Survey Methodology, December 1996
Vol. 22, No. 2, pp. 175-183
Statistics Canada

Applications of Spatial Smoothing to Survey Data

ANN COWLING, RAY CHAMBERS, RAY LINDSAY and BHAMATHY PARAMESWARAN[1]

ABSTRACT

In this paper we present two applications of spatial smoothing using data collected in a large scale economic survey of Australian farms: one a small area and the other a large area application. In the small area application, we describe how the sample weights can be spatially smoothed in order to improve small area estimates. In the large area application, we give a method for spatially smoothing and then mapping the survey data. The standard method of weighting in the survey is a variant of linear regression weighting. For the small area application, this method is modified by introducing a constraint on the spatial variability of the weights. Results from a small scale empirical study indicate that this decreases the variance of the small area estimators as expected, but at the cost of an increase in their bias. In the large area application, we describe the nonparametric regression method used to spatially smooth the survey data as well as techniques for mapping this smoothed data using a Geographic Information System (GIS) package. We also present the results of a simulation study conducted to determine the most appropriate method and level of smoothing for use in the maps.

KEY WORDS: Kernel estimation; Mapping survey data; Small area estimation; Survey weighting.

1. INTRODUCTION

The Australian Bureau of Agricultural and Resource Economics (ABARE) is the applied economic research organisation attached to the Department of Primary Industries and Energy. Amongst its information gathering activities, ABARE conducts annual surveys of selected Australian agricultural industries which provide a broad range of information on the economic and physical characteristics of farm business units.

The largest survey is the Australian Agricultural and Grazing Industries Survey (AAGIS), which covers farm establishments with an estimated value of agricultural operations (EVAO) of $A22,500 or more in the last agricultural census that are classified to one of the broadacre industries – that is, cereal crop production, beef cattle production, and sheep and wool production. For the last two years, around 1650 farms have been included in the AAGIS sample, which is stratified by geographic area, industry, and EVAO. The sample farms are located throughout Australia with a non-uniform density. The latitude and longitude of the sample farms (defined in terms of the location of the farm "gate") is recorded as a regular part of the collection. This knowledge of the location of the surveyed farms enables the spatial smoothing techniques described in this paper to be used.

Traditionally, AAGIS estimates have been presented only as tables of numbers showing averages for all Australia, each state, and industries within states. However, the concern of rural industry and government about the combined impact of drought in some areas of Australia and the decline in certain commodity prices has highlighted the need for timely and detailed information on regional trends in farm performance. In particular, there has been a perceived need for information which portrays the spatial distribution of farm performance, reflecting actual variability in climate and production across Australia.

A highly effective way of presenting information on a spatial basis is to map the regional variation in economic performance of the surveyed farms. We use a nonparametric regression method to spatially smooth the farm level survey data, which is then presented in the form of a map. Recent improvement in computing power and the availability of high quality and affordable GIS packages have made this form of presentation a practical alternative to the traditional tabular method of presenting survey results.

Maps have been found to be a successful form of exposition for a number of reasons. First, estimates presented in a map are easily interpreted; when presented with too many tables it is very easy for a client to overlook local variations or be "swamped" by numbers. Next, maps make it easy for a client to relate the geographic variation in one variable with that of another. Finally, a colour map has great visual impact.

This demand for information on a spatial basis has resulted in an increased emphasis on small area estimates. One method of small area estimation (which originated naturally from smoothing survey data for presentation in maps) is to spatially smooth the sample weights. This reduces the variability of the small area estimates.

In Section 2, we examine a method of integrating geographical location into ABARE's survey weighting methods in order to make our small area estimates less variable. It is applied to sub-regional estimation within two Agricultural Regions in Section 3. In Section 4, we describe how kernel regression techniques can be used to produce

[1] Ann Cowling, CSIRO Division of Fisheries, GPO Box 1538, Hobart TAS 7001, Australia and Australian Bureau of Agricultural and Resource Economics; Ray Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton S017 1BJ, United Kingdom; Ray Lindsay and Bhamathy Parameswaran, Australian Bureau of Agricultural and Resource Economics, GPO Box 1563, Canberra ACT 2601, Australia.

maps which give a good indication of the local geographic variation of a surveyed variable. Two methods of mapping the smoothed data are discussed, both of which use ARC/INFO, a GIS software package. The results of a simulation study comparing various kernel regression methodologies for use in ABARE's maps are summarised in the Appendix.

2. SMALL AREA ESTIMATION BY SPATIALLY SMOOTHING SAMPLE WEIGHTS

The standard method used to compute sample weights at ABARE is described in Bardsley and Chambers (1984). It rests on the assumption that at some appropriate level of aggregation (say, Agricultural Region) the variable Y follows a linear model of the form

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{V} \tag{2.1}$$

where $\boldsymbol{Y}$ is the N-vector of values of Y at this level of aggregation, $\boldsymbol{X}$ is a $N \times p$ matrix of values of a set of p benchmark variables, β is an unknown p-vector of regression coefficients and $\boldsymbol{V}$ is a N-vector of errors satisfying $E(\boldsymbol{V}) = \boldsymbol{0}$ and $\text{var}(\boldsymbol{V}) = \sigma^2\Omega$, where σ is an unknown scale parameter and Ω is a known $N \times N$ diagonal matrix having as its elements the measure of size of each farm, EVAO, introduced in the previous section.

Since this model is a multipurpose model, with the same set of benchmark variables used for each survey variable, the column dimension, p, of $\boldsymbol{X}$ is usually large. Typically, $\boldsymbol{X}$ consists of between 3 and 7 variables related to the main agricultural commodities produced by farms in the region together with dummy variables indicating industry strata within the region. Best linear unbiased estimation of the population total of a survey variable on the basis of such an overspecified model typically results in weights that are highly variable and often negative.

As discussed in Bardsley and Chambers (1984), negative weights are highly undesirable in a multi-purpose survey like AAGIS. In particular, such weights can lead to negative estimates of intrinsically positive quantities. This problem has been pointed out in the literature a number of times (see for example, Deville and Särndal 1992; Bankier, Rathwell and Majkowski 1992; and Fuller, Loughin and Baker 1994). The method used at ABARE to control for strictly positive sample weights is based on the ridge-type modification to the best linear unbiased weights suggested by Bardsley and Chambers (1984).

Given a sample of size n from a particular region, the ridge weighting approach determines the sample weight vector $\boldsymbol{w}$ by minimising the mean squared error criterion

$$Q = \lambda^{-1}\boldsymbol{B}^T\boldsymbol{C}\boldsymbol{B} + (\boldsymbol{w} - \boldsymbol{1})^T\omega(\boldsymbol{w} - \boldsymbol{1}). \tag{2.2}$$

Here $\boldsymbol{B} = \boldsymbol{T} - \boldsymbol{x}^T\boldsymbol{w}$ is a p-vector of benchmark biases, corresponding to the differences between the (known) population totals $\boldsymbol{T}$ of the p benchmark variables making up $\boldsymbol{X}$ and the corresponding survey estimates $\boldsymbol{x}^T\boldsymbol{w}$ of these totals, $\boldsymbol{C}$ is a $p \times p$ diagonal matrix of non-negative relative "costs" associated with these biases, ω is the sample component of Ω, $\boldsymbol{x}$ is the sample component of $\boldsymbol{X}$, $\boldsymbol{1}$ is a n-vector of ones and λ is a scaling constant which is chosen by the survey analyst. The value of $\boldsymbol{w}$ minimising Q is

$$\boldsymbol{w} = \boldsymbol{1} + \omega^{-1}\boldsymbol{x}(\lambda\boldsymbol{C}^{-1} + \boldsymbol{x}^T\omega^{-1}\boldsymbol{x})^{-1}(\boldsymbol{T} - \boldsymbol{x}^T\boldsymbol{1}). \tag{2.3}$$

The scale constant λ is called the ridge parameter associated with these weights. As λ increases from zero, the sample weights in $\boldsymbol{w}$ move away from their best linear unbiased values under the model (2.1) (namely, their values at $\lambda = 0$) and become less and less variable. That is, as λ increases, the variances of the survey estimates based on these weights decrease. On the other hand, as λ increases, these estimates become more biased under (2.1), so the components of $\boldsymbol{B}$ move away from their zero values at $\lambda = 0$ (where the sample weights define unbiased estimates under (2.1)). These components become larger and larger (in absolute terms) as λ increases.

The survey analyst makes a tradeoff between these two competing sources of "error" by choosing the smallest value of λ such that the sample weights in $\boldsymbol{w}$ stabilise at strictly positive values as close as possible to their best linear unbiased values under (2.1). This ensures that the components of $\boldsymbol{B}$ are as small as possible subject to this stability requirement. At ABARE, the value of λ is chosen so that the sample weights are at least unity.

Recent small area estimation research in ABARE has focussed on a method of modifying this ridge weighting procedure to create sample weights that are less spatially variable. We achieve this by modifying the mean squared error criterion Q in (2.2) to include a constraint on spatial variability, while continuing to regard the elements of the variable $\boldsymbol{Y}$ as being independent.

Let $\boldsymbol{K}$ be an $n \times n$ matrix reflecting Euclidean distance between sample farms, such that $\boldsymbol{K}$ is symmetric and non-negative, $\boldsymbol{K}_{ii} = 1$ for all i and $\boldsymbol{K}_{ij} \downarrow 0$ as the distance between farm i and farm j increases. Put $\boldsymbol{u} = \boldsymbol{w} - \boldsymbol{1}$. The aim is then to choose $\boldsymbol{u}$ so that when $\boldsymbol{K}_{ij}$ is large, the difference between $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ is small. That is, we seek to minimise a quantity of the form

$$\sum_{i\epsilon s}\sum_{j\epsilon s}\boldsymbol{K}_{ij}(\boldsymbol{u}_i - \boldsymbol{u}_j)^2 = 2(\boldsymbol{u}^{(2)})^T\boldsymbol{K}\boldsymbol{1} - 2\boldsymbol{u}^T\boldsymbol{K}\boldsymbol{u} \tag{2.4}$$

where $(\boldsymbol{u}^{(2)})_i = (\boldsymbol{u}_i)^2$. An appropriate modification to the mean squared error criterion (2.2) leads to minimisation of

$$Q^* = \lambda^{-1}\boldsymbol{B}^T\boldsymbol{C}\boldsymbol{B} + \boldsymbol{u}^T\omega\boldsymbol{u} + (\boldsymbol{u}^{(2)})^T\boldsymbol{K}\boldsymbol{1} - \boldsymbol{u}^T\boldsymbol{K}\boldsymbol{u}.$$

Minimising with respect to $\boldsymbol{u}$ leads to

$$\boldsymbol{u} = \eta^{-1}\boldsymbol{x}(\lambda\boldsymbol{C}^{-1} + \boldsymbol{x}^T\eta^{-1}\boldsymbol{x})^{-1}(\boldsymbol{T} - \boldsymbol{x}^T\boldsymbol{1})$$

provided η^{-1} exists, where

$$\eta = \text{diag}(\boldsymbol{K}\mathbf{1}) - \boldsymbol{K} + \omega. \tag{2.5}$$

Clearly, then,

$$\boldsymbol{w} = \mathbf{1} + \eta^{-1}\boldsymbol{x}(\lambda \boldsymbol{C}^{-1} + \boldsymbol{x}^T\eta^{-1}\boldsymbol{x})^{-1}(\boldsymbol{T} - \boldsymbol{x}^T\mathbf{1}). \tag{2.6}$$

It can be seen that the modified mean squared error criterion Q^* equally weights the spatial smoothness criterion given in (2.4), and the term corresponding to the variance of the prediction error of the sample estimates, $\boldsymbol{u}^T\omega\boldsymbol{u}$. As the scale of $\boldsymbol{K}$ was arbitrarily specified, the comparative weighting of the two criteria must be modified by "scaling up" the spatial matrix $\{\text{diag}(\boldsymbol{K}\mathbf{1}) - \boldsymbol{K}\}$ by a factor ϕ in order to make it comparable in size with the heteroscedasticity matrix ω, and by adding a parameter α, $0 \le \alpha \le 1$, to the expression for η in equation (2.5), so that

$$\eta = (1 - \alpha)\phi\{\text{diag}(\boldsymbol{K}\mathbf{1}) - \boldsymbol{K}\} + \alpha\omega.$$

These spatially smoothed sample weights can be derived in a second way, providing deeper insight into how they should be interpreted. This follows from noting that

$$\eta = \begin{bmatrix} \sigma_1^2 + \sum_{m\neq 1} \boldsymbol{K}_{1m} & -\boldsymbol{K}_{12} & \cdots & -\boldsymbol{K}_{1n} \\ -\boldsymbol{K}_{21} & \sigma_2^2 + \sum_{m\neq 2} \boldsymbol{K}_{2m} & \cdots & -\boldsymbol{K}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{K}_{n1} & -\boldsymbol{K}_{n2} & \cdots & \sigma_n^2 + \sum_{m\neq n} \boldsymbol{K}_{nm} \end{bmatrix}$$

can be expressed as $\eta = \boldsymbol{S}\boldsymbol{R}\boldsymbol{S}$, where $\boldsymbol{S}$ is a diagonal matrix with $\boldsymbol{S}_{ii} = (\sigma_i^2 + \sum_{m\neq i}\boldsymbol{K}_{im})^{1/2}$, and $\boldsymbol{R}$ is a correlation matrix with

$$R_{ij} = \begin{cases} 1 & \text{if } i = j \\ -\boldsymbol{K}_{ij}\left\{\left(\sigma_i^2 + \sum_{m\neq i}\boldsymbol{K}_{im}\right)\left(\sigma_j^2 + \sum_{m\neq j}\boldsymbol{K}_{jm}\right)\right\}^{-1/2} & \text{if } i \neq j. \end{cases}$$

Thus the spatially smoothed sample weights can alternatively be derived as ridge-type regression weights based on the assumption that the variable $\boldsymbol{Y}$ follows a linear model of the form (2.1), with $\boldsymbol{V}$ redefined as satisfying $E(\boldsymbol{V}) = \mathbf{0}$, $\text{var}(\boldsymbol{Y}_i) = \sigma_i^2 + \sum_{m\neq i}\boldsymbol{K}_{im}$, and $\text{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = -\boldsymbol{K}_{ij}$ for $i \neq j$. The usual ridge weighting procedure then leads directly to (2.6) with η defined by (2.5). Note that under this implied model neighbouring farms are negatively correlated.

This second method of derivation shows clearly that the introduction of spatial smoothness for the survey weights is at odds with standard concepts of statistical efficiency as far as estimation at the aggregate level is concerned. Since the spatial correlation between neighbouring farms will typically be positive, efficient survey estimation at the aggregate level will involve weighting based on (2.3) with ω replaced by a non-diagonal variance/covariance matrix reflecting this positive spatial correlation. These are not the weights that result when one imposes as spatial similarity constraint. Consequently, one could expect that such "large area efficient" weights would tend to be more dissimilar for neighbouring farms than they would be for farms that are far apart. That is, there is a price to pay in weighting – if less variable aggregate level estimates are required, then this tends to lead to more variable small area estimates. Conversely, if (2.6) is adopted as the method of weighting because of its desirable small area properties, then it can be expected that aggregate level estimates obtained by summing these small area estimates will be less efficient.

The spatially smooth sample weights (2.6) have been implemented using

$$\boldsymbol{K}_{ij} = \exp(-d\|z_i - z_j\|), \tag{2.7}$$

where $\|z_i - z_j\|$ is the distance between farm i and farm j and d is a constant controlling the radius of circle around the i-th farm within which spatial smoothing is applied. The smaller the value of d, the larger the radius of spatial smoothing. At present, the "scaling up" constant ϕ is computed as the ratio of the determinants of the $\boldsymbol{K}$ and ω matrices, raised to the power n^{-2}. An empirical evaluation of this method is described in the following Section.

3. AN APPLICATION OF SPATIALLY SMOOTHED SAMPLE WEIGHTING

Initial results from an evaluation of the first method of spatially smoothed ridge weighting described in the previous section are set out in Tables 1 to 3. These results are for two Agricultural Regions. The first, Region A, is in New South Wales. In spatial terms, this region is relatively homogeneous, being located in the southwestern corner of the state. The principal agricultural activities are wheat and rice production and wool and lamb production. The second, Region B, is in Western Australia. This region is more spatially heterogeneous, ranging from established cropping and wool production farms in the central west of the state to much larger livestock and cropping farms on marginal farming land in the south east of the state. The principal agricultural activities are wheat and legumes production and wool production.

Six variations of the spatially smoothed ridge weights (2.6) with $\boldsymbol{K}$ given by (2.7) were used in the evaluation, defined by values of $d = 0.05$ (weak spatial effects) and $d = 0.005$ (strong spatial effects), and values of $\alpha = 0.9$ (most emphasis on the standard ridge weights), $\alpha = 0.5$ (equal emphasis on standard ridge weights and spatially smooth weights) and $\alpha = 0.1$ (most emphasis on spatially smooth weights).

Table 1
Values (in relative percentage terms) of the biases associated with estimation of the benchmark variables corresponding to the principal agricultural commodities produced in Region A (sample size $n = 101$ farms) and Region B (sample size $n = 85$ farms) using the standard ridge weights (2.3) and the spatially smooth ridge weights (2.6)

		Wheat	Sheep	Rice
Region A				
Standard ridge weights		-0.50	5.0	13.0
Spatially smoothed ridge weights				
$d = 0.05$	$\alpha = 0.9$	-0.50	4.6	11.9
	$\alpha = 0.5$	-0.46	4.7	12.4
	$\alpha = 0.1$	0.07	6.2	17.4
$d = 0.005$	$\alpha = 0.9$	-0.40	4.9	12.7
	$\alpha = 0.5$	0.80	8.9	28.0
	$\alpha = 0.1$	9.20	25.0	60.0
		Wheat	Sheep	Legumes
Region B				
Standard ridge weights		0.43	-1.25	1.49
Spatially smoothed ridge weights				
$d = 0.05$	$\alpha = 0.9$	0.42	-1.16	1.37
	$\alpha = 0.5$	0.44	-1.14	1.40
	$\alpha = 0.1$	0.69	-1.25	2.53
$d = 0.005$	$\alpha = 0.9$	0.50	-1.20	1.68
	$\alpha = 0.5$	1.51	1.14	9.73
	$\alpha = 0.1$	26.57	19.61	45.46

Table 1 shows the relative biases associated with estimation of the population totals of the main commodity related benchmarks for each region under these different weighting systems, as well as the corresponding biases associated with the standard ridge weights. The increase in these biases as the amount of spatial smoothing in the weights is increased is evident. Since these production benchmarks are positively correlated with most of the economic variables measured in the survey, these benchmark biases can be expected to be translated into a corresponding upward bias in survey estimates based on these weights.

Figures 1 to 4 show the difference between the smoothed weights and the standard ridge weights for the two "extreme" combinations of α and d in both regions changes as the size (measured in terms of the logarithm of the estimated value of agricultural operations, or log(EVAO)) of the sample farms changes.

Observe that for relatively strong spatial smoothing (Figures 1 and 3), the effect of smoothing is to increase the weights of most of the larger sample farms, while dramatically decreasing the weights of a small number of smaller sample farms. Weak spatial smoothing (Figures 2 and 4) changes the weights much less, and there is little relationship between the size of the farm and the direction of weight change. Consequently, an upward shift in survey estimates for these regions could be expected with the introduction of

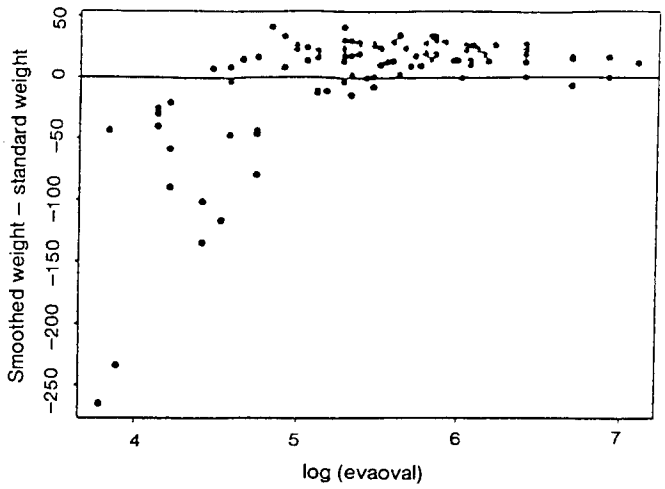

Figure 1. Difference between smoothed weight with $\alpha = 0.1$ and $d = 0.005$ and standard ridge weight, Region A

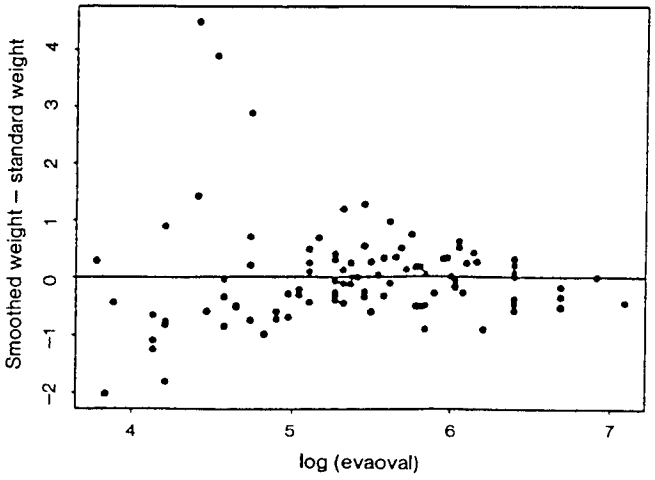

Figure 2. Difference between smoothed weight with $\alpha = 0.9$ and $d = 0.05$ and standard ridge weight, Region A

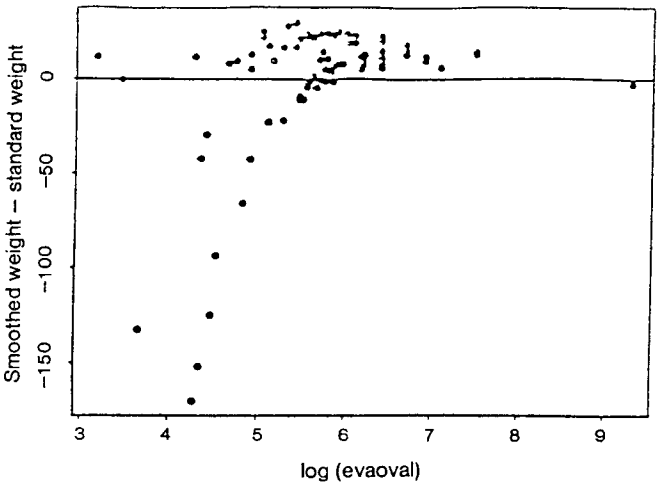

Figure 3. Difference between smoothed weight with $\alpha = 0.1$ and $d = 0.005$ and standard ridge weight, Region B

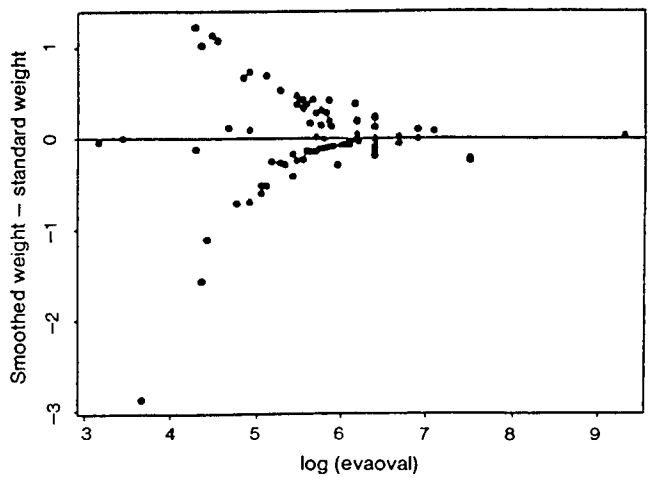

Figure 4. Difference between smoothed weight with $\alpha = 0.9$ and $d = 0.05$ and standard ridge weight, Region B

strongly spatially smoothed sample weights. Given the increased positive biases indicated in Table 1, this upward shift would be expected to be essentially due to the introduction of a positive bias in these estimates.

Is this increased bias compensated for by a lower standard error? To evaluate this question, survey estimates and estimated standard errors were computed for a key financial variable, total cash costs. These estimates are set out in Table 2 (Region A) and Table 3 (Region B). Estimates are provided both for each region and for small areas within each region, denoted SR-i in the table, with the index i ranging between 1 and 6 for Region A and between 1 and 7 for Region B.

Table 2
Estimates (with corresponding estimated standard errors in parentheses) of the average value of Y = total cash costs in subregions SR–1 to SR–6, making up Region A (sample size n = 101 farms), using the standard ridge weights (2.3) and the spatially smooth ridge weights (2.6)

	Standard weights	Spatially smoothed ridge weights					
		$d = 0.05$			$d = 0.005$		
		$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$
SR–1	100,618	100,453	101,297	107,263	102,059	112,635	135,419
	(24,551)	(24,511)	(23,906)	(20,487)	(23,474)	(18,923)	(18,011)
SR–2	115,320	115,417	116,002	120,362	116,917	126,165	153,707
	(26,754)	(26,661)	(26,448)	(25,637)	(26,423)	(25,990)	(27,975)
SR–3	167,524	167,453	167,486	168,257	167,709	170,781	187,683
	(28,479)	(28,467)	(28,473)	(28,426)	(28,175)	(26,471)	(24,211)
SR–4	182,940	180,317	177,838	163,556	176,257	174,077	192,296
	(106,471)	(105,485)	(101,012)	(74,418)	(97,823)	(69,109)	(43,651)
SR–5	132,050	132,083	132,389	134,786	132,490	136,369	151,046
	(25,089)	(25,096)	(25,154)	(25,475)	(25,173)	(24,410)	(23,110)
SR–6	132,493	132,184	133,204	141,623	133,763	147,652	192,781
	(44,385)	(44,546)	(44,757)	(46,736)	(45,078)	(46,953)	(53,105)
Region A	134,114	133,807	134,141	137,080	134,506	142,040	166,432
	(15,691)	(15,655)	(15,426)	(13,845)	(15,199)	(13,494)	(12,815)

Table 3
Estimates (with corresponding estimated standard errors in parentheses) of the average value of Y = total cash costs in subregions SR–1 to SR–7, making up Region B (sample size n = 85 farms), using the standard ridge weights (2.3) and the spatially smooth ridge weights (2.6)

	Standard weights	Spatially smoothed weights					
		$d = 0.05$			$d = 0.005$		
		$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.9$	$\alpha = 0.5$	$\alpha = 0.1$
SR–1	183,194	183,262	183,528	186,151	184,287	195,138	257,652
	(64,851)	(64,325)	(64,051)	(64,967)	(64,132)	(69,859)	(59,518)
SR–2	261,952	261,487	261,119	261,182	261,938	276,912	331,805
	(70,989)	(70,601)	(70,502)	(73,131)	(70,723)	(79,751)	(67,356)
SR–3	113,499	113,441	113,742	116,847	114,631	125,525	157,007
	(30,304)	(30,289)	(30,255)	(30,731)	(30,377)	(31,507)	(32,500)
SR–4	242,220	242,182	242,208	242,221	242,163	242,439	250,871
	(26,160)	(25,671)	(26,159)	(26,160)	(26,154)	(24,244)	(24,836)
SR–5	134,524	134,970	135,700	139,122	134,734	131,448	148,629
	(32,420)	(32,528)	(32,432)	(30,607)	(32,202)	(27,867)	(27,942)
SR–6	176,540	176,977	175,708	163,241	172,076	148,434	171,856
	(60,377)	(60,703)	(59,214)	(46,361)	(55,925)	(36,218)	(39,527)
SR–7	205,287	205,644	205,433	202,039	204,519	194,998	219,959
	(44,137)	(44,008)	(43,963)	(44,044)	(43,972)	(45,434)	(51,690)
Region B	176,283	176,342	176,397	176,822	176,294	179,998	216,445
	(19,039)	(18,869)	(18,874)	(18,213)	(18,511)	(18,540)	(17,099)

It is seen that, in general, the answer to the question posed above is yes. The estimated standard errors of the survey estimates decrease as the degree of spatial smoothness of the weights increases (from left to right across the tables). However, as expected, the estimates themselves also increase in size, becoming more and more positively biased. Overall, the gain due to reduced standard error seems to cancel out the increase in bias, except for the heaviest spatial smoothing ($\alpha = 0.1$, $d = 0.005$). In this latter case the increase in bias outweighs the reduction in standard error. The choice $\alpha = 0.1$ and $d = 0.05$ seems a good compromise, leading to reasonable (but not spectacular) bias-variance tradeoffs in Region A, and little change in the estimates in Region B.

4. ESTIMATION AND MAPPING OF LOCAL AVERAGES

A survey data map is a two-dimensional surface which estimates the spatial mean function of the survey variable in the population. In practice, such a map is obtained by applying a nonparametric regression technique to the weighted unit record data obtained in the survey.

At ABARE, we use kernel regression (a nonparametric technique) to produce maps which show the spatial variation of the estimated spatial mean function surfaces of key survey variables. These surfaces are obtained by replacing the observed sample values of these variables by locally weighted averages. In addition, for each local average map, a

corresponding map is produced which shows an estimate of the local variability of the variable of interest. We give below a brief outline of the technique: for clarity of exposition we deal only with the univariate case. See Ruppert and Wand (1994), Wand and Jones (1995, p140), and the references therein, for discussion of the multivariate case.

We assume that the finite population is generated as an iid sample $\{(Z_i, Y_i), i = 1, ..., N\}$ from a super population where Y_i is the value of a response variable Y observed at location Z_i. We suppose that the observations follow the model

$$Y_i = m(Z_i) + \epsilon_i, \qquad i = 1, ..., N$$

where $m(z) = E(Y \mid Z = z)$ is the conditional mean of Y given Z, and the ϵ_i are independent random variables with zero mean and variance $\sigma^2(z)$. Suppose that the error terms ϵ_i are independent of the process by which the sample is selected, so that the sample values $\{(Z_i, Y_i), i = 1, ..., n\}$ follow the same model, and write f for the density of $Z_1, ..., Z_n$.

A natural choice for the local average at any point z is then the mean of the values of the response variable for those observations with locations close to z, since observations from points far away will tend to have very different mean values. The local average is defined as a weighted mean

$$\hat{m}(z) = n^{-1} \sum_{i=1}^{n} W_i(z) Y_i$$

where the weights $\{W_i(z)\}$ depend on the locations $\{Z_i\}$ of the sample observations, and $\hat{m}(z)$ estimates $m(z)$.

The weights are constructed using a function K known as the kernel, which is continuous, bounded, symmetric and integrates to one. Various weight sequences have been proposed: the traditional Nadaraya-Watson weights (Nadaraya 1964 and Watson 1964) are

$$W_i(z) = h^{-1} K\{(z - Z_i)/h\} \Big/ \left[(nh)^{-1} \sum_{j=1}^{n} K\{(z - Z_j)/h\} \right],$$

where h is a scale factor known as the bandwidth. The kernel function K gives an observation close to z relatively more influence on the local average at this location than it gives to an observation further from z.

Where observations are sparse, a fixed-bandwidth window may contain few points and the corresponding estimator may therefore have a very high variance. This may be avoided by using the k-nearest-neighbour method in which a different bandwidth is used at each estimation point z. The bandwidth at z is the distance to the k-th nearest neighbour of z, so that there are always exactly k points in the bandwidth window. Let h_k be the distance between z and its k-th nearest neighbour. The k-nearest-neighbour Nadaraya-Watson weights are

$$W_{ih_k}(z) = h_k^{-1} K\{(z - Z_i)/h_k\} \Big/ \left[(nh_k)^{-1} \sum_{j=1}^{n} K\{(z - Z_j)/h_k\} \right].$$

We show in Table 4 the asymptotic mean squared error (MSE) properties of the usual (fixed-bandwidth) and k-nearest-neighbour estimators as given in Härdle (1990, p. 46).

Table 4
Asymptotic bias and variance of Nadaraya-Watson estimators; $c_K = \int K^2(u)du$, $d_K = \int u^2 K(u)du$

	Fixed-bandwidth	k-nearest-neighbour
Bias	$h^2 \frac{(m''f + 2m'f')(x)}{2f(x)} d_K$	$\left(\frac{k}{n}\right)^2 \frac{(m''f + 2m'f')(x)}{8f^3(x)} d_K$
Variance	$\frac{\sigma^2(x)}{nhf(x)} c_K$	$\frac{2\sigma^2(x)}{k} c_K$

Clearly, the bias of the estimated regression function can be reduced by using a smaller bandwidth h (number of nearest-neighbours k), but this leads to a noisy estimate $\hat{m}$ with local detail masking global features of the curve ($\hat{m}$ has high variance). If $h(k)$ is large, $\hat{m}$ is smoother but the global features are dampened ($\hat{m}$ has high bias and low variance). The bias, then, can only be reduced at the expense of variance and vice versa, with the bandwidth h determining the ratio of (squared) bias to variance.

In reality, the survey design and the spatial distribution of a survey variable Y will not be independent, so simple local averages for Y derived from the sample data will be misleading as estimates of the local population means of this variable. To overcome this problem the kernel weights are multiplied by the survey weights to get the final smoothing weights used for calculating the local average. This is equivalent to estimating the local population mean $m(z)$ of Y under the assumption that it is locally linear in the same benchmark variables as those used to model the overall population mean of Y.

A wide array of alternative kernel smoothing procedures have been discussed in the literature. As well as various sequences of smoothing weights $\{W_i\}$, there are different types of bandwidths, and several automatic bandwidth selection methods. A simulation study was therefore conducted to determine the most appropriate kernel methodology for use in ABARE's maps. This is described in the Appendix.

Uncertainty about the estimate of the spatial mean derived via kernel-based spatial smoothing can be represented by mapping the local variability of the variable of interest. Areas of high local variability correspond to areas where the map of the mean function is less precise and vice versa for areas of low local variability.

The usual method of determining confidence regions for a kernel curve estimate is the bootstrap; see Härdle (1990), Hall (1992), and references therein. However, for computational efficiency, we use the expectiles (Newey and Powell 1987) of the spatial distribution of Y to describe this

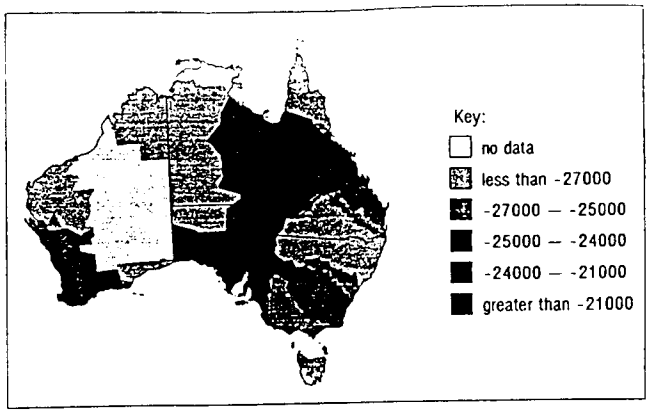

Figure 5. Polygon map of farm business profit in 1991-1992, all broadacre farm ($)

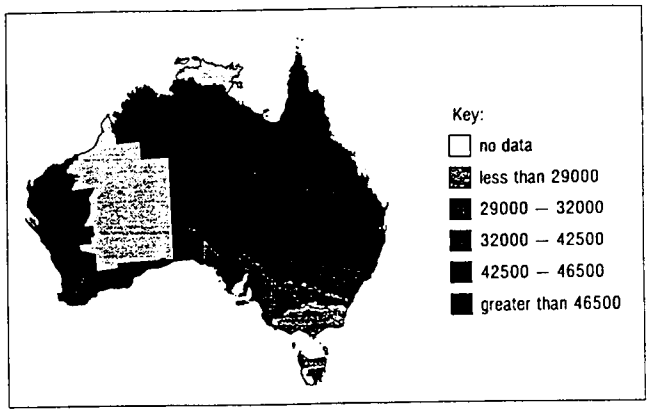

Figure 6. Polygon map of interexpectile range of farm business profit in 1991-1992, all broadacre farms ($)

local variability. An expectile bears the same relationship to the mean as the corresponding quantile does to the median. In particular, the difference between the 75th and 25th expectiles of a distribution is a measure of the spread of the distribution in the same way as the interquartile range is a measure of this spread. The smoothing program contains a module for non-parametric M-quantile regression (Breckling and Chambers 1988) which is used to fit a smooth surface to the expectiles of the Y-distribution at any location. The difference between the smoothed 75th and 25th expectile surfaces (the smooth expectile analogue of the interquartile range) is then mapped to show areas of high and low variability in the data.

Not surprisingly, this smooth interexpectile range tends to be highest in areas where the farms are sparsely located and the farm-to-farm variability in Y is therefore highest. The interexpectile range map corresponding to Figure 5 is shown in Figure 6. Note that these smoothed interexpectile range maps provide similar information to confidence bands at any particular point on the map. However, they do not have the same repeated sampling interpretation as confidence intervals, and hence should be treated as guides to, rather than measures of, the uncertainty associated with a particular map.

For confidentiality reasons, care must be taken when mapping the smoothed data for publication to ensure that the locations of the surveyed farms are not revealed. Another requirement is output quality compatible with desktop publication packages. Two procedures for generating the final maps that satisfy these requirements have been developed using ARC/INFO.

In the first method, a Thiessen polygon is constructed around each farm. The polygon defines the area closer to that farm than to any other farm. The farm location is not in the centre of its polygon, and the polygon shape does not resemble the shape of the farm, so the polygons conceal the locations of the survey farms, as shown in Figure 7. The whole of each polygon is coloured according to the smoothed value of Y at the farm location in that polygon. Usually ten colours are used in each map and the estimated population deciles of the smoothed data are used as boundaries for the colour area. The maps shown in this paper are black-and-white analogues of these colour maps.

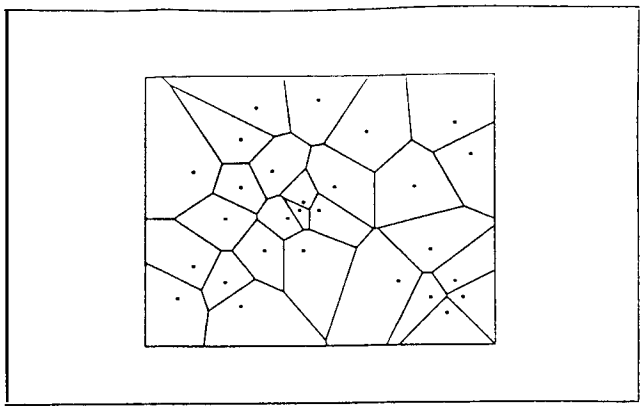

Figure 7. Thiessen polygons constructed around selected ABARE survey farms. Farm location is shown as a small square within each polygon

In the second method, smoothed values on a dense rectangular grid are used in place of smoothed values at the farm locations, and a further minor interpolation of the data is carried out in ARC/INFO. A continuous 3-dimensional surface which passes through the smoothed values at the grid points is built in two steps. As a first approximation, a faceted surface of triangles obtained by Delauney triangulation is constructed, and then a bivariate fifth degree polynomial is fitted within each triangle using Akima's algorithm (Akima 1978). The resulting continuous surface is then contoured using the estimated population deciles. Figure 8 is an example.

In this second method of presentation, the locations of the survey farms are not used in any way, thereby completely concealing the location of each survey farm. It also gives smooth contours, and the result is not as patchy as the polygon based map. Moreover, it is preferred by ABARE's graphics staff because it reduces the number of areas to be

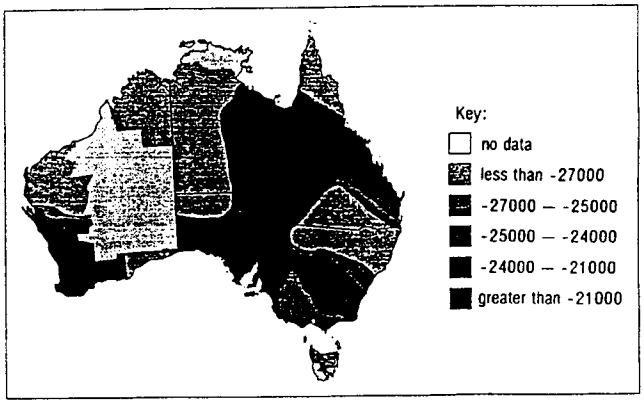

Figure 8. Contour map of farm business profit in 1991-1992, all broadacre farms ($)

separately coloured and has lower storage requirements, enabling the maps to be more readily manipulated in desktop publishing packages. Its disadvantage is that it uses more computing time in the ARC/INFO stage.

Since the above procedures interpolate across all of Australia, including areas where there is no agricultural activity, the final stage of the map production in ARC/INFO is the "blanking out" of those areas of Australia where there are few or no farms involved in the particular broadacre industry represented by the map. As Figure 9 shows, different areas are blanked out for different industries.

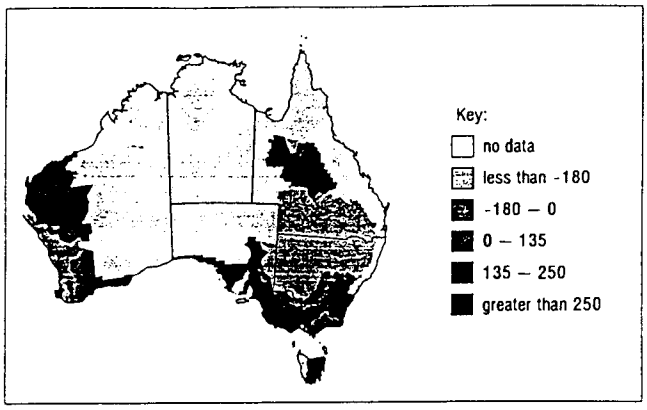

Figure 9. Polygon map showing expected change in wool production, 1991-92 to 1992-93, farms with 100 or more sheep in 1991-92 (kg)

5. DISCUSSION

In this paper we have demonstrated that when survey data has a spatial dimension, as in the case of the AAGIS, spatial smoothness concepts may be useful to the analyst. The concept can be used to modify survey weights to ensure less variable small area survey estimates. It may also be used to smooth the data along spatial dimensions before mapping the spatial mean function.

Because we describe mapping in this paper, we have only considered smoothing along spatial dimensions. However, it is clearly possible to use the same techniques to smooth along other dimensions. Thus, if there is reason to expect the presence of strong serial correlation when the underlying population is ordered according to some variable, then one can consider applying the methods described in this paper to mapping the "change" in the survey variables relative to the change in this variable. In doing so, it should be noted that such "maps" are nothing more than nonparametric estimates of the conditional means of the survey variables given this "ordering" or "smoothing" variable. The analyst should, however, remember the "curse of dimensionality": the effective sample size drops sharply with each additional smoothing variable used in these nonparametric techniques.

Finally, in mapping the survey data, we have used kernel-based estimation techniques. However, spline smoothing, or even parametric methods could also be used. We regard the choice of smoothing technology as somewhat subjective and purpose specific, as there are no definitive objective reasons for preferring one method over another.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their helpful comments which have greatly improved the presentation of the paper.

APPENDIX

In the last few years a number of optimality properties have been established for the locally-linear kernel weights (see for example Wand and Jones (1995) and references therein). We therefore compared Nadaraya-Watson (NW) and locally-linear (LL) weight sequences using fixed (FBW) and k-nearest-neighbour (NN) bandwidths with each weight sequence. For each of these combinations, we selected the bandwidth using least-squares cross-validation (CV), and an *ad hoc* method (detailed in the last paragraph of this section) aimed at reducing the speckledness of a map (SF).

Two criteria were used to evaluate the performance of each methodology. The first, MSE, is the obvious statistical criterion for assessing a biased estimator. The second criterion is more ABARE specific. As estimates are produced both in tables (by State) and in maps, the impression of the state average given by the map should be close to the tabulated value. We therefore used a weighted sum of the squared differences between the state averages of the raw and smoothed survey data (SB^2). This measure was also calculated at regional rather than state level (RB^2; there are between one and nine regions in each state).

Data were generated at the survey farm locations using three smooth functions with varying degrees of smoothness (measured by $\int m''$) and normal mixture errors. For example,

$$m_1(z) = 6.25 \times 10^4 \times \cos\left(\frac{z_1 - 132.5}{2.25}\right)\cos\left(\frac{z_2 + 27.5}{1.75}\right),$$

where z_1 and z_2 are the longitude and latitude of the point z. The functions $m_i(z)$ were scaled to have the same range as the smoothed values of a key survey variable, and the errors were scaled to have the same range as the residuals of the same variable after smoothing. Large variances were generated at locations with high residuals, and small variances at locations with low residuals. The simulation results based on the smooth function are given in Table 5.

Using MSE as the criterion for assessing methodology, the results were not consistent for the three functions $m_i(z)$. However, when either RB^2 or SB^2 was used as the performance measure, the LL estimator with k-nearest-neighbour bandwidth selected using SF outperformed the other methods by at least ten percent for each function $m_i(z)$, and is therefore the currently preferred methodology for producing ABARE's maps.

Table 5

Comparison of locally-linear (LL) and Nadaraya-Watson (NW) weight sequences, using fixed (FBW) and k-nearest-neighbour (NN) bandwidths selected by least-squares cross-validation (CV) and the criterion detailed below (SF). The results were obtained from 400 independent samples with mean function and normal mixture errors. The MSE values were calculated using the average over the finite population of $(y - \hat{m}(z))^2$

		$MSE \times 10^{-7}$		$RB^2 \times 10^{-7}$		$SB^2 \times 10^{-7}$	
		CV	SF	CV	SF	CV	SF
LL	FBW	39.64	93.93	4.44	1.67	1.33	0.39
	NN	20.50	22.83	2.22	1.35	0.37	0.14
NW	FBW	41.91	52.78	3.29	1.77	0.34	0.17
	NN	21.77	22.22	3.03	2.33	0.62	0.41

The bandwidth selection method aimed at reducing the speckledness of a map (SF) is a measure of the smoothness of the map: it measures how similar the smoothed value is at any farm to that of its neighbours. Let $p(i)$ be the survey estimate of the percentile of the smoothed variable at the i-th farm. Let S_i be the set of indices of the six farms closest to the i-th farm. In this method, the value of

$$SF(h) = (6n)^{-1}\sum_{i}\sum_{k \in S_i} |p(i) - p(k)|$$

is calculated. It is scale-free, and decreases monotonically as the bandwidth decreases. The chosen bandwidth is the smallest bandwidth with a sufficiently small ($< \epsilon$) rate of decrease of SF. The value of ϵ was chosen subjectively following detailed examination of maps of five key variables for five values of ϵ.

REFERENCES

BANKIER, M.D., RATHWELL, S., and MAJKOWSKI, M. (1992). Two step generalised least squares estimation in the 1991 Canadian Census. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*. Statistics Sweden, Örebro, October 5-7.

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

BRECKLING, J., and CHAMBERS, R.L. (1988). *M*-quantiles. *Biometrika*, 75, 761-771.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

FULLER, W.A., LOUGHIN, M.M., and BAKER, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

NADARAYA, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 10, 186-190.

NEWEY, W.K., and POWELL, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55, 819-847.

RUPPERT, D., and WAND, M.P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22, to appear.

WAND, M.P., and JONES, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

WATSON, G.S. (1964). Smooth regression analysis. *Sankhyā*, Series A, 26, 101-116.