Space-time mixture modelling of public health data

Dankmar Böhning *, Ekkehart Dietz and Peter Schlattmann

Department of Epidemiology, Institute of Social Medicine, Free University of Berlin,
Fabeckstr. 60–62, 14195 Berlin, Germany

SUMMARY
This paper aims to enlarge the usual scope of disease mapping by means of dynamic mixtures (DMDM) in case a time component is involved in the data. A special mixture model is suggested which looks for space-time components (clusters) simultaneously. The idea is illustrated using data on female lung cancer from the East German cancer registry for 1960–1989. The conventional mixed Poisson regression model is used as a third model for comparison. The models are discussed in terms of their benefits, difficulties and ease in interpretation, as well as their statistical meaning. Some ideas on evaluation of these models are also included. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Analysis of the spatial variation of disease and its subsequent representation on a map has become an important topic in epidemiological research. Identification of spatial heterogeneity of disease risk gives valuable indications of possible exposure and targets for analytical studies.

Another important use of disease mapping may be seen in disease surveillance and health outcome research. It has become widely accepted that a potentially fruitful way to monitor the disease status of a community is to look at health data in time and space. The Centers for Disease Control (CDC) defines public health surveillance as the ongoing systematic collection, analysis and interpretation of health data essential to the planning, implementation and evaluation of public health practice. This objective is almost impossible to achieve if temporal and spatial dimensions are ignored. For further general discussion see Teutsch and Churchill [1] and for a case study see Kelly [2]. Maps are used particularly in cancer registries to facilitate reporting of the public health situation and frequently maps are a starting point for cluster investigations.

These uses of disease mapping are frequently constrained to a single time period. Typically, data in public health are available for time windows of 5 or 10 years. Thus a natural extension of disease mapping approaches would be to investigate the dynamics of a certain disease over time. This may give valuable indications of emerging patterns over time. Modelling space-time interaction

*Correspondence to: Dankmar Böhning, Department of Epidemiology, Institute of Social Medicine, Free University of Berlin, Fabeckstr. 60–62, 14195 Berlin, Germany.
Contract/grant sponsor: German Research Foundation
Contract/grant sponsor: BIOMED2

Copyright © 2000 John Wiley & Sons, Ltd.
Figure 1. Mortality rate in Germany and incidence from East German Cancer Registry for lung cancer, men and women.

has found considerable interest recently. We call disease mapping with mixture models which also incorporates time period aspects disease mapping by means of dynamic mixtures (DMDM).

2. THE DATA SET ON LUNG CANCER

The health region under study consists of the 215 counties of the former German Democratic Republic (the five new states of Germany since 1990). The regions do not include the western part of Berlin. Lung cancer incidence data (ICD 162) for women were available from 1960 to 1989, in three time periods: 1960–1969; 1970–1979; 1980–1989. From a public health point of view the investigation of the space-time distribution of lung cancer mortality in women may be fruitful for a number of reasons. First, in contrast to men the incidence and mortality of lung cancer for women in Germany is on the rise [3]. Figure 1 presents mortality data and morbidity data (the latter based on the East German cancer registry) on lung cancer for men and
women. The major risk factor for lung cancer is tobacco smoking [4], thus the mortality of lung cancer is closely related to the increasing prevalence of smoking in women. Estimates of the attributable risk of lung cancer due to smoking range from 57 to 83 per cent in women. Thus lung cancer is frequently addressed as a disease which belongs to the category of avoidable death, that is, deaths which may be avoided by medical and/or preventive intervention [5]. Since the prognosis of patients suffering from lung cancer is poor and at the present time the effectiveness of screening measures is controversial [6], the major preventive action available are smoking cessation programmes as a means of primary prevention. Thus identification of high risk areas in disease maps could provide targets for preventive action such as smoking cessation programmes.

3. MAP CONSTRUCTION FOR A SINGLE TIME PERIOD

The first step in the construction of disease maps is usually the choice of an epidemiological measure which shall be presented on the map. A frequently used measure is the standardized mortality ratio \( SMR = \frac{O}{E} \) where the expected cases \( E \) are calculated based on a reference population and \( O \) denotes the observed cases.

A common approach in map construction is the chloropleth method [7]. This method implies categorizing each area and then shading or colouring the individual regions accordingly. Frequently the categorization of the individual region is of particular importance.

Traditional approaches of categorization are based on the percentiles of the SMR distribution. Most cancer atlases use this approach usually based on quartiles, quintiles or sextiles. Figure 2 contains a map for female lung cancer based on quartiles. Now it is widely accepted that these traditional methods do not provide a valid method of map construction and should at least be accompanied with a smoothed map based on (empirical) Bayesian methods [8].

A more flexible approach is given in terms of hierarchical modelling leading to simple random effects models, that is, models where the distribution of relative risks \( \theta_i \) between areas is assumed to have a probability density function \( g(\theta) \) as has been suggested by Clayton and Kaldor [9]. Given the area or tract \( i \) of interest, the number of incidence or mortality cases \( O_i \) are assumed to be Poisson distributed conditional on \( \theta \) with expectation \( E(O_i) = \theta_i E_i \). It is conventionally assumed that there are \( n \) areas available with associated values for \( O_i \) and \( E_i \), \( i = 1, \ldots, n \).

Parametric modelling of \( g(\theta) \) with the gamma–Poisson model has been done extensively in the context of disease mapping by Martuzzi and Hills [10,11]. If we assume that our population under scrutiny consists of subpopulations with different levels of disease risk \( \theta_j \), we are led to a non-parametric mixture approach. Namely, each of these subpopulations with disease risk \( \theta_j \) represents a certain proportion \( p_j \) of all regional units. Statistically, this means that the mixing distribution reduces to a finite mass point distribution. Here we face the problem of identifying the level of risk for each subpopulation and the corresponding proportion of the overall population. One can think of this situation as a hidden (or latent) structure, since the subpopulation to which each area belongs remains unobserved. These subpopulations may have different interpretations. For example, they could indicate that an important covariate has not been taken into account. Consequently, it is straightforward to introduce an unobserved or latent random vector \( Z \) of length \( k \) consisting of only 0s besides one 1 at some position (say \( j \)th), which then indicates that the area belongs to the \( j \)th subpopulation. Taking the marginal density over the unobserved random variable \( Z \) we are led to a discrete semi-parametric mixture model. If we assume a non-parametric
Figure 2. Disease maps based on quartiles for female lung cancer; quartiles are determined from data of all three time periods.
distribution

\[ P = \begin{bmatrix} \theta_1 & \cdots & \theta_k \\ p_1 & \cdots & p_k \end{bmatrix} \]

for the mixing density \( g(\theta) \) (whose maximum likelihood estimate can be shown to be always discrete in its nature), we obtain the mixture density as the weighted sum of Poisson densities for each area \( i \):

\[
f(o_i, P, E_i) = \sum_{j=1}^{k} p_j f(o_i, \hat{\theta}_j, E_i), \quad \text{with} \quad \sum_{j=1}^{k} p_j = 1 \text{ and } p_j \geq 0
\]

\( j = 1, \ldots, k \). Here, \( f(o, \theta, E) = \text{Po}(o | \theta E) = e^{-\theta E} (\theta E)^o / o! \). In the following we will denote the random variable of observed cases by \( O \), whereas its actual observation will be denoted by \( o \). The expected number of cases is assumed to be known and non-random and will always be denoted as \( E \).

Please note that the model consists of the following parameters: the number of components \( k \); the \( k \) unknown relative risks \( \theta_1, \ldots, \theta_k \), and \( k - 1 \) unknown mixing weights \( p_1, \ldots, p_{k-1} \). The likelihood associated with this model can be provided as

\[
L(P) = \prod_{i=1}^{n} f(o_i, P, E_i) \quad (1)
\]

There are no closed form solutions available for finding the maximum likelihood estimates; suitable algorithms are given by Böhning et al. [12]. An overview about reliable algorithms may be found in Böhning [13]. Public domain software to estimate the parameters of the mixture is available with the package C.A.MAN [12, 14]. For the special case of disease mapping the package DismapWin [15] may be used. A general strategy implies calculating the non-parametric maximum likelihood estimator (NPMLE) and then applying a backward selection strategy to determine the number of components by means of the likelihood ratio statistic [15, 16]. For our data in the time frame from 1980–1989 we obtain a mixture model with four components, that is, \( \hat{k} = 4 \), and the following estimate of the mixing distribution:

\[
\begin{pmatrix} \theta_1, \theta_2, \ldots, \theta_k \\ p_1, p_2, \ldots, p_k \end{pmatrix} = \begin{pmatrix} 1.33 & 0.98 & 0.78 & 0.56 \\ 0.12 & 0.27 & 0.51 & 0.10 \end{pmatrix}
\]

Applying Bayes theorem and using the estimated mixing distribution as a prior distribution we are able to compute the probability for each region to belong to a certain component:

\[
\hat{p}_{ij} = \Pr(Z_{ij} = 1 | o_i, \hat{\theta}_j, E_i) = \frac{\hat{p}_j f(o_i, \hat{\theta}_j, E_i)}{\sum_{l=1}^{k} \hat{p}_l f(o_i, \hat{\theta}_l, E_i)}
\]

The \( i \)th area is then assigned to that subpopulation \( j \) for which it has the highest posterior probability of belonging to. In terms of the latent vector \( Z \) the Bayes theorem gives us its posterior distribution.

The posterior expectation for this model may be computed as follows:

\[
\hat{\theta}_i(o_i, \hat{\theta}, E_i) = E(\theta_i | o_i, \hat{\theta}, E_i) = \frac{\sum_{j=1}^{k} \hat{p}_j f(o_i, \hat{\theta}_j, E_i) \hat{\theta}_j}{\sum_{l=1}^{k} \hat{p}_l f(o_i, \hat{\theta}_l, E_i)}
\]
Figure 3. Disease maps for female lung cancer using a separate mixture model for each time period.
For the construction of the actual disease map the posterior probability may be utilized; area \( i \) is classified into that subpopulation \( j \) for which the posterior probability is largest, \( \hat{\rho}_{ij} = \max \hat{\rho}_{ij} \). The corresponding maps for the three periods of lung cancer data for East Germany are given in Figure 3. Note that each period has its own mixture model.

4. MAP CONSTRUCTION FOR SEVERAL TIME PERIODS

The development of methods for the spatial analysis of disease occurrence has been paralleled by the developments for methods where time is implicitly included in the analysis. The information on time can have various forms. For example, we might know for a case the duration of an illness, the time of first diagnosis etc.

In this contribution we assume that we have different fixed time windows, such as 5-year periods or 10-year periods, for which the number of disease counts are available. This situation is usually called the fixed time period case [17]. There have been some developments for this situation including Knorr-Held and Besag [18], Waller et al. [19] and Bernardinelli et al. [20]. These approaches define typically a Poisson (binomial) likelihood for the disease counts with a log-linear model for the spatial and spatio-temporal components.

The basic idea here is to consider the time-space data as one data set and model a single mixture distribution to the space-time data set. More precisely, let \( O_{it} \) and \( E_{it} \) be the observed and expected cases for time \( t \), \( t = 1, \ldots, T \) and area \( i, i = 1, \ldots, n \). We assume the existence of \( k \) possibly unknown time-space components (clusters) which we model with one nonparametric distribution

\[
P = \begin{bmatrix} \theta_1 & \cdots & \theta_k \\ \rho_1 & \cdots & \rho_k \end{bmatrix}
\]

for the mixing density \( g(\theta) \). We obtain the mixture density as weighted sum of Poisson densities for each area \( i \) and each time period \( t \):

\[
f(o_{it}, P, E_{it}) = \sum_{j=1}^{k} p_j f(o_{it}, \theta_j, E_{it}), \quad \text{with} \quad \sum_{j=1}^{k} p_j = 1 \quad \text{and} \quad p_j \geq 0
\]

\( j = 1, \ldots, k \). The likelihood is in this case

\[
L(P) = \prod_{t=1}^{T} \prod_{i=1}^{n} f(o_{it}, P, E_{it}) \quad (2)
\]

It might be helpful in understanding this model to compare the likelihood (2) with the likelihood (1) of the previous section if it is applied for \( T \) time periods:

\[
\prod_{t=1}^{T} L(P^t) = \prod_{t=1}^{T} \prod_{i=1}^{n} f(o_{it}, P^t, E_{it}) \quad (3)
\]

where \( P^t \) is the mixture model for the \( t \)th time period. Both models offer flexibility since the number of components \( k \) is not fixed in both models. However, it can be expected that the latter modelling approach will need fewer parameters. For example, suppose the population is homogenous in time and space. Then the second approach would need only one parameter, whereas the first approach would still estimate \( T \) parameters. In addition, the model leading to the likelihood (2) is easier
Classification of the areas into the $T$ maps would be done again with the posterior probability $\hat{p}_{ij}$ with

$$\hat{p}_{ij} = \Pr(Z_{ij} = 1|o_{it}, \hat{P}, E_{it}) = \frac{\hat{P}_{i}f(o_{it}, \hat{\theta}_{i}, E_{it})}{\sum_{l=1}^{k} \hat{P}_{l}f(o_{it}, \hat{\theta}_{l}, E_{it})}$$

so that area $i$ in time period $t$ is classified in that component $j$ for which $\hat{p}_{ij}$ is largest (comparisons are done over $l, l = 1, \ldots, k$). We apply this model to the lung cancer data set we have considered before. The results are shown in Figure 4. A four component mixture model is found. About 32 per cent of the area-time data are classified into the non-elevated component with relative risk of 1.01. About 52 per cent are classified into the component with decreased relative risk of 0.76, about 6 per cent into the low relative risk component of 0.53; 10 per cent are classified into the elevated component with relative risk of 1.31. It is quite clear that this component is weakly represented in the period 1960–1969, and becomes stronger over the years with concentration in the south-west of the capital (Berlin) of Germany.
5. MODELLING TIME DEPENDENCY WITH COVARIATES

Now the question arises how to include known covariates, in this case the covariate time period into the mixture model. The inclusion of covariates leads into the area of ecologic studies; for a detailed discussion see the contribution by Biggeri et al. [21]. A natural extension of the homogenous Poisson regression model is given by the mixed Poisson regression model [22–24]. We are interested in extending the univariate Poisson mixture model \( O_{it} \sim p_1 \text{Po}(o_{it}, \theta_1, E_{it}) + \cdots + p_k \text{Po}(o_{it}, \theta_k, E_{it}) \) to a multivariate version which takes into account a potential covariation in time. Consider the joint distribution of all time periods for area \( i \): \( f(o_{i1}, o_{i2}, \ldots, o_{iT}) \). Let us assume that the \( O_{i1}, O_{i2}, \ldots, O_{iT} \) become independent conditional on the covariate information on time and the information on the component membership, in other words

\[
f(o_{i1}, o_{i2}, \ldots, o_{iT} | \text{LP}_{itj}, t = 1, \ldots, T) = \prod_{t=1}^{T} f(o_{it} | \text{LP}_{itj})
\]  

(4)

Here, the linear predictor \( \text{LP}_{itj} \) in the \( j \)th component is given as

\[
\text{LP}_{itj} = \alpha_j + \beta_j t + \log E_{it}
\]

where we think of \( \alpha \) as the intercept and \( \beta \) denoting the period effect. As usual, \( f(o_{it} | \text{LP}_{itj}) = \text{Po}(o_{it}, \exp(\text{LP}_{itj})) \).

The marginal distribution of \( (O_{i1}, O_{i2}, \ldots, O_{iT}) \) is then given as

\[
\sum_{j=1}^{k} p_j f(o_{i1}, o_{i2}, \ldots, o_{iT} | \text{LP}_{itj}, t = 1, \ldots, T) = \sum_{j=1}^{k} p_j \prod_{t=1}^{T} f(o_{it} | \text{LP}_{itj}) = \sum_{j=1}^{k} p_j \prod_{t=1}^{T} \text{Po}(o_{it}, \exp(\text{LP}_{itj}))
\]

Again, estimation may be done by maximum likelihood. Note that the likelihood is in this case

\[
L(P) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j \prod_{t=1}^{T} f(o_{it} | \text{LP}_{itj})
\]  

(5)

where now \( P \) is giving weights \( p_j \) to the effect vectors \( \beta_j \):

\[
P = \begin{bmatrix}
\beta_1 \\
\vdots \\
\beta_k
\end{bmatrix}
\begin{bmatrix}
p_1 \\
\vdots \\
p_k
\end{bmatrix}
\text{with } \beta_j = (\alpha_j, \beta_j)^{T}
\]

Again, there are no closed form solutions available for maximum likelihood estimates. An adaptation of the EM algorithm of Dempster et al. [25] has been developed by Dietz [22]; see also Mallet [26] for a discussion of maximum likelihood estimation. The areas are classified again into the various components (of the regression model) using the posterior distribution. Note that now each regression model corresponds to a colour (grey pattern) in the map. A detailed description can also be found in Schlattmann et al. [23]. The computations involved may be done with the program DismapWin [15].

In our example of female lung cancer, we include time period as a covariate into the model. Here we consider a full random effects model, that is, the effect of the covariate time may differ
in each component of the mixture model. For our data we obtain a mixture model with four components, that is, $k = 4$, and the following estimate of the mixing distribution:

\[
\begin{pmatrix}
    (\lambda_1, \beta_1) & (\lambda_2, \beta_2) & (\lambda_3, \beta_3) & (\lambda_4, \beta_4) \\
    p_1 & p_2 & p_3 & p_4
\end{pmatrix}
\]

\[
= \begin{pmatrix}
    (-0.27, -0.03) & (0.00, -0.01) & (0.20, 0.05) & (0.68, 0.04) \\
    0.53 & 0.38 & 0.07 & 0.01
\end{pmatrix}
\]

The results are also shown in Figure 5. There are two low-weighted components with elevated baseline risks of 1.22 and 1.97 and a positive time effect (8 per cent of all areas are categorized into these components). The associated areas can be found again south-west of Berlin.

6. EVALUATION AND DISCUSSION

The area of disease mapping already has some traditional approaches [7, 17, 27] with many interesting aspects such as the empirical [28, 29], full Bayesian perspective [28–30], the local and global shrinkage aspect [31], and the construction of disease maps by means of mixture models [32]. It has become of more recent interest, however, to evaluate the diverse methods available and reach some conclusions about which of these methods should be used [33]. Here we want to give only some elementary consideration as to how the mixture models we have considered in this paper, perform.

In Table I we have provided some likelihood-based performance criteria. It has become common practice to provide the Bayesian information criterion as a device for discriminating across a diversity of models. Suppose $M_i$ is the model of interest (single mixture model for each time period (I), mixture model for time and space components (II), Poisson regression model (III)).
and $M_0$ the associated *homogeneous* model. Then

$$BIC = 2[\log(L_{M_1}) - \log(L_{M_0})] - (p_1 - p_0) \log(nT)$$  

(6)

where $p_1$ is number of parameters under $M_1$ and $p_0$ is the number of parameters under $M_0$. For example, if $M_1$ is model I, then there are two components in the mixture model for the first time period (3 parameters), three components for the second time period (5 parameters), and four components for the last period (7 parameters), which totals to $p_1 = 15$ parameters. In this case $M_0$ corresponds to three one-component mixture models, leading to $p_0 = 3$ parameters in total. It is quite evident that the mixture model which models time and space simultaneously is performing quite well for this data set.

**ACKNOWLEDGEMENTS**

The work of Dankmar Böhning and Ekkehart Dietz is under the support of the German Research Foundation. Dankmar Böhning and Peter Schlattmann are supported by the BIOMED2-grant Disease Mapping and Risk Assessment.

**REFERENCES**