AUSTRIAN JOURNAL OF STATISTICS
Volume 28 (1999), Number 1, 21–28

Bayesian Smoothing of Lung Cancer Data in Tirol, Salzburg and Vorarlberg

Rose-Gerd Koboltschnig
Institut für Mathematik, Statistik und Didaktik der Mathematik
Universität Klagenfurt

Abstract: Due to the high variability of ML-estimates of relative risk in low-population areas incidence ratios have to be smoothed before mapping. We fit a Bayesian hierarchical model where the posterior distribution of relative risks is simulated via a Markov Chain Monte Carlo technique.

Zusammenfassung: Wegen der hohen Variabilität von ML-Schätzungen des relativen Risikos in Regionen mit niedrigen Bevölkerungszahlen müssen diese vor Verwendung in Krebsatlanten geglättet werden. Am Beispiel von Lungenkarzinomdaten wird die Anwendbarkeit eines Bayes'schen Modells unter Verwendung spezieller Software zur Simulation der a-posteriori-Verteilungen gezeigt.

Keywords: Cancer Maps, Smoothing, Incidence Rates, Bayesian Model, BUGS and CODA, MCMC.

1 Introduction

Disease maps are a valuable tool in descriptive epidemiology. They are helpful in obtaining insight into the geographical distribution of disease occurrence and in identifying clusters of unusual high or low risk. Such maps can provide useful suggestions about the underlying causes of diseases (see, e.g. Zatonski et al., 1996, Pesch et al., 1994).

If the disease is rare or the geographical regions are small, the observed incidences O_i in a region i can be thought to be mutually independent outcomes from a Poisson distribution.

Mapping of maximum-likelihood estimates of relative risk can be misleading due to the fact that the most extreme estimates in such maps are those based on low population.

One of several strategies for dealing with this problem is to use a Bayesian hierarchical model which combines the information contained in the data and prior information on the relative risks, expressed in form of a prior distribution.

Full Bayesian analysis is done via Gibbs sampling, a Markov Chain Monte Carlo technique that can be used to simulate the posterior distribution of the relative risk.

2 Motivation, Data and Model

2.1 Motivation

In 1998 Oberaigner et al. (1998) have published a disease atlas for western Austria, containing data on cancer incidences and mortality ratios for the municipalities of Salzburg, Tirol and Vorarlberg. For each region i they computed E_i, the number of incidences one

could expect in region i, if the cancer rate would be the same as throughout the country. The standardised incidence ratio then is

$$SIR_i = \frac{O_i}{E_i}, \tag{1}$$

where O_i is the total number of incidences in region i (Fisher and Van Belle, 1993).

The "raw" *SIRs* show high variability: they range from 0, when no cancer incidences were observed, to over 30 in a region where only one case occurred. Ratios can be stabilised when using larger geographical units, e.g. districts, but then regions with higher risk will not be identifiable any more, because municipalities with lower rates in the same district will hide these effects. So they decided to map the ratios for the municipalities, but also to produce smoothed maps which are more reliable. To smooth the extreme values, they used a local mean estimator (Bowman and Azzalini, 1997) where neighbouring regions correct the value in region i towards a local mean.

These data were placed to our disposal by Dr. Oberaigner, so that we could try a different approach: a Bayesian approach (see, e.g. Besag et al., 1991). Our aim was to take a simple Bayesian model to estimate relative risks, which lie between the *SIR*s and a local mean, and to test whether this approach is suitable for a high number of regions where the observations are rare and - as an additional problem - the neighbourhood structure is rather inhomogeneous. The municipalities of Austria do not show a geometrical pattern as, e.g. the departments in France.

The results represented here are part of my doctoral thesis (Koboltschnig, 1998).

2.2 The Data

We focused on lung-cancer data for men, which consist of observed incidence counts per municipality for 18 age-groups for the period from 1988 to 1992.

In Figure 1 the Standardised Incidence Ratios (see formula (1)) for each of the 493 municipalities are shown. Due to the fact that lung-cancer is fairly rare, a lot of regions have a *SIR* of zero; but some municipalities have a high *SIR*. The highest value (5.69) can be found in one municipality of Tirol: four cases of lung-cancer occurred, but only 0.7 were expected.

2.3 The Model

Let θ_i denote the true but unknown relative risk in region i. The posterior distribution of relative risks, given the data, is proportional to the likelihood function of relative risks given the observations, times the prior distribution of the relative risks,

$$\pi(\theta|O) \sim L(O|\theta)p(\theta). \tag{2}$$

We use a conditionally independent Poisson model with conditional expectation

$$\mu_i = E_i\theta_i \tag{3}$$

for municipality i, $i = 1, \ldots, N$. Taking the logarithm, this model can be written as

$$\log \mu_i = \log E_i + r_i, \tag{4}$$

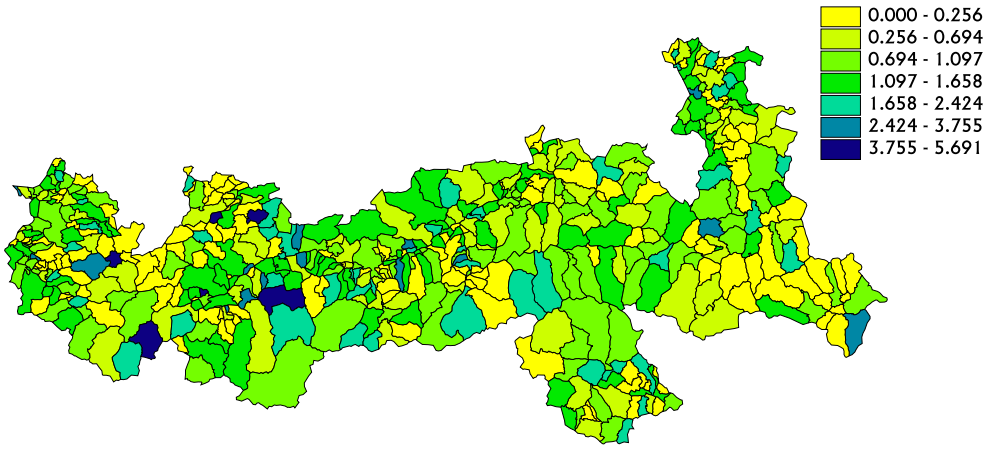

Figure 1: SIRs for lung-cancer for men in Western Austria, 1988-1992

with $r_i = \log \theta_i$.

Since the O_i's can be thought to be conditionally independent given the risk θ_i we get

$$L(O|\theta) = \prod_{i=1}^{N} L(O_i|\theta_i). \tag{5}$$

Our prior belief in the risk is formulated in terms of a prior distribution. We take a normal prior distribution with hyperparameters μ and σ^2 for the log-relative risks r_i. This prior distribution reflects the assumption of exchangeable log-relative risks

$$p(r|\mu, \sigma^2) = \prod_{i=1}^{N} p(r_i|\mu, \sigma^2), \tag{6}$$

where the prior distribution is the same for every region. This distribution can be easily generalised to allow for the possibility of spatial correlation (Clayton and Kaldor, 1987).

If prior knowledge gives reason to assume that adjacent regions tend to have similar risks we can express this in using a normal prior distribution on the log-relative risks where the relative risks have a locally dependent error structure. As noted in Besag et al. (1991) or Mollié (1996), Gaussian intrinsic autoregression, which allows for a non-constant conditional variance, is appropriate when the number of neighbouring regions is varying throughout the map as is the fact in Western Austria.

In our case where two regions have weight 1 if they are adjacent (= if they share a common boundary) and weight zero else, the normal conditional prior distribution of r_i given all other r_j and the hyperparameter σ^2, has mean

$$\mathbb{E}(r_i|r_{-i}, \sigma^2) = \bar{r}_i, \tag{7}$$

where $\bar{r}_i$ is the mean of the r_j in regions adjacent to region i, and r_{-i} denotes the log-relative risk in all areas $j \neq i$ and variance

$$Var(r_i|r_{-i}, \sigma^2) = \frac{\sigma^2}{n_i}, \tag{8}$$

where n_i denotes the number of regions adjacent to region i.

3 Simulating the Posterior Distribution

Because of the complexity, full Bayesian modelling seldom can be done directly, but Markov Chain Monte Carlo simulation can be used to simulate the posterior distribution.

Monte Carlo integration evaluates the expectation of a function g with respect to a density p

$$\mathbb{E}_p(g) = \int g(\theta)p(\theta)d\theta \tag{9}$$

by drawing samples $\theta^{(t)}$, $t = 1, \ldots, n$, from the posterior distribution $p(\theta)$ and then approximating the population mean through a sample mean

$$\frac{1}{n}\sum_{t=1}^{n} g(\theta^{(t)}). \tag{10}$$

If the samples are independent, the laws of large numbers ensure that this sum can be made as accurate as wanted (Gilks et al., 1996).

If $p(\theta)$ is too complex for direct simulation, the $\theta^{(t)}$s can be generated through a Markov Chain having $p(\theta)$ as its stationary distribution. A Markov Chain is a sequence of random variables $\{\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots\}$ such that at each time $t \geq 0$, given $\theta^{(t)}$, the next state $\theta^{(t+1)}$ does not depend on the history of the chain, i.e. if

$$P(\theta^{(t+1)}|\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(t-1)}, \theta^{(t)}) = P(\theta^{(t+1)}|\theta^{(t)}). \tag{11}$$

Although samples will not be necessarily independent, convergence to the required expectation is ensured by the ergodic theorem (Roberts, 1996).

Markov Chains can be constructed in various ways. One well known algorithm is the Gibbs sampler, a special case of the single-component Metropolis-Hastings algorithm (Metropolis et al., 1953, Hastings, 1970). The Gibbs sampler requires the full conditional density of a parameter given all other parameters in the model:

$$p(\theta_j|\theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_m). \tag{12}$$

This can be deduced from the joint posterior distribution

$$p(\theta_j|\theta_{-j}) = \frac{p(\theta)}{\int p(\theta_j, \theta_{-j})d\theta_j}, \tag{13}$$

where θ_{-j} denotes all parameters except θ_j.

In our model the full conditional density for the log-relative risk in region i can be written as

$$p(r_i|r_{-i}, O, \sigma^2) \propto p(O_i|r_i)p(r_i|r_{-i}, \sigma^2) \tag{14}$$

(Mollié, 1996), and for the precision $\tau = 1/\sigma^2$ we get

$$p(\tau|r, O) \propto p(r|\tau)p(\tau). \tag{15}$$

The Gibbs sampler has the great advantage that good software[1] is available. BUGS can be obtained via Internet[2]. BUGS carries out Bayesian inference on problems where the model consists of a defined joint distribution. It is intended for complex models for which conditional independence assumptions can be used, but for which no exact analytic solution exists. BUGS does not only carry out the numerical integrations using simulations, it also provides tools for monitoring and summarising the simulated values. BUGS is distributed as compiled code and is available for various computer platforms (SPARC, DOS, HP, DEC ALPHA, LINUX, ...). There also exists a version for WINDOWS. WinBUGS has a graphical interface to construct the model, a standard 'point-and-click' windows interface for controlling the analysis as well as graphical tools for monitoring.

When simulating the posterior distribution of interest one has to ensure that convergence has been reached before computing an approximation. Convergence to the stationary distribution can be checked by various diagnostics implemented in CODA[3], which runs under S-Plus[4] or R (Ihaka and Gentleman, 1996).

4 Results

Using this Bayesian model with Gibbs sampling for a map with a high number of regions and few cases can be difficult. In contrast to other models, the prior distribution for the precision $\tau = \frac{1}{\sigma^2}$ in our model has more influence on the convergence. Use of a $\Gamma(0.1, 0.1)$-prior for τ does not influence the resulting posterior distribution for examples with a lower number of regions and more observations, but convergence fails for our data. As an additional problem we have an inhomogeneous neighbourhood structure with varying numbers of neighbours. We tested different parameters for the prior distribution of τ and found out, that a proper gamma distribution $\Gamma(1, 1)$ yielded the best results with respect to convergence.

Convergence was then checked in the following way: we computed a pilot chain with 5000 updates and applied Raftery and Lewis (1992)´s convergence diagnostic to get out how many iterations were necessary to reach a required accuracy of estimated quantiles. We re-ran the chain with this number which was about 10000 updates and then checked convergence with Heidelberger and Welch (1983)´s, Gelman and Rubin (1992)´s and Geweke (1992)´s diagnostics which are implemented in CODA. After several trials

[1] BUGS: Bayesian Inference Using Gibbs Sampling. © MRC Biostatistics Unit 1997

[2] http://www.mrc-bsu.cam.ac.uk/bugs/Welcome.html

[3] CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output. © MRC Biostatistics Unit 1995

[4] © Mathsoft Inc., 1988-97

where convergence was rejected we found out, that the chains for every parameter has to be run for 55000 iterations with a burn-in of 5000 updates and a thinning interval of ten, that means, to report only every tenth update. Updating and monitoring the simulated values took about 90 min. on a Pentium 90 with 96 MB RAM. All of the implemented diagnostics indicated that convergence could be reached for the resulting 5000 iterations per municipality.

Using `CODA`, we estimated the posterior mean

$$\widehat{SIR}_i = \frac{\mu_i}{E_i} \tag{16}$$

of the i-th *SIR* as an average over the repeated samples generated by the Gibbs sampler, and used those estimates to produce a smoother map (see Figure 2) of lung-cancer incidences.

We also computed other statistical summaries such as the standard deviation, the posterior median or credible intervals (C.I.). A 95%-C.I. is delimited by the 2.5%- and the 97.5%-quantile, derived from the sampled values. The highest smoothed *SIR* is 1.720 and

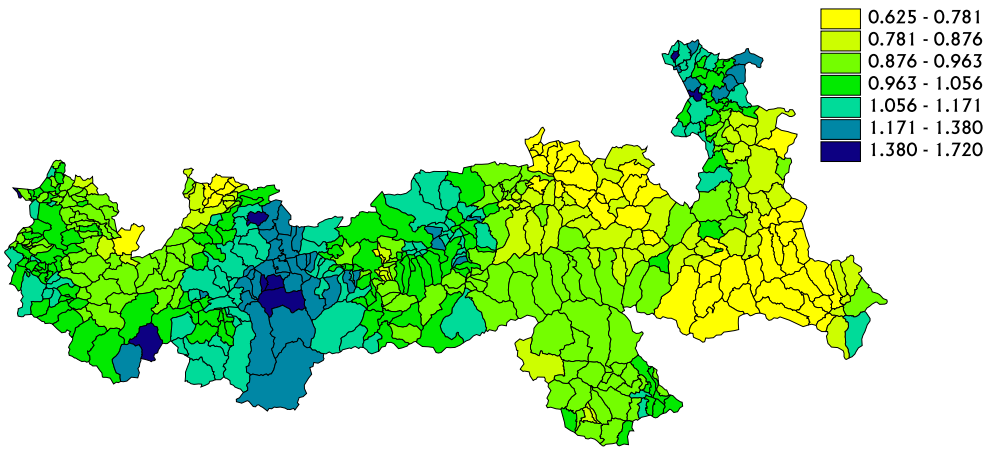

Figure 2: Smoothed relative risk ratios (mean) for lung-cancer for men in Western Austria, 1988-1992

can be found in Umhausen, Tirol. The posterior median in this municipality ($\tilde{x} = 1.677$) was also found to be higher than in all the other regions. Looking at the 95%-credible interval, we found out, that for this municipality the lower bound of the credible interval was 1.177, so we can be sure that this region has higher risk to obtain lung cancer than others. This result corresponds with the fact that Umhausen has a higher radon concentration, which may be a reason for increased lung-cancer risk (Wirth, 1994).

As a conclusion we can say that the Bayesian model is a valuable tool for smoothing cancer ratios, but of course both convergence diagnostics and interpretation have to be done carefully.

References

J. Besag, J. York, and A. Mollié. Bayesian Image Restoration, with two Applications in Spatial Statistics (with discussion). *Ann. Inst. Statist. Math.*, 43(1):1–59, 1991.

A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. Clarendon Press, Oxford, 1997.

D. Clayton and J. Kaldor. Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping. *Biometrics*, 43:671–681, 1987.

L. D. Fisher and G. Van Belle. *Biostatistics, A Methodology for the Health Sciences*. Wiley, New York, 1993.

A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 1992.

J. Geweke. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 4. Oxford University Press, Oxford, 1992.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. Introducing Markov Chain Monte Carlo. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 1. Chapman and Hall, London, 1996.

W. K. Hastings. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57:97–109, 1970.

P. Heidelberger and P. Welch. Simulation run lengths control in the presence of an initial transient. *Operations Research*, 31:1109–33, 1983.

R. Ihaka and R. Gentleman. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

R. Koboltschnig. *Anwendung Bayesscher Modelle in der Räumlichen Epidemiologie am Beispiel von Lungen-Karzinomdaten in Westösterreich*. Doctoral Thesis, Universität Klagenfurt, 1998.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21: 1087–1091, 1953.

A. Mollié. Bayesian mapping of disease. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 20. Chapman and Hall, London, 1996.

W. Oberaigner, H. Cronich, and H. Hausmaninger. *Krebsatlas Westösterreich, 1988-1992, Salzburg, Tirol, Vorarlberg*. Verein Arbeitsgemeinschaft regionaler Tumorregister Österreichs, Innsbruck, 1998.

B. Pesch, U. Halekoh, M. Richter, and F. Pott. *Krebsatlas Nordrhein-Westfalen*. Ministerium für Arbeit, Gesundheit und Soziales des Landes Nordrhein-Westfalen, Düsseldorf, 1994.

A. E. Raftery and S. Lewis. How many Iterations in the Gibbs Sampler? In J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 4, pages 763–773. Oxford University Press, Oxford, 1992.

G. O. Roberts. Markov Chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 3. Chapman and Hall, London, 1996.

V. Wirth. Serie Umhausen, 1. Teil: Radon und Krebs. Konzept einer aktiven Tumorprophylaxe. *Curriculum Oncologicum*, 2, 1994.

W. Zatonski, M. Smans, J. Tyczynski, and P. Boyle. *Atlas of Cancer Mortality in Central Europe*. International Agency for Research on Cancer, Lyon, 1996.

Author's address:

Dr. Rose-Gerd Koboltschnig
Institut für Mathematik, Statistik und Didaktik der Mathematik
Universität Klagenfurt
Universitätsstr. 65-67
A-9020 Klagenfurt

E-Mail: rose.koboltschnig@uni-klu.ac.at