Journal of Official Statistics
Vol. 4. No. 3. 1988, pp. 223–240

Regional Presentation–Choice of Region Sizes

Daniel Thorburn[1]

Abstract: In many surveys data are presented for different regions. If these regions are too large the geographical variation is not reflected. If they are too small, the geographical variation cannot be separated from the random variation. We discuss, from a mathematical point of view, the choice of the number and size of regions when the sample is given. We also consider the choice of regions in the sampling phase. Asymptotic expressions for the optimal size of regions are derived for one, two, and higher dimensional regions. It is shown that the number of regions increases as the third root of the sample size in the one-dimensional case and as the square root in the two-dimensional case. Formulas are given for proportional and optimal allocations. The results are applied to three existing and quite different statistical surveys.

Key words: Class size; cross-tabulation; geographical partitioning; optimal allocation; region size; spatial statistics; statistical presentation; statistical reporting; stratification.

1. Introduction

Many statistical quantities vary geographically, e.g., unemployment, nativity, precipitation, and housing standards. These variations are often interesting and it is the object of statistics to describe them as well as possible. The variation is usually described by presenting figures for different regions. An interesting and very important question is the best number and size of these regions. This question has no theoretical basis and is usually solved quite ad hoc. In this paper we shall give a mathematical foundation for this problem. We shall also give some practical examples showing that the method can be used in real surveys.

The standard procedure today is to decide the presentation regions in advance. The statistician has to answer questions on the cost of producing accurate figures for counties, for municipalities and for other geographic partitions. He or she computes the number of observations needed to give an acceptable precision. The number of regions is then decided and the survey is conducted. Sometimes the presentation regions are not decided until the survey is completed. In this case the regions are decided so that the precision should be sufficient. However, there is, á priori, no reason to believe that these standard procedures will lead to an optimal solution. A parti-

[1] Professor of Statistics, University of Stockholm, 106 91 Stockholm.

This work herein was partially supported by the Swedish Council for Research in the Humanities and Social Sciences.

Acknowledgement: I want to express my sincere gratitude to the editor Eva Elvers and the referees. The paper benefitted very much from their comments.

tion satisfying a precision constraint may be far from optimal.

In this paper we will use a new approach and view the choice of regions as an optimization problem. If too many regions are chosen the random errors will become too large, but if too few regions are chosen, the regional variation will not be adequately described. What is the "best" partition given the sample and the prior knowledge of the subject matter under study?

An important question is what should be meant by "best." We use the criterion smallest average mean square error (AMSE). This can be explained in the following way. If a random element is represented by a region's average instead of its true value, an error is made. The variance of this error is minimized when the AMSE is minimized. When the regions are too large to describe the regional variation, the representation error often becomes large as does the AMSE. On the other hand, if the areas are small the average in every region will be badly estimated. The representation error and the AMSE will thus be large for this reason. Somewhere between these two extremes is the optimal region size. In order to calculate the AMSE, rough estimates of the sampling variance and the geographical variation are used. These can be obtained from the sample itself or from prior knowledge.

Our approach is similar to those used in the estimation of probability densities using histograms or kernel estimates (see Devroye and Györfi (1985), Silverman (1986) and Prakasa and Rao (1983)). Previous articles on optimal stratification and regionalization have mostly minimized the variance of the population total (e.g., Dalenius and Gurney (1951)). Three other references on related problems are the domains of study in Cochran (1977), some work in Matern (1960) and in Ripley (1981).

We mostly discuss geographical regions, but our results can also be applied to other classification schemes like age groups and income classes. We will, for instance, obtain optimal classification schemes for the margins of two-way tables.

Our results must, of course, be modified for use in practical work. One ought, for instance, to use existing traditional or administrative regions and boundaries. One should use prior knowledge on other relevant variables, like climatic zone or population density, when deciding where the actual boundaries should be drawn. However, we feel that the derived formulas should be useful as rules of thumbs or as warnings for when the data are broken down into too small regions.

In Section 2, the problem is formulated in mathematical terms. This section also contains the solution in the simple case of a one-dimensional country. By this we mean an oblong country, which is so long and narrow that its width can be neglected. The problem is formulated in model-based terms, but we could also have used ordinary design-based terms. Section 3 contains a thorough discussion of this and of our other assumptions.

In Sections 4–7, the model is developed in different ways. We consider optimal allocation in stratified sampling, two and higher dimensional countries, small area or kernel estimation and estimation of functions. The problems of stratified sampling now involves both the choice of sampling intensities and presentation regions simultaneously with the object of minimizing the AMSE. In the last case, estimation of functions, the regional figures are only intermediate results, which will be used in further calculations by the user of the statistics.

Sections 4–7 are more technical than Sections 1–3. We go through a lot of different cases. In order to make the paper reasonably short we have omitted some detail and extensive justifications. In places, a full page mathematical proof has been replaced by a single-line verbal explanation. We hope that this will not cause any serious problems for the reader.

2. One Dimension

2.1. Formulation

We assume that every element in the population can be associated with a point in the area of study. In applications, it may be a person's home or the point where a thermometer is placed. We use a model-based approach and assume that the answer from the individual or the measurement on the element is a random variable, whose distribution depends on the corresponding point x. The mean value is $h(x)$ and the standard deviation is $\sigma(x)$. When needed, we assume that h and σ are smooth functions with one or two derivatives.

Initially we assume that all elements are evenly spread over the whole area, i.e., the corresponding points follow a uniform distribution. We want to estimate $h(x)$ as well as possible. Let $\hat{h}(x)$ denote the estimate for the region containing x. The mean square error at point x is a measure of how good the estimate is at that point. As an overall measure we use the average mean square error (AMSE)

$$\int [h(x)-\hat{h}(x)]^2 dx \,/\, \int dx,$$

where the average is taken over all points in the area. We believe that this is the most natural choice even though there are other candidates. Thus, our goal is to find the partition having the smallest AMSE.

In this section we assume that the area is an interval on the real line. In later sections the results are generalized to higher dimensional regions.

2.2. Solution

Example 2.1. We start by solving our problem with a very simple case where the mean value is a straight line $h(x)=a+bx$ and the variance σ^2 is constant. We let the whole area be the unit interval and assume that it shall be divided into equally large regions. We denote the total sample size by N and shall determine the number of regions m and the region sample size n (which equals N/m) in an optimal way.

This example is illustrated in Fig. 1, where the true mean value $h(x)$ is represented by the sloping thin line. The thick lines are the true averages in the regions. The broken lines are the sample estimates (a realization of the random outcome with this sample).

The squared bias due to the use of a constant in each region instead of the line $h(x)$ is

$$\sum_{i=0}^{m-1} \int_{i/m}^{(i+1)/m} \left(a+bx-a-b\frac{i+1/2}{m}\right)^2 dx = \frac{b^2}{12m^2}.$$

This is the integrated squared difference between the thick and thin lines in the figure.

The variance in each region is

$$\frac{\sigma^2}{n}+\frac{b^2}{12nm^2}.$$

The first term here is the common variance of the mean of n independent random variables. The second term is included because a random element may be located anywhere in the region. We will call this term the random location error. These two terms correspond to the expected squared distance between the thick and broken lines in a region in the figure.

Using $n=N/m$, our total average mean square error (AMSE) is

$$\frac{b^2}{12m^2}+\frac{\sigma^2 m}{N}+\frac{b^2}{12mN}. \tag{2.1}$$

Expression (2.1) is minimized for a given N by

$$m=\left(\frac{(2N+m)b^2}{12\sigma^2}\right)^{1/3} \approx (Nb^2/6\sigma^2)^{1/3}. \tag{2.2}$$

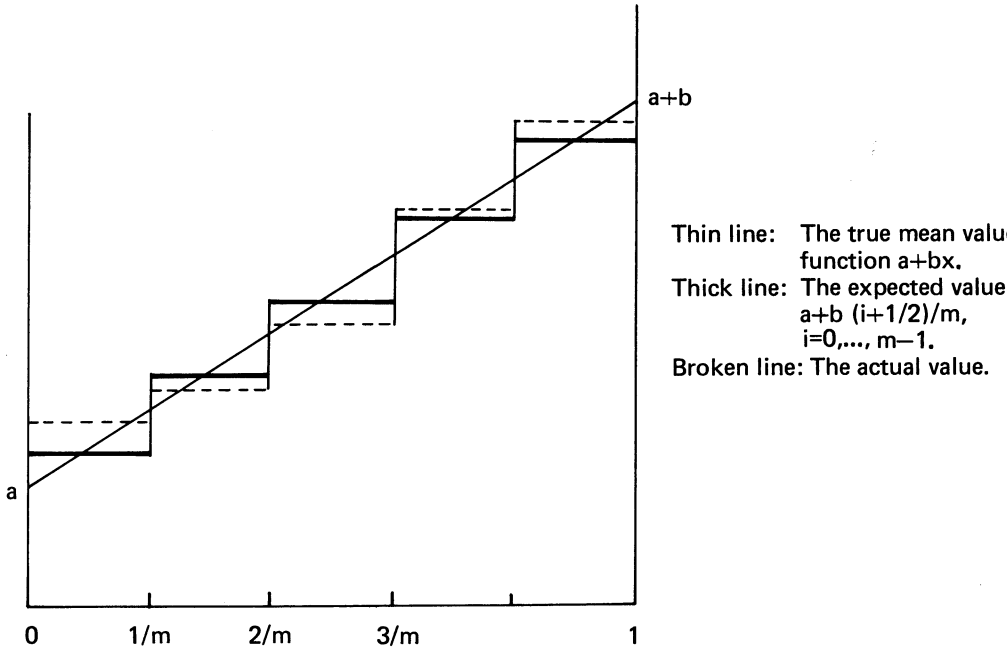

Fig. 1. The simple case with five regions

The number of observations in a region is

$$n = N/m \approx 6^{1/3}|N\sigma/b|^{2/3}. \quad \square$$

Remark 2.1. The expected mean difference between two adjacent regions is for this case

$$b/m \approx \sigma\sqrt{6/n},$$

where the optimal value of m has been inserted.
The standard deviation of the difference between any two region estimates is

$$\sigma\sqrt{2/n}.$$

The coefficient of variation of the difference between two adjacent regions should thus be the ratio $1/\sqrt{3}$. We see that the optimal regions should be so small that the differences seldom become statistically significant. $\square$

In practical work the true function $h(x)$ is seldom a straight line. Example 2.1 can easily be generalized to general differentiable curves.

Theorem 2.1. Let $h(x)$ have a continuous derivative and $\sigma(x)$ be continuous in the area which is given by $(c, c+d)$. Furthermore, let the elements be uniformly distributed in the area and let the loss function be the average mean square error. If all regions are required to be equally large, the optimal number of

observations in each region is

$$[6N^2\int\sigma^2(x)dx/d^2\int(h'(x))^2)dx]^{1/3}+o(N^{2/3})$$

$$\text{as } N\to\infty. \tag{2.3}$$

If the regions are allowed to have different sizes, the optimal number of observations in the region containing x is

$$[6N^2\sigma^2(x)/d^2(h'(x))^2]^{1/3}+o(N^{2/3})$$

$$\text{as } N\to\infty. \tag{2.4}$$

Proof: We prove only formula (2.3). Formula (2.4) is proved along the same lines (c.f. also the proof of Theorem 4.1.).

The total mean square error is

$$\sum_{i=0}^{m-1}\int_{c+id/m}^{c+(i+1)d/m}(h(x)-\bar{h}(x))^2dx\left(1+\frac{1}{n}\right)+\frac{\int\sigma^2(x)dx}{nd}, \tag{2.5}$$

where $\bar{h}(x)$ is the average value of $h(x)$ in the interval containing x. The factor $\left(1+\frac{1}{n}\right)$ is inserted because the squared bias and the random location error are combined here. The expression (2.5) tends to

$$\frac{d}{12m^2}\int(h'(x))^2dx+\frac{\int\sigma^2(x)dx\cdot m}{Nd}\quad\text{as } N\to\infty.$$

The standard differentiation procedure proves formula (2.3). □

Remark 2.2. If $h(x)$ has two continuous derivatives, the error term can be replaced by $o(N^{-1/3})$ both in formula (2.3) and (2.4). The error term in the number of regions is then $o(1)$. □

Remark 2.3. It is easily seen that the random location error is asymptotically small compared to the squared bias. In the following, the random location error will sometimes be omitted without our calling this omission to the reader's attention. □

Remark 2.4. The AMSE is usually rather flat at the minimum and practical considerations must also be taken into account when deciding the actual number of regions. This means that the best choice is not always the exact figure obtained by these formulas. If the number of regions lies between half and twice this number, it is mostly acceptable. On the other hand, it is seldom justified to have more than twice the number of regions obtained by the formulas. □

Example 2.2. A labour force survey is performed in Sweden (Statistics Sweden (1987)) with a sample size of 16 000. The sample is proportionally allocated to 24 counties. The unemployment level varied between 0.5% and 10% in the different counties.

We shall first use the simple model with a straight line and assume a fixed variance of 0.03, which corresponds to the total unemployment level of around 3%. The optimal number of regions is calculated to

$$\left(\frac{16\,000\cdot(0.10-0.005)^2}{6\cdot 0.03}\right)^{1/3}\approx 9.$$

A more realistic model is that the unemployment level can be expected to follow an approximately U-shaped curve like $(22x^2-15x+3)/100$, where $x=0$ in the southern most part of Sweden and $x=1$ at the northern end. This curve was obtained as a free hand parable through the 24 county figures from 1985 ordered from south to north. The model is still unrealistic since we have not yet discussed how to handle non-uniform populations. With this free hand curve, b^2 is replaced by $\int[h'(x)]^2dx\approx 0.021$ and σ^2 by

$\int h(x)(1-h(x))dx \approx 0.027$. The optimal number of equally large regions is then 13.

This substantial change in $h(x)$ resulted only in a modest change in the optimal number of regions (c.f. Remark 2.4). A more detailed curve will only change this result slightly. The optimal number of regions is in most cases fairly insensitive to minor changes in $h(x)$. This is discussed more in Example 3.1.

Unemployment is usually reported for both sexes and three age groups. The optimal number of regions is then decreased by a factor of $6^{1/3}$ to 9, if the differences in size and employment between the groups are disregarded.

If the regions may have different sizes, computations based on (2.4) show that the regions should contain 1 250, 1 350, 5 800, 1 400, 1 300, 1 250, 1 250, 1 200, and 1 200 sampled individuals. The optimal number of regions is thus nine instead of thirteen. The reason is that one large region can be made where the mean value curve, $h(x)$, is flat $(0.16<x<0.52)$. □

3. Discussion of the Assumptions

3.1. *Average mean square error*

If an element is taken at random and its value is represented by the region average, there will be an error. The size of this error is

$$h(x)+\varepsilon(x)-\hat{h}(x),$$

where $\varepsilon(x)$ is a random variable with mean zero and variance $\sigma^2(x)$. The variance of this error is

$$\int (h(x)-\hat{h}(x))^2 dx + \int \sigma^2(x)dx.$$

It is now easily seen that this variance and the AMSE are minimized simultaneously.

The average mean square error is thus a natural measure of how close the estimated and the true curves are. There are other possible choices. The average mean absolute error is less sensitive to large deviations. The optimal region size increases as $N^{2/3}$ in that case too, but it is normally a little larger than the size given by (2.3).

Another possible criterion is maximum mean square error which minimizes the mean square error at its largest point. With this choice the region sizes are determined by the part which is most difficult to describe. This often gives an unnecessarily bad description in a large part of the total area. However, a formula corresponding to (2.3) can easily be derived

$$n = (2N^2\sigma^2(x_0)/d^2(h'(x_0))^2)^{1/3} + o(N^{2/3}),$$

where x_0 is the point with maximum value of $(h'(x))^2 \cdot \sigma^4(x)$. With the criterion, maximum mean square error, it is impossible to derive an analogy to formula (2.4) for areas of different sizes, without further specification, since the maximum mean square criterion deals only with the point x_0. With this criterion one has a wide range of possible region sizes where these deviations are small.

All criteria so far measure the fit of the level and not of differences between neighbouring regions. One criterion that does measure differences between adjacent regions is

$$\frac{d}{d-a} \int_{c+a}^{c+d} (h(x)-h(x-a)-\hat{h}(x)+\hat{h}(x-a))^2 dx,$$

where a is a fixed small number. If h is twice differentiable, $h'(x)$ is replaced by $a \cdot h''(x)$ in the optimal formulae of Theorem 2.1. The regions should thus be larger if the differences between points, which are rather close, are important. This follows from the fact that in this case a is a small number. However, it is our experience that the level is always an essential property. Thus this loss function may possibly be used in combination with the average mean square error, but may not be used as the only criterion.

Moreover, if one is interested in changes over time in different regions, our method can easily be modified. The variance $\sigma^2(x)$ is replaced by the variance of the difference which is $2\sigma^2(x)$ if the two samples are independent. The function $h(x)$ should be replaced by a guess of the magnitude of the change. For unemployment statistics, a magnitude proportional to $h(x)$ may be a fair guess. We use the proportionality factor 0.5, but the choice should, of course, depend on the time scale. With this proportionality factor, the optimal presentation regions become twice as large. This was to be expected, since larger areas are usually needed to reveal significant changes.

In many surveys, several variables are presented simultaneously for the same regions. If the criterion is a weighted sum of the average mean square errors, formulas (2.2), (2.3) and (2.4) can still be used if $\sigma^2(x)$ and $(h'(x))^2$ are replaced by the corresponding weighted sums. However, we think that it is a better practice to do the computations separately for each variable. If the calculations give similar partitions, anyone of these would seem a sensible choice. If the calculations give different partitions, then different presentation regions should be considered for different variables.

There exist other goodness-of-fit criteria than those mentioned here. But we believe that the AMSE is a good choice that has most of the desirable properties.

3.2. *Mean value function*

In Theorem 2.1 we assumed that $h(x)$ is absolutely continuous. If it is known that $h(x)$ is very smooth or even belongs to a certain parametric family, one might argue that the data ought to be presented in a way that reflects this knowledge rather than as constant averages in a couple of regions. However, we do not believe that the user of the statistics would be interested in the unemployment level given by a spline function of the latitude and the longitude. We believe that smoothness properties may be used in the estimation phase, but that often regional figures are the only acceptable presentation form.

If $h(x)$ is not absolutely continuous, the regions should be smaller. For instance, let $h(x)$ be a self-similar fractal of dimension p (Mandelbrot (1983)). This can be interpreted as saying that $h(x)$ and $b^{p-2}h(b(x+a))$ have the same type of irregularities for all positive a and b. A differentiable function is self-similar with dimension $p=1.0$, a Wiener process with $p=1.5$, and white noise with $p=2.0$. The optimal number of regions is then asymptotically of the order $N^{(1/(5-2p))}$. This can be shown using the self-similarity property in (2.5).

In practice $h(x)$ is unknown and must be guessed. This can often be done from previous knowledge but also from the sample itself. Theorem 2.1 is fairly robust against misspecification of $h(x)$. This is illustrated in the following example. It shows that details smaller than the interval width should not be included in $h(x)$. Hence, even if the true $h(x)$ is quite irregular one could safely assume it to be smooth.

Example 3.1. Consider the labour force survey once more. Suppose that the true function is

$$h(x)=\begin{cases}(22x-15x+3+0.3\sin(20\pi x)), & 0.1<x\leqslant 0.6\\ (22x-15x+3+\sin(20\pi x)), & x\leqslant 0.1 \text{ or } x>0.6.\end{cases}$$

The asymptotically optimal formula (2.3) now gives 13 regions if the age-sex grouping is considered. The sine curve has a period of 0.1 which means that at least 20 regions are necessary to describe the small fluctuations. Thus the analysis in Example 2.2 is more adequate. □

Note that Remark 2.1 for adjacent regions holds also when formula (2.4) is used for the

optimal number of observations in regions of different sizes. If the regions are chosen so that the coefficients of variation for the differences are close to $1/\sqrt{3}$, a reasonable partitioning is obtained.

3.3. *Population distribution*

In Theorem 2.1 we assumed that the sampling points were evenly spread over the whole area. If the population density $p(x)$ is non-uniform, the calculations may be changed for two reasons.

1. The density of the sample can also be $p(x)$.
2. The average mean square error may be taken over the population rather than the area, i.e.,

$$\int (h(x)-\hat{h}(x))^2 p(x)\, dx.$$

If both these changes are made $(h'(x))^2$ in formulae (2.3) and (2.4) must be multiplied by $p(x)$. If only the first change is made and the initial average mean square error formula is used, one should divide $\sigma^2(x)$ by $p(x)$ in (2.3) and leave (2.4) unchanged.

It is also possible to take into account the importance of different areas with other weights than $p(x)$. However, the optimal formula (2.4) will not change at all if this change is made. The results given in subsequent sections would change, but we will not discuss that problem in this paper.

3.4. *Nongeographical partitioning and other estimates than the average*

Our methods can be used for other background variables than geographical regions.

Table 1. Salary per month (SEK) according to date of birth for JUSEK-members employed by the public sector (born 1936–1959)

Date of birth	Government		Municipalities	
	number	median salary	number	median salary
1936	200	12 545		
1937	208	12 732		
1938	195	12 350	311	13 437
1939	242	12 345		
1940	247	12 796		
1941	234	11 954	84	13 437
1942	329	11 493	116	12 755
1943	403	11 578	123	13 085
1944	464	11 578	125	12 755
1945	434	11 178	125	12 437
1946	398	10 889	121	12 153
1947	456	10 512	130	11 877
1948	391	10 325	115	11 611
1949	351	10 024	100	11 115
1950	320	9 712	84	10 741
1951	269	9 380	82	10 741
1952	250	9 435	55	10 225
1953	220	8 948	48	10 067
1954	218	8 807	45	10 067
1955	238	8 469	44	9 789
1956	217	8 257	45	9 789
1957	226	8 257	36	9 156
1958	258	8 082	30	8 876
1959	239	7 924	34	8 614

This is best illustrated with an example.

The example will also illustrate that our method is not confined to estimates of an average level or proportion. It deals with monthly salaries and it is well known that the median works better for that variable, since it is not so influenced by extreme elements.

Example 3.2. JUSEK, a Swedish labour organization, produces statistics for its members on the median monthly salary according to age (JUSEK (1986)). For governmental employees the statistics are based on about 300 members in each one year age group with year of birth between 1936 and 1959. The figures are given in Table 1.

A rough, data-based estimate for $h(x)$ is $(550+290x)$ SEK, where x is the age. The standard deviation of the median can be estimated to be $(90x-1\,800)/u$ SEK (based on data not presented here) where u is the number of persons in that age group. Restricting the partitioning to regions of equal size, Formula (2.3) gives the optimal $n=265$. In other words, it is quite sensible to report the median salary for each age group.

JUSEK also reports the monthly salary for municipal employees in a similar table. In that table there are only 70 persons in each age group. The optimal formula now gives the sample size 165 in the optimal classes. Thus, that presentation is not adequate. □

3.5. Finite populations and design-based versus model-based sampling

To describe a true mean value function as well as possible, we have used a model-based approach where the model parameter is the true value. These methods can be used for finite populations where the object is to estimate the average over the whole population. One must, however, add different corrections for finite populations to the three components of the error. In practical examples the method works well without the correction factor. If the correction is used when the sampling fraction is 100%, the minimum is obtained when $m=N$. To have as many regions as there are individuals is, however, not realistic. We recommend that the method be used without corrections even when the sampling fraction is considerable.

In this paper we used a model for an observation at a single point. The main reason for this is mathematical convenience. With integrals and differentiable functions, all discrete boundary problems can be avoided. But similar results can also be proved in a classic design-based case.

Let the total population contain Q objects. Each object is characterized by its location x_i and its value h_i, $i=1, 2, \ldots, Q$. As before $\hat{h}(x)$ denotes the estimate in the region containing x. The average mean square error criterion is in this case

$$\sum_{1}^{Q}(h_i-\hat{h}(x_i))^2.$$

In other words, we want to minimize the variance of the error that is generated by representing the individual values h_i by the corresponding region estimates.

The results in this paper are still approximately true if all regions are connected (and, in higher dimensions, convex). The function $h(x)$ must then be interpreted as a suitable smooth moving average of the h_i. With this interpretation our results are independent of whether the statistician has used model or design-based sampling.

4. Stratified sampling

So far, simple random sampling has been used and the sample size has been fixed. The optimal size of the regions was then derived. It is

possible to derive similar formulae for other fixed sampling schemes (e.g., pps or cluster sampling) and other estimation techniques (e.g., regression estimates). The derivations and the results are similar, but the formulae and notations become more complicated. Here we will not derive such formulae.

In this section we shall assume that neither the sampling fractions nor the design are given. Instead we shall use a flexible stratification scheme. We may change not only the presentation regions but also the stratum widths and the sampling fractions. Only the total sample size N is fixed and we still assume that the loss function is the AMSE

$$\int_0^1 (h(x)-\hat{h}(x))^2 dx,$$

where $\hat{h}(x)$ is the estimated level in the stratum that contains x. For simplicity, and without loss of generality, we assume that the whole area is the unit interval.

Let $m(x)$ be the number of intervals per length unit at x and $Nt(x)$ be the sampling intensity. The length of the presentation region containing x is thus around $1/m(x)$. The optimal function $m(x)$ varies with N, but it will turn out that $m(x)\cdot N^{-1/3}$ tends to a limit as $N \to \infty$. An approximate expression for the average mean square error is

$$\int \frac{(h'(x))^2}{12m^2(x)} dx + \int \frac{\sigma^2(x)m(x)}{Nt(x)} dx \qquad (4.1)$$

with the restriction $t(x)dx = 1$. These terms correspond to the first two terms in (2.1). This expression is minimized in the appendix. The following theorem is obtained.

Theorem 4.1. Let $h'(x)$ and $\sigma(x)$ exist and be continuous. Let the loss be the average mean square error. The asymptotic results on the optimal sampling intensity $Nt(x)$ and the number of presentation intervals are given through

$$t(x) = |h'(x)|^{2/5}\sigma(x)^{4/5} / \int |h'(y)|^{2/5}\sigma(y)^{4/5} dy$$

and

$$N^{-1/3}m(x) = \frac{|h'(x)|^{4/5}}{\sigma(x)^{2/5}6^{1/3}\left(\int |(h'(y)|^{2/5}\sigma(y)^{4/5}dy\right)^{1/3}}.$$

□

Example 4.1. (Continued from Example 2.2.) Suppose that the unemployment level approximately follows the model given in Example 2.2 with $h(x) = (22x^2-15x+3)/100$ and $\sigma^2(x) = h(x)(1-h(x))$. The invariant functions $m(x)N^{-1/3}$ and $t(x)$ and the region width $1/m(x)$ are calculated in Table 2.

Table 2. The optimal asymptotic region size when $h(x) = (22x^2-15x+3)/100$, $N = 16\,000$

Location, x	$m(x)N^{-1/3}$	$t(x)$	$1/m(x)$
0	0.53	1.19	0.074
0.1	0.45	0.83	0.088
0.2	0.34	0.51	0.118
0.3	0.14	0.24	0.281
0.4	0.19	0.30	0.213
0.5	0.36	0.57	0.110
0.6	0.47	0.90	0.085
0.7	0.55	1.26	0.072
0.8	0.61	1.64	0.065
0.9	0.67	2.03	0.059
1.0	0.72	2.43	0.055

By interpolating in Table 2, we may easily find that a suitable size of the southern most stratum is roughly (0, 0.08) with a sample size of about $16\,000\times0.08\times1.1 \approx 1\,400$ individuals. In this way it is seen that a good partitioning is the following (0, 0.08, 0.19, 0.38, 0.50, 0.60, 0.68, 0.75, 0.82, 0.88, 0.94, 1.00) with 11 regions. The corresponding optimal stratum sample sizes are 1 400, 1 200, 1 000, 800, 1 150, 1 300, 1 450, 1 750, 1 750, 2 000, and 2 200, respectively.

These results make sense. It is more efficient to use stratified sampling compared to simple random sampling. It is thus possible to get a more informative presentation than in Example 2.2, where nine intervals were best.

□

5. Two or More Dimensions

So far we have considered only one dimension. It has mostly been a geographical dimension such as latitude but also non-physical dimensions such as age (Example 3.2) have been mentioned. We now turn to two and more dimensions. When more than two dimensions are studied at least one of them must be non-geographical.

In geographical problems with two dimensions the regions may have any shape. On the other hand, non-geographical regions are often one-dimensional and it is natural to use rectangular regions. The most common example is ordinary cross tabulation between two variables. In this section we first concentrate on rectangular regions, but in the last theorem we shall also mention arbitrary shapes.

If only one variable is studied, the optimal regions will be long bands along the isolines where $h(x)$ remains constant. In order to partition the country so that the optimal regions do not form long and narrow bands, we shall assume that several variables are studied and shall be presented for the same regions. We shall assume that there are at least as many variables as there are dimensions, even though this restriction is necessary only for arbitrary shapes.

Example 5.1. Let us first consider the simple case where k mean value functions are each a linear function of one distinct coordinate on the k-dimensional unit cube. Furthermore we also assume that the elements are uniformly spread in the unit cube. Our objective is to divide the cube into $m_1 \times m_2 \times \ldots \times m_k$ rectangular areas.

The average mean square error loss function is

$$\sum_{j=1}^{k} \left(\frac{b_j^2 \omega_j}{12 m_j^2} + \frac{\sigma_j^2 \omega_j}{N / (\prod_{i=1}^{k} m_i)} \right), \tag{5.1}$$

where b_j, ω_j, and σ_j^2 are the slope, the importance (weight), and the variance of the jth variable, respectively. Here we have omitted the random location error corresponding to the last term of (2.1) since that term is negligible.

A maximization of (5.1) leads to the optimal number of regions

$$\prod_{j=1}^{k} m_j = 6^{-k/(k+2)} (\Pi b_j^2 \omega_j)^{1/(k+2)} (\Sigma \sigma_j^2 \omega_j)^{-k/(k+2)} N^{k/(k+2)} \tag{5.2}$$

and

$$\frac{m_i}{m_j} = \frac{b_i (\omega_i)^{1/2}}{b_j (\omega_j)^{1/2}} . \tag{5.3}$$

□

The rectangular areas are not optimal, but may sometimes be natural to use, e.g., in cross tabulation. It may be shown that in two dimensions the asymptotically optimal shape is an affine image of a regular hexagon. In higher dimensions we believe that the optimal shape is an affine image of a regular polyhedron with $k(k+1)$ sides, but we have not been able to prove that.

Remark 5.1. In Section 2.1 we noted that the coefficient of variation of the difference between two one-dimensional adjacent regions should be $1/\sqrt{3}$. The same result also holds here if the shapes are rectangular. The coefficient of variation is slightly smaller for optimal region shapes. □

Example 5.2. All Swedish farms are annually registered in the Swedish farm register and classified according to size (Statistics Sweden (1985)). A transition matrix for flows between size groups is calculated.

In a transition matrix many entries are reported, so this case is certainly not one-dimensional. A typical transition probability may vary between 0.10 in southern Sweden and 0.15 in northern Sweden. The differences in the transition probabilities between east and west can be estimated to about half this size. There are about 25 000 farms in a typical class. Using the simple formula (5.2) we get that the optimal number of regions is

$$\left(\frac{25\,000\cdot(0.15-0.10)(0.15-0.10)/2}{6\cdot(0.125\cdot 0.875+0.125\cdot 0.875)}\right)^{1/2}\approx 5\,.$$

Statistics Sweden reports transition matrices for eight production areas. In Remark 2.4 we said that a ratio in the interval 1/2 to 2 is quite acceptable. The present partition is thus probably rather good. □

Formulas (5.2) and (5.3) are generalizations of Example 2.1. It is also possible to extend Theorem 2.1 to more general cases. We still concentrate on rectangular areas and assume that the elements are uniformly distributed. The following theorem can easily be proved in the same way as (5.2) and (5.3).

Theorem 5.1. Let the total region have the area d and let there be l variables whose true values are differentiable functions $h_j(\underline{x})$ and suppose that the variance function $\sigma_j^2(\underline{x})$ are continuous. Further suppose that each marginal must have a constant interval length. Let the interval length for the kth marginal be $1/m_k$.

The optimal number of regions is then given by

$$\prod_{j=1}^{k} m_j=(N/(6\sum_{i=1}^{l}S_i^2))^{k/(k+2)}(d^2\Pi \mathrm{B}_j^2)^{1/(k+2)}$$

and

$$m_i/m_j=\mathrm{B}_i/\mathrm{B}_j\,,$$

where

$$\mathrm{B}_j^2=\sum_{i=1}^{l}\left(\int\!\!\int\left[\frac{\partial h_i(\underline{x})}{\partial(x_j)}\right]^2\prod_{k=1}^{l}dx_k\right)\omega_i$$

and

$$S_i^2=\int\!\!\int\sigma_i^2(\underline{x})\prod_{k=1}^{l}dx_k\omega_i\,. \qquad \square$$

It is possible to generalize this result to rectangular areas where the marginals may have varying interval lengths. We will not do so. Instead we will state Theorem 5.2, which is a generalization of Theorem 4.1. Theorem 4.1 treated stratified sampling with varying probabilities from a population which was uniformly spread over the whole area. The proof of Theorem 5.2 is omitted but can be performed along the same lines as in the Appendix, but it is more intricate.

Theorem 5.2. Suppose that a survey shall measure l variables on a k-dimensional background space with the area d ($l\geq k$). The observation of variable i at location $\underline{x}=(x_1,\ \ldots,\ x_k)$ is a random variable with a differentiable mean $h_i(\underline{x})$ and variance $\sigma_i^2(\underline{x})$. The loss function is

$$\sum_i\omega_i\int|h_i(\underline{x})-\hat{h}_i(\underline{x})|^2\underline{dx}.$$

The asymptotically optimal sampling intensity, $Nt(x)$, is given by

$$t(\underline{x})=\frac{|\det H(\underline{x})|^{1/(4+k)}(\Sigma\omega_j\sigma_j^2(\underline{x}))^{2/(4+k)}}{\int|\det H(\underline{y})|^{1/(4+k)}(\Sigma\omega_j\sigma_j^2(\underline{y}))^{2/(4+k)}dy},$$

where $|\det H(x)|$ is the absolute determinant of a matrix $H(x)$ with the element at place p, q equal to

$$\sum_i\frac{\partial h_i(\underline{x})}{\partial x_p}\cdot\frac{\partial h_i(\underline{x})}{\partial x_q}\cdot\omega_i\,.$$

The asymptotically optimal number of presentation regions per unit area is

$$m(\underline{x}) = \frac{\left(\frac{N}{dc_k}\right)^{2k/(4+k)} |\det H(x)|^{2/(4+k)}}{[\Sigma\omega_j\sigma^2(\underline{x})]^{k/(4+k)}[\int|\det H(\underline{y})|^{1/(4+k)}(\Sigma\omega_j\sigma_j^2(\underline{y}))^{2/(4+k)}d\underline{y}]^{2/(4+k)}},$$

where c_k is a constant depending on the shape of the region. When we restrict ourselves to rectangles or generalized parallel epipeds its value is 6. The optimal value of c_2 is $18\sqrt{3}/5$. It is also true that

$$6 < c_k < \frac{k+2}{2}\left(\pi^{(k/2)}2^{[(k+1)/2]/k!!}\right)^{2/k}$$

for the optimal c_k $(k>1)$, where $[\cdot]$ denotes the integer part. □

Theorem 5.2 seems to model a more realistic case than Theorem 5.1. However, it is more difficult to use and larger samples are needed before the asymptotics work. It is also seen that the formal gain with nonrectangular shapes is small for all dimensions encountered in practice.

Theorem 5.1 is probably too simple for many cases where the elements are not uniformly distributed and where the variances and derivatives vary considerably. It is, of course, possible to derive theorems corresponding to most practical situations. We believe that in many situations where several variables are encountered, it is not evident that all variables should be presented for the same regions. Thus we believe that rectangular areas often are sufficient, but that the sizes of the rectangles may vary considerably between different parts of the area and different variables.

6. Smoothing Between Small Adjacent Regions–Kernel Estimators

When the regions are small it is often possible to improve the estimators by using adjacent regions. If the function $h(x)$ is smooth it may be a good idea to incorporate the sampled elements from adjacent regions in the estimators but with smaller weights. This technique is called kernel estimation and is sometimes used in small area estimation.

In this section we shall only consider the very simple one-dimensional case on the interval (0,1) where statistics are reported for the intervals $(i/m, (i+1)/m)$. The estimates for these regions are the simple averages over the interval $((i-p)/m, (i+p+1)m)$. (In this simple example we neglect all boundary effects.) The mean value function is assumed to be $h(x)$.

The average mean square error is given by

$$\sum_{i=0}^{m-1}\int_{i/m}^{(i+1)/m}\left(h(x)-\frac{m}{2p+1}\int_{(i-p)/m}^{(i+p+1)/m}h(z)dz\right)^2dx$$

$$+\frac{m}{N(2p+1)}\left[\sigma^2+\sum_{i=0}^{m-1}\frac{m}{2p+1}\int_{(i-p)/m}^{(i+p+1)/m}(h(x)\right.$$

$$\left.-\frac{m}{2p+1}\int_{(i-p)/m}^{(i+p+1)/m}h(z)dz\right)^2dx\Big]. \qquad (6.1)$$

The last term corresponding to the random location error is, as usual, negligible compared to the random reporting error. The expression within brackets in the first term can be written

$$[\{h(x)-h((i+1/2)/m)\}+\{h((i+1/2)/m)$$

$$-\int_{(i-p)/m}^{(i+p+1)/m}h(z)dz\}]^2.$$

Expanding $h(x)$ in a Taylor series using only first order terms in the first part and second order terms in the second part gives after some calculation that (6.1) is approximately equal to

$$\frac{1}{12m^2}\int_0^1 (h'(x))^2dx+\frac{(2p+1)^4}{576m^4}\int_0^1 (h''(x))^2dx + \frac{\sigma^2 m}{N(2p+1)}. \quad (6.2)$$

This expression is maximized for fixed N by

$$(2p+1)/m=\left(\frac{144\sigma^2}{N\int(h''(x))^2dx}\right)^{1/5} \quad (6.3)$$

with m as large as possible, for instance $N^{0.5}$ or $N^{0.8}$. The width of the wider intervals should thus decrease as $N^{-0.2}$.

This expression is based on the fact that all regions are equally long. However, the asymptotic order will be the same if the regions may have different lengths, if the variance σ^2 may vary in the interval, and if stratified sampling is allowed. The relevant formulas can in that case be shown to be

$$\frac{m(x)}{2p(x)}=\frac{|h''(x)|^{4/9}N^{1/5}}{144^{1/5}(\sigma(x))^{2/9}(\int|h''(y)|^{2/9}\sigma(y)^{8/9}dy)^{1/5}}$$

and

$$t(x)=\frac{|h''(x)|^{2/9}\sigma(x)^{8/9}}{\int|h''(y)|^{2/9}\sigma(y)^{8/9}dy},$$

where $p(x)$ and $m(x)$ are the obvious generalizations of p and m when they are allowed to depend on x.

These results are further improved if distant observations carry less weight than observations close to each other, i.e., more bell-shaped kernèls are used instead of

$$K(x)=\begin{cases}1/2 \text{ if } |x|<1\\ 0 \quad \text{if } |x|\geq 1,\end{cases}$$

which was used in (6.1).

These one-dimensional results can be generalized to k dimensions with rectangular areas in the same way as in previous sections. One must, however, remember that the boundary effects seldom may be totally neglected, since formula (6.3) often give large values for $(2p+1)/m$.

7. Estimation of Functions

7.1. *General*

If one is interested only in a total figure for the whole population, there is no need to report statistics for regions. Sometimes, however, it is better to use regions in the calculations and to sum the figures in the final report. This can be illustrated by an example.

Example 7.1. Suppose that we want to make forecasts using a Markov chain technique. The forecast for the total region is $X\hat{P}^T$, where X is a row vector with the observed numbers in each state, $\hat{P}$ is the observed transition matrix for the whole country based on previous years, and T is the time horizon of the forecast. If the forecast is based on m regions, the formula will be $\sum_{i=1}^{m} X_i\hat{P}_i^T$, where $\hat{P}_i$ is the observed matrices for the regions. We will see that this expression sometimes yields a smaller mean square error than $X\hat{P}^T$. □

In other words, we want to estimate $\int g(h(x))dx$ with mean square loss, where g is a known function and where an observation at point x gives the results $h(x)$ plus a random quantity.

7.2. *One-dimensional regions*

Suppose that the expected value is $h(x)$ for elements observed at the point x and that the variance is σ^2. We shall estimate the integral $\int_0^1 g(h(x))dx$, where g is a twice differentiable

function, by the sum $\frac{1}{m}\sum_{i=0}^{m-1} g(\bar{z}_i)$ where $\bar{z}_i$ is the mean of the observations in the region $(i/m, (i+1)/m)$.

The mean square error is given by

$$\{E[\Sigma \int_0^1 g(h(x))dx - \frac{1}{m}\Sigma g(\bar{z}_i)]\}^2$$

$$+\Sigma \mathrm{Var}[g(\bar{z}_i)/m]$$

$$\approx \{\Sigma \frac{1}{2} g''(h(\frac{i}{m}))\left[\int_{i+m}^{(i+1)/m} (h(x)-\bar{h}_i)^2 dx\right.$$

$$\left.+\frac{1}{m}\mathrm{Var}(\bar{z}_i)\right]\}^2 + \Sigma \frac{1}{m^2} g'(h(\frac{i}{m}))^2 \frac{\sigma^2 m}{N},$$

where $\bar{h}_i = E(\bar{z}_i) = m \int_{i/m}^{(i+1)/m} h(x)dx$. To obtain this approximate equality, $g(x)$ was expanded in a Taylor series with second order terms. After approximating also $h(x)$ and performing the integral we get

$$\{\Sigma \frac{1}{2m} g''(h(\frac{i}{m}))\left[\frac{(h'(x))^2}{12m^2} + \frac{\sigma^2 m}{N}\right]\}^2$$

$$+\Sigma_i [g'(h(\frac{i}{m}))]^2 \frac{\sigma^2}{Nm}$$

$$\approx \left[\int_0^1 g''(h(x))(h'(x))^2 dx \frac{1}{24m^2}\right.$$

$$\left.+\int_0^1 g''(h(x))dx \frac{\sigma^2 m}{N}\right]^2 + \int (g'(h(x))^2 dx \frac{\sigma^2}{N}$$

$$= \left[\frac{K_1}{m^2} + \frac{K_2 m}{N}\right]^2 + \frac{K_3}{N}. \qquad (7.1)$$

Note here that the first term of (7.1) corresponds to the square of the ordinary AMSE which was introduced in Section 2 when $g''(x)$ is a constant (i.e., if g is a second order polynomial).

If the number of regions increases with a rate between $N^{1/4}$ and $N^{1/2}$, the second term will dominate asymptotically. All such choices will thus lead to asymptotically optimal solutions. This holds in particular for formulas (2.2)–(2.4). It is also possible to generalize the results of Section 6 to this problem.

7.3 Two or more dimensional regions

In the previous subsection, the number of regions could vary considerably without affecting the asymptotic optimality. This result is special to one-dimensional problems. In the simple case with a function of k independent variables and rectangular regions, the formula corresponding to (5.1) is

$$(\sum_{i=1}^{k} \frac{K_i}{m_i^2} + K_p \frac{\Pi m_i}{N})^2 + \frac{K_0}{N}, \qquad (7.2)$$

where $K_1, \ldots, K_k$, K_p, and K_0 are constants which can be expressed in g, h_i, and σ_i, $i=1, \ldots, k$. The first term here is the square of the ordinary AMSE given by formula (5.1) for a suitable choice of g.

If $k=2$ and if the m_i are chosen optimally the first and second terms of (7.2) are of the same order. If $k>2$ the first term dominates asymptotically. The optimal number of regions are in both cases proportional to $N^{k/(k+2)}$ which corresponds to (5.2) These results on the asymptotic order hold also for other cases, such as non-rectangular regions or stratified sampling with varying shapes and sizes. Let us collect our results so far in a theorem.

Theorem 7.1. Suppose that we shall estimate $\int g(h(x))dx$ by $\frac{1}{m}\Sigma g(\bar{z}_i)$ where $\bar{z}_i$ is the mean of the observations in the ith of m re-

gions. If g is twice differentiable and h continuous the optimal region sizes in Theorems 2.1, 3.1, 5.1, and 5.2 are still asymptotically optimal for this estimation problem. □

Remark 7.1. We noted above that in one dimension all solutions with rate increases between $N^{1/2}$ and $N^{1/4}$ are asymptotically optimal. In higher dimensions the optimal sizes are essentially unique. □

Remark 7.2. In Section 6 we found that smoothed or kernel estimates improved the average mean square error. When estimating functions, as we do here, kernel estimates are unnecessary in one-dimensional problems, since the error is determined by the last term of (7.2). In higher dimensions, however, they can improve the mean square error. Instead of minimizing (7.2), the following formula should then be minimized

$$(\sum_{i=1}^{k} \frac{C_i P_i^4}{m_i^4} + \frac{C_p m_i}{N\Pi P_i})^2 + \frac{C_0}{N},$$

where $C_1, \ldots, C_n$, C_p, and C_0 are constants. If this expression is minimized, the sizes, $\Pi(2p+1)/m_i$, of the regions are of the order $N^{-k/(k+4)}$. The order of the first terms of the mean square error are then $N^{-8/(k+4)}$. This is smaller than N^{-1}, as long as the number of dimensions is smaller than four.

One might believe that kernel estimates are not useful in this context since estimates from adjacent areas are added anyway. As we noticed above, the asymptotic mean square error is of the order $N^{-4/(k+2)}$ without kernel estimates when k is at least 2. For kernel estimates, the corresponding error is $N^{-\min(1,8/(k+4))}$. When there are more than two dimensions, the gain with smoothed estimates is substantial. □

Example 7.2. (Continued from Examples 5.2 and 7.1.) Suppose that one wants to forecast the Swedish farm structure using the forecast $\sum_{i=1}^{m} X_i \hat{P}_i^T$, where X_i is the actual structure this year and $\hat{P}_i$ the observed transition matrix. The forecast is linear in X_i but non-linear in $\hat{P}_i$, so the situation in this section applies.

Sweden is rather oblong, which means that the problem is essentially one-dimensional. The number of regions can thus vary within wide limits. When T is two, the function g is second order and the first term of (7.1) is proportional to the square of (2.1). If the appropriate values are inserted into the optimal function (2.2), it is seen that between 5 and 10 regions are suitable. If the problem was considered two-dimensional the same number of regions would follow from Example 5.2. □

8. Final Comments

We have developed a method of finding the optimal number of regions that describe as well as possible the geographical variation.

To give the best description of the regional variation the number of regions should increase with the sample size. If the region or stratum is k-dimensional, the number of regions should increase as $N^{k/(k+2)}$ (Section 5). For example, if the sample size in a two-dimensional problem is doubled, the optimal number of regions only increases by 41%. This result holds true when simple random sampling, optimal stratification or other sampling plans are used even though the proportionality factor varies.

Although the formulas are asymptotic in nature, the results work fairly well for the sample sizes of most surveys. For one or two dimensions it is seldom justified to use more than twice the number given by the optimal

formula. However, when partitioning after many background variables, the boundary problems may become serious. In those cases, both three and even four times as many regions may be good for common sample sizes. But the boundary problem may have the opposite effect too, so the best number may be only one third of that given by the formula.

We have calculated the optimal number of regions for many surveys. Some examples were given in this article where the optimal value agreed fairly well with that actually used. We have also identified many surveys where there obviously were too many regions, but these examples have not been discussed here.

In this paper, we have generalized the method in a number of ways. Other possible generalizations remain to be done. We have not discussed what to do when supplementary information exists as in ratio or regression estimates. Another problem is the generalization of Sections 6 and 7 to macro databases where the regions are not used by themselves but as buildingstones in larger regions specified by different users.

Appendix 1

The minimization of $\int \frac{(h'(x))^2}{12m^2(x)}dx + \int \frac{\sigma^2(x)m(x)}{Nt(x)}dx$ *subject to the restriction* $\int t(x)dx = 1$.

It is well-known that in optimal allocation $t(x) = \sigma(x)\sqrt{m(x)} / \int \sigma/(y)\sqrt{m(y)}dy$. (A1)
It remains now to minimize

$$\int \frac{(h'(x))^2}{12m^2(x)}dx + \frac{1}{N}(\int(\sigma(y)\sqrt{m(y)})dy)^2.$$

Using variational calculus we differentiate with respect to $m(x)$ $(x \in (0,1))$ and get

$$-\frac{(h'(x))^2}{6m^3(x)} + \frac{1}{N}\int \sigma(y)\sqrt{m(y)}\,dy \cdot \sigma(x)/\sqrt{m(x)}. \tag{A2}$$

If this is set equal to zero, we get

$$m(x)^{5/2} = \frac{N(h'(x))^2}{6\int \sigma^2(y)\sqrt{m(y)}dy \cdot \sigma(x)}$$

and

$$\sigma(x)\sqrt{m(x)} = \frac{N^{1/5}|h'(x)|^{2/5}\sigma(x)^{4/5}}{6^{1/5}(\int \sigma(y)\sqrt{m(y)}dy)^{1/5}}. \tag{A3}$$

An integration shows that

$$(\int \sigma(y)\sqrt{m(y)}dy)^{6/5} = \frac{N^{1/5}}{6^{1/5}} \int |h'(y)|^{2/5} \sigma(y)^{4/5} dy.$$

When this is inserted into (A1) and (A3) we get

$$m(x) = \frac{N^{1/3} \cdot |h'(x)|^{4/5}}{\sigma(x)^{2/5} 6^{1/3} (\int |h'(y)|^{2/5} (\sigma(y))^{4/5} dy)^{1/3}} \tag{A4}$$

and

$$t(x) = \frac{|h'(x)|^{2/5} (\sigma(x))^{4/5}}{\int |h'(y)|^{2/5} (\sigma(y))^{4/5} dy}. \tag{A5}$$

Differentiating Formula (A2) once more gives

$$\frac{(h'(x))^2}{2m^4(x)} - \frac{1}{2N} \int \sigma(y)\sqrt{m(y)}dy \cdot \sigma(x) \cdot m^{-3/2}(x)$$

$$= \frac{1}{2m(x)} \cdot [\frac{(h'(x))^2}{6m^3(x)} - \frac{1}{N} \int \sigma(y)\sqrt{m(y)}dy \cdot \sigma(x)/\sqrt{m(x)}] + \frac{5(h'(x))^2}{12m^3(x)} > 0.$$

The inequality follows from (A2) which is set equal to zero.

This shows that (A4) and (A5) give the unique minimum.

9. References

Cochran, W.G. (1977): Sampling Techniques. Third Edition, Wiley, New York.

Dalenius, T. and Gurney, M. (1951): The Problem of Optimum Stratification II. Skandinavisk Auktarietidskrift 34, pp. 133–148.

Devroye, L. and Györfi, L. (1985): Nonparametric Density Estimation, The L_1-view. Wiley, New York.

JUSEK (1986): Lönestatistik 1986, Offentliga sektorn. Report, JUSEK, Stockholm, Sweden. (In Swedish.)

Mandelbrot, B. (1983): The Fractal Geometry of Nature. Freeman and Co, New York.

Matern, B. (1960): Spatial Variation. Medd. Statens skogsforskningsinst., 49(3), Stockholm, Sweden. Second Edition (1986), Lecture Notes in Statistics, Springer-Verlag, Berlin.

Prakasa, F. and Rao, B.L.S. (1983): Nonparametric Functional Estimation. Academic Press, New York.

Ripley, B.D., (1981): Spatial Statistics. Wiley, New York.

Silverman, B.W. (1986): Density Estimation in Statistics and Data Analysis. Arrowsmith, Bristol.

Statistics Sweden (1985): Reports from the Farm Register 1984. J30 SM 8502 and 8504. (In Swedish with English summary.)

Statistics Sweden (1987): The Labour Force Survey. Am 10 SM 8702 and Am 11 SM 8702. (In Swedish with English summary.)

Received November 1986
Revised September 1988