Medical Infobahn for Europe
A. Hasman et al. (Eds.)
IOS Press, 2000

Mixture models and modelling heterogeneity of the regional distribution of avoidable death in Germany 1995

P. Schlattman, Dept. of Psychiatry, Free University Berlin, Germany

1. Introduction

The analysis of the spatial variation of disease and its subsequent representation on a map has become an important topic in epidemiological research. One important question in disease mapping is to test the hypothesis that the cases of disease occur at random within the study region presented on the map. In assessing the non-randomness of a map two particular mechanisms are frequently distinguished.

The first mechanism deals with heterogeneity of disease risk, i. e. different levels of disease risk are present in the study region due to geographical variation of unknown or unobserved risk factors. Identification of spatial heterogeneity of disease risk thus may give valuable hints for possible exposure and targets in subsequent analytical studies.
Another phenomenon frequently used to address the non-randomness of a map is spatial autocorrelation, i. e. neighboring regions are said to have similar values of disease risk. Autocorrelation may either reflect a contagious mode of disease transmission or unobserved risk factors common to neighboring regions. Frequently these phenomena are difficult to entangle, heterogeneity may impose as autocorrelation and vice versa.
Usually construction of disease maps starts with choosing geographic sub-units and calculating an appropriate epidemiological measure (Standardized Mortality Ratio, rates, etc.) for each sub-unit.
A measure often used is the Standardized Mortality Ratio (SMR) or Standardized Incidence Ratio (SIR) for incidence data. For each area the SIR_i is defined as

$$SMR_i = \frac{O_i}{E_i}, \text{ with } E_i = \sum_{j=1}^{J} P_{ij}\mu_j \text{ , and J is the number of age groups}$$

where O_i are the observed cases in the i-th regional area, E_i the expected cases in the i-th region based on an external reference and P_{ij} the person years in the i-th area and j-th age stratum. μ_j denotes the age-specific mortality rate of the j-th age stratum in reference, which is assumed to be known.

1.1 Disease Mapping and "avoidable deaths"

Here we look at the example of the carcinoma of the mamma in women in Germany in 1995. Frequently this disease is addressed as belonging to the category of "avoidable death" (Holland, 1993)[1]. This health indicator comprises untimely death cases which might be preventable by medical intervention, preventive measures such as primary prevention or screening or a combination of these actions. The mortality of breast cancer in women depends on access to adequate treatment and the availability and acceptance of screening measures. (Chamberlain, 1996)[2]. An analysis of the regional distribution of the mortality of breast cancer may give valuable hints towards deficits in medical care or regional differences of the acceptance of screening measures. For such an analysis the inclusion of known risk factors for breast cancer such as parity, age at menarche etc. would be most desirable.

In our example we use mortality data from the census offices of the 16 states of Germany. The data are based on the spatial resolution of "Landkreise".

2. Modelling heterogeneity of disease risk

2.1 Introduction

A common approach used in map construction is the choropleth method (Howe, 1990)[3]. This implies categorizing each area and then shading or coloring the individual regions accordingly. Traditional methods of map construction face serious methodological problems: Classification according to the percentiles of the distribution of the epidemiological measure is likely to reflect only chance fluctuations in the corresponding small counts.
Probability maps based on a Poisson assumption face the problem of misclassification as well. It can be shown (Schlattmann[4] and Böhning, 1993) that probability maps do not provide a consistent estimate of heterogeneity of disease risk.
A more flexible approach is given in *random effects models,* i. e. models where the distribution of relative risks θ_i between areas is assumed to have a probability density function $g(\theta)$. The O_i are assumed to be Poisson distributed conditional on θ_i with expectation $\theta_i E_i$ .Several parametric distributions like the Gamma-distribution or the log-normal distribution have been suggested for $g(\theta)$. For details see Clayton[5] and Kaldor (1987) or Mollie[6] and Richardson (1991).

2.2 The Poisson mixture model

In the mixture model setting, we assume that the population under scrutiny consists of subpopulations with different levels of disease risk θ_j, j=1,...k.. Statistically we face the problem to identify the level of risk for each subpopulation and the corresponding proportion of the overall population. This leads to a random effects model where we assume a *discrete* parameter distribution P for $g(\theta)$ with $P = [\theta_1....\theta_k;p_1....p_k]$. P is the discrete probability distribution which gives mass p_j to parameter θ_j.
This model therefore assumes that O_i comes from a nonparametric mixture density of the form:

$$f(o_i, P, E_i) = \sum_{j=1}^{k} p_j f(o_i, \theta_j, E_i) \text{ , with } j=1,...,k \text{ and } p_j \geq 0,\ i = 1 ,..., n \text{ (number of areas)},$$

where f(.) denotes the Poisson-density with $f(o_i, \theta, E_i) = e^{-\theta E_i}(\theta E_i)^{o_i}/o_i!$ Please note that the model consists of the following parameters: the unknown number of components k, the k unknown (relative) risks $\theta_1,...,\theta_k$ and k-1 unknown mixing weights $p_1,.., p_k$. The term E_i denotes the population at risk or the expected cases where SMRs are used.

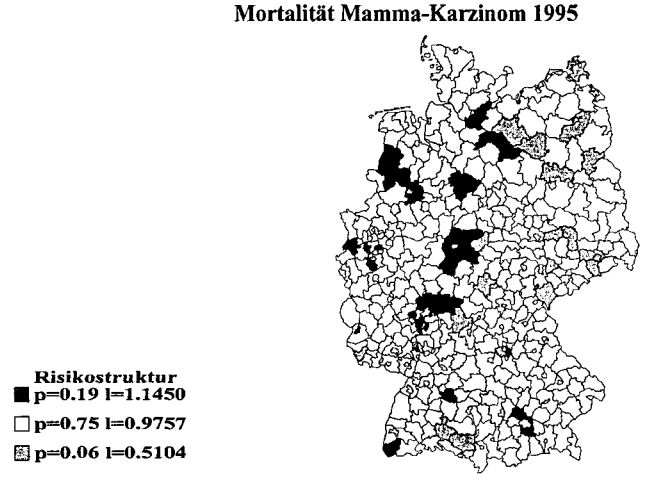
There are no closed form solutions available for finding the maximum likelihood estimates. Suitable algorithms are given by Böhning, Schlattmann and Lindsay (1992)[7]. One principle algorithmic strategy is given in the mixture algorithm, which consists of two steps:
1. In the first step a flexible support size is assumed, i. e. the number of potential subpopulations is assumed to be unknown.
2. The second step involves calculating a solution with a fixed support size, i. e. the number of components is assumed to be known. This second step makes use of the results of the flexible support size solution as starting values for the EM-algorithm.

To estimate the nonparametric maximum likelihood estimator with DismapWin (Schlattmann[8], 1996) we start with the first step of the mixture algorithm. This involves defining a grid containing the parameter values $\theta_1,...,\theta_g$ over which the corresponding

population proportions that maximize the likelihood function have to be found. For details on the algorithmic approach see Böhning[6,9] et al. or Schlattmann[10] et al. (1996).
In order to use this mixture distribution for map construction, its classification aspects are applied. This can be done by defining an unobserved indicator variable z_{ij} for membership in the j-th component. For example $\mathbf{z}_i = (0,1, 0, 0, ..., 0)$ indicates that the i-th region belongs to the second component. Using the estimated mixing distribution as a priori distribution and applying Bayes's theorem leads to

$$\text{pr}(Z_{ij} = 1 | o_i, \hat{P}) = \frac{\hat{p}_j f(o_i, \hat{\theta}_j, E_i)}{\sum_{l=1}^{k} \hat{p}_l f(o_i, \hat{\theta}_l, E_i)}$$

for the probability of belonging to the j-th component. The i-th area is assigned to that subpopulation j for which it has the highest posterior probability of belonging.
Our map of breast cancer mortality shows three levels of disease risk with some marked patterns.



This map shows clearly an east-west and a urban-rural difference. Mainly some rural areas of the eastern states show a decreased risk of 50% whereas urban areas mainly of the western states show an increased risk of dying of about 14 percent.

2.3 The mixture Poisson model with covariates

Once heterogeneity is detected, the question arises on how to address remaining spatial dependency and how to include known covariates $x_1, .,x_M$ into the model.
In the homogenous case covariates are included through Poisson regression (Breslow[11] and Day, 1975).
This leads to a log-linear model, where the Poisson parameter is given by $\theta_i = \exp(LP_i)$, with the linear predictor $LP_i = \alpha+\beta_1 x_{i1} +....+ \beta_M x_{iM}+\log E_i$. With raising $\theta_i = \alpha+\beta 1 x_{i1} +....+ \beta_M x_{iM}$

and log E_i to the power of e (Euler's constant 2. 718) we have a generalization for the Poisson model $O_i \sim Po(\theta E_i)$.

In order to perform an ecological study for the breast cancer data we include the covariates "East-West" and "Urban-Rural"

Poissonregression

Parameter		s. e.	RR = exp (Param.)	95% KI
Intercept	0.063	0.026		
Urban	0.099 (Urban = 1, Rural = 0)	0.015	1.10	1.07, 1.14
East_west	-0.184 (East = 1, West = 0)	0.015	0.83	0.81, 0.86

The homogenous poisson regression model confirms the visual impression we find a significant east-east and urban rural difference. Of course these variables are only surrogate measures indicating differences in lifestyle patterns such as diet, parity etc.

A natural extension of the homogenous Poisson regression model is given by the mixed Poisson regression model (Dietz[12] (1992), Schlattmann[9] et al.(1996)). An extension of the univariate Poisson mixture density $O_i \sim p_1 Po(o_i,\theta_1,E_i) + ... + p_k Po(o_i,\theta_k,E_i)$ is given by a random effects model where the random parameter P is discrete finite with

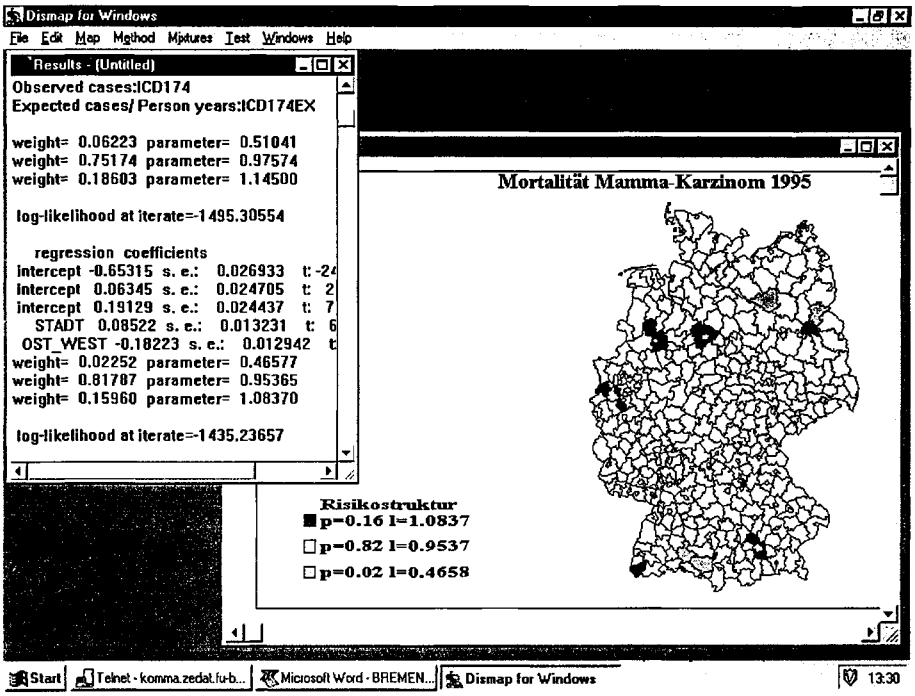
$$P = [\beta_1....\beta_k;p_1....p_k] \text{ with } \beta_j = (\alpha_j, \beta_{1j},...,\beta_{Mj}), j = 1, ..,k ,$$

where M denotes the number of covariates in the Poisson regression model. The conditional distribution of O_i is given by $O_i \sim \sum_{j=1}^{k} p_j f(o_i,\exp(LP_{ij}))$, with linear predictor

$LP_{ij} = \alpha_j+\beta_{1j}x_{1i}+....+ \beta_{Mj}x_{Mi} + \log E_i$, where f(.) again denotes the Poisson density. The number (M+1) of parameters in the Poisson regression is the same for each subpopulation. The univariate mixture model approach may be considered as a special case with mixing only over the intercepts α_j and $\beta_{1j} =...=\beta_{Mj}=0$, j =1, .., k, where k denotes the number of components and M denotes the number of covariates.

Again estimation may be done by maximum-likelihood. If the indicator variables z_{ij} were known, then the maximum-likelihood estimators for the parameters would simply be the MLE's from each component groups. Again, there are no closed form solutions available for maximum likelihood estimates. An adaptation of the EM-algorithm by Dempster[13] et al. (1977) has been developed by Dietz (1992). A detailed description can be found in Schlattmann et al. (1996) as well. The computations involved may be done with the program DismapWin

The next figure shows a screen dump of DismapWin with a covariate adjusted mixture model for the breast cancer data. Clearly after adjusting for the covariates there is still residual heterogeneity present.



3. Discussion

From a practical point of view, especially from the viewpoint of the Public Health practioner the mixture model approach described here has several attractive features. First it provides relatively easy computation and implementation. Second, with packages such as DismapWin, there is free software[14] available, which directly produces maps based on these methods.
Third empirical Bayes methods do not require a difficult convergence diagnostic such as the full Bayesian approach. This relative simplicity is mainly due to the fact, that this approach models unstructured heterogeneity of disease risk and ignores structured heterogeneity. Certainly the full Bayes[15] approach offers the most flexible approach to the data, since any aspect of structured and unstructured heterogeneity and trend can easily be modelled. Also complete inference for any part of the model may be obtained from the posterior distribution. A case study comparing empirical and full Bayesian methods for disease mapping may be found in Schlattmann et al. (1999)[16]. However a rigorous comparison of these various methods of disease mapping is called for.

References

[1] Holland, W. W. (Ed.) 'European Community Atlas of Avoidable Death', Oxford University Press, 1993

[2] Chamberlain, J. S. Moss: Evaluation of Cancer Screening, Springer-Verlag Berlin-Heidelberg-New York, 1996.

[3]Howe, G. M. 'Historical Evolution of Disease Mapping in General and Specifically of Cancer Mapping' in Boyle, P., Muir, C. S. and Grundmann, E. (Eds.) Cancer Mapping. Springer, Berlin, p. 1-21. 1990

[4]Schlattmann, P and Böhning, D. 'Mixture Models and Disease Mapping', *Statistics in Medicine*, **12**, p. 1943-50 (1993)

[5]Clayton, D. and Kaldor, J. 'Empirical Bayes estimates for age-standardized relative risks', *Biometrics*, **43**, p. 671-681 (1987)

[6]Mollie, A. and Richardson, S. 'Empirical Bayes Estimates of Cancer Mortality Rates Using Spatial Models', *Statistics in Medicine*, **10**, p. 95-112 (1991)

[7]Böhning, D., Schlattmann, P. and Lindsay, B. G. 'C.A.MAN- Computer Assisted Analysis of Mixtures: Statistical Algorithms', *Biometrics*, **48**, p. 283-303 (1992)

[8] Schlattmann, P. 'The computer package DismapWin', *Statistics in Medicine*, **15**, p. 931 (1996)

[9] Böhning, D. 'A Review of Reliable Maximum Likelihood Algorithms for Semiparametric Mixture Models', *Journal of Statistical Planning and Inference*, **47**, p. 5-28 (1995)

[10]Schlattmann, P., Dietz, E. and Böhning, D. 'Covariate adjusted mixture models with the program DismapWin', *Statistics in Medicine*, **15**, p. 919-929 (1996)

[11]Breslow, N. E., Day, N. E. 'Indirect standardization and the multiplicative model for rates with reference to the age adjustment of cancer incidence and relative frequency data', *Journal of Chronic Diseases*, **28**, p. 289-303 (1975)

[12] Dietz, E. 'Estimation of Heterogeneity - A GLM-Approach', Fahrmeir, L. , Francis F., Gilchrist, R., Tutz, G. (Eds.), Advances in GLIM and Statistical Modeling, Lecture Notes in Statistics, Springer Verlag Berlin, 1992, p. 66-72

[13]Dempster, A. P., Laird, N. M., Rubin, D. B. 'Maximum likelihood estimation from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B*, **39**, p.1-38 (1977)

[14] DismapWin may be obtained from the URL: www.medizin.fu-berlin.de/sozmed/DismapWin.html.

[15] Besag, J., York, J. and Mollié, A. 'Bayesian image restoration with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics*, **43**, p. 1- 59 (1991)

[16] P. Schlattmann, D. Böhning, A. Clark and A. Lawson (1999): Lung cancer mortality in women in germany 1995 - A case study in disease mapping. In: A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.F. Viel (Eds.) "Disease Mapping and Risk Assessment for Public Health decision making", Wiley, Chichester,400-111