A Comparison of Mixed Effect Models for Spatially Correlated Data

Rafida M. Elobaid

Department of Statistics, Faculty of Science
Sudan University of Science and Technology 407, Khartoum, Sudan
rofaobaid@gmail.com

Abstract

Over the past decade, a variety of hierarchical models have been proposed to model the random effects of a certain phenomenon. The hierarchical models can be used in the epidemiology field to examine the trend of the relative risks associated with the disease spread in lattice grid. Several factors such as covariate effects, interregional heterogeneity and the spatial dependence affect the relative risks among the counties within the lattice grid and lead to over-dispersion. In this article, we provide two types of hierarchical models within a Bayesian paradigm to classify the over-dispersion within different counties after taking into consideration the spatial autocorrelation. When spatial autocorrelation exists we might suspect that random effect corresponding to counties in closer proximity to each other might also be similar in magnitude. A well known set of lip cancer data was used to compare the performance of these models. The model selection was achieved by applying the Deviance Information Criterion. Disease maps were also used to show the trend of the disease spread by inferring the geographical distribution of the rates for each model. Fitting the hierarchical models to the data shows that there exists variability which was not explained by the classical approach. Fitting the global spatial model (GSM) shows that large amount of this variability appears to be from the spatial term. The DIC shows that the mix effect models are much better than the fixed effect model. Clustering trend in the disease map was observed when using the classical approach. The model with fixed and both random effects (GSM) was selected as the best fitted model. It also provides a map of smooth relative risk concluding that this model performs and fit the data well.

Keywords: Hierarchical models, Over-dispersion, Relative risks, Spatial autocorrelation, MCMC, Deviance Information Criterion
1. Introduction

In epidemiology researchers are often interested in mortality or disease occurrences. Usually, they wish to compare mortality or disease rates at two different locations or at two different scales in order to identify whether there are spatial factors that may influence these rates. For example, in different developing regions, we might suspect that random effects corresponding to counties in closer proximity to each other might also be similar in magnitude [22].

Relative risk is one of the several measures which are commonly used to summarize comparisons of disease rates between populations. This measure of association is often used by researchers. One reason for this is that relative risk can be estimated by a wider range of study designs. Furthermore, this measure also works well in most of the cases in epidemiology and clinical fields.

The Standardized Mortality Rate (SMR) is used to estimate the relative risk parameter. SMR is frequently used in occupational epidemiology as a measure of risk. It is commonly used in comparing mortality experiences of different populations at a fixed point of time. SMR can also be used to compare mortality of the same population at different time periods. This is usually done by measuring the amount of relative risk in each population.

Disease incidents estimation was usually analyzed by using classical approach. However, two major problems cited in the literature are firstly, the classical method failed to explain the over-dispersion [5, 15, 21] which is due to random variability and leads to extra heterogeneity in the counties population. Secondly, classical method did not take into consideration the effect of the covariate(s) which have a substantial impact on the risk of the disease. To overcome such problems, Bayesian models were used by including single random effect as unstructured variability term. Besag et al. [13] described a Bayesian approach with spatial random effect.

In this article two main sources are identified as cause of the relative risk heterogeneity. The first source is the over-dispersion between counties that have spatial neighbouring structure, which in turn separated into two main effects, interregional variability exists due to different environmental (local, natural, or built) factors and spatial variability exists due to the spatial autocorrelation among the different counties. The other source is the covariate effect which influences the occurrences of the disease in the different counties.

With the development of Markov Chain Monte Carlo (MCMC) methods, Bayesian approaches are being applied to the analysis of many epidemiological and health problems in addition to disease mapping and modelling. The purpose of the present study is to reveal the advantage of the hierarchical Bayesian analysis over the Bayesian approach which allow for the over-dispersion and covariate effects to be explained. It also develops models that simultaneously link covariate, interregional and spatial effects. Moreover, this paper uses appropriate sets of prior densities such as Gamma, normal and ICAR to produce robust posterior densities.
The present study also reports the results of fitting these random effect models to Scotland lip cancer data after taking into consideration the spatial autocorrelation. Model fit comparison is accomplished using the deviance information criterion introduced by Spiegelhalter et al. [7]. The Bayesian method based on the tools of the geographical information system (GIS) [8, 9, 23] is also used in this paper to examine the clustering trend in the disease maps due to the different source of variability and to show whether these models give more stable estimates of the pattern of the disease risk. Such analysis will inform the growing literature within public health that is focused on the geospatial relation between aspects of the social environment and population and the disease aetiology.

2. Method

This section provides a brief discussion in using the likelihood estimation to identify regions of higher and lower risk [18] via classical approach. The hierarchical approach in this section introduces two models with different sources of heterogeneity to allow for over-dispersion and covariate effects to be explained.

2.1 Classical Approach

Consider a spatial data that are available as summary counts or rates for defined regions (counties) such as a country, district or census tract. Suppose further that for each region, counts or rates are collected within a fixed time interval. In general notation an observation can be denoted by $Y_i$, where $i = 1, ..., N$ indexes the regions [17], represents the observed cases of a disease from $N$ geographical regions, and $E_i$ indicates the expected counts based on known risk factors. The expected counts $E_i$ represent a form of standardization of the data. Assuming that there are no confounding factors, the expected value is computed as follows:

$$E_i = \frac{n_i \sum Y_j}{\sum_j n_j},$$

where $n_i$ is the population in region $i$ and $\sum_j Y_j/\sum_j n_j$ is the overall disease rate. Introducing $\psi_i$ as a vector of relative risk parameter in region $i$, and conditional on the expected counts and the relative risk parameters, the likelihood model assumes that the observed cases $Y_i$ are modelled as independent Poisson random variables,

$$Y_i, \psi_i, E_i \sim PO(\psi_i, E_i), \quad i = 1, ..., N$$

(1)

and the maximum likelihood estimate (MLE) of the relative risk is shown to be,

$$\hat{\psi}_i = SMR_i = \frac{Y_i}{E_i}$$

(2)
where $SMR_i$ is the standardized mortality or morbidity rate in region $i$, i.e. the ratio of the observed to expected disease cases (or deaths).

2.2 Hierarchical Models
Unlike the conventional statistical inference which derives the average estimates of the parameters, hierarchical Bayesian modelling produces parameter estimates for each individual by borrowing information from all analysis units. The ability to incorporate prior knowledge without the restriction of classical distributional assumptions makes the Bayesian inference a potent forecasting tool in a wide variety of fields.

The Bayesian paradigm is based in specifying a probability model for the observed data with specific likelihood function, given a vector of unknown parameters $\theta$, then we assume that $\theta$ is random and has a prior distribution [10, 11]. Inference concerns $\theta$ are based on the posterior distribution, which is obtained by the Bayes theorem. The posterior distribution which involved a contribution from the observed data through the likelihood function and a contribution from the prior information quantified through the prior distribution is used to estimate the $SMR_i$ [3, 20].

To investigate whether there is variability between the different counties due to over-dispersion, we may seek to estimate and map the underlying relative risk surface $\psi = (\psi_1, \ldots, \psi_T)^T$, so we might think of a random effect model for the true relative risk assuming that all the true risks come from a common underlying distribution. Using a random effect model will allow the procedure to borrow strength across the various counties and explain whether the spatial specification (if exist) shows a better estimation with obvious trend in the disease map.

One approach that leads to exploratory relative risk mapping [26] follows the concept of high risk areas [14]. If there is spatial variation in risk, due to the presence or absence of unknown risk factors, then the disease will tend to randomly cluster. Thus, the random cluster trend is an expression of elevated risk due to the risk factors and locates a particularly exposed subpopulation on the map. By extension, the population outside the cluster could be viewed as a less or even unexposed population.

To do such analysis we adopted two hierarchical models that account for the heterogeneity. This achieved by incorporating mix effects that describe the sources of over-dispersion and covariate effects. The models are classified as follows. The interregional variability model (IVM) which includes the covariate effect as a fixed effect, while the interregional variability which is due to non-spatial characteristics is included as a random effect. After counting for spatial autocorrelation using Moran’s $I$ test statistics, the global spatial model (GSM) accommodates the entire fixed and random effects, in which the effects of the covariate and the over-dispersion (interregional and spatial variability) are estimated and exploratory investigated.
A comparison method between the hierarchical Bayesian models IVM and GSM was carried out. The model selection was achieved by using a generalization of Akaike Information Criterion [12], namely the Deviance Information Criterion (DIC) which is given as follows,

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\theta) ,$$

where $\bar{D}$ is the model fit and summarized by the posterior deviance, $D = E_{\theta}(D)$, which has often been used to compare models informally [1, 4, 24, 25, 27, 28]. The effective number of parameters $p_D$ is defined as the posterior mean of the deviance minus deviance evaluated at the posterior mean of the parameters given, as shown in the following:

$$p_D = E_{\theta}(D) - D(E_{\theta}(\theta)) = \bar{D} - D(\theta) .$$

The large value of the effective number of the parameters and the small value of DIC indicate a better fitting model. The sets of the relative risks posterior means were then used to create maps to visualize the high or low risk of disease using the hierarchical Bayesian models. Crude map was also developed from the likelihood model (classical approach) and compared with the maps created via the Bayesian approach.

3. Statistical Analysis

3.1 Interregional Variability Model (IVM)

In the Interregional Variability Model (IVM) we investigate the covariate effect and the internal variability of the relative risk estimation in each region (county) when the variability is not spatially structured i.e. there is only non-spatial heterogeneity. This non-spatial heterogeneity might refer to any social or environmental characteristics which affected the relative risk of the disease. As the relative risk should be positive, a log link function was used as this is usually adopted in the epidemiological studies. The $SMR$ is then defined as follows:

$$\psi_i = SMRE_i = e^{\beta_1 + \beta_2 X_i + \theta_i} ,$$

where $\beta_1$ is the intercept term, $\beta_2$ is the coefficient for the covariate $X_i$ that is expected to have an effect in the study and $\theta$ is the random effect which represents the interregional variability in regions $i = 1, ..., N$.

As in the classical approach, we again assumed that the observed cases in each region follow Poisson distribution (Eq. 1) and the likelihood of $Y_i$ observations is proportional to

$$L(Y_i | \lambda_i) \propto e^{-\lambda_i} \lambda_i^{Y_i} ,$$

where $\lambda_i = \psi_i E_i$. Then, the full model can be written by specifying the prior densities of the parameters, in three stages.
Stage (1): without prior expectations about the direction and magnitude of the covariate effects and conditional on known hyper-parameters, the intercept $\beta_1$ and the covariate coefficient $\beta_2$ were assumed to have vague but informative normal prior, that is:

$$f(\beta_1 | a, b) \propto e^{-(\beta_1 - a)^2/2b},$$
$$f(\beta_2 | r, d) \propto e^{-(\beta_2 - r)^2/2d},$$

(4)

where the hyper-parameters $a, b, r, d$ are fixed known values. The previous research findings showed that the normal prior could give better results when it is used with the Poisson likelihood. This might be due to its conjugacy and the flexibility of computing and integrating the normal density.

Stage (2): conditional on the hyper-parameters $\alpha_1$ and $\alpha_2$, the random effect parameter of the interregional variability was assumed as gamma prior, that is:

$$f(\theta_1 | \alpha_1, \alpha_2) \propto \theta_1^{\alpha_1-1} e^{-\alpha_1 \theta_1},$$

where $\alpha_1 / \alpha_2$ is the prior expectation for $\theta_1$, and $\alpha_1 / \alpha_2$ is the prior variance.

Adopting Poisson likelihood and gamma prior density will produce a negative binomial posterior density which is used to obtain the statistical inferences for the interregional variability. Gamma prior was also chosen as a prior distribution because it is an appropriate density for positive variables, so it is usually used with Poisson likelihood. It is also conjugate prior and flexible in presenting the skewness.

Stage (3): the hyper-parameters $\alpha_1$ and $\alpha_2$ of the gamma density were also given priors. Due to its conjugacy; the most appropriate prior selected was the exponential distribution with constant second order hyper-parameters $c$ and $m$ for $\alpha_1$ and $\alpha_2$, respectively, that is:

$$f(\alpha_1 | c) \propto e^{-c \alpha_1},$$
$$f(\alpha_2 | m) \propto e^{-m \alpha_2}.$$

From these priors and based on the Bayes theory the joint posterior density for IVM can be presented as:

$$p(\theta, \beta_1, \beta_2, \alpha_1, \alpha_2 | Y_i, \Phi) \propto \prod_{i=1}^{n} L(Y_i | \theta, \beta_1, \beta_2) p(\theta | \alpha_1, \alpha_2) p(\beta_1 | a, b) p(\beta_2 | r, d) p(\alpha_1 | c) p(\alpha_2 | m),$$

where $\Phi = a, b, r, d, c, m$ represents the fixed hyper-parameters and the second order hyper-parameters of the prior distributions. The posterior inferences for the parameters of the model were obtained by simulating the joint posterior distribution using the Gibbs sampling within the Markov Chain Monte Carlo (MCMC) method.
3.2 Global Spatial Model (GSM)

In this model, the variables of interest were considered to be spatially auto-correlated with respect to their geographical locations. The model was constructed in such a way that it included both over-dispersion and covariate effects, enabling it to establish a possible association between all the expected risk factors and disease incidences. The Global Spatial Model was defined as:

$$
\hat{\psi}_i = SMR_i = e^{\theta_i + \beta_1 X_i + \beta_2 Y_i + \phi},
$$

where $\beta_1, \beta_2$ and $\theta_i$ defined as before, and $\phi$ is a random effect that reflects the spatial variability.

As in IVM, the model assumes that conditional on the underlying relative risk, the observed cases in each region follow a Poisson distribution. Hence, the likelihood was represented as in equation (3). The prior implementation in this model was structured so as to accommodate the covariate, interregional and spatial heterogeneity effects. The prior densities for the parameters and hyper-parameters are specified as follows.

Stage (1): conditional on the hyper-parameter $\tau_\theta$, which controls the variability that is not due to the spatial structure, the random effect parameter of the interregional variability $\theta_i$ was assumed as a conjugate normal independent prior that is

$$
f(\theta_i|\tau_\theta) = \sqrt{\frac{\tau_\theta}{2\pi}} \exp\left(-\frac{\tau_\theta \theta_i^2}{2}\right),
$$

where $\tau_\theta = 1/\sigma^2_\theta$ is the interregional precision term.

Stage (2): conditional on the hyper-parameter $\tau_\phi$, which controls the variability of $\phi$, the random effect parameter of the spatial variability $\phi_i$ was assumed as Intrinsic Conditional Autoregressive (ICAR) model, where $\phi_i$ has a normal distribution with conditional mean $\bar{\phi}_i$ and conditional variance, that is

$$
\phi_i | \theta_i \sim N(\bar{\phi}_i, \frac{1}{\tau_\phi \omega_i}),
$$

where $\tau_\phi = 1/\sigma^2_\phi$ is the spatial precision term, $\bar{\phi}_i = \sum \phi_{i',j}/\omega_i$ denotes the average of the neighbouring $\phi_{i',j}$ that is adjacent to $\phi_i$ and $\omega_i$ is the number of the adjacent counties. The most common density of this prior [16] has the joint distribution proportional to

$$
\tau_\phi^{\omega_i-2} \exp\left[-\frac{\tau_\phi}{2} \sum \phi_i (\phi_i - \bar{\phi}_i)^2\right] \propto \tau_\phi^{\omega_i-2} \exp\left[-\frac{\tau_\phi}{2} \sum \omega_i (\phi_i - \bar{\phi}_i)\right].
$$

Stage (3): due to the absence of the prior information, the vague flat prior distribution was used for the intercept term $\beta_i$ to give the desired properties of the improper ICAR prior which was used for the spatial random effect. The covariate coefficient $\beta_2$ was assumed to have a normal prior distribution as in equation (4) with fixed hyper-parameters $r$ and $d$. 

Stage (4): priors were chosen for the hyper-parameters \( \tau_\theta \) and \( \tau_\phi \) of the interregional and spatial variability. The most appropriate priors were the vague Gamma distribution with constant second order hyper-parameters \( k_1, k_2 \) for \( \tau_\theta \) and second order hyper-parameters \( l_1, l_2 \) for the spatial precision term \( \tau_\phi \) as shown below,

\[
f(\tau_\theta \mid k_1, k_2) \propto \tau_\theta^{k_1-1} e^{-k_2 \tau_\theta},
\]
\[
f(\tau_\phi \mid l_1, l_2) \propto \tau_\phi^{l_1-1} e^{-l_2 \tau_\phi}.
\]

These stages produce the following joint posterior density, which is the product of the Poisson likelihood and the prior densities,

\[
p(\theta, \phi, \beta_1, \beta_2, \tau_\theta, \tau_\phi \mid Y, \Phi) \propto \prod_{i=1}^{N} \frac{\Gamma(Y_i) \phi^Y_i e^{-\phi_i}}{Y_i!} \frac{1}{\Gamma(\tau_\theta) \Gamma(\tau_\phi)} \frac{1}{\Gamma(k_1) \Gamma(k_2) \Gamma(l_1) \Gamma(l_2)} \cdot
\]

where \( \Phi = r, d, k_1, k_2, l_1, l_2 \) are constant hyper-parameters.

### 3.3 Applications of the Classical Approach and the Hierarchical Models to Lip cancer Data

The classical approach and the two hierarchical Bayesian models were applied to the lip cancer disease in 56 counties in Scotland over the period 1973 to 1980. The data have been originally analyzed by Clayton and Kaldor [6] and Breslow and Clayton [19]. The form of the data includes the name and the number of the 56 counties, the observed number of male lip cancer cases, population size of males, a covariate measuring the percentage of the county’s population engaged in agriculture, forestry and fishing (AFF), and the position of each county expressed as a list of adjacent counties.

Morans I spatial correlation coefficient was calculated for the data. The value of z score test was 1.982, with p-value 0.0474 indicating that the coefficient is significant at 0.05 level. Based on the results, it can be concluded that there is a spatial autocorrelation among the counties.

The value of the estimated relative risk was obtained directly from equation (2) in the classical approach while Markov Chain Monte Carlo (MCMC) method was used to obtain the joint and the conditional posterior densities of the parameters via the WinBUGS software. The inferences of the relative risk were then obtained from the posterior densities of each model in the hierarchical approach.

For each model different sets of initial values were used. Rapid convergence to the posterior distribution was observed using two chains within one thousand iterations followed by two thousand iterations. The first five hundred iterations were discarded from the chains as pre-convergence burn in. The convergence was assessed by checking the trace plots and the Gelman-Rubin convergence statistic [2] of the samples.
4. Results
The estimation of the parameters was obtained from the hierarchical Bayesian models and given in Table 1. The table shows the values of the intercept, the covariate coefficient, the standard deviation of the interregional variability $\sigma_\theta$ and the standard deviation of the spatial variability $\sigma_\phi$ for each model. Fitting the hierarchical models to the data shows that there exists variability which was not explained by the classical approach. Fitting the GSM shows that large amount of this variability ($\sigma_\phi = 0.546$) appears to be from the spatial term. Table 1 also shows that there is a positive relationship ($\beta_2 = 0.062$ and $\beta_2 = 0.050$ using the IVM and the GSM respectively) between the AFF population and the disease relative risk. Such relation could not be assessed without using the Bayesian approach.

<table>
<thead>
<tr>
<th>model</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\sigma_\theta$</th>
<th>$\sigma_\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>IVM</td>
<td>-1.118</td>
<td>0.0625</td>
<td>0.527</td>
<td>-</td>
</tr>
<tr>
<td>GSM</td>
<td>-0.338</td>
<td>0.0501</td>
<td>0.320</td>
<td>0.546</td>
</tr>
</tbody>
</table>

Table 1 Parameter Estimation Using the Hierarchical Models

<table>
<thead>
<tr>
<th>District name</th>
<th>classical</th>
<th>IVM</th>
<th>GSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Skye-Lochalsh</td>
<td>6.52</td>
<td>5.393</td>
<td>5.006</td>
</tr>
<tr>
<td>8 Shetland</td>
<td>3.04</td>
<td>2.379</td>
<td>1.511</td>
</tr>
<tr>
<td>11 Western Isles</td>
<td>2.95</td>
<td>2.648</td>
<td>1.725</td>
</tr>
<tr>
<td>12 Sutherland</td>
<td>2.79</td>
<td>2.234</td>
<td>3.344</td>
</tr>
<tr>
<td>13 Nairn</td>
<td>2.78</td>
<td>1.836</td>
<td>2.494</td>
</tr>
<tr>
<td>16 Kincardine</td>
<td>1.98</td>
<td>1.696</td>
<td>2.131</td>
</tr>
<tr>
<td>17 Badenoch</td>
<td>1.87</td>
<td>1.36</td>
<td>2.026</td>
</tr>
<tr>
<td>55 Annandale</td>
<td>0</td>
<td>1.081</td>
<td>0.7851</td>
</tr>
</tbody>
</table>

Table 2 SMR Estimation for Selected Counties

Different values of estimated relative risk were observed when using the classical approach, the IVM and the GSM. The results summarized in Table 2 show the effect of including the covariate and the over-dispersion in the estimated SMR values for selected counties.

A comparison between the hierarchical Bayesian models (IVM and GSM) was carried out. The selection was done based on the DIC values computed for
each model, using the MCMC method. To see the impact of the random term in the mix effect models, the analysis also included the result of fitting a fixed effect model that contained the covariate term only. Table 3 lists the deviance summaries for each model.

The result of computing the DIC shows that the mix effect models (IVM and GSM) are much better than the fixed effect model which has the largest DIC value among all models. This finding suggests that a random effect term is needed to represent the over-dispersion in the relative risk among the different counties. Using more than 1000 iterations, the model with spatial and interregional random effects (GSM) had better average fit to the data, $D = 272.846$, and best fitting point estimate $D(\bar{\Theta}) = 238.999$ than the model with the interregional variability (IVM). Among the three models GSM had the smallest DIC value (306.692) and extra number of effective parameters, $p_D = 33.846$. As a result, it is suggested that this is the most robust and appropriate model to be used.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\bar{D}$</th>
<th>$D(\bar{Theta})$</th>
<th>$p_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Model</td>
<td>454.700</td>
<td>452.983</td>
<td>1.717</td>
<td>456.417</td>
</tr>
<tr>
<td>IVM</td>
<td>320.330</td>
<td>293.839</td>
<td>26.491</td>
<td>346.821</td>
</tr>
<tr>
<td>GSM</td>
<td>272.846</td>
<td>238.999</td>
<td>33.84</td>
<td>306.692</td>
</tr>
</tbody>
</table>

Table 3 Deviance Summaries for Hierarchical Bayesian Models. The table shows the posterior mean of the deviance $\bar{D}$, the deviance evaluated at the posterior mean of the parameters $D(\bar{Theta})$, the effective number of the parameters $p_D$ and the DIC.

This can be clearly seen in Fig. 1 and Fig. 2. The maps in the figures were created using the ArcGIS software. In Fig. 1 the map shows that the risks of the disease tend to randomly cluster (areas of high and low risk due to over-dispersion) when using the classical approach (see map of crude SMR). The fact which suggests that there are unexplained factors which might affect the risks associated with the disease. The inconsistency and clustering trend strongly suggest the need for statistical smoothing to overcome these disadvantages.

Fig. 2 introduces the maps of the posterior SMR estimation for the lip cancer data in the 56 counties of Scotland, using the IVM and the GSM results, respectively. The maps show the impact of including the covariate term and the interregional (unstructured) heterogeneity via IVM (Fig. 2a) and the impact of
Fig. 1 Crude SMR via Classical Approach

Fig. 2 SMR Estimation Using Hierarchical Bayesian Approach:
(a) Estimated SMR via IVM
(b) Estimated SMR via GSM.
including the covariate term and interregional heterogeneity together with the spatial random effect via GSM (Fig 2b).

In Figure 2a the range of the posterior relative risk has reduced in some counties. The random variability, due to the small count, was also removed from the data; thus, no county is now assigned a value of exactly zero, as shown in county 55 (see also Table 2). However, there are still small counties with extremely high rates, as in counties 4 and 14.

Similarly, a trend of random disease clustering can also be identified when using the IVM. This clustering trend suggests that there is still variability existing, which might be due to the spatial correlation among the neighbouring counties.

In Figure 2b, the GSM map shows the pattern of the local spatial clustering of similar values in the adjacent neighbours. The map showed less variation as compared to the map of the crude SMR. Some extreme SMR estimates have disappeared and much of the map has been smoothed indicating that the GSM shows fewer extremes in the relative risk estimates.

5. Conclusions

This article provides a discussion on the analysis of the disease incidences. The purpose of this article was to use a hierarchical Bayesian approach to investigate different sources that might affect the relative risk estimation. Two types of hierarchical Bayesian models were introduced. These models allow for the covariate(s) and the over-dispersion of the relative risk to be investigated. The covariate(s) effect and the over-dispersion which in turn classified as interregional and spatial effects were the main factors that cause the relative risk heterogeneity.

The relative risk inferences for the hierarchical models were obtained from the posterior densities using the MCMC method, while the relative risk estimation using the classical approach was obtained via the maximum likelihood method. We discussed the results of fitting both approaches (classical and Bayesian) to the Scotland lip cancer data, after taking into consideration the spatial autocorrelation.

The result gathered from the maximum likelihood estimation showed that the SMRs in Scotland counties were found to vary around their overall mean. Compared to the classical approach, large amounts of variability appeared when using the hierarchical models. Moran’s I test statistic showed that the spatial autocorrelation exists among the counties. Introducing the covariate and the interregional terms allow for amount of the variability to be explained using the IVM. Using ICAR as prior density in the GSM showed that the data had a large amount of spatial heterogeneity, which did not appear in the IVM.

The study was also concern about the most appropriate model which could fit the data well. This was achieved by conducting the goodness of fit test via the DIC. Three hierarchical models that included/excluded the two random effects were compared. The model with fix and both random effects (GSM) was selected as the best fitted model.
Disease maps were conducted in this study to carry out the exploratory investigations on the spread of the disease. For this purpose, the output obtained from the classical approach was compared with those obtained from the Bayesian models using the GIS tools. Clustering trend in the disease map was observed when using the classical approach. Using the IVM, it explains amount of the clustering trend which occurred due to the interregional variability effects as shown in Figure 2a, while map of smooth relative risk was obtained using the GSM (Fig. 2b) concluding that this model performs and fit the data well.

References


Received: March, 2012