International Journal of Probability and Statistics 2014, 3(1): 15-22
DOI: 10.5923/j.ijps.20140301.03

Small Area Estimation with Application to Disease Mapping

E. P. Clement

Department of Mathematics and Statistic University of Uyo, Uyo, Nigeria

Abstract Small Area Estimation is important in survey analysis when domain (subpopulation) sample sizes are too small to provide adequate precision for direct domain estimators. Small Area Estimation (SAE) is a mathematical technique for extracting more detailed information from existing data sources by statistical modeling. The estimates are often mapped, so the technique is often generically called mapping. These maps and estimates (together with estimates of accuracy) are key information in aid allocation within a country. They are also increasingly important inputs to negotiations on allocation of international aid to particular countries. This paper provides a critical review of the main advances in small area estimation (SAE) methods in recent years with application to disease mapping. The review discusses in detail earlier developments of small area estimation methods in the field of disease mapping which serve as a necessary background for the new studies in disease mapping of small areas which we termed "Extensions". Illustrative examples of the application of Small Area Estimation (SAE) to disease mapping are presented.

Keywords Disease Mapping, Empirical Bayes (EB), Hierarchical Bayes (HB), Relative Risk (RR), Small Area Estimation (SAE), Standardized Mortality Ratios (SMRs)

1. Introduction

As with the analysis of any data set, it is always good practice to begin by producing and inspecting graphs. A feel for the data can then be obtained and any outstanding features identified. In spatial epidemiology this is called disease mapping.

Disease mapping is considered as exploratory analysis used to get an impression of the geographical or spatial distribution of disease or the corresponding risk. Disease mapping is an epidemiological technique used to describe the geographic variation of disease and to generate etiological hypotheses about the possible causes for apparent differences in risks. A disease map is used for reporting the results of a geographical correlation study or to highlight geographical areas with high and low incidence, prevalence and mortality rates of specific disease and the variability of such rates over a spatial domain (small area). They can also be used to detect spatial clusters which may be due to common environmental, demographical, or cultural effects shared by neighbouring regions. The correct geographical allocation of health care resources would be greatly enhanced by the development of statistical models that allow a more accurate depiction of "true" disease occurrence and prevalence.

The Millennium Development Goals (MDGs) provide a context for small area estimation, since local estimates (small area estimates) of disease rates and their updates have potential to provide fine-detailed national monitoring information against which progress can be measured. Small Area Estimation is a statistical technique involving the estimation of parameters for small sub-populations (small areas) where a sample has insufficient or no sample for the sub-population (small area) to be able to make accurate estimates for them. The term "small area" may refer strictly to a small geographical area such as a county, but may also refer to a "small domain", that is, a particular demographic within an area. Small area estimation methods use models and additional data sources (such as census data) that exist for these small areas in order to improve estimates' for them.

Small area estimation is important in survey analysis when domain (sub-population or small area) sample sizes are too small to provide adequate precision for direct domain estimators. It is a mathematical technique for extracting more detailed information from existing data sources by statistical modeling. The estimates are often mapped to obtain and identify any outstanding features, so the technique is often generically called mapping. These maps and estimates (together with estimates of accuracy) are key information in aid allocation within a country. They are also increasingly important inputs to negotiations on allocation of international aid to particular countries by the World Health Organization (WHO) and other International Aid Agencies

* Corresponding author:
epclement@yahoo.com (E. P. Clement)
Published online at http://journal.sapub.org/ijps

(IAA).

The purpose of this paper is to provide a critical review of the main advances in small area estimation with application to Health surveys in general and disease mapping in particular. Four statistical models: Poisson-gamma, log-normal, conditional autoregression normal (CAR-normal) and two-level models are discussed. The empirical bayes (EB) and the hierarchical bayes (HB) models are also discussed as extensions to the four basic models. The review discusses in detail earlier developments of small area estimation (SAE) methods in the field of disease mapping which serve as a necessary background for the various extensions of the disease mapping models proposed in recent literature. Illustrative examples of studies so far proposed are presented. The paper ends with a brief summary and some concluding remarks.

2. Application of Small Area Estimation to Health

Small area estimation of health related characteristics has attracted a lot of attention in the Western countries like the U.S, Britain, U. K and Canada because of a continuing need to assess health status, health practices and health resources at both the national and sub-national levels. Reliable estimates of health-related characteristics help in evaluating the demand for health care and the access individuals have to it. Health care planning often takes place at the state and sub-state levels because health characteristics are known to vary geographically.

Health System Agencies in the United States, mandated by the National Health Planning Resource Development Act of 1994, are required to collect and analyse data related to the health status of the residents and to the health delivery systems in their health service area [1].

The U.S. National Centre for Health Statistics pioneered the use of synthetic estimation, based on implicit linking models, developing state estimates of disability and other health characteristics for different groups from the National Health Interview Survey (NHIS) [2]. [3] studied HB estimation of overweight prevalence for adults by states, using data from NHANES III, [4] produced survey-weighted HB estimates of small area prevalence rates for states and age groups, for up to 20 binary variables related to drug use, using data from pooled National Household Surveys on Drug Abuse. [5] studied EB estimates of state-wide prevalence of the use of alcohol and drugs (e.g. Marijuana) among civilian non-institutionalized adults and adolescents in the United States. These estimates are used for planning and resource allocation, and to project the treatment needs of dependent users.

Direct (or crude) estimates of rates, called standardized mortality ratios (SMRs) can be very unreliable, and a map of crude rates can badly distort the geographical distribution of disease incidence or mortality because the map tends to be dominated by areas of low population. Disease mapping, using model-based estimators, has received increased attention in recent years. Typically, sampling is not involved in disease mapping applications.

3. Mortality and Disease Rates Models

Mortality and disease rates of small area in a region or a county are often used to construct disease maps such as cancer atlases. Such maps are used to display geographical variability of a disease and identify high-rate areas warranting intervention. A simple small area model is obtained by assuming that the observed small area counts y_i are independent Poisson variables with conditional mean

$$E(y_i|\lambda_i) = n_i\lambda_i \tag{1}$$

and that $\lambda_i \sim_{iid} gamma\,(\alpha,\tau)$.

where λ_i is the true rate, n_i is the number exposed in the ith area, and (α,τ) are the scale and shape parameters of the gamma distribution under this model, smoothed estimates of λ_i are obtained using EB or HB methods[6],[7].

If A_i denotes a set of "neighbouring" areas of the ith area, then a conditional autoregression (CAR) spatial model assumes that the conditional distribution of $b_i v_i$ given $(v_l : l \neq i)$ is given by

$$b_i v_i|\{v_l: l \neq i\} \sim N\left(\rho \textstyle\sum_{l\epsilon A_i} q_{il} b_l v_l\,, b_i^2\sigma_v^2\right) \tag{2}$$

where $\{q_{il}\}$ are known constants satisfying $q_{il}b_\ell^2 = q_{li}b_i^2\ (i<l)$, and $\delta = (\rho,\sigma_v^2)^T$ is the unknown parameter vector.

CAR spatial models of the form (2) on log rates

$$\theta_i = log(\lambda_i) \tag{3}$$

have also been proposed by[6].

The model on λ_i can be extended to incorporate area level covariates z_i, for example:

$$\theta_i = z_i^T\beta + v_i \quad \text{with} \quad v_i \sim_{iid} N(O,\sigma_v^2) \tag{4}$$

[8] studied regression models on age-specific log rates

$$\theta_{ij} = log\lambda_{ij} \tag{5}$$

involving random slopes, where j denotes age.

Joint mortality rates (y_{1i}, y_{2i}) can also be modeled if (y_{1i}, y_{2i}) are asummed independently distributed conditional on $(\lambda_{1i}, \lambda_{2i})$ and

$$\theta_i = (log\lambda_{1i}, log\lambda_{2i})^T \sim_{iid} N_2(\mu,\Sigma) \tag{6}$$

Further, $y_{1i}\ and\ y_{2i}$ are assumed to be conditionally independent Poisson variables with

$$\mathrm{E}(y_{1i}|\,\lambda_{1i}) = n_{1i}\lambda_{1i} \tag{7}$$

and $\mathrm{E}(y_{2i}|\,\lambda_{2i}) = n_{2i}\lambda_{2i}$

where y_{1i} and n_{1i} denote the number of deaths due to a disease (say malaria) and the population at risk at site (or area) 1 respectively and y_{2i} and n_{2i} denote the number of deaths due to the same disease and the population at risk at site (or area) 2 respectively. [9] showed that the bivariate model leads to improved estimates of the rate $(\lambda_{1i}, \lambda_{2i})$ compared to estimates based on separate univariate models.

4. Disease Mapping

Mapping of small area mortality (or incidence) rates of disease such as cancer, malaria is a widely used tool in Public Health research. Such maps permit the analysis of geographical variation in rates of diseases which may be useful in formulating and assessing etiological hypotheses, resource allocation, and identifying areas of unusually high risk warranting intervention.

The following are examples of studies of disease rates in the literature: [6] studied lip cancer rates in the 56 counties (small areas) of Scotland; [7] studied the incidence of leukemia in 281 census tracts (small areas) of upstate, New York. [8] studied all cancer mortality rates for white males in health service areas (small areas) of the United States; [10] studied prostrate cancer rates in Scottish counties; and [11] studied infant mortality rates for local health areas (small areas) in British Columbia, Canada.

Worthy of note is the fact that sampling is not used in disease mapping only administrative data on event counts and related auxiliary variables are used in disease mapping.

4.1. Disease Mapping Models

Suppose that the country (or the region) used for disease mapping is divided into m non-overlapping small areas. Let θ_i be the unknown relative risk (RR) in the ith area. A direct (or crude) estimator of θ_i is given by the standardized mortality ratio (SMR)

$$\widehat{\theta_i} = y_i/e_i \tag{8}$$

where y_i and e_i denote the observed and expected number of deaths (cases) over a given period $(say\ i = 1,2,\dots,m)$ respectively.

Where:

$$e_i = n_i(\textstyle\sum_i y_i/\sum_i n_i) \tag{9}$$

where is the number of person – years at risk in the ith area, and then treated as fixed. Some authors use mortality (event) rates τ_i as parameters instead of relative risks, and a crude estimator of τ_i is then given by $\hat{\tau}_i = y_i/n_i$. However, the two approaches are equivalent since $\sum_i y_i/\sum_i n_i$ is treated as a constant.

A common assumption in disease mapping is that $y_i\,|\theta_i \sim_{iid}$ Poisson $(e_i\theta_i)$. Under this assumption, the maximum likelihood (ML) estimator of θ_i is the SMR, $\widehat{\theta_i} = y_i/e_i$

However, a map of crude rates $\{\widehat{\theta_i}\}$ can badly distort the geographical distribution of disease incidence or mortality because it tends to be dominated by areas of low population, e_i exhibiting extreme SMR's that are least reliable.

$$Var\ (\widehat{\theta_i}) = \theta_i/e_i \tag{10}$$

is large if is small.

Empirical Bayes (EB) or hierarchal bayes (HB) methods provide reliable estimators of relative risk (RR) by borrowing strength across areas. As a result, maps based on empirical Bayes (EB) or hierarchical bayes (HB) estimates are more reliable compared to crude maps. We will give account of empirical Bayes (EB) and hierarchical bayes (HB) methods for each of the disease model discussed based on simple linking models.

4.1.1. Poisson-Gamma Model

Given a two-stage model, at the first stage, assume $y_i \sim_{ind}$ Poisson $(e_i\theta_i)$, i = 1, 2,…, m. A conjugate model linking the relative risks θ_i is assumed in the second stage: $\theta_i \sim_{iid}$ gamma (ν,α) denotes the gamma distribution with shape parameter $\nu(>0)$ and scale parameter $\alpha(>0)$. Then we have

$$\mathrm{f}(\theta_i \mid \alpha,\nu) = \frac{\alpha^{\nu}}{\Gamma(\upsilon)}\, e^{-\alpha\theta_i}\ \theta_i^{\nu-1} \tag{11}$$

and

$$E(\theta_i) = \upsilon/\alpha = \mu, \quad Var\ (\theta_i) = \nu/\alpha^2 \tag{12}$$

where:

$\theta_i|y_i,\alpha,\nu \sim_{ind} gamma(y_i+\nu, e_i+\alpha)$, the bayes estimators of θ_i and posterior variance of θ_i are obtained from (3) by changing α to $e_i+\alpha$ and ν to $y_i+\nu$ such that:

$$\hat{\theta}_i^B(\alpha,\nu) = E(\theta_i \mid y_i,\alpha,\nu) = (y_i+\nu)/(e_i+\alpha) \tag{13}$$

and

$$Var\ (\theta_i \mid y_i,\alpha,\nu) = g_{1i}(\alpha,\nu,y_i) = (y_i+\nu)/(e_i+\alpha)^2 \tag{14}$$

The maximum likelihood (ML) estimator of α and ν from the marginal distribution, $y_i \mid \alpha,\nu \sim_{iid}$ negative binomial, using the log likelihood is:

$$l(a,v) = \sum_{i=1}^{m}\left[\sum_{k=0}^{y_i-1}\log(v+k) + v\log(a)\right] - (y_i+v)\log(e_i+a) \tag{15}$$

Closed form expressions for $\hat{\alpha}_{ML}$ and $\hat{\nu}_{ML}$ do not exist. However, [12] obtained simple moment estimators by equating the weighted sample mean $\hat{\theta}_{e.} = \frac{1}{m}\sum_i(e_i/e.)\,\hat{\theta}_i$ and the weighted sample variance $S_e^2 = \frac{1}{m}\sum_i(e_i/e.)\left(\hat{\theta}_i - \theta e.2\right.$ to their expected values and then solving the resulting moment equations for α and ν, where $e. = \sum_i(e_i/m)$. This leads to moment estimators, $\hat{\alpha}$ and $\hat{\nu}$, given by:

$$\frac{\hat{\nu}}{\hat{\alpha}} = \hat{\theta}_{e.} \tag{16}$$

$$\frac{\hat{\nu}}{\hat{\alpha}^2} = s_e^2 - (\hat{\theta}_{e.}/e.) \tag{17}$$

[13] provided more efficient movement estimators. The moment estimators may also be used as starting values for maximum likelihood (ML) iterations.

Substituting the moment estimators $\hat{\alpha}$ and $\hat{\nu}$ into (13) we get the empirical Bayes (EB) estimator of θ_i as

$$\hat{\theta}_i^{EB} = \hat{\theta}_i^B(\hat{\alpha},\hat{\nu}) = \hat{\gamma}_i\widehat{\theta_i} + (1-\hat{\gamma}_i)\hat{\theta}_{e.}\ , \tag{18}$$

where $\hat{\gamma}_i = e_i/(e_i+\hat{\alpha})$. It should be noted that $\hat{\theta}_i^{EB}$ is a weighted average of the direct estimator (SMR) $\widehat{\theta_i}$ and the synthetic estimator $\hat{\theta}_{e.}$, and more weight is given to $\widehat{\theta_i}$ as the area expected deaths, e_i, increase. If $S_e^2 < \left(\hat{\theta}_{e.}/e.\right)$ then $\hat{\theta}_i^{EB}$ is taken as the synthetic estimator $\hat{\theta}_{e.}$. The empirical bayes (EB) estimator is nearly unbiased for θ_i in the sense that its bias is of order m^{-1}, for large m.

The Jackknife method may be used to obtain a nearly unbiased estimator of MSE $\left(\hat{\theta}_i^{EB}\right)$. The jackknife estimator is given by

$$MSE_J\left(\hat{\theta}_i^{EB}\right) = \hat{m}_{1i} + \hat{m}_{2i} \quad (19)$$

where

$$\hat{m}_{1i} = g_{1i(\hat{\alpha},\hat{v},y_i)} - \frac{m-1}{m}\sum_{l=1}^{m}[g_{1i}(\hat{\alpha}_{-l},\hat{v}_{-l},y_i) - g_{1i}(\hat{\alpha},\hat{v},y_i)]$$

$$\hat{m}_{2i} = \frac{m-1}{m}\sum_{l=1}^{m}\left(\hat{\theta}_{i,-l}^{EB} - \hat{\theta}_i^{EB}\right)^2$$

$$\hat{\theta}_i^{EB} = k_i(y_i,\hat{\alpha},\hat{v}\ and\ \hat{\theta}_{i,-l}^{EB} = k_i(y_i,\hat{\alpha}_{-l},\hat{v}_{-l})$$

where $\hat{\alpha}_{-l}\ and\ \hat{v}_{-l}$ are the delete $-l$ moment estimators obtained from $\{(y_i,e_i), i \neq 1,2,\dots m\}$. Note that $MSE_J\left(\hat{\theta}_i^{EB}\right)$ is area-specific in the sense that it depends on y_i. [13] obtained a Taylor expansion estimator of MSE, using a parametric bootstrap to estimate the covariance matrix of $(\hat{\alpha},\hat{v})$.

The linking gamma model on the $\theta_i's$ can be extended to allow for area -level covariates z_i, such as degree of relative risk (RR). [6] allowed varying scale parameters, α_i, and assumed a loglinear model on

$$E(\theta_i) = v/\alpha_i : log\left(E(\theta_i)\right) = z_i^T\beta \quad (20)$$

Empirical bayes (EB) estimators for this extension are given by:

$$\hat{\theta}_i^{EB}(\alpha,\beta) = E[\theta_i|y_i,\alpha,\beta] = [(y_i+\beta)/(e_i+\alpha_i)] \quad (21)$$

and

$$Var(\theta_i|y_i,\alpha_i,\beta) = g_{1i}(\alpha_i,\beta,y_i) = (y_i+\beta)/(e_i+\alpha_i)^2 \quad (22)$$

[14] studied the Poisson-gamma regression model in detail and proposed accurate approximations to the posterior mean and the posterior variance of θ_i. The posterior mean approximation is used as the empirical bayes (EB) estimator and the posterior variance approximation as a measure of its variability.

(a) Hierarchical Bayes (HB) Estimation

Let $\theta_i, y_i and\ e_i$ respectively denote the relative risk (RR), observed and expected number of cases (deaths) over a given period in the ith area $(i = 1,2,\dots m)$. A hierarchical bayes (HB) estimation of the Poisson-gamma model, is given by:

$$\begin{aligned} &\text{(i) } y_i|\theta_i \sim_{ind} Poisson\ (e_i\theta_i)\\ &\text{(ii) } \theta_i|\alpha,v \sim_{iid} G(v,\alpha)\\ &\text{(iii) } f(\alpha,v) \propto f(\alpha)f(v) \end{aligned} \quad (23)$$

with $f(\alpha) \propto \frac{1}{\alpha}; v \sim G\left(a=\frac{1}{2},b\right) b > 0$

See [7].

The joint posterior $f(\theta,\alpha,v|y)$ is proper if at least one y_i is greater than zero. The Gibbs conditionals are given by:

$$\begin{aligned} &\text{(i) } [\theta_i|\alpha,v,y] \sim_{ind} G(y_i+v,e_i+\alpha)\\ &\text{(ii) } [\alpha|\theta,v,y] \sim G(mv,\textstyle\sum_i\theta_i)\\ &\text{(iii) } f(v|\theta,\alpha,y) \propto (\textstyle\prod\theta_i^{v-1})^i\ exp(-bv)\alpha^{vm}/\Gamma^m(v) \end{aligned} \quad (24)$$

Monte Carlo Markov Chain (MCMC) samples can be generated directly from (i) and (ii) of (24), but we need to use the Metropolis-Hastings (M-H) algorithm to generate samples from (iii) of (24). Using the Monte Carlo Markov Chain (MCMC) samples $\left\{\theta_i^{(k)},\dots,\theta_m^{(k)},v^{(k)},\alpha^k; k = d+1,\dots,d+D\right\}$, posterior quantities of interest may be computed, in particular, the posterior mean $E[\theta_i|y]$ and posterior variance $Var(\theta_i|y)$ for each area $i = 1,2,\dots,m$.

4.1.2. Log-Normal Model

Log-normal two-stage models have also been proposed. The first-stage model is not changed, but the second-stage linking model is changed to $\xi_i = log(\theta_i) \sim_{iid} N(\mu,\sigma^2), i = 1,2,\dots,m$.

As in the case of logit-normal models, implementation of empirical bayes (EB) is more complicated for the log-normal model because no closed-form expression for the bayes estimator, $\hat{\theta}_i^B(\mu,\sigma^2)$, and the posterior variance, $Var(\theta_i|y_i,\mu,\sigma^2)$ exist. [6] approximated the posterior density, $f(\xi|y,\mu,\sigma^2)$, by a multivariate normal which gives an explicit approximation to $\hat{\xi}_i^B$, where $\xi = (\xi_i,\dots,\xi_m)^T$ and $y = (y_1,\dots,y_m)^T$. Maximum likelihood estimators of model parameters μ and σ^2 were obtained using the EM algorithm and then used in the approximate formula for $\hat{\xi}_i^B$ to get EB estimators $\hat{\xi}_i^{EB}$ of ξ_i and $\hat{\theta}_i^{EB} = exp(\hat{\xi}_i^{EB})$ of θ_i.

The empirical bayes (EB) estimator $\hat{\theta}_i^{EB}$, however, is not nearly unbiased for θ_i. Moment estimators of μ and σ proposed by [15] may be used to simplify the calculation of Jackknife estimator of $MSE\left(\hat{\theta}_i^{EB}\right)$.

The basic log-normal model readily extends to the case of covariates z_i: $\xi_i = log\theta_i \sim_{ind} N(z_i^T\beta,\sigma^2)$.

(a) Hierarchical Bayes Estimation

A hierarchical bayes (HB) estimator of the basic log-normal model is given by:

$$\begin{aligned} &\text{(i) } y_i|\theta_i \sim_{ind} Poisson\ (e_i\theta_i)\\ &\text{(ii) } \xi_i = log(\theta_i)|\mu,\sigma^2 \sim_{iid} N(\mu,\sigma^2)\\ &\text{(iii) } f(\mu,\sigma^2) \propto f(\mu)f(\sigma^2) \end{aligned} \quad (25)$$

with $f(\mu) \propto 1; \sigma^{-2} \sim G(a,b); a \geq 0, b > 0$

The joint posterior $f(\theta,\mu,\sigma^2|y)$ is proper if at least one y_i is greater than zero, it is easy according to[2] to verify that the Gibbs conditionals are given by:

$$\begin{aligned} &\text{(i) } f(\theta_i|\mu,\sigma^2,y) \propto \theta_i^{y_i-1}exp\left[-e_i\theta_i - \tfrac{1}{2\sigma^2}(\xi_i-\mu)^2\right]\\ &\text{(ii) } [\mu|\theta,\sigma^2,y] \sim N\left(\tfrac{1}{m}\textstyle\sum_i\xi_i, \tfrac{\sigma^2}{m}\right)\\ &\text{(iii) } [\sigma^2|\theta,\mu,y] \sim G(\tfrac{m}{2}+a, \tfrac{1}{2}\textstyle\sum_i(\xi_i-\mu)^2+b) \end{aligned} \quad (26)$$

See [16]

Monte Carlo Markov Chain (MCMC) samples can be generated directly from (ii) and (iii) of (26), but we need to use Metropolis-Hastings (M-H) algorithm to generate samples from (i) of (26). We can express (i) as:

$$f(\theta_i|\mu,\sigma^2,y) \propto k(\theta_i)h(\theta_i|\mu,\sigma^2)$$

Where $k(\theta_i) = exp(-e_i\theta_i)\theta_i^{y_i}$

and $h(\theta_i|\mu,\sigma^2) \propto g^1(\theta_i)exp\{-(\xi_i-\mu)^2/2\sigma^2\}$

with $g^1(\theta_i) = \partial g(\theta_i)/\partial\theta_i$ and $g(\theta_i) = log(\theta_i)$

We can use $k(\theta_i|\mu,\sigma^2)$ to draw the candidate, θ_i^*, noting that $\theta_i = g^{-1}(\xi_i)$ and $\xi_i|\mu,\sigma^2 \sim N(\mu,\sigma^2)$. The acceptance probability used in the M-H algorithm is then given by $a\left(\theta^{(k)},\theta_i^*\right) = min\left\{k(\theta_i^*)/k\left(\theta_i^{(k)},1\right)\right\}$. The basic log-normal model with Poisson counts y_i as noted earlier, readily extends to the case of covariates z_i where (ii) and (iii) of (25) become respectively:

$$\text{(i) } \xi_i = \xi_i|\beta,\sigma^2 \sim N(Z_i^T\beta,\sigma^2)$$
$$\text{(ii) } f(\beta,\sigma^2) \propto f(\beta)f(\sigma^2) \quad (27)$$

With $f(\beta) \propto 1, and\ \sigma^{-2} \sim G(a,b)$

4.1.3. Car-Normal Model

The basic log-normal can be extended to allow spatial correlations; mortality data sets often exhibit significant spatial relationship between the log relative risks, $\xi_i = log(\theta_i)$. A simple conditional autoregression (CAR)-normal model on ξ assumes that ξ is a multivariate normal specified by:

$$E(\xi_i|\xi_l, l \neq i) = \mu + \rho\sum_{l(\neq i)} q_{il}\,(\xi_l - \mu) \quad (28)$$
$$Var(\xi_i|\xi_l, l \neq i) = \sigma^2 \quad (29)$$

where ρ is the correlation parameter and $Q = (q_{il})$ is the "adjacency" matrix of the map given by $q_{il} = 1$ if i and l are adjacent areas and $q_{il} = 0$ otherwise. It follows from [17] that ξ is multivariate normal with mean $\mu = \mu I$ and covariance matrix $\Sigma = \sigma^2(I - \rho Q^{-1})$, where ρ is bounded above by the inverse of the largest eigenvalue of Q. [6] approximated the posterior density, $f(\xi|y,\mu,\sigma^2,\rho)$ similar to the log-normal case.

The assumption of equation (29) of a constant conditional variance for the $\xi_i^{'}s$ results in the conditional mean of (28) proportional to the sum of the neighbouring $\xi_i^{'}s$ rather than the mean of the neighbouring $\xi_i^{'}s$ (local mean). [18] Proposed an alternative joint density of the $\xi_i^{'}s$ given by:

$$f(\xi) \propto (\sigma^2)^{-m/2} exp\ \left[-\sum_{i\neq l}\sum(\xi_i - \xi_l)^2 q_{il}/(2\sigma^2)\right] \quad (30)$$

This specification leads to:

$$E(\xi_i|\xi_l, l \neq i) = \Sigma_l q_{il}\xi_l/\Sigma_l q_{il} \quad (31)$$

And

$$Var(\xi_i|\xi_l, l \neq i) = \sigma^2/\sum_l q_{il} \quad (32)$$

Here the conditional variance is inversely proportional to $\Sigma_l q_{il}$, the number of neighbours of area i and the conditional mean is equal to the mean of the neighbouring values ξ_l. In the context of disease mapping, the alternative specification may be more appropriate [2].

(a) Hierarchical Bayes Estimation

As noted earlier, the basic log-normal can be extended to allow spatial covariates. A hierarchical bayes (HB) estimator of the spatial CAR-normal model is given by:

$$\text{(i) } y_i|e_i \sim Poisson(e_i\theta_i)$$
$$\text{(ii) } \xi_i|\xi_{j(j\neq i)},\rho,\sigma^2 \sim N[\mu + \rho\sum_l q_{il}\,(\xi_l - \mu),\sigma^2]$$
$$\text{(iii) } f(\mu,\sigma^2,\rho) \propto f(\mu)f(\sigma^2)f(\rho) \quad (33)$$

with $f(\mu) \propto 1;\ \sigma^{-2} \sim G(a,b); a \geq 0, b > 0; \rho \sim U(0,\rho_0)$ where ρ_0 denotes the maximum value of ρ in the CAR-model, and $Q = (q_{il})$ is the "adjacency" matrix of the map with $q_{il} = q_{li}$, $q_{il} = 1$ if i and l are adjacent areas and $q_{il} = 0$ otherwise.

[16] obtained the Gibbs conditionals. In particular,

$$[\mu|\theta,\sigma^2,\rho,y] \sim normal$$
$$[\sigma^{-2}|\theta,\mu,\rho,y] \sim gamma$$
$$[\rho|\theta,\mu,\sigma^2,y] \sim truncated\ normal$$

and $\left[\theta_i|\theta_{j(j\neq i)},\mu,\sigma^2,\rho,y\right]$ does not admit a closed form in the sense that the conditional is known only up to a multiplicative constant. Monte Carlo Markov Chain (MCMC) samples can be generated directly from the three conditionals, but we need to use the M-H algorithm to generate samples from the conditionals $\left[\theta_i|\theta_{j(j\neq i)},\mu,\sigma^2,\rho,y\right]$, $i = 1,2,\ldots m$ [2].

4.2. Two-Level Models

Let y_{ij} and n_{ij} denote the number of cases (deaths) and the population at risk in the jth age class in the ith area $(j = 1,\ldots,J; i = 1,\ldots,m)$ respectively. Using the data $\{y_{ij}, n_{ij}\}$ it is of interest to estimate the age-specific mortality rates τ_{ij} and the age-adjusted rates $\sum_j a_j\tau_{ij}$ where the $a_j^{'}s$ are specified constant.

The basic assumption is

$$y_{ij}\,|\tau_{ij} \sim_{ind} Poisson\,(n_{ij}\,\tau_{ij}) \quad (34)$$

[8] studied HB estimation under different linking models.

$$log(\tau_{ij}) = z_{ij}^T\beta + v_i, \vee_i|\sigma_{ij}^2 \sim_{iid} N(0,\sigma_v^2 \quad (35)$$
$$log(\tau_{ij}) = z_j^T\beta_i, \beta_i|\beta,\Delta \sim_{iid} N_p(\beta,\Delta) \quad (36)$$
$$log(\tau_{ij}) = z_j^T\beta_i + \delta_j, \beta_i|\beta,$$
$$\Delta \sim_{iid} N_p(\beta,\Delta), \delta_j \sim_{iid} N(0,\sigma^2) \quad (37)$$

where z_j is a $p \times 1$ vector of covariance and δ_j is an "offset" corresponding to age class j. [1] assumed that flat prior $f(\beta) \propto 1$ and proper diffuse (that is, proper with very large variance) priors for σ_v^2, Δ and σ^2. For model selection, they used the posterior expected predictive deviance, the posterior predictive value and measures based on the cross-validation productive densities.

5. Examples

We now present some illustrative examples of the application of Small Area Estimation (SAE) in health – disease mapping.

(i) *Lip Cancer*

[16] modeled $\theta_i = log\lambda_i$ as $N(\mu,\sigma^2)$. He also considered a CAR spatial model on the $\theta_i^{'}s$ which relates each θ_i to a set of neighbourhood areas of area i. He developed model-based estimates of lip cancer incidence in Scotland for each of 56 counties.

[6] applied empirical bayes (EB) estimation to data on observed cases, y_i, and expected cases, e_i, of lip cancer

registered during the period 1975 – 1980 in each of 56 counties (small area) of Scotland. They reported the SMR, the empirical bayes (EB) estimate of θ_i based on the Poisson-gamma model $\left(\hat{\theta}_i^{EB}(1)\right)$ and the approximate empirical bayes (EB) estimates of θ_i based on the log-normal model and the CAR-normal model (denoted $\hat{\theta}_i^{EB}(2),\ and\ \hat{\theta}_i^{EB}(3)$) for each of the 56 counties (all values multiplied by 100). The SMR-values varied between 0 and 652 while the empirical bayes (EB) estimates showed considerably less variability across counties as expected: $\hat{\theta}_i^{EB}(1)$ varied between 31 and 422 (with cv = 0.78) and $\hat{\theta}_i^{EB}(2)$ varied between 34 to 495 (with cv=0.85), suggesting little difference between the two sets of empirical bayes (EB) estimates. Ranks of empirical bayes (EB) estimates differed little from the corresponding ranks of the SMRs for most counties, despite less variability exhibited by the empirical bayes (EB) estimates.

For the CAR-normal model, the adjacency matrix, Q, was specified by listing adjacent counties for each county i. The maximum likelihood (ML) estimates of ρ was 0.174 compared to the upper bound of 1.175, suggesting a high degree of spatial relationship in the data set. Most of the CAR estimates, $\hat{\theta}_i^{EB}(3)$, differed little from the corresponding estimates $\hat{\theta}_i^{EB}(1)$ and $\hat{\theta}_i^{EB}(2)$ based on the independence assumption. Counties with few cases, y_i and SMRs differing appreciably from adjacent counties are the only counties affected substantially by spatial correlation. For instance, county number 24 with $y_{24} = 7$ is adjacent to several low-risk counties, and the CAR estimate $\hat{\theta}_{24}^{EB}(3) = 83.5$ is substantially smaller than $\hat{\theta}_{24}^{EB}(1) = 127.7$ and $\hat{\theta}_{24}^{EB}(2) = 123.6$ based on the independence assumption.

[16] applied hierarchical bayes (HB) estimator to the same data using the log-normal and the CAR-normal models. The hierarchical bayes (HB) estimates $E(\theta_i|y)$ of lip cancer incidence are very similar for the two models, but the standard errors, $\sqrt{Var(\theta_i|y)}$, are smaller for the CAR-normal as it exploits the spatial structure of the data.

[19] proposed a spatial log-normal model that allows covariates z_i. It is given by:

$$\text{(i) } y_i|e_i \sim_{ind} Poisson\ (e_i\theta_i)$$
$$\text{(ii) } \xi_i = z_i^T\beta + u_i + v_i \qquad (38)$$

where $z_i^T\beta$ does not include an intercept term.

$v_i \sim_{iid} N(0, \sigma_V^2)$ and the $u_i's$ have joint density of (30):

$f(u) \propto (\sigma_u^2)^{-m/2} exp[-\sum\sum_{i\neq l}(u_i - u_l)^2 q_{il}/(2\sigma^2)]$, where $q_{il} \geq 0$ for all $1 \leq i \neq l \leq m$.

(iii) β, σ_u^2 *and* σ_V^2 *are mutually independents with* $f(\beta) \propto 1$

$$\sigma_u^{-2} \sim G(a_u, b_u) \text{ and } \sigma_V^{-2} \sim G(a_v, b_v) \qquad (39)$$

[19] showed that the Gibbs conditionals except $\left[\theta_i|\theta_{l(l\neq i)}, \beta, \mu, \sigma_u^2, \sigma_V^2, y\right]$, admit closed forms. They also established conditionals for the propriety of the joint posterior, in particular, we need $b_u > 0, b_v > 0$.

(ii) Leukemia Incidence

[19] applied the HB method based on the model given in (39), to leukemia incidence estimation for m=281 census tracts (small area) in an eight-county region of upstate New York. Here $q_{il} = 1$ if i and l are neighbours and $q_{il} = 0$ otherwise, and z_i is a scalar ($\rho = 1$) variable z_i denoting the inverse distance of the centroid of the ith census tract from the nearest hazardous waste site containing trichloroethylene (TCE), a common contaminant of ground water (See [19] for details).

(iii) *Mortality Rates*

[8] applied the hierarchical bayes (HB) method to estimate age-specific and age-adjusted mortality rates for U.S. Health Service Areas (HSAs). They studied one of the disease categories, all cancer for white males, presented in the 1996 Atlas of United States Mortality. The number of HSAs (small areas), m, is 789 and the number of age categories, J, is 10:0-4, 5-14, …, 75 – 84, 85 and higher, coded as 0.25, 1, …,9. The vector of auxiliary variables is given by

$$z_j = [1, j-1, (j-1)^2, (j-1)^3, max\{0, ((d-1) - knot)^3\}]^T$$

for $j \geq 2$ and

$$z_1 = [1, 0.25, (0.25)^2, (0.25)^3, max\{0, (0.25 - knot)^3\}]^T,$$

where the value of the "knot" was obtained by maximizing the likelihood based on marginal deaths $y_i = \sum_i y_{ij}$ and population at risk, $n_j = \sum_i n_{ij}$, where $y_j|n_j, \tau_j \sim_{ind} Poisson\ (n_j\tau_j)$ with $log\tau_j = z_i^T\beta$. The auxiliary vector z_j was used in the Atlas model based on a normal approximation to $log(\gamma_{ij})$ with mean log (τ_{ij}) and matching linking model given by (36) where $\gamma_{ij} = y_{ij}/n_{ij}$ is the crude rate.

[8] used unmatched sampling linking model, based on the Poisson sampling model of (34) and the linking models of (35) – (37). We denote these models as models 1,2 and 3 respectively. Also they used the Monte Carlo Markov Chain (MCMC) samples of generated from the three models to calculate the values of the posterior expected predictive deviance.

$E\{\Delta(y; y_{obs})|y_{obs}\}$ using the chi-square measure $\Delta(y, y_{obs}) = \sum_i\sum_j (y_{ij} - y_{ij,obs})^2/(y_{ij} + 0.5)$. They also calculated the posterior predictive p-values, using $T(y, \tau) = \sum_i\sum_j (y_{ij} - n_{ij}\tau_{ij})^2/(n_{ij}\tau_{ij})$, the standardized cross-validation residuals

$$d_{2,ij}^* = \frac{\tau_{ij,obs} - E(\gamma_{ij}|y_{(ij),obs})}{\sqrt{Var\,(\gamma_{ij}|y_{(ij)obs})}} \qquad (40)$$

where $y_{(ij),obs}$ denotes all elements of y_{obs} except $y_{ij,obs}$ (See [2]; section 10.2.28). The residuals $d_{2,ij}^*$ were summarized by counting:

(a) the number of (i,j) *such that* $|d_{2,ij}^*| \geq 3$, *called "outliers", and*

(b) the number of HSAs is such that $d_{2,ij}^* \geq 3$ *for at least one* j, *called "ǂ of HSAs".*

[20] used models and methods similar to those in[19] to estimate age-specific and age-adjusted mortality rates for chronic obstructive pulmonary disease for white males in HSAs.

6. Extensions

Various extensions of the disease mapping models, studied so far have been proposed in recent literature. [9] proposed a two-stage, bivariate logit-normal model to study joint relative risks (or mortality rates), θ_{1i} and θ_{2i}, of two cancer sites (e.g. lung and large bowel cancers) or two groups (e.g. lung cancer in males and females) over several geographical areas. Denote the observed and expected number of deaths at the two sites as (y_{1i}, y_{2i}) and (e_{1i}, e_{2i}) respectively for the ith area $(i = 1, \dots, m)$. The first stage assumes that $(y_{1i}, y_{2i})|(\theta_{1i}, \theta_{2i}) \sim_{ind} Poisson\,(e_{1i}\theta_{1i}) * Poisson\,(e_{2i}\theta_{2i}), i = 1, \dots, m$, where * denotes that $f(y_{1i}, y_{2i}|\theta_{1i}, \theta_{2i}) = f(y_{1i}|\theta_{1i})(y_{2i}|\theta_{2i})$. The joint risks $(\theta_{1i}, \theta_{2i})$ are linked in the second stage by assuming that (logit (θ_{1i}), logit $(\theta_{2i})) \sim_{ind}$ bivariate normal with means μ_1, μ_2, standard deviations σ_1 and σ_2 and correlation ρ, denoted $N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. Bayes estimators of θ_{1i} and θ_{2i} involve double integrals which may be calculated numerically using Gauss-Hermite quadrature. Empirical bayes (EB) estimators are obtained by substituting maximum likelihood (ML) estimators of model parameters in the bayes estimators. [9] applied the bivariate empirical bayes (EB) method to two data sets consisting of cancer mortality rates in 115 counties of the State of Missouri during 1972 – 1981.

(i) Lung and large bowel cancers

(ii) *Lung cancer in males and females*

The empirical bayes (EB) estimates based on the bivariate model lead to improved efficiency for each site (group) compared to the empirical bayes (EB) estimates based on the univariate logit-normal model, because of significant correlation: $\hat{\rho} = 0.54$ for data set (i) and $\hat{\rho} = 0.76$ for data set (ii). [21] first-order approximation to the posterior variance was used as a measure of variability of the empirical bayes (EB) estimates.

[22] extended the bivariate model by introducing spatial correlations (via CAR) and covariates into the model. They used a hierarchical bayes (HB) approach instead of the empirical bayes (EB) approach. They applied the bivariate spatial model to male and female lung cancer mortality in the State of Missouri, and constructed disease maps of male and female lung cancer mortality rates by age group and time period.

[11] extended the Poisson-gamma model to handle nested data structures, such as a hierarchical health administrative structure consisting of local health districts, i, in the first level and local health areas, j, within districts in the second level $j = 1, \dots, n_i,\ i = 1, \dots m)$. The data consist of incidence or mortality counts y_{ij} and the corresponding population at risk counts, n_{ij}.[11] derived empirical bayes (EB) estimates of the local health area rates θ_{ij}, using a nested error Poisson-gamma model. The bayes estimator of θ_{ij} is a weighted combination of the crude local area rate, y_{ij}/n_{ij}, the correspond crude district rate, $y_{i.}/n_{i.}$ and the overall rate $y_{..}/n_{..}$, where $y_{i.} = \sum_j y_{ij}$ and $y_{..} = \sum_i y_{i.}$, and $n_{i.}, n_{..}$) similarly defined.

[11] used the [21] first-order approximation to posterior variance as a measure of variability. They applied the nested error model to infant mortality data from the province of British Columbia, Canada.

7. Concluding Remarks

Small area estimation of health related characteristics has attracted a lot of attention in the Western countries like the U.S., U.K., and Canada because of a continuing need to assess health status, health practices and health resources at both the national and sub national levels. Reliable estimates of health related characteristics help in evaluating the demand for health care and the access individuals have to it. Mortality and disease rates of small area in a region or a county are often used to construct disease maps which are used to display geographical variability of a disease and identify high rate and, or high risk areas warranting interventions. This article attempts to overview the main advances in small area estimation methods in the field of disease mapping and some relevant statistical models in disease mapping. A critical review of earlier developments of small area estimation methods in the field of disease mapping which serve as a necessary background for the various extensions of disease mapping models proposed in recent literature with illustrative examples are presented.

Two important issues not considered are model selection and model diagnostics. As mentioned earlier; small area estimation is one of the few fields in survey sampling, where it is widely recognized that the use of model dependent is often inevitable. Given the growing use of small area estimates and their immense importance, it is imperative to develop efficient tools for the selection of their goodness of fit.

A further issue which certainly deserves consideration is the objective comparison of the different statistical models for disease mapping and an evaluation of the quality of their forecasts. These will be our focus in a forthcoming article.

REFERENCES

[1] Nandram, B 1999. "An empirical bayes prediction interval for the finite population mean of small area". Statistica Sinica, (9), 325-343.

[2] J. N. K. Rao. Small Area Estimation, New York, Willey, 2003.

[3] Malec, D., Davis, W. W. and Cao, X. 1999. Model-based small area estimates of overweight prevalence using sample selection adjustment. Statistics in Medicine, (18), 3189-3200.

[4] R. Folsom, B.V Shah and A. Vaish. Substance Abuse in States: A Methodological Report on Model Based Estimates from 1994-1996 National Household Surveys on Drug Abuse. In Proceedings of the Section on Survey Research Methods: American Statistical Association. 1999. Washington, D. C. 371-375.

[5] M. Chattopadhyay, P. Lahiri, M. Larsen and J. Reimnitz. 1999. Composite Estimators of Drug Prevalences for Sub-State Areas. Survey Methodology, (25), 81-86.

[6] D. Clayton and J. Kaldor. 1987. Empirical Bayes Estimates of Age-Standardised Relative Risks for Use in Disease Mapping. Biometrics, (43), 671-681.

[7] G. S. Datta, M. Ghosh and L. A. Waller, "Hierarchal and Empirical Bayes Methods for Environmental Risk Assessment," in Handbook of Statistics ,P. K. Sen and C. R. Rao (eds.), Volume 18, Elsevier Science B. V., Amsterdam, pp. 223-245, 2000.

[8] B. Nandram, J. Sedransk and L. Pickle. 1999. Bayesian Analysis of Mortality Rates for U. S. Health Service Areas. Sankhyā, Series B, (61), 145-165.

[9] C. M. DeSouza. 1992. An Appropriate Bivariate Bayesian Method for Analysing Small Frequencies. Biometrics, (48), 1113-1130.

[10] I. H. Langford, A. H. Leyland, J. Rasbash and H. Goldstein. 1999. Multilevel Modelling of the Geographical Distribution of Diseases. Applied Statistics. (48), 253-268.

[11] C. B. Dean and Y. C. MacNab. 2001. Modeling of Rates over a Hierarchical Health Administrative Structure. Canadian Journal of Statistics. (29), 405-419.

[12] R. J. Marshall. 1991. Mapping Disease and Mortality Rates using Empirical Bayes Estimators. Applied Statistics, (40), 283-294.

[13] P. Lahiri, and T.Maiti. Empirical Bayes Estimation of Relative Risks in Disease Mapping. Technical Report, Department of Statistics, University of Nebraska, Lincoln. 1999.

[14] C. L. Christiansen and C. N. Morris. 1997. Hierarchical Poisson Regression Modeling. Journal of the American Statistical Association, (92), 618-632.

[15] J. Jiang and W. Zhang. 2001. Robust Estimation in Generalised Linear Mixed Models. Biometrika, (88), 753-765.

[16] T. Maiti, "1998. Hierarchical Bayes Estimation of Mortality Rates for Disease Mapping. Journal of Statistical Planning and Inference, (69), 339-348.

[17] J. E. Besag. 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion). Journal of the Royal Statistical Society, Series B, (35), 192-236.

[18] D. Clayton and L. Bernardinelli, "Bayesian Methods for Mapping Disease Risk," in Geographical and Environmental Epidemiology: Methods for Small-Area Studies, P. Elliot, J. Cuzick, D. English and R. Stern (eds.), Oxford University Press, London, 1992.

[19] M. Ghosh, K. Natarajan, L. A. Waller and D. H. Kim.1999. Hierarchical Bayes GLMs for the Analysis of Spatial Data: An Application to Disease Mapping. Journal of Statistical Planning and Inference, (75), 305-318.

[20] B. Nandram, J. Sedransk and L. Pickle. 2000 Bayesian Analysis and Mapping of Mortality Rates for Chronic Obstructive Pulmonary Disease. Journal of the American Statistical Association, (95), 1110-1118.

[21] R. E. Kass and D. Steffey.1989. Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)," Journal of the American Statistical Association, (84), 717-726.

[22] H. Kim,H D. Sun and R. K. Tsutakawa. 2001. A Bivariate Bayes Method for Improving the Estimates of Mortality Rates with a Twofold Conditional Autoregressive Model. Journal of the American Statistical Association, (96),1508-152.