

A Hierarchical Bayesian Model for Relative Risk Estimation in presence of Spatially Misaligned Data¹

Un Modello Bayesiano Gerarchico per la Valutazione del Rischio Relativo in presenza di Dati Non Allineati

Fedele Greco

Dipartimento di Scienze Statistiche “Paolo Fortunati”

Università di Bologna

greco@stat.unibo.it

Riassunto: In questo lavoro si propone un approccio di tipo bayesiano gerarchico per la valutazione del rischio per la salute derivante da esposizioni ambientale. In particolare viene considerato il caso in cui i dati per l'esposizione e quelli per i conteggi di eventi malattia/decesso sono rilevati con una differente risoluzione spaziale e devono quindi essere riallineati per rendere possibile la stima di un modello di regressione.

Keywords: Spatially Misaligned Data, Kernel Smoothing, Hierarchical Bayesian Models, Relative Risk, Disease Mapping.

1. Introduction

Ecological regression studies are often performed in geographical epidemiology to increase the understanding of disease aetiology, investigating relationships between health hazard and putative risk factors.

The available data usually include diseases/deaths counts, measurements on pollutants or other risk factors and the reference population at risk, along with other covariates describing socio-demographic features. While disease counts are often measured at small area level because of confidentiality restriction, environmental exposure is often measured by a monitoring network producing point referenced data, i.e. independent variable (exposure) and response variable (counts) are spatially misaligned. The relationship between relative risk and environmental exposure can be investigated only when such misalignment is resolved estimating exposure levels over the small areas.

A natural approach to deal with the misalignment problem, taking proper account of the variability induced from the covariate estimation, is to apply hierarchical Bayesian models where the exposure and the relationship exposure/relative risk are modeled on different levels of the hierarchy. Hierarchical Bayesian models are often used in the relatively young literature about the misalignment problem (Zhu et al. 2003, Mugglin et al. 2000). In these models the exposure field is modeled under the hypotheses of stationarity and isotropy that usually do not hold in empirical data. We propose an alternative solution based on the kernel smoothing technique: no hypothesis such as

¹ The research leading to this paper has been partially funded by a 2002 grant (Sector 13: Economics and Statistics) for research project of national interest by MIUR. The author thanks the South Carolina Department of Health and Environmental Control and Tom Temples for data availability. I am grateful to Daniela Cocchi and Andrew B. Lawson for their suggestions and supervision.

stationarity or isotropy is required and the model is computationally much simpler than other models proposed in literature.

2. The relative risk model

Bayesian hierarchical models are widely used in disease mapping: prior information on the disease rates variability are considered allowing to take proper account of unstructured heterogeneity as well as spatially structured variations in the counts.

Let λ_i , o_i , e_i , u_i be respectively the log-relative risk, the observed counts, the expected counts and the region-specific explanatory covariate measurement in the i -th small area, $i = 1, \dots, n$. The counts o_i are modelled as Poisson random variables with mean $e_i \exp(\lambda_i)$. The following linear model is employed for the log-relative risk:

$$\lambda_i = \alpha + \beta u_i + v_i + s_i \quad (1)$$

The regression coefficient β captures the relationship between log-relative risk and exposure and is the parameter of interest if the model is built for evaluating the exposure/relative risk relationship. The terms v_i and s_i control the structured and unstructured heterogeneity respectively. For s_i , independent identically distributed normal priors with zero mean and precision σ_s^{-2} are usually chosen. The structured heterogeneity term v_i is introduced when the count data show spatial patterns. Prior knowledge can be expressed via a nearest neighbour Markov Random Fields (MRF). The MRF we employ in the model is the intrinsic Gaussian Conditional Autoregression (CAR) (Besag et al. 1991) parameterised by the precision parameter σ_v^{-2} .

In this work we consider the misalignment problem arising when exposure measurements are available on a point-referenced grid while counts are available at the small area level. To estimate the parameters of the regression model (1) we need the pairs (λ_i, u_i) , $i=1, \dots, n$, but the exposure data consist in a set of p measurements on p points over the whole study region. Let u_j^o and u_i^a be respectively the value of the putative risk factor measurement in the j -th monitoring site, $j=1, \dots, p$ and the unknown covariate level in the i -th area, $i=1, \dots, n$. For obtaining the realignment, the predictive distribution of the exposure level in the small area given the observed values $[u^a | u^o]$ will be used.

A model is needed to describe the spatial process underlying the exposure field. Let θ be the set of parameters of such model. The target predictive distribution is obtained by integrating out the model parameters from the distribution $[u^a | u^o, \theta]$. Such procedure allows taking proper account of the variability induced by the estimation of the vector θ . The variables $u_i^a | u^o$ are the covariates in the regression model (1) that becomes:

$$\lambda_i = \alpha + \beta(u_i^a | u^o) + v_i + s_i \quad (2)$$

In the next paragraph a kernel-based procedure to obtain the predictive distribution $[u_i^a | u^o]$ is described.

3. A kernel-based model

Let $u(s)$ denote the exposure level in a generic point s and let A_i denote the i -th small area. The average exposure level we seek to estimate can be obtained by integrating the covariate level over the i -th area as follows:

$$u_i^a = \bar{u}(A_i) = \frac{1}{|A_i|} \int_{A_i} u(s) ds \quad s \in A_i \quad (3)$$

where $|A_i|$ is the i -th region area. Estimation of the unknown values $u(s)$ is performed via the Nadaraya-Watson estimator (Watson 1964): each estimate is a weighted mean of the data u^o with weights decreasing with the distance. Let $u^{NW}(s)$ be the Nadaraya Watson estimate of the exposure level in the point s .

The integral (3) cannot be computed analytically and is approximated via Monte Carlo integration drawing a set of points independently and uniformly distributed over each small area in the study region. Let $\{s_{ik}\}_{k=1, \dots, P_i}$ be the set of simulated points in the area A_i . The mean estimated exposure level for this area is:

$$u_i^{aNW} = \bar{u}^{NW}(A_i) = \frac{1}{P_i} \sum_{k=1}^{P_i} u^{NW}(s_{ik}) \quad s_{ik} \in A_i \quad (4)$$

To obtain the Nadaraya Watson estimates, the smoothing parameter has to be selected; this parameter has the role of θ . We follow a model-based approach based on cross-validation for bandwidth estimation. Let $\{u_{-j}^o\}$ be the set of observed covariate levels in all monitoring points but the j -th. Each observation u_j^o is supposed to be generated from the kernel density based on all other observations. The cross-validation likelihood is specified as $\prod_{j=1}^p [u_j^o | \{u_{-j}^o\}, h]$, i.e. conditional independence is assumed. The likelihood is specified as a normal distribution with mean equal to the expectation of Nadaraya-Watson estimator and variance σ_{NW}^2 .

$$\prod_{i=1}^p [u_j^o | \{u_{-j}^o\}, h] = \prod_{i=1}^p N(u_j^{oNW}, \sigma_{NW}^2) \quad (5)$$

The model-based procedure for the bandwidth selection is completed by the specification of a uniform prior probability for $\eta = 1/h$ with parameters d and r : $\eta \sim U(d, r)$. The posterior distribution for the bandwidth turns out to be proportional to the product of the normal likelihood and the uniform prior. Once the kernel function has been specified, the distribution of the random variables u_i^a depends only on the bandwidth and the observed data. Coherently with the likelihood specification, we use the predictive distribution:

$$[u_i^a | u^o, \eta] \sim N(u_i^{aNW}, \sigma_{NW}^2) \quad (6)$$

The inverse bandwidth is integrated out to obtain the predictive distribution $[u_i^a | u^o] = \int [u_i^a | u^o, \eta] [\eta] d\eta$.

4. Application

We illustrate the model by assessing the relationship between exposure to uranium present in drinkable waters and cancer incidence in South Carolina, USA. The data set encloses 39 small areas (zip code) in the north of South Carolina. Counts of total cancer o from 1996 to 1999 are available for each area. In addition the dataset includes measurements of uranium levels in drink water (u^o) in 281 monitoring sites. We assume that hyperparameters σ_{NW}^{-2} , σ_s^{-2} , σ_v^{-2} are independently distributed *a priori* as *Gamma* distributions: $\sigma_{NW}^{-2} \sim \text{Gamma}(0.01, 0.01)$, $\sigma_s^{-2} \sim \text{Gamma}(3.8, 0.011)$ and $\sigma_v^{-2} \sim \text{Gamma}(1.24, 0.02)$. For the inverse bandwidth $\eta = 1/h$ a *Uniform*(0.3, 100) prior has been chosen. This prior should have the effect of penalising against high values of the bandwidth: values greater than 3.33 for h have 0 prior probability. Such prior seems not too restrictive because of the low variability in the uranium measurements: a value of $h > 3$ produces total smoothing of the uranium surface. The model has been implemented using the WinBugs software. Convergence in Monte Carlo Markov Chain (MCMC) sampling algorithm has been quite fast for all the parameters.

Table 1: *Posterior distributions descriptive statistics*

	Mean	St. Dev.	2.50%	Median	97.50%
h	0.0973	0.0352	0.0527	0.0902	0.1828
β	-0.0061	0.0060	-0.0179	-0.0060	0.0053
σ_{NW}	1.3139	0.0574	1.2090	1.3110	1.4320

Bayesian credibility interval for the regression parameter β includes 0, showing no evidence of correlation between relative risk of cancer and exposure to uranium in drinkable waters. The model is computationally less expensive than other model proposed in literature, mainly because no matrix inversion is required for estimation. Because computation is fast the model can be implemented with a fully Bayesian approach, taking account of all the sources of variability.

References

- Besag J., York J., Mollié A. (1991) Bayesian image restoration, with two applications in spatial statistics, *Ann. Inst. Statist. Math.*, 43, 1-21.
- Mugglin A.S., Carlin B.P., Gelfand A.E., (2000) Fully model-based approaches for spatially misaligned data, *Journal of American Statistical Association*, 95, 877-887.
- Watson G., (1964) Smooth regression analysis, *Sankhya, series A*, 26, 359-372.
- Zhu L., Carlin B.P., Gelfand A.E. (2003) Hierarchical regression with misaligned spatial data: relating ambient ozone and paediatric asthma ER visits in Atlanta, *Environmetrics*, 14, 537-557.