Space-Time Mixture Model of Infant Mortality in Peninsular Malaysia from 1990-2000

NUZLINDA ABDUL RAHMAN
Universiti Kebangsaan Malaysia
Faculty of Science & Technology
School of Mathematical Sciences
43600 UKM Bangi, Selangor
MALAYSIA

ABDUL AZIZ JEMAIN
Universiti Kebangsaan Malaysia
Faculty of Science & Technology
School of Mathematical Sciences
43600 UKM Bangi, Selangor
MALAYSIA

Abstract: Disease mapping is a method used to display the geographical distribution of disease occurrence. Recently, this method has received much attention from many researchers including epidemiologists, biostatisticians and medical demographers. Some traditional methods of classification for detection of high or low risk area such as traditional percentiles method and significant method have been used in disease mapping for map construction. However, as described by several authors, the classifications based on these traditional methods have some disadvantages for describing the spatial distribution of the risk of the disease concerned. To overcome these limitations, an approach using mixture model within an empirical Bayes framework is described in this paper. The aim of this study is to investigate the geographical distribution of infant mortality in Peninsular Malaysia from the year 1991 to 2000 using space-time mixture model. The analysis showed that in the early year of 1990's the spatial heterogeneity effect was more prominent; however, towards the end of 1990's this pattern tends to disappear. Indirectly, this may indicate that the provisions of health services throughout the Peninsular Malaysia are uniformly distributed over the period of the study, particularly towards the year 2000.

Key–Words: Disease mapping, Space-time mixture model, Infant mortality, Geographical distribution, Spatial heterogeneity, Pattern

1 Introduction

Child health is a central issue amongst the public in many developing countries. Infant mortality rate is one of the most common measure used to describe the level of services relating to health, socio-economic and education of a country. Since independence in 1957, Malaysia had experienced a very remarkable decline in infant mortality from the rate of around 100 per thousand to around 13 per thousand by the late 1980's. It was reported that this rate has been reduced to 9 per thousand in 2004. This achievement is nearly equal to the rate experienced by developing countries such as United States and Britain, with 7 and 6 death per thousand respectively. The decline in infant mortality rate in Malaysia could possibly be due to the prosperous socio-economic situation where the average incomes have increased over the years. Moreover, the basic facilities such as water supply, electricity, sewage, sanitation and health services have been improved provided to the wider population of the country. Apart from that, the levels of education and health consciousness have increased among Malaysians and other factors that directly and indirectly influence the infant mortality in Malaysia such as ethnicity, mother' education, preceding birth interval, birth place, etc [1].

The basic concept of mapping is to group the information in the data for all regions in the study area into several components or exclusive groups, where individual region in the same component has a similar risk. One way of displaying the variability of disease or mortality rate is by a widely used technique called disease mapping. It is very useful to produce such maps especially for government agencies in resource allocation or identifying hazards that contribute to the disease [2]. Usually, in the context of health sector, the authority in charged aims to identify whether the risks for a particular disease concerned are uniformly distributed or homogeneous for different regions of the country. For example, as mentioned earlier, it is fortunate that the infant mortality rate in Malaysia have improved over the last few decades, but the issue concerned is whether the improvement is uniformly distributed throughout the country. Does every district experience the same level of improvement or reduction of the risks? Does the improvement only occur in certain areas while the other areas still

remain in the high risk areas category? If there is a huge gap between the high risk areas and the low risk areas, the disease risks can be divided to several categories or considered as heterogeneous. This is the main issue that will be addressed in this paper in the context of disease mapping context of infant mortality in Malaysia.

In disease mapping, let us divide the study area to be mapped into M mutually exclusive districts $(i = 1, 2, ..., M)$. Each district has its own observed number of cases, O_i and expected number of cases, E_i. The expected number of cases is calculated as;

$$E_i = N_i \sum o_i / \sum N_i \tag{1}$$

where N_i is the population for area i [3]. Here the standardization is done on the total population at risk. The standardization can be done on other factors such as age, gender, etc and this method have been discussing by several authors [4, 5].

It is common to assume that the observe number of cases, O_i, follows a Poisson distribution with expectation $E_i\theta$ and the probability density function is defined as:

$$\begin{aligned} Pr(O_i = o_i) &= \frac{\exp(-\theta E_i)(\theta E_i)^{O_i}}{o_i!} \\ &= f(o_i, \theta, E_i) \end{aligned} \tag{2}$$

where θ is the relative risk of disease concern over the study area. Using O_i and E_i as obtained based on the data, we can have one of the most common indexes to estimate the relative risk for region i, θ_i i.e Standardized Mortality Ratio (SMR) defined as:

$$\hat{\theta}_i = SMR_i = \frac{O_i}{E_i} \tag{3}$$

In map construction, the important elements are obtaining smoothed estimators of relative risk and categorizing or classification of all districts into several components using shading or colouring to differentiate the level of risks for each component. Although, SMR has been used commonly as an index to measure relative risk; however, it has some weaknesses where the variance of SMR, θ_i/E_i depends on E_i. The variance will be large when the expected value is small, as contributed by the small population size and the variance will be small when the expected value is large due to the large population size. If the observed value is zero such as in the case of rare disease, the SMR and the standard deviation will be zero. Another limitation of SMR is the instability of the relative risk estimation due to the presence of extreme SMR when rare diseases are investigated in small population areas [6].

To overcome the drawbacks of the SMR a Bayesian approach had been used which allow for the risk to vary between the different districts as given by the assumption:

$$O_i \sim Poisson(E_i\theta_i) \tag{4}$$

and the probability density function is define as:

$$\begin{aligned} Pr(O_i = o_i) &= \frac{\exp(-\theta_i E_i)(\theta_i E_i)^{O_i}}{o_i!} \\ &= f(o_i, \theta_i, E_i) \end{aligned} \tag{5}$$

For example, the empirical Bayes of the relative risks where a random effects (or mixture) model that assumes a parametric probability density function (pdf), denoted as $f(\theta)$ for the distribution of relative risks between districts were adopted [7]. This modeling approach has been used in many fields including disease data by applying several empirical Bayes methods of estimation to smooth the SMR [8, 9, 10]. Some authors have provided discussion on hierarchical Bayesian approach with structured and unstructured spatial random effects [11]. Although the Bayesian methods can provide estimate on relative risks for each district, the number of optimum classification for categorizing the districts cannot be obtained based on them. The most common approach that is widely used by many researchers for categorizing areas in disease mapping is classification based on quartiles. However, this method is rather arbitrary and has no guarantee in detecting the classification of high or low risk areas. Another disadvantage of the Bayesian relative risks estimation is the usage of assumption in Eq. (4) will give the number of parameters and the number of districts is the same. If the numbers of districts are large, there might be difficulties in estimating parameters consistently because of too many parameters to be estimated. As an alternative approach, a method have been suggested to overcome these drawbacks which include the time factor and consider spatial heterogeneity effect will be discussed in this paper known as space-time mixture model within non-parametric approach for map construction. This method has appeared to be very attractive approach for practical applications and become a more flexible tool [12]. Infant mortality data in Peninsular Malaysia from 1991 to 2000 will be applied using this approach to examine the geographical distribution of the disease concern.

2 Methodology

Space-time mixture model is an extension model of mixture model by including the time factor in order

to study the disease pattern in certain period of time. This model gives a valuable indication of an emerging pattern over time because it looks simultaneously for all space-time components [6]. The basic idea of space-time mixture model approach in the context of disease mapping is to consider all space-time data as a single data set. The same steps of modeling the mixture model will be applied in estimating parameters, so in this paper, the discussion on mixture model will be presented in application to space-time data. In mixture model, we assume that the population comes from several heterogeneous components where every component consists of different risk levels of disease. This assumption will give a more heterogeneous case. Assume that the mixture model consists of c components and each component has a disease risk, θ_j $j = (1, ..., c)$. Let p_j denotes the proportion of regional areas having θ_j risk. This discrete parameter distribution P for describing the level of risk can be given as:

$$P = \begin{bmatrix} \theta_1, ..., \theta_c \\ p_1, ..., p_c \end{bmatrix} \tag{6}$$

Accordingly, we may assume that observed data in district i of a particular year t, o_{it} comes from a non-parametric mixture density identified in the following form:

$$f(o_{it}, P, E_{it}) = \sum_{t=1}^{T} \sum_{j=1}^{c} P_j f(o_{it}, \theta_j, E_{it}) \tag{7}$$

where $p_1 + ... + p_c = 1,\ p_j \geq 0,\ j = 1, ..., c$ and $t = 1, ..., T$. E_{it} is the expected number of cases in district i of a particular year t. The numbers of parameters to be estimated in the model with c components considered above are $2c - 1$ which consists of c unknown relative risks $\theta_1, ..., \theta_c$ and $c-1$ unknown mixing weights $p_1, ..., p_{c-1}$ where $f(.)$ denotes the Poisson density taken from previous assumption [13].

One of the basic issue in mixture model is whether the number of components c, is unknown or assumed to be known [14]. They called the two cases as flexible support size and fixed support size respectively. However, in both cases, the maximum likelihood approach can be applied for the parameter estimation. In the estimation based on flexible support size, a grid containing θ_j's is defined and the corresponding p_j that maximized the likelihood function is determined. However, in this paper, the fixed support size is considered and outline of the algorithms used for this estimation is the EM algorithm [15].

In EM algorithm, the first step of mixture model involves estimating θ_j and p_j in each component by giving their initial values. These initial values and the number of component to be estimated can be obtained from the histogram of relative risks or SMR where the height of the bars maybe used as the estimate for proportion corresponding to relative risks while the number of bars maybe used as the number of components. We are interested to determine the membership of each district to which particular component. Let us denote the full data as $(o_{it}, E_{it}, x_{i1t}, x_{i2t}, ..., x_{ict})$ where x_{ijt} indicate the membership of district i in the jth component for the year t. For example, if region i in the year t belongs to the third component, can be written as $x_{it} = (0, 0, 1, 0, ..., 0)^T$. Based on the information of the initial weights, $\hat{p}_j$ and relative risks, $\hat{\theta}_j$ obtained, we can execute the EM algorithm which consist of E-step and M-step. The E-step consists of the calculation of the probability of each district belonging to jth component while the M-step consists of the calculation of the weights and relative risks. These two steps will be repeated alternately until the convergence criterion is met and can be summarized as below:

E-step:

$$\begin{aligned} w_{ijt}^{(r)} &= Pr\left(X_{ijt} = 1 | o_{it}, P, E_{it}\right) \\ &= \frac{\hat{p}_j^{(r)} f\left(o_{it}, \theta_j^{(r)}, E_{it}\right)}{\sum_{j=1}^{c} \hat{p}_j^{(r)} f\left(o_{it}, \theta_j^{(r)}, E_{it}\right)} \end{aligned} \tag{8}$$

M-step:

$$\hat{p}_j^{(r+1)} = \frac{\sum_{i=1}^{M} w_{ijt}^{(r)}}{M} \tag{9}$$

and

$$\theta_j^{(r+1)} = \frac{\sum_{i=1}^{M} w_{ijt}^{(r)} \frac{o_{it}}{E_{it}}}{\sum_{i=1}^{M} w_{ijt}^{(r)}} \tag{10}$$

When the convergence is obtained, the next step is to compute the non-parametric maximum likelihood estimator (NPMLE) that maximizes the log-likelihood function which is defined as:

$$\begin{aligned} l_c &= \sum_{t=1}^{T} \sum_{i=1}^{M} \log f(o_{it}, P, E_{it}) \\ &= \sum_{t=1}^{T} \sum_{i=1}^{M} \log \left\{ \sum_{j=1}^{c} p_j f(o_{it}, P, E_{it}) \right\} \end{aligned} \tag{11}$$

Further step is to determine the most suitable number of components by computing the difference between the log-likelihood for c components and $c+1$ components, which is known as Likelihood Ratio Statistics (LRS) and is defined as:

$$LRS = -2(l_c - l_{c+1}) = -2 \log \theta \tag{12}$$

The purpose of calculating the LRS is to test this hypothesis:

H_o : number of components is c

H_a : number of components is $c+1$

A problem arises in determining the number of components when the solution consists of the log-likelihood values that are nearly the same for every component. Conventionally, the LRS test has an asymptotic chi-square distribution with degrees of freedom equal to the difference between the number of parameters under the alternative and null hypothesis. However, this conventional results for LRS do not hold for mixture and a method proposed to obtain the critical values in determining the number of components is via a simulation technique for example by parametric bootstrap [14].

Once the optimum number of components is obtained, the final step in mixture model approach is to classify the membership of each district to which component. Classification can be obtained by applying Bayes' theorem which involves computing the probability of each district belonging to each component with the posterior probability given by:

$$Pr\left(X_{ijt}=1\right)=\frac{\hat{p}_j f(o_{it},\theta_j,E_{it})}{\sum_{j=1}^{c}\hat{p}_j f(o_{it},\theta_j,E_{it})} \qquad (13)$$

for $i=1,...,M,\ j=1,...,c$ and $t=1,...,T$. The ith district in the year t will belong to the component or subpopulation j if the posterior probability of this belonging is highest.

3 Result

Data analysis based on space-time mixture model is illustrated using the infant mortality data in Peninsular Malaysia from the year 1991 to 2000. Comparisons of the results obtained by this method throughout the study period become easier as all maps for each particular year have the same categorization. Table 1 shows the result on how to determine the optimum number of components based on log-likelihood, l_c mentioned above. From this table, the models with four and five components have the lowest log-likelihood value, which is the same. Since there is no improvement in log-likelihood value for space-time mixture model with five components and by considering the parsimony factor, we choose a model with four components as the best model to fit the space-time data used in this study. Although it has been suggested by some studies that bootstrap method should be applied in deciding either to choose between c or $c+1$ components, we based our decision on the previous argument. From the fitted model with four components, the first category had the lowest risk with mean of 0.726 and weight of 0.324 and the highest risk category with mean of 2.199 and weight of 0.014. The analysis of the space-time infant mortality data in Peninsular Malaysia over the last decade leads to the mixture density with four components given by:

$$\begin{aligned} f\left(o_{it},\hat{P},J_{it}\right) &= f(o_{it},0.726,J_{it})\times 0.324 \\ &+ f(o_{it},1.131,J_{it})\times 0.550 \\ &+ f(o_{it},1.630,J_{it})\times 0.112 \\ &+ f(o_{it},2.199,J_{it})\times 0.014 \end{aligned} \qquad (14)$$

Table 1: Result of space-time mixture model for infant mortality data in Peninsular Malaysia from 1991 to 2000.

Number of components (c)	Mean relative risks ($\hat{\theta}_j$)	Weight ($\hat{p}_j$)	log-likelihood (l_c)	LRS
$c=5$	0.726	0.324	-	
	1.131	0.550	2625.094	
	1.630	0.112		0.000
	2.199	0.012		
	2.199	0.002		
$c=4$	0.726	0.324		
			-	
			2625.094	
	1.131	0.550		
	1.630	0.112		72.420
	2.199	0.014		
$c=3$	0.749	0.370	-	
			2661.304	
	1.175	0.547		330.504
	1.810	0.083		
$c=2$	0.834	0.564	-	
			2826.556	
	1.374	0.436		1515.370
$c=1$	1.070	1.000	-	
			3584.241	

Corresponding to the results given in the table, we can summarize the geographical distribution of infant mortality throughout the study period as given in Figures 1, 2 and 3 by providing the space-time maps for the year 1991, 1996 and 2000, respectively. As each map obtained throughout the study period have the same classification, the comparison and fluctuation of the disease concern over time is easier to compare and interpret. In the early year of 1990's, it can be seen

that only about 6% of the districts fall in lowest risk areas, however in the middle and late 1990's, the infant mortality had improves with almost 50% or more of the districts were in this category. These figures also shown that none of the districts were in the highest risk category in the year 1996 and 2000 but four districts were in this category in 1991. These changes indicate that infant mortality in Peninsular Malaysia had improved over the last decade and tends to be more homogeneous towards the end of the study period.

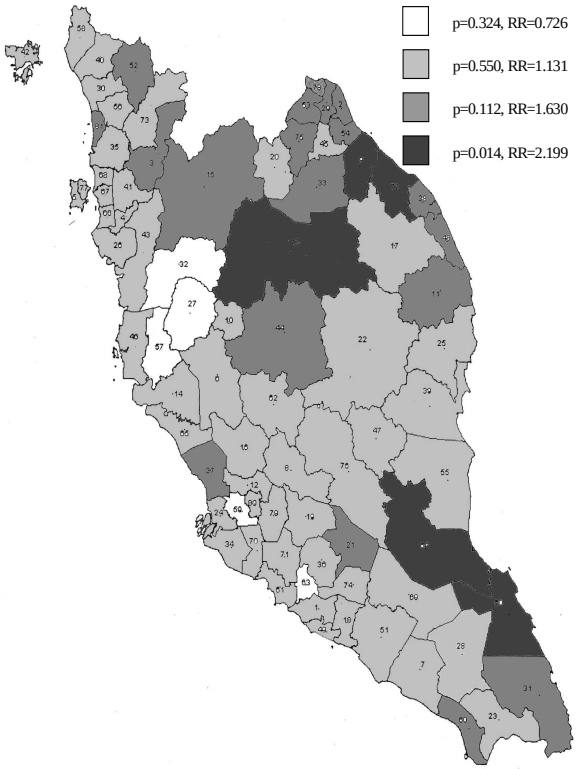

Figure 1: Infant mortality maps in Peninsular Malaysia based on space-time mixture model for the year 1991

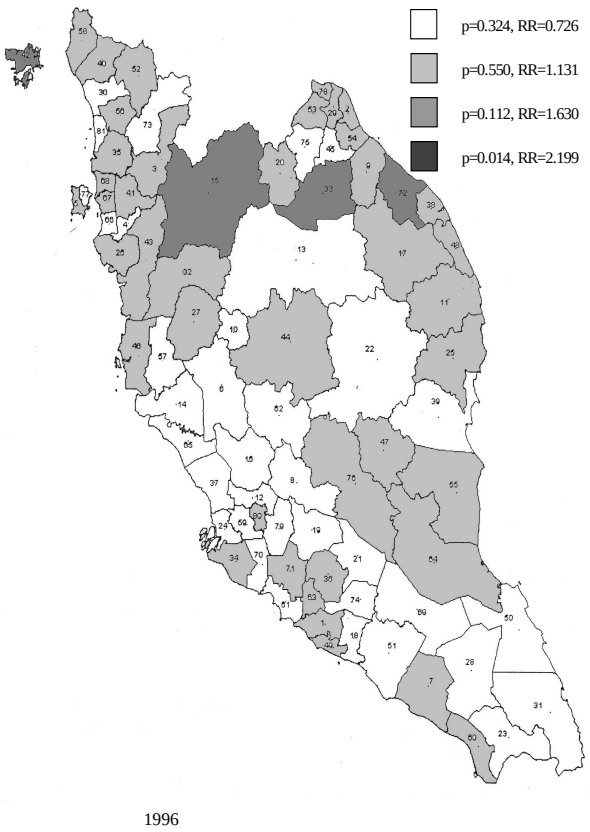

Figure 2: Infant mortality maps in Peninsular Malaysia based on space-time mixture model for the year 1996

4 Discussion

For quite some time, many researchers have conducted various studies in disease mapping using the traditional methods of classification such as percentiles method and significant method. However, these methods have some deficiencies and potential of misrepresenting the graphical distribution and question regarding whether these classifications give a correct interpretation may be raised [14]. An alternative approach suggested is the mixture model that could produce a smoother map where the random variability has been extracted from the data. Other main advantages of using the mixture distribution are its discreteness making the map construction is straightforward and provides the optimum number of components. The inclusion of space-time factor in mixture model satisfactorily produces maps that easier to interpret compare by looking the maps separately since every map for each year throughout the study period have the same classification.

Based on the three maps in Figure 1–3, it is very clear that the space-time mixture model has removes random variability from the map and provides a better and clearer picture of classification for high and low risk areas. For the period of 10 years i.e from 1991-2000, we can conclude that the classifications tend to be more homogeneous implying that the random variability has reduced with time. Furthermore, towards the end of the study period, maps obtained shown that more districts have fallen into the low risk categories which indicate that the infant mortality in Peninsular Malaysia have improved within the last decade.

There are many factors that contribute to the re-

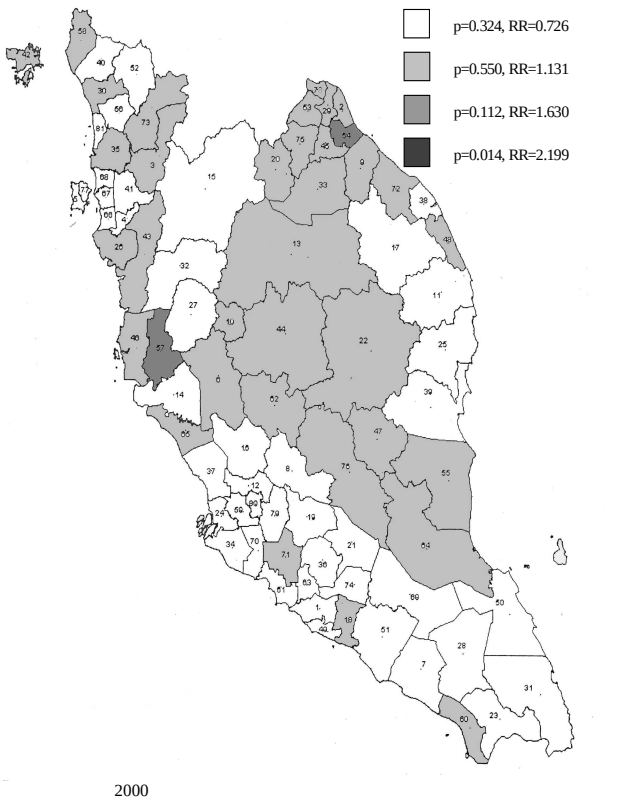

Figure 3: Infant mortality maps in Peninsular Malaysia based on space-time mixture model for the year 2000

duction of infant mortality. Some literatures stated that infant mortality is more likely to be related to the socio-economic level, health behavior, quality of antenatal care, support during delivery, postnatal care, nutritional status, education level, unemployment and birth intervals [16, 17, 18]. These factors were addressed in the case of Malaysia as shown by the increased in health of RM17.30 per capita in 1970 to RM248 per capita in the year 2000 [19]. The number of hospitals was increased from 84 public hospitals in 1965 to 116 public hospitals in 2002 along with many private hospitals, health clinics and rural clinics being built throughout the country to provide better health system in Malaysia [20]. As the number of hospitals increased, more facilities were upgraded such as providing more hospital beds while at the same time the number of registered doctors, trained nurses and midwives have also been increased [20]. In general, the health and medical services in Malaysia have significantly improved in the past four decades since independence contributing to the improvement in infant mortality rates. The government has put in a lot of effort especially in terms of the quality of service, the advancement of medicine and medical technologies, the resolution of the issue of unbalanced distributions of medical resources between rural and urban areas, the establishment of collaborations among government and private hospitals or medical institutions and etc. A lot of campaigns and programmes have been done by the local government and the Ministry of Health to educate and increase the health consciousness among Malaysians.

In conclusion, as discussed before, even though the space-time mixture model have some advantages in estimating the disease risks and provide a better and clearer picture of categorization, this approach still has a weakness in which the relative risk for different districts could possibly be correlated, i.e dependent on geographical proximity. An example of model that can be used which includes the neighboring factor among the areas is the parametric conditional autoregressive model [4].

Acknowledgements: The first author acknowledge Universiti Sains Malaysia and Ministry of Higher Education, Malaysia, for the financial support received throughout the course of this study.

References:

[1] W.–N. Mohamed, I. Diamond and W.–F.–P. Smith, The determinants of infant mortality in Malaysia: a graphical chain modeling approach, *J. R. Statistical Society A* 161, 1998, pp. 349-366.

[2] A.–B. Lawson and F.–L.–R. Williams, *An Introductory Guide to Disease Mapping*, Wiley, New York, 2001.

[3] I.–H. Langford, A.–H. Leyland, J. Rasbash and H. Goldstein, Multilevel modeling of the geographical distributions of diseases, *Applied Statistics*, 48, 1999, pp. 253–268.

[4] N. Mantel and C.–R. Stark, Computation of indirect-adjusted rates in the presence of confounding, *Biometrics*, 24, 1968, pp. 997–1005.

[5] A.–H. Pollard, F. Yusuff and G.–N. Pollard, *Demographic techniques*, Pergamon Press, Sydney, 1981.

[6] S. Rattanasiri, D. Bohning, P. Rojanavipart and S. Athipanyakom, A mixture model application in disease mapping of malaria, *Southeast Asian Journal Trop Med Public Health*, 35, 2004, pp. 38–47.

[7] H. Robbins, The empirical Bayes approach to statistical decision problems, *Annals of Mathematical Statistics*, 35, 1964, pp. 1–20.

[8] D. Clayton and J. Kaldor, Empirical Bayes estimates of age-standardized relative risks for

use in disease mapping, *Biometrics*, 43, 1987, pp. 671–681.

[9] R.–J. Marshall, Mapping disease and mortality rates using empirical Bayes estimators, *Appl. Statist.*, 40, 1991, pp. 283–294.

[10] J.–L. Meza, Empirical Bayes estimation smoothing of relative risks in disease mapping, *J. of Statistical Planning and Inference*, 112, 2003, pp. 43–62.

[11] A.–B. Lawson, W.–J. Browne and C.–L.–V. Rodeiro, *Disease Mapping with WinBUGS and MlwiN*, Wiley, New York, 2003.

[12] A. Biggeri, E. Dreassi, C. Lagazio and D. Bohning, A transitional non-parametric maximum pseudo-likelihood estimator for disease mapping, *Computational Statistics & Data Analysis*, 41, 2003, pp. 617–629.

[13] B.–S. Everitt and D.–J. Hand, *Finite mixture distributions*, *Chapman and Hall*, New York, 1981.

[14] P. Schlattmann and D. Bohning, Mixture models and disease mapping, *Statistics in Medicine*, 12, 1993, pp. 1943–1950.

[15] P. Schlattmann, E. Dietz and D. Bohning, Covariate adjusted mixture models and disease mapping wih the program Dismapwin, *Statistics in Medicine*, 15, 1996, pp. 919–929.

[16] S.–B. Adebayo, L. Fahrmeir and S. Klasen, Analyzing infant mortality with geoadditive categorical regression model: a case study for Nigeria, *Economics and Human Biology*, 2, 2004, pp.229–244.

[17] S.–O. Rutstein, Effects of preceding birth intervals and neonatal, infant and under-five years mortality and nutritional status in developing countries: evidence from the demographic and health survey, *International Journal of Gynecology and Obstetrics*, 89, 2005, pp. 7–24.

[18] G. Turrell and K. Mengersen, Socioeconomic status and infant mortality in Australia: a national study of small urban areas, 1985-89, *Social Science & Medicine*, 50, 2000, pp. 1209–1225.

[19] *Estimates of Malaysia Federal Revenue and Expenditure*, Ministry of Finance Malaysia, 1970-2000.

[20] *Social Statistics Bulletin Malaysia*, Department of Statistics Malaysia, 1965-2002.