STATISTICS IN MEDICINE
Statist. Med. **18**, 3221–3234 (1999)

ALL MAPS OF PARAMETER ESTIMATES ARE MISLEADING

ANDREW GELMAN[1] AND PHILLIP N. PRICE[2] *

[1] *Department of Statistics, Columbia University, 618 Mathematics Building, New York, New York 10027, U.S.A.*
[2] *Lawrence Berkeley National Laboratory, LBNL 90-3058, Berkeley, California 94720, U.S.A.*

SUMMARY

Maps are frequently used to display spatial distributions of parameters of interest, such as cancer rates or average pollutant concentrations by county. It is well known that plotting observed rates can have serious drawbacks when sample sizes vary by area, since very high (and low) observed rates are found disproportionately in poorly-sampled areas. Unfortunately, adjusting the observed rates to account for the effects of small-sample noise can introduce an opposite effect, in which the highest adjusted rates tend to be found disproportionately in well-sampled areas. In either case, the maps can be difficult to interpret because the display of spatial variation in the underlying parameters of interest is confounded with spatial variation in sample sizes. As a result, spatial patterns occur in adjusted rates even if there is no spatial structure in the underlying parameters of interest, and adjusted rates tend to look too uniform in areas with little data. We introduce two models (normal and Poisson) in which parameters of interest have *no* spatial patterns, and demonstrate the existence of spatial artefacts in inference from these models. We also discuss spatial models and the extent to which they are subject to the same artefacts. We present examples from Bayesian modelling, but, as we explain, the artefacts occur generally. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

1.1. Background

When a spatially-varying parameter of interest is subject to substantial uncertainty, maps of predicted values can differ in important and systematic ways from the spatial distribution of true values. A standard method for correcting for these artefacts – Bayes shrinkage estimation – introduces new and opposite artefacts of its own.

We will illustrate this point with generic statistical models, but it is helpful to keep a specific example in mind. Consider the mapping of cancer mortality rates by county in the United States. Much of the variation in observed cancer death rates by county is attributable to statistical noise due to the small number of (observed and expected) cancer deaths in low-population counties. Because of this stochastic noise, a disproportionate fraction of low-population counties are observed to have extremely high (or low) cancer rates when compared to typical counties in the United States. Thus

* Correspondence to: Phillip N. Price, Lawrence Berkeley National Laboratory, LBNL 90-3058, Berkeley, California 94720, U.S.A. E-mail: pnprice@lbl.gov

Contract/grant sponsor: U.S. National Science Foundation
Contract/grant numbers: DMS-9404305, SBR-9708424, DMS-9457824
Contract/grant sponsor: U.S. Department of Energy
Contract/grant numbers: DE-AC03-76SF00098

CCC 0277–6715/99/233221–14$17.50

when counties with very high observed rates are highlighted on a map of the U.S., almost all of the highlighted counties are low-population counties.[1, 2] Since the Central and Western U.S. contain a great many such counties, a much higher fraction of counties in the Central and Western U.S. is highlighted than in the rest of the country.

Manton *et al.*[1] and Riggan *et al.*[3] use a Bayesian procedure to estimate underlying cancer rates by county. This procedure is now common, with minor variations, for U.S. cancer maps.[4, 5] The posterior mean estimate for each county is a compromise between the observed county cancer mortality rate and the mean cancer mortality rate for the entire U.S. (or for a region of the U.S.[5]), with the relative weighting of these rates being dependent on the county population. The estimated underlying cancer death rate for a high-population county with a given observed rate is close to the observed value, and this estimate has a small standard error. A low-population county with the same observed rate has a posterior mean somewhere between the observed rate and the U.S. mean rate, with a larger standard error.

Manton *et al.*[1] quite reasonably suggest that the posterior mean estimates are more appropriate for mapping than are the observed death rates, since the observed rates are subject to systematic effects related to county population, and since Bayes and empirical Bayes methods tend to yield more accurate predictions than do raw rates.[6–8] Unfortunately, the posterior means are subject to a similar type of systematic artefact related to county population, but in the opposite direction, as we will show. (Similar problems with the ensemble of posterior mean estimates are noted by Louis.[9]) We also show that most other mapping methods have artefacts associated with populations or sample sizes.

We quantify these artefacts in this paper, using the examples of standard models for continuous and discrete data to demonstrate that maps of point estimates can introduce spurious spatial patterns. This occurs even when the model being fit is appropriate, and even when there is *no* underlying spatial structure in the parameter of interest. In Section 2 we consider an example with normally-distributed parameters and measurements. In Section 3, we examine a Poisson/gamma model with parameters taken from cancer data. In Section 4, we discuss the occurrence of artefacts in fitting spatial models to data that *do* have underlying spatial structure.

1.2. Theoretical approach to examining statistical artefacts

If one fits a statistical model that is inappropriate to the data being analysed, then inferences will be incorrect and maps of predictions might well show spurious spatial patterns. This is *not* what we mean by 'spatial artefacts' in this paper. Instead, we consider a spurious spatial pattern to be an 'artefact' if it occurs even when inferences are based on the *correct* statistical model.

We analyse mapping artefacts in the context of a theoretical model with *no* spatial effects. This approach allows us to illustrate our points with simple and easily interpreted statistical models, and makes it easy to see the effects of the artefacts in our sample maps since any apparent spatial pattern is an artefact. As we discuss in Section 4, the same sorts of artefacts occur when the parameter of interest varies spatially, even if the correct spatial model is fit.

Under our model, each of J counties, $j = 1, \ldots, J$, has an unknown parameter θ_j. The ensemble of parameters, $\{\theta_1, \ldots, \theta_J\}$, follows some distribution, $p(\theta_j)$, assumed known. In each county j, we have n_j independent measurements y_{ij}, $i = 1, \ldots, n_j$, with a known sampling distribution: $y_{ij}|\theta_j \sim p(y_{ij}|\theta_j)$. We further assume, in this theoretical model, that the sample sizes n_j are statistically independent of the true parameter values θ_j (so that the values of n_j do not convey information about the θ_j's), and that the parameters θ_j are spatially uncorrelated.

We now suppose that a statistical analysis is performed and then a map is drawn to indicate the estimated value of θ_j in each county. Although this paper applies to parameter mapping in general, we will focus on maps that highlight only the counties with the highest point estimates, so that we can use black and white maps as illustrations. Colour or greyscale maps would manifest the same artefacts – for example, using a colour map to search for 'hot spots' of a parameter would be equivalent to looking at the highlighted counties in our black and white maps.

In our analysis, we ignore the difference between (a) highlighting the top x per cent of counties and (b) highlighting the counties that exceed a threshold that, in expectation, exceeds all but x per cent of the counties. In practice, both procedures are used.[1, 10] The two procedures give essentially the same result. For example, there is little difference between highlighting 27 out of 274 counties or using a fixed threshold so that the expected number of counties highlighted is 27·4. For mathematical simplicity, we consider procedure (b) in this paper.

If the map of extreme values is simply based on point estimates $\hat{\theta}_j$, this means that some threshold c is set so that the counties j for which $\hat{\theta}_j > c$ are highlighted. More generally, if one attempts to adjust for sample size then a function $h(\cdot,\cdot)$ is chosen and a threshold c is set so that the counties for which $h(y_j, n_j) > c$ are highlighted. The main point of this paper is that, for most mapping methods – that is, for most choices of $h(\cdot,\cdot)$ – the probability that a county is highlighted, $\Pr(h(y_j, n_j) > c | n_j)$, depends on the sample size n_j, so that the map of highlighted counties will display patterns based on the sample sizes.

2. CONTINUOUS MEASUREMENTS

We first work out the basic results for the relatively simple problem of continuous measurements with normally-distributed errors. For counties $j = 1, \ldots, J$, let θ_j be the true value of a parameter in county j. We assume that the true values of the county parameters, θ_j, follow a normal distribution:

$$\theta_j \sim \mathrm{N}(\mu, \tau^2). \tag{1}$$

By assuming the county parameters follow a common distribution, we are *not* assuming that the counties are identical – that would correspond to $\tau = 0$. The data from each county constitute n_j independent, identically distributed measurements

$$y_{ij} | \theta_j \sim \mathrm{N}(\theta_j, \sigma^2). \tag{2}$$

2.1. Problems with mapping the sample means

The direct estimate of each θ_j is the observed county mean, which we label y_j. It is well known that, if the n_j's vary, the procedure of selecting the counties with the highest observed means tends to yield counties with few observations; we will quantify this artefact. An observed county mean y_j based on n_j observations is distributed as

$$y_j \sim \mathrm{N}(\mu, \tau^2 + \sigma^2/n_j). \tag{3}$$

Under our model, the probability that the observed mean y_j exceeds a threshold c, for a county with sample size n_j, is

$$\Pr(y_j > c | n_j) = \Phi\left[\frac{\mu - c}{\tau}\left(1 + \frac{\sigma^2}{n_j \tau^2}\right)^{-1/2}\right]. \tag{4}$$

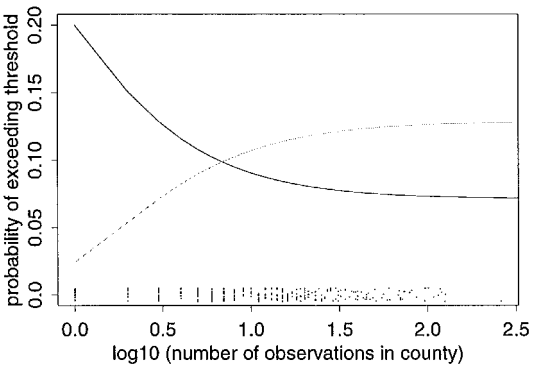

Fig. 1. Solid line: probability that an observed county mean, y_j, will exceed a specified cut-off point, c_1. Dotted line: probability that the posterior mean estimate for a county, $E(\theta_j|y_j)$, will exceed a specified cut-off point, c_2. Both lines are plotted as a function of the log (base 10) of n_j, the number of observations in the county. Each cut-off point is set to catch an average of 10 per cent of the counties. Curves are derived from the values of n_j in the radon data structure and from the variance ratio $\tau^2/\sigma^2 = 0{\cdot}49$ estimated from the radon data. The points at the bottom of the figure show the 274 values of $\log_{10} n_j$; they are jittered (see Chambers *et al.*[12]) so that duplicate values are visible

For any given threshold c, one can compute the expected number of counties that will be shaded under the model by summing the probabilities (4), for a given set of n_j's.

If the threshold c is to be set so that some small fraction of counties (for example, 5, 10, or 20 per cent) is expected to exceed it, then c will almost certainly be larger than the grand mean, μ, and the probability of exceeding it is a decreasing function of n_j. The variation of this probability with n_j depends on both the variance ratio σ^2/τ^2 and the value of c, which itself depends on the distribution of the J values of n_j.

To illustrate, we use the example of home radon levels in the mid-Atlantic region of the U.S., which comprises 277 counties, including some independent cities in Virginia. In this region, the Environmental Protection Agency and the state health departments randomly sampled 5677 homes;[11] three of the counties had no homes surveyed, and of the remaining counties, the number n_j of homes surveyed ranged from 1 to 261. The measurements y_{ij} are the natural logarithms of the measured radon levels, and the parameter θ_j is the average log radon level in county j (that is, the log geometric mean radon measurement that would be obtained if every home in the county were to be measured). We fit a hierarchical normal model[13, 14] to these data and obtained estimates of 1·0 and 0·7 for the within- and between-county standard deviations, σ and τ, respectively.

We study the artefacts created by the mapping procedure for the radon example by working out what would happen if the hierarchical normal-normal model were true, with hyperparameter values $\sigma = 1{\cdot}0$ and $\tau = 0{\cdot}7$. That is, we construct a model in which the statistical distribution of county radon levels is similar to that from the actual data, but in which (unlike the actual radon data) the county parameters are distributed randomly, with *no* spatial correlation. Under this model and the given set of 274 values of n_j, the cut-off value to highlight the top 10 per cent of counties is $c_1 = \mu + 1{\cdot}468\tau$.

The solid line on Figure 1 shows the probability that any given county mean y_j will exceed c_1, as a function of $\log_{10} n_j$. The points at the bottom of the figure show the values of $\log_{10} n_j$ in the data set. (Ignore the dotted line on the figure for now.) Counties with fewer than about six measurements are much more likely to exceed the threshold than are more heavily sampled

counties. This statistical artefact manifests itself as a spatial artefact in a map of county means, because the sample sizes themselves vary spatially.

2.2. Problems with mapping the posterior point estimates

It has been suggested[1] that one should map the county posterior mean estimates, $E(\theta_j|y_j,n_j)$, to avoid the artefact discussed above. Unfortunately, mapping county posterior means or highlighting the counties with highest posterior means leads to new problems. Under the normal model above, the posterior mean (and mode) estimate for a county is

$$E(\theta_j|y_j,n_j)=\frac{\dfrac{1}{\tau^2}\mu+\dfrac{n_j}{\sigma^2}y_j}{\dfrac{1}{\tau^2}+\dfrac{n_j}{\sigma^2}}. \tag{5}$$

Averaging over the marginal distribution of y_j, we find that, for a county with sample size n_j, the probability that $E(\theta|y_j,n_j)$ exceeds a fixed value c is

$$\begin{aligned}\Pr(E(\theta_j|y_j,n_j)>c|n_j)&=\Pr\left(y_j>\frac{\sigma^2}{n_j}\left[\left(\frac{1}{\tau^2}+\frac{n_j}{\sigma^2}\right)c-\frac{1}{\tau^2}\mu\right]\middle|\,n_j\right)\\&=\Phi\left[\frac{\mu-c}{\tau}\left(1+\frac{\sigma^2}{n_j\tau^2}\right)^{1/2}\right]\end{aligned} \tag{6}$$

an expression which is similar to (4) but is now an increasing, rather than decreasing, function of n_j (assuming $c>\mu$, which will be the case if the cut-off is set so that a small fraction of counties will be highlighted).

Mapping county posterior mean estimates (5) still leads to artefacts related to sample sizes, since $E(\theta_j|y_j,n_j)$ depends on n_j. For example, in the radon data much of West Virginia was sparsely sampled (values of n_j were low), so that a map of the posterior estimates of county means in West Virginia will appear quite uniform even if the true county levels θ_j are highly variable.

Under the assumed model, the threshold c that leads to an expected 10 per cent of the counties being highlighted is $c_2=\mu+1{\cdot}132\tau$, a lower value than the cut-off c_1 for the raw county means, which makes sense since the posterior mean estimates are shrunken towards the grand mean. (We computed c_2 by iteratively trying different values of c until the average value of (6), averaging over the counties j, was 10 per cent.) The dotted line in Figure 1 displays the probability of a county's posterior mean estimate exceeding c_2, as a function of $\log_{10}n_j$. Clearly, a map highlighting the posterior means has a strong artefact in the opposite direction to the map of the observed means; the counties with fewer than six observations are disproportionately *unlikely* to have notably high posterior means.

2.3. Problems with maps based on statistical significance

Other natural methods of mapping extreme counties also suffer from artefacts so that the probability of a county being highlighted depends on the number of observations in the county. For example, one could highlight the counties with the highest posterior probability of exceeding some specified level, $\mu+x\tau$. Under the normal model, this is equivalent to choosing the counties with the highest 'posterior z-scores', $z_j=(E(\theta_j|y_j,n_j)-(\mu+x\tau))/\mathrm{sd}(\theta_j|y_j,n_j)$. The resulting probability that a

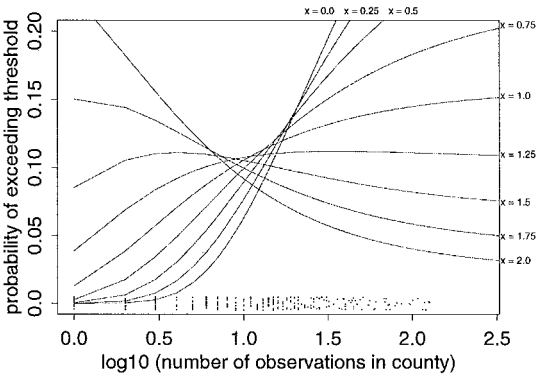

Fig. 2. Probability that a county will be in the top 10 per cent of counties as ranked by $\Pr(\theta_j > \mu + x\tau|y_j)$. Curves shown for $x = 0, 0{\cdot}25, 0{\cdot}5, \ldots, 2{\cdot}0$; each curve is plotted as a function of the log (base 10) of n_j, the number of observations in the county. Curves are derived from the values of n_j in the radon data structure and from the variance ratio $\tau^2/\sigma^2 = 0{\cdot}49$ estimated from the radon data. The points at the bottom of the figure show the 274 values of $\log_{10} n_j$

county with sample size n_j is highlighted is

$$\Pr(z_j > z_c|n_j) = \Phi\left[-\frac{z_c\sigma}{n_j^{1/2}\tau} - x\left(1 + \frac{\sigma^2}{n_j\tau^2}\right)^{1/2}\right] \tag{7}$$

where z_c is the cut-off z-score level set so that 10 per cent (say) of counties are highlighted. Expression (7) is dependent on n_j in a relatively complicated manner; note that z_c can be either positive or negative, depending on x and the data structure.

To illustrate, Figure 2 displays the probability that a county is highlighted, as a function of $\log_{10} n_j$, for each of several values of x, from $x = 0$ (corresponding to selecting the 10 per cent of counties with the highest posterior probability of $\theta_j > \mu$) to $x = 2$ (corresponding to selecting the 10 per cent of counties with the highest posterior probability of $\theta_j > \mu + 2\tau$), for the radon data structure. None of these curves is a constant function of n_j, but for $x = 1{\cdot}5$ the curve is close to flat, corresponding to a mapping procedure that is relatively free of artefacts due to sample sizes.

Should we, then, construct a map based on the rankings of the counties in terms of $\Pr(\theta_j > \mu + 1{\cdot}5\tau|y_j, n_j)$? We think not, because this is not a natural measure or ranking. In fact, the '1·5' depends on the structure of the data and would change if σ/τ or the set of n_j's were changed, so maps of different data (death rates from different cancer types, for instance) would require disparate ranking methods to avoid spatial artefacts. In addition, it is not clear what relevance such a measure as $\Pr(\theta_j > \mu + 1{\cdot}5\tau|y_j, n_j)$ would have to any questions of inherent scientific interest. Using such a measure would reduce artefacts due to sample sizes, but only at the expense of the ease of interpretability that is one of the reasons for producing maps in the first place.

A related approach to weeding out the highly variable small counties is to highlight the counties that are statistically significantly greater than the overall mean – in the normally-distributed case, this would mean $y_j > \mu + 2\sigma/n_j^{1/2}$. This method can be an improvement on merely mapping y_j (see, for example, Tufte,[15] pp. 16–19, who displays maps from Mason *et al.*[16] indicating both extreme values of y_j and statistical significance, and Schlattmann *et al.*[17], who map Bayes estimates indexed by statistical significance). However, as with all the other methods we have considered

so far, maps highlighting statistical significance do not eliminate artefacts based on sample size; if the sample size in a county is extremely large, even a small difference between the county's observed rate y_j and the mean rate μ will be statistically significant, so again this method is more likely to include a high population county than a low-population one with the same true parameter value.[18]

In the normal model, artefacts based on sample size can be eliminated by highlighting the counties for which the quantity $z_{\text{marg}} = (y_j - \mu)/(\tau^2 + \sigma^2/n_j)^{1/2}$, is highest. We label this the *marginal z-score*, because it measures the discrepancy of the county mean y_j with respect to its marginal distribution, averaging over the unknown county parameter θ_j. Under the assumed model, $\Pr(z_{\text{marg}} > c|n_j)$ is just the cumulative standard normal distribution evaluated at c and does not depend on n_j – thus, no artefacts due to sample size. (Incidentally, this works only for continuous data; any discreteness in the distribution of y_j causes the probabilities to vary with n_j.) A map of the extreme values of z_{marg} could be a useful kind of 'standardized residuals' plot. However, such a map still has the same problem as the other proposals mentioned in this section; the mapped values have no direct interpretation as estimates of θ_j. For example, the low-sample-size counties highlighted on such a map will have lower values of θ_j, on average, than the highlighted counties with high sample size.

2.4. Multiple imputation of posterior parameters

An alternative method of producing maps is to multiply impute the vector of posterior parameters. Multiple imputation[19, 20] is a method of accounting for the posterior uncertainty in a vector, $\theta = (\theta_1, \ldots, \theta_J)$ by drawing L simulations of the vector, θ^l, $l = 1, 2, \ldots, L$. This means drawing each vector θ from the posterior distribution $p(\theta|y)$. A map based on one simulated vector of county parameters $\theta^l = (\theta_1^l, \ldots, \theta_J^l)$ represents just one 'possible' reality. A multiple imputation yields several such maps, each based on a different draw of the vector of county parameters.

For example, if the highest counties were of interest, one could highlight on each map the 10 per cent of counties with highest values of θ_j in that simulation draw. Variation from map to map would show posterior uncertainty. Thus, a county for which no information is available would be highlighted on 1/10 of the maps (after all, it *could* be in the top 10 per cent of true county means); a county with many observations and a very high observed value would be highlighted in nearly all the maps; a county with few observations and a very high observed value would be highlighted on more than 1/10, but perhaps not most, of the maps; and so forth.

A multiply-imputed map does not suffer from the artefacts described in the previous sections. More precisely, *if* the model being applied is correct, and a map is made highlighting all counties with imputed θ_j values higher than some cut-off c, *then* the probability that a county is highlighted in any given imputation does not vary with the sample size, n_j. To see why this is so, notice that the probability that county j is highlighted in a single randomly-produced map, given the data y_j from that county, is just the posterior probability $\Pr(\theta_j > c|y_j, n_j)$. The probability that a particular county with sample size n_j is highlighted in a map, obtained by averaging over the marginal distribution of y_j, is

$$\int \Pr(\theta_j > c|y_j, n_j)\, p(y_j|n_j)\, \mathrm{d}y_j = \Pr(\theta_j > c|n_j) \tag{8}$$

which depends only on the distribution of true county parameters, $p(\theta_j)$, and not on the number of observations n_j. (Recall that we have assumed that θ_j and n_j are statistically independent.)

Of course, use of multiple imputation requires the production of multiple maps if one wishes to examine the spatial distribution and uncertainties of quantities of interest; any single map based on multiple imputation gives no indication of which spatial features are due to chance and which are strongly supported by the data, as we will discuss below in the context of multiple imputation of cancer maps.

3. COUNTED DATA

The occurrence of artefacts related to the amount of information in each map unit is a general result, but the details vary with the model and data structure. We illustrate the case of counted data with the Poisson/gamma model, which is commonly used in small area estimation with data such as cancer incidences; similar results would be obtained, with somewhat more computational effort, under the other standard family of models,[8] the Poisson/log-normal. For counties $j = 1, \ldots, J$, let θ_j be the underlying rate parameter, n_j be the population in county j, and

$$y_{ij} = \begin{cases} 1 & \text{if individual } i \text{ is affected} \\ 0 & \text{otherwise.} \end{cases}$$

Finally, we label the observed number of incidences in county j as $y_j = \sum_{i=1}^{n_j} y_{ij}$, so that the observed rate for the county is y_j/n_j. It is then standard to model y_j as a Poisson random variable with parameter $n_j\theta_j$. We further assume that the county parameters θ_j follow a gamma(α, β) distribution.

We illustrate with the data structure of the Manton *et al.*[1] example of ten-year kidney/ureter cancer rates in counties of the United States, with n_j equal to county populations, and with $\alpha = 20$ and $\beta = 20/(4{\cdot}65 \times 10^{-5})$. We chose these parameters so that the mean and variance of the gamma(α, β) distribution would approximately match the mean and variance of the county parameters in the Manton *et al.* paper. From this distribution, we draw a 'true' cancer rate for each county. We also assign an 'observed' rate, drawn from the Poisson $(n_j\theta_j)$ distribution for each county. As before, we do not use the data y_j from Manton *et al.*; rather, we model what would happen if the true values county parameters were drawn from a gamma distribution with the (approximately) correct scale and shape but independently of any spatial or other variables.

We consider the effects of highlighting counties based on the raw means, y_j/n_j, or the posterior means, which are given by

$$E(\theta_j | y_j, n_j) = \frac{\alpha + y_j}{\beta + n_j}. \tag{9}$$

As with the normal model, the mapping artefacts depend on the sample sizes (in this case, the populations), n_j, and the distribution of county rates, θ_j.

Figure 3 is the analogy, under the Poisson-gamma model, to Figure 1. The solid line in Figure 3 shows the probability that any given county mean, y_j/n_j, will exceed c_1, as a function of $\log_{10} n_j$, where $c_1 = 11{\cdot}2 \times 10^{-5}$ is the cut-off set so that one expects 10 per cent of the counties to be highlighted. (For any given threshold c, one can compute the expected number of counties that will be shaded under the model by simulation from the gamma and Poisson distributions. We arrived at the value 11·2 by iteratively altering c until the expected proportion of shaded counties was 10 per cent.) The dotted line shows the probability that any given posterior mean, $(\alpha + y_j)/(\beta + n_j)$, will exceed $c_2 = 5{\cdot}0 \times 10^{-5}$, the cut-off set so that one would expect 10 per cent of the counties

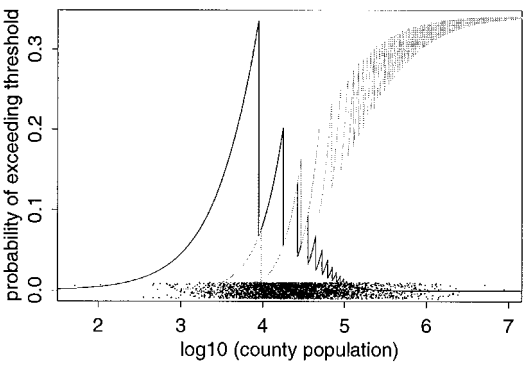

Fig. 3. Solid line: probability that an observed county cancer death rate, y_j, will exceed a specified cut-off point, c_1. Dotted line: probability that the posterior mean estimate for a county, $E(\theta_j|y_j)$, will exceed a specified cut-off point, c_2. Both lines are plotted as a function of the log (base 10) of n_j, the population in the county. Each cut-off point is set to catch an average of 10 per cent of the counties. Curves are derived from the values of n_j in U.S. counties and from the gamma $(20, 4{\cdot}3 \times 10^5)$ distribution fit to the Manton *et al.*[1] data. The points at the bottom of the figure show the 3082 values of $\log_{10} n_j$

to be highlighted under this method. (Recall that the grand mean of the θ_j's is assumed to be $4{\cdot}65 \times 10^{-5}$.) Given the cut-offs c_1 and c_2, we computed probabilities for the solid and dotted lines based on the marginal distribution for y_j, which is negative binomial. The points at the bottom of the figure show the 3082 values of $\log_{10} n_j$ for U.S. counties.

The sawtooth pattern of Figure 3 arises from the discrete nature of the data; for example for a map based on observed rates, a county with n_j in the range $[0, 1/c_1)$ will be highlighted if $y_j \geqslant 1$, whereas if n_j is in the range $[1/c_2, 2/c_1)$, at least two occurrences of cancer are required, and so forth. In addition to the sawtooth pattern, Figures 1 and 3 show different behaviours at the limits of small and large n.

Maps based on observed rates overemphasize the counties with small populations, but maps based on posterior mean have the reverse problem that the more populous counties are more likely to be highlighted. For the model discussed above, the average county population is 80,000, but the expected average population of the highlighted counties is 16,000 if highlighting is based on raw means or 190,000 if highlighting is based on posterior means.

Figure 4 displays the top 10 per cent of counties according to θ_j, for our simulated data; this is equivalent to a random sample of 10 per cent of U.S. counties. Figures 5(a) and (b) display the top 10 per cent of counties according to the observed rates and posterior means, respectively. The patterns – most notably, the presence of many counties from the Mountain and Plains states in the highest 10 per cent based on the observed rates, and the very small fraction of counties in those states in the maps of posterior means – are similar to Figures 1 and 2 of Manton *et al.*,[1] which plot the counties with highest observed rates and posterior means for kidney/ureter cancer death rates. This similarity suggests that many of the spatial patterns in that paper, and in maps of Bayes-smoothed cancer rates in general, are artefactual.

As in the previous example, we can avoid mapping artefacts by creating multiply imputed maps from the posterior distribution. Under the assumptions of the model, equation (8) holds – that is, the probability that a county is highlighted is independent of its population. We illustrate with the simulated-data example above; we sample from the posterior distribution of the vector of county

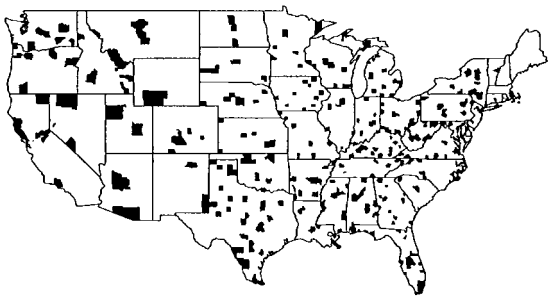

Fig. 4. Shaded counties are those in which the true county parameters θ_j are in the top 10 per cent of U.S. counties. Values of θ_j are drawn independently from a common distribution; this is thus equivalent to a selection of U.S. counties chosen at random. This map is the 'truth' that is estimated in Figures 5 and 6

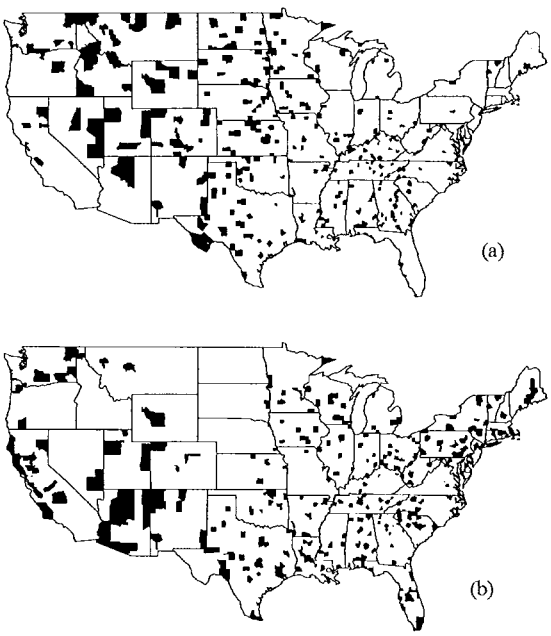

Fig. 5. (a) Shaded counties are those in which the observed rates, y_j/n_j, are in the top 10 per cent of U.S. counties. (b) Shaded counties are those in which the posterior means, $E(\theta_j|y_j) = (\alpha + y_j)/(\beta + n_j)$, are in the top 10 per cent. Compare these maps to the map of the highest true county parameters in Figure 4. The map of the observed rates highlights too many low-population rural counties, whereas the map of the posterior means includes to many high-population urban counties. These effects are perhaps most easily seen in the generally low-population counties of the Plains states

parameters – which, for the Poisson-gamma model described above, happens to be a gamma(α', β') distribution for each county with α' and β' given by the numerator and denominator of equation (9), respectively. Figure 6 displays four maps of independent multiple imputations of the vector θ, each displaying the counties with highest imputed values of θ_j. These maps differ from each other, and from Figure 4, because of the Poisson variability in the data.

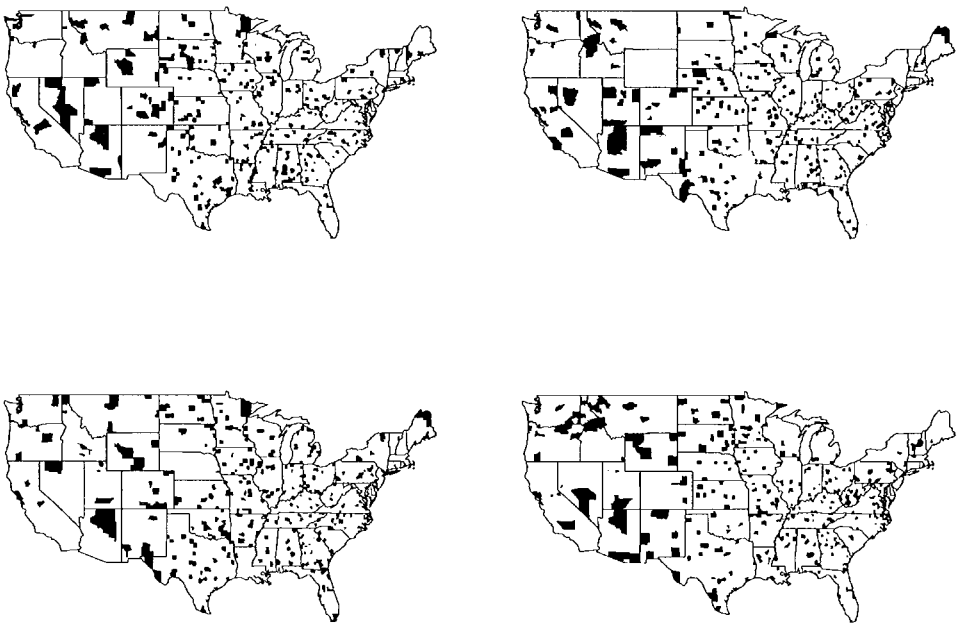

Fig. 6. Four multiple imputations. For each map, the shaded counties are those in which the imputed rates, θ_j, drawn from their posterior distribution, are in the top 10 per cent of U.S. counties, for that imputation. Compare these maps to the map of the highest true county parameters in Figure 4. These maps have no systematic artefacts due to variation in the county populations

The variation among the four maps gives some indication of the posterior uncertainty in the county parameter estimates. For example, in the map on the upper right, the western state of Wyoming has no highlighted counties, whereas in the other maps several Wyoming counties are highlighted. This implies that, given the model and the data, the true rates in those counties could be mostly low, or mostly high, or a mixture, and the maps show various of these possible realities. No strong conclusions can be drawn from any single map – in the presence of statistical uncertainties there is no way to map reality, just possible realities given the model and the data. Instead, one must look for spatial patterns that persist over most of the maps. We have no hard and fast rule for how many maps to make; in this case it seems unnecessary to display more than six or seven (or fewer than three). We suggest starting by making as many as can comfortably fit on a page while allowing sufficient resolution to discern spatial patterns if they are present.

4. MORE COMPLICATED MODELS

The basic cause of the mapping artefacts is that the posterior uncertainties in the counties are unequal, and this inequality can lead to spatial patterns in maps of point estimates. More sophisticated modelling will tend to reduce the variation of uncertainties among counties but will not, in general, equalize the uncertainties altogether. Thus the artefacts described in this paper should remain, in qualitatively similar form. Here, we briefly consider the effects of three forms of

added model sophistication: accounting for uncertainty in the hyperparameters; adding regression predictors, and spatial modelling.

Our examples would gain realism by considering the hyperparameters – (μ, σ, τ) in the normal model and (α, β) in the Poisson-gamma model – as unknown and estimated from the data rather than fixed. In general we agree with Clayton and Bernardinelli[10] that it is best to average over posterior uncertainty. This would not change the essential pattern of the figures or our main results, but it would cause the lines in Figure 3 to lose sharpness in their sawtooth pattern. The multiply imputed maps would still have no sample size artefacts.

It is standard practice to include explanatory variables (such as demographics in the analysis of cancer rates,[1] or geologic indicators in the analysis of radon levels[14]) and to explicitly model spatial correlation, typically to account for missing or poorly measured spatially-correlated covariates (see, for example, Clayton and Kaldor[18] and Mollie and Richardson[21] for Bayesian examples in disease mapping, and Cressie[22] for a general review). Unfortunately, spatial modelling does not remove the artefacts discussed in this paper, although it can sometimes reduce them by diminishing parameter uncertainties. Rather than choose specific spatial models to illustrate this point, we merely point out two extreme cases for which the presence of artefacts is readily apparent.

First, the non-spatial examples in the previous sections can be thought of as spatial models in the limit of zero spatial correlation, so if there is some *small* amount of spatial correlation, the artefacts will be nearly the same as those described above.

Second, consider an opposite extreme; suppose correlation is fairly high at small spatial scales but decreases with distance. For simplicity of exposition, suppose we are interested in a large region and that some areas around the perimeter of the region are very heavily sampled, but a large interior portion has no measurements at all. Any spatial estimation or modelling procedure we are aware of (including interpolation, splines, kriging and hierarchical Bayesian methods) will tend to generate predictions for the interior that are too smooth – subunits in the interior will have predictions that are very close to one another, since there is no information that allows them to be distinguished (see Nobre and De Macedo[23] for an example with contour maps).

Details of the artefacts in more elaborate models will obviously depend on the exact nature of the models and the data. Our point is that mapping artefacts due to spatial variation in parameter uncertainties are nearly ubiquitous, whether the mapped quantities are measured values, predictions from conventional regressions, Bayesian posterior predictions, or whatever, and whether the models are spatial or not.

5. DISCUSSION

Mapping raw data can lead to spurious spatial features. For example, regions can appear highly variable because of small sample sizes in spatial sub-units (as in the radon example) or small populations (as in the cancer example), and these apparently variable regions contain a disproportionate number of very high (or low) observed parameter values. Mapping posterior means leads to the reverse problems: areas that appear too uniform because of small sample sizes or populations. Moulton *et al.*[24] discuss some other problems with maps of posterior means. Similar problems occur with mapping counties based on statistical significance, as discussed in this article.

One way to avoid these artefacts is to produce multiple maps based on imputations from the posterior distribution (of county means, for example); spatial correlation in these maps must come from some other source. In a typical application, one might make maps of imputations from the posterior distribution of *residuals* from predictions based on covariates. Substantial spatial

correlation in the residuals that occurs in all or most of the imputed maps would indicate the presence of un-included covariates that are themselves spatially correlated, such as geologic or house construction features in the radon example. When used in this manner, multiply imputed maps can be thought of as posterior predictive checks.[25, 26]

Unfortunately, multiply imputed maps are not suitable for presenting final results (estimated cancer rates, mean radon concentrations, etc.) to most audiences, who would likely just be confused by them. Furthermore, maps really do make convenient look-up tables (what is the cancer rate, or mean radon level, in my county?). Unfortunately, even maps that are intended to be used only as look-up tables are almost sure to be used for identifying spatial features – we find it very hard to suppress this instinct ourselves. For example, a state Department of Health might map posterior estimates of county mean radon concentrations and choose to focus public education efforts on the areas of the state that appear to have high radon levels. If some contiguous group of counties is sparsely sampled – a common occurrence in practice – then these counties are likely to have near-average posterior estimated levels even if some of the counties have quite high radon levels. Therefore the group of counties will appear both average and uniform on the map, which may lead to seriously incorrect inference if the visual appearance of a large, uniform area on the map is interpreted as evidence of spatial smoothness of county mean radon levels in the area.

To the extent that some of the features identified by conventional mapping methods may be (in some cases are likely to be) artefacts, the natural tendency to associate uniformity on the map with uniformity in reality is unfortunate. Perhaps hatching or shading can be used to indicate not only the point estimates of the quantities of interest but also their uncertainties (for example, see Carlin and Louis[27]); or two maps can be presented, one of posterior means and one of posterior standard deviations; but this is a graphical design issue rather than a statistical one.

Our main goal in this paper has been to illustrate and quantify the extent to which statistical artefacts lead to misleading maps. It is clear that there are serious drawbacks to using spatial distributions of mapped point estimates to gauge the spatial distribution of quantities of interest. Multiple imputation can help avoid this problem in exploratory analysis and model checking, but we know of no satisfactory solution to the problem of generating maps for general use.

ACKNOWLEDGEMENTS

We thank Donald Rubin and Hal Stern for helpful comments. This work was supported in part by the U.S. National Science Foundation grants DMS-9404305, SBR-9708424, and Young Investigator Award DMS-9457824, and the Director, Office of Energy Research, Office of Health and Environmental Research, Environmental Services Division of the U.S. Department of Energy, under contract DE-AC03-76SF00098.

REFERENCES

1. Manton, K. G., Woodbury, M. A., Stallard, E., Riggan, W. B., Creason, J. P. and Pellom, A. C. 'Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates', *Journal of the American Statistical Association* **84**, 637–650 (1989).
2. Smans, M. and Esteve, J. 'Practical approaches to disease mapping', in Elliot, P., Cuzick, J., English, D. and Stern, R. (eds), *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, Oxford University Press, Oxford, 1992, pp. 141–150.
3. Riggan, W. B., Creason, J. P., Nelson, W. C., Manton, K. G., Woodbury, M. A., Stallard, E., Pellom, A. C. and Baubier, J. *U.S. Cancer Mortality Rates and Trends*, 1950–1979 *Vol.* IV: Maps, U.S. Government Printing Office, Washington, D.C., 1987.

4. Devine, O. J., Louis, T. A. and Halloran, M. E. 'Empirical Bayes methods for stabilizing incidence rates before mapping', *Epidemiology*, **5**, 622–630 (1994).
5. Pickle, L. W. and White, A. A. 'Effects of the choice of age-adjustment method on maps of death rates', *Statistics in Medicine*, **14**, 615–627 (1995).
6. Efron, B. and Morris, C. 'Data analysis using Stein's estimator and its generalizations', *Journal of the American Statistical Association*, **70**, 311–319 (1975).
7. Rubin, D. B. 'Using empirical Bayes techniques in the law school validity studies (with discussion)', *Journal of the American Statistical Association*, **75**, 801–827 (1980).
8. Clayton, D. and Kaldor, J. 'Empirical Bayes estimates of age-standardized relative risks for use in disease mapping', *Biometrics*, **43**, 671–681 (1987).
9. Louis, T. A. 'Estimating a population of parameter values using Bayes and empirical Bayes methods', *Journal of the American Statistical Association*, **79**, 393–398 (1984).
10. Clayton, D. and Bernardinelli, L. 'Bayesian methods for mapping disease risk', in Elliott, P., Cuzick, J., English, D. and Stern, R. (eds), *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, Oxford University Press, Oxford, 1992, pp. 205–220.
11. Wirth, S. *et al.* 'National radon database documentation: the EPA state/residential radon surveys', Sanford Cohen and Associates. Prepared for the U.S. Environmental Protection Agency, Washington, D.C., 1992.
12. Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. *Graphical Methods for Data Analysis*, Wadsworth, Pacific Grove, California, 1983.
13. Price, P. N. 'Predictions and maps of county mean indoor radon concentrations in the mid-atlantic states', *Health Physics*, **72**, 893–906 (1997).
14. Price, P. N., Nero, A. V. and Gelman, A. 'Bayesian prediction of mean indoor radon concentrations for Minnesota counties', *Health Physics*, **71**, 922–936 (1996).
15. Tufte, E. R. *The Visual Display of Scientific Information*, Graphics Press, Cheshire, Connecticut, 1983.
16. Mason, T. J., McKay, F. W., Hoover, R., Blot, W. J. and Fraumeni, J. F. *Atlas of Cancer Mortality for U.S. Counties*: 1950–1969, Public Health Service, National Institutes of Health, Washington, D.C., 1975.
17. Schlattmann, P., Dietz, E. and Bohning, D. 'Covariate adjusted mixture models and disease mapping with the program Dismapwin', *Statistics in Medicine*, **15**, 919–929 (1996).
18. Muir, C. S. 'Cancer mapping: overview and conclusions', *Recent Results in Cancer Research*, **114**, 269–273 (1989).
19. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, 1987.
20. Rubin, D. B. 'Multiple imputation after 18+ years', *Journal of the American Statistical Association*, **91**, 473–489 (1996).
21. Mollie, A. and Richardson, S. 'Empirical Bayes estimates of cancer mortality rates using spatial models', *Statistics in Medicine*, **10**, 95–112 (1991).
22. Cressie, N. A. C. *Statistics for Spatial Data*, revised edition, Wiley, New York, 1993.
23. Nobre, F. F. and De Macedo, M. M. A. 'Feasibility of contour mapping epidemiological data with missing values', *Statistics in Medicine*, **14**, 605–613 (1995).
24. Moulton, L. H., Foxman, B., Wolfe, R. A. and Port, F. K. 'Potential pitfalls in interpreting maps of stabilized rates', *Epidemiology*, **5**, 297–301 (1994).
25. Rubin, D. B. 'Bayesianly justifiable and relevant frequency calculations for the applied statistician', *Annals of Statistics*, **12**, 1151–1172 (1984).
26. Gelman, A., Meng, X. L. and Stern, H. S. 'Posterior predictive assessment of model fitness via realized discrepancies (with discussion)', *Statistica Sinica*, **6**, 733–807 (1996).
27. Carlin, B. P. and Louis, T. A. *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall, London, 1996.

