Bayesian Analysis of Geographical Variation in Disease Risk

C.L. Vidal-Rodeiro
Norman J. Arnold School of Public Health, USA

M.I. Santiago-Perez
Public Health Department of Galicia, Spain

E. Vazquez-Fernandez
Public Health Department of Galicia, Spain

M.E. Lopez-Vizcaino
Galician Statistical Institute, Spain

X. Hervada-Vidal
Public Health Department of Galicia, Spain

INTRODUCTION

Disease mapping is a big focus of interest in the area of public health (Bailey, 2001; Moore & Carpenter, 1999), and the geographical distribution of a disease has an important role in understanding its origin, its causes or its evolution.

In recent years, there have been many efforts to map mortality or incidence from diseases (see, for example, Lopez-Abente, Pollán, Escolar, Errezola & Abraira, 2001). The most widely used indicator in geographical representation is the standardized mortality ratio (SMR); this offers the advantage of eliminating the confounding effect of the variable by which it is adjusted, usually age, but presents certain drawbacks when the population size varies over the map (Breslow & Day, 1975). In such a case, estimators of different accuracy are obtained in each area; areas having small populations—and thus fewer cases—tend to register very extreme estimations of risk, which then dominate the map and hinder epidemiological interpretation. This is a particular problem for rare diseases where thousands of individuals are needed before a single case is expected to occur. This makes the necessity to utilize information from neighbouring areas in order to produce better estimates.

Alternative risk measures may be obtained by applying other modelling techniques that take the sources of spatio-temporal variation into account (Lawson, 2001). A simple technique consists of adjusting a Poisson regression model that displays a log-linear relationship between risk and space-time variables. While successful in reducing the variability in risk, this method continues to pose a number of drawbacks. First, in geographical areas having few cases, this model yields unstable estimations due to extra-Poisson variation. Furthermore, if the hypothesis of spatial independence between risks does not hold, the model is not appropriate, as it takes no account of a possible correlation between areas.

A possible solution to these problems is the Bayesian extension of the model introduced by Clayton and Kaldor (1987) and further developed by Besag, York, and Mollié (1991). Basically, this approach provides a way to integrate, in the estimation of the unknown relative risk, local information consisting of the observed and expected number of cases in each area and prior information on the overall variability of the relative risk, their potential similarity in neighbouring areas, and their connection with geographically defined covariates.

In this article, we compare the behaviour of the mentioned techniques with the purpose of estimating and mapping mortality relative risks in small geographical areas; this is illustrated by the analysis of the geographical variation in men lung cancer mortality in Galicia (Spain).

THEORETICAL MODELS

Classical Approach

Let O_i denote the number of observed cases, E_i the number of expected cases, calculated by using the population broken down by age for each area or geographical unit, plus the specific mortality rates, and let ξ_i be the relative risk (RR).

The classical approach to disease mapping is based on the assumption that, conditional on the E_i's being known, the ξ_i's are mutually independent. Moreover, each O_i follows a Poisson distribution with mean $E_i\xi_i$:

$$[O_i \mid E_i, \xi_i] \sim Poisson(E_i \xi_i)$$

Under these assumptions, the maximum likelihood estimate of ξ_i, denoted by $\hat{\xi}_i$, is the *SMR*:

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

Hierarchical Bayesian Model

Bayesian methods estimate the risk of an area by incorporating information from adjacent areas, so as to reduce the effect of random fluctuations unrelated to the risk. Furthermore, on taking account of spatial correlation between adjoining areas, the resulting smoothed maps prove more informative (Banerjee, Carlin & Gelfand, 2004; Richardson, 2003; Wakefield, Best & Waller, 2000).

In the Bayesian approximation, Poisson variation is modelled at a first level, and a model for the relative risks is specified at a second level, with area-specific random effects further decomposed into two components: a spatially structured component that takes into account the effects that vary in a structured manner in space (clustering) and a component that models the effects that vary in an unstructured manner between one area and the next (heterogeneity).

The spatio-temporal hierarchical model is formulated as follows (Bernardinelli, Clayton, Pascutto, Montomoli, Ghislandi & Songini, 1995):

Let O_{ij} be the number of observed cases, E_{ij} the number of expected cases, and ξ_{ij} the *RR* in area *i* and period *j*. A likelihood model is specified for the vector of observed cases, given the risk vector

$$[O_{ij} | \xi_{ij}] \sim \text{Poisson}(E_{ij}\xi_{ij})$$

and the *RR* is modelled as

$$\log \xi_{ij} = \alpha + \phi_i + \theta_i + (\beta + \delta_i)t_j + \Sigma \gamma_k X_{ki}$$

where α is the mean of the logarithm for *RRs* over all areas, ϕ_i the clustering component, θ_i the heterogeneity component, t_j the time, β the mean of the time trend across all areas, δ_i the space-time interaction effect, and X_k indicate the covariates, with coefficients γ_k.

Estimation of the risk across time in each area is given by $\exp(\beta + \delta_i)$. Following the Bernardinelli, Clayton, Pascutto, Montomoli, Ghislandi, and Songini (1995) notation, δ_i is named the differential trend for area *i*: a value of $\delta_i < 0$ indicates that the trend in area *i* is below the mean, while a value of $\delta_i > 0$ implies that the trend in area *i* is above the mean.

Bayesian modelling requires specification of prior distributions for random effects. Several prior distributions can be considered, and next, we will describe the specifications used in the practical application.

The distribution model for the heterogeneity component is

$$[\theta_i | \theta_j, i \neq j, \sigma_\theta^2] \sim \text{normal}(\bar{\theta}_{-i}, \sigma_\theta^2)$$

where

$$\bar{\theta}_{-i} = \frac{1}{I-1}\sum_{j \neq i} \theta_j, \quad I = \text{number of areas.}$$

By virtue of this prior distribution, it is assumed that variation in risk between areas is independent and, as a consequence, posterior estimations of the area effect will tend towards an overall mean.

For the clustering component, the estimations of the risk in any area depend on neighbouring areas; this was achieved by allocation of weights. Specifically, weights were taken equal to one in cases where the areas were adjacent (i.e., share a common boundary) and zero in cases where they were not. The conditional autoregressive (CAR) model proposed by Besag, York, and Mollié (1991) was used:

$$[\phi_i | \phi_j, i \neq j, \sigma_\phi^2] \sim \text{normal}(\bar{\phi}_i, \sigma_i^2)$$

where

$$\bar{\phi}_i = \frac{1}{\sum_j w_{ij}} \sum_j \phi_j w_{ij}$$

$$\sigma_i^2 = \frac{\sigma_\phi^2}{\sum_j w_{ij}}$$

$w_{ij} = 1$ if *i*,*j* are adjacent (or 0 if they are not).

Parameters σ_θ^2 and σ_ϕ^2 control the variability of θ and ϕ. Following the recommendations of Bernardinelli,