MULTIVARIATE STRATIFICATION FOR A MULTIPURPOSE SURVEY

by

Garth Bode
B.Sc. (Hons) (Adelaide)

Thesis submitted as part
requirement for the degree of

Master of Science
(by coursework in Statistics),

Australian National University,

January 1983.

C O N T E N T S

S U M M A R Y

This report is the outcome of a study into the problem of stratification in a multipurpose sample survey. Theoretical aspects of the problem were investigated, and an empirical study made using population data. The sample survey in question is an annual crop survey for New South Wales conducted by the Australian Bureau of Statistics, and involves nine estimation variables. For each of wheat, oats and barley, the variables are (i) area sown for all purposes; (ii) area sown for grain; and (iii) area likely to be harvested for grain.

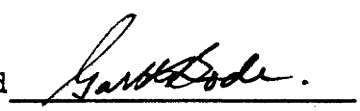
Supplementary (prior) information, both qualitative and quantitative, is available for all population units (land holdings), and is used mainly to increase the precision of estimates. Quantitative prior information consists of predicted values for the variable "area sown for all purposes" for each crop and each holding, and is obtained from census data. Thus each holding has associated with it a known 3- vector which may be used in a variety of ways such as size stratification, unequal probability sampling and ratio estimation. The geographic location of each holding is also known, and this may be used for further stratification, thus enabling estimates to be made over subsets of the population, such as statistical divisions.

In this study, major interest was centred on the use of the quantitative supplementary information for size

stratification. Methods of multivariate size stratification and allocation were investigated and a comparative study of various designs made. Population data was analysed and used to provide an efficient design for future surveys.

M E M O R A N D U M

This report contains no material which has been accepted for the award of any other degree or diploma in any university, and to the best of my knowledge and belief contains no material previously published or written by another person, except where due reference is made in the text.

Signed 

Garth Bode

A C K N O W L E D G E M E N T S

I thank Phil Hughes of The Australian Bureau of Statistics for helpful discussions during my work on this project, and Mr. J.H.T. Morgan and Dr. T.J. O'Neill of the Department of Statistics, Australian National University, for their help and encouragement. I also thank Mrs. I. Scholtens for her excellent typing.

CHAPTER 1

INTRODUCTION

1.1 The Multipurpose Crop Acreage Survey

The objective of the crop acreage survey is the production of timely statistics relating to total areas sown for wheat, oats and barley in New South Wales. The sample survey enables estimates to be produced 12 months in advance of the actual totals resulting from the annual Agricultural Census. Data are collected on the following items for each crop:

(i) total area sown for all purposes;

(ii) area sown for grain;

(iii) area likely to be harvested for grain.

Estimates of totals for each of the nine variables are required for the state and for statistical divisions. The design relative standard errors to be attained are 2%, 3% and 4% for wheat, barley and oats respectively, for the variable "area sown for grain".

Questionnaires are despatched each year in August. The sampling frame is the list of holdings from the annual Agricultural Census taken on March 31 of the year in which the survey is run. The census also provides benchmark information for the survey and stratum population numbers. (See the next section for more details.)

1.2 Current Design for the Crop Survey

In surveys conducted in the past, a single stage stratified design has been used, with the sampling and

estimation methods used being dependent on the stratum type.

1.2.1 Stratification

As estimates are required for statistical divisions, any geographic stratification must be a refinement of statistical divisions, and in this case there were 27 geographic strata covering the 7 statistical divisions.

A further stratification was made on the basis of whether or not a unit (land holding) had useful quantitative supplementary information. (Quantitative supplementary information consisted of predicted values for the variable "area sown for all purposes" for each crop and, as mentioned previously, was obtained from the Agricultural Census. Thus a known 3-vector denoted by $\underset{\sim}{F}$ was associated with each unit of the population.) A unit was designated a "forecaster" if $\underset{\sim}{F} \neq \underset{\sim}{0}$, and a "non-forecaster" if $\underset{\sim}{F} = \underset{\sim}{0}$. Thus the population was partitioned into 54 strata.

1.2.2 Sample Selection

Independent samples were selected in each of the 54 strata. In the 27 non-forecaster strata, simple systematic sampling was used, whereas in the 27 forecaster strata, systematic PPS sampling was used. (See the Appendix for a description of these techniques.)

The size variable used for PPS sampling was a linear combination of the predicted areas

$$Z = \underset{\sim}{\alpha} \cdot \underset{\sim}{F} \quad ,$$

using an optimum $\underset{\sim}{\alpha}$. The method by which $\underset{\sim}{\alpha}$ was determined for past surveys is now described.

Data from two successive Agricultural Censuses were used, the first giving $\underset{\sim}{F}$ values for those units in the forecaster strata, and the second giving actual values for the variable "area of crop sown for grain" for the same units. Let

C_i = area of crop C sown for grain for the i^{th} forecaster unit.

Ignoring the finite population correction factor and the geographic stratification,

$$\text{var}\ (\hat{T}_c) \doteq \frac{1}{n_c} \sum_{1}^{N} K_i\ (\underset{\sim}{\alpha}) \left[C_i - R\ (FC)_i \right]^2 \quad ,$$

$$\text{where } R = \sum_{1}^{N} C_i \ / \sum_{1}^{N} (FC)_i \quad ,$$

$$K_i(\underset{\sim}{\alpha}) = \sum_{1}^{N} Z_i\ (\underset{\sim}{\alpha}) \ / \ Z_i\ (\underset{\sim}{\alpha}) \quad ,$$

$$\underset{\sim}{F} = (FW, FO, FB).$$

For each crop the desired value for var $(\hat{T}_c)$ was substituted and the above equation solved for $n_c\ (\underset{\sim}{\alpha})$. The optimum $\underset{\sim}{\alpha}$ was that vector giving the smallest valid sample size (i.e. minimum cost).

$$n^* = \min_{\underset{\sim}{\alpha}} \left[\max_{c} \left| n_c\ (\underset{\sim}{\alpha}) \right| \right] .$$

1.2.3 Estimation

Number raised estimation was used in the 27 non-forecaster strata (since $\underset{\sim}{F} = \underset{\sim}{Q}$ there) whereas ratio estimation was used in the 27 forecaster strata. Separate benchmark variables were used for each crop in the ratio estimation (i.e. FW was the benchmark for the 3 wheat variables, FB for the 3 barley variables and FO for the 3 oats variables).

Let T be the state total for one of the 3 variables (call it C) pertaining to crop C. The estimate of T used was

$$\hat{T} = \sum_{s=1}^{S} \sum_{h=1}^{H_s} (\hat{T}_{f,sh} + \hat{T}_{\bar{f},sh}) \quad ,$$

where S = number of statistical divisions,

H_s = number of geographic strata in statistical division s

and f, $\bar{f}$ refer to forecaster and non-forecaster units respectively.

$\hat{T}_{\bar{f},sh}$ is the number raised estimate for non-forecasters in geographic stratum h within statistical division s, and $\hat{T}_{f,sh}$ is the PPS ratio estimate for forecasters in geographic stratum h within statistical division s.

1.3 Proposals for Future Surveys

For various reasons it has been decided to discontinue PPS selection and to replace it with simple

systematic sampling. In future the quantitative supplementary information will be used for size stratification instead of PPS selection. Ratio estimation using separate benchmark variables will be retained.

1.4 Outline of Present Study

The main elements of this study are:

(i) An investigation into the theoretical aspects of multivariate stratification for multi-purpose surveys (Chapter 2);

(ii) An analysis of population data (Chapter 3);

(iii) A comparison of various size stratified designs and an assessment of their efficiencies relative to the PPS design (Chapter 4).

There is also an appendix containing a summary of some of the theory and methods of sample surveys relevant to this study.

1.5 Data Available for Analysis

Population data in the form of matched data from two successive Agricultural Censuses will be used. The first census gives the quantitative supplementary information $\underset{\sim}{F}$ (predicted values of "area sown for all purposes") for each holding, and the second gives the actual values for the variables "area sown for grain" for the three crops for corresponding holdings. Information is also available on the geographic location of each holding.

CHAPTER 2

STRATIFICATION WITH MANY VARIABLES IN A MULTIPURPOSE SURVEY

Multivariate stratification refers to the situation where strata are determined by the values of two or more variables, called stratification variables. The usual type is stratification by cross-classification, where M stratification variables, with the i^{th} variable having H_i intervals, form $H = \prod_{i=1}^{M} H_i$ strata. Another type of multivariate stratification uses clustering techniques to partition the population of units according to an appropriate clustering criterion. A third type uses a real valued function of the stratification variables, such as a linear combination, to effect a univariate stratification.

2.1 Multivariate Stratification

The basic design parameters to be determined when there are many variables available for stratification are:

(i) number of stratification variables (and choice of variables);

(ii) number of intervals from each variable (for stratification by cross-classification);

(iii) choice of stratum boundaries.

The decision on which available variables to use for stratification is partly determined by the correlations

among these variables and the estimation variables. Ideally we want stratification variables to have low correlations with each other, but for each to be strongly correlated with at least one estimation variable.

The number of stratification variables chosen and the population size help determine the number of intervals from each variable in the cross-classification case. There is generally more gain from the use of coarser divisions of several variables than from finer divisions of one.

An optimum choice of boundaries is affected by the method of allocation to be used, such as Neyman allocation or proportional allocation. It will be seen that the problem of optimum stratification by cross-classification is largely intractable, even for a simple method of allocation such as proportional allocation. Hence multivariate stratification methods tend to be ad hoc with emphasis on the sample allocation as a way of obtaining good gains.

2.2 A Recent Study

Kish and Anderson (1978) made a study of multivariate stratification for multipurpose surveys with the primary aim of quantifying the advantage of stratification by cross-classification over one-way stratification. In particular they compared a two-way cross-classification (H_1 by H_2) with a one-way stratification ($H = H_1H_2$ strata).

Here one-way refers to strata determined by the values of a single variable. The single variable could be one of the stratification variables or a combined variable such as a linear combination (for example the first principal component).

A 4-variate normal distribution was assumed for the two stratification variables and two estimation variables, and symmetry was imposed to reduce the number of correlation parameters to three. They used proportional allocation and number raised estimation, and stratum boundaries were chosen by applying the cum $\sqrt{f}$ rule to each variable. (This is an ad hoc procedure, but seems reasonable given the difficulty of computing the optimum points.) See A.5.3 of the appendix and 2.4 for more details.

Within their idealized theoretical framework they demonstrated that two-way stratification gave considerable gains over one-way with the same number of strata. Also stratification by the principal component performed poorly compared with two-way stratification.

2.3 Stratification by Clustering

Stratification by clustering is not suitable for the multipurpose crop survey with which this study is concerned, mainly because of the large population size. The technique is particularly suited to first-stage units in a multi-stage survey, which are not too numerous and

which have good supplementary information from many variables. Hence this method will not be considered further. (See Jarque (1981) for a solution to this problem.)

2.4 Stratification by Cross-classification

The problem of optimum stratification for a multipurpose survey was solved for the case of two stratifiers and two estimation variables by Ghosh (1963) who extended Dalenius's theory for optimum univariate stratification. He assumed infinite populations with continuous distributions and used simple random sampling with proportional allocation over a fixed number of strata to estimate the two population means. The measure of precision used was the generalized variance G, the determinant of the covariance matrix of the sample means.

Only cross-classification stratification was considered, and within this class he found the stratification points which minimized G. To find an explicit solution, a set of equations must be solved using bivariate iteration, with Dalenius's univariate solutions as starting points.

Ghosh's solution is not a practical one, and as yet no "rule of thumb" giving approximately optimum boundaries has been proposed. This explains the widespread use of ad hoc procedures.

2.5 Multipurpose Allocation

Consider a fixed multivariate stratification of the population. The simplest method for allocation of the sample to the strata is proportional allocation in which n_h is proportional to N_h. This method is particularly suitable if the stratum variances do not differ greatly and has the great advantage of requiring no information at the design stage other than the N_h values. However it is unresponsive to the demands of a multipurpose survey in which specified accuracies must be attained for each variable.

If the population consists of units of widely varying size (according to an appropriate measure of size), the stratum variances S_h^2 tend to be larger for strata with the larger units, making proportional allocation inefficient. For this type of population some sort of "optimal" allocation is highly desirable.

Suppose the N_h are known and that good estimates of the S_h^2 are available for all survey variables. Then for fixed total sample size n, an optimum (Neyman) allocation can be computed for each variable. (See A.5.1 of the appendix). The optimum allocation for one variable may be far from optimum for another, but if the estimation variables are positively correlated, the Neyman allocation for one will be reasonably efficient for the others. In this situation any one of the individual Neyman allocations may be appropriate, although a simple average of the

allocations would be a good compromise. In neither case is the method sensitive to the prescribed standard errors on the variables; the allocation may not achieve the desired precision for all variables, whereas another allocation with the same sample size may.

A compromise allocation which incorporates the desired precisions as constraints is the obvious solution. The "best" compromise allocation is that which minimizes the total cost subject to the required constraints on standard errors, and to the minimum sample size constraint $m \leq n_h \leq N_h$. For example, for K variables and H strata, and cost proportional to total sample size, the problem is to determine n_h, $h = 1,2,\ldots,H$ to minimize $\sum_1^H n_h$ subject to

$$\text{(i)} \quad \text{var}\,(\hat{T}_i) \leq \alpha_i, \qquad i = 1,\ldots,K \;;$$

$$\text{(ii)} \quad m \leq n_h \leq N_h \;, \qquad h = 1,\ldots,H \;.$$

This is a problem in non-linear programming, and there are algorithms for solving it, such as the one given by Kokan and Khan (1967). The complexity of this optimization problem necessitates the use of a computer programme, and this is its main disadvantage.

A different approach, based on the minimization of the generalized variance of the estimates was used by Ghosh (1958) to obtain an optimum multipurpose allocation. The explicit solution is complicated and must be obtained by iterative methods. The criterion of minimum generalized

variance does not take into account the required accuracy for each variable and as a result the solution is not very useful.

2.6 Completely Enumerated Strata

For a single purpose survey with one stratification variable, large gains in efficiency can be made by completely enumerating large units. (See A.5.2 of the appendix for more details.)

When there is multivariate stratification in a multipurpose survey, the situation is no longer clearcut. Firstly there are many possible criteria for determining the membership of the CE stratum, and secondly the gains will be largely influenced by the correlations among the estimation variables. In general, low correlations should lead to only moderate gains, whereas high correlations may produce substantial gains, since then units which are large for one variable will tend to be large in the others.

2.7 Post Stratification

Stratification may also be performed after the sample is selected, and this post stratification is particularly useful in multipurpose surveys since it enables a different stratification variable to be used for each estimation variable. Post stratification can also be used in tandem with ordinary stratification.

CHAPTER 3

ANALYSIS OF MATCHED CENSUS DATA

3.1 Non-forecasters

Of the total population of 24,108 holdings, 6,436 were non-forecasters; i.e. had predictions of zero for all three crops. Totals, means and variances of the actual area sown for grain for these units are given below.

	Wheat	Oats	Barley
Totals	167,427	13,793	23,519
Means	26.01	2.14	3.65
Variances	9,235.4	224.5	753.9

The quantities sown are quite substantial, and justify the inclusion of non-forecasters in the survey.

However, as these units are not able to be considered for size stratification, our attention will focus on the forecasters.

3.2 Forecasters

The data for the 17,672 forecasters were analysed using frequency tables, two-way tables of area sown versus area forecast, scatter plots and correlation tables.

3.2.1 Frequency Tables

Frequency tables for the benchmark and estimation variables for each crop are given in figures 3.1-3.6. The

relative strengths of the three crops are apparent with, for example, the average area sown for grain being 170.24, 13.40 and 24.20 for wheat, oats and barley respectively.

It is also obvious that a large proportion of oats is used for non-grain purposes. The ratios of area sown for grain to area forecast for all purposes are .93, .47 and .85 for wheat, oats and barley respectively.

The frequency tables also show that a CE stratum could be established. The distributions are highly positively skewed with, for example, the variable "wheat sown for grain" having just 36 out of 17,672 observations (.074%) spread over the upper 75% of its range.

Although all forecaster holdings had, by definition, a non-zero forecast for at least one crop, it is interesting to note that the proportions of holdings with zero forecasts were wheat 11.87%, oats 41.23% and barley 57.66%. The high proportion of zeroes for oats and barley creates problems for design and estimation which would not exist for a univariate survey. In a univariate (single purpose) survey all units with zero forecasts could be placed in a separate stratum.

FREQUENCY TABLES FOR WHEAT

(i) Wheat forecast for all purposes (benchmark variable)

HECTARES	FREQUENCY	CUM. PERCENT
Zero	2097	11.87
0 - 50	2733	27.33
50 - 100	2887	43.67
100 - 200	4382	68.46
200 - 300	2491	82.56
300 - 400	1044	88.47
400 - 500	795	92.97
500 - 1000	966	98.43
1000 - 2000	229	99.73
2000 - 3000	30	99.90
3000 - 4000	7	99.94
4000 - 5000	4	99.96
5000 - 6000	3	99.98
6000 - 7000	3	99.99
7000 - 8000	0	99.99
8000 - 9000	1	100.00

Total Area	=	3,241,233
Number of Holdings	=	17,672
Mean Per Holding	=	183.41
Variance	=	73,805.34
Coefficient of Variation	=	148.12

Figure 3.1

(ii) Wheat sown for grain (estimation variable)

HECTARES	FREQUENCY	CUM. PERCENT
Zero	2648	14.98
0 - 50	2757	30.59
50 - 100	2889	46.93
100 - 200	4251	70.99
200 - 300	2303	84.02
300 - 400	1005	89.71
400 - 500	713	93.74
500 - 1000	884	98.74
1000 - 2000	186	99.80
2000 - 3000	23	99.93
3000 - 4000	4	99.95
4000 - 5000	2	99.96
5000 - 6000	3	99.98
6000 - 7000	2	99.99
7000 - 8000	1	99.99
8000 - 9000	1	100.00

Total Area	=	3,008,443
Number of Holdings	=	17,672
Mean per Holding	=	170.24
Variance	=	65,903.89
Coefficient of Variation	=	150.80

Figure 3.2

FREQUENCY TABLES FOR OATS

(i) Oats forecast for all purposes (benchmark variable)

HECTARES	FREQUENCY	CUM. PERCENT
Zero	7286	41.23
0 - 50	6954	80.58
50 - 100	2228	93.19
100 - 200	997	98.83
200 - 300	141	99.63
300 - 400	41	99.86
400 - 500	14	99.94
500 - 1000	10	99.99
1000 - 2000	1	100.00

Total Area	=	500,328
Number of Holdings	=	17,672
Mean per Holding	=	28.31
Variance	=	2,055.83
Coefficient of Variation	=	160.15

Figure 3.3

(ii) Oats sown for grain (estimation variable)

HECTARES	FREQUENCY	CUM. PERCENT
Zero	11338	64.16
0 - 50	4875	91.74
50 - 100	1053	97.70
100 - 200	345	99.66
200 - 300	47	99.92
300 - 400	7	99.96
400 - 500	5	99.99
500 - 1000	2	100.00
1000 - 2000	0	100.00

Total Area	=	236,854
Number of Holdings	=	17,672
Mean per Holding	=	13.40
Variance		879.46
Coefficient of Variation	=	221.27

Figure 3.4

FREQUENCY TABLES FOR BARLEY

(i) Barley forecast for all purposes (benchmark variable)

HECTARES	FREQUENCY	CUM. PERCENT
Zero	10189	57.66
0 - 50	4042	80.53
50 - 100	1952	91.57
100 - 200	1080	97.69
200 - 300	263	99.17
300 - 400	72	99.58
400 - 500	45	99.84
500 - 1000	23	99.97
1000 - 2000	6	100.00

Total Area	=	500,577
Number of Holdings	=	17,672
Mean per Holding	=	28.33
Variance	=	3,727.64
Coefficient of Variation	=	215.54

Figure 3.5

(ii) Barley sown for grain (estimation variable)

HECTARES	FREQUENCY	CUM. PERCENT
Zero	11056	62.56
0 - 50	3718	83.60
50 - 100	1681	93.11
100 - 200	913	98.28
200 - 300	200	99.41
300 - 400	42	99.65
400 - 500	32	99.83
500 - 1000	24	99.97
1000 - 2000	6	100.00

Total Area	=	427,671
Number of Holdings	=	17,672
Mean per Holding	=	24.20
Variance	=	3,261.94
Coefficient of Variation	=	236.00

Figure 3.6

3.2.2 Correlations

The 15 correlations among the 6 variables are shown in the following table (Figure 3.7):

CORRELATIONS

		AREA FORECAST			AREA SOWN		
		Wheat	Oats	Barley	Wheat	Oats	Barley
AREA FORECAST	Wheat	1.000	0.124	0.168	0.919	0.070	0.187
	Oats		1.000	0.071	0.116	0.551	0.081
	Barley			1.000	0.189	0.081	0.810
AREA SOWN	Wheat				1.000	0.080	0.203
	Oats					1.000	0.087
	Barley						1.000

Figure 3.7

The correlation between area forecast (for all purposes) and area sown (for grain) is high for wheat and barley (.919 and .810 respectively) but only moderate for oats (.551). This is to be expected from the results presented in 3.2.1, from which it was concluded that nearly half the oats was sown for non-grain purposes.

The correlations among the three benchmark variables and among the three estimation variables are low, and this lends weight to the argument for including all three benchmarks for size stratification.

Scatter plots were obtained to illustrate the relationships among the variables, and two examples are shown in Figures 3.8 and 3.9. The latter shows barley forecast versus oats forecast, and the former shows wheat sown versus oats sown, both for statistical division 5. There is a noticeable tendency for holdings to specialise.

3.2.3 Area Sown Versus Area Forecast

Figures 3.10 and 3.11 give two-way frequency tables of area sown for grain and area forecast for all purposes, for each crop. Also some scatter plots for statistical divisions 3 and 5 are given in Figures 3.12 - 3.14. (Plots for the other statistical divisions showed similar trends.) A strong linear relationship is evident, particularly for wheat, but there are many points along the axes, especially for oats and barley, which could create problems for estimation. Points on the horizontal axis tend to reflect the diverting of crops to non-grain purposes, and perhaps some optimistic forecasting. It is believed that those on the vertical axis arise mainly from non-response in the Agricultural Census (which provides the forecasts of areas sown), and the subsequent imputation of a zero forecast.

WHEAT SOWN VERSUS OATS SOWN

STATISTICAL DIVISION 5

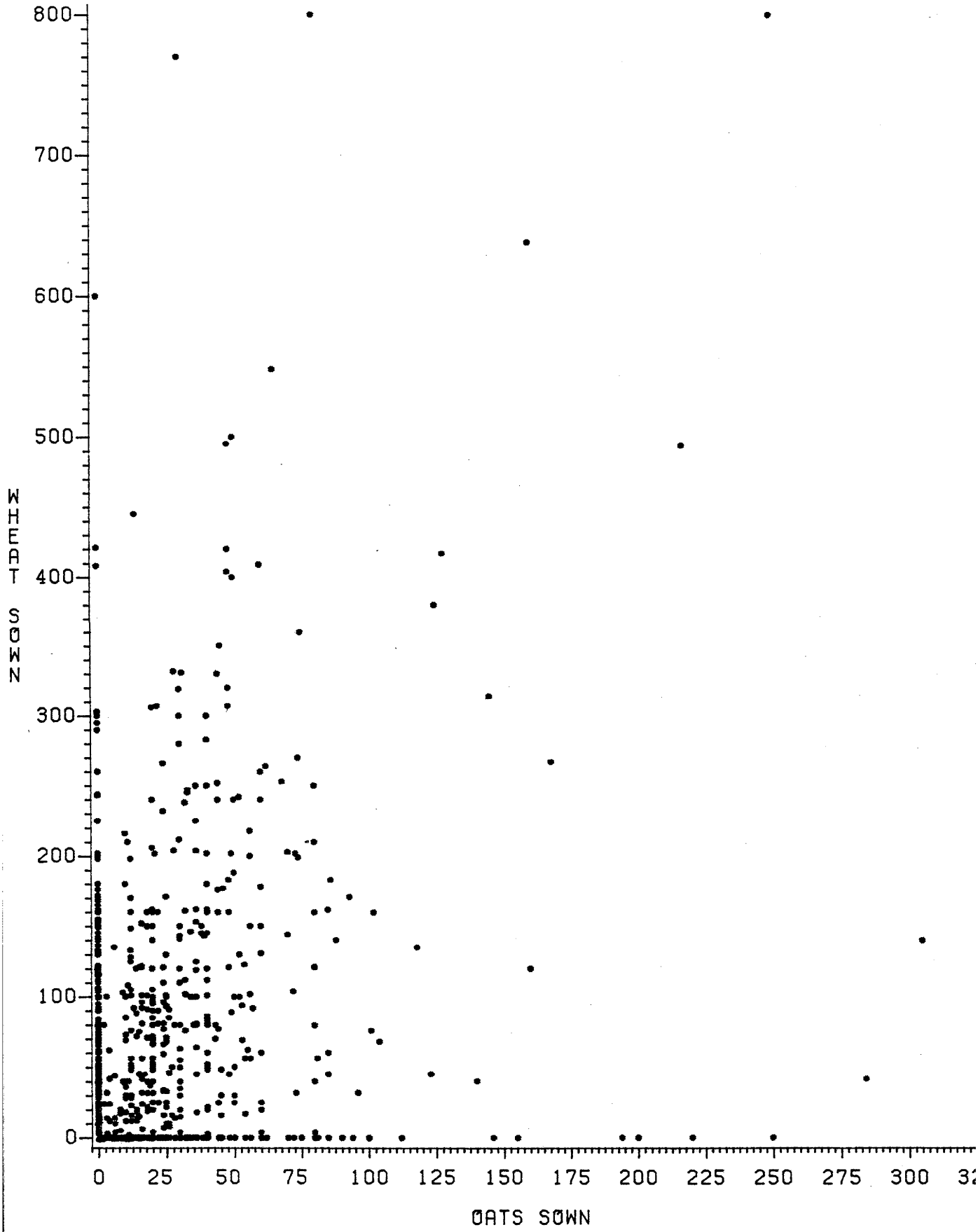


Figure 3.8

BARLEY FORECAST VERSUS OATS FORECAST

STATISTICAL DIVISION 5

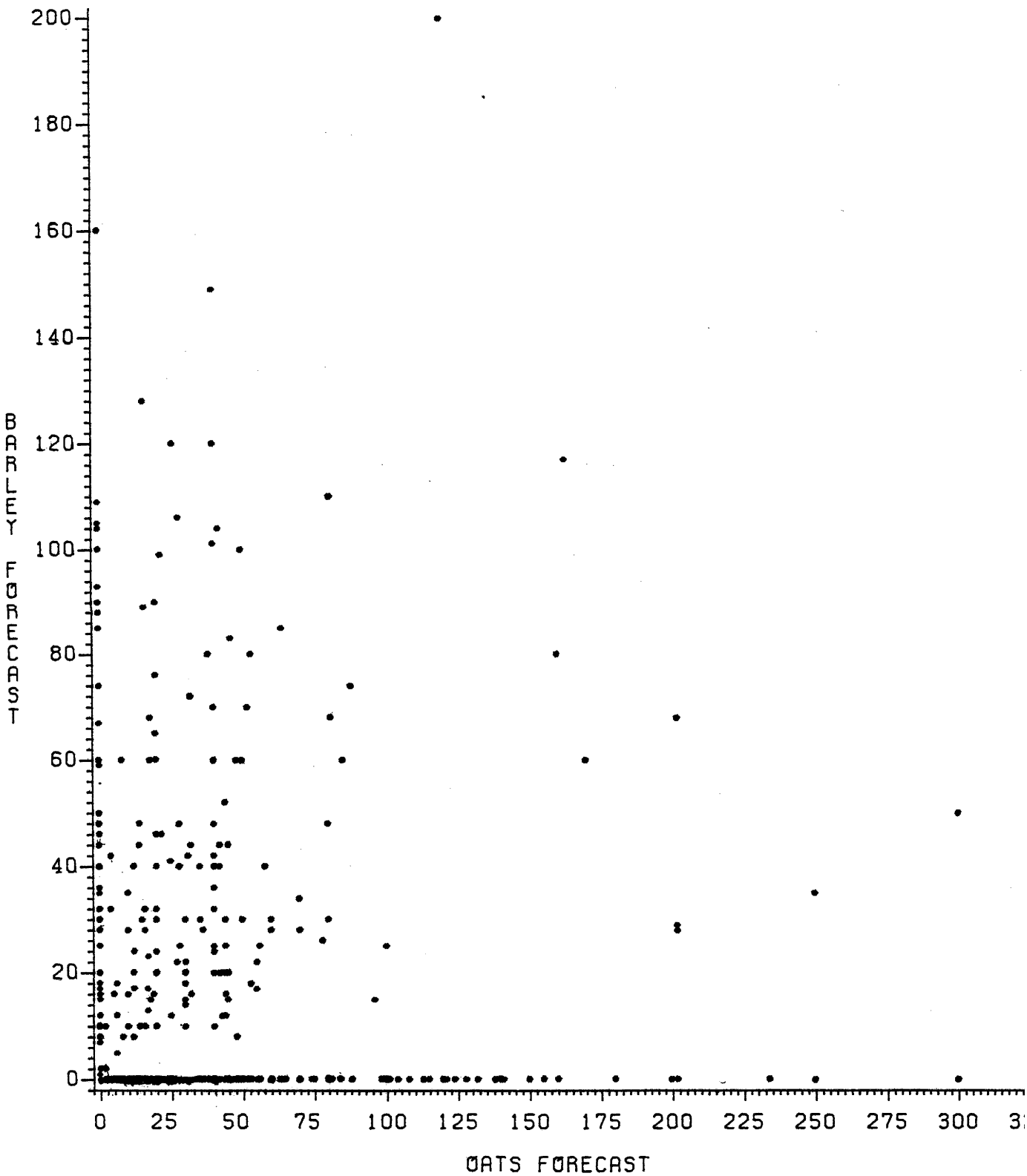


Figure 3.9

WHEAT SOWN VERSUS WHEAT FORECAST

WHEAT SOWN (FOR GRAIN)

wheat forecast	Zero	0-50	50-100	100-200	200-300	300-400	400-500	500-1000	1000-2000	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000	7000-8000	8000-9000	Total
Zero	1846	163	43	23	12	3	3	4									2097
0-50	417	1960	286	51	13	2	2	2									2733
50-100	161	495	1859	327	34	8	0	3									2887
100-200	122	115	618	3176	296	29	17	8	1								4382
200-300	52	17	59	559	1533	207	49	11	4								2491
300-400	14	3	16	70	291	516	103	30	1								1044
400-500	17	0	5	26	85	181	387	88	5	1							795
500-1000	15	4	2	17	33	55	142	661	37	0							966
1000-2000	2		1	1	6	3	10	73	129	4							229
2000-3000	2			1		0		4	8	14	1						30
3000-4000						0			0	4	3						7
4000-5000						1			1			2					4
5000-6000													3				3
6000-7000														2	1		3
7000-8000															0		0
8000-9000																1	1
TOTAL	2648	2757	2889	4251	2303	1005	713	884	186	23	4	2	3	2	1	1	17672

Figure 3.10

OATS SOWN VERSUS OATS FORECAST

OATS SOWN (FOR GRAIN)

OATS FORECAST	Zero	0-50	50-100	100-200	200-300	300-400	400-500	500-1000	1000-2000	TOTAL
Zero	6829	389	42	21	5					7286
0-50	3318	3451	174	11	0					6954
50-100	791	742	641	51	3					2228
100-200	327	257	170	224	16	3				997
200-300	48	24	17	30	19	2	1			141
300-400	12	8	8	6	3	2	2			41
400-500	7	3	1	2	1		0			14
500-1000	6	1					1	2		10
1000-2000							1			1
TOTAL	11338	4875	1053	345	47	7	5	2	0	17672

BARLEY SOWN VERSUS BARLEY FORECAST

BARLEY SOWN (FOR GRAIN)

BARLEY FORECAST	Zero	0-50	50-100	100-200	200-300	300-400	400-500	500-1000	1000 2000	TOTAL
Zero	9449	555	113	52	13	1	3	2	1	10189
0-50	1192	2544	264	35	6	1	0	0	0	4042
50-100	265	523	999	158	5	2	0	0	0	1952
100-200	118	82	279	545	49	5	2	0	0	1080
200-300	20	14	20	97	96	10	4	2	0	263
300-400	8		5	15	22	15	2	5	0	72
400-500	2		1	8	6	6	18	4	0	45
500-1000	2			2	3	2	2	11	1	23
1000-2000				1			1		4	6
TOTAL	11056	3718	1681	913	200	42	32	24	6	17672

Figure 3.11

WHEAT SOWN VERSUS WHEAT FORECAST

STATISTICAL DIVISION 5

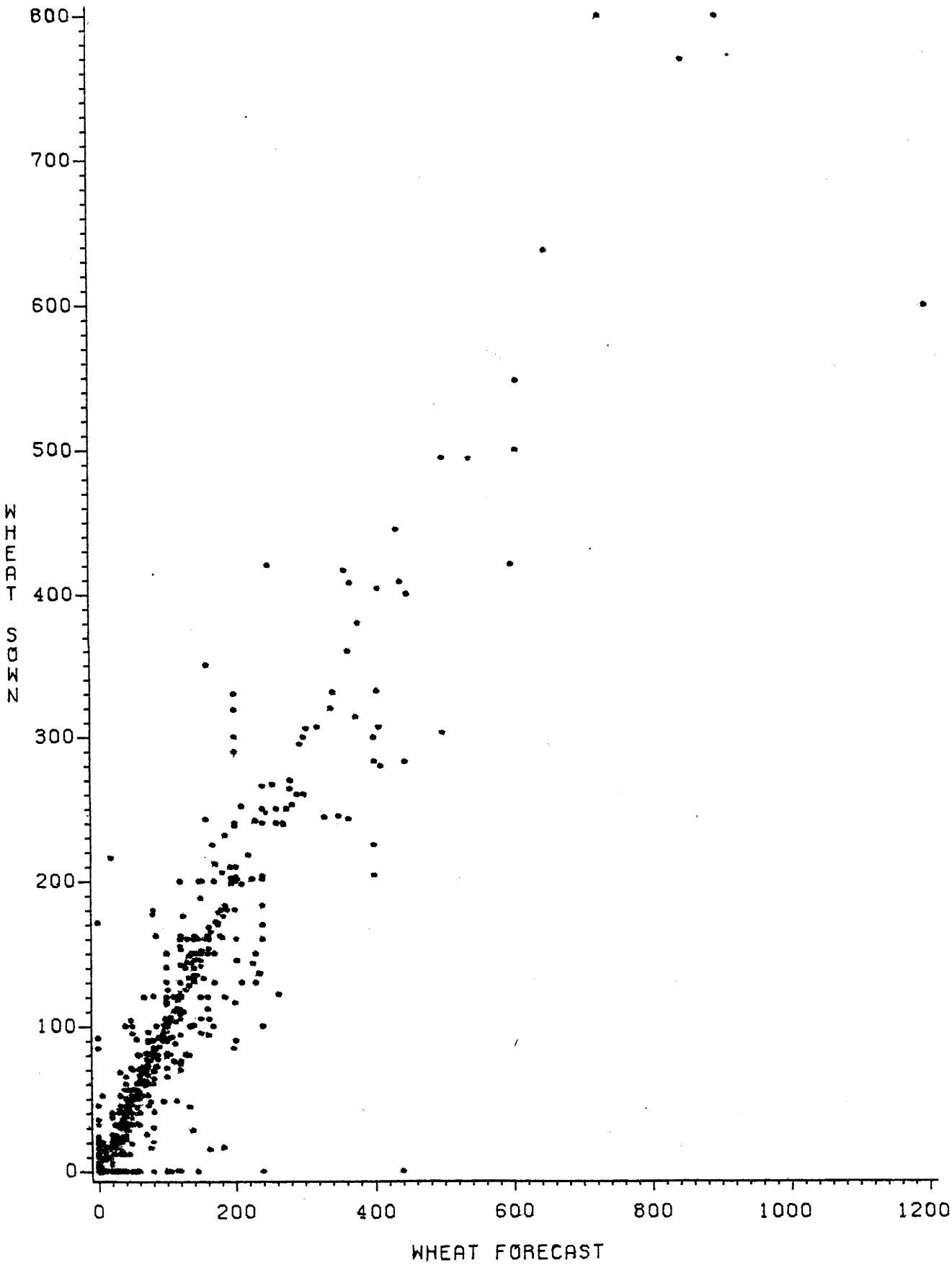


Figure 3.12

OATS SOWN VERSUS OATS FORECAST

STATISTICAL DIVISION 5

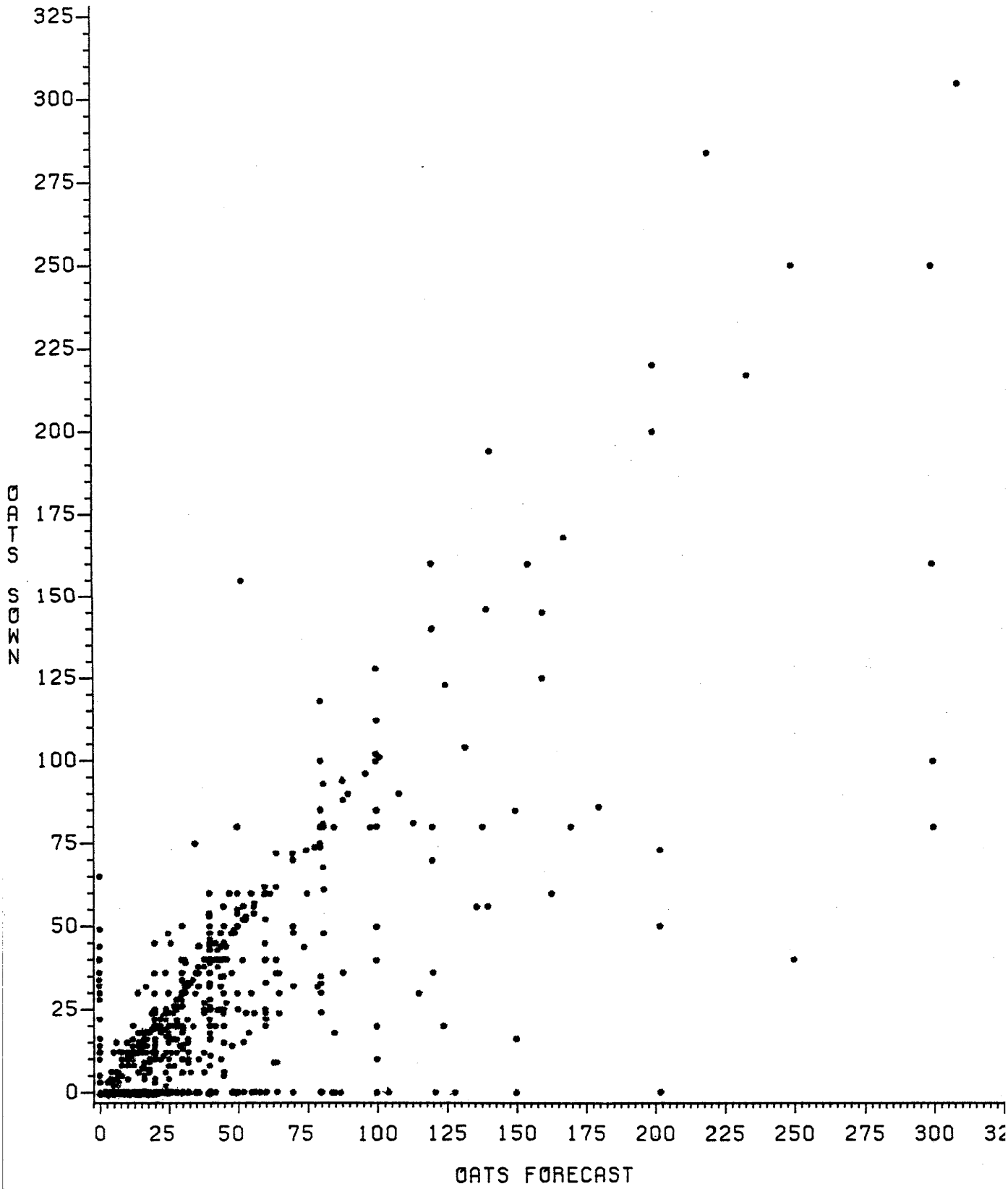


Figure 3.13

BARLEY SOWN VERSUS BARLEY FORECAST

STATISTICAL DIVISION 3

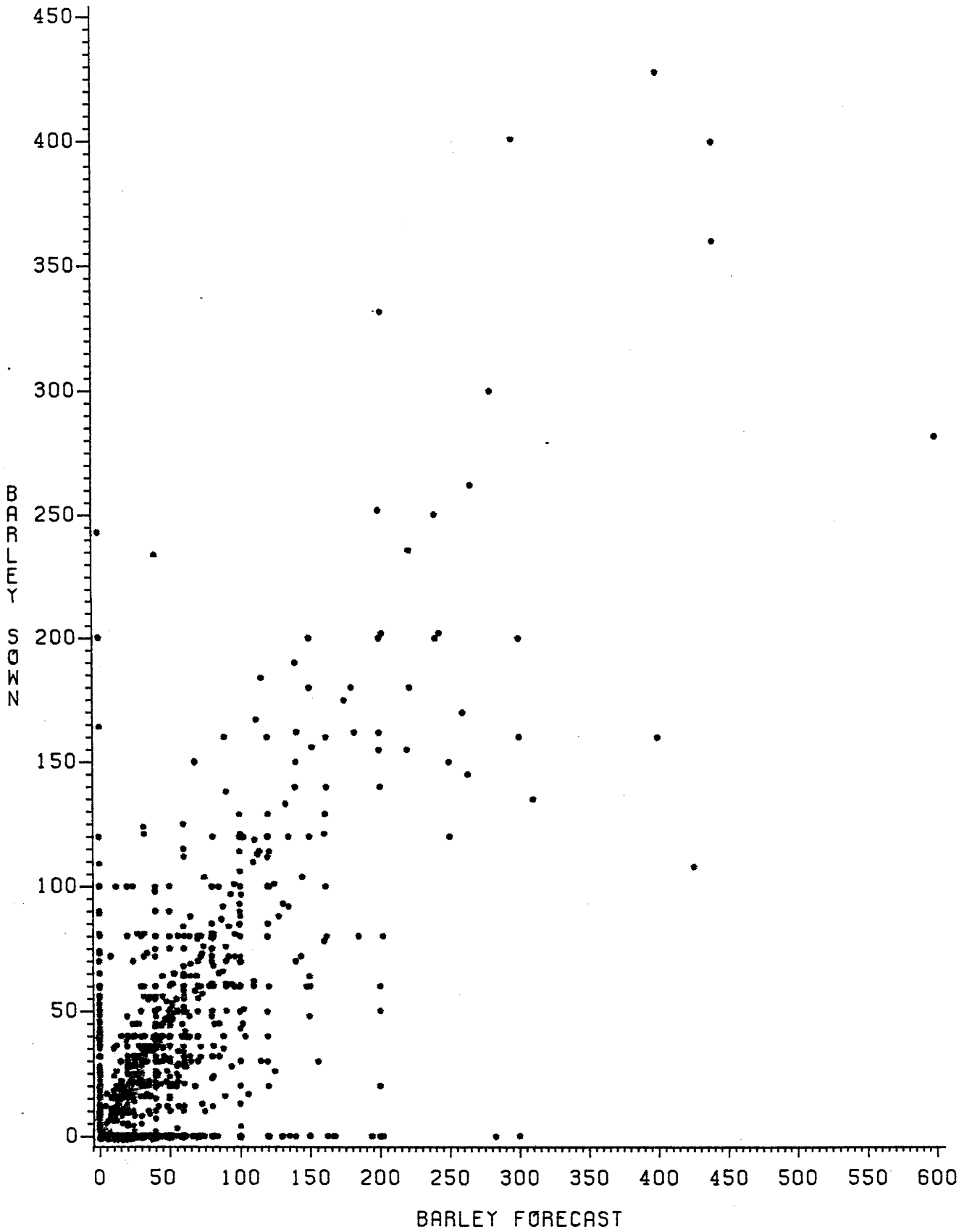


Figure 3.14

3.2.4 Differences among Statistical Divisions

The following table (Figure 3.15) summarises the differences in sizes and levels among the 7 statistical divisions:

	STATISTICAL DIVISION						
	1	2	3	4	5	6	7
Number of Holdings	261	3629	3259	3235	912	3660	2716
Wheat forecast per unit	154	249	201	209	73	155	121
Wheat sown per unit	141	230	182	198	67	147	113
ρ(wheat sown, wheat forecast)	.96	.95	.90	.86	.93	.86	.86
Oats forecast per unit	32	25	33	27	33	30	25
Oats sown per unit	7	4	13	15	21	19	15
ρ(oats sown, oats forecast)	.46	.33	.58	.54	.75	.65	.66
Barley forecast per unit	11	26	17	31	10	33	43
Barley sown per unit	10	23	13	26	9	29	38
ρ(barley sown, barley forecast	.75	.82	.77	.78	.80	.82	.81

Figure 3.15

From the table it is seen that statistical divisions 1 and 5 are much smaller than the others, with statistical division 1 being too small to support a size stratification involving all three benchmark variables.

There is also considerable variation in level among statistical divisions which is not uniform in the crops. This is probably due to a combination of particular areas favouring certain crops, and some areas having large holdings.

Correlations are consistently strong for wheat and barley but are quite variable for oats.

3.2.5 Implications for Estimation Methods

From A.2.2 of the appendix, it is seen that, to a first approximation, ratio estimation is more efficient than number raised estimation if $\rho_{xy} > \frac{1}{2}C_y/C_x$, where Y is the benchmark variable, X is the estimation variable and C represents the coefficient of variation. The figures for these quantities, for the state as a whole, are shown below.

	ρ_{xy}	$\frac{1}{2}C_y/C_x$
Wheat	.919	.491
Oats	.551	.362
Barley	.810	.457

Hence based on the above criterion, ratio estimation should be more efficient than number raised estimation, providing the inequality still holds after stratification.

Because of the high proportion of holdings with a benchmark of zero for the crops oats and barley (41.23% for oats and 57.66% for barley), ratio estimation may not be as effective as is indicated by correlations. In some strata the chance of a bad sample (such as all or almost all sample units having a zero benchmark) may be high. Also points along the axes in figures 3.12 - 3.14 can cause problems for ratio estimation. These issues will be examined in Chapter 4.

CHAPTER 4

SIZE STRATIFIED DESIGNS FOR THE MULTIPURPOSE CROP SURVEY

In this chapter, a description will be given of some possible size stratified designs applied to the population data described in Chapter 3. An assessment of their performances will be given, and some recommendations made. A problem which is specific to multipurpose surveys will be investigated, and suggestions made of ways to overcome it.

4.1 Stratification by Cross-classification for the Crop Survey

As a result of the findings reported in Chapter 3, it was decided to use all three benchmark variables for size stratification. The parameters needed to define the stratification are:

(i) number of intervals from each stratification variable;

(ii) boundary points for intervals;

(iii) CE cut-off points if CE strata are to be formed.

The findings in Chapter 3 also indicate that the stratification should not be uniform across statistical divisions. For example, statistical division 1 is too small to support as fine a stratification as the others, and statistical division 5 will require separate consideration

with respect to the location of boundary points because of its relatively low wheat figures.

4.1.1 Stratum Boundary Points

In all statistical divisions except the first it was decided to use two intervals from each variable giving $2^3 = 8$ sampled strata. In statistical division 1 only wheat was split giving 2 sampled strata. As mentioned in Chapter 2, there is no useful optimality theory for determining stratum boundary points and so it must be done subjectively. The "objective" approach of using the Dalenius-Hodges method for each stratification variable is ad hoc. It was tried for this population but did not produce a workable stratification. It is important to ensure that none of the strata has too few units, and this objective was not attained with the Dalenius-Hodges method due mainly to the large number of units with small or zero values. The main reasons for ensuring that strata are not too small are:

(i) there is a minimum sample size requirement of 8 for each stratum, causing very small strata to have excessive sampling rates;

(ii) in a stratum where ratio estimation is to be used, a sample of considerably more than 8 units is desirable, and this will only occur if the stratum population size is fairly large.

4.1.2 CE Strata

The frequency analysis of Chapter 3 suggests the use of CE strata, but as pointed out in Chapter 2, the anticipated gains may not materialise in the multivariate case. It was stated in Chapter 3 that the estimation variables (area sown for grain for wheat, oats and barley) are uncorrelated and so we would expect only modest gains from CE strata.

It was decided to form a CE stratum in each statistical division, with a unit being placed in a CE stratum if its benchmark value was "large" for at least one variable. The cut-off points would be determined subjectively.

4.1.3 Final Cross-classified Stratification

For each of numerous stratifications tried, stratum numbers, stratum variances for each crop and stratum correlations for each crop were computed, and a sample allocated to the strata three times using Neyman allocation on each variable. Stratum boundary points were varied around the initial subjective values and it was found that the variances of the estimates of state totals were not very sensitive to the position of the boundary points. It was also found that the use of small CE strata lowered variances, but the trend soon reversed when the cut-offs were lowered further. Thus the CE cut-offs had to be set surprisingly high, giving very

small CE strata. (The modest size of the gain from complete enumeration may not be enough to justify its use, as there are practical estimation problems when using CE strata caused by such ever-present phenomena as non-response.)

After much trial and error the following stratification was produced.

Statistical Division 1

Stratum Description	Stratum Number
$0 \leq FW < 150$ and $0 \leq FO < 200$ and $0 \leq FB < 200$	1
$150 \leq FW < 500$ and $0 \leq FO < 200$ and $0 \leq FB < 200$	2
$FW \geq 500$ or $FO \geq 200$ or $FB \geq 200$	3 (CE)

Statistical Division 5

Stratum Description	Stratum Number
$0 \leq FW < 100$ and $0 \leq FO < 30$ and $0 \leq FB < 40$	1
$0 \leq FW < 100$ and $30 \leq FO < 200$ and $0 \leq FB < 40$	2
$0 \leq FW < 100$ and $0 \leq FO < 30$ and $40 \leq FB < 200$	3
$0 \leq FW < 100$ and $30 \leq FO < 200$ and $40 \leq FB < 200$	4
$100 \leq FW < 500$ and $0 \leq FO < 30$ and $0 \leq FB < 40$	5
$100 \leq FW < 500$ and $30 \leq FO < 200$ and $0 \leq FB < 40$	6
$100 \leq FW < 500$ and $0 \leq FO < 30$ and $40 \leq FB < 200$	7
$100 \leq FW < 500$ and $30 \leq FO < 200$ and $40 \leq FB < 200$	8
$FW \geq 500$ or $FO \geq 200$ or $FB \geq 200$	9 (CE)

Statistical Divisions 2,3,4,6,7

Stratum Description	Stratum Number
$0 \leq FW < 250$ and $0 \leq FO < 40$ and $0 \leq FB < 50$	1
$0 \leq FW < 250$ and $40 \leq FO < 350$ and $0 \leq FB < 50$	2
$0 \leq FW < 250$ and $0 \leq FO < 40$ and $50 \leq FB < 450$	3
$0 \leq FW < 250$ and $40 \leq FO < 350$ and $50 \leq FB < 450$	4
$250 \leq FW < 2000$ and $0 \leq FO < 40$ and $0 \leq FB < 50$	5
$250 \leq FW < 2000$ and $40 \leq FO < 350$ and $0 \leq FB < 50$	6
$250 \leq FW < 2000$ and $0 \leq FO < 40$ and $50 \leq FB < 450$	7
$250 \leq FW < 2000$ and $40 \leq FO < 350$ and $50 \leq FB < 450$	8
$FW \geq 2000$ or $FO \geq 350$ or $FB \geq 450$	9 (CE)

Figure 4.1

The total number of units in all CE strata is 150, which is only about 7% of the sample that is allocated to the 17,672 forecaster units. However, many of these 150 units have a zero forecaster in at least one variable. The actual figures are as follows:

	FW = 0	FO = 0	FB = 0
No. of CE units	25	61	84
Percentage	$16\frac{2}{3}$ %	$40\frac{2}{3}$ %	56%

This is a good illustration of the special problems encountered with CE strata in multipurpose surveys. For example, over half the 150 CE units in the above design are wasted from the point of view of barley. It is not surprising that the gains are modest, and disappear quickly when the number of CE units is increased.

Figure 4.2 is a table of stratum sizes and correlations between area forecast and area sown for the sampled strata. Comparison with Chapter 3 shows that the correlations have decreased considerably with the increased stratification. This was to be expected.

Statistical Division 1

Stratum	N_h	$\rho(W)$	$\rho(O)$	$\rho(B)$
1	164	.64	.35	.68
2	85	.56	.43	.68

Statistical Division 2

Stratum	N_h	$\rho(W)$	$\rho(O)$	$\rho(B)$
1	1652	.80	.19	.56
2	464	.82	.25	.54
3	292	.54	.16	.61
4	103	.49	.08	.68
5	656	.82	.17	.08
6	205	.76	.12	.09
7	140	.80	.30	.65
8	70	.86	.17	.52

(continued)

Statistical Division 3

Stratum	N_h	ρ(W)	ρ(O)	ρ(B)
1	1427	.73	.39	.47
2	668	.80	.30	.52
3	194	.76	.19	.67
4	99	.83	.17	.64
5	505	.73	.15	.30
6	231	.86	.39	.55
7	52	.74	.34	.53
8	66	.81	.43	.72

Statistical Division 4

Stratum	N_h	ρ(W)	ρ(O)	ρ(B)
1	1574	.83	.53	.65
2	374	.72	.30	.66
3	281	.77	.23	.57
4	104	.71	.54	.67
5	345	.78	.22	.30
6	223	.84	.40	.52
7	167	.77	.28	.71
8	152	.84	.49	.68

Statistical Division 5

Stratum	N_h	ρ(W)	ρ(O)	ρ(B)
1	404	.89	.53	.55
2	191	.78	.54	.73
3	29	.66	.46	.62
4	17	.77	.61	.31
5	84	.60	.22	.59
6	114	.83	.47	.70
7	23	.67	.72	.38
8	26	.68	.64	.73

Statistical Division 6

Stratum	N_h	ρ(W)	ρ(O)	ρ(B)
1	1890	.84	.52	.65
2	599	.82	.44	.59
3	372	.73	.38	.67
4	160	.63	.35	.67
5	179	.65	.23	.21
6	147	.58	.47	.19
7	131	.65	.35	.61
8	167	.82	.52	.66

Statistical Division 7

Stratum	N_h	ρ(W)	ρ(O)	ρ(B)
1	1437	.78	.51	.54
2	395	.86	.67	.47
3	431	.73	.66	.58
4	135	.64	.42	.74
5	83	.81	.43	.07
6	21	.77	.07	.25
7	104	.75	.17	.67
8	90	.82	.52	.47

Figure 4.2

4.1.4 Estimation Methods

Traditionally, correlations have been used as the main criteria for determining the method of estimation to be used (e.g. ratio, number raised). The stratum population variances are given by

$$S^2_R \doteq (N/n)(N-n)(N-1)^{-1} \sum_1^N (X_i - RY_i)^2 \quad \text{for ratio estimation,}$$

and $S^2 = (N/n)(N-n)(N-1)^{-1} \sum_1^N (X_i - \bar{X})^2$ for number raised estimation, where the summation is over the units in a stratum. As the stratum variances are used for determining optimum allocations and producing design standard errors (at the planning stage), it is important that the formulae give the correct values.

The correlation between two variables is not an adequate summary of their bivariate structure and the data must be examined closely before deciding on the estimators to be used and their variance formulae. A look at the population data presented in frequency tables and graphs in Chapter 3 shows that a considerable proportion of units have zero benchmarks and/or zero grain sown for at least one crop. This could be disastrous for ratio estimation, even though the correlation may satisfy $\rho_{xy} > \frac{1}{2} C_y/C_x$. Size stratification would result in the problem of zero benchmarks being concentrated in particular size strata for each variable.

4.2 The Problem of Zero Benchmarks in a Multipurpose Survey

As stated in the appendix, an estimator $\hat{T}$ of $T = \sum_{1}^{N} X_i$ is a real valued function on Ω. The sample space Ω is the set of all data points $\omega = (s, X_i : i \in s)$, $s \in S$, $\theta = (X_1, \ldots, X_N) \in \Theta$. For simple random samples of size n without replacement, S consists of the ${}^{N}C_{n}$ subsets of size n from the set $\{1,2,\ldots,N\}$. The parameter space Θ is a subset of R^N.

Since $\hat{T}$ must map almost all $\omega \in \Omega$, the class of possible estimators depends on Ω and p, the randomization distribution on S.

The ratio estimator is given by

$$\hat{T}_R = \left(\sum_{i \in s} X_i \Big/ \sum_{i \in s} Y_i \right) \sum_{1}^{N} Y_i \quad ,$$

where Y is the benchmark variable and X is the estimation variable. Clearly $\hat{T}_R$ cannot be defined on Ω when there exists at least one sample $s \in S$ such that $p(s) > 0$ and $Y_i = 0$ for all $i \in s$. In other words the ratio estimator cannot theoretically be used for a variable when it is possible to select a sample with all benchmarks equal to zero for that variable. The problem of zero benchmarks would only occur in conjunction with a multipurpose survey, since they could easily be stratified out in a single purpose survey.

For simple random samples of size n, there only needs to be n zero benchmarks in the population to

invalidate the ratio estimator, but this should not necessarily prevent its being used if the estimation and benchmark variables are "roughly proportional". One practical, but theoretically unsound policy would be to use ratio estimation if possible, and number raised estimation if the sample selected was a zero-benchmark sample. (A zero-benchmark sample is defined to be one in which all benchmark values for one of the estimation variables are zero.) It would be difficult to obtain an expression for the variance of this "adaptive" estimator. Its use is not recommended.

In practice, the problem might be ignored if the probability of selecting a zero-benchmark sample is very small, but the usual formula for the variance of $\hat{T}_R$ may no longer be approximately true. With a large proportion of zero benchmarks in the population, the probability of selecting a zero-benchmark sample depends on the population size and sampling fraction. For example, with N = 21, n = 8 and 18 out of 21 units (86%) having zero benchmarks, the probability of a zero-benchmark sample is .215. With N = 205, n = 53 and 181 out of 205 units (88%) having zero benchmarks, the probability is .0005.

The presence of a substantial number of zero benchmarks in the population will also produce poor estimates from ratio estimation when samples contain many but not all zero benchmarks for which the corresponding value of the estimation variable is positive. This could result in

over estimates of population totals, and the large sample variance formula $\text{var}(\hat{T}_R) \doteq (N/n)(N-n)(N-1)^{-1} \sum_{1}^{N} (X_i - RY_i)^2$ may give a value considerably less than the true mean square error.

Note that the above variance formula can always be evaluated irrespective of the X and Y values, even though it is meaningless when there is a large number of zero benchmarks, and the ratio estimator is not defined.

4.2.1 Solutions to the Problem of Zero Benchmarks

The severity of the problem and the quality of the non-zero benchmark values should determine the appropriate solution.

The simplest solution is to use number raised estimation for the affected variables in the appropriate strata. This simple solution (call it the "N-R" solution) is appropriate if the non-zero benchmarks are not very useful for ratio estimation. From a practical point of view it is very effective, but it overlooks the potential gains from using the benchmark information in ratio estimation, if there are gains to be made.

If the non-zero benchmarks are good, post stratification would be the appropriate measure to adopt. (See A.5.4 of the appendix for details on post stratification.) Number raised estimation would be used in the post stratum with zero benchmarks for the variable in question, and ratio estimation in the other. (Call this solution the "P-S" solution.)

This illustrates the advantage of post stratification over stratification before selection, in a multipurpose survey. Stratification variables can be chosen in different ways for different estimation variables, thus leading to a more efficient design.

4.2.2 Frequency of Zero Benchmarks in the Crop Data

Frequencies of zero benchmarks were determined for all size strata in all statistical divisions for the multivariate stratified design described in 4.1.3. Some of the results are presented in Figure 4.3 below. ($P_0(W)$ represents the proportion of units having a zero benchmark for wheat, etc.) CE strata are omitted.

Statistical Division	Size Stratum	N_h	$P_0(W)$ %	$P_0(O)$ %	$P_0(B)$ %
1	1	164	28	33	82
	2	85	0	55	81
2	1	1652	14	59	70
	2	464	25	0	75
	3	292	20	72	0
	4	103	13	0	0
	5	656	0	87	90
	6	205	0	0	88
	7	140	0	87	0
	8	70	0	0	0
3	1	1427	13	51	79
	2	668	19	0	76
	3	194	21	58	0
	4	99	14	0	0
	5	505	0	84	94
	6	231	0	0	82
	7	52	0	77	0
	8	66	0	0	0

(continued)

Statistical Division	Size Stratum	N_h	P_0(W) %	P_0(O) %	P_0(B) %
4	1	1574	7	52	64
	2	374	9	0	69
	3	281	13	64	0
	4	104	8	0	0
	5	345	0	70	78
	6	223	0	0	74
	7	167	0	72	0
	8	152	0	0	0
5	1	404	49	25	83
	2	191	59	0	88
	3	29	34	55	0
	4	17	41	0	0
	5	84	0	45	79
	6	114	0	0	79
	7	23	0	48	0
	8	26	0	0	0
6	1	1890	11	44	60
	2	599	17	0	63
	3	372	9	52	0
	4	160	4	0	0
	5	179	0	75	80
	6	147	0	0	67
	7	131	0	60	0
	8	167	0	0	0
7	1	1437	15	53	57
	2	395	29	0	73
	3	431	10	60	0
	4	135	6	0	0
	5	83	0	82	81
	6	21	0	0	86
	7	104	0	72	0
	8	90	0	0	0

Figure 4.3

From Figure 4.3 we see that there are very few problems with wheat, with the possible exception of statistical division 5 where between one-half and one-third of the units in size strata 1, 2, 3 and 4 had zero benchmarks.

For barley and oats the proportions of zero benchmarks are very high in size strata 1, 3, 5 and 7 for oats, and 1, 2, 5 and 6 for barley. From Figure 3.11 we see that 94% of zero forecasts for oats actually sowed zero oats for grain, with the corresponding figure for barley being 93%. Thus the vast majority of the oats points in strata 1, 3, 5 and 7 are at the "origin", as are the barley points in strata 1, 2, 5 and 6.

Notwithstanding these facts, we see from the plots in Figures 3.12-3.14 that some points on the vertical axes are far removed from the origin, and their being in separate post strata should produce much better estimators.

The conclusion arrived at from the above figures is that ratio estimation is inappropriate in many strata for at least one of the variables, irrespective of the values of correlation coefficients. Either number raised estimation (the "N-R" solution) or post stratification (the "P-S" solution) should be used for the relevant variables in these strata.

As an example, consider size stratum 7 of statistical division 4, defined by

$$(250 \leq FW < 2000) \cap (0 \leq FO < 40) \cap (50 \leq FB < 450).$$

There are 167 units in this stratum of which 72% have FO = 0. Clearly ratio estimation could not be used for oats. For the "N-R" solution we would use ratio estimation for wheat and barley in this stratum, and number raised estimation for oats. For the "P-S" solution, ratio estimation would be used for wheat and barley, and the post stratified estimator used for oats. The "P-S" solution utilizes the information on the selected units with a non-zero oats benchmark, whereas the "N-R" solution does not.

The suggested estimation schemes for both the "N-R" solution and the "P-S" solution, in conjunction with the 2 x 2 x 2 size stratification are given below in Figure 4.4.

Statistical Division 1

Size Stratum	"N-R" Solution Method of Estimation Wheat	Oats	Barley	"P-S" Solution Method of Estimation Wheat	Oats	Barley
1	R	NR	NR	R	PS	PS
2	R	NR	NR	R	PS	PS

(continued)

Statistical Division 5

Size Stratum	"N-R" Solution Method of Estimation Wheat	Oats	Barley	"P-S" Solution Method of Estimation Wheat	Oats	Barley
1	NR	NR	NR	PS	PS	PS
2	NR	R	NR	PS	R	PS
3	NR	NR	R	PS	PS	R
4	NR	R	R	PS	R	R
5	R	NR	NR	R	PS	PS
6	R	R	NR	R	R	PS
7	R	NR	R	R	PS	R
8	R	R	R	R	R	R

Statistical Divisions 2, 3, 4, 6, 7

Size Stratum	"N-R" Solution Method of Estimation Wheat	Oats	Barley	"P-S" Solution Method of Estimation Wheat	Oats	Barley
1	R	NR	NR	R	PS	PS
2	R	R	NR	R	R	PS
3	R	NR	R	R	PS	R
4	R	R	R	R	R	R
5	R	NR	NR	R	PS	PS
6	R	R	NR	R	R	PS
7	R	NR	R	R	PS	R
8	R	R	R	R	R	R

NR = number raised estimation

R = ratio estimation

PS = post stratified estimation

Figure 4.4

4.2.3 Simplified Post Stratification Estimators

The population need not necessarily be size stratified before applying the post stratification estimator. Indeed, post stratification on top of the 2^3 size stratification may create as many problems as it solves. Post stratification is effective when the sample size is moderately large, thereby giving all post strata a fair representation; this may not eventuate if the pre-sample stratification is too fine.

Another difficulty encountered when post stratification is applied on top of an ordinary size stratification is the "polarizing" effect whereby some size strata consist predominantly of one post-stratum type. In our case, using the 2^3 size stratification before post stratification produces size strata with as many as 90% of their units in a zero-benchmark post stratum, thus leading to a small sample for the post stratum in which ratio estimation is used.

In view of the above points, we should try applying the "P-S" estimator to coarser stratifications. Initially we shall apply it to statistical division stratification, since this is the coarsest stratification allowable.

The proportions of zero-benchmark units for statistical divisions are given in Figure 4.5.

Proportions of Zero Benchmarks

Statistical Division	N_h	P_0(W) %	P_0(O) %	P_0(B) %
1	261	18	41	81
2	3629	12	53	63
3	3259	11	40	71
4	3235	6	42	53
5	912	37	19	74
6	3660	10	34	48
7	2716	14	43	44

Figure 4.5

Another possibility is to use the finer geographic stratification, although gains in efficiency from further geographic stratification tend to be modest.

There are many other possibilities such as deeper post stratification or the "P-S" scheme in tandem with a coarser size stratification than the one above. An attempt to evaluate the efficiency of the "P-S" scheme will be made in 4.4.

4.3 Relative Standard Errors for the "N-R" Estimation Scheme

Strata were formed according to Figure 4.1, and stratum variances computed. Forecasters and non-forecasters were adjoined and a Neyman allocation of the sample of 2600 determined for each of wheat, oats and barley. The final allocation was found by the arithmetic mean of the three Neyman allocations, and design relative standard errors

were then computed using this compromise allocation. Relative Standard Errors (as percentages) for the three Neyman allocations and for the compromise (average) allocation are given in Figure 4.6. Design specifications are in parentheses, and an asterisk signifies a value that failed to meet the specifications. Fortunately the compromise allocation appears to work well.

Relative Standard Errors For "N-R" Scheme

	Neyman alloc. on Wheat	Neyman alloc. on Oats	Neyman alloc. on Barley	Compromise allocation
Wheat (2)	1.051	1.538	1.287	1.140
Oats (4)	3.831	2.894	3.699	3.229
Barley (3)	3.007 *	3.603 *	2.465	2.642

Figure 4.6

It is interesting to compare these relative standard errors with the figures from some cruder designs. Results are presented here for the following designs:

(a) Full size stratification for forecasters (as given in Figure 4.1, fine geographic stratification for non-forecasters, and number raised estimation in all strata.

(b) Stratification by statistical division only, and number raised estimation in all strata.

An optimum (Neyman) allocation for each crop was determined, and its relative standard error computed. The results, together with the corresponding values for the "N-R" scheme, and approximate relative efficiencies, are given in Figure 4.7. The relative efficiency is given by $2600 / n$, where n is the approximate sample size in the (a) or (b) design required to give the same performance as the "N-R" scheme.

Relative Efficiencies

	"N-R"	(a)	(b)
RSE (%) Wheat (2)	1.050	1.481	2.700
RSE (%) Oats (4)	2.894	3.015	4.254
RSE (%) Barley (3)	2.465	2.661	4.538
Relative efficiency	1	.80	.35

Figure 4.7

The above figures give some idea of the respective gains from ratio estimation and size stratification, and both are obviously substantial.

4.4 Gains from Post Stratification

Suppose there is ordinary stratification (geographic or geographic/size) forming H strata. Within each of the H strata and for each variable, the post stratified estimator of total is

$$\hat{T}_{PS} = (N_0/n_0) \sum_{i\epsilon s_0} X_i + \left(\sum_{i\epsilon s_1} X_i \Big/ \sum_{i\epsilon s_1} Y_i\right) \sum_{1}^{N} Y_i$$

where s_0 = sampled units having zero benchmark, and $s_1 = s - s_0$. N_0 is the number of zero-benchmark units in the stratum, and n_0 the number in the stratum sample.

As mentioned previously, this estimator will tend to underestimate the stratum total T because of points along the horizontal axis. (See Figures 3.12-3.14.) Although contributing nothing to the total, these points will drag the "estimated regression line" towards the horizontal. Refinements of the above estimator which attempt to overcome this problem are discussed in 4.6.

There is little doubt that post stratification in an already size stratified design will produce significant gains. Apart from the evidence presented in Holt and Smith (1979) for number raised estimation, we have the added bonus of being able to use ratio estimation instead of number raised estimation in post strata not containing zero benchmarks. A separate numerical study will be required to quantify the gains from post stratification in our multipurpose survey.

4.4.1 Post Stratification Following Geographic Stratification

To illustrate the power of post stratification in our multipurpose survey, we applied it to the population

stratified by statistical division only, which is the coarsest possible stratification. For this stratification the only valid alternative to the post stratification estimator is to use number raised estimation everywhere. Figure 4.8 presents a comparison of the two estimation schemes. Unconditional variances are used, and the relative standard errors for each crop are for a Neyman allocation based on that crop. (The "number raised" figures are from (b) of Figure 4.7 above.)

Relative standard errors

	Number raised	Post stratified
Wheat	2.700	1.315
Oats	4.254	3.491
Barley	4.538	2.863

Figure 4.8

4.5 Comparison of Size Stratified Designs with the Original PPS Design

An attempt will be made to compare the efficiencies of the new size stratified designs with the old probability proportional to size (PPS) design.

The methodology for the PPS design was described in Chapter 2, and formulae for estimators and sampling variances given. Ratio estimation was used in all geographic strata for all variables.

Using the matched census data, approximate PPS ratio estimation stratum variances were computed for the three crops and a sample of size 2600 allocated to the combined forecaster and non-forecaster populations using Neyman allocation for each crop. A simple average of the three Neyman allocations was found and design relative standard errors calculated for each crop using this allocation. The Hartley-Rao approximate PPS variance formula was used. (See A. 4 of the appendix.) This is all in accord with the methodology and design work used in actual past surveys. The relative standard errors obtained were 1.23%, 3.13% and 2.47% for wheat, oats and barley respectively.

This analysis, however, does not take into account the large number of zero benchmarks, and it is obvious from 4.2 and Figure 4.5 that ratio estimators cannot be defined.

Figure 4.9 below gives data on zero benchmarks for the geographic stratification used. The highest proportions of zero benchmarks are 37%, 78% and 89% for wheat, oats and barley respectively.

Proportions of Zero Benchmarks

Statistical Division	Geographic Stratum	N_h	$P_0(W)$ %	$P_0(O)$ %	$P_0(B)$ %
1	1	261	18	41	81
2	1	794	23	28	59
	2	865	17	46	62
	3	517	6	68	54
	4	327	6	58	64
	5	496	3	65	82
	6	630	6	28	65
3	1	1037	23	22	73
	2	750	6	40	59
	3	669	8	37	60
	4	484	4	72	88
	5	292	3	61	89
	6	27	15	59	78
4	1	719	2	39	52
	2	801	10	34	65
	3	613	3	47	51
	4	613	4	45	53
	5	489	9	52	37
5	1	912	37	19	74
6	1	1045	15	23	47
	2	1011	14	24	53
	3	725	2	34	36
	4	879	5	60	55
7	1	1128	18	19	45
	2	938	12	57	40
	3	595	10	63	48
	4	55	13	78	47

Figure 4.9

Although the proportions of zero benchmarks are generally high, the chances of zero benchmark samples being selected will be very small in most strata. For example, geographic stratum 5 of statistical division 3 has 89% of its 292 units with zero benchmarks for barley, but the chance of a zero-benchmark (barley) sample of size 45 being selected under SRS is approximately .002. (The sampling scheme is actually PPS, but it is the small value of the probability that is surprising.) The figure of 45 for the sample size resulted from the average Neyman allocation described above.

Thus ratio estimation can be (and was) "fairly safely" used even though the ratio estimator is undefined. Unfortunately, the usual variance formula will be meaningless and our design relative standard errors of 1.23%, 3.13% and 2.47% are irrelevant.

In heuristic terms, a ratio estimator attempts to estimate the unsampled values by fitting a regression line through the origin for X (estimation variable) against Y (benchmark variable). Points which are zero in both X and Y contribute nothing towards the estimation of the line and so the "effective" sample size is much smaller than the actual sample size. Hence the usual variance formula will give a serious underestimate of the sampling variance. (The actual sampling variance is a conditional variance given that the sample is not a zero benchmark sample.)

Clearly it is impossible to compare the multivariate size stratified designs with the PPS design, owing to the type of estimation used in the PPS design.

4.6 Ratio Estimation and Outliers

As mentioned in the appendix, ratio estimation is used when the estimation and benchmark variables are "roughly proportional", and is said to be more efficient than number raised estimation (in terms of variances under repeated sampling, ie the randomization distribution) when the correlation is high enough. Although a "regression through the origin" model is implicit when ratio estimation is applied, the randomization model does not permit an explicit formulation, since the population values $X_1,\ldots,X_N$ of the estimation variable are fixed arbitrary constants under the model. In particular, X cannot be stochastically related to the benchmark Y.

In the model based approach to sample survey theory, a stochastic model describing the relationship between X and Y is assumed explicitly. For example, if $\{E(X_i) = \beta Y_i,\ \text{var}(X_i) \propto Y_i\}$ is the regression model assumed, then the ratio estimator is the best linear unbiased predictor of $T = \sum_1^N X_i$ for a given sample.

A closer examination of data than merely looking at correlations will often lead to a plausible stochastic model which explains the data well. For example, the scatter plots in Figures 3.12-3.14 tell us much more about

the relationship between X and Y than do the correlations in Figure 3.15. There is clearly a "noisy" underlying linear relationship in each case, with outliers along both axes.

As we are interested in estimating population totals and not regression slopes per se, we cannot simply identify and remove the outliers. However we still need an accurate estimate of the slope of the regression line, and points along the axes that appear in the sample could seriously bias this estimate. This conflict can be resolved by treating the three groups of points (those along the vertical axis, those along the horizontal axis and those along the regression line) separately.

Since the number of points N_0 on the vertical axis (including the origin) is known, they can be handled by post stratification. (See 4.2.1). On the other hand, the number of points N_H on the horizontal axis (excluding the origin) is not known, so post stratification cannot be used for them. However, any points on this axis that are selected in the sample must not be included in the ratio estimator, since they will cause the estimate of total to be biased downwards. The difficulty lies in adjusting the ratio estimator when these points are removed.

Let s_H be the set of sampled units having a positive value for the benchmark and zero for the estimation variable, and $s_R = s - s_o - s_H$. A post stratification estimator which adjusts for points in s_H is

$$\hat{T}^{*}_{PS} = (N_0/n_0) \sum_{i\varepsilon s_0} X_i + \left(\sum_{i\varepsilon s_R} X_i / \sum_{i\varepsilon s_R} Y_i \right) \hat{Y}_R \quad ,$$

where $\hat{Y}_R$ is an estimator of the benchmark total for units off the horizontal axis. One such estimator would be

$$\hat{Y}_R = \sum_{1}^{N} Y_i - \frac{N - N_0}{n_1} \sum_{i\varepsilon s_H} Y_i \quad , \text{ which is a type}$$

of number raised estimator.

The plots in Figures 3.12-3.14, together with plots for other statistical divisions, show that points on the horizontal axis are concentrated near the origin, but with some large outliers. A further refinement would be to scale down any selected large outliers in the number raised estimator $\hat{Y}_R$. Hidiroglou and Srinath (1981) considered the effect of large outliers on number raised estimation, and compared the performances of a number of scaled estimators. It may be possible to modify them for our purposes.

4.7 Conclusions

Multivariate size stratification was applied successfully to the population data. A number of approaches were investigated, and the most promising reported in 4.1-4.6. Attention focused on the problem of zero benchmarks, and remedies were suggested.

The 2^3 size stratified design with the "N-R" estimation scheme is simple and practical, and appears to be very efficient. Optimality results for design parameters reported in the literature were not found very useful.

Post stratification was found to be a powerful technique for a multipurpose survey of this nature, but needs to be investigated further in numerical studies. A combination of pre-and post stratification offers the most promise.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques.* John Wiley & Sons, New York, third edition.

Dalenius, T. (1957). *Sampling in Sweden.* Contributions to the methods and theories of sample survey practice. Almqvist and Wicksell, Stockholm.

Dalenius, T., and Hodges, J.L., Jr. (1959). Minimum variance stratification. *Jour. Amer. Stat. Assoc.*, 54, 88-101.

Ghosh, S.P. (1958). A note on stratified random sampling with multiple characters. *Calcutta Stat. Assoc. Bull.*, 8, 81-89.

Ghosh, S.P. (1963). Optimum stratification with two characters. *Ann. Math. Statist.*, 34, 866-872.

Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Ann. Math. Statist.*, 33, 350-374.

Hidiroglou, M.A., and Srinath, K.P. (1981). Some estimators of a population total from simple random samples containing large units. *Jour. Amer. Stat. Assoc.*, 76, 690-695.

Holt, D., and Smith, T.M.F. (1979). Post stratification. *Jour. Roy. Stat. Soc.*, A142, 33-46.

Hughes, P. (1982). Optimum complete enumeration (CE) cut-offs. Australian Bureau of Statistics, Canberra.

Jarque, C.M. (1981). A solution to the problem of optimum stratification in multivariate sampling. *Appl. Statist.*, 30, 163-169.

Kish, L., and Anderson, D.W. (1978). Multivariate and multipurpose stratification. *Jour. Amer. Stat. Assoc.*, 73, 24-34.

Kokan, A.R., and Khan, S. (1967). Optimum allocation in multivariate surveys: an analytical solution. *Jour. Roy. Stat. Soc.*, B29, 115-125.

APPENDIX

ESTIMATION IN SAMPLE SURVEYS

A.1 MATHEMATICAL MODEL

In the conventional sample survey model, there is a population of N identifiable units, where N is finite and known. For convenience we number them $1,2,\ldots,N$. Associated with the i^{th} unit is an unknown quantity X_i which for simplicity we take to be a real number, although it would often be a vector.

A sample of size n is a subset s of the N units. Units are randomly selected according to a probability function $p(\cdot)$ on the set S of all samples. The probability function $p(\cdot)$ is often referred to as the sample design, or the randomization distribution. The sample data ω consists of the randomly chosen subset s together with the observations X_i, $i \in s$. The sample space Ω is the totality of the ω's. For each point $\theta = (X_1,\ldots,X_N)$ of the parameter space Θ there is a discrete probability measure P_θ on the sample space Ω determined by the randomization distribution p on S. Θ is a subset (often the positive region) of R^N.

An estimator $\hat{T}$ of a real parametric function $T(\theta)$ is a real valued function on Ω, and at the point $\theta \in \Theta$ has expectation

$$E_\theta[\hat{T}] = \int_\Omega \hat{T}\, dP_\theta = \sum_{s \in S} \hat{T}(\omega)\, p(s).$$

By definition, $\hat{T}$ is a design unbiased estimator of $T(\theta)$ if $E_\theta [\hat{T}] = T(\theta)$ identically in θ.

Note that under the above model the X_i are fixed constants, and that the probability distribution p due to the artificial randomization is the only probabilistic part of the model. The interpretation of probability is the longrun relative frequency for repeated sampling under p from the same finite population.

Although the model is simple and is known no hold exactly, its generality (the parameter space is N dimensional) does not allow the existence of best estimates such as MVUE's.

Alternative formulations of the sample survey model treat $\underset{\sim}{X} = (X_1, \ldots, X_N)$ as a random vector with a distribution belonging to a parametric family. This approach will not be taken in this study.

A.2 SIMPLE RANDOM SAMPLING

$$p(s) = \binom{N}{n}^{-1} \text{ for all } \binom{N}{n} \text{ possible samples of size n.}$$

A.2.1 NUMBER RAISED ESTIMATION

The number raised or simple expansion estimator of $T = \sum_{1}^{N} X_i$ is

$$\hat{T} = (N/n) \sum_{i \in s} X_i \quad .$$

$E[\hat{T}] \equiv T$, and so $\hat{T}$ is design unbiased.

$$\text{var } (\hat{T}) = (N^2/n)(1 - n/N)S^2, \text{ where } S^2 = (N-1)^{-1} \sum_{1}^{N} (X_i - \bar{X})^2.$$

$$\hat{\text{var}} \ (\hat{T}) = (N^2/n)(1 - n/N)\hat{S}^2, \text{ where } \hat{S}^2 = (n-1)^{-1} \sum_{i \in s} (X_i - \bar{X}_s)^2.$$

A.2.2 RATIO ESTIMATION

Suppose we have known supplementary (prior) information $Y_1, \ldots, Y_N$ associated with the N units. If prior information indicates that X and Y are "roughly proportional", then the ratio estimator

$$\hat{T}_R = \left(\sum_{i \in s} X_i \Big/ \sum_{i \in s} Y_i\right) \cdot \sum_{1}^{N} Y_i$$

is often used as an estimator of $T = \sum_{1}^{N} X_i$.

The variable Y is called a benchmark variable. It may be a previous value of X from a census, or a related quantity known from records or a census.

To a first order approximation in n,

$$E\,[\hat{T}_R] \doteq T\ ;$$

$$\text{var}\,(\hat{T}_R) \doteq (N^2/n)(1-n/N)\,(N-1)^{-1} \sum_1^N (X_i - RY_i)^2,$$

$$\text{where } R = \sum_1^N X_i \Big/ \sum_1^N Y_i .$$

Hence with the same approximation,

$$\text{var}\,(\hat{T}_R) < \text{var}\,(\hat{T}) \quad \text{iff } \rho_{xy} > \tfrac{1}{2}\, C_y/C_x$$

where C denotes coefficient of variation.

A.3 SYSTEMATIC SAMPLING

In practice, systematic and not simple random sampling (SRS) is often used. The N units are randomly ordered and then the ordered units numbered k, k+ N/n, k+ 2N/n, ..., k+ (n-1)N/n are chosen, where k is a randomly chosen integer between 1 and N/n. (N/n is called the skip interval.) SRS variance formulae are used as approximations for systematic sampling variance formulae.

A.4 PROBABILITY PROPORTIONAL TO SIZE (PPS) SAMPLING

With each unit in the population is associated a known "measure of size" Z_i, and a sample of distinct units is chosen such that the probability of a unit being selected is proportional to its measure of size. The size variable Z

should be highly correlated with the survey variable X, and could be a benchmark variable as mentioned in A.2.2.

Let Π_i = P {i^{th} population unit selected}

$$= \sum_{s \ni i} p(s).$$

Then $\Pi_i \propto Z_i \Rightarrow \Pi_i = nZ_i / \sum_1^N Z_i, \quad i = 1, \ldots, N.$

Note that the measures of size must satisfy $Z_i \leq \sum_1^N Z_i/n$.

The design unbiased Horvitz-Thompson estimator $\hat{T}_p = (\sum_1^N Z_i/n) \sum_{i \in s} (X_i/Z_i)$ is used to estimate $T = \sum_1^N X_i$.

The variance of $\hat{T}_p$ is

$$\text{var } (\hat{T}_p) = \tfrac{1}{2} \sum_{i \neq j}^N (\Pi_i \Pi_j - \Pi_{ij}) \left[\frac{X_i}{\Pi_i} - \frac{X_j}{\Pi_j} \right]^2 ,$$

where Π_{ij} is the joint probability of selection of the i^{th} and j^{th} units. PPS sampling is considerably more efficient than SRS when X and Z are "roughly proportional".

$$\text{The estimator } \hat{\text{var}} \ (\hat{T}_p) = \tfrac{1}{2} \sum_{i \neq j \in s} \frac{\Pi_i \Pi_j - \Pi_{ij}}{\Pi_{ij}} \left[\frac{X_i}{\Pi_i} - \frac{X_j}{\Pi_j} \right]^2$$

is an unbiased estimator of var $(\hat{T}_p)$ providing $\Pi_{ij} > 0$ for all $i \neq j \in \{1,2,\ldots,N\}$, but requires the knowledge of the Π_{ij}. For any given $\{Z_1,\ldots,Z_N\}$, and hence $\{\Pi_1,\ldots,\Pi_N\}$, there correspond many arrays (Π_{ij}), depending on which PPS sampling

scheme is used. The Π_{ij} are difficult to evaluate for most schemes, and are usually not known. Another disadvantage of $\hat{\text{var}}\ (\hat{T}_p)$ is that it can take negative values.

For systematic PPS sampling from a random list, Hartley and Rao (1962) gave approximate expressions for Π_{ij} valid for large N and small n/N. These give

$$\text{var}\ (\hat{T}_p) \doteq \frac{1}{n} \sum_{1}^{N} \frac{Z_i}{Z_T} \left(1 - (n-1)\frac{Z_i}{Z_T} \right) \left[\frac{Z_T}{Z_i} X_i - X_T \right]^2$$

$$\hat{\text{var}}\ (\hat{T}_p) = \frac{Z_T^{\,2}}{n^2\,(n-1)} \sum_{i<j \varepsilon s} \left(1 - \frac{n}{Z_T}\,(Z_i+Z_j) + \frac{n}{Z_T^2} \sum_{1}^{N} Z_i^{\,2} \right) \left[\frac{X_i}{Z_i} - \frac{X_j}{Z_j} \right]^2 ,$$

$$\text{where } Z_T = \sum_{1}^{N} Z_i \ .$$

Systematic PPS sampling is effected by using ordinary systematic sampling on a list of cumulative Z totals with skip interval $\sum_{1}^{N} Z_i/n$.

A.4.1 RATIO ESTIMATION WITH PPS SAMPLING

As in A.2.2, the ratio estimator and its variance are given by

$$\hat{T}_R = \left(\sum_{i \varepsilon s} X_i/Z_i \right) \cdot \left(\sum_{i \varepsilon s} Y_i/Z_i \right)^{-1} \cdot \sum_{1}^{N} Y_i \quad ;$$

$$\text{var}\ (\hat{T}_R) \doteq \frac{1}{n} \sum_{1}^{N} \frac{Z_T}{Z_i} \left(1 - (n-1)\ \frac{Z_i}{Z_T} \right) (X_i - R\ Y_i)^2 .$$

(Here Y is the benchmark variable.)

A.5 STRATIFIED SAMPLING

The population of N units is partitioned into H subsets with corresponding X values

$X_{11}, X_{12}, \ldots, X_{1N_1}$	-	Stratum 1
$X_{21}, X_{22}, \ldots, X_{2N_2}$	-	Stratum 2
.	.	.
.	.	.
.	.	.
$X_{H1}, X_{H2}, \ldots, X_{HN_H}$	-	Stratum H

Independent samples of sizes $n_1, n_2, \ldots, n_H$ are selected by one of the above sampling schemes (not necessarily the same type of random sampling for all strata) and the estimate of

$$T = \sum_{h=1}^{H} \sum_{i=1}^{N_h} X_{hi}$$

obtained by summing individual estimates of stratum totals. For example, with SRS and number raised estimation in all strata,

$$\hat{T}_s = \sum_{h=1}^{H} (N_h/n_h) \sum_{i \epsilon s_h} X_{hi}$$

is the usual unbiased estimator of T.

$$\text{var } (\hat{T}_s) = \sum_{h=1}^{H} (N_h^2/n_h)(1 - n_h/N_h)\ S_h^2$$

and $$\hat{\text{var}}\ (\hat{T}_s) = \sum_{h=1}^{H} (N_h^2/n_h)(1 - n_h/N_h)\ \hat{S}_h^2.$$

A.5.1 OPTIMUM ALLOCATION

Consider a fixed stratification of the population and fixed total sample size n.

If $\hat{T}_s = \sum_{h=1}^{H} \hat{T}_h$ is the estimator of T, and

$\text{var } (\hat{T}_s) = K + \sum_{h=1}^{H} V_h^2/n_h$, where K is independent of the n_h, then the sample allocation which minimizes $\text{var } (\hat{T}_s)$ is $n_h = nV_h / \sum_{1}^{H} V_h$. This is known as Neyman allocation.

For example with SRS,

$$\text{var } (\hat{T}_s) = \sum_{1}^{H} (N_h^2/n_h)(1 - n_h/N_h)\ S_h^2$$

$$= \sum_{1}^{H} (N_h S_h)^2/n_h - \sum_{1}^{H} N_h S_h^2$$

and the Neyman allocation is

$$n_h = n\ N_h S_h / \sum_{1}^{H} N_h S_h.$$

Thus the sampling fraction n_h/N_h is proportional to S_h, giving a heavier sampling rate in the more variable strata. This is in contrast to proportional allocation in which n_h/N_h is constant. Clearly proportional allocation will be inefficient when the S_h vary substantially.

A.5.2 COMPLETE ENUMERATION

If a size variable is used to stratify a population, strata corresponding to large values of the size variable tend to have large variances, and complete enumeration of the largest size stratum substantially reduces the variance of $\hat{T}_s$. Hughes (1982) describes a method for determining the optimum cut-off point for the completely enumerated (CE) stratum and reports that analysis of actual survey data shows that the CE stratum should often contain as much as 45% of the sample.

A.5.3 OPTIMUM STRATIFICATION

Under a given method of sample allocation (Neyman, proportional, equal etc.), var $(\hat{T}_s)$ is a function of the stratum boundaries and of H. For fixed H, and strata formed by a single size variable, Dalenius (1957) presents a solution to the optimum boundary problem. A method of finding approximately optimum stratum boundaries is the Dalenius-Hodges cum $\sqrt{f}$ rule (Dalenius and Hodges (1959)), which is valid for SRS with number raised estimation and Neyman allocation. (See Cochran (1977) for more details.)

For ratio estimation, no simple rule exists, although for want of something better, the cum $\sqrt{f}$ rule is sometimes used.

A.5.4 POST STRATIFICATION

In some surveys where stratification is desirable, the stratum to which a sampled unit belongs is not known until after the survey is conducted. Post-stratification is stratification of the sampled units after selection of the sample. The usual stratified estimate is used and so the stratum sizes N_h must be known. (Note that the stratum sample sizes n_h are now random variables.) Poststratification is almost as efficient as proportional stratified sampling.

Consider estimation of the total $T = \sum_{1}^{N} X_i$ from a simple random sample of size n by the poststratified estimator

$$\hat{T} = \sum_{h=1}^{H} (N_h/n_h) \sum_{i \in s_h} X_{hi} , \quad \sum_{1}^{H} n_h = n.$$

For $\underset{\sim}{n}$ such that $n_h > 0$ for all h,

$E[\hat{T}|\underset{\sim}{n}] = T.$ (See A.5 of this appendix.)

Note that $\hat{T}$ is undefined when some $n_h = 0$ unless its definition is altered to include these cases; for example, adjacent strata could be amalgamated. Hence

$$E[\hat{T}] = E[E[\hat{T}|\underset{\sim}{n}]]$$

$$= E[T]$$

$$= T,$$

providing $P\{n_h = 0 \text{ for some } h\} = 0$.

In general, some samples will have $n_h = 0$ for some h, and so $E[\hat{T}] \neq T$. However if we make the assumptions

(i) H is small;

(ii) n is large,

then $E[\hat{T}] \doteq T$.

Similarly, $\text{var}\,(\hat{T}|\underset{\sim}{n}) = \sum_{1}^{H} (N_h^2/n_h)(1 - n_h/N_h)\, S_h^2$

for $\underset{\sim}{n}$ such that $n_h > 0$ for all h. Using

$$\text{var}\,(\hat{T}) = E[\text{var}\,(\hat{T}|\underset{\sim}{n})] + \text{var}\,(E[\hat{T}|\underset{\sim}{n}]),$$

it is easy to show that (see Cochran (1977))

$$\text{var}\,(\hat{T}) \doteq (N^2/n)(1-n/N) \sum_{1}^{H} (N_h/N) S_h^2 + n^{-2} \sum_{1}^{H} (1-N_h/N) S_h^2 .$$

The first term is the variance under proportional stratified sampling, and the second term is relatively small under assumptions (i) and (ii) above, and if either the N_h or the S_h^2 are not very variable.

Thus post stratification is nearly as efficient as proportional stratified sampling, as was asserted above.

The sampling variance of $\hat{T}$ includes the variation of the sample configuration $\underset{\sim}{n}$, but as Holt and

Smith (1979) point out, the appropriate framework for statistical inference after the sample is drawn is the conditional distribution given $\underset{\sim}{n}$. That is, var $(\hat{T}|\underset{\sim}{n})$ should be used for determining confidence intervals for T. However at the planning stage, the unconditional variance var $(\hat{T})$ should be used to determine the design variances.

The unconditional variance of the post stratified estimator indicates that only modest gains will be made over the ordinary self-weighting estimator. However Holt and Smith found that when the *conditional* variances were compared for a number of data sets, the post stratified estimator was often very much better, and seldom worse than the self-weighting estimator.