

Bayesian Hierarchical Model for Small Area Disease Mapping: a Breast Cancer Study

Stefano Marchetti, Claudia Dolci, Samantha Riccadonna and Cesare Furlanello

Abstract In this paper we propose a hierarchical Bayesian method to estimate the relative risk for female breast cancer at the municipality level in the province of Trento. To model the relative risk we use the so called convolution model for count data, which takes into account random and spatially correlated random effects (uncorrelated and correlated heterogeneity). The method is adopted to obtain reliable estimates of the relative risk in those areas where the low number of observations makes the *rough* relative risk estimates unstable (where rough means a relative risk estimated using only observations and target population data at a given area level). This Bayesian method can be applied to a wide range of problems. Here we apply this method to epidemiological data aiming at an on-line computational solution.

Key words: Small Area Estimation, Relative Risk, Hierarchical Bayesian Model, Province of Trento

1 Introduction

A much larger awareness of the connections between disease and environmental factors and the availability of computerised geographic information system (GIS) technology to manage and overlay geocoded data have certainly supported a growing interest for disease mapping methods. While collections of curated, longitudinal data are increasingly available for disease surveillance and control, it is still hard to downscale statistical models to a level closer to that of the local public admin-

Stefano Marchetti

Department of Statistics and Mathematic Applied to Economics, Via Ridolfi 10 - Pisa (PI) Italy,
e-mail: s.marchetti@ds.unifi.it

Claudia Dolci, Samantha Riccadonna and Cesare Furlanello

Fondazione Bruno Kessler, Via Sommarive 18 - Povo (TN) Italy, e-mail: Claudia Dolci -
dolci@fbk.eu, Samantha Riccadonna - riccadonna@fbk.eu and Cesare Furlanello - furlan@fbk.eu

istration. At the same time, a large number of environmental and socio-economic information from administrative databases could be automatically associated to epidemiological indicators and easily distributed as risk maps through Internet-based technologies.

In this paper, as a prerequisite to the development of an on line Cancer Atlas at the municipality scale, we focus on relative risk estimate at a small area level. Relative risk is intended as the risk to acquire a certain disease in a given area with respect to a common risk underlying the set of all the areas, e.g. the relative risk of a municipality with respect to their province. According to [6], we can approach the identification of epidemiological risk at the municipality level as a small area problem. Even major killers such as cancer or cardiovascular diseases can have high variability in a real small population: the number of observations relative to a disease in an area can be small, or even zero, while just few cases can show a highly concentrated burden. Low numbers of observations impact on rough estimates of the relative risk, so other solutions are needed besides resorting to upscaling. Bayesian methods can be used to handle these problems. In our study we present a prototype technical solution and its computational implementation for cancer mapping.

In Section 2 we present the convolution model for count data; we use this hierarchical Bayesian estimator to model small area relative risk. In Section 3 we describe an application to a case study for breast cancer data and we simulate its use at the municipality level starting from a well curated, longitudinal database from the Cancer Registry of the Province of Trento in which data are available at an higher level of aggregation (“comprendori”: departments of 10-15 municipalities).

The convolution model has been already applied to similar epidemiological data (see for example [3]). Our proposal implements the convolution model for breast cancer relative risk with a general model approach, that can be applied to other cancers and sub-domains (e.g. female or male, certain age classes, etc.). Moreover, we include the possibility of introducing auxiliary variables at the area level.

In the last Section, we discuss the role of heterogeneity in our model and a possible extension.

2 The Bayesian Estimation of the Relative Risk in a Small Area Framework

Data related to diseases are often given as count data and they are often modelled with the Poisson distribution. Here we consider the case where count data are available, for a given disease, at a given area level. In addition, we also assume to know auxiliary variables at the same area level.

An approximation of the risk to contract a disease in a given area is easily computed as the count of new cases of the disease in a specific time period divided by the population exposed to risk:

$$r = \frac{y}{N}, \quad (1)$$

where y is the count of new cases of the disease in the area and N is the population exposed to risk; y and N are referred to a specific time period. We suppose that the area considered is divided in m sub-areas (partitions). We are interested in the relative risk of the sub-areas with respect to the common risk in (1). An estimator, $\tilde{\theta}$, of the relative risk in a given sub-area is:

$$\tilde{\theta}_i = \frac{y_i}{E_i}, \quad (2)$$

where y_i is the observed count of a disease in the i th sub-area and E_i is the expected count of the same disease in the same sub-area. The expected count E_i is computed with respect to the common risk r , computed as in Eq. 1; so $E_i = rN_i$ (N_i is the population exposed to risk in the sub-area i).

If in the sub-areas we observe too few cases, then the estimated relative risk in Eq. (2) is very unstable and can lead to wrong conclusions. When a population exposed to risk is small the number of observed cases of a disease can be affected by a large variability due only to chance, and not due to environmental variables or particular behaviour of the target population. For example let us consider two similar sub-areas, a and b , with 3000 people exposed to risk ($N_a = N_b = 3000$). We assume that areas are close and people are exposed to the same environmental risks and have the same behaviour (e.g. smoke, alcohol, drug abuse, etc.). Suppose that the common risk for the considered disease present in the main area is 0.05%, so the expected cases in each area are $0.0005 \times 3000 = 1.5$. Imagine that in sub-area a we observe 2 cases while in sub-area b we observe 0 cases. The relative risk of sub-area a is $\tilde{\theta}_a = 2/1.5 = 1.33$ while in sub-area b is $\tilde{\theta}_b = 0/1.5 = 0$. So, in sub-area a there is a higher risk than in the main area (relative risk greater than one) and in sub-area b there is no risk. The difference is due only to chance since we supposed identical conditions in the two sub-areas.

To handle this problem we need a risk model that allows for reliable relative risk estimator (in our example relative risks of area a and b must be similar).

Bayesian estimators are widely used in order to obtain reliable estimate for the relative risk when there are sub-areas with small population and rough (traditional) estimates of relative risk (as in (2)) lead to unreliable results.

In what follow we refer to the set of all areas as main area and to the m partitions as small areas.

Let y_i be the number of count of a given disease in the small area i and let N_i be the population (exposed to risk) in small area i .

We assume that y_i is Poisson distributed with mean μ_i and likelihood:

$$P(y_i|\mu_i) = \prod_{i=1}^m \frac{\mu_i^{y_i} \exp\{-\mu_i\}}{y_i!}.$$

Under the assumption that data are independently distributed and using the relationship expressed in Eq. (2) we can write:

$$E[y_i] = \mu_i = E_i \theta_i, \quad (3)$$

where θ_i is the true and unknown relative risk of small area i (with respect to the main area). E_i is the same as in Eq. (2).

Our objective is to estimate θ_i and provide a confidence interval that allows for testing the hypothesis that θ_i is equal to one or not.

Under the assumption that the relative risk varies smoothly in the space we use the Bayesian convolution model to estimate the relative risk. This is a Bayesian hierarchical model for count data that takes into account for correlated and uncorrelated heterogeneity between small areas.

The [2] parametrization is:

- $y_i | \theta_i \sim \text{Poisson}(\theta_i E_i)$, conditional distribution of y_i ,
- $\eta_i = \log \theta_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i + v_i$, link function,
- $\mathbf{x}_i' \boldsymbol{\beta}$, trend of fixed covariate component,
- $u_i \sim N(0, \tau_u^2)$, uncorrelated heterogeneity,
- $v_i | v_{-i} \sim N(\sum_{j \sim i} v_j / \delta_i, \tau_v^2 / \delta_i)$, correlated heterogeneity,
- $\tau_u \sim \text{Gamma}(\psi_{u1}, \psi_{u2})$, prior distribution on τ_u ,
- $\tau_v \sim \text{Gamma}(\psi_{v1}, \psi_{v2})$, prior distribution on τ_v ,
- δ_i , number of neighbours of the area i ,
- E_i , expected count for the area i .

The spatially correlated heterogeneity follow an *intrinsic* Conditional Autoregressive model, a particular specification of the Conditional Autoregressive model (CAR):

$$\mathbf{v} \sim N_m(\boldsymbol{\mu}, \tau_v^2 \boldsymbol{\Sigma}),$$

where $\mathbf{v} = (v_1, \dots, v_m)$ is the vector of the random effects, $\boldsymbol{\mu}$ is the mean vector and $\tau_v^2 \boldsymbol{\Sigma}$ is the variance-covariance matrix. τ_v^2 controls the overall variability of the v_i and $\boldsymbol{\Sigma}$ is an $m \times m$ positive definite matrix.

The variance-covariance matrix has the following structure:

$$\tau_v^2 \boldsymbol{\Sigma} = \tau_v^2 (\mathbf{I} - \gamma \mathbf{C})^{-1} \mathbf{M},$$

where $\mathbf{I}$ is an $m \times m$ identity matrix, $\mathbf{M}$ is an $m \times m$ diagonal matrix, with elements M_{ii} proportional to the conditional variance $v_i | v_{-i}$. $\mathbf{C}$ is an $m \times m$ weight matrix, with elements C_{ij} reflecting spatial association between areas i and j . γ controls the overall strength of spatial dependence. Giving these definitions it is possible to show that the joint multivariate Gaussian model can be expressed in the form of a set of conditional distributions

$$v_i | v_{-i} \sim N(\mu_i + \sum_{j \sim i} \gamma C_{ij} (v_j - \mu_j), \phi M_{ii}).$$

The use of this conditional specification requires only to specify matrices $\mathbf{C}$ and $\mathbf{M}$ and the spatial dependence parameter γ ¹.

The intrinsic CAR model, proposed by [2], sets $C_{ij} = 1/\delta_i$ if areas i and j are adjacent and $C_{ij} = 0$ otherwise (with $C_{ii} = 0$), $M_{ii} = 1/\delta_i$ and $\gamma = 1$. Giving these settings we obtained our proposed intrinsic CAR model specification.

¹ Various constraints are needed on the values of $\mathbf{C}$, $\mathbf{M}$ and γ in order to ensure that $\boldsymbol{\Sigma}$ is symmetric positive definite. See [1] for details

Using the MCMC method we can obtain estimates and predictions with their confidence intervals for each of the parameters present in the model, according to the Bayesian inference technique.

3 A Breast Cancer Study in the Province of Trento

The application that motivated this approach is the potential mapping of oncology risk for the province of Trento at high resolution. In this study, we consider an estimate of the breast cancer relative risk for women for all municipalities. The province of Trento has about 500,000 inhabitants subdivided into 223 municipalities. Amongst them, 210 municipalities have a population that is less than 5,000 inhabitants. We consider small area methods to obtain a reliable estimate of the relative risk. Note that as an additional complication, several municipalities are further divided into two or three parts (administrative islands), for a total of 262 different small areas. For our purposes, a small area is a municipality or a part of a municipality. This definition must be taken into account to create the contiguity matrix used for the intrinsic CAR model.

To estimate breast cancer relative risk for women we used data provided by the Epidemiological Observatory together with demographic data collected by the ISTAT. However, these downscaled data can not be used for clinical applications but only for simulation purposes. These data give us the number of female breast cancer registered in each municipality from 1999 to 2002 and the female population in the same years. We consider these data as population data measured without errors.

We estimate the relative risk following the model presented in Section 2. In order to obtain a comparison with the estimated relative risks we also compute the rough relative risks following the formula in Eq. 2.

Let y_i be the number of female breast cancer cases registered in the i th small area and let N_i be the female population in the small area i , $i = 1, \dots, m = 262$. First of all we compute the risk of female breast cancer in the province of Trento, that is the underlying common risk: $r = \sum_{i=1}^m y_i / \sum_{i=1}^m N_i$. Given the common risk we compute the expected number of cancer cases in each small area: $E_i = rN_i$. To estimate the relative risks, θ_i , we specified the Bayesian convolution model as suggested by [6]:

- $y_i | \theta_i \sim \text{Poisson}(\theta_i E_i)$, the likelihood of y_i
- $\log \theta_i = \alpha + u_i + v_i$, the link function,
- $u_i \sim N(0, \tau_u^2)$, the uncorrelated heterogeneity,
- $v_i | v_{-i} \sim N(\sum_{j \sim i} v_j / \delta_i, \tau_v^2 / \delta_i)$, the correlated heterogeneity,
- $\tau_u \sim \text{Gamma}(\psi_{u_1} = 10, \psi_{u_2} = 10)$, the hyperprior distribution on τ_u ,
- $\tau_v \sim \text{Gamma}(\psi_{v_1} = 1000, \psi_{v_2} = 1000)$, the hyperprior distribution on τ_v .

According to the Bayesian inference technique we obtain the marginal posterior distribution for the parameters of interest, the relative risks. We use the MCMC method based on Gibbs sampling with single chain with 30,000 iterations and 20,000 burn-in. According to the Geweke ([4]), Heidelberg-Welch ([5]) and Raftery-Lewis ([7]) tests, the chains converged for all the relative risks and for most of the parameters present in the model as shown in Table 1. All computations were run

in the R statistical environment, calling the the WinBugs framework for the Gibbs sampling algorithm. For the point estimates of the relative risks, we use the mean of the θ values generated by the converged MCMC (as usual), while for confidence intervals we take the appropriate quantiles from the converged chain.

Table 1 Test statistics for MCMC convergence. Percentage of tests passed (789 parameters).

Geweke (Z-value)	Heidelberg-Welch	Raftery-Lewis (I)
75.67	98.76	88.78

Main results are summarised in Table 2, which shows the distribution over the 262 small areas of the Bayesian estimates of the breast cancer relative risk for women in the province of Trento. The rough estimates obtained by Eq. 2 are listed for comparison.

Relative risk was not estimated for 39 non inhabited small areas (all administrative areas). For the remaining 223 small areas, Bayesian estimates of the relative risk range from a minimum of 0.59 to a maximum of 1.23 while the rough estimates range from 0 to 4.4. The rough relative risks are unacceptable estimates for several areas. A value of 0 means that in that area there is no risk at all for a female to have a breast cancer. The rough relative risk is null in 39 of the 223 inhabited small areas, so practitioners who use these estimates can be misled². Indeed, these 39 areas are sparse over the province and there is no evidence to think that relative risk shall be different from neighbours areas. On the other side, a relative risk of 4.4 means that females living in that area have 4.4 time the risk with respect to a female living in the province of Trento (the main area). Although specific genetic isolates can be present and cannot be excluded in an Alpine population, such risk peaks are also unlikely and should be in general corrected.

An advantage of the Bayesian approach is that it is easy to compute confidence intervals for the estimates. Confidence intervals for the relative risks estimates are useful to check if in a given small area the relative risk is different from one. In our application (simulated data) only two small areas have a relative risk statistically different from one at a 95% confidence level: the municipalities of “Canazei” and “Rovereto”. The estimated relative risk for Canazei has a confidence interval that ranges from 0.31 to 0.96 with a punctual estimate of 0.59 while Rovereto estimated

Table 2 Bayesian and rough (Eq. 2) relative risks estimates for female breast cancer in the municipalities of the province of Trento. Results are summarised over 223 inhabited areas. Data simulated by downscaling according to ISTAT 2000.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Relative Risk (Bayes)	0.59	0.90	0.96	0.96	1.02	1.23
Relative Risk (rough)	0.00	0.45	0.86	0.91	1.22	4.42

² These 39 small areas are not the 39 uninhabited small areas.

relative risk ranges from 1.02 to 1.37 with a point estimate of 1.19. Confidence intervals and point estimates are shown in figure 3.

The map in Figure 2 shows the rough and the Bayesian estimates of the relative risk for the breast cancer in the province of Trento. In Figure 2 relative risks are higher where colour is darker.

4 Concluding Remarks

We used a hierarchical Bayesian method to estimate the female breast cancer relative risk in the municipalities of the province of Trento. Statistical models are needed in order to obtain reliable estimates of the relative risk in a small area.

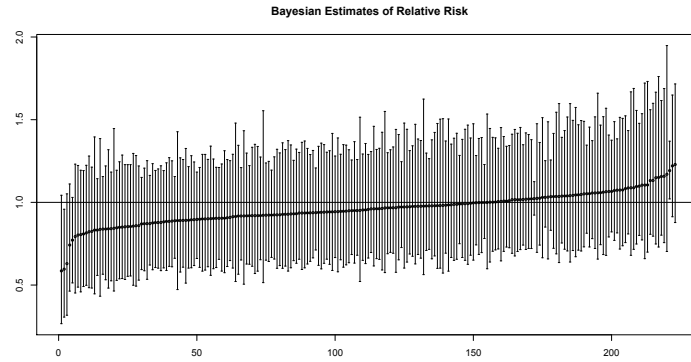


Fig. 1 Bayesian point estimates and confidence intervals (95%) for the relative risk of the female breast cancer in the province of Trento. Small areas are ordered from minimum to maximum estimated relative risk. Data simulated by downscaling according to ISTAT 2000.

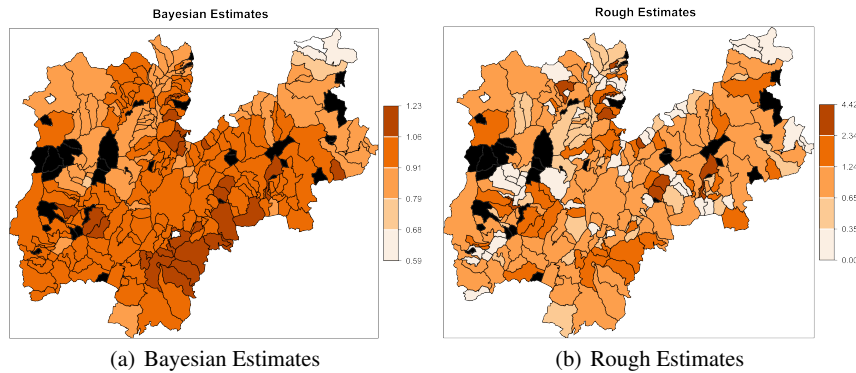


Fig. 2 Bayesian (a) and rough (b) estimates for the relative risk of the female breast cancer in the province of Trento. Data simulated by downscaling according to ISTAT 2000.

The model took into account correlated and uncorrelated heterogeneity. Instead of a CAR model with simple contiguity matrix used in this application, a significant improvement should include time effects and the use a different area correlation structure or distance function. Moreover, a sensitivity analysis on hyperpriors specifications for our application should be added to the system.

The application is designed to support an online cancer atlas framework. It has been designed to be (i) as general as possible in order to be applied to other cancers and target populations, and (ii) having computational times compatible with an on-line application. A complete demo prototype has been developed that includes the statistical application discussed in this paper. The statistical computing components are currently available in batch mode, but a on-line prototype is under construction at the moment of writing this note ³.

Acknowledgements The research described in this paper is part of the *Oncocure* project, funded by Caritro Foundation (Trento, Italy). We thank E. Galligioni and S. Piffer (APSS Trentino) for providing the motivating example and data, C. Eccher and S. Forti (FBK) for supporting the use of spatial data within Oncocure. We sincerely thanks colleagues at the FBK-MPBA Research Unit with a special thank to Shamar Droghetti, Riccardo De Filippi, Giuseppe Jurman.

References

1. Besag, J., Kooperberg, C.: On conditional and intrinsic autoregressions. *Biometrika* **82**, 733–746 (1995)
2. Besag, J., York, J., Mollié, A.: Bayesian image restoration with two application in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1–59 (1991)
3. Cressie, N., Chan, N.: Spatial modelling of regional variables. *Journal of the American Statistical Association* **84**, 393–401 (1989)
4. Geweke, J.: Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: J. Bernardo, J. Berger, A. Dawid, A. Smith (eds.) *Bayesian Statistics*, vol. 4. Oxford University Press (1992)
5. Heidelberg, P., Welch, P.: Simulation run length control in the presence of initial transient. *Opns Res.* **31**, 1109–1144 (1983)
6. Lawson, A.: *Bayesian Disease Mapping. Hierarchical Modeling in Spatial Epidemiology*. CRC Press (2009)
7. Raftery, A., Lewis, S.: The number of iterations, convergence diagnostics and generic metropolis algorithms. In: W. Gilks, D. Spiegelhalter, S. Richardson (eds.) *Practical Markov Chain Monte Carlo*. Chapman and Hall, London (1995)

³ Further information on the online application are available upon request.