Geographical & Environmental Modelling, Vol. 5, No. 1, 2001, 9–22

A Bayesian Approach to Schools' Catchment-based Performance Modelling

CHRIS BRUNSDON

ABSTRACT *A Bayesian model for predicting schools' performance in SATs is proposed, in which geographical explanatory variables are averaged over a radius centred on each school. The coefficients for the explanatory variables, and the averaging (or smoothing) radius are treated as unknown variables, whose prior and posterior distributions will be considered in the Bayesian analysis. Posterior distributions cannot be determined analytically, and will be investigated using simulation via Gibbs sampling.*

1. Introduction

It is sometimes asserted that performance of schools can be assessed in terms of pupils' success in exam results, and in the 'league tables' that may be generated from these. Whilst this is true to some extent, one cannot reasonably judge a school on this evidence alone. It is also well known that the social background of pupils attending the school will have some bearing on exam performance. Given the differences between social backgrounds of pupils attending different schools, the competition for a high placing in the league tables does not take place on a level playing field. It would perhaps be more helpful to take this into account when assessing schools. One way of achieving this is by modelling the relationship between social background and schools' performance in a probabilistic way—that is, one would like to obtain a probability distribution of some indicator of school exam performance *given* a set of social and economic indicators relating to the catchment area of the school.

However, at this stage a new problem is faced. It is wished to measure social conditions in the catchment area of the school, but it is not known exactly where this catchment area is. In fact, catchment areas pose some interesting problems in the British context. The principle of parental choice implies that it is quite possible for two children living in the same area to attend different schools. One cannot simply divide the map into zones, and then assuming that all children living in a given zone attend a particular school. In some countries this is a reasonable approach, as this is the basis for the assignment of pupils to schools, but in Britain, although distance from home does play a practical role in school choice, catchment areas are fuzzy, and overlap.

A typical approach to modelling the linkages between school exam performance and other explanatory variables is to use multi-level models (Goldstein, 1995). This

Chris Brunsdon, Department of Geography, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK. E-mail: chris.brunsdon@newcastle.ac.uk

1361-5939 print/1469-8323 online/01/010009-14
DOI: 10.1080/13615930120032590

allows variables measured at different hierarchical levels, such as individual pupil, class, school, education and authority to be used to predict exam performance at any of these levels. While this has proved extremely useful, particularly in modelling performance at the pupil level, it is felt that this approach adopts an unrealistic approach to the geography of school catchment area characteristics. Essentially, the discrete, hierarchical nature of the explanatory variables sits uneasily with the fuzzy, overlapping nature of catchment geography described above. In this paper, an attempt to address this shortcoming will be made.

This will be done by using spatially smoothed census ward data to predict school performance. The smoothing window applied to the data is intended to reflect the catchment area around each school. The window will take a kernel form, as this reflects the fuzziness of catchment area boundaries. The performance itself is measured by the number of pupils in each school reaching a certain level of achievement in attainment tests in English.[1] This can be thought of as a binomial variable with the number of trials corresponding to the number of pupils taking the test, and the number of successes corresponding to the number of pupils achieving the required grade. Thus, the model will take the general form:

$$z_i \sim \mathrm{Bin}(\mathrm{logit}(\mathbf{S}\mathbf{x}_i'\beta), n_i)$$

where $\mathbf{X}$ is a matrix of census variables, so that $\{\mathbf{X}\}_{jl}$ is the lth census variable for zone j, and $\mathbf{x}_j$ is the jth column of this matrix, $\mathbf{S}$ is a smoothing matrix, z_i is the school performance for school i and β is a vector of regression coefficients. Assume also that there are m_1 schools, m_2 census zones and m_3 census variables. Thus, $\mathbf{S}$ has m_1 rows and m_2 columns, $\mathbf{X}$ has m_2 rows and m_3 columns, and therefore $\mathbf{SX}$ has m_1 rows and m_3 columns.

The smoothing matrix $\mathbf{S}$ has rows that provide a vector of weights, summing to unity, which are used to provide a weighted mean of the census variables. Typically s_{ij} will be a monotone decreasing function of the distance between the centroid of census zone i and school j. Note that the dimensions of $\mathbf{X}$ and $\mathbf{p}$ (the vector of binomial probabilities) need not match, as they do in a standard general linear model. This is because the smoothing matrix, $\mathbf{S}$, is not square (unless the number of schools equals the number of census zones), so that $\mathbf{SX}$ plays the role of the usual design matrix in a general linear model. In a more informal notation, the model used here takes the form:

school performance $= f$(smoothed census variables, random error)

In this paper, $\mathbf{S}$ will be assumed to depend on a number of parameters in a vector $\mathbf{k}$ say, where $\mathbf{k}$ is unknown. Thus, in the following analyses it is intended to make inferences about both β and $\mathbf{k}$. These correspond to the relationships between the census variables and schools' performance, and the size of the catchment areas (or more directly the area of smoothing) over which the latter should be measured. In its simplest form $\mathbf{k}$ will be replaced by a scalar k, a universal smoothing bandwidth for all schools.

A Bayesian approach will be used for this analysis, using Monte-Carlo Markov Chain (MC)[2] methods to simulate the posterior distributions for β and $\mathbf{k}$. There are a number of reasons for adopting this approach. Firstly, one may wish to investigate posterior distributions of some functions of β and $\mathbf{k}$, and secondly, one may wish to replace *fixed effects* of β and $\mathbf{k}$ by random coefficients—this will be discussed in the next section. Although this may be addressed in a non-$(MC)^2$ framework, this approach allows the exploration of models which do not provide convenient analytical forms for the posterior distributions of unknown coefficients.

In the following sections, a more detailed formulation of the model will be given, followed by a brief outline of the $(MC)^2$ methodology. The ideas in these sections will then be combined in a subsequent analysis of the data. Finally, the implications of the analysis and related ideas will be discussed.

2. Model Formulation

There are three factors that complicate the issue of school performance assessment. Firstly, although a school catchment area can be defined in terms of the home addresses of pupils, it has also been suggested that it is not only the home characteristics of pupils that affects performance, but also that of *their neighbourhood* (Coombes & Raybould, 1997). For the furthest-flung pupils, then, it is possible that social characteristics of places even further away from schools than the outside edges of the catchment areas might have some effect on exam performance—see Figure 1.

This simply reinforces the 'fuzziness' of the situation—it is known that overlapping fuzzy zones are linked to school performance, but not how large these zones are. In

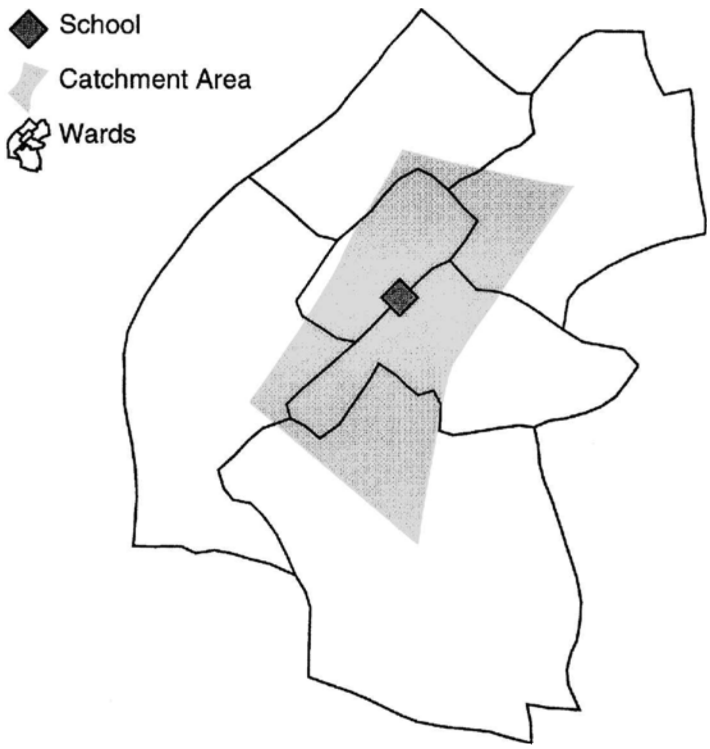

Figure 1. How areas beyond schools' catchments could affect school exam performance: social conditions in the neighbourhoods of pupils could affect their performance. If the school's catchment above is the light grey zone, it is seen that neighbourhoods whose social conditions could affect school performance (black boundaries) extend beyond this.

set theoretical terms, the union of the areas in Figure 1 is the area likely to influence school performance, but one does not know the extent of this area. This is something that will need to be considered in the model.

This has more implications for the interpretation of the model than for the calibration. Suppose that initially one assumes that a universal smoothing parameter can be specified, say k. Then treat k as an unknown model parameter and calibrate it in the analysis. This way, some information about the size of the area around a school whose social conditions affect performance will be gained. However, one does not know whether this size corresponds to the schools' catchment area, or to an area extending beyond this, as in Figure 1. A calibrated value for k gives a general idea of the overall area size, but not of its decomposition into catchment area and 'outer ring'. This could only be understood with further data about home address for the pupils of each school. Essentially, the catchment boundary on Figure 1 cannot be explicitly determined without this data.

The second issue relates to the *distribution* of pupils in the catchment area and surrounding neighbourhoods. It is unlikely that the students are *uniformly* distributed within their catchment area. Although parental choice is the guiding principle for school allocation, in part this choice will be based on the distance between home and school—so it seems reasonable that there will be some fall-off of density of pupils' home addresses as one moves away from any school.

This is reflected in the shape of the kernel smoothing function. It is unlikely that the spatial distribution of a school's pupils will be uniform over a catchment area—it is more likely that as one moves away from a school, there will a decrease in the number of pupils per unit area. Thus it seems reasonable to weight the census variables closer to a school more highly than those further away. Essentially, this justifies a kernel smoothing approach. For any given school, an individual performance predictor will be a weighted average of all values of a single census variable, with the weighting greater for census wards close to the school. Suppose the census variable is x_{lj}, where j runs from 1 to m_2. Then the lth predictor for school i is

$$\sum_{j=1}^{m_2} s_{ij} x_{ij} \tag{2}$$

A kernel-based approach is achieved if one sets, for example

$$s_{ij} = \alpha \exp(-d_{ij}^2/k^2) \tag{3}$$

where α is chosen so that $\sum_{j=1}^{m_2} s_{ij} = 1$, and d_{ij} is the distance between the centroid of zone j to school i. A model of this form ensures that the lth predictor for school i depends on the values of nearby values of x_{jl}. Although a Gaussian form for the kernel is chosen here, other choices may work equally well. Much research into smoothing operators has shown that specific functional form for the kernel is unimportant, so long as a monotone decrease with d_{ij} occurs.

3. Bayesian Inference and $(MC)^2$

Models of the above kind may both be calibrated using Bayesian techniques and $(MC)^2$ methods. Since these approaches are less commonly used than standard classical inference and maximum likelihood techniques, this section provides a brief overview.

Bayesian analysis is essentially simple. Suppose one has a probability model for a set of observations $\{z_1 \dots z_n\}$, dependent on some unknown parameter h. For example, $\{z_1 \dots z_n\}$ could be drawn from a binomial distribution with unknown parameter h and known parameter n_i, as in the schools' performance example. Although one does not know the exact value of h, suppose one can at least provide a probability distribution for this value. Then there are two probability statements to consider: the probability distribution for h, $p(\mathrm{h})$ say, and the probability distribution for the $\{z_1 \dots z_n\}$ *given* h. Refer to the latter as $f(z_1 \dots z_n | \mathrm{h})$. What one would like to do is to use $\{z_1 \dots z_n\}$ to find out more about h. That is, one would like to find out the probability distribution of h *given* $\{z_1 \dots z_n\}$, written as $p(\mathrm{h} | z_1 \dots z_n)$. Bayes' theorem gives exactly this:

$$p(\mathrm{h} | z_1 \dots z_n) \propto p(\mathrm{h}) f(z_1 \dots z_n | \mathrm{h}) \tag{4}$$

Here the convention of denoting probability distributions for the parameters as $p(.)$, whilst probability distributions for the *observations* are denoted as $f(.)$ will be followed. The constant of proportionality is chosen to ensure that $p(\mathrm{h} | z_1 \dots z_n)$ is normalized—that is, that it integrates (or sums, if h is discrete) to unity over the allowable range of h. On many occasions, finding the value of this constant can be difficult or even impossible to solve analytically. However, this problem can be circumnavigated by simulation. In particular the technique of *rejection sampling* (Gelman *et al.*, 1997) allows random sampling from an unnormalized distribution.

At this stage, the analytical and computational basis to make inferences about a *single* scalar parameter h given a set of observations is provided. It is also possible to compute point estimates of h based on the mean or mode of the posterior distribution. However, in real-life problems one usually needs to consider several parameters. For example, in (6) one needs to estimate l and the elements of β. Theoretically, the Bayesian framework can be extended to this situation. One simply needs to consider the prior distribution and the data model as multivariate: $p(\mathrm{h}_1 \dots \mathrm{h}_m)$ and $p(z_1 \dots z_n | \mathrm{h}_1 \dots \mathrm{h}_m)$. As in the univariate case, simulation provides a useful way to overcome the often difficult problems of normalization. Since it is convenient to consider non-normalized probability expressions, this suggests that rejection sampling should again be used.

However, although this is theoretically possible, for high-dimensional probability distributions this is extremely computationally inefficient. A more workable alternative is to make use of the *Gibbs sampler* (Geman & Geman, 1984). This provides a technique for simulating draws from the posterior distribution of the parameters, $p(\mathrm{h}_1 \dots \mathrm{h}_m | x_1 \dots x_n)$ given expressions for the univariate conditional distributions of each h_i, say $p(\mathrm{h}_i | \mathrm{h}_1 \dots \mathrm{h}_{i-1}, \mathrm{h}_{i+1} \dots \mathrm{h}_m, x_1 \dots x_m)$. This expression is essentially the posterior distribution for h_is assuming all of the other parameters are known. Clearly one can use rejection sampling to simulate random h_is in this case, as this distribution is univariate. Given that this can be done for each $i \in \{1 \dots m\}$, the Gibbs sampling procedure is set out below:

1. Supply initial guesses for $\{\mathrm{h}_1 \dots \mathrm{h}_m\}$. Call these $\{\mathrm{h}_1^{(0)} \dots \mathrm{h}_m^{(0)}\}$.
2. For each h_i: Simulate a random h_i from the distribution

   $$p(\mathrm{h}_i | \mathrm{h}_1^{(0)} \dots \mathrm{h}_{i-1}^{(0)}, \mathrm{h}_{i+1}^{(0)} \dots \mathrm{h}_m^{(0)}, z_1 \dots z_m).$$

3. Call the new simulated parameter set $\{\mathrm{h}_1^{(1)} \dots \mathrm{h}_m^{(1)}\}$.
4. Return to step (2) a number of times, each time simulating a parameter set $\{\mathrm{h}_1^{(j)} \dots \mathrm{h}_m^{(j)}\}$ from the previous set $\{\mathrm{h}_1^{(j-1)} \dots \mathrm{h}_m^{(j-1)}\}$.

If this cycle is repeated a sufficient number of times, it can be shown (Geman & Geman, 1984) that the process 'forgets' the initial guesses for $\{h_1 \ldots h_m\}$ and the final $\{h_1^{(j)} \ldots h_m^{(j)}\}$ is distributed as $p(h_1 \ldots h_m | x_1 \ldots x_m)$. Thus, by a series of univariate simulations, the multivariate posterior distribution for $\{h_1 \ldots h_m\}$ can be simulated. Furthermore, for future Gibbs samples, the output of the previous run, $\{h_1^{(j)} \ldots h_m^{(j)}\}$ can be used as the initial guess for the next run. After this, there should be some checking to see when the initial guesses for $\{h_1 \ldots h_m\}$ have been 'forgotten'—see for example Gelman *et al.* (1997). The early simulation cycles over which the initial conditions have influence are known as the *burn-in*. The simulated values from the burn-in should be discarded. It is sometimes also useful to check for serial correlation in the simulations. If this is apparent, the final simulated values should be produced by *thinning* the series, that is selecting only every qth simulation, where q is some small positive integer.

Once the above procedures have been followed, it is possible to investigate the properties of multivariate posterior distributions empirically, by examining the output of the simulation. Typically, one can consider the joint probability distribution of a subgroup of parameters, or one can consider the marginal distribution of each parameter in turn.

4. Analysis of Schools' Data

The following section outlines how the above methodology may be applied to the schools' data, and lists results. The analysis takes place in three stages. Firstly the universal bandwidth model is fitted. Using the output from this, geographically smoothed patterns in explanatory variables can be derived. Finally, a variable bandwidth model is considered. Results are considered for each stage in turn.

4.1. The Basic Approach

For an initial analysis of the schools' data model, a universal catchment model will be considered. The performance data considered is the 1997 Primary School (Key Stage 2) Performance Results in English for schools in Tyne and Wear (*The Times*, 1998). As a predictor variable, ward-based male unemployment rates (as proportions between 0 and 1) in Tyne and Wear are used. There are 120 wards in the study area, and 403 schools. The locations of schools and wards are shown in Figure 2.

In this instance, β is a two-element column vector, $(b_0, b_1)'$, where b_0 is an intercept term and b_1 is a regression coefficient for (smoothed) unemployment rates. As discussed earlier, it is these coefficients together with k, the universal smoothing bandwidth, which must be estimated. Using simulation, it is intended to draw from the distribution $p(b_0, b_1, k | z_1 \ldots z_{m1})$ to make inferences about the three parameters. The results presented below are based on $(MC)^2$ simulation techniques using a Gibbs sample r as discussed in the previous section.

4.2. Results of Basic Model

For the schools' data, 1050 simulations were run based on the methodology stated above. The first 50 simulations were discarded as 'burn-in'. Thus, 1000 random draws from the joint posterior distribution of b_0, b_1 and k were produced.

For the basic analysis, one is interested in the marginal posterior distributions of

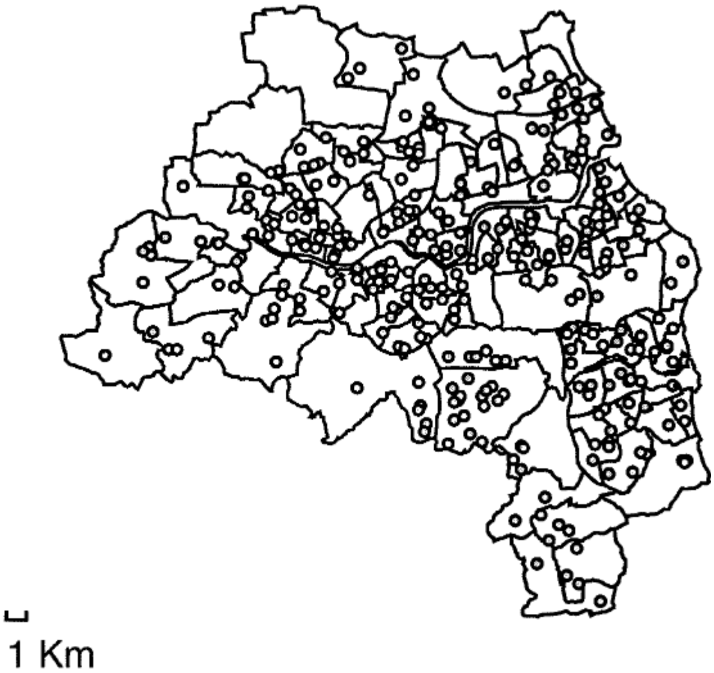

Figure 2. Relationship between census wards and school locations in Tyne and Wear: ward boundaries are shown as outlines and each circle represents a school.

b_0, b_1 and k. These can be summarized by the means of the simulations (giving point estimates of the parameters), and by the standard deviations of the parameters (giving posterior standard deviations, the Bayesian equivalent of standard errors). These are shown in Table 1.

From this it can be seen that it is unlikely that any of the parameters are equal to zero on the basis of posterior distributions. This is a Bayesian statement similar to stating that all parameters differ significantly from zero in a classical framework. Note also that b_1 is negative—suggesting that higher surrounding levels of unemployment tend to affect school performance adversely, as one might expect. Finally, the universal catchment area bandwidth is about three-quarters of a kilometre. How may this figure be interpreted? Consider the smoothing process as taking a weighted average of unemployment rates. If a census ward is a distance d from a school, its weight is $\exp(-d^2/2k^2)$. If the weight is 0.01, then d is about three times k. Thus, it is reasonable to state that the effective extent of the kernel is about three times k.

Table 1. Results of the basic model calibration using $(MC)^2$. Estimate 1 and Std. Dev. 1 use an upper limit of 10 km on the prior for k. Estimate 2 and Std. Dev. 2 use 20 km. k is measured in metres. Note that results agree to two significant digits

Parameter	Estimate 1	Std. Dev. 1	Estimate 2	Std. Dev. 2
k	764	10.6	767	10.6
b_0	1.4	0.06	1.4	0.06
b_1	−5.0	0.30	−5.0	0.3

Here, the extent is about 2.2 km. This may seem like a very low figure, but it should be considered that the data here relate to *primary schools* and typically these tend to serve relatively small neighbourhoods. It is speculated that for secondary schools the bandwidth would be larger.

More detailed information about the marginal posterior distributions of the parameters can be obtained by drawing histograms. This will produce images of the estimated posterior distributions. These histograms are shown in Figure 3. The production of this kind of plot is one of the advantages of a Bayesian approach.

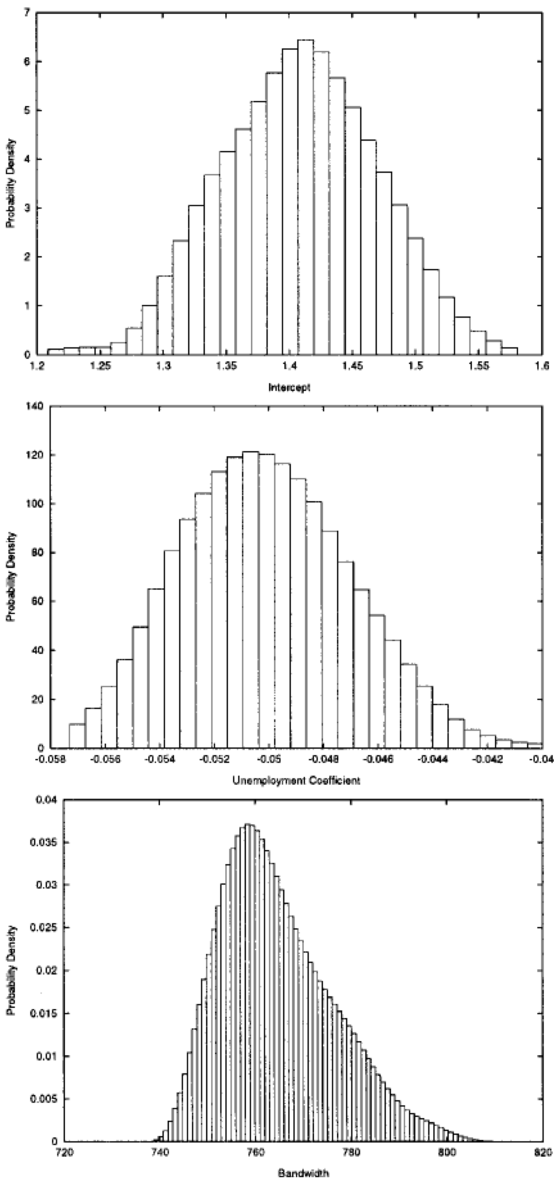

Figure 3. Posterior distributions of model parameters: from top to bottom b_0, b_1 and k.

Rather than simply obtaining a point estimate or confidence intervals for the unknown parameters, richer information is provided by the posterior distribution. For example, it can be seen here that the posterior distribution for k is skewed to the right, whilst those for b_0 and b_1 are closer to symmetry.

Finally, the goodness of fit of the model is assessed. To do this, standardized residuals are mapped. Since the values of z_i for each school are assumed to follow a binomial distribution, we can compare the actual proportion of pupils attaining the desired grade in English z_i/n_i, against p_i, the predicted proportion using the basic model. The variance of z_i/n_i is $p_i(1 - p_i)/n_i$, so the standardized residual is

$$r_i = \frac{z_i/n_i - p_i}{\sqrt{p_i(1 - p_i)/n_i}} \qquad (5)$$

The pattern of r_is is mapped in Figure 4, showing where either $r_i < -2$ or $r_i > 2$. Here it can be seen that high positive and high negative values of r_i tend to cluster (suggesting there is further spatial grouping than explained in the basic model). The high number of residuals whose standardized absolute value exceeds 2 also suggests that there is some extra-binomial variation in the model. One approach to modelling this is to include a random intercept coefficient in the basic model, allowing for an individual school's effect in addition to the environmental effect modelled above. An alternative approach, which may also address the geographical grouping, is discussed in the following section.

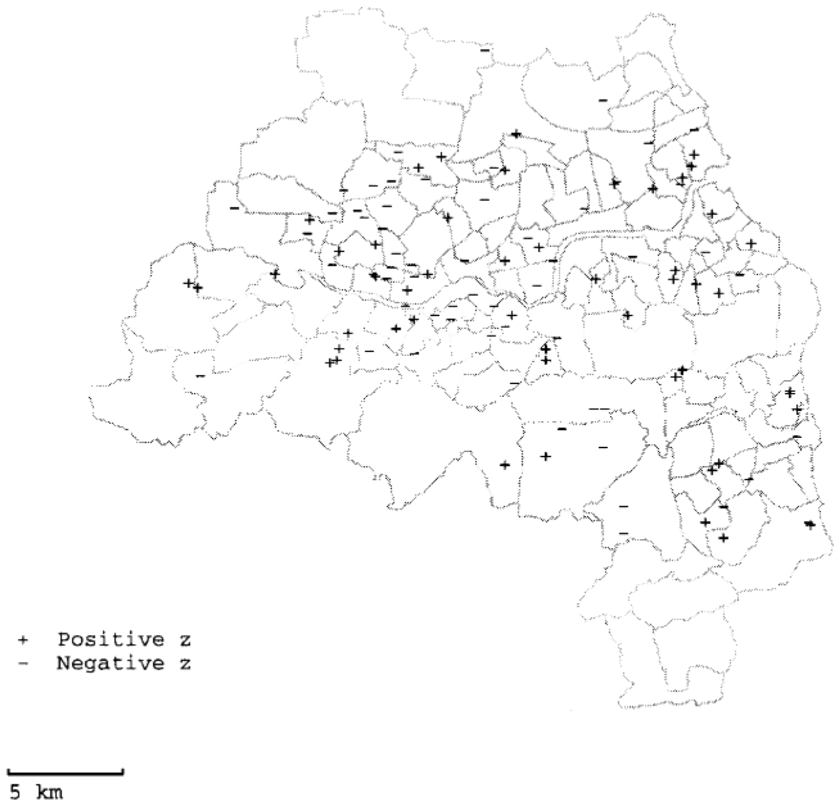

Figure 4. Standardized residuals for the basic model.

4.3. Mapping Geographical Effects

In the universal catchment model the smoothing parameter is the same for any location. This suggests that one could pick any location (u, v) in the study area and obtain a predictive distribution of the performance for a hypothetical school sited at this location. More specifically, one can obtain estimates of the mean of this distribution. Since the mean is a scalar, it may be considered as a function of the two location variables, say $\mathsf{f}(u, v)$. This function can then be mapped, illustrating expected school performance patterns given the geographical data relating to unemployment. Alternatively, one can consider the smoothed unemployment index evaluated at the hypothetical school location (u, v). This can also be thought of as a function, say $\mathsf{u}(u, v)$. A map of this shows the 'best' smoothed unemployment surface, in the sense that it provides the best predictor of performance. The latter approach is useful if there are several indicators, since the geographical effect of each indicator may be mapped in turn.

All of the above would be straightforward if one knew the values of k, b_0 and b_1. However, we only have probabilistic information about these—as shown in Figure 3. This leads to 'fuzziness' in our estimate of $\mathsf{f}(u, v)$ and $\mathsf{u}(u, v)$—and analytical difficulties in finding a closed form expression for $E(\mathsf{u}(u, v))$ or $E(\mathsf{f}(u, v))$. However, the fact that an $(\mathrm{MC})^2$ simulation approach is being used can be of help here. To estimate $E(\mathsf{u}(u, v))$ one simply computes the smoothed indicator $\mathsf{u}(u, v)$ for each simulated k value, and takes the average. Doing this over a grid of (u, v) points leads to a surface estimate. A similar approach may be used to estimate $E(\mathsf{f}(u, v))$, where the simulated values of b_0 and b_1 are also used. To illustrate the method, estimates of $E(\mathsf{u}(u, v))$ are computed, for (u, v) points positioned on a hexagonal grid and shown in Figure 5. Notable features in the map are the higher levels of unemployment around the city centres of Newcastle and Sunderland, particularly following the line of the River Tyne.

4.4. The Local Catchment Area Model

Having considered the basic model above, and noted the degree of extra-binomial variation, one has to consider other possible sources of variation in the model. For example, one could ask whether the same catchment area characteristics apply to *all* schools. It is quite possible that some more popular schools have a wider catchment area. This could be due to a number of reasons—for example Roman Catholics may be more willing to send their children to a Roman Catholic school in favour of a closer non-denominational school. Similarly, one might expect rural schools to have a wider catchment area than urban schools. For this reason, one must seriously consider the possibility of applying different catchment areas to each school.

This can be tackled in a number of ways. In all of these, one must replace the scalar k in the model by an m_1-dimensional vector $\mathbf{k}$, where $\{k\}_i$ is a kernel bandwidth for the ith school. If some details of each school were available, such as whether the school was independent or state-funded, or whether it was associated with a particular religious denomination, then it might be possible to model each k_i as a function of these factors. This would lead to a number of subgroups of the data each having a separate value of k estimated. Unfortunately, such information is not available in the data set supplied. It is also possible that appropriate values of k might vary geographically—for example rural state schools are likely to have larger catchment areas than urban ones.

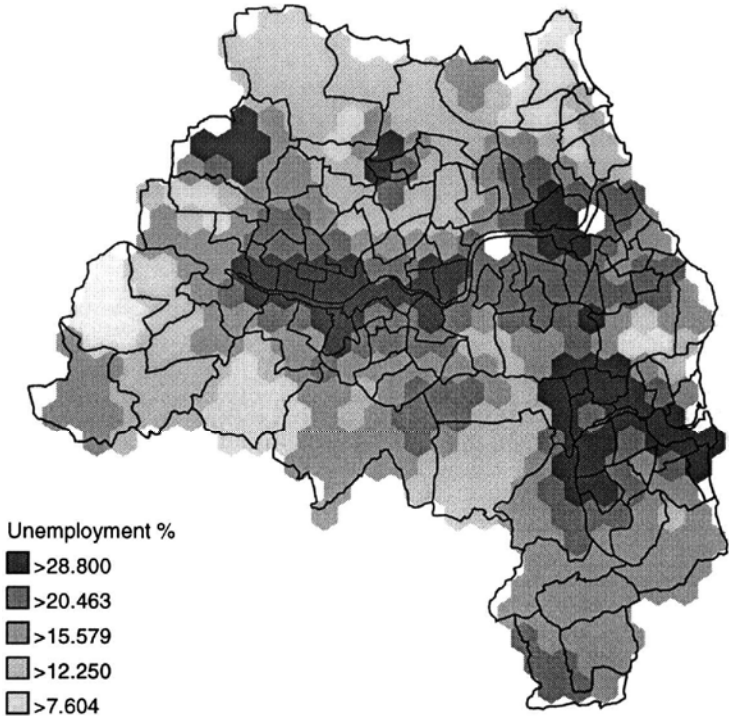

Figure 5. Smoothed unemployment figures. Note the use of a hexagonal grid, rather than the usual rectangular grid. This was found to give a more easily readable map.

These complications make it necessary to assume that each school has its own catchment area. In this case, there could be m_1 distinct values of k_i in **k**. This leads to complications in itself—there are now a very large number of parameters to estimate.[2] One way of overcoming this is to adopt a *hierarchical model* (Good, 1965). In a model of this sort, the elements of **k** are assumed to be random variates from some distribution, and the parameters for this distribution are estimated. Typically, there will be only one or two parameters in this distribution, which will greatly reduce the dimensionality of the parameter estimation problem. Here k_i is modelled as exponentially distributed, with mean l. Thus, model (1) is extended to become

$$z_i \sim \text{Bin}(\text{logit}(\mathbf{S}_k \mathbf{x}_i'\beta), n_i) \qquad \text{where } k_i \sim \text{Exp}(l) \tag{6}$$

Note the smoothing matrix is written here as $\mathbf{S}_k$ to emphasize its dependence on **k**. There is now only one **k**-related parameter to estimate, l. It is, however, possible to obtain conditional distributions for the individual k_is given the data and β and l—so that conditional estimates of individual catchment area sizes can be computed.

The original model (1) is thus extended to (6). $(\text{MC})^2$ analysis applied to the extended model is quite similar to the simpler case. To simulate a draw of the random vector β given l, one firstly generates a set of m_1 k-parameters. Each of these produces a set of weights for each of the m_1 schools. Using these, a set of m_1 unemployment indicators is computed. On the basis of this, using the technique in Section 4.1, a simulated $\{\mathsf{b}_0, \mathsf{b}_1\}$ pair is generated. The second part of the $(\text{MC})^2$ cycle is to simulate l given $\{\mathsf{b}_0, \mathsf{b}_1\}$. This is done using the inverse-CDF method, in

Table 2. Results of the local catchment model calibration using $(MC)^2$. $\mathtt{l}$ is measured in metres. Note that posterior SDs are greater than for the universal catchment model

Parameter	Estimate	Std. Dev.
$\mathtt{l}$	978.0	63.2
$\mathtt{b}_0$	1.5	0.10
$\mathtt{b}_1$	−5.7	0.53

the same way that k is generated in Section 4.1, with the posterior density distribution modified to take into account the random coefficients.

Carrying out simulations in the above manner allows the investigation of $\mathtt{b}_0, \mathtt{b}_1$ and $\mathtt{l}$ in the local catchment model. However, it is also useful to estimate the individual bandwidths for each school, $\{k_i\}$. It is possible to derive the expression for $p(k_i | \mathtt{l}, \mathtt{b}_0, \mathtt{b}_1, z_i)^3$ and to generate a draw of k_i from this distribution for each i during each cycle of the simulation. Averaging these gives individual estimates of k_i for each school.

The results of the $(MC)^2$ simulation are given in Table 2.

Note that the regression coefficients both have slightly higher estimated values, and that posterior standard deviations are now greater for all parameter estimates. This is perhaps unsurprising as the random catchment assumption adds greater uncertainty to the model.

The individual k_i estimates are now shown in map form (Figure 6). From this it can be seen that although there is some tendency for schools further from urban

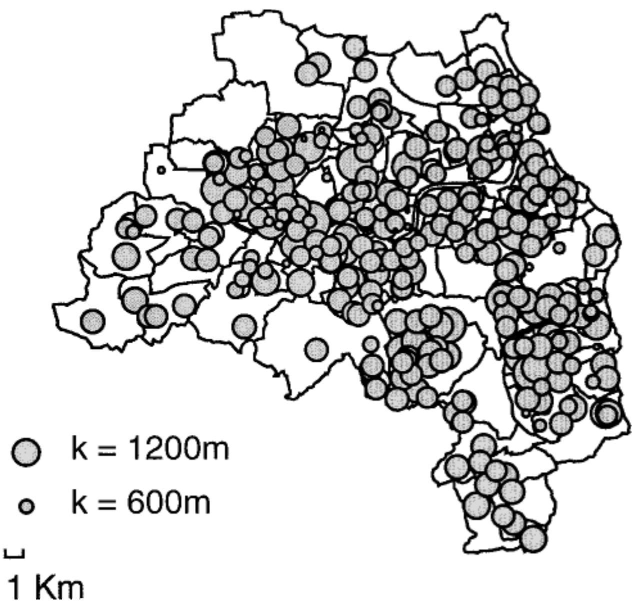

Figure 6. Local catchment area results. Note that size of catchment does not seem to relate to geography in a simple urban/rural way. This suggests catchment can vary greatly even between proximal schools.

centres to have larger catchment areas, this is by no means the only factor to affect catchment area. It is possible, for example, that some relatively remote schools have their catchment in one small village, and conversely, that some more popular urban schools attract pupils from relatively large distances.

5. Conclusions

In this paper, a model for schools' performance has been proposed, based on social and economic characteristics of a geographical area surrounding the school. Clearly, the technique need be applied exclusively to modelling school performance—there are a wide number of geographical problems that could be addressed using this kind of model. There are a number of further ways in which the model could be extended. For example, if they were available, some variables measured at the individual school level (say **Y**) could also be used to predict performance. This could be quite simply incorporated into the Bayesian framework, by replacing **SX** with

$$\begin{pmatrix} \mathbf{SX} \\ \mathbf{Y} \end{pmatrix}$$

in model (1) or (6). There are also a wealth of Bayesian diagnostics and model comparison methods (Kass & Raftery, 1995; Rubin, 1981) which could be used in this context.

Rather than prescribing the extent of the area, this quantity is treated as a parameter in the model to be inferred from the data. This is believed to be important in a methodological sense. In most quantitative geographical studies, the zones over which quantities are aggregated are fixed before statistical analysis takes place. However, an effect first noted by Gehlke and Biehl (1934), later popularized as the modifiable areal unit problem (MAUP) (Openshaw, 1984) implies that geographical patterns can change when these zones of aggregation are altered. This variability introduces a degree of uncertainty into the model not acknowledged when the standard practice is adopted. It seems reasonable that statistical models of geographical processes should make some attempt to analyse the geography itself, rather than treating it as an a priori certainty. It is hoped that this and related techniques will provide a means of addressing this issue.

Notes

1. It might be preferable to measure the actual test scores, but the data are officially released in this format.
2. The exact number is $m_1 + m_3$, for only m_3 observations.
3. Note that k_i only depends on the observed data $\{z_1 \ldots z_{m_1}\}$ only through z_i since the values of z and k are both independently distributed.

References

Coombes, M. & Raybould, S. (1997) Modelling the influence of individual and spatial factors underlying variations in the levels of secondary schools examination results. *Environment and Planning A*, 29, 641–658.

Gehlke, C. & Biehl, K. (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29(Suppl.), 169–170.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (1997) *Bayesian Data Analysis*. London: Chapman & Hall.

Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Goldstein, H. (1995) *Multilevel Statistical Models*. New York: Halstead.

Good, I. (1965) *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

Kass, R. & Raftery, A. (1995) Bayes factors and model uncertainty. *Journal of the American Statistical Association*, 90, 773–795.

Openshaw, S. (1984) *CATMOG 38: The Modifiable Areal Unit Problem*. Norwich: GeoAbstracts.

Rubin, D. (1981) Estimation in parallel randomised experiments. *Journal of Educational Statistics*, 6, 377–401.

The Times (1998) *The Times* newspaper home page http://www.the-times.co.uk/.