ELSEVIER Computational Statistics & Data Analysis 41 (2003) 617–629

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

A transitional non-parametric maximum pseudo-likelihood estimator for disease mapping

A. Biggeri[a,*], E. Dreassi[a], C. Lagazio[b], D. Böhning[c]

[a] *Department of Statistics "G. Parenti", University of Florence, Viale Morgagni 59, 1-50134, Florence, Italy*

[b] *Department of Statistical Science, University of Udine, Italy*

[c] *Department of Epidemiology, Free University, Berlin, Germany*

Received 1 March 2002

Abstract

Non-parametric maximum likelihood estimators of relative risk have been proposed as an alternative to empirical Bayes or full Bayes approaches to disease mapping. They have the advantage of being relatively simple, the EM algorithm assures convergence and area classification is straightforward. However, they do not take into account spatial autocorrelation and have higher mean square error when the true underlying risk pattern is strongly spatially structured. Furthermore, the EM algorithm is sensible to starting values and could converge to local maxima. We review the transitional generalized linear models and propose a transitional non-parametric maximum pseudo-likelihood estimator for disease mapping. The usual kernel likelihood of the mixture models is replaced by the conditional density of the observed response for a single area given the values observed in adjacent areas. The estimation of the parameters is based on the EM algorithm, appropriately modified to handle the problem of local maxima and to estimate the number of components of the mixture. A simulation study shows that the transitional non-parametric maximum pseudo-likelihood estimator performs similarly to full Bayes estimators.

Keywords: Disease mapping; Non-parametric maximum likelihood; Transitional models; Autoregressive models; Empirical Bayes

1. Introduction

Descriptive epidemiology focuses on the variation of disease occurrence among populations, and has been applied in etiologic research, to screen for environmental

* Corresponding author. Tel.: +39-0554237252; fax: +39-0554223560.
E-mail address: abiggeri@ds.unifi.it (A. Biggeri).

0167-9473/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-9473(02)00189-5

exposures, and in health service research, to compare institutional performances or to allocate resources.

Smoothed estimators of relative risk have been proposed in the context of descriptive epidemiology studies in the mid-1970s by Efron and Morris (1975), specifically to map prevalence rates of an infectious disease in El Salvador. In the 1980s, they have been used to study hospital variation in death rates due to surgical procedures. The literature in the last decades reported several examples of disease mapping at small scale of geographical resolution. Most recent applications are based on a full Bayesian approach. However, finite mixtures models appear to be very attractive for practical applications.

Data (disease counts and population denominators) typically present substantial overdispersion and James–Stein type estimators are used for that reason. The distribution of rates or rate ratios depends not only on the true risk variation but also on population size variation. Areas with small population size have small expected counts (even less than one) and the estimated rate ratio would assume only discrete values. Reporting rates or rate ratios produces a map which is dominated by the less populated areas which rank at the extreme of the used scale.

A better solution would be to map a weighted average between the maximum likelihood estimator and a general mean:

$$\mathrm{EBMR} = w\,\mathrm{SMR} + (1-w)\mu,$$

where the weights w should be directly proportional to the precision of the SMR estimate and to the extent of the variability of the true rate ratios. We can assume a given distribution of the true rate ratios among areas, and focus the analysis on the estimation of its mean and standard deviation. Several models of this kind are reported in the literature: Gaussian–Gaussian empirical Bayes procedure (Efron and Morris, 1975); parametric Poisson-Gamma (Manton et al., 1981); empirical Bayes approach based on moments (McPherson et al., 1982); empirical Bayes and full Bayes approach using non-conjugate priors (Tsutakawa, 1985); Poisson-Gamma and Poisson-logNormal, conditionally autoregressive (CAR) priors and non-parametric maximum likelihood (Clayton and Kaldor, 1987); full Bayes approach with structured and unstructured spatial random effects (BYM: Besag et al., 1991); mixture models (Schlattmann and Böhning, 1993).

Few papers have addressed the evaluation of methods by simulation studies. Marshall (1991) showed that CAR-type estimators are more biased than simpler overdispersion-adjusted estimators. Lawson et al. (2000) concluded that the BYM model outperforms other approaches; Militino et al. (2001) addressed specifically mixture models, showing that they are valid but can give biased results with strongly spatially structured risk patterns.

In the present paper we will focus on finite mixture models. A preliminary paper gave some suggestions and real examples (Biggeri et al., 2000). Here we formalize the proposed approaches and present a simulation study having the BYM full Bayesian approach as benchmark.

Finite mixture models have the advantage of being less computer intensive, more friendly for applied researchers, since convergence is assured, and not dependent on the specification of the mixing distribution for the random effects (Aitkin, 1999). Inaccuracies in the estimation of the latent distribution have been claimed by Carlin and Louis (1996), but this is a minor point when the emphasis is on the posterior estimates, as in disease mapping applications.

We propose a transitional non-parametric maximum pseudo-likelihood estimator for disease mapping which takes into account the spatially structured random effects. Section 2 reviews transitional regression models and their extension to disease mapping, Section 3 introduces transitional finite mixture models, Section 4 is dedicated to computational aspects, Section 5 reports the results of a simulation study, discussion and conclusions are presented in Section 6.

2. Transitional models

Autoregressive models have been used in the context of time series and spatial analysis. The basic idea is that a given observation y_i of a response variable Y_i (the current observation in time, the observation in a given space location) is a function of potential covariates $\mathbf{X}_i$ and other responses $H_i=\{Y_j\}_{j\in S_i}=\{Y_{j\sim i}\}$ (where S_i denotes the set of observations considered to be "close" in space or time to the ith one). Usually, attention is restricted to one class of automodels, Markov random chains or fields, for which $Y_i \mid H_i$ are independent. The model order refers to the space or time proximities considered in H_i (for time series, for example, $f(H_i) = \sum_{j=1}^{q} \alpha_j Y_{i-j}$ represents an autoregressive model of order q). In the following we will refer to models of order 1, with a single interaction parameter α. This is justified since we are interested in spatial processes which could be represented as Markov random fields.

A transitional generalized linear model (TGLM) has conditional density

$$f(y_i \mid H_i) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}$$

with known functions $a(\cdot)$ and $b(\cdot)$. The conditional mean and variance are

$$E(Y_i \mid H_i) = \mu_i = \frac{\partial b}{\partial \theta_i}, \quad \mathrm{Var}(Y_i \mid H_i) = \frac{\partial^2 b}{\partial \theta_i^2}\, a(\phi).$$

The linear predictor is related to the mean through the link function $g(\cdot)$:

$$g(\mu_i) = \mathbf{x}_i'\beta + \sum_j f_j(H_i, \alpha),$$

where $f_j(\cdot)$ is an appropriate function for the autoregressive terms. These terms are treated as additional explanatory variables. In case of *linear models* we have the important result of exact correspondence between a transitional model and a model with autoregressive error terms. Actually the transitional linear model is

$$Y_i = \mathbf{x}_i'\beta + \alpha \sum_j (Y_{j\sim i} - \mathbf{x}_j'\beta) + \varepsilon_i$$

with $\varepsilon \sim \mathrm{Gaussian}(0, a(\phi))$, $g(\mu_i)=\mu_i$, $\partial^2 b/\partial\theta_i^2=1$, and it is equal to the autoregressive linear model:

$$Y_i = \mathbf{x}_i'\beta + \varepsilon_i^{\star}$$

with $\varepsilon_i^\star = \alpha \sum_j \varepsilon_{j\sim i}^\star + \varepsilon_i$. Unfortunately, this is no more true with non-linear link functions $g(\cdot)$.

In general a class of transitional regression models can be defined with two distinct cases (Brumback et al., 2000): Generalized Linear Models with AutoRegressive error terms (GLM-AR)

$$\mu_i = g^{-1}(\mathbf{x}_i'\beta) + \sum_j f_j(H_i, \alpha)$$

and transitional generalized linear models

$$\mu_i = g^{-1}\left[\mathbf{x}_i'\beta + \sum_j f_j(H_i, \alpha)\right].$$

In the case of Poisson random variables the GLM-AR does not guarantee non-negativity of expected values. Therefore, we restrict our attention only to transitional GLMs.

Several specifications of the function $f_j(H_i, \alpha)$ have been proposed. Besag (1974) introduced the AutoPoisson model on a two dimensional lattice

$$\mu_i = \exp\left(\mathbf{x}_i'\beta + \alpha \sum_j y_{j\sim i}\right)$$

which has not been widely used since for positive autocorrelations the conditional expectation grows in space or time and stationary processes imply only negative autocorrelations. Simple modifications of this model use a difference (Brumback et al., 2000) or a ratio (Biggeri et al., 2000) between the responses and the inverse link function of the systematic component $\exp\{\mathbf{x}_i'\beta\}$.

Zeger and Qaqish (1988) considered longitudinal data and proposed a transitional log-linear model using the residual at time $i-1$ in the scale of the link function:

$$\mu_i = \exp\{\mathbf{x}_i'\beta + \alpha[\ln(y_{i-1}) - \mathbf{x}_{i-1}'\beta]\}.$$

2.1. Transitional models in disease mapping

We propose to model the conditional expectation of disease counts observed in a given region as function of population denominators and observed responses in the neighboring areas.

The general family of models of this form is the Gibbs distribution defined on a finite lattice of locations. The area centroids are the nodes of the lattice and the joint probability distribution of the observed counts is specified from the conditional probability $f(O_i \mid \{O_j\}_{j\in S_i}) = f(O_i \mid \{O_{j\sim i}\})$. The model is

$$O_i \mid \{O_{j\sim i}\} \sim \text{Poisson}(E_i\lambda_i)$$

with

$$\lambda_i = \exp\left[\beta_0 + \beta_1 \ln E_i + \alpha \ln\left(\frac{\sum_j O_{j\sim i}}{\sum_j E_{j\sim i}}\right)\right]. \quad (1)$$

The population denominator E_i is the number of person years at risk or the expected count under indirect standardization. It is treated as an *offset* fixing $\beta_1 = 1$. The adjacent areas are defined on the basis of some suitable distance function: e.g. two areas are defined adjacent if they share the boundary (for other specifications see Cressie, 1993). The adjacency matrix obtained is therefore symmetric having entries either 0 or 1.

Fitting transition models requires some considerations. First, the transitional GLM is specified using conditional distributions. In this case, the estimates maximize the quantity

$$PL = \prod_{i} f(O_i \mid \{O_{j\sim i}\})$$

which represents the well-known pseudo-likelihood (PL) approximation to the full likelihood (Besag, 1975).

Second, the function for the autoregressive terms contains both (β_0, β_1) and α parameters. In our case we model standardized mortality ratios with expected counts given by internal standardization. Then $\beta_0 = 0$ and $\beta_1 = 1$, as previously said. We do not need any specific iterative process between the estimation of the systematic component and the autoregressive coefficients (Brumback et al., 2000). In fact the autoregressive component can be calculated *before* fitting the model to the data, since it depends only on the observed disease counts and the known population denominators. The model then reduces to a standard GLM with an offset and an additional covariate which can be fitted by IRLS.

3. Transitional finite mixture models

In disease mapping we face the problem of overdispersion as a consequence of the wide variability in population denominators among small areas. The distribution of the observed disease counts is viewed as a marginal distribution and the true distribution of relative risks is considered latent. This marginal distribution is the integral of the likelihood function over the latent distribution, where $i = 1, \ldots, n$ denotes the areas in the region of interest:

$$f(O_i) = \prod_{i=1}^{n} \int f(O_i \mid \lambda_i) f(\lambda_i)\, \mathrm{d}\lambda_i.$$

Since the likelihood for the observed disease counts is Poisson, a very popular approach assumes that λ_i, $i = 1, \ldots, n$, are a sample from a Gamma prior. The use of the conjugate distribution leads to the negative binomial marginal density.

There is no subject specific justification for the choice of the Gamma distribution. To avoid the arbitrariness in selecting the prior density we can approximate it by assuming a discrete prior distribution with probability π_k at mass points λ_k, $k = 1, \ldots, K$; the integrated likelihood contribution of the ith area becomes a sum and the full likelihood takes the form

$$L = \prod_{i=1}^{n} \sum_{k=1}^{K} f(O_i \mid \lambda_k) \pi_k.$$

The rational is that the data contain information not only about the parameters of the prior density but also on its form (Laird, 1978; Aitkin, 1999). The first application to disease mapping can be found in Clayton and Kaldor (1987). Other simple examples are provided by Aitkin (1996b, 1999).

Böhning et al. (1992), Schlattmann and Böhning (1993) gave a complete description of this approach. Schlattmann et al. (1996) extended the model to include covariates.

In the specification above, the kernel likelihood is given by a generalized linear model. We instead propose to define a pseudo-likelihood kernel based on the transitional generalized linear model (1). The marginal likelihood is then approximated by a finite mixture of transitional Poisson pseudo-likelihoods with autocorrelation parameter which varies among the different K components. The model is then defined in the following way:

$$PL = \prod_{i=1}^{n} \sum_{k=1}^{K} f[O_i \mid \lambda_i(\beta_k, \alpha_k)]\pi_k$$

and substituting the Poisson kernel

$$PL = \prod_{i=1}^{n} \sum_{k=1}^{K} [(\lambda_i(\beta_k, \alpha_k))^{O_i} \mathrm{e}^{-E_i \lambda_i(\beta_k, \alpha_k)}]\pi_k \qquad (2)$$

with $\lambda_i(\beta_k, \alpha_k)$ given by the formula

$$\lambda_i(\beta_k, \alpha_k) = \exp\left[\beta_k + \ln E_i + \alpha_k \ln\left(\frac{\sum_j O_{j\sim i}}{\sum_j E_{j\sim i}}\right)\right].$$

This approach is close to the NPML for random effects models of Aitkin (1999) where the support points lie in the plane defined by the random intercept and the random slope.

The main difference with the simpler NPML case is that we have defined only a pseudo-likelihood approximation to the marginal likelihood. On the other side, the model includes also a random slope for the autoregressive term that describes spatial correlation of observed values. Moreover, it is assumed that the force of interaction between adjacent areas is not necessarily independent from the area absolute level of risk or rate ratio. Suppose we have only two components. In one case it could be that low risk areas group each other showing high autocorrelation; while high risk areas could appear isolated from each other being close to low risk areas, exhibiting zero or negative autocorrelation. In a second case, the areas could cluster, high risk areas close each other and low risk areas close each other, resulting in equal autocorrelation parameter in the two components.

The transitional non-parametric maximum pseudo-likelihood (TNPMPL) estimates of β_k, α_k and π_k can be obtained using the EM algorithm for fixed number of components. The smoothed relative risks $\{\lambda_i\}$ are obtained from the fitted model:

$$\tilde{\lambda}_i = \sum_k \hat{w}_{ik} \lambda_i(\hat{\beta}_k, \hat{\alpha}_k),$$

where $\hat{w}_{ik}$ is an estimate of the latent assignment variable w_{ik} ($w_{ik} = 1$ if the ith observation belongs to component k and $w_{ik} = 0$ otherwise), that represents the weight of the kth component for the ith observation (details are given below).

4. Computational aspects

To obtain estimates of support points, probability masses and observation weights we used the following algorithm (we will refer only to the transitional mixture model, the same can be done also for the simpler NPML approach):

(1) Choose initial values of the number K of components ($k=1,\dots,K$), of the support points (β_k,α_k) and of the probability masses π_k of the mixing distribution according to Gaussian quadrature (Aitkin, 1996a)

$$Q_K=\begin{pmatrix}(\beta_1,\alpha_1) & \dots & (\beta_K,\alpha_K)\\ \pi_1 & \dots & \pi_K\end{pmatrix}.$$

(2) Use the EM algorithm to estimate Q_K. The complete log-pseudo-likelihood is

$$\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\{\ln\pi_k+\ln f[O_i\,|\,\lambda_i(\beta_k,\alpha_k)]\}, \qquad (3)$$

where w_{ik} are the weights previously defined. The E-step gives the unknown weights estimates

$$\hat{w}_{ik}=\hat{\pi}_k f[O_i\,|\,\lambda_i(\hat{\beta}_k,\hat{\alpha}_k)]\Big/\sum_{l}\hat{\pi}_l f[O_i\,|\,\lambda_i(\hat{\beta}_l,\hat{\alpha}_l)]$$

and the M-step maximizes (3) (with w_{ik} replaced by $\hat{w}_{ik}$) giving new estimates $(\hat{\beta}_k,\hat{\alpha}_k)$ of (β_k,α_k) and

$$\left\{\hat{\pi}_k=\sum_{i}^{n}\hat{w}_{ik}/n\right\}.$$

(3) Add one component, choose initial values as in step 1 and iterate with EM until convergence. Compare the log-pseudo-likelihood with that previously obtained; if no improvement is recorded then stop, take $K-1$ component and go to step 4. Otherwise repeat step 3.

(4) Evaluate the gradient function, i.e. the path derivative of the mixture log-pseudo-likelihood $\ln(PL(Q))=\sum_i \ln\sum_k \pi_k f_{ik}(O_i\,|\,\lambda_i(\beta_k,\alpha_k))$ in the direction of a single component (Lindsay, 1995, Chapter 2), which is in our case

$$D(\lambda,Q)=\frac{1}{n}\sum_{i}\frac{\mathrm{e}^{-\lambda}\lambda^{O_i}}{\sum_k \pi_k \mathrm{e}^{-\lambda_i(\beta_k,\alpha_k)}\lambda_i(\beta_k,\alpha_k)^{O_i}}$$

on a fine grid over $[\lambda\in 0.5$–$3.0]$. If the maximum of $D(\lambda,Q)\leqslant 1+\varepsilon$ then stop (ε=0.001), otherwise substitute the value corresponding to the maximum of $D(\lambda,q)$ to the support point with lowest probability mass and perform step 2 (repeat utmost ten times). If the gradient function criterion is not yet satisfied add one more component and repeat step 4.

The EM algorithm is easy to use and it gives the same estimates as by direct maximization of the score functions (see Böhning, 2000, pp. 59–66). It has however two drawbacks to be considered. First, the EM is based on a fixed number K of components, while in our approach K is one of the parameters to be estimated. The proposed algorithm follows a forward strategy, starting with a few components (e.g. $K = 2$) and adding one more component at time. The stopping rule is based on the evaluation of the gradient function, the directional derivative at Q of the mixture log-pseudo-likelihood $\ln(L(Q))$ on the path to Q'. Indeed, the general mixture maximum likelihood theorem (Lindsay, 1983a, b) gives us an appropriate stopping rule (see step 4 of the algorithm). The theorem states that $\hat{Q}$ is NPMLE if and only if, for all λ,

$$D(\lambda, Q) = \frac{1}{n} \sum_i \frac{e^{-\lambda} \lambda^{O_i}}{\sum_k \pi_k e^{-\lambda_k} \lambda_k^{O_i}} \leqslant 1 \ \forall \lambda.$$

Following Lindsay (1995, pp. 132–135) we set a tolerance of 0.001 and half-grid-width of $0.005\sigma_\lambda$ for n=341. The standard deviations for the four true situations chosen in the simulation study were no greater than 0.5. We decided to use a fixed half-grid-width of 0.001.

The second point is that the EM algorithm does not guarantee convergence to a global maximum. This is of particular concern since the finite component likelihood has significant multimodality (Lindsay, 1995, p. 65). Again, the general mixture maximum likelihood theorem assures the convergence of our algorithm to a global maximum. Moreover, the gradient function has been also used to identify good starting values for the EM algorithm and therefore avoid local maxima. To achieve this goal, we used the value of λ with maximum gradient function as a new initial value when adding one more component in step 4 or, alternatively, we exchanged "bad" with "good" estimated support points (again in step 4). These ideas are adapted from the Vertex Exchange Algorithm (Böhning, 2000, p. 50).

5. Simulation study

To evaluate the proposed estimators we conducted a simulation study comparing the non-parametric maximum likelihood (NPML) estimator (Schlattmann and Böhning, 1993), the Transitional NPMPL presented in this paper, the simpler empirical Bayes Poisson-Gamma (Clayton and Kaldor, 1987), the full Bayesian estimator of Besag et al. (1991) and SMR (maximum likelihood estimator).

We used four different true risk maps (each map with $n = 341$ areas) taken from appropriate real examples (Biggeri et al., 2000):

(1) HET: high and low risk areas not spatially structured, intermediate-low average number of events (heterogeneity);

(2) HET-CLUS: a mixed pattern with high and low risk areas not spatially and spatially structured, intermediate-low average number of events (heterogeneity and clustering);

(3) CLUS: high and low risk areas strongly spatially structured, high-intermediate average number of events (clustering);

(4) MIXT: high and low risk areas from a four components mixture model, intermediate-low average number of events (mixture).

The four patterns range from a pure heterogeneous Poisson map with sparse data to a highly clustered one (Fig. 1).

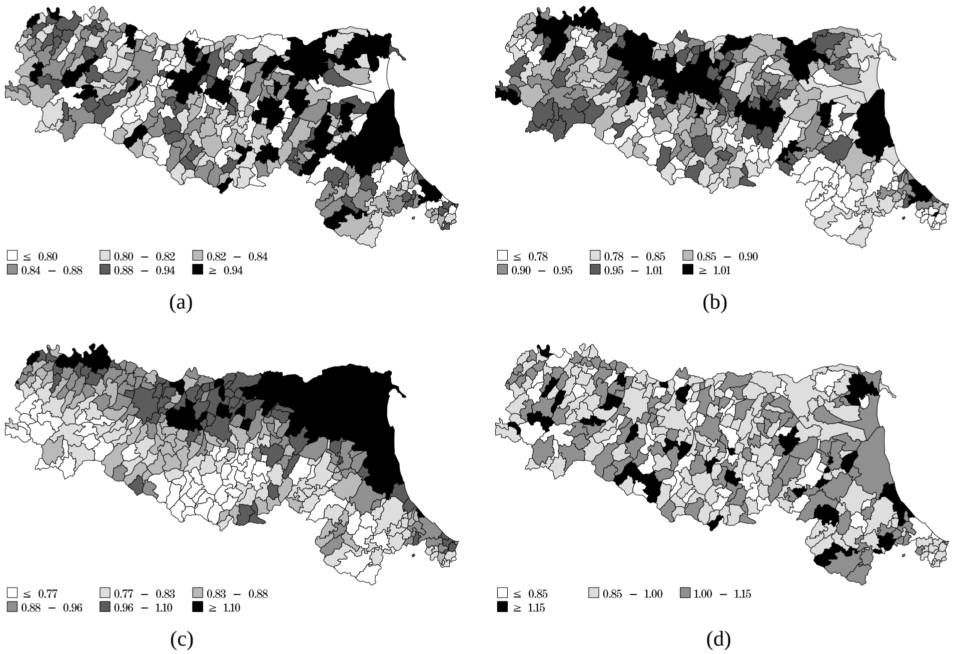

Fig. 1. Spatial distribution of relative risk for (a) HET, (b) HET-CLUS, (c) CLUS and (d) MIXT.

Table 1
Descriptive statistics of the four true patterns of relative risks (RR) used in the simulation

	mean RR	s.d.	min	median	max
HET	0.87	0.105	0.66	0.84	1.61
HET-CLUS	0.89	0.142	0.37	0.90	1.44
CLUS	0.94	0.229	0.67	0.88	2.027
MIXT	0.98	0.133	0.80	0.90	1.20

One hundred data sets were generated from each pattern using a Poisson law and population denominators as described (Table 1).

The estimates obtained using the four methods were compared using the average bias, the average variance, the average mean square error (Table 2) and the correlation coefficients between average ranks based on the estimates given by each method and the true ranks (Tables 3–6).

The average bias is $\sum_i(\bar{\bar{\lambda}}_{Mi} - \lambda_i)/n$, where λ_i is the true value of risk in the ith area and $\bar{\bar{\lambda}}_{Mi}$ is the mean value of the estimates obtained with method M in the 100 simulated data sets. It was lower for the SMR, as expected, and both the non-parametric estimators.

The average variance is $\sum_i [\sum_j(\hat{\lambda}_{Mij} - \bar{\bar{\lambda}}_{Mi})^2/(J-1)]/n$, where j is a simulated dataset ($j = 1,\ldots,J$) and $\hat{\lambda}_{Mij}$ is the estimate of risk using method M, in the ith area

Table 2
Mean bias, mean variance and mean square error (MSE), for the considered estimators: SMR, Poisson-Gamma (PG), non-parametric maximum likelihood (NPML), transitional NPML (TNPML) and Bayesian (BYM)

				Estimators		
	Risk pattern	SMR	PG	NPML	TNPMPL	BYM
Bias	HET	0.04480	0.07090	0.03171	0.03084	0.04619
	HET-CLUS	0.00994	0.05087	0.03213	0.02516	0.04682
	CLUS	−0.00032	0.01989	0.01025	−0.00200	0.01306
	MIXT	0.00469	−0.00830	−0.01732	−0.01665	−0.00999
Variance	HET	0.72771	0.00262	0.00530	0.01385	0.00790
	HET-CLUS	0.29183	0.00193	0.00459	0.00988	0.00443
	CLUS	0.11833	0.01004	0.00872	0.01154	0.00694
	MIXT	0.34208	0.00301	0.00309	0.00789	0.00362
MSE	HET	0.79307	0.01533	0.01231	0.02072	0.01536
	HET-CLUS	0.29392	0.02156	0.02083	0.02079	0.01942
	CLUS	0.11826	0.03350	0.03301	0.01915	0.01250
	MIXT	0.34281	0.01723	0.01802	0.02238	0.01747

Table 3
Correlation matrix of mean ranks estimated with the considered estimators. Heterogeneous risk pattern (HET)

	TRUE	SMR	PG	NPML	TNPMPL	BYM
TRUE	1.000	0.566	0.571	0.503	0.534	0.542
SMR	0.566	1.000	0.718	0.564	0.642	0.661
PG	0.571	0.718	1.000	0.954	0.953	0.793
NPML	0.503	0.564	0.954	1.000	0.946	0.723
TNPMPL	0.534	0.642	0.953	0.946	1.000	0.728
BYM	0.542	0.662	0.793	0.723	0.728	1.000

for the jth dataset. The maximum value was always observed for SMR and, generally speaking, the average variance was greater for estimators based on autoregressive models (TNPMPL and BYM). These findings reflect the trade-off between bias and precision.

The average mean square error $\sum_i [\sum_j (\hat{\lambda}_{ij} - \lambda_i)^2/J]/n$ was lower for the Bayesian estimator.

For the heterogeneity risk pattern the NPML behaves better than the other estimators, for clustered risk patterns the TNPMPL was very similar to the BYM model, while for the four component mixture the parametric Poisson-Gamma or BYM model appeared slightly better than NPML. This could be attributed to the difficulties in estimating the number of components in the mixture.

The correlation coefficients between the average estimated ranks and true ranks were consistent with the average bias: the SMR is generally better. The transitional NPMPL

Table 4
Correlation matrix of mean ranks estimated with the considered estimators. Heterogeneous and clustered risk pattern (HET-CLUS)

	TRUE	SMR	PG	NPML	TNPMPL	BYM
TRUE	1.000	0.851	0.714	0.731	0.814	0.710
SMR	0.851	1.000	0.828	0.853	0.843	0.718
PG	0.714	0.828	1.000	0.997	0.783	0.744
NPML	0.731	0.853	0.997	1.000	0.802	0.746
TNPMPL	0.814	0.843	0.783	0.802	1.000	0.797
BYM	0.710	0.718	0.744	0.746	0.796	1.000

Table 5
Correlation matrix of mean ranks estimated with the considered estimators. Clustered risk pattern (CLUS)

	TRUE	SMR	PG	NPML	TNPMPL	BYM
TRUE	1.000	0.913	0.854	0.833	0.912	0.930
SMR	0.913	1.000	0.956	0.937	0.913	0.921
PG	0.854	0.956	1.000	0.996	0.852	0.872
NPML	0.833	0.937	0.996	1.000	0.829	0.852
TNPMPL	0.912	0.913	0.852	0.829	1.000	0.962
BYM	0.930	0.921	0.872	0.852	0.962	1.000

Table 6
Correlation matrix of mean ranks estimated with the considered estimators. Mixture pattern (MIXT)

	TRUE	SMR	PG	NPML	TNPMPL	BYM
TRUE	1.000	0.887	0.773	0.769	0.744	0.714
SMR	0.887	1.000	0.828	0.808	0.795	0.750
PG	0.773	0.828	1.000	0.995	0.975	0.922
NPML	0.769	0.808	0.995	1.000	0.973	0.917
TNPMPL	0.744	0.795	0.975	0.973	1.000	0.900
BYM	0.714	0.750	0.922	0.918	0.900	1.000

is highly correlated with the Bayesian estimator when the underlying true risk pattern is spatially structured. The Poisson-Gamma estimator has a good performance with heterogeneous risk and mixture patterns.

6. Discussion and conclusions

Clayton and Kaldor (1987) stated: “[The NPML approaches] ignore any spatial correlation and assume [that the relative risks are] iid random variables with density of unknown parametric form. (...) for mapping diseases in very small areas (...) clearly it will be necessary to allow for spatial autocorrelation.” Also Aitkin (1999) remarked:

"A limitation of the NPML in [the disease mapping example] is that it does not allow for spatial dependence between neighboring units." Böhning (2000) noted the difficulties of including an adjacency matrix into mixture models.

All these claims are even more important since Militino et al. (2001) reported that NPML performs worse than other methods when spatial autocorrelation is present. In the present paper we confirmed these findings by a simulation study.

We used a pseudo-likelihood approach to derive a transitional NPMPL estimator. The method proposed includes spatial neighbourhood dependence but assumes that the observed data are independent. The pseudo-likelihood is likely to be valid only under weak correlation.

A simulation study which covered a broad range of realistic models is then conducted to evaluate the performance of the Transitional NPMPL estimator: overall it provides estimates that are close to those obtained by Bayesian autocorrelated models.

However, the user should be warned against the possibility of local maxima and the difficulty in detecting the optimal number of components. Actually we encountered this kind of problem in 9% of the simulated datasets using TNPMPL and NPML for HET, 7% and 9%, respectively, for HET-CLUS, 5% and 17%, respectively, for CLUS and 16% and 22%, respectively, for MIXT. Special software (such as CAMAN, Böhning et al., 1992, Böhning et al., 1998) has been developed and should be recommended mainly to inexperienced users.

Another drawback of the non-parametric methods is that it is very difficult to evaluate the standard errors of estimates, that are not provided by the EM algorithm.

Moreover, for the transitional approach, standard likelihood theory is not applicable, since it uses a pseudo-likelihood approximation.

Extension to ecological regression is not straightforward, since the autoregressive term is pre-computed when it clearly must be estimated when other covariates than an offset term are to be considered (Brumback et al., 2000).

In conclusion we reviewed the proposed non-parametric maximum likelihood estimators for disease mapping and presented a transitional NPMPL approach. The performance of non-parametric estimators was compared with that of the Bayesian hierarchical estimator using a simulation study. Overall, the transitional NPMPL estimates were closer to the Bayesian estimates than exchangeable NPML estimates. This formulation addresses the point raised by Aitkin (1999) and Militino et al. (2001).

Acknowledgements

The present research has been made possible by funds from the Vigoni Programme 1999–2000 (A. Biggeri and D. Böhning) and COFIN 1999 (C. Lagazio).

References

Aitkin, M., 1996a. A general maximum likelihood analysis of overdispersion in generalized linear models. Statist. Comput. 6, 251–262.

Aitkin, M., 1996b. Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In: Forcina, A., Marchetti, G., Hatzinger, R., Galmacci, G. (Eds.), Statistical Modelling, Graphos, Città di Castello, pp. 85–94.

Aitkin, M., 1999. A general maximum likelihood analysis of variance. Biometrics 55, 117–128.

Besag, J., 1974. Spatial interactions and the statistical analysis of lattice systems. J. Roy. Statist. Soc. B 36, 192–236.

Besag, J., 1975. Statistical analysis of non-lattice data. The Statistician 24, 179–195.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration with two applications in spatial statistics. Ann. Inst. Statist. Math. 43, 1–59.

Biggeri, A., Marchi, M., Lagazio, C., Böhning, D., Martuzzi, M., 2000. Non-parametric maximum likelihood estimators for disease mapping. Statist. Med. 19, 2539–2554.

Böhning, D., Schlattmann, P., Lindsay, B.G., 1992. Computer assisted analysis of mixtures (C.A.MAN): statistical algorithms. Biometrics 48, 283–303.

Böhning, D., Dietz, E., Schlattmann, P., 1998. Recent developments in computer assisted analysis of mixtures. Biometrics 54, 525–536.

Böhning, D., 2000. Computer Assisted Analysis of Mixtures. Chapman & Hall/CRC, Boca Raton.

Brumback, B.A., Ryan, L.M., Schwartz, J.D., Neas, L.M., Stark, P.C., Burge, H.A., 2000. Transitional regression models, with application to environmental time series. J Amer. Statist. Assoc. 95, 16–27.

Carlin, B., Louis, T., 1996. Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall/CRC, Boca Raton.

Clayton, D., Kaldor, J., 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43, 671–681.

Cressie, N., 1993. Statistics for Spatial Data, (rev. edn.). Wiley, New York.

Efron, B., Morris, C., 1975. Data Analysis using Stein's estimation and its generalization. J. Amer. Statist. Assoc. 70, 311–319.

Laird, N.M., 1978. Nonparametric maximum likelihood estimation of a mixing distribution. J. Amer. Statist. Assoc. 73, 805–811.

Lawson, A.B., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J-F, Clark, A., Schlattmann, P., Divino, F., 2000. Disease mapping models: an empirical evaluation. Statist. Med. 19, 2217–2242.

Lindsay, B.G., 1983a. The geometry of mixture likelihoods: a general theory. Ann. Statist. 11, 86–94.

Lindsay, B.G., 1983b. Efficiency of the conditional score in a mixture setting. Ann. Statist. 11, 486–497.

Lindsay, B.G., 1995. Mixture models: theory, geometry and applications. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5. Hayward, Institute of Mathematical Statistics.

Manton, K.G., Woodbury, M.A., Stallard, E., 1981. A variance components approach to categorical data models with heterogeneous cell populations: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties. Biometrics 37, 259–269.

Marshall, R.J., 1991. Mapping disease and mortality rates using empirical Bayes estimators. J. Roy. Statist. Soc. C 40, 283–294.

McPherson, K., Wennberg, J.E., Hovind, O.B., Clifford, P., 1982. Small-area variations in the use of common surgical procedures: an international comparison of New England, England, and Norway New England J. Med. 307, 1310–1314.

Militino, A.F., Ugarte, M.D., Dean, C.B., 2001. The use of mixture models for identifying high risks in disease mapping. Statist. Med. 20, 2035–2049.

Schlattmann, P., Böhning, D., 1993. Mixture models and disease mapping. Statist. Med. 12, 1943–1950.

Schlattmann, P., Dietz, E., Böhning, D., 1996. Covariate adjusted mixture models and disease mapping with the program DismapWin. Statist. Med. 15, 919–929.

Tsutakawa, R., 1985. Estimation of cancer mortality rates: a Bayesian analysis of small frequencies. Biometrics 41, 69–79.

Zeger, S., Qaqish, B., 1988. Markov regression models for time series: a quasi-likelihood approach. Biometrics 44, 1019–1031.