The Economic and Social Review, Vol. 27, No. 4, July, 1996, pp. 341-363

Mapping Disease Risk Estimates Based on Small Numbers: An Assessment of Empirical Bayes Techniques*

D.G. PRINGLE
St. Patrick's College, Maynooth

Abstract: Choropleth maps are frequently used to analyse spatial variations in the risk of a disease. In such maps the relative risk is typically quantified by dividing some measure of the number of cases of the disease by some measure of the population at risk. The resulting rates may be regarded as maximum likelihood estimates of individual risk. These estimates may be unstable if the areas are very small or if the disease is rare. In such situations, the highest and lowest values on the map will display a tendency to be concentrated in the areas with the smallest populations. The traditional solution to this problem is to supplement maps based on ratios with probability maps. However, probability maps display an opposite bias — i.e., they tend to highlight the areas with the largest populations. Several statisticians have suggested a compromise between these two extremes using empirical Bayes techniques. This paper outlines the rationale underlying empirical Bayes techniques, and assesses their usefulness using case studies of neo-natal mortality and cancer mortality.

I INTRODUCTION

Maps of the prevalence or incidence of a disease often provide useful aetiological information about the possible causes of that disease. A heightened incidence of a particular disease in certain areas could reflect the effects of either "contextual" factors such as local environmental risks (e.g., a pollution or radiation source), or "compositional" factors such as a higher percentage of people who as individuals have a higher risk because of genetic

*The author would like to thank the unknown referees of an earlier draft of this paper for their many useful comments.

factors or high risk behaviours (Duncan *et al.*, 1993). Conversely, the absence of a heightened incidence may help to alleviate unjustified public concern about a non-dangerous facility. Disease maps can also aid decision making by health service policy makers, whether their primary concern is disease prevention or the provision of services for medical treatment. Some health risk factors, such as pollution sources or high risk behaviours (e.g., smoking), may be amenable to preventive strategies; other risk factors (such as age, race or gender) obviously are not, but a knowledge of the spatial distribution of these factors may facilitate the provision of a more efficient health delivery system.

Disease mapping is by no means unproblematic. There are a large number of statistical problems which have to be addressed — some of which are by no means obvious. There are also a large number of cartographic pitfalls awaiting the inexperienced map-maker. Maps can create a powerful visual image, but they can also create a very misleading impression either by accident or design (Monmonier, 1991). The meaningful interpretation of disease maps is therefore contingent upon an awareness of the potential problems.

There are at present comparatively few examples of disease mapping in Ireland, although it is envisaged that this situation will change rapidly with increased adoption of geographical information systems (GIS). Dawson's (1911) study of "insanity" provides an interesting early example of a disease map (especially when compared with maps of reported schizophrenia morbidity at present), but most of the more recent examples of disease mapping tend to be based on mortality data. The Central Statistics Office normally includes maps of mortality in the *Report on Vital Statistics* which it compiles each year for the Department of Health. However, these maps are based on data which are standardised for variations in population structure using the direct method. This tends to amplify minor stochastic fluctuations, thereby disguising aetiologically important regularities in the underlying variations in risk. Indirect standardisation, as used for example by Howell *et al.* (1993), is preferable, although the spatial variations displayed in such maps tend to be dominated by variations in the number of deaths in the older age cohorts. Pringle (1986) attempted to overcome this particular problem, with mixed success, by calculating an "Unfulfilled Life Index".

Most Irish studies map regional variations at a national level — mainly because of the absence of published data at larger scales. However, a few studies have mapped mortality at an intra-urban level using data directly extracted from death certificates. These studies are particularly susceptible to the "small numbers" problem (Congdon, 1990; Diehr, 1984; King, 1979). Johnson and Dack (1989) in a study of Dublin attempted to reduce the problem by aggregating data for several consecutive years, whilst Pringle

(1983; 1987) in studies of Belfast and Dublin aggregated adjoining areas. Both strategies provide only partial solutions.

The "small numbers" problem is probably the most pervasive problem in disease mapping. The problem arises when the absolute number of cases of the disease in each area is small, either because the areas are very small or the disease is comparatively rare or both. When used as the numerators to calculate rates, small numbers may provide very misleading impressions of the underlying risks. Similar problems are associated with other social and economic indicators expressed as rates or ratios for small areas (e.g., unemployment rates, per capita income). Possible solutions, based on empirical Bayes techniques, have been proposed by various statisticians in recent years. These techniques form the focus of the present paper. The paper begins by explaining the small numbers problem in more detail and then outlines some of the strategies which have traditionally been adopted to try to circumvent the problem. Alternative approaches based on empirical Bayes techniques are then introduced. The final part of the paper reports an empirical assessment of the effectiveness of these techniques using case studies of neonatal mortality and deaths from cancer.

II THE SMALL NUMBERS PROBLEM

Normally when one maps the incidence or prevalence of a disease one is primarily interested in the spatial distribution of individual risks — i.e., the question we wish to address is: do people living in some areas have a higher risk of disease than people in other areas? Age specific rates provide maximum likelihood estimates of the individual risk in a given age cohort — i.e., the maximum likelihood estimate of the individual risk θ_{ij} for people in age group j in area i is given by dividing the number of cases of the disease d_{ij} amongst people in age group j in area i by the corresponding population at risk p_{ij}:

$$\hat{\theta}_{ij} = d_{ij} / p_{ij} \tag{1}$$

Age standardised rates and ratios (e.g., standardised mortality ratios) are essentially weighted sums of age specific rates and are consequently subject to similar problems. However, in the interests of clarity, discussion will be confined here to age specific rates for a single age group.

The small numbers problem arises if the values of the numerator d_{ij} are small and discrete. The numerators will be discrete because they are counts of the number of people who are sick or die (i.e., you cannot have half a person); and they may be small, either if the areas are small or if the

incidence or prevalence of the disease is small (i.e., if it is a rare disease).

If the numerators are small, the resulting estimates of risk will become very unstable — i.e., one case more or one case less will make a large percentage difference. For example, consider a disease with an average incidence of one case in 1,000, in two areas A and B having populations of 10,000 and 1,000 respectively. The expected number of cases in area A would be 10 (i.e., 10,000/1,000). If there were actually 11 cases, the estimated risk for that area would increase to 1.1 per 1,000, whereas if the number of cases was one below the expected value the rate would be 0.9 per 1,000 — either way, there would be comparatively little change. However, the expected number of cases in area B is only 1 (i.e., 1,000/1,000). If there was one case less than expected in a given year, the rate for that area would fall to 0.0; whilst if there was one case more than expected, the rate would increase to 2.0. In other words, the rates for the smaller area could fluctuate from extremely low to extremely high, possibly reflecting nothing more than stochastic variations. The net effect of this is that the extreme values on a choropleth map of estimated risks (whether extremely low or extremely high) will display a tendency to be concentrated in the smaller areas. Given that the extreme values are the ones which attract most attention when interpreting the map, there is an obvious danger of creating a very misleading impression of the spatial distribution of disease risk.

Table 1: *The Small Numbers Problem*

	Area A	*Area B*
Population At Risk	10,000	1,000
Expected Cases	10	1
Rate If +1 Case	1.1	2.0
Rate If –1 Case	0.9	0.0

Most researchers attempt to minimise the problem by aggregating data in ways which will increase the expected numbers of cases. There are three main strategies in this regard. The first is to extend the study period (e.g., from 1 year to 10 years). The second is to aggregate adjoining areas to form a smaller number of larger areas, each with a larger expected number of cases. The third is to aggregate data on a particular disease with data on diseases believed to have a similar aetiology. Each strategy has obvious drawbacks: the first will disguise temporal trends within the data, some of which could be aetiologically significant; the second will disguise local variations, some of which could again be aetiologically significant; whilst the third could result in

different diseases with different causal processes being mixed and confused.

A further response to the small numbers problem (often in addition to one or more of the strategies outlined above) is to supplement the choropleth map of estimated risks with a probability map indicating which areas have a significantly high or a significantly low number of cases. These probabilities are usually calculated based on the assumption that the underlying risks are the same in all areas and that the number of cases in each area is the product of a Poisson process. If the cumulative probability calculated under these assumptions is very large or very small, then the observed number of cases may be designated "significantly high" or "significantly low" (Choynowski, 1959; Giggs *et al.*, 1980; White, 1972). However, this "solution" also runs into problems if the numbers are small. It is quite possible for there to be no cases at all in an area, yet if the area has a small population at risk even zero cases may not be low enough to be regarded as "significant". Likewise, a much higher prevalence is required in a small area for the number of cases to be regarded as significantly high. The net effect of this is that probability maps tend to create an impression which is biased in favour of the larger (i.e., more populous) areas (Kaldor and Clayton, 1989). Urban areas are therefore more likely to be identified as having a significantly high or a significantly low number of cases than less populous rural areas.

Thus, to summarise, choropleth maps of the estimated risk tend to focus attention on the smaller areas, whereas probability maps tend to focus attention on the larger areas. Clearly we need some sort of compromise between these two extremes. Several biostatisticians have suggested that empirical Bayes techniques offer a possible solution.

III THE EMPIRICAL BAYES APPROACH

The observed number of cases of the disease d_{ij} in age group j in area i may be regarded as the outcome of a Poisson process with an expectation $\theta_{ij}p_{ij}$ where θ_{ij} is the underlying risk and p_{ij} is the number of people at risk. However, the observed number of cases may be either higher or lower than the expected number, due to stochastic variations. The basic objective is to estimate and map the unknown underlying risk θ_{ij} for each area using the information available on the observed numbers of cases. The problem is that, for any given area, one does not know whether the observed number of cases is higher or lower than the expected number (which is governed by the unknown underlying risk) or by how much.

In the absence of any other information about the underlying risks, the risks could be estimated using formula (1). This gives the maximum likelihood estimate of the underlying risk, but (as explained in the previous

section) it may be subject to very large stochastic fluctuations as a result of the small numbers problem. The basic premise of the empirical Bayes approach is that we do in fact have access to additional information which may be incorporated in the estimation process to produce more realistic estimates of the underlying risks.

To understand the nature of this additional information, it is instructive to consider an extreme hypothetical situation in which we have perfect information for every area except one — for which we have no information whatsoever (Kaldor and Clayton, 1989). Applying the maximum likelihood formula we would be unable to say anything at all about the risk in the area with the missing data. However, it does not seem unreasonable to assume that the risk in this area is probably of the same order of magnitude as the risks in the areas for which we do have information. Indeed, if the risks in the other areas display a high degree of spatial order, we may even be justified in concluding that the risk in the area with the missing data is similar to the risks in the areas which border upon it. In other words, we could use our knowledge of the distribution of risks in general to say something about an area for which we have no empirical information whatsoever.

A similar logic may be applied in more normal circumstances to areas for which we do have empirical data. If the maximum likelihood estimate of the risk in an area is extremely high or extremely low then it would seem reasonable to modify it in the light of what we know about the distribution of risks in other areas. The resulting estimate of the underlying risk for each area may therefore be regarded as a compromise between the maximum likelihood estimate based upon the observed number of cases and the overall risk for the entire study area. The balance between these two components should reflect the reliability of the empirical data: if the area is large, then the maximum likelihood estimate should receive most weight; however, if the area is small, and the maximum likelihood estimates could be subject to large stochastic fluctuations, then the overall mean should receive more weight.

These objectives may be achieved within a Bayesian framework. Bayesian statistics are based upon the observation that the joint probability of two events A and B is equal to the probability of the first times the probability of the second conditional upon the first (Iverson, 1984):

$$p(AB) = p(A).p(B|A) \tag{2}$$

Thomas Bayes, an eighteenth century clergyman, used this relationship to develop a theorem which may be written as:

$$p(\theta|x) = \frac{p(x|\theta).p(\theta)}{\int p(x|\theta).p(\theta).d\theta} \tag{3}$$

where x is the observed data and θ is an unknown parameter to be estimated (in our case, the risk of disease). The left hand side of the equation is referred to as the *posterior* distribution, and it describes the distribution of possible values for the parameter conditional upon the observed data. The term $p(\theta)$ on the right hand side, referred to as the *prior* distribution, describes what we know about the distribution of possible parameter values before observing the data. The bottom line on the right hand side is a normalising constant which integrates to unity, whilst the term $p(x \mid \theta)$ may be interpreted as the likelihood of θ given data x. Bayes theorem may therefore be summarised as:

$$\text{posterior distribution} \propto \text{likelihood function} \times \text{prior distribution}$$

In the context of disease mapping, the prior distribution expresses our initial belief about the distribution of underlying risks between areas, the likelihood function is the Poisson distributed number of cases conditional on the true risk in each area, and the posterior distribution is the distribution of possible values for the underlying risk in each area conditional upon the observed number of cases.

The researcher must make assumptions about the nature of the prior distribution of the underlying risks and the likelihood function. In a full Bayesian approach the researcher would be required to specify the prior distribution in its entirety, but in an empirical Bayes approach the researcher need only make an assumption about the nature of the prior distribution. The parameters of the prior distribution are then estimated from the observed data (Bailey and Gatrell, 1995).[1] This prior distribution is then combined with the likelihood function to give the posterior distribution. The mean of the posterior distribution provides an empirical Bayes estimate of the underlying risk.

Applications of the empirical Bayes approach vary in their choice of a prior distribution. A Gamma distribution is often preferred because it is intuitively reasonable and results in comparatively simple calculations, due to the fact that a Gamma distribution combined with a Poisson distribution gives a Negative Binomial distribution (e.g., Clayton and Kaldor, 1987; Langford, 1994; Manton *et al.*, 1987; Stone, 1988; Tsutakawa, 1988). Other published studies assume a Normal distribution of the logits of the relative risks (e.g., Tsutakawa *et al.*, 1985); a Normal distribution of the Freeman-Tukey

1 Empirical Bayes techniques make use of the observed data twice: once to estimate the parameters of the prior distribution; and once to estimate the parameters of the posterior distribution — i.e., the "prior" information to be "refined" is to some extent dependent upon the same set of empirical observations which are used to refine it. There is therefore an element of "double counting" which runs counter to the Bayesian ethos. This is difficult to justify on theoretical grounds, but it may be defended on practical grounds by virtue of the fact that it seems to work.

transformed risks (Cressie, 1993); a Log-Normal distribution (e.g., Clayton and Kaldor, 1987); a Beta distribution (e.g., Miyawaki and Chen, 1981); a uniform distribution (e.g., Heisterkamp, 1993); or even a non-parametric model (e.g., Clayton and Kaldor, 1987; Heisterkamp, 1993). The choice of prior distribution clearly influences the method of calculation and the values of the resulting estimates: the empirical Bayes approach should therefore be regarded as a family of techniques rather than a single method.

Most of these models are unfortunately mathematically complex. The case studies below use the Gamma model developed by Clayton and Kaldor (1987). This is probably the best known model and has the advantage that it can be expanded to accommodate situations where it is necessary to take account of variations in age composition (as in the second case study reported below). Clayton and Kaldor's model assumes that the relative risks θ_i follow a Gamma distribution with a scale parameter α, and a shape parameter ν (i.e., mean ν/α and variance ν/α^2) and that the observed number of cases O_i are Poisson variates with an expectation $\theta_i E_i$, where E_i is the number of cases would be expected in area i given its population at risk. Clayton and Kaldor derive two equations which can be used recursively to estimate the parameters α and ν:

$$\frac{\hat{\nu}}{\hat{\alpha}} = \frac{1}{N}\sum_i \frac{O_i + \hat{\nu}}{E_i + \hat{\alpha}} = \frac{1}{N}\sum_i \hat{\theta}_i \tag{4}$$

$$\frac{\hat{\nu}}{\hat{\alpha}^2} = \frac{1}{N-1}\sum_i \left(1 + \frac{\hat{\alpha}}{E_i}\right)\left(\theta_i - \hat{\nu}/\hat{\alpha}\right)^2 \tag{5}$$

The author is unaware of any commercially available software to solve for α and ν, so it is necessary to write one's own. The procedure requires one to substitute estimates of α and ν into the right hand sides of Equations (4) and (5). A new estimate for α may then be obtained by dividing the answer for Equation (4) by the answer for Equation (5). This may then be substituted into Equation (4) to get a new estimate of ν. The whole procedure is then repeated until the estimated values of α and ν stabilise between iterations. Having solved for α and ν, the posterior expectation of θ_i conditional on O_i is given by:

$$E(\theta_i | O_i; \alpha, \nu) = \frac{O_i + \nu}{E_i + \alpha} \tag{6}$$

This provides the empirical Bayes estimates of the relative risks for each area. These may be converted into rates by simply multiplying them by the overall rate.

IV CASE STUDIES

Most published studies of empirical Bayes techniques provide a detailed mathematical argument in favour of a particular approach, but provide little concrete evidence to support the contention that empirical Bayes techniques are superior to more traditional maximum likelihood techniques. This section of the paper therefore attempts to provide an objective assessment of their utility based upon case studies of neo-natal deaths and female deaths from cancer in Ireland in 1989 and 1990 — the last two years for which the *Report on Vital Statistics* is available at the time of writing. In both studies the data are disaggregated by county and county borough.

Case Study 1 : Neo-Natal Mortality, 1989

Neo-natal mortality (i.e., deaths before 4 weeks) provides a suitable subject for assessment purposes. The neo-natal mortality rate is calculated by dividing the total number of neo-natal deaths in a given year by the total number of live births in the same year.[2] Both figures may be directly obtained from the *Report on Vital Statistics*, thereby eliminating the possibility of additional complications associated with other mortality or morbidity rates arising from the need to estimate the population at risk for inter-censal years or from the need to take account of inter-county variations in age composition. The number of neo-natal deaths is also small enough to make estimates of the underlying risk in most counties extremely problematic, thereby providing a difficult challenge to the methods under scrutiny.

The total number of live births in 1989 was 52,018; whilst the total number of neo-natal deaths was 249, giving a national death rate of 4.8 per thousand. The absolute numbers of deaths per county varied from a low of 0 (Roscommon) to a high of 37 in Dublin County. The number of births varied from a low of 355 (Leitrim) to a high of 8,825 in Dublin County. The resulting rates per county ranged from a low of 0.0 (Roscommon) to a high of 9.46 (Limerick CB).

If we graph the calculated neo-natal mortality rates against the number of births (Figure 1), we can see that the neo-natal rates in counties with less

2. This gives a period rate. A cohort rate, calculated by dividing the total number of deaths under the age of 4 weeks amongst children born in a given year by the total number of births in that year, would be preferable (Pressat, 1978). However, the period rate is normally used in Ireland and is sufficiently accurate for present purposes.

than 2,000 births (i.e., the majority of counties) vary from very low to very high. The rates for the larger areas (i.e., areas with most births) tend to be more in line with the national average. This may be because the underlying risks for the larger areas just happen to fall in the middle range, but a more likely explanation is that the observed variations in the smaller counties are a function of the small numbers problem (i.e., one extra death in a small county could move it from the "low risk" category into a "high risk" category).

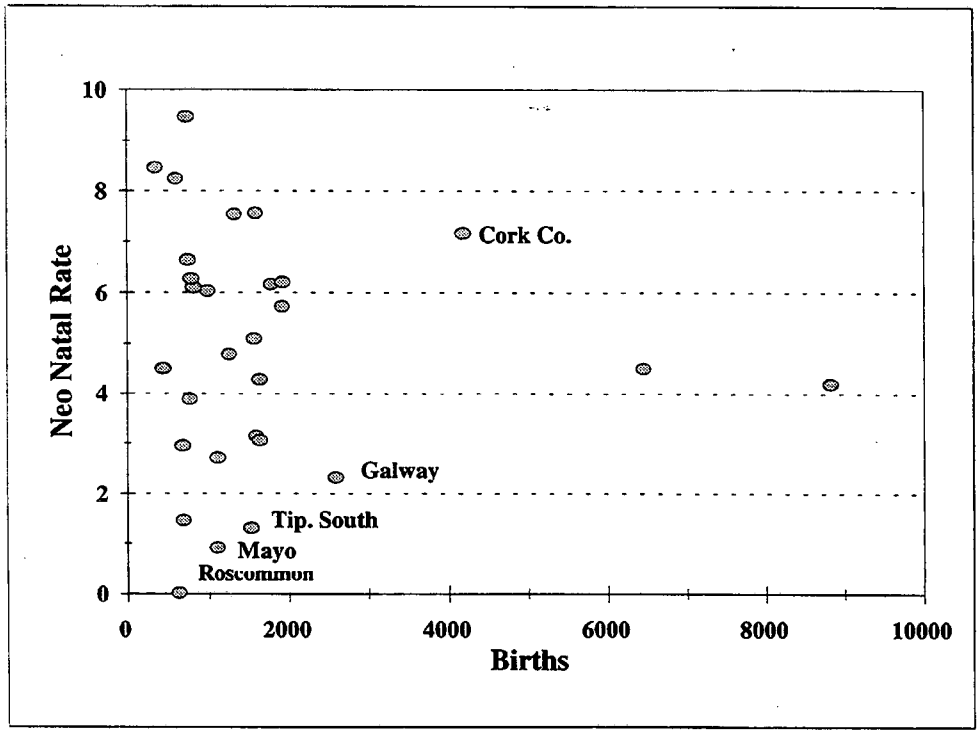

Figure 1: *Maximum Likelihood Estimates of Neo-Natal Mortality Rates in 1989 Graphed Against the Number of Births.*

As a result, the spatial distribution of mortality rates (Figure 2) cannot be regarded as a reliable indicator of the spatial distribution of risks, given the small numbers of deaths in most of the counties. The normal procedure would be to test the rates in each area for statistical significance. Doing this, it is found that Cork County has a significantly high number of deaths, and four counties (Tipperary South, Galway, Mayo and Roscommon) have a significantly low number of deaths.

If we look at the location of these counties on the graph (Figure 1), it will be noted that although Cork County is the only county to have a significantly

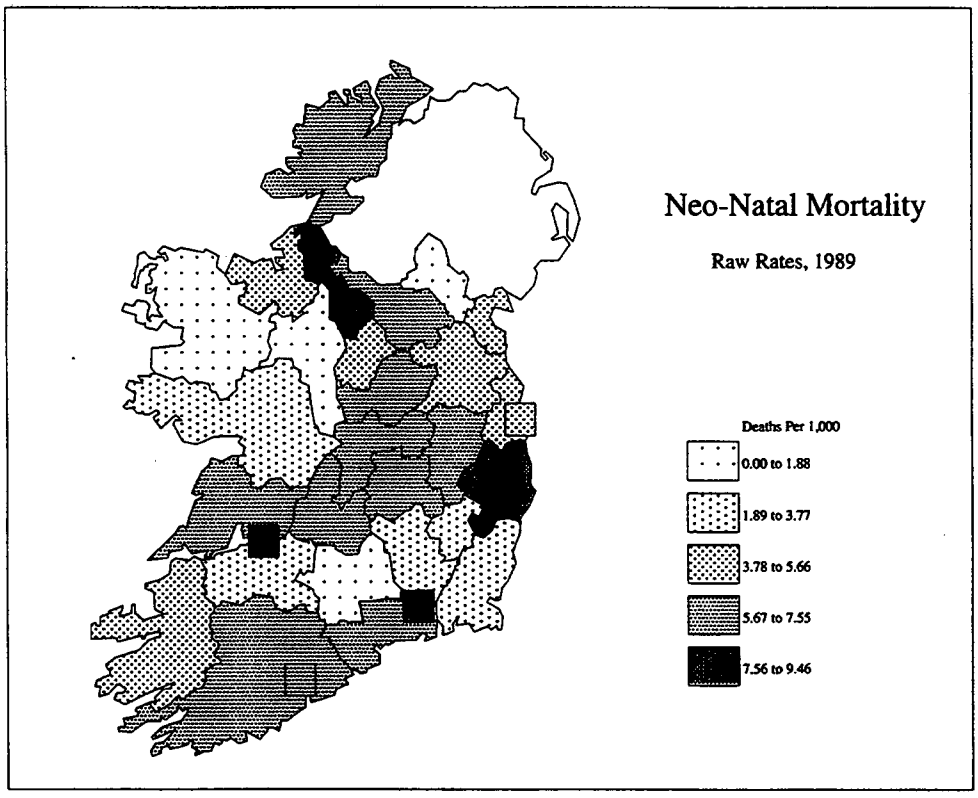

Figure 2: *Spatial Distribution of Maximum Likelihood Estimates of Neo-Natal Mortality Rates in 1989.*

high number of deaths, it had only the sixth highest rate. This indicates that the rates for the 5 counties having a higher rate probably need to be interpreted with caution due to the fact that they are not significant. However, it may also reflect the tendency for significance tests to favour the more populous areas (such as Cork County). The significance tests identify the three areas with the lowest rates as having a significantly low number of deaths, but the area with the fourth lowest rate (Monaghan) is not significant, whereas Galway with a higher death rate is significant. Again there may be a tendency for the test to favour the larger areas (i.e., Galway, as opposed to Monaghan).

In an attempt to strike a balance between these contradictory tendencies to highlight either the smaller areas or the larger areas, empirical Bayes estimates of the underlying risk were calculated using the approach developed by Clayton and Kaldor (1987). The original range of values from 0.0 to 9.46 was found to be compressed into a much smaller band, ranging

Table 2: *Maximum Likelihood and Empirical Bayes Estimates of Neo-natal Mortality in 1989 (Deaths Per 1,000 Live Births)*

County	*Maximum Likelihood Estimate*	*Empirical Bayes Estimate*
Carlow	2.95	4.61
Dublin CB	4.49	4.64
Dublin County	4.19	4.43
Kildare	5.72	5.07
Kilkenny	2.69	4.45
Laois	6.27	5.02
Longford	4.49	4.80
Louth	4.78	4.82
Meath	5.08	4.89
Offaly	6.12	5.00
Westmeath	6.02	5.02
Wexford	3.13	4.43
Wicklow	7.56	5.47
Clare	7.53	5.38
Cork CB	6.15	5.17
Cork County	7.15	5.87
Kerry	4.27	4.69
Limerick CB	9.46	5.41
Limerick County	3.05	4.40
Tipperary NR	6.08	5.00
Tipperary SR	0.90	4.14
Waterford CB	8.24	5.19
Waterford County	6.65	5.06
Galway	2.30	3.98
Leitrim	8.45	5.06
Mayo	1.31	4.03
Roscommon	0.00	4.30
Sligo	3.88	4.70
Cavan	6.26	5.02
Donegal	6.20	5.20
Monaghan	1.44	4.43

from a low of 3.98 (Galway) to a high of 5.87 (Cork County) using Clayton and Kaldor's method (Table 2). Graphing the empirical Bayes estimates against births, it is clear that the rates for the larger counties remained more or less unchanged, whereas those for the smaller counties were shrunk towards the national average (Figure 3). The empirical Bayes estimates can be regarded as a weighted mean between the information available on each county and that available for the whole country, with more weight being given to the county information when the county is large.

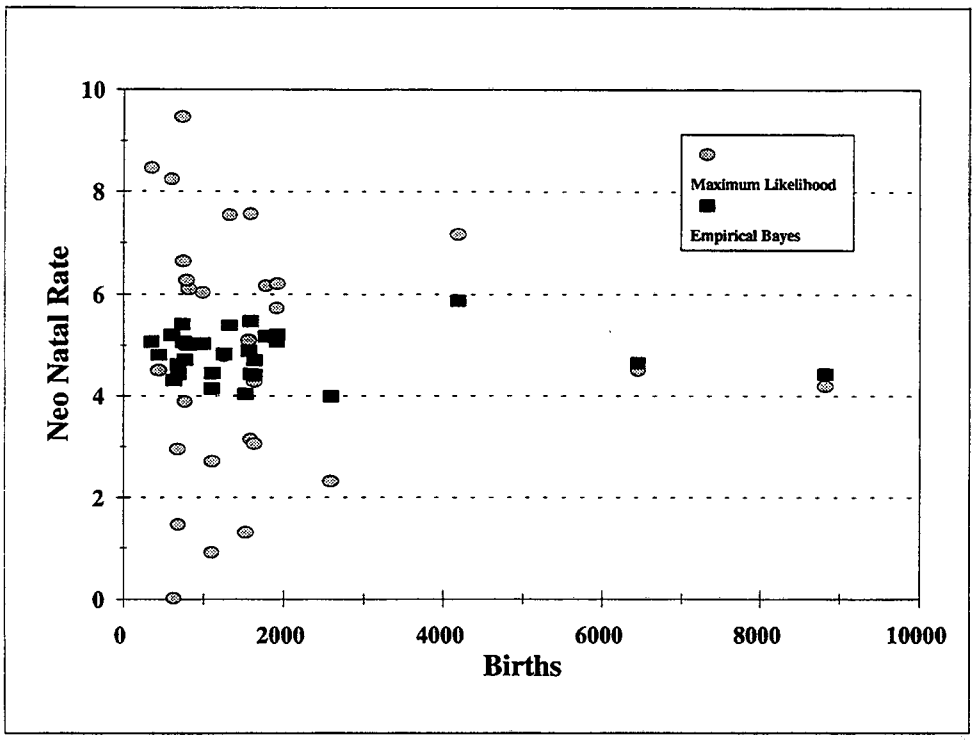

Figure 3: *Maximum Likelihood and Empirical Bayes Estimates of Neo-Natal Mortality Rates in 1989 Graphed Against the Number of Births.*

The pattern of empirical Bayes estimates is, as one might expect, broadly similar to the pattern of raw rates (or maximum likelihood estimates) shown in Figure 2. However, there are also some notable differences (Figure 4). It will be noted, for example, that Cork has now emerged as the area of highest risk, overtaking 4 counties with a much smaller population. Likewise Galway, a county with a large population, has moved into the lowest risk category, whilst Leitrim, a county with a low population, has moved from a high risk category into a medium risk category. Overall, there is a higher degree of spatial autocorrelation in the distribution of the empirical Bayes estimates — i.e., there is a greater degree of regional patterning in the empirical Bayes estimates, possibly indicating a reduction in the "noise to signal" ratio.

Figure 4 intuitively appears to provide a more reliable indication of the underlying risks than Figure 2. However, appearances can be deceptive, so we ideally require a more objective assessment of the relative efficiency of the two methods. The underlying risks are by definition unknown, so it is impossible to know with complete certainty which method provides the better

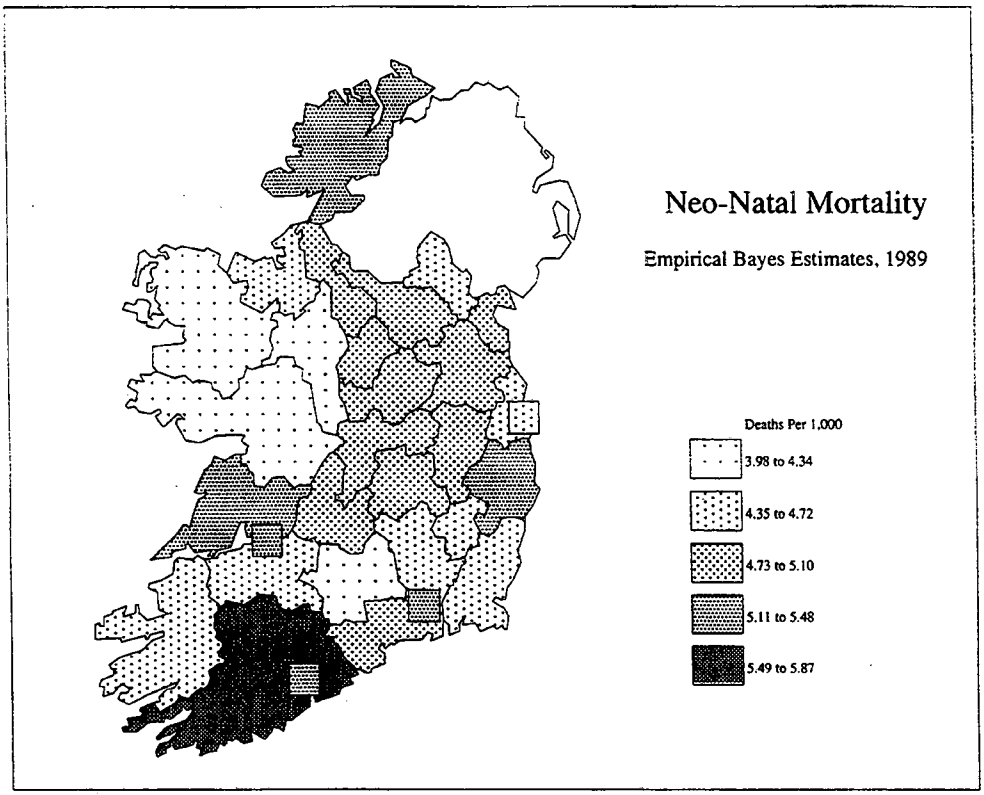

Figure 4: *Spatial Distribution of Empirical Bayes Estimates of Neo-Natal Mortality Rates in 1989.*

estimates. However, if it is assumed that the pattern of risks remains much the same from one year to the next, then one way of gauging which method provides the better estimates of the risks in 1989 is to see which set of estimates provides the better prediction of the actual numbers of deaths in 1990. The expected number of deaths in 1990 were therefore calculated using three alternative approaches:

(a) The expected number of deaths in each county was calculated assuming an equal risk in each area — this is equivalent to assuming that there is no systematic variation in risk and that the observed pattern of neo-natal mortality simply reflects stochastic variations;
(b) The expected number of deaths in each county was calculated assuming the relative risks in 1990 were distributed in the same way as the raw (i.e., maximum likelihood) neo-natal mortality rates in 1989; and
(c) The expected number of deaths in each county was calculated

assuming the relative risks in 1990 were distributed in the same way as the empirical Bayes estimates in 1989.

The number of deaths in each county predicted by each method was then compared with the actual number of deaths in 1990. The performance of each method is summarised by calculating the mean of the absolute differences between the observed and expected number of deaths in each area.[3] The results are shown in Table 3.

Table 3: *Relative Accuracy of Three Predictions of Neo-natal Deaths in 1990*

	Mean Absolute Error
Method 1 — Equal Risk	2.56
Method 2 — Maximum Likelihood Estimates	2.75
Method 3 — Empirical Bayes Estimates	2.45

Comparison of the first two methods rather surprisingly indicates that one could predict the number of deaths in each county in 1990 more accurately assuming that all counties had exactly the same risk rather than using the information on the maximum likelihood estimates of the neo-natal rates in 1989 as a guide. This would suggest that most of the observed variations in neo-natal rates in 1989 are actually stochastic noise generated by the small numbers problem, rather than a reliable indication of variations in the underlying risk.

The 1989 empirical Bayes estimates, in contrast, provide a better prediction of the 1990 rates than the method assuming equal risk. The improvement is admittedly quite small. However, if one examines Figure 3, it will be noted that most of the empirical Bayes estimates have been "shrunk" to a value close to the national mean (or, to put it another way, close to the values which would pertain under an assumption of equal risk). The "signal" detected by the empirical Bayes estimates is still very weak, but the fact that it produces an improvement in the accuracy of the predictions suggests that the technique is in fact successfully detecting real (and persistent) variations in the underlying risk.

Case Study 2 : Female Cancer Mortality

Female deaths from malignant neoplasms are examined in the second case study. This would appear, in some respects, to provide a less difficult challenge than neo-natal mortality. There were a total of 3,481 deaths amongst

3. The performance of each of the three methods of prediction is summarised here by the mean absolute error. Measures based on mean squared errors are used in many statistical methods, but they tend to give greater weight to a small number of areas with large errors. The method used here provides a better indication of the average error in each area.

women from cancer in 1989, compared with only 249 neo-natal deaths, so one might assume that the small numbers problem would be less pronounced and that the underlying spatial variations in risk would be revealed by conventional maximum likelihood methods. However, cancer mortality provides a more difficult challenge in other respects. Cancer is probably not a single disease but a family of diseases with different aetiologies and possibly different geographies, making it more difficult to identify regular spatial variations in risk. Also, the risk of mortality from cancer increases with age. This creates two complications. First, a high mortality rate in a particular area amongst the elderly does not necessarily indicate a health concern: it could indicate that a higher proportion of women are living to an old age before eventually succumbing to cancer (i.e., it could in a paradoxical sort of way indicate low risk). Second, it becomes necessary to take account of variations in age composition between areas, due to the fact that an area having an elderly population would be expected, all other things being equal, to experience more fatalities from cancer than an area with a younger population — i.e., the crude death rate does not necessarily provide a reliable indication of the underlying risk.

The first problem is circumvented in this study by considering only "premature" deaths (defined here as deaths below the age of 65). This reduces the number of deaths under consideration from 3,481 to 1,061. This solves the interpretation problems caused by deaths amongst the elderly, but intensifies the difficulties caused by the small numbers problem. The second problem is normally resolved by calculating either age-specific rates for each age-group or by calculating a weighted age-standardised index. Clayton and Kaldor (1987) outline a procedure for extending the simple empirical Bayes method used in the first case study to take account of variations in age composition. This may be regarded as an empirical Bayes equivalent of an indirectly stanardised rate. The procedure requires a relative risk to be estimated for each age group in addition to the estimation of a relative risk for each area. The two sets of relative risks are estimated simultaneously using an iterative approach.

Figure 5 graphs the standardised mortality ratio (calculated by the indirect method) for each area against the female population aged less than 65 years.[4] It also shows the empirical Bayes estimate of the relative risk for each area. As before, the empirical Bayes estimates for each area are

4. Standardised mortality ratios are conventionally multiplied by a scaling factor of 100 to facilitate easier interpretation. Values larger than 100 indicate above average mortality, whilst values lower than 100 indicate below average mortality. However, this convention is not followed here to facilitate direct comparisons with the empirical Bayes estimates. Values larger than 1.0 indicate above average mortality.

invariably closer to the overall mean than the corresponding standardised mortality ratio. The largest differences are generally in the areas with the smallest populations. The standardised mortality ratios range from a low of 0.29 (Limerick CB) to a high of 1.61 (Monaghan), whereas values for the empirical Bayes estimates range from a low of 0.79 (Limerick CB) to a high of 1.18 for Dublin CB (Table 4). The empirical Bayes estimates intuitively seem more plausible. The Limerick values, for example, are based on only 5 deaths in 1989, whereas there were 19 deaths in Limerick in 1990, suggesting that the exceptionally low number of deaths (and correspondingly low standardised mortality ratio) in 1989 may have been a freak occurrence.

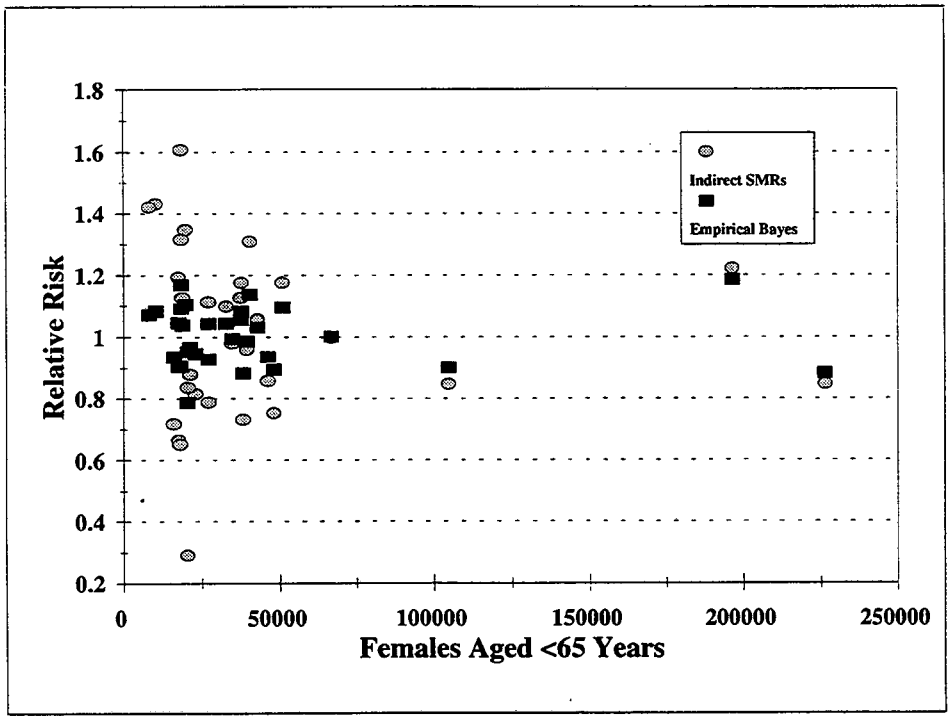

Figure 5: *Standardised Mortality Ratios and Empirical Bayes Estimates of Relative Risk For Female Cancer Deaths in 1989 Graphed Against the Population at Risk.*

The spatial distribution of the standardised mortality ratios is shown in Figure 6. The spatial distribution of the empirical Bayes estimates is shown in Figure 7. Neither map exhibits a convincing degree of spatial order, suggesting either that more data may be required to identify the underlying spatial regularities in risk or else that such regularities do not in fact exist for cancer taken as a whole.

Using the 1989 results to predict the expected number of cases in 1990

Table 4: *Standardised Mortality Ratios and Empirical Bayes Estimates of Female Cancer Mortality in 1989 (Relative Risks)*

County	*Standardised Mortality Ratio*	*Empirical Bayes Estimate*
Carlow	1.19	1.05
Dublin CB	1.22	1.18
Dublin County	0.85	0.88
Kildare	0.75	0.89
Kilkenny	0.79	0.93
Laois	1.13	1.04
Longford	1.43	1.08
Louth	0.98	0.99
Meath	0.96	0.98
Offaly	0.88	0.97
Westmeath	0.81	0.94
Wexford	1.18	1.08
Wicklow	1.12	1.06
Clare	1.10	1.04
Cork CB	1.17	1.09
Cork County	0.85	0.90
Kerry	1.05	1.03
Limerick CB	0.29	0.79
Limerick County	1.31	1.14
Tipperary NR	0.83	0.95
Tipperary SR	1.11	1.04
Waterford CB	0.72	0.93
Waterford County	1.32	1.09
Galway	1.00	1.00
Leitrim	1.42	1.07
Mayo	0.73	0.88
Roscommon	0.66	0.90
Sligo	1.35	1.10
Cavan	0.65	0.90
Donegal	0.86	0.93
Monaghan	1.61	1.17

results in similar conclusions to the previous case study (Table 5). As before, the conventional method produces poorer predictions than the assumption of no systematic variation in risk between areas, highlighting the vulnerability of conventional indices to stochastic fluctuations.[5] However, the empirical Bayes estimates represent a substantial improvement over each of the other two methods.

5. The "equal risk" predictions assume an equal risk between areas, but do not assume an equal risk between age groups — i.e., the predicted number of deaths in 1990 takes account of variations in age composition between areas.

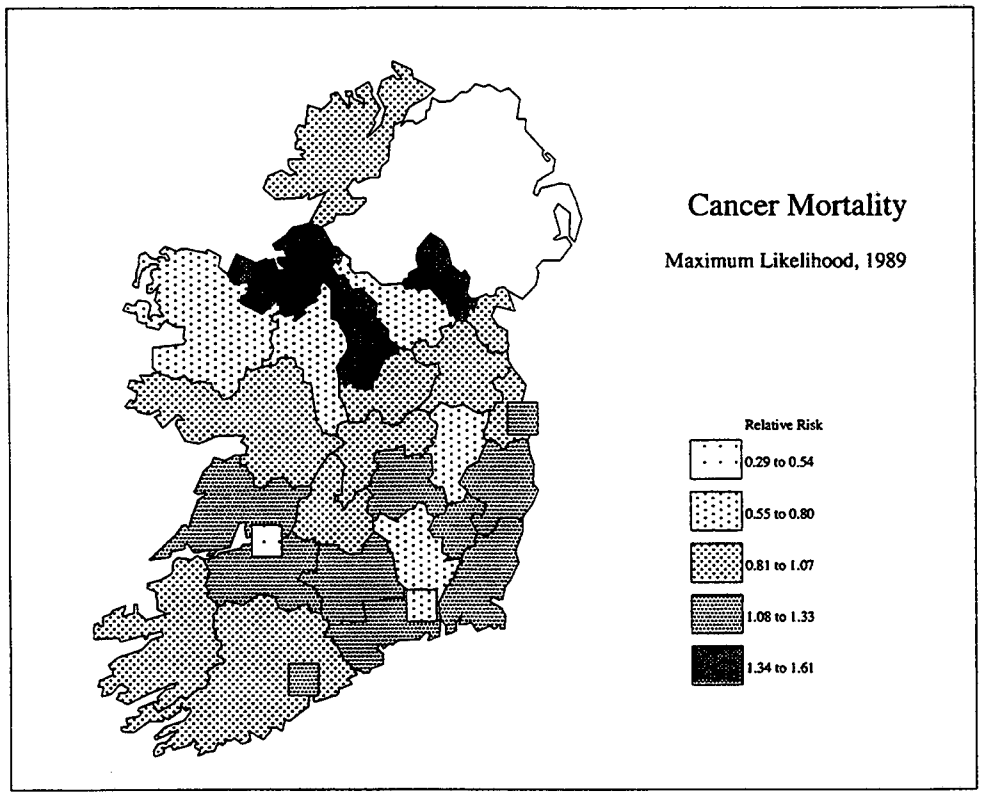

Figure 6: *Spatial Distribution of Stanardised Mortality Ratios for Female Cancer Deaths in 1989.*

Table 5: *Relative Accuracy of Three Predictions of Female Cancer Deaths in 1990*

	Mean Absolute Error
Method 1 — Equal Risk	6.35
Method 2 — Standardised Mortality Ratios	6.46
Method 3 — Empirical Bayes Estimates	5.09

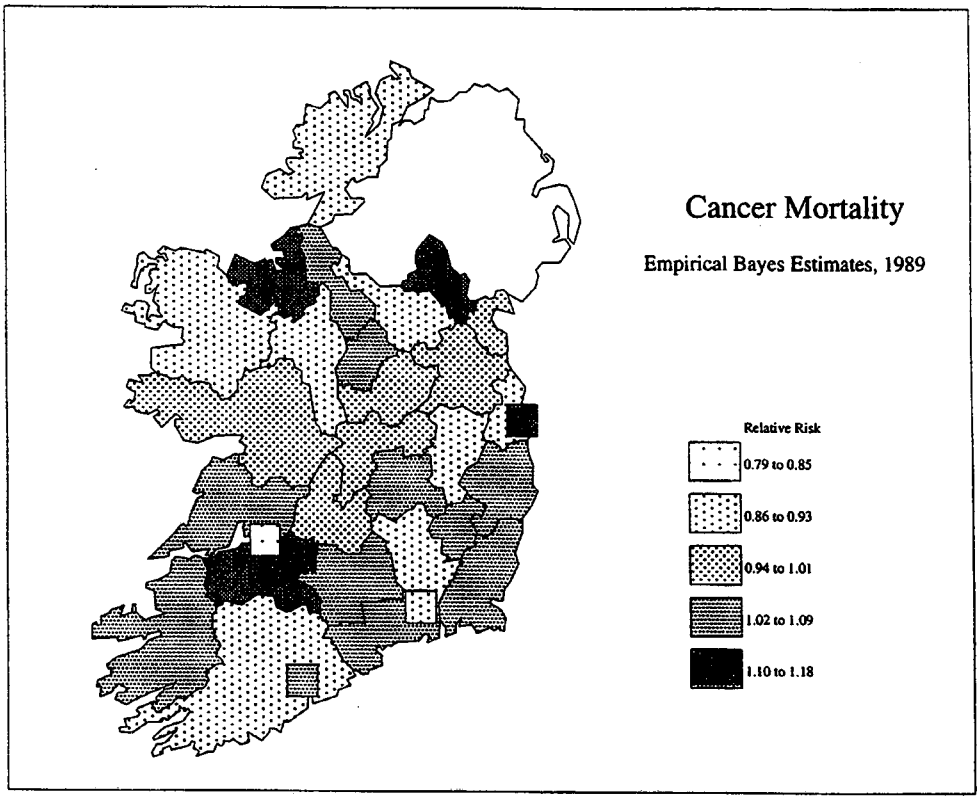

Figure 7: *Spatial Distribution of Empirical Bayes Estimates of Relative Risk for Female Cancer Deaths in 1989.*

V DISCUSSION

Epidemiologists often have to estimate the risk of disease from limited information. Conventional maximum likelihood methods provide reliable estimates if there are a large number of cases of the disease, but they can provide very misleading estimates when the number of cases is small because of stochastic fluctuations. Several biostatisticians have suggested that empirical Bayes techniques may provide more reliable estimates. The two case studies reported here would appear to support these claims: the estimates derived by empirical Bayes methods *subjectively* appear to provide more plausible estimates than their maximum likelihood equivalents. More importantly, they *objectively* provide better predictions of the expected numbers of deaths in 1990 than the method using maximum likelihood estimates or the method assuming equal risk. Indeed, the poor performance

of the maximum likelihood estimates in these tests highlights just how serious the small numbers problem can be.

These findings suggest that, where possible, empirical Bayes estimates should be used as an alternative (or at least as a supplement) to conventional maximum likelihood estimates of disease risk. Although the present paper discusses the problem of estimating disease risk in the context of disease mapping, similar arguments apply to estimates of disease risk for other purposes (e.g., investigations of temporal trends). Likewise, Bayesian approaches may provide improved estimates for non-medical data. Cressie (1995), for example, outlines methods for estimating the true rates for socio-economic variables (e.g., unemployment rates, per capita incomes) from "noisy" observed data, paying particular attention to the problem of estimating census undercounts. The major difference between these situations and the medical examples discussed above is that the source of uncertainty is assumed to be measurement errors rather than stochastic variability: the uncertainty is therefore assumed to be Gaussian (i.e., Normally distributed) rather than conforming to a Poisson process.

The case studies reported above assume comparatively simple models. It should perhaps be noted, however, that more complex models have been proposed to take account of spatial structures, such as autocorrelation and/or regionalisation (e.g., Clayton and Kaldor, 1987; Marshall, 1991). In other words, they allow one to build an assumption into the empirical Bayes estimates that the disease risks will exhibit spatial autocorrelation (i.e., that adjoining areas will tend to have more similar risks than distant areas) or that areas within defined regions will tend to have similar risks. Several authors have also suggested ways to incorporate covariates into the models to take account of suspected risk factors, such as pollution or urbanisation (e.g., Tsutakawa *et al.*, 1985; Clayton and Kaldor, 1987; Heisterkamp *et al.*, 1993; Cressie, 1995). The usefulness of these enhancements clearly depends upon whether the objective is to examine the data to identify empirical regularities or, having determined that these regularities exist, to model the data to make more accurate predictions, possibly for planning purposes.

REFERENCES

BAILEY, T.C., and A.C. GATRELL, 1995. *Interactive Spatial Data Analysis*, London: Longman.

CHOYNOWSKI, M., 1959. "Maps Based on Probabilities", *Journal of the American Statistical Association*, Vol. 54, pp. 585-588.

CLAYTON, D., and J. KALDOR, 1987. "Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping", *Biometrics*, Vol. 43, pp. 671-681.

CONGDON, P. 1990. "Issues in the Analysis of Small Area Mortality", *Urban Studies*, Vol. 27, pp. 519-536.

CRESSIE, N., 1993. "Regional Mapping of Incidence Rates Using Spatial Bayesian Models", *Medical Care*, Vol. 31, No. 5, Supplement, YS60-YS65.

CRESSIE, N., 1995. "Bayesian Smoothing of Rates in Small Geographic Areas", *Journal of Regional Science*, Vol. 35, No. 4, pp. 659-673.

DAWSON, D.F., 1911. "The Relation Between the Geographical Distribution of Insanity and that of Certain Social and Other Conditions in Ireland", *Journal of Mental Science*, Vol. 57, pp. 571-597.

DIEHR, P., 1984. "Small Area Statistics: Large Statistical Problems", *American Journal of Public Health*, Vol. 74, pp. 313-314.

DUNCAN, C., K. JONES, and G. MOON, 1993. "Do Places Matter? A Multi-level Analysis of Regional Variations in Health-related Behaviour in Britain", *Social Science and Medicine*, Vol. 37, pp. 725-733.

GIGGS, J.A., D.S. EBDON, and J.B. BOURKE, 1980. "The Epidemiology of Primary Acute Pancreatis in the Nottingham Defined Population Area", *Transactions of the Institute of British Geographers*, Vol. 5, pp. 229-242.

HEISTERKAMP, S.H., G. DOORNBOS, and M. GANKEMA, 1993. "Disease Mapping using Empirical Bayes and Bayes Methods on Mortality Statistics in The Netherlands", *Statistics in Medicine*, Vol. 12, pp. 1,895-1,913.

HOWELL, F., M. O'MAHONY, J. DEVLIN, O. O'REILLY, C. BUTTANSHAW, 1993. "A Geographical Distribution of Mortality and Deprivation", *Irish Medical Journal*, Vol. 86, pp. 96-99.

IVERSON, G.D., 1984. *Bayesian Statistical Inference*, Beverly Hills: Sage Publications.

JOHNSON, Z., and P. DACK, 1989. "Small Area Mortality Patterns", *Irish Medical Journal*, Vol. 82, pp. 105-108.

KALDOR, J., and D. CLAYTON, 1989. "Role of Advanced Statistical Techniques in Cancer Mapping", *Recent Results in Cancer Research*, Vol. 114, pp. 87-98.

KING, P.E., 1979. "Problems of Spatial Analysis in Geographic Epidemiology", *Social Science and Medicine*, Vol. 13D, pp. 249-252.

LANGFORD, I.H., 1994. "Using Empirical Bayes Estimates in the Geographical Analysis of Disease Risk", *Area*, Vol. 26, pp. 142-149.

MANTON, K.G., E. STALLARD, M.A. WOODBURY, W.B. RIGGAN, J.P. CREASON, and T.J. MASON, 1987. "Statistically Adjusted Estimates of Geographic Mortality Profiles", *Journal of the National Cancer Institute*, Vol. 78, pp. 805-815.

MARSHALL, R.J., 1991. "Mapping Disease and Mortality Rates Using Empirical Bayes Estimators", *Applied Statistics*, Vol. 40, pp. 283-294.

MIYAWAKI, N., and S.C. CHEN S.C., 1981. "A Statistical Consideration on the Mapping of Disease", *Social Science and Medicine*, Vol. 15, D, pp. 93-101.

MONMONIER, M., 1991. *How To Lie With Maps*, Chicago: University of Chicago Press.

PRESSAT, R., 1978. *Statistical Demography*, London: Methuen.

PRINGLE, D.G., 1983. "Mortality, Cause of Death and Social Class in the Belfast Urban Area", *Ecology of Disease*, Vol. 2, pp. 1-8.

PRINGLE, D.G., 1986. "Premature Mortality in the Republic of Ireland, 1971-1981", *Irish Geography*, Vol. 19, pp. 33-40.

PRINGLE, D.G., 1987. "Health Inequalities in Dublin", in A.A. Horner and A.J. Parker, (eds.), *Geographical Perspectives On The Dublin Region*, Dublin: Geographical Society of Ireland.

STONE, R.A., 1988, "Investigations of Excess Environmental Risks Around Putative Sources: Statistical Problems and a Proposed Test", *Statistics in Medicine*, Vol. 7, pp. 649-660.

TSUTAKAWA, R.K., G.L. SHOOP, and C.J. MARIENFELD, 1985. "Empirical Bayes Estimation of Cancer Mortality Rates", *Statistics in Medicine*, Vol. 4, pp. 201-212.

TSUTAKAWA, R.K., 1988. "Mixed Model for Analyzing Geographic Variability in Mortality Rates", *Journal of the American Statistical Association*, Vol. 83, No. 401, pp. 37-42.

WHITE, R.R., 1972. "Probability Maps of Leukaemia Mortalities in England and Wales", in N.D. McGlashan (ed.), *Medical Geography: Techniques And Field Studies*, London: Methuen.