

# USE OF SPLINE FUNCTIONS FOR PREMIUM RATING BY GEOGRAPHIC AREA

BY G. C. TAYLOR

## ABSTRACT

The paper gives details of a case study in the premium rating of a Householders Contents insurance portfolio. The rating is performed by the fitting of bivariate spline functions to a version of operating ratio described in Section 3.

The use of bivariate splines requires a small amount of mathematical equipment, which is developed in Section 4. The fitting of splines, using regression is carried out in Sections 5 and 6, where the numerical results are given, including some assessment of goodness-of-fit.

Contour maps of the spline surfaces are also given, and used for the selection of geographic areas used for premium rating purposes. These are compared with the areas, past and present, actually used by the insurer concerned.

## 1. INTRODUCTION

It is common in insurance of domestic property lines, e.g. motor vehicle (collision damage) and householders, to find that the risk premium per unit exposure varies with geographic area when all other risk factors are held constant.

Such variation may or may not be continuous as a function of spatial coordinates. In either event, it will be necessary for practical purposes to divide the total area for which premium rates are required into a relatively small number of regions of reasonable size such that, all other risk factors equal:

- (i) premiums vary as between region,
- (ii) premiums do not vary within region.

Henceforth, such regions will be referred to as *rating regions*.

This raises the question of how such regions should be determined. The present paper considers this in a context in which the determination is to be made solely on the basis of data. In practice, of course, it may be necessary to modify the conclusions reached in this way in order to make suitable allowance for available anecdotal or circumstantial evidence.

Thus it is assumed that claims and exposure data are available in respect of a number (possibly a large number) of subdivisions of the total area for which premium rates are required. In the specific example considered here these subdivisions are postal areas (*postcodes* in Australian terminology). The problem consists of identifying the appropriate aggregations of the postcodes into rating regions satisfying the two conditions set out above.

It is possible to regard the problem as one of cluster analysis, suitable clusters of postcodes being sought. However, it is evident that clustering must be carried out according to the criteria of both geographic clustering and clustering by

premium. Definition of a metric for the clustering algorithm which incorporates both criteria satisfactorily is not easy.

This paper follows a different path. Risk premium, with all factors except geographic coordinates held constant, is envisaged as a continuous function of these coordinates. This function is then estimated and examined for steep gradients which would lead to the definition of rating areas satisfying conditions (i) and (ii) above.

Thus, the fitting of the premium function becomes the major task. The mathematical form of this function is quite unknown. A natural way to fit it smoothly to the available data points is to make use of spline fitting. Note that the fitted spline function is bivariate.

This fitting is carried out in Section 3 and the results presented in Section 4.

## 2. DATA AND NOTATION

### 2.1. Data

Data available in respect of each metropolitan postcode related to the experience of the Householders Contents portfolio of a large Australian insurer in the financial year 1985/86 and in the state of New South Wales. The data comprised:

- (i) postcode identifier and geographic coordinates;
- (ii) years of exposure to risk,
- (iii) number of claims;
- (iv) average cost per claim;
- (v) average sum insured;
- (vi) average jewellery penetration (i.e. the proportion of policies carrying jewellery insurance, this risk requiring a separate coverage);
- (vii) average jewellery sum insured, the average being taken over those cases which carry a non-zero sum insured;
- (viii) average earned premium per year of exposure;
- (ix) *average gross experience premium* per year of exposure, consisting of:

average observed risk premium per year of exposure

*plus*

administration expense loading consisting of a charge per policy, a charge per claim, a percentage of claim payments, and a percentage of premium,

where *average observed risk premium* per year of exposure is defined as:

number of claims per year of exposure

×

average cost per claim,

- (x) the Company's present system for premium rating the relevant postcode, consisting of:
- (a) the rating region to which that postcode is currently assigned;
  - (b) premium rating formula for that rating region, of the form:

base premium +  $\text{const}_1 \times \text{sum insured} + \text{const}_2 \times \text{jewellery sum insured}$ ,

the terms "base premium" and "const." each varying with rating region.

A small sample of these data is displayed in Appendix A

## 2.2. Notation

The remainder of this paper uses the following notation. Suppose there are  $N$  postcodes. Without loss of generality, they can be treated as numbered  $1, 2, \dots, N$  (although in fact they are not). In the following definition of notation a subscript  $i$  denotes postcode  $i$ .

Let

- $E_i$  = number of years exposure;
- $n_i$  = number of claims;
- $x_i$  = average cost per claim;
- $S_i$  = average sum insured;
- $p_i$  = average jewellery penetration;
- $J_i$  = average jewellery sum insured;
- $P_i^E$  = average earned premium per year of exposure before allowance for no claim discounts (NCD);
- $P_i^R$  = average risk premium per year of exposure  $n_i x_i / E_i$ , as defined in Section 2.1,
- $P_i^G$  = average gross experience premium per year of exposure;
- $k$  = average NCD in the portfolio, expressed as a proportion of premium payable net of NCD;
- $R_i^{(P)}$  = rating region to which the postcode is assigned in the present premium rating system,
- $b_R$  = base premium in rating region  $R$ ;
- $\pi_R^{(B)}$  = premium rate per \$ 1000 of sum insured in rating region  $R$ ,
- $\pi_R^{(J)}$  = premium rate per \$ 1000 of jewellery sum insured in rating region  $R$ ;
- $\mathcal{A} = UR_i^{(P)}$  = total region covered by postcodes.

According to this notation, the jewellery premium rate for postcode  $i$  is denoted by  $\pi_{R_i^{(P)}}^{(J)}$ . This very cumbersome expression is abbreviated to  $\pi_i^{(J)}$ . Similarly,  $b_{R_i^{(P)}}$  and  $\pi_{R_i^{(P)}}^{(B)}$  are abbreviated to  $b_i$  and  $\pi_i^{(B)}$  respectively.

## 3. ISOLATION OF GEOGRAPHIC AREA AS A RISK FACTOR

## 3.1. General context

As Section 1 explains, the objective of this paper is to fit a function to the "geographic area effect". This requires controlling for any other factor affecting risk premium.

Strictly, all risk factors should be fitted to the data simultaneously. This, however, would be extremely difficult. It is assumed here that risk premiums are estimated by a three-stage procedure:

- *Stage 1.* Fit all factors simultaneously, but with only a rough fit of the "geographic area effect". This effect could be roughly incorporated in the model using rating regions taken from an existing premium rating system or even chosen by guesswork.
- *Stage 2.* Control for all risk factors other than geographic area by calculating, for each postcode, an index of risk (the following uses a version of operating ratio) based on standardized values of all other risk factors.
- *Stage 3.* Treating this index as function of geographic coordinates  $x, y$ , fit a function  $I(x, y)$ . Then estimate risk premium at  $(x, y)$  as proportional to  $I(x, y)$ .

In the present application, Stage 1 was taken as being carried out by the existing premium rating system, which was believed to be reasonably accurate.

In a general context Stage 2 proceeds as follows.

The operating ratio for postcode  $i$  is:

$$(3.1.1) \quad \rho_i = (1+k)P_i^G/P_i^E.$$

A more accurate version of this formula would have been

$$\rho_i = (1+k_i)P_i^G/P_i^E,$$

with an NCD factor  $k_i$  specific to postcode  $i$ . Unfortunately, the factors  $k_i$  were not available and it has been necessary to use the compromise formula (3.1.1).

This would result in a tendency to reduce high observed operating ratios, and increase low ones. Such attenuation of the data would have little effect on the present exercise, since the selection of geographic rating regions depends largely on the risk *ordering* of postcodes rather than the *magnitudes* of their risks. It is, however, a factor which would need to be taken into account in the subsequent exercise of determining premium rates for the rating regions selected.

Suppose that the premium rate depends on  $m$  factors  $F_1, \dots, F_m$ , with  $F_1$  representing rating region (of the present system). Let  $\pi(f_1, \dots, f_m)$  be the present tabular premium when  $F_1 = f_1, \dots, F_m = f_m$ .

Now let  $\hat{f}_j, j = 2, \dots, m$ , denote the average value of  $F_j$  observed in postcode  $i$ ; and let  $\bar{f}_j, j = 2, \dots, m$  denote the corresponding average over the entire portfolio. Averages here are weighted averages with years of exposure used as weights.

Next select a “standard” rating region, preferably not near the extremes of high or low risk, and denote it by  $R_*^{(P)}$ . With a slight abuse of notation, the tabular premium in this region on the basis of average values of risk factors  $F_2, \dots, F_m$  taken over postcode  $l$  may be written as  $\pi(*, \bar{f}_{i2}, \dots, \bar{f}_{im})$ .

Then, to a reasonable degree of approximation, the factor:

$$(3.1.2) \quad \phi_l = \pi(l, \bar{f}_{i2}, \dots, \bar{f}_{im}) / \pi(*, \bar{f}_{i2}, \dots, \bar{f}_{im})$$

is the factor by which premium actually received in respect of postcode  $l$  has been increased relative to the premium which would have been received had rates of the “standard” rating region been applied to all postcodes.

Removal of the geographic area effect adjusts  $P_i^E$  to  $P_i^E / \phi_l$ , and hence  $\rho_l$  to:

$$(3.1.3) \quad \rho_l^{(*)} = \rho_l \phi_l,$$

by (3.1.1). By (3.1.2) and (3.1.3),

$$(3.1.4) \quad \rho_l^{(*)} = \rho_l \pi(l, \bar{f}_{i2}, \dots, \bar{f}_{im}) / \pi(*, \bar{f}_{i2}, \dots, \bar{f}_{im}),$$

which is the operating ratio which would have been observed in postcode  $l$  had that postcode been subject to the premium rates of the “standard” region. It is essentially an estimate, subject to sampling error, of relative costs of the various postcodes. This will be referred to as the *operating ratio adjusted for region*.

Now let  $(\bar{x}_l, \bar{y}_l)$  denote the centroid of the polygonal postcode area  $l$ , i.e. the average of the vertex coordinates (see Appendix A1). Then a first approximation to the index of risk required in Stage 2 is:

$$(3.1.5) \quad \hat{I}(\bar{x}_l, \bar{y}_l) = \rho_l^{(*)}.$$

An alternative version of  $\phi_l$  may be considered. This is:

$$(3.1.6) \quad \phi'_l = \pi(l, \bar{f}_2, \dots, \bar{f}_m) / \pi(*, \bar{f}_2, \dots, \bar{f}_m).$$

Usually (3.1.2) would be preferred to (3.1.6) since the former takes into account any unusual variation in the risk factors  $F_2, \dots, F_m$  as between different postcodes.

### 3.2. Specific context

In the specific context of the premium rating system set out in Section 2.1, item (x) (b),  $m = 4$  with

- $F_1$  = rating region;
- $F_2$  = sum insured,
- $F_3$  = jewellery penetration;
- $F_4$  = jewellery sum insured;

where, for an individual policy,

- $F_3 = 1$ , if the policy carries a jewellery sum insured,
- $= 0$ , otherwise.

Thus, the numerator of  $\phi_i$ , as defined in (3.1.2), is the premium for a policy written in postcode  $i$  carrying the average sum insured and average jewellery sum insured of that postcode and with the jewellery component of the premium scaled down by the jewellery penetration factor. The denominator of  $\phi_i$  is the same except based on premium rates of the "standard" region instead of the region containing postcode  $i$ .

A small sample of the results of these calculations is set out in Appendix B. As appears there, the two versions  $\phi_i$  and  $\phi'_i$  of operating ratio adjusted for region produce quite similar results and only  $\phi_i$  has been used subsequently.

## 4. SPLINE FITTING

### 4.1. General

According to the preamble in Section 1, the operating ratio adjusted for region is envisaged as a function  $I(x, y)$ ,  $x, y \in \mathcal{A}$ . By (3.1.5), estimates of  $I(\bar{x}_i, \bar{y}_i)$  are available. The function  $I(x, y)$  is to be estimated as a spline which fits these estimates adequately.

### 4.2. Analytical

It is first necessary to clarify what is meant by a spline with multi-dimensional domain. The present application is concerned with a 2-dimensional domain, and discussion will be restricted to that dimensionality. The concepts generalize readily to higher dimensions, but at the cost of more complex definitions which would represent impedimenta in the present context.

**DEFINITION.** Let  $\mathcal{A}$  be the 2-dimensional domain  $[0, \infty) \times [0, \infty)$ , and  $(u, v): \mathcal{A} \rightarrow \mathcal{A}$  a  $C^\infty$  bijection. Let  $0 < h_1 < h_2 < \dots < h_m < \infty$  and  $0 < k_1 < k_2 < \dots < k_n < \infty$ . Define  $\gamma_j$  to be the curve with parametric form:

$$(4.2.1) \quad \gamma_j(t) = [u(t, k_j), v(t, k_j)], \quad j = 1, \dots, n,$$

and similarly define the curve  $\delta_i$ :

$$(4.2.2) \quad \delta_i(t) = [u(h_i, t), v(h_i, t)], \quad i = 1, \dots, m.$$

The curves  $\gamma_j$  and  $\delta_i$  will be called *hinges*. They are sometimes called *knot lines* in the literature. A subset of  $\mathcal{A}$  bounded only by hinges will be called a *hinged subset*. A hinged subset which does not contain any other hinged subset as a proper subset will be called a *minimal hinged subset*. A real function defined on  $\mathcal{A}$  is a *spline function of degree  $p$*  if, when restricted to any minimal hinged subset, it is a bivariate polynomial of degree  $\leq p$ , at least one such polynomial having degree equal to  $p$ , and all derivatives (including mixed derivatives) of order  $\leq p-1$  are continuous on the whole of  $\mathcal{A}$ .

REMARKS. The function  $(u, v)$  establishes curvilinear coordinates in  $R$ . The set of curves  $\{\gamma_j, \delta_i\}$  are coordinate lines in these new coordinates. The remainder of the definition generalizes concepts involved in spline functions of a single variable. The correspondences between univariate and bivariate splines are as follows.

Univariate	Bivariate
Knot	Hinge
Interval between knots	Minimal hinged subset

Note that, because the map  $(u, v)$  is a bijection, i.e. a coordinate transformation, its inverse (call it  $g$ ) transforms the hinges into coordinate lines in  $R$ . Thus, instead of fitting a spline function with hinges  $\gamma_j, \delta_i$  to observations  $I(\bar{x}_i, \bar{y}_i)$  one might fit a spline function to observations  $\hat{I}(g(\bar{x}_i, \bar{y}_i))$  with hinges (in the  $xy$ -plane)  $x = h_i, i = 1, \dots, m$  and  $y = k_j, j = 1, \dots, n$ .

In this coordinate system, all minimal hinged subsets will be rectangular. In all subsequent analytical development, therefore, this rectangular structure will be assumed without any loss of generality.

It should be noted that, since the coordinate transformation is in general non-linear, a spline function fitted in the coordinate system in which hinges are rectilinear will not necessarily induce a spline function in the coordinate system of curvilinear hinges. Nevertheless, the function fitted in this latter coordinate system will be a reasonable interpolating function.

DEFINITION. Let  $(z)_+$  denote  $\max(z, 0)$ , and read  $(z)_+^p$  as  $[(z)_+]^p$ . An  $M$ -spline of degree  $p$  is a function (defined on the Euclidean  $xy$ -plane) which assumes one of the two forms  $(x-h)_+^p$  for some constant  $h$  or  $(y-k)_+^p$  for some constant  $k$ .

The  $M$ -splines are of use in constructing spline functions as the following result shows. Their univariate version is discussed by GREVILLE (1969, pp. 2-3), though not under that name.

PROPOSITION. Any spline function of degree  $p$  with hinges  $x = h_1, \dots, h_m$  cutting the  $x$ -axis and hinges  $y = k_1, \dots, k_n$  cutting the  $y$ -axis can be decomposed into a sum of:

- (i) a polynomial of degree  $\leq p$  on  $\mathcal{R}$ , and
- (ii) constant multiples of the  $M$ -splines  $(x-h_i)_+^p, i = 1, \dots, m$ ,
- (iii) constant multiples of the  $M$ -splines  $(y-k_j)_+^p, j = 1, \dots, n$ .

PROOF. See Appendix C

### 4.3. Choice of hinges

Certain criteria can be established for the determination of the hinges of bivariate spline functions to be fitted to a particular data set.

First, the more hinges are specified, the more parameters are to be estimated, since the spline function changes its polynomial form each time a hinge is crossed. Thus, the choice of too many hinges will lead to over-fitting, i.e. a “wrinkly” fitted surface. For this reason the number of hinges used should be minimized subject to adequate fit to data.

Second, the choice of hinges should be related in some way to the “flatness” of the data points. For example, if all data points assumed the same value, a constant polynomial would fit over the entirety of the relevant domain. There would be no purpose in choosing any hinges to allow the polynomial form to vary from one hinged subset to another.

Strictly, it is not “flatness” which matters but rather conformity of data points with a simple polynomial form. For example, if a quadratic spline is to be fitted and all data points lie on a quadratic surface, there would again be no purpose in choosing any hinges. However, this type of conformity of the data will often be difficult to verify by simple inspection.

Conversely to the second point, hinges will need to occur more densely in those subregions where the surface to be fitted is evidently changing more rapidly.

Thus, in general terms, hinges should be chosen to delineate “essentially different” parts of the surface, with greatest (resp. least) density in those areas where the surface is changing most (resp. least) rapidly. Parts of the surface which are apparently different can be identified from the map in Appendix G1. It can be seen in the following diagram that the hinges have been chosen to approximate the boundaries of these regions, isolating for example:

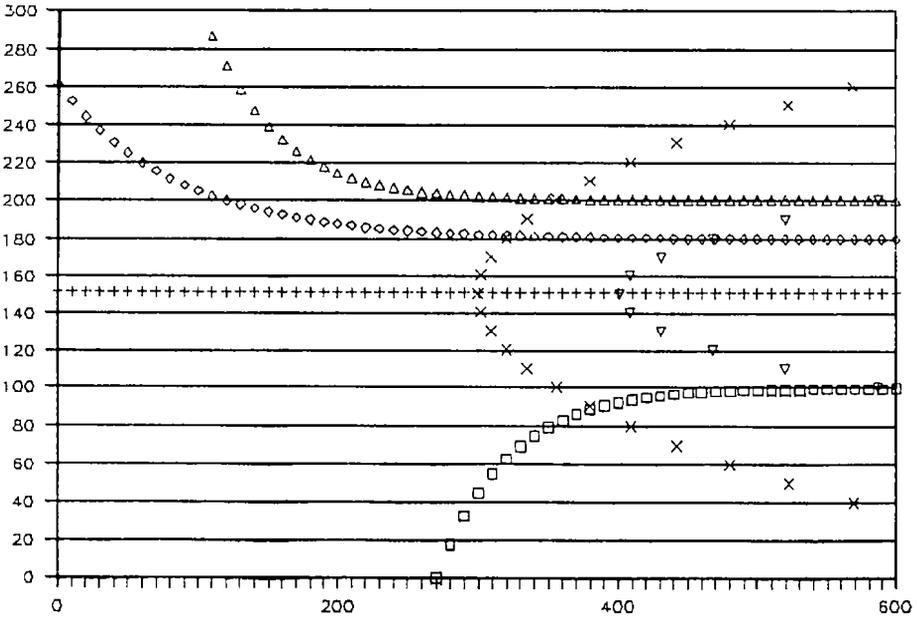
- (i) the north-east and south-east corners;
- (ii) the central east region around the harbour;
- (iii) the far west region.

With all this, however, it must be said that the choice of hinges actually adopted on any particular occasion remains very much a subjective one.

The following diagram provides a schematic representation of the (curvilinear) hinges chosen in relation to the fitting of operating ratio adjusted for region detailed in Appendix B. In fact, in the actual surface fitting procedure, some of these hinges were very slightly distorted, as described in Section 5.2.

The rectangular region covered by these coordinates is the region appearing in the diagrams of Appendix G, though the horizontal scale has been distorted relative to the vertical scale in the schematic representation.

The precise definition of these hinges is given in the numerical detail of Section 5.1. Initially, two further hinges, like flatter versions of the seemingly parabolic hinges to the right of the diagram, were considered to the left of those. However, the numbers of observations in the minimal hinged subsets so created were sufficiently small that the additional hinges were dropped.



**4.4. Fitting by regression**

The hinges illustrated in Section 4.3 were transformed to rectilinear hinges by means of the coordinate transformation set out in Section 5.2. The rectilinear hinges were:

$$\begin{matrix}
 & x = 300, & x = 400; \\
 y = 100, & y = 150, & y = 180, & y = 200.
 \end{matrix}$$

Then, by the spline decomposition result quoted in Section 4.2, quadratic and cubic spline functions can be written in the forms:

Quadratic spline:  $f(x, y) = \sum_{\substack{k, l=0 \\ k+l \leq 2}}^2 a_{kl} x^k y^l + \sum_{i=1}^m b_i (x-h_i)_+^2 +$   
 (4.4.1)  $\quad + \sum_{j=1}^n c_j (y-k_j)_+^2;$

Cubic spline.  $f(x, y) = \sum_{\substack{k, l=0 \\ k+l \leq 3}}^3 a_{kl} x^k y^l + \sum_{i=1}^m b_i (x-h_i)_+^3 +$   
 (4.4.2)  $\quad + \sum_{j=1}^n c_j (y-k_j)_+^3.$

The splines are fitted by estimation of the coefficients  $a_{kl}$ ,  $b_l$ ,  $c_j$ . Since  $f$  is linear in these unknowns, regression can be used to carry out the fit.

Regressions have been performed using the GLIM (Generalised Linear Interactive Modelling) system (PAYNE, 1985). Regression fits are carried out in this system by the method of maximum likelihood. It is assumed that each observation represents a drawing from a *gamma* distribution. This recognises the essential positivity of the sampled variable, operating ratio adjusted for region. Each observation is assigned a *weight*  $E_i$ . This means that the coefficient of variation of the gamma distribution associated with postcode  $i$  is taken to be

$$\text{const.}/E_i^{\frac{1}{2}},$$

where the const. term is independent of  $i$ .

The difficulties arising in the choice of an error distribution deserve some discussion. It may be reasonable to regard the amount of claims in each cell as a generalized Poisson variate. For a large expected number of claims, i.e. large  $E_i$ , this is known to approximate a gamma variate (SEAL, 1977). Hence operating ratio adjusted for region  $\rho_i^{(*)}$ , a scalar multiple of this claims amount (see (3.1.1) and (3.1.4)), will also be a gamma variate approximately.

Difficulties arise when  $E_i$  is small. In this case the distribution of  $\rho_i^{(*)}$  consists of a spike at zero, together with a continuous distributions on strictly positive support. No standard distribution provides a model for this

It is evidently extremely difficult to find an error distribution which provides an adequate representation of  $\rho_i^{(*)}$  at both large and small exposures, and is also computationally manageable for regression purposes. Certainly, the standard regression packages do not appear to provide for this

In the event, only a small minority of cells contained small  $E_i$ . The great majority contained  $E_i$  of at least some hundreds, probably sufficient to justify the adoption of the gamma error distribution.

When the gamma error distribution is used, it is natural that the *reciprocal* of the linear model (i.e. the *reciprocal of the spline function*) be fitted to the data. This is done by the GLIM package.

## 5. RESULTS

### 5.1. Hinges

The reasoning governing the selection of hinges is set out in Section 4.3, as is the general shape of those actually selected. The precise forms of the hinges, special cases of (4.2.1) and (4.2.2), are as follows:

$$(5.1.1) \quad \gamma_j(t) = \{t, k_j + \text{sgn}(k_j - 150) \exp[-0.0004 |k_j - 150| (t - 500)] + 5(k_j - 100)/3\}, \quad 0 \leq t \leq 600,$$

for  $j = 1, 2, 3, 4$ , with  $k_1 = 100$ ,  $k_2 = 150$ ,  $k_3 = 180$ ,  $k_4 = 200$ ; and

$$(5.1.2) \quad \delta_i(t) = \{h_i + 600000(t - 150)^2 / (600 - h_i)^3, t\}, \quad 0 \leq t \leq 300,$$

for  $i = 1, 2$ , with  $h_1 = 300$ ,  $h_2 = 400$ , where  $\text{sgn}(\cdot)$  is defined by:

$$\begin{aligned} \text{sgn}(x) &= +1, x > 0, \\ &= 0, x = 0; \\ &= -1, x < 0. \end{aligned}$$

The functional forms needed to produce hinges of the right shape are evidently complicated, as they will be in most practical implementations. Discovery of these forms, and production of the associated coordinate transformations (Section 5.2), are the only non-routine, and hence difficult, parts of the whole fitting procedure.

### 5.2. Coordinate transformations

As remarked just after the definition of a spline function in Section 4.2, the hinges  $\delta_i$  are the images under the coordinate transformation  $(u, v)$  of the coordinate lines  $x = h_i$  in the  $xy$ -plane. Similarly, the hinges  $\gamma_j$  are the images of the coordinate lines  $y = k_j$ .

Comparison of (4.2.2) with (5.1.2) indicates that, along the hinge  $\delta_1$ ,

$$u(x, y) = x + 600000 [v(x, y) - 150]^2 / (600 - x)^3.$$

For convenience, write  $x', y'$  for the transformed coordinates induced by  $x, y$ . Then, along  $\delta_1$ ,

$$(5.2.1) \quad x' = x + 600000 (y' - 150)^2 / (600 - x)^3.$$

Similarly, along  $\gamma_1$ ,

$$(5.2.2) \quad y' = y + \text{sgn}(y - 150) \exp[-0.0004 |y - 150| (x' - 500 + 5(y - 100)/3)].$$

Now (5.2.1) and (5.2.2) together do not give a coordinate transformation in a convenient form since they give:

$$\begin{aligned} (x, y') &\mapsto x' \\ (x', y) &\mapsto y', \end{aligned}$$

and not

$$(5.2.3) \quad (x, y) \mapsto (x', y'),$$

as required.

In the present case this difficulty can be overcome by using the fact (from (5.2.2)) that  $y' \sim y$  for large  $x'$ . The coordinate transformation chosen is therefore:

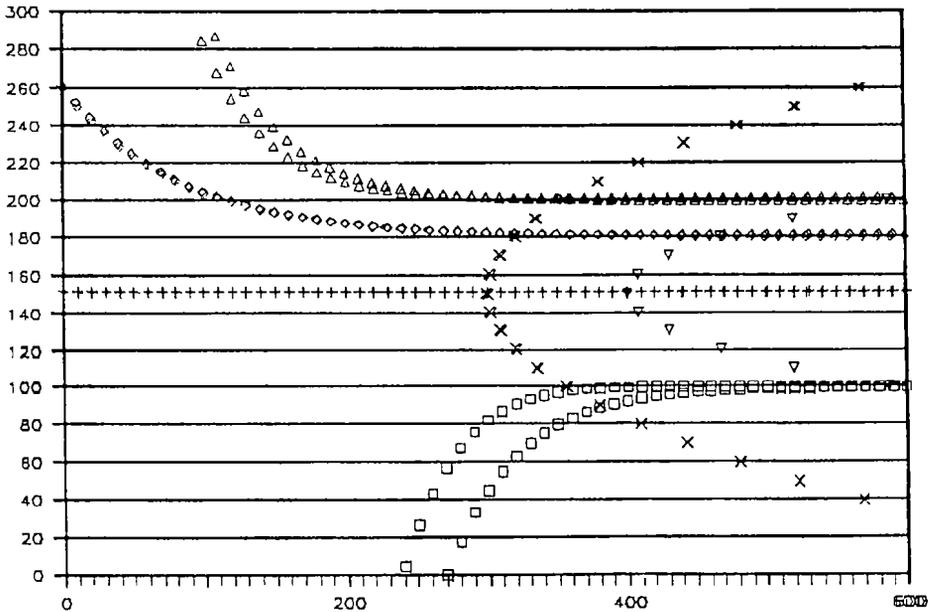
$$(5.2.4) \quad x' = x + 600000 (y - 150)^2 / (600 - x)^3;$$

$$(5.2.2) \quad y' = y + \text{sgn}(y - 150) \exp[-0.0004 |y - 150| (x' - 500 + 5(y - 100)/3)].$$

Equation (5.2.4) is of the form required by (5.2.3). If  $x'$  in (5.2.2) is expressed in terms of  $x, y$  by means of (5.2.4), then (5.2.2) is also in the form required by (5.2.3).

The means of converting each  $(x', y')$  pair used in defining a postcode (Appendix A) to a point  $(x, y)$  is given in Appendix D2.

Naturally, the approximation of  $y'$  by  $y$  in (5.2.4) will distort the hinges defined by  $x = h_i$  and  $y = k_j$  respectively. However, for the points  $x', y'$  with larger values of  $x'$  and smaller values of  $y'-150$ , which are primarily the ones where the definition of the hinges needs to be reasonably precise (see the diagram in Section 4.3), the distortion will be small. This is illustrated by the following diagram which displays the hinges  $x = 300, 400$  and  $y = 100, 150, 180, 200$  in the  $x' y'$ -plane, as produced by the coordinate transformation (5.2.4) and (5.2.2), and overlays them on the diagram of the desired hinges illustrated in Section 4.3. The difference between the two sets of hinges is very small for practical purposes, and in a number of respects the two are quite indistinguishable.



It should be pointed out here that (5.2.4) and (5.2.2) do not in fact provide a coordinate transformation of  $\mathcal{A}$ . The appearance of the terms  $\text{sgn}(y-150)$  and  $|y-150|$  in (5.2.2) produces discontinuities in the gradients of the expression given there for  $y'$ . However, (5.2.4) and (5.2.2) do provide separate coordinate transformations of the two subregions of  $\mathcal{A}$  defined by the constraints  $y \geq 150$  and  $y \leq 150$  respectively. It is evident from (5.2.2) that these subregions are mapped to  $y' \geq 150$  and  $y' \leq 150$  respectively. It follows that the line  $y = 150$  is mapped to  $y' = 150$ .

### 6. THE SPLINE SURFACES

Section 4.4 gives the algebraic forms of the quadratic and cubic splines whose reciprocals are to be fitted to the data [see (4.4.1) and (4.4.2)]. These are written

in terms of  $(x, y)$  coordinates indicating that the independent coordinates are from the domain in which the hinges are rectilinear, as assumed in the formulas given. Details of the regression model used in the fit are also given in Section 4.4

The function to which the reciprocal spline is fitted in Section 4.4 is taken as  $I(x, y)$ , the operating ratio adjusted for region, as experienced at  $(x, y)$ . As noted in (3.1.5), observations  $\hat{I}(x, y)$  on this function are taken as available at the centroids  $(\bar{x}_i, \bar{y}_i)$  of the various postcodes. A sample of observations is listed in Appendix B.

Note that these centroids must be expressed in terms of the  $(x, y)$  coordinates. This has been done by means of the following procedure:

- (i) for each vertex  $(x', y')$  of postcode  $i$  listed in Appendix A1, calculate the corresponding coordinates  $(x, y)$ ,
- (ii) calculate  $(\bar{x}_i, \bar{y}_i)$  as the average of all the vertex coordinates  $(x, y)$  related to postcode  $i$ .

The results of these computations are sampled in Appendix E. Note that, because the transformation between the  $(x, y)$  and  $(x', y')$  coordinate systems is non-linear, postcode polygons in the  $(x', y')$  coordinates will not have rectilinear sides in the  $(x, y)$  representation. Therefore, the evaluation of a centroid as the average of the vertex coordinates will involve some error. Since most postcodes are small in area, especially where the curvature of the coordinate transformation is greatest, this error will be small and probably negligible.

Appendix E summarizes a small sample of the data used in the spline-fitting regressions.

The results of the regressions are as follows.

QUADRATIC SPLINE

Term	Spline coefficient	
	Fitted value	Standard error
const	1 555	1 659
$x$	$-6\ 247 \times 10^{-3}$	$2\ 830 \times 10^{-3}$
$y$	$4\ 236 \times 10^{-2}$	$3\ 725 \times 10^{-2}$
$x^2$	$1\ 297 \times 10^{-2}$	$5\ 095 \times 10^{-6}$
$xy$	$1\ 387 \times 10^{-5}$	$1\ 253 \times 10^{-5}$
$y^2$	$-3\ 648 \times 10^{-4}$	$2\ 142 \times 10^{-4}$
$(x-300)^2$	$-4\ 326 \times 10^{-5}$	$1\ 996 \times 10^{-5}$
$(x-400)^2$	$1\ 312 \times 10^{-5}$	$5\ 250 \times 10^{-5}$
$(y-100)^2$	$5\ 741 \times 10^{-4}$	$3\ 003 \times 10^{-4}$
$(y-150)^2$	$-1\ 512 \times 10^{-4}$	$2\ 483 \times 10^{-4}$
$(y-180)^2$	$4\ 544 \times 10^{-4}$	$4\ 319 \times 10^{-4}$
$(y-200)^2$	$-5\ 708 \times 10^{-4}$	$3\ 876 \times 10^{-4}$

It may be noted that a number of the coefficients here are not statistically significant. This fact is taken no further here, but will be referred to again below in relation to the fitted cubic spline

## CUBIC SPLINE

Term	Spline coefficient	
	Fitted value	Standard error
const	-1 681	4 916
$x$	$1\ 300 \times 10^{-3}$	$2\ 098 \times 10^{-2}$
$y$	$1\ 287 \times 10^{-2}$	$1\ 519 \times 10^{-1}$
$x^2$	$1\ 198 \times 10^{-4}$	$4\ 694 \times 10^{-5}$
$xy$	$-2\ 965 \times 10^{-4}$	$1\ 610 \times 10^{-4}$
$y^2$	$5\ 945 \times 10^{-4}$	$1\ 837 \times 10^{-3}$
$x^3$	$-2\ 016 \times 10^{-7}$	$5\ 861 \times 10^{-8}$
$x^2y$	$9\ 316 \times 10^{-8}$	$1\ 352 \times 10^{-7}$
$xy^2$	$7\ 772 \times 10^{-7}$	$3\ 412 \times 10^{-7}$
$y^3$	$-3\ 670 \times 10^{-6}$	$7\ 013 \times 10^{-6}$
$(x-300)_+^3$	$5\ 835 \times 10^{-7}$	$2\ 167 \times 10^{-7}$
$(x-400)_+^3$	$-1\ 744 \times 10^{-6}$	$8\ 728 \times 10^{-7}$
$(y-100)_+^3$	$7\ 390 \times 10^{-6}$	$9\ 671 \times 10^{-6}$
$(y-150)_+^3$	$-3\ 353 \times 10^{-6}$	$7\ 737 \times 10^{-6}$
$(y-180)_+^3$	$5\ 155 \times 10^{-6}$	$1\ 270 \times 10^{-5}$
$(y-200)_+^3$	$7\ 855 \times 10^{-6}$	$1\ 102 \times 10^{-5}$

As was the case with the fitted quadratic spline, many of the terms in the cubic spline are not statistically significant. It is possible to eliminate these from the fit. Experimentation with the elimination of insignificant variables led to the following cubic spline function.

Term	Spline coefficient	
	Fitted value	Standard error
const	3 742	0 4087
$x$	$-3\ 229 \times 10^{-2}$	$6\ 474 \times 10^{-3}$
$x^2$	$1\ 695 \times 10^{-4}$	$3\ 306 \times 10^{-5}$
$x^3$	$-2\ 542 \times 10^{-7}$	$5\ 132 \times 10^{-8}$
$y^3$	$-5\ 215 \times 10^{-7}$	$6\ 686 \times 10^{-8}$
$(x-300)_+^3$	$6\ 871 \times 10^{-7}$	$2\ 051 \times 10^{-7}$
$(x-400)_+^3$	$-1\ 857 \times 10^{-6}$	$8\ 625 \times 10^{-7}$
$(y-100)_+^3$	$3\ 401 \times 10^{-6}$	$5\ 307 \times 10^{-7}$
$(y-150)_+^3$	$-3\ 884 \times 10^{-6}$	$9\ 090 \times 10^{-7}$

Some further statistics related to the regression models, particularly concerning goodness of fit, are of interest. These appear in the following table.

The meaning of the estimated coefficients of variation is as follows. For the largest postcodes, with exposures in excess of 4000, the coefficient of variation of the adjusted operating ratio is about 15%. For a relatively small postcode with 100 years of exposure, the coefficient of variation is about 100%.

Spline surface	Estimated coefficient of variation of operating ratio adjusted for region (a)	Coefficient of determination of regression (b)	
		Unadjusted %	Adjusted %
Quadratic	10.7	39	35
Cubic			
full model	10.1	46	41
reduced model	10.1	44	41

Notes (a) This coefficient of variation relates to a single year of exposure. The corresponding coefficient for  $E_i$  exposure years is this figure adjusted by a factor of  $E_i^{-1}$   
 (b) The adjusted coefficient of determination is defined as (SEBER, 1977, pp 362-363)

$$1 - nS^2/(n-p),$$

where

$$S^2 = \frac{\text{residual sum of squares of the regression model fit}}{\text{residual sum of squares of a constant model fit}}$$

$$= 1 - \text{unadjusted coefficient of determination}$$

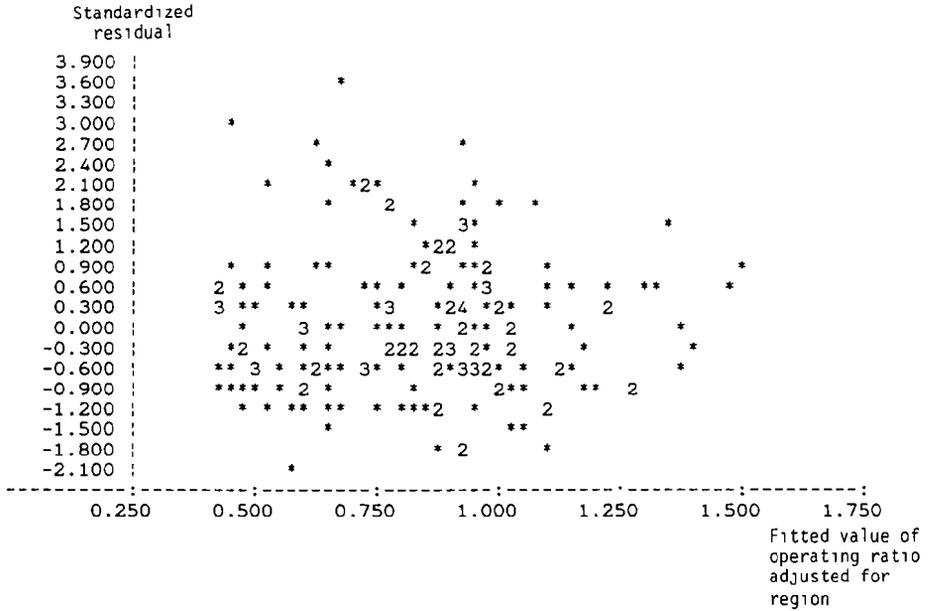
These coefficients of variation seem large, but perhaps not unrealistically so. For example, if each cell were Poisson distributed the coefficient of variation for a single year's exposure would be 1. Since it is fair to assume that there is variation in the mean claim frequency associated with individuals within a cell, it may be inferred that the coefficient of variation of claim frequency associated with a single year's exposure will be larger than 1 (see e.g. SEAL, 1969, p. 25). It is also known that the distribution of Contents insurance claim sizes tends to be long tailed. When additional allowance is made for this component of variation, it may be that the actual coefficient of variation of the adjusted operating ratio approaches the value estimated from the data. Otherwise, the conclusion would be that the spline fit is inadequate, and its deviation from the true underlying adjusted operating ratio manifests itself as a spurious increase in random noise.

The unadjusted coefficients of determination show the proportion of variation in the data explained by the fitted spline surfaces.

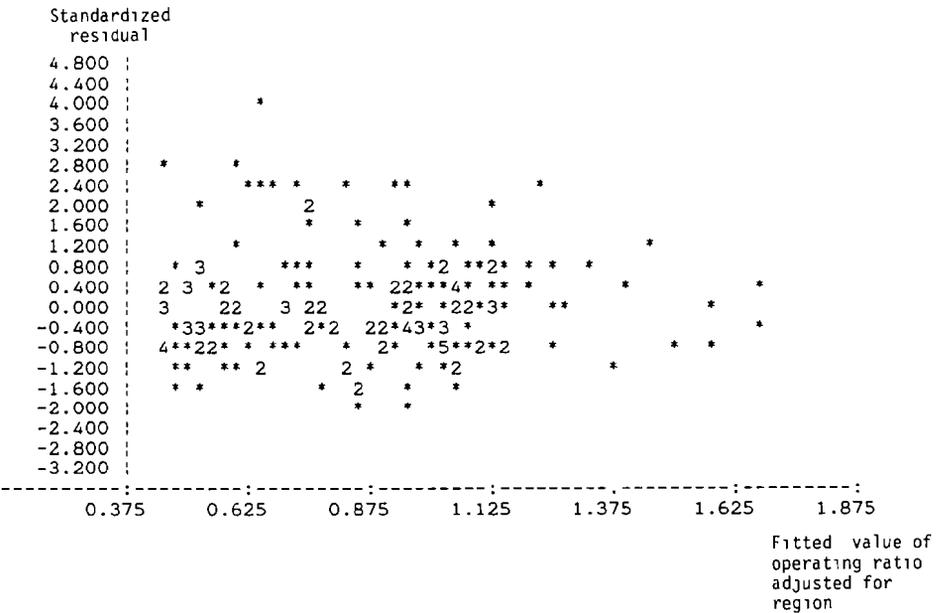
As explained by SEBER (1977, p. 363), the unadjusted coefficients of determination of regressions involving different numbers of regressors are not comparable. The adjusted coefficient of determination is intended to make for comparability.

On the basis of these statistics there seems little to choose between the two cubic splines, both of which appear somewhat superior to the quadratic spline. The quadratic spline was dropped from the final process of selecting rating regions.

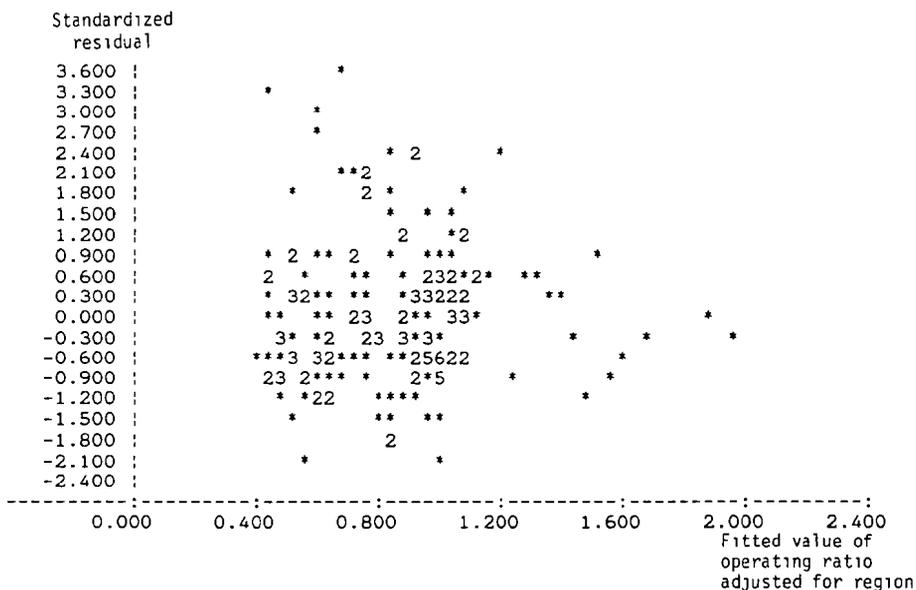
QUADRATIC SPLINE



FULL CUBIC SPLINE



REDUCED CUBIC SPLINE



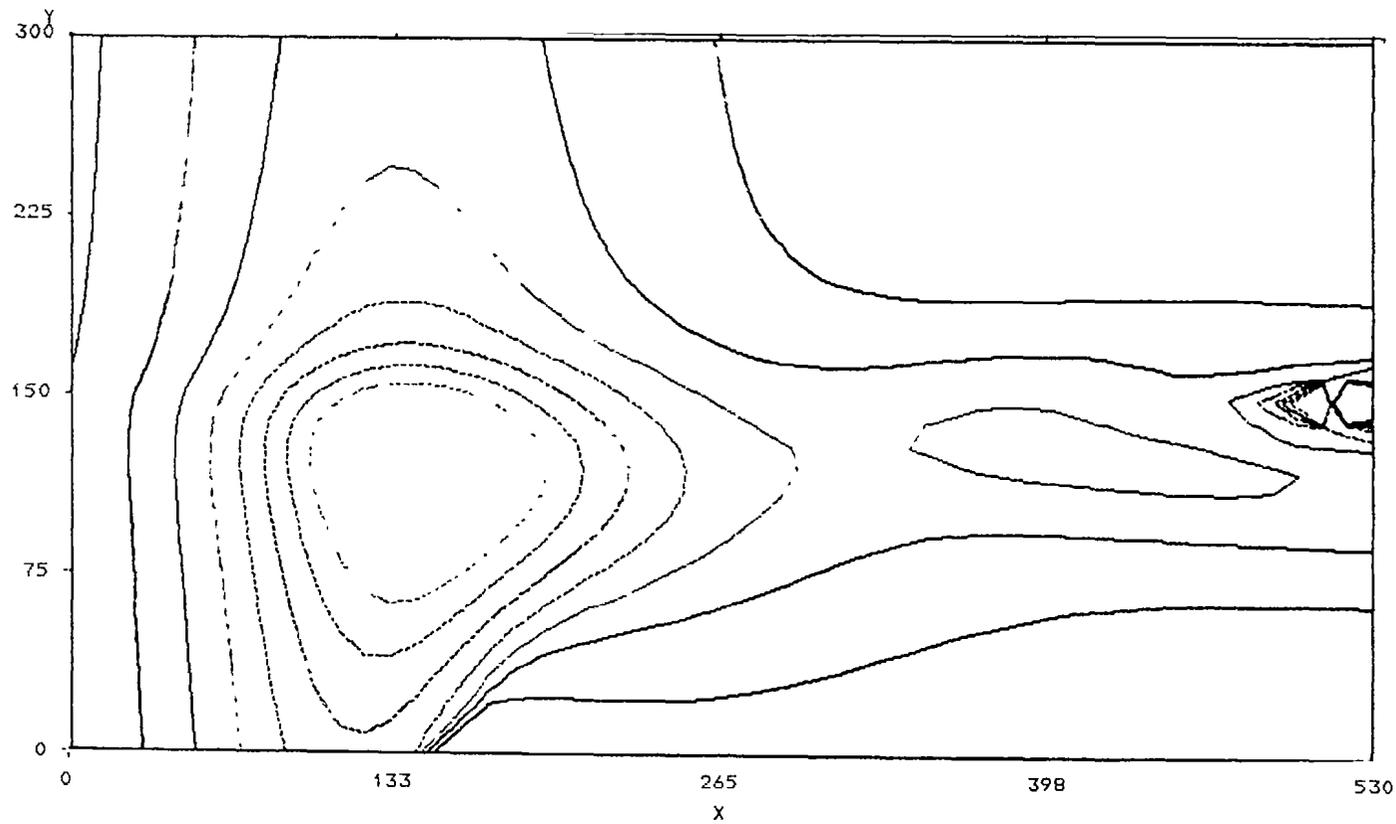
A small sample of values of operating ratio adjusted for region fitted by each of the three splines is displayed in Appendix F, together with the standardized residuals in each case. Plots of these standardized residuals against the fitted values appear above. In these plots, a \* in the (x, y) position denotes occurrence of a standardized residual of y in the case that operating ratio adjusted for region is x; a 2 in the (x, y) position is equivalent to two \*s there; a 3 equivalent to three \*s, etc.

The plots appear reasonable. There is perhaps a hint that, for constant exposure, coefficient of variation decreases with increasing operating ratio. It might have been feared that the spline surface would tend to flatten out real eccentricities in operating ratio. There is, however, no evidence that the spline surfaces tend to under-estimate (resp. over-estimated) at the upper (resp. lower) extreme of operating ratios.

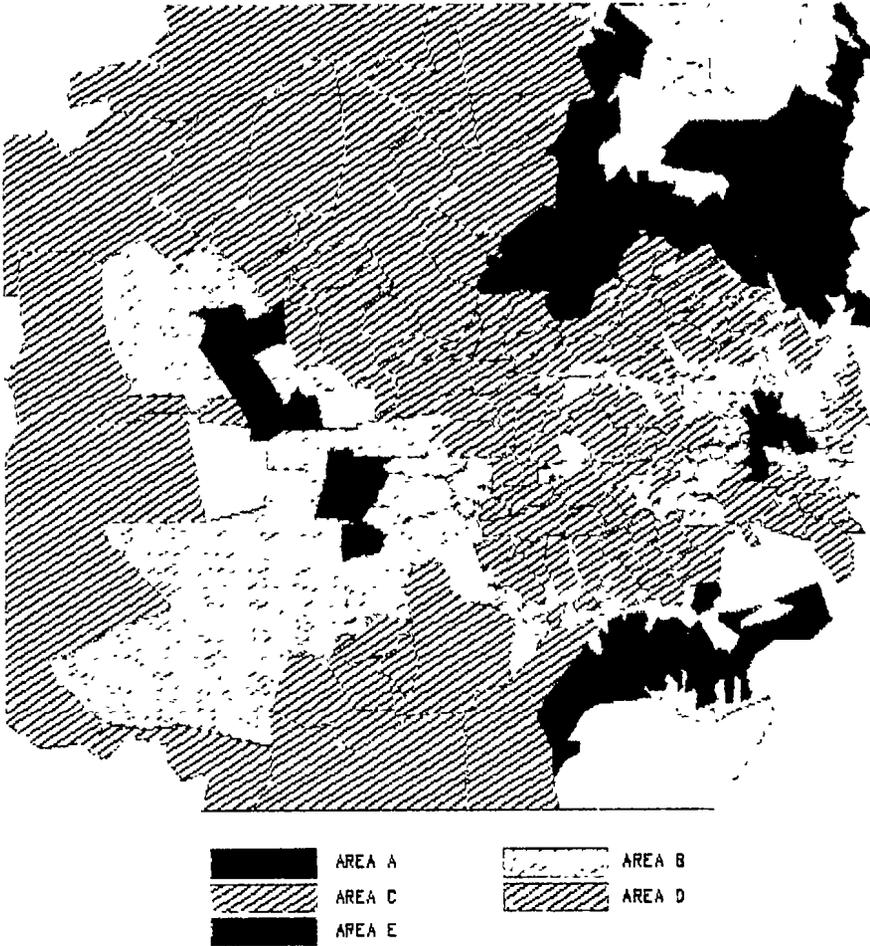
In the case of each of the cubic splines, maps of the total rating region were produced showing the division into different ranges of adjusted operating ratio as estimated by the spline function in question. The map for the reduced cubic spline appears in Appendix G2. This may be compared with the corresponding map in Appendix G1 which shows the division into different ranges of adjusted operating ratio as observed.

Contour maps of the two cubic spline surfaces were also produced. These may be used to select rating regions. The map relating to the reduced cubic spline is reproduced below.

# RATIO-TWO



## NEW RATING AREAS



The contours indicate five regions of steep gradient which divide off five clear rating areas:

- (i) the north-eastern suburbs (low risk);
- (ii) the south-eastern suburbs (low risk);
- (iii) the far western suburbs (low risk);
- (iv) a small pocket of certain eastern suburbs (high risk);
- (v) the central western suburbs (high risk).

The remainder of the total region would then provide a sixth rating region.

In practice the task would be completed by using the contours to determine

boundaries of the six rating regions, which would then be treated, for rating purposes, as homogeneous with respect to operating ratio.

It is of interest to compare these suggested rating regions with those actually used by the insurer concerned. Appendix G3 displays the regions in use during 1985/86, the period to which the data of the present paper relate, the immediately preceding map, in which Areas A to E are in descending order to risk, displays the regions currently in use.

This map does indeed identify most of the rating regions suggested by the splines. Moreover, a comparison with Appendix G3 shows that in the recent past the insurer concerned has considerably expanded its high risk region in the central west. The splines identified the need for this from 1985/86 data, i.e. at least 17 months before it actually occurred.

The main differences between the rating regions suggested by the splines and those actually currently in use are:

- (i) the actual regions do not identify any of the eastern suburbs as of particularly high risk, whereas the splines do;
- (ii) the actual regions identify Sydney city as high risk, whereas the splines do not;
- (iii) the actual regions identify a corridor of relatively high risk inner western suburbs, whereas the splines interpret this in a relatively minor way.

Reference to Appendix G1 (the data) can assist in resolving these disparities, although one must remember that Appendix G1 gives no indication of the exposure, and therefore the statistical significance, of each of the postcodes mapped.

However, such a comparison suggests the following conclusions.

First, the splines are probably correct in identifying a very high risk pocket of eastern suburbs.

Second, the splines are probably wrong in their treatment of Sydney city and some of the innermost suburbs. This may be indicative of splines' failure, as locally low degree polynomials, to respond to highly localized steep gradients.

## 7. CONCLUSION

*Spline functions can provide an effective means of determining geographic regions for premium rating. Most of the implementation is routine. The exception to this is the determination of a suitable set of curvilinear coordinates, and the transformations which take these coordinates to and from rectangular coordinates. This step can be difficult and time-consuming.*

## 8. ACKNOWLEDGEMENTS

The great bulk of the computational work involved in this paper was carried out by Dr. K. H. CHEN, whose assistance is gratefully acknowledged.

Thanks are also due to NRMA Insurance Ltd for the use of their SAS package.

APPENDIX A  
DATA

**A1. Geographic coordinates of postcodes**

Postcode boundaries have been approximated by polygons. These have been defined by the geographic coordinates of the vertices of the polygons.

These coordinates take the form of a list of  $(x, y)$ -coordinates for each postcode, a small sample of which follows.

Post code	Coordinates		Post code	Coordinates		Post code	Coordinates		
	$x$	$y$		$x$	$y$		$x$	$y$	
2000	456	162	2011	477	150	2020	446	178	
	460	157		473	148		433	188	
	463	153		471	148		434	193	
	467	150		468	150		442	193	
	468	145		467	150		449	202	
	464	147		463	153		455	201	
	464	145		2015	444		172	453	199
	462	144			451		173	451	199
	461	146			446		178	447	193
	459	146			457		178	456	192
	456	143			460		177	464	186
	451	145			460		173	464	181
	455	154			460		171	457	181
	451	153			456		165	457	178
	451	155			454		165	446	178
456	159	448	167		2021	473	164		
456	162	451	169			475	167		
2006	447	164	2016			444	172	478	168
	451	160				456	165	485	167
	442	161				469	165	489	166
	443	162				469	162	489	163
447	163	464		162		481	159		
2007	451	155		456		162	483	156	
	456	159		456		162	479	155	
	451	160		454		164	473	155	
	447	155		454		165	468	158	
	447	155		456		165	473	164	
	451	155							

**A2. Claims data (sample only)**

Post code	Years of exposure	Average sum insured	Jewellery penetration	Average jewellery sum insured	Number of claims	Average claim cost	Average earned premium	Average observed risk premium	Average gross experience premium
		\$ 000's	%	\$		\$	\$	\$	\$
2006	3	44 667	0 00		2	1300	290 80	865 92	935 98
2008	66	21 388	10 67	1734	9	4293	267 73	585 13	608 65
2749	45	23 329	9 52	1562	6	2974	103 51	395 17	427 29
2027	340	42 688	10 51	15048	35	2865	208 52	295 54	316 26
2157	226	28.816	5 22	3435	19	2189	119 60	184 30	208 78
2171	841	25 920	5 75	2978	106	1510	140 31	190 84	216 02
2759	393	24 307	9 18	1973	42	1414	110 34	151 80	179 95
2115	381	24 129	8 09	2305	37	1608	108 28	156 80	176 94
2177	314	22 368	4 86	2690	41	1595	183 19	209 02	232 24
2025	274	33 654	9 67	5551	32	1671	168 03	196 00	218 04
2761	115	22 253	4 03	2404	18	1243	182 60	194 81	228 10
2026	1153	22 432	11 03	3540	102	1921	144 28	170 68	190 13

**A3. Existing premium rating system**

Each metropolitan postcode is assigned to one of 5 rating regions. For these regions, the existing premium formula has been taken as the following:

Rating region	Basic (i.e. non-jewellery) premium		Jewellery premium
	Base premium	Premium per \$ 1000 basic sum insured	Premium per \$ 1000 jewellery sum insured
	\$	\$	\$
A	130	3 60	20 00
B	94	2 70	20 00
C	72	2 00	15 00
D	49	2 00	15 00
E	27	2 00	10 00

In fact, some 5% to 10% of policies were subject to a loading of 33% on these rates, but this fact has been ignored in the following.

APPENDIX B  
OPERATING RATIO ADJUSTED FOR REGION

The following results relate to the same sample of postcodes as appears in Appendix A2.

Rating region	Post code	Operating ratio	Adjustment to operating ratio for region				Loss ratio adjusted for region		
			1st version (3 1 2)		2nd version (3 1 6)		1st version based on (3 1 2)	2nd version based on (3 1 6)	Ratio of 1st version to 2nd
			Numerator	Denominator (Area C)	Numerator	Denominator (Area C)			
a	2006	386.2%	290.80	161.33	225.76	128.26	696.2%	679.9%	1.024
a	2008	272.8%	210.70	117.55	225.76	128.26	489.0%	480.2%	1.018
d	2749	495.4%	97.89	120.89	103.25	128.26	401.1%	398.8%	1.006
c	2027	182.0%	181.10	181.10	126.25	128.26	182.0%	179.2%	1.016
d	2157	209.5%	109.32	132.32	103.25	128.26	173.1%	168.6%	1.026
c	2171	184.8%	126.41	126.41	126.25	128.26	184.8%	181.9%	1.016
d	2759	195.7%	100.33	123.33	103.25	128.26	159.2%	157.5%	1.011
d	2115	196.1%	100.06	123.06	103.25	128.26	159.4%	157.9%	1.010
b	2177	152.1%	157.01	118.70	167.17	128.26	201.2%	198.3%	1.015
c	2025	155.7%	147.36	147.36	126.25	128.26	155.7%	153.3%	1.016
b	2761	149.9%	156.02	117.96	167.17	128.26	198.3%	195.4%	1.015
c	2026	158.1%	122.72	122.72	126.25	128.26	158.1%	155.7%	1.016

APPENDIX C  
PROOF OF SPLINE DECOMPOSITION

PROOF OF THE PROPOSITION IN SECTION 4.2 Consider any particular spline function  $f(x, y)$ . Define the polynomial ( $i$ ) in the statement of the Proposition to be the extension to  $\mathcal{A}$  of  $f(x, y)$  for  $0 \leq x \leq h_1, 0 \leq y \leq k_1$ . Call this polynomial  $p(x, y)$ . Now consider the spline function for  $h_1 \leq x \leq h_2, 0 \leq y \leq k_1$ . It is a polynomial of degree  $\leq p$  on this region, and therefore so is  $f(x, y) - p(x, y) = q(x, y)$ , say. Then  $q(x, y)$  can be written as a linear combination of terms  $x^a y^b, a + b \leq p$ . By a change of origin (on the  $x$ -axis),  $q(x, y)$  can be written alternatively as a linear combination of terms  $(x - h_1)^a y^b$ . Thus,  $q(x, y)$  as a function over the region  $0 \leq x \leq h_2, 0 \leq y \leq k_1$  is a linear combination of terms  $(x - h_1)_+^a y^b$ .

Now recall the continuity requirements on the derivatives of a spline function. These imply continuity of all derivatives of  $q(x, y)$  of order  $< p$ . Suppose  $a < p$  and consider  $(\partial^a / \partial x^a) [(x - h_1)_+^a y^b]$ . It is simple to verify that this derivative does not exist at  $x = h_1, y > 0$ . Thus  $q(x, y)$  reduces to a multiple of  $(x - h_1)_+^p$ .

The traversal of other hinges can be dealt with in precisely the same way. Traversal of each hinge  $x = h_i$  (in the positive direction) adds a multiple of  $(x - h_i)_+^p$  to the spline function. Traversal of each hinge  $y = k_j$  (in the positive direction) adds a multiple of  $(y - k_j)_+^p$ .

APPENDIX D  
COORDINATE TRANSFORMATIONS

**D1. Methodology**

The coordinates against which actual claims data are recorded are denoted by  $(x', y')$ . This is the coordinate system in which the postcode boundaries of Appendix A are defined; and also in which the hinges illustrated in Sections 4.3 and 5.2 are defined.

An alternative coordinate system in which these hinges are rectilinear uses coordinate pairs denoted by  $(x, y)$  (Sections 4.1 and 5). The transformation  $(x, y) \mapsto (x', y')$  is given by (5.2.4) and (5.2.2)

The inverse transformation  $(x', y') \mapsto (x, y)$  for each of the pairs  $(x, y)$  listed in Appendix A1 is given in Appendix D2. The inversion has been carried out numerically, as follows.

For convenience, let  $f$  denote the function  $(u, v)$ , such that

$$(x', y') = f(x, y).$$

Suppose that it is necessary to solve for  $x, y$  in:

$$(D1.1) \quad f(x, y) = (x'_0, y'_0),$$

for particular values of  $x'_0, y'_0$ . Note that, for another coordinate pair  $(x^*, y^*)$  in the  $xy$ -plane,

$$(D1.2) \quad f(x^*, y^*) - (x'_0, y'_0) = [(x^* - x)(\partial u / \partial x) + (y^* - y)(\partial u / \partial y), (x^* - x)(\partial v / \partial x) + (y^* - y)(\partial v / \partial y)],$$

to first order, where all derivatives are evaluated at  $(x, y)$ .

It would be possible at this point to use Newton's algorithm to obtain a sequence of iterations of  $(x^*, y^*)$  converging to the required  $(x, y)$ . However, this algorithm would involve the partial derivatives of  $f$ , rather messy expressions obtained from (5.2.2) and (5.2.4). To avoid this messiness, Newton's algorithm has been very slightly modified by replacing the partial derivatives by discretized versions of them as follows:

$$(D1.3) \quad f(x+a, y) - (x'_0, y'_0) = [a(\partial u / \partial x), a(\partial v / \partial x)];$$

$$(D1.4) \quad f(x, y+b) - (x'_0, y'_0) = [b(\partial u / \partial y), b(\partial v / \partial y)].$$

Now substitution of the right sides of (D1.3) and (D1.4) in (D1.2) yields

$$(D1.5) \quad (u^*, v^*) = [(x^* - x)(u_a, v_a)/a + (y^* - y)(u_b, v_b)/b],$$

where

$$(u^*, v^*) = f(x^*, y^*) - (x'_0, y'_0),$$

$$(u_a, v_a) = f(x+a, y) - (x'_0, y'_0);$$

$$(u_b, v_b) = f(x, y+b) - (x'_0, y'_0).$$

Equation (D1.5) represents two simultaneous equations,

$$(D1.6) \quad (bu_a)(x^* - x) + (au_b)(y^* - y) = abu^*$$

$$(D1.7) \quad (bv_a)(x^* - x) + (av_b)(y^* - y) = abv^*$$

It is now possible to obtain the solution  $x^* - x$ ,  $y^* - y$  to (D1.6) and (D1.7)

$$(D1.8) \quad x^* - x = a(u^*v_b - v^*u_b)/\Delta,$$

$$(D1.9) \quad y^* - y = -b(u^*v_a - v^*u_a)/\Delta,$$

where

$$(D1.10) \quad \Delta = \begin{vmatrix} u_a & u_b \\ v_a & v_b \end{vmatrix}.$$

Still working to first order only, (D1.8) and (D1.9) yield

$$(D1.11) \quad x = x^* - a(u^*v_b - v^*u_b)/\Delta,$$

$$(D1.12) \quad y = y^* + b(u^*v_a - v^*u_a)/\Delta.$$

The whole procedure is made iterative, by letting  $(x^{(n)}, y^{(n)})$ , the  $n$ -th approximation to  $(x, y)$ , replace  $(x^*, y^*)$ . Then the adaptation of (D1.11) and (D1.12) yields:

$$(D1.13) \quad x^{(n+1)} = x^{(n)} - a[u^{(n)}v_b^{(n)} - v^{(n)}u_b^{(n)}]/\Delta^{(n)},$$

$$(D1.14) \quad y^{(n+1)} = y^{(n)} + b[u^{(n)}v_a^{(n)} - v^{(n)}u_a^{(n)}]/\Delta^{(n)},$$

where

$$(D1.15) \quad (u^{(n)}, v^{(n)}) = f(x^{(n)}, y^{(n)}) - (x'_0, y'_0),$$

$$(D1.16) \quad (u_a^{(n)}, v_a^{(n)}) = f(x^{(n)} + a, y^{(n)}) - (x'_0, y'_0);$$

$$(D1.17) \quad (u_b^{(n)}, v_b^{(n)}) = f(x^{(n)}, y^{(n)} + b) - (x'_0, y'_0).$$

The recursion is initialized by:

$$(D1.18) \quad (x^{(0)}, y^{(0)}) = (x'_0, y'_0).$$

It should also be noted, as remarked in Section 5.2, that (5.2.4) and (5.2.2) do not provide a proper coordinate transformation, but rather separate coordinate transformations of the  $xy$ -subplanes defined by the constraints  $y \geq 150$  and  $y \leq 150$  respectively. When  $y' = 150$ , the inverse transformation gives  $y = 150$ , indeed  $(x, y) = (x', y')$ .

## D2. Transformed coordinates

For each of the pairs of coordinates  $(x', y')$  defining the polygonal boundary of a postcode (given as  $(x, y)$  in Appendix A1), the inverse transformed coordinates  $(x, y)$  were obtained by the recursive algorithm (D1.13)–(D1.18).

Occasionally, when the algorithm failed to produce convergence, or failed to produce a sufficient rate of convergence, manual intervention reinitialized the procedure. For example, if the sequence of  $(x^{(n)}, y^{(n)})$  oscillated with period 2, the procedure was restarted with

$$(x^{(0)}, y^{(0)}) = [(x^{(n)} + x^{(n+1)})/2, (y^{(n)} + y^{(n+1)})/2].$$

## APPENDIX E

## DATA FOR SPLINE-FITTING REGRESSIONS

Section 6 explains how the postcode vertex coordinates of Appendix A1 are used to compute postcode centroids in the alternative coordinate systems. These results are listed below for various postcodes. The additional data required in respect of each postcode for input to the regression have also been listed. They are drawn from the same source as Appendix B.

Note that the first set of coordinates appearing in the following table, while given for interest, are not used in the regression

Postcode $i$	Centroid coordinates				Years of exposure $E_i$	Observed loss ratio adjusted for region $l(\bar{x}_i, \bar{y}_i)$
	Given coordinate system $\bar{x}_i$ $\bar{y}_i$		Coordinate system of rectilinear hinges $\bar{x}_i$ $\bar{y}_i$			
2000	458.8	150.9	453.7	151.0	83	0.275
2006	446.0	162.0	430.5	161.2	3	6.962
2007	450.5	156.5	444.9	155.6	48	1.343
2008	452.4	162.0	435.4	161.2	66	4.890
2009	448.4	152.1	447.7	151.6	26	0.337
2010	465.2	158.8	452.2	158.0	319	2.164
2011	471.0	151.5	469.8	151.2	329	1.529
2015	452.6	171.7	412.3	171.1	123	1.411
2016	459.3	163.6	436.6	162.8	134	1.112
2017	462.1	168.6	425.4	168.0	49	0.262
2018	466.5	178.7	403.3	178.4	337	0.769
2019	470.3	193.0	374.4	192.9	284	0.608

APPENDIX F  
OPERATING RATIOS FITTED BY SPLINE SURFACES

The following table displays, for each postcode appearing in Appendix E, the observed operating ratio adjusted for region. This is accompanied by the corresponding operating ratio fitted by each of the spline surfaces described in Section 6 and the standardized residual.

The standardized residuals (for a gamma error term) are calculated according to the formula

$$\text{standardized residual} = \frac{(\text{observed value} - \text{fitted value}) \times \text{weight}}{\text{fitted value} \times \text{coefficient of variation}}$$

Postcode	Observed	Full quadratic spline		Full cubic spline		Reduced cubic spline	
		Fitted value	Standardized residual	Fitted value	Standardized residual	Fitted value	Standardized residual
2000	0.275	1.0768	-0.635330	0.9417	-0.635527	0.9179	-0.629456
2006	6.962	1.0122	0.953465	1.0101	1.005597	0.9777	1.045798
2007	1.343	1.0519	0.179540	0.9661	0.266327	0.9372	0.295957
2008	4.890	1.0297	2.852375	0.9986	3.119427	0.9692	3.241843
2009	0.337	1.0510	-0.324423	0.9385	-0.322000	0.9126	-0.317242
2010	2.164	1.0931	1.638832	0.9842	2.109650	0.9595	2.211845
2011	1.529	1.1696	0.521963	1.0260	0.876103	1.0065	0.928970
2015	1.411	0.9783	0.459386	1.0629	0.357886	1.0458	0.382018
2016	1.112	1.0382	0.077088	0.9985	0.129678	0.9730	0.163161
2017	0.262	1.0100	-0.485528	1.0234	-0.513134	1.0052	-0.510541
2018	0.769	0.9661	-0.350791	1.0390	-0.470091	1.0555	-0.491562
2019	0.608	0.8735	-0.479695	0.9034	-0.542916	0.9803	-0.631311

APPENDIX G  
RISK PLOTS

**G1. Data**

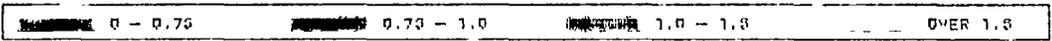
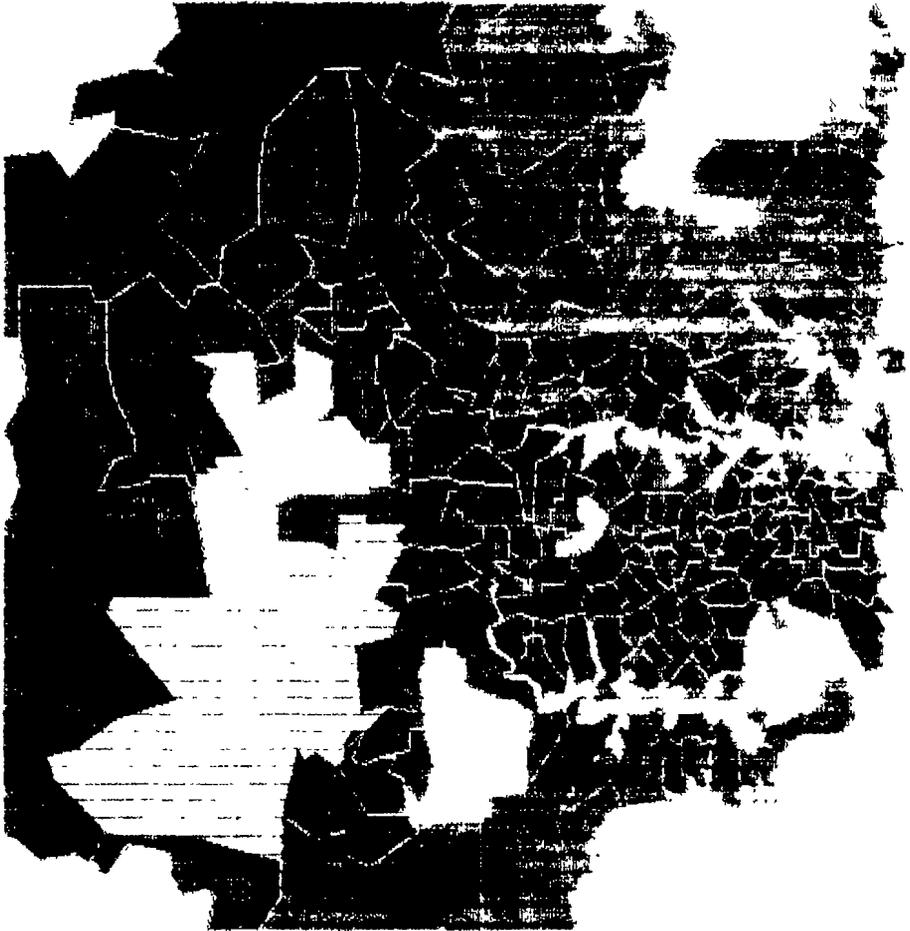
The following map plots the operating ratio adjusted for region, (3.1.4), different colours designating broad bands of values of this ratio.



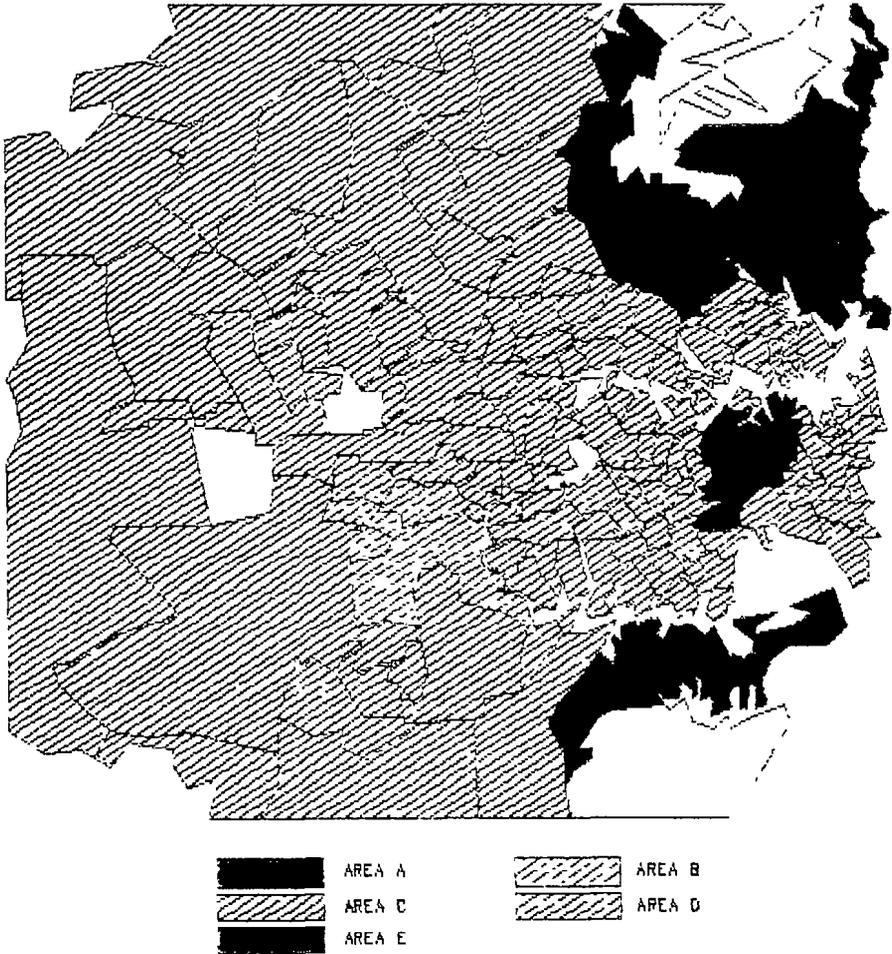
### G2. Reduced cubic spline function

The following map repeats the one appearing in Appendix G1 but with the observed operating ratio replaced by that fitted using the reduced cubic spline of Section 6.

## RATIO-TWO



### G3. Rating areas used in practice from 1/4/85 to 1/2/88



#### REFERENCES

- GREVILLE, T N E (1969) (ed) *Theory and applications of spline functions* Academic Press, Inc New York.
- PAYNE, C D (1985) (ed) *The Generalised Linear Interactive Modelling System Manual Release 3 77* Numerical Algorithms Group Ltd, Oxford, UK
- SEAL, H L (1969) *Stochastic theory of a risk business* John Wiley & Sons, Inc. New York
- SEAL, H L (1977) Approximations to risk theory's  $F(x, t)$  by means of the gamma distribution *ASTIN-Bulletin*, **9**, 213-218
- SEBER, G A F (1977) *Linear Regression analysis* John Wiley & Sons, Inc New York

G. C. TAYLOR

WILLIAM M MERCER CAMPBELL COOK & KNIGHT

*The American Express Tower, 388 George Street, Sydney, NSW 2000, Australia*