 10, 11<sup>th</sup> in from the external edge) the values of the relative risk do not change too much. In cases where information in the neighbour counties is not considered, no prior knowledge model, the differences in the relative risk in the most internal areas are very close to zero. The Gamma-Poisson and the lognormal models are the ones where the differences are bigger in the internal counties. In those cases, differences increase until the 8<sup>th</sup> Set and then decrease a bit but they are still higher than in the 6<sup>th</sup> and 7<sup>th</sup> Sets.

#### 4.1 Results of the simulation study

The previous results are limited to the fact that we were examining a particular data set and some of the features may be related to the uniqueness of the data set rather than to general applications. The simulation study described in section 3.0 was set up to make the results more rigorous. Here, we present some preliminary results concerning the goodness of fit of a variety of disease mapping methods making comparisons regardless to the edge effects (Figures 5 and 6).

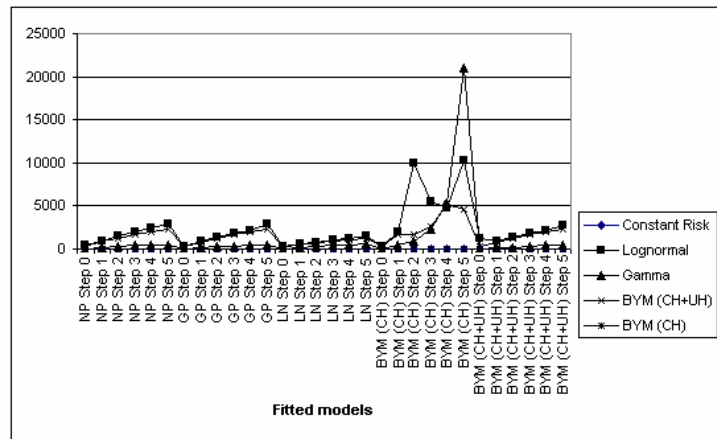


Figure 5: Residual sum of squares (RSS): Relative Risk comparison, all counties.

As we are examining the closeness of the fitted models to the true models, we should expect that, in general, the fitted models do well when recovering their equivalent true models.

In both edge and internal regions, as the number of counties with simulated counts increases, the residual sum of squares applied to the relative risks increases, that is, the less information we have on the edge, the worse the estimations are. Also, it appears that the BYM models (CH+UH) have the best behaviour across all the situations considered here.

## 5. CONCLUSIONS

In mapping geographical variation of risk or occurrence of disease in count data, information about boundary areas is often incomplete. As we showed in this work, this lack of information could distort the estimates of the relative risk in the study region. However, some results are limited to the fact that we are examining one data set and some of the features may be related to the uniqueness of the data set rather than general applications. The results of the simulation study can help us to be more rigorous when we try to make statements about edge effects.

Further work is needed to examine in depth the implications of the edge effects in the estimation of the relative risks. It will be worth to examine in detail the changes in the internal regions and derive some theoretical results concerning the edge effects found in this preliminary study.

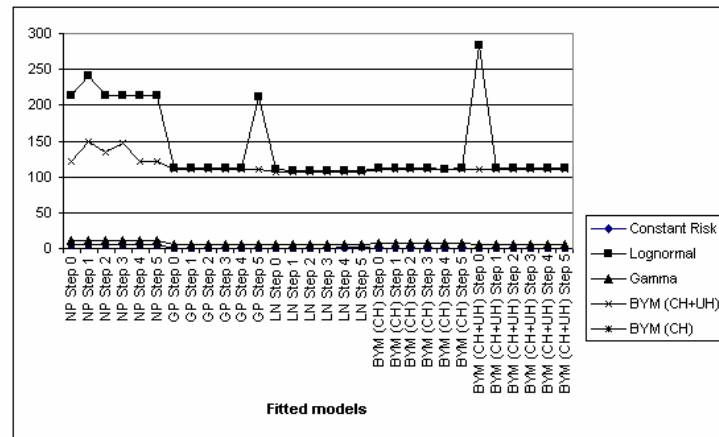


Figure 6: Residual sum of squares (RSS): Relative Risk comparison, internal counties.

Spatial censoring can also occur due to the inability to fully observe the complete time sequence of a disease. That is, some events could occur outside the spatial window during the time period studied but be unobserved. This is known as spatio-temporal censoring and should be taken into account in future works.

## ACKNOWLEDGEMENTS

This work was made possible by the support of the National Institutes of Health grant no: 5R01CA092693-2.

## REFERENCES

- Besag J., York J. and Mollié A. (1991) A Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43: 1-59.
- Clayton D. and Kaldor J. (1987) Empirical Bayes Estimates of Age-standardised Relative Risks for use in disease mapping. *Biometrics*, 43: 671-681.
- Gilks W.R., Richardson S. and Spiegelhalter D.J. (eds) (1996) *Markov Chain Monte Carlo in Practice*. Chapman&Hall, London.
- Griffith D.A. (1983) The boundary value problem in spatial statistical analysis. *Journal of Regional Science*, 23: 377-387.
- Lawson A.B. (2001) *Statistical Methods in Spatial Epidemiology*. Wiley, Chichester.
- Lawson A.B., Biggeri A., Dreassi A.E. (1999) Edge effects in disease mapping. In Lawson A.B., Boehning D., Lasaffree E., Biggeri A., Viel J.F. and Bertollini R. (eds). *Disease Mapping and Risk Assessment for Public Health*. Wiley, Chichester.
- Ripley B.D. (1988) *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge.
- Spiegelhalter D.J., Thomas A. and Best N.G. (2000) *WinBugs. Version 1.3. User Manual*. MRC Biostatistics Unit, Cambridge.
- Tanner M.A. (1996) *Tools for Statistical Inference*. Springer-Verlag, New York.