LETTERS TO THE EDITOR

Use of deprivation indices in small area studies

SIR - I would like to make a contribution to the wide ranging debate currently focussed on small area health data and deprivation indices. This contribution is in response to the supplement *Use of deprivation indices in small area studies of environment and health*,¹ which focussed on these issues.

I have a number of concerns about issues discussed in three of the papers in this supplement²⁻⁴ and my comments mainly focus on statistical issues related to general ecological modelling and, in particular, to putative source modelling. The statistical issues of concern are of great relevance to epidemiological interpretation and hence I feel that this Journal is an appropriate forum for these concerns.

A recent comprehensive review of these statistical issues appears in Lawson and Waller.⁵

Autocorrelation and heterogeneity

The issue of whether autocorrelation or spatially correlated heterogeneity should be included in the analysis of ecological problems is of great concern. John Bithell² noted that Lawson⁶ used spatial correlation in the analysis of small area counts. In other discussions and papers, a tacit assumption is made that case counts or region counts can be modelled as independent events.^{3,4}

My concern over the lack of focus on autocorrelation and heterogeneity is as follows:

- 1 Individuals may have independent responses to non-infectious diseases, and hence it could be appropriate to assume an independent likelihood model for events.

- 2 It is possible that in two ways, the individual are only conditionally independent, that is:

- (a) Unobserved heterogeneity in the environment, eg, unobserved covariates could lead to "apparent" correlation or clustering in the observed data.

- (b) Disease which "naturally" clusters due to possible genetic or even viral aetiology will display unconditional dependence, ie, cases will be more likely to be found near others.

However, once the cluster structure is known, then the events could be regarded as conditionally independent.

The impact of both of these considerations, after adjustment for "expectation", is that we should expect autocorrelation in ecological studies. This particularly applies to putative source studies where only a small set of explanatory spatial variables is examined (eg, distance only, as in Diggle and Elliott³). This argument applies equally to regionalised counts and case events.

In addition, it may be thought more important to include random effects in case event data, for two reasons. Firstly, the use of a residential location as a surrogate for exposure has many drawbacks. It is unlikely that the bulk of any population spends even the majority of its daily hours at the place of residence. Aggregation to counts at least

yields larger sample sizes for inference purposes, and some safeguard against such "random" effects as journey to work/school/shop. Most analyses of case events data only examine the locations and do not look at case histories. Hence, any benefit of analysing "exact" locations is lost by not exploiting the individual information which could make allowance for the inevitable individual variability in frailty. The approach proposed in Diggle's paper³ further suggests that disease point maps should be used as controls for such case event analysis. The Diggle and Rowlingson approach, which has significant advantages over earlier approaches which used background smoothing, is only available when such point maps are available. This approach cannot be advocated as a general panacea as it forces the use of point maps for both cases and controls without the use of case specific information.

Mixed aggregate levels

When mixed levels of aggregation are available (eg, case point map, but "expected" number of cases in census regions) then a hybrid model is possible. Lawson and Williams⁷ proposed two approaches to mixed levels of aggregation.

General putative source analysis

Most workers who have to deal with putative source problems consider a range of evidence which could be indications of a link between small area data and a putative source. A short list of items might be:

- 1 Decline of disease intensity with distance from source.
- 2 Distance to maximum intensity.
- 3 Angular effects (related to wind direction).
- 4 Topographic effects.

In the papers and discussion in this issue only 1. is considered. While evidence of distance decline is of great importance, in most applications, the other effects are of great significance, particularly when dealing with putative air pollution sources.

This emphasis on 1. alone can lead to serious problems of statistical interpretation. I will allude to this in the next section.

Comments in the section on heterogeneity are particularly important in putative source applications for the simple reason that when single effects (such as distance only) are fitted there is likely to remain considerable structure in the data unexplained by the model. This unexplained variation can considerably reduce the goodness-of-fit of any model (whether for case events or regionalised counts). In addition, the possibility that clustering is found in the case disease has not been addressed. Leukaemias are known to cluster and have been studied around putative sources (nuclear power stations).

Risk models in putative source analysis

There are specific issues concerning the approach to risk models that are advocated in Diggle and Elliott,³ which have not been raised, which are related to the concerns discussed above, and can have a significant effect on epidemiological interpretation of statistical results.

In Diggle and Elliott,³ the model specified to describe health risk around a source is given as:

$$f(d) = 1 + \Theta \exp(-\Phi d^2)$$

where d is the distance from the source and Θ and Φ are parameters. This model represents a monotone distance decline from

the source and with the squared distance component, makes the distance relation "normal-like".

The authors state that: "We emphasise that the particular algebraic form of $f(d)$ is neither crucial nor compelling". However, this form is used throughout the paper, and further, no practical or theoretical reasons are given for its use. On the other hand there are a number of reasons why this model is *not* appropriate.

- 1 The model does not include any directional effects. These have been included in alternative models (see, eg Lawson and Waller⁵ for a review of such models).

- 2 The fitting of a two parameter model (with only one spatial parameter $\{\Phi\}$, could lead to further problems:

Firstly, the true exposure risk may not be monotone, but could increase before declining with distance. This result is both predicted by the theory of diffusion around stacks,⁸ and by a variety of empirical studies of atmospheric dispersion.⁹ Both these study areas suggest that the resulting distribution of outfall will form a concentric ring whose structure in any given direction will depend on the prevailing wind regime, and will not produce a monotonic decline in outfall. Rather an outfall peak occurs at some distance from the source, the location and size of the peak depending on direction. Given that in most retrospective putative sources studies, the exact extent of outfall effects is unknown, then the assumption of monotonic risk is inappropriate.

Even if a monotone outfall model were appropriate, the relation between dispersal of pollution and health risk may not be linear. It is possible that a distance threshold exists, which has to be reached before risk exists. In that case a marked peak of risk would occur at a distance from the source.

In addition, there is a potentially serious consequence of the assumption of monotonic risk and/or independence. It has been noted by Besag and Newell¹⁰ and also Lawson¹¹ that disease which naturally cluster will render tests for monotonic effects with low power. This is also true when the underlying risk is not monotone.

Score tests and related maximum likelihood estimation for a distance-only model (whether modelled as squared distance or distance) will not yield an accurate depiction of the risk behaviour around a source. The tests will be very conservative and the ML estimates likely to have unnecessarily high variance. This lack of accuracy could be improved by more detailed modelling of exposure risk based on outfall (in the air pollution case), or other characteristics.

ANDREW LAWSON
School of Informatics,
University of Abertay Dundee,
Dundee.

- 1 St Ledger S, ed. Use of deprivation indices in small area studies of environment and health. *J Epidemiol Community Health* 1995;49(Suppl 2):S2-S87.
- 2 Bithell JF, Dutton SJ, Neary NM, Vincent TJ. Controlling for socioeconomic confounding using regression methods. *J Epidemiol Community Health* 1995;49(Suppl 2):S15-S19.
- 3 Diggle P, Elliott P. Disease risk near point sources: statistical issues for analyses using individual or spatially aggregated data. *J Epidemiol Community Health* 1995;49(Suppl 2):S20-S27.
- 4 Bentham G, Eimermann J, Haynes R, Lovett A, Brainard J. Limiting long term illness and its associations with mortality and indicators of social deprivation. *J Epidemiol Community Health* 1995;49(Suppl 2):S57-64.
- 5 Lawson AB, Waller L. A review of point pattern methods for spatial modelling of events around

- sources of pollution. *Environmetrics* 1996 (in press).
- 6 Lawson AB. On using spacial Gaussian prior to model heterogeneity in environmental epidemiology. Proceedings of the Bayesian statistics conference, Nottingham. *The Statistician* 1994;43:69–76.
 - 7 Lawson AB, Williams FLR. Arndale a case study in environmental epidemiology. *Journal of the Royal Statistical Society A* 1994;157:28.
 - 8 Panofsky MA, Dutton JA. *Atmospheric turbulence*. Chichester: Wiley, 1984.
 - 9 Pasquill F. *Atmospheric diffusion*. 2nd ed. Chichester: Wiley, 1974.
 - 10 Besag J, Newell J. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society A* 1991;154:143–55.
 - 11 Lawson AB. On the analysis of mortality events around a prespecified fixed point. *Journal of the Royal Statistical Society A* 1993;156:363–77.

Reply

Lawson's letter makes a number of points concerning the importance of spatial autocorrelation in the analysis of geographical disease data with which I cannot agree. This is an area where generalisation may be dangerous, so I would certainly not wish to dispute the possibility that such autocorrelation may be important in some instances. But by the same token, I believe that there are others where it is not of major concern. I restrict my remarks to diseases for which cases may be taken to occur independently of one another conditionally on any underlying variation in risk; the appropriate underlying error model must then be the Poisson process, although we will not necessarily know the true risk at a given point.

In the first place, unrecognised spatial autocorrelation presumably has the effect – at least within the framework of frequentist inference and modelling – of leading to heterogeneity that might be recognised, for example, by a high deviance residual to a fitted model. However, in this respect this heterogeneity is like that resulting from any other factor that has not been properly taken into account in a given model. The crucial question is surely whether this over dispersion leads to misleading conclusions about the factors that have been fitted. Whether it does so will, broadly speaking, depend on whether these factors are themselves associated with whatever spatially varying factors are responsible for the over dispersion. If they are, then of course such factors may be regarded as confounding the effects we are interested in. However, we can never eliminate the possibility of confounding factors anyway; it seems to me to be of limited inferential value to know that such confounding – if it occurs – is mediated through an unidentified spatially varying variable. It is, of course, scientifically much more valuable to try and identify all the factors that matter.

Lawson claims that the heterogeneity induced by spatial autocorrelation is “particularly important in putative source applications”. I believe this to be most unlikely. For here the contributions of cases to neighbouring intervals on the distance axis will come in large measure from small areas which are relatively far apart, so that the serial correlation on this axis should be expected to be substantially less than the spatial autocorrelation per unit distance prevailing in two dimensions.

Likewise, the claim that “Leukaemias are known to cluster . . . around putative sources (nuclear power stations)” is highly suspect. There is very little published evidence relating to adult leukaemias, while, for childhood

leukaemia in England and Wales, the evidence was recently reviewed in the *British Medical Journal*¹ and found to be largely negative, with the exception of the data relating to Sellafield. The latter excess is extremely hard to interpret because of the post hoc nature of the observation; numerous papers have addressed the possible aetiology, but it seems clear that a purely geographical explanation is far from convincing.

In fact, contrary to numerous claims in the literature, childhood leukaemia does not exhibit strong spatial auto-correlation. The residual deviance exhibited in my own contribution to the session addressed in Lawson's letter suggests that there is very little residual heterogeneity to explain,² while methodologically sound and intelligible attempts to demonstrate a spatial explanation for any such heterogeneity have generally produced only equivocal results.³ For adults, there tends to be substantially more heterogeneity, but the small scale spatial component of this again appears to be very weak.⁴

To assume that spatial auto-correlation is important and to build this assumption into a statistical analysis may at first sight appear to be a sensible defensive strategy. But there is a high price to pay for an unnecessarily elaborate model. Its interpretation becomes more difficult and it is likely that estimates of the parameters of primary interest become less precise and stable as we attempt to gain more information from modest data sets. Important questions of procedure optimality⁵ are also much harder to address. These are familiar ideas in applied statistics; the fact that they are hard to make precise in any given case does not make them less important. The onus is on the builder of a more complex model to demonstrate that a simpler one will not suffice.

J BITHELL
Department of Statistics,
University of Oxford

- 1 Bithell JF, Dutton SJ, Draper GJ, Neary NM. Distribution of childhood leukaemias and non-Hodgkin's lymphomas near nuclear installations in England and Wales. *BMJ* 1994; 309:501–5.
- 2 Bithell JF, Dutton SJ, Neary NM, Vincent TJ. Use of regression methods for control of socioeconomic confounding. *J Epidemiol Community Health* 1995;49(Suppl 20):S15–S19.
- 3 Draper G, ed. *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966–8*. Studies on medical and population subjects, 53. London: HMSO, 1991.
- 4 Cartwright RA, Alexander FE, McKinney PA, Rickets TJ. *Leukaemia and lymphoma: an atlas of distribution within areas of England and Wales, 1984–1988*. London: Leukaemia Research Fund, 1990.
- 5 Bithell JF. The choice of test for detecting raised disease risk near a point source. *Stat Med* 1995; 14:21.

Reply

Lawson raises a number of interesting issues, but in our opinion misrepresents what we say in our paper in several respects.

1 We agree that the assumption of an underlying Poisson point process may be violated in practice, which is why we refer to this model as an idealised point process model. We discuss this issue explicitly in the last two paragraphs of our paper.

2 We agree that residential location is an imperfect surrogate for the epidemiologically relevant location or locations of an individual, but again we say this explicitly in the paper.

Also, dealing with counts in regions hides, rather than resolves, the difficulty while introducing other potential problems, most obviously the well documented one of ecological bias. Nowhere in our paper do we suggest that our approach (or any other single approach) is a “general panacea”; we seek only to clarify some of the relevant practical issues by deliberately working within an idealised theoretical framework.

3 The problem of “mixed aggregate levels” is much more subtle than simply being able to deal with point cases and aggregated population data (controls). In our opinion, much work remains to be done in developing practical solutions to this problem.

4 Models more complex than our simple equation (8) would undoubtedly be justified if sufficient, and sufficiently precise, data were available for their validation. It does not follow that such models would give better inferences for the necessarily limited and imprecise data available in many applications. We have proposed a general model for decline in risk with distance around the source that approximates to the “threshold” or “peaked exposure” (eg around a stack) scenarios discussed by Lawson (Diggle *et al*, unpublished data).

While recent methodological developments in complex statistical modelling open up exciting possibilities, it should not be forgotten that all models are wrong, and that parsimony is a sound guiding principle.

P DIGGLE
Department of Mathematics and Statistics,
University of Lancaster.
P ELLIOTT
London

The decline in sex ratios at birth, England and Wales, 1973–90

SIR – Dickinson and Parker entitled the letter which was published in this journal,¹ “Why is the sex ratio falling in England and Wales?”. The short answer to this question is that the sex ratio does not seem to be falling still. It fell more or less continuously from 1973–90, but in the four succeeding years it rose; and in 1993 it reached the highest level since 1983. It is interesting, however, to consider the causes of the fall during the 1970s and 1980s. Dickinson and Parker¹ mention three possibilities – coital rate, pollution, and hormones – and I should like to comment on these.

Coital rate

The sex ratios of white births in the USA between 1915 and 1988 seemed to show unexplained oscillations across intervals of very roughly 20–30 years.² In particular, they rose during the years 1965–75 and fell during the years 1975–88. In that paper,² I cited evidence that age standardised reported marital coital rates increased by 22% between 1965 and 1975 and fell by 27% between 1975 and 1988. Moreover, I also reviewed there the hypothesis that parental coital rate (weakly) controls the offspring sex ratio.² The evidence seems strong so it appears that in the USA from 1965–88 variations in the sex ratio might be explained by secular variations in the coital rate.

The recent decline in the sex ratio in England and Wales parallels a decline in the USA across the same years. It is possible that the