R - Statistical Methods

An Empirical Study of the Area Frame Stratification

by

Nicholas J. Ciancio

Dwight A. Rockwell

Robert D. Tortora

Sampling Studies Section
Sample Survey Research Branch
Research Division
Statistical Reporting Service
U.S. Department of Agriculture
Washington, D.C.

July 1977

SF 77-07

INTRODUCTION

The area frame used by the Statistical Reporting Service (SRS) consists of the total land area of the U.S. This land area is divided into the land use classes (areas) such as agricultural, recreational, urban, etc. In particular, the agricultural class is stratified into four strata using percent cultivated as the stratification variable. The strata definitions are:

Stratum 11: more than 75% of land is cultivated,

Stratum 12: between 50% and 75% cultivation,

Stratum 20: between 15% and 50% cultivation,

Stratum 40: less than 15% of land is cultivated.

Aerial photographs are the means of delineating the strata according to land use and count units within strata. Each count unit is delineated by identifiable boundaries and is approximately 10 to 15 square miles in size. Segments are smaller land areas within count units and form the primary sampling units. The sum of the land areas in the count units divided by the desired size of each segment gives the population of potential sample segments for the stratum. Random numbers indicate which count units contain the selected sample of segments. Each of these count units are then subdivided into the proper number of segments and a segment is selected at random.

There is little analytic or empirical evidence to prove or disprove the use of the current stratification variable, the current number of strata or the current stratum boundary values.

There are five major problems that must be solved when stratified sampling is going to be used for a survey design. These problems are the choice of the design within strata, the allocation of the sample size to each strata, the choice of the stratification variable, the best values of the stratification variables for stratum boundaries, and the choice of the number of strata. Although the solutions to these problems are interrelated, this paper deals

empirically with the latter three for the area frame. Before developing the methodology a descriptive analysis of the area frame points out problem areas.

Descriptive Analysis

Preliminary descriptive analysis is an important part of the overall analysis of a project. In this light we proceed to explore certain aspects of the area frame. All data used is from the 1975 June Enumerative Survey for Ohio, Illinois, Kansas and Minnesota.

In the four main agricultural strata (11, 12, 20 and 40) the average number of agricultural tracts per sample segment was 5.1 for Ohio, 6.2 for Illinois, 4.3 for Kansas, and 4,8 for Minnesota. These figures do not constitute a large workload for the enumerators. According to the JES Supervising and Editing Manual, a problem segment is one in which there are 20 or more tracts of which 10 or more are agricultural tracts. Illinois has the highest average of 6.2 which is not close to 10.

Potential sample segments in a given count unit are made as nearly equal in size as possible following identifiable natural boundaries and selected with equal probability. An examination of their size (in square miles) could indicate a problem with the basic rules used in frame construction. Table 1 presents the expected and observed average segment sizes by stratum for each State. The expected segment size is that which was decided upon before the count units were divided into segments during frame construction. Observed segment size was derived from that reported by tract operators during the 1975 JES and is subject to error.

Table 1: Expected and Observed Average Segment Size

	Ohio		Illinois		Kansas		Minnesota	
Stratum	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.
11	.5	.5	1.0	1.0	1.0	.99	1.0	.98
12	.5	.54	1.0	1.01	1.0	1.0	1.0	1.01
20	1.0	.91	1.0	.98	1.0	1.01	2.00	2.05
40	1.0	1.0	2.0	2.09	4.0	4.02	3.09	3.09

The largest deviation from the expected is 0.09 which seems tolerable.

We now examine the average compositon of the area frame segments. Table 2 considers: a) the average percent of segment area in permanent pasture, b) the average percent of segment area in non-agricultural use, and c) the average percent of segment area under cultivation for the four main agricultural strata.

Table 2

	Ohio			Illinois			Kansas			Minnesota		
	a	b	c	a	b	c	a	b	c	a	b	c
11	8.2	5.6	76.7	3.7	2.7	88.0	11.4	0.8	84.8	4.3	2.4	85.9
12	11.1	17.2	53.7	13.1	9.5	63.0	27.8	1.7	66.0	16.9	10.3	51.6
20	23.3	24.3	29.3	18.0	8.9	50.9	50.9	2.5	42.1	13.4	28.9	32.1
40	23.3	49.0	7.1	16.6	58.4	12.9	73.1	7.1	18.7	3.2	87.0	4.8

These percents are very interesting. Of the sixteen average percents cultivated in agricultural strata, only Illinois - stratum 20 and Kansas - stratum 40 do not conform to current stratum definitions. This would suggest that our current definitions are adequate. The problem with these averages are the ones near the endpoints of the boundary value, e.g., Ohio - stratum 11 has an average percent cultivated of 76.7%. This indicates that there must have been quite a few segments that fell below the stratum boundary value. This might indicate that a change in the number of strata or the stratum boundary values would reduce the number of falling outside the boundaries. However, caution is needed because it is often difficult to discern permanent pasture from cultivated land during photo interpretation to establish stratum boundaries.

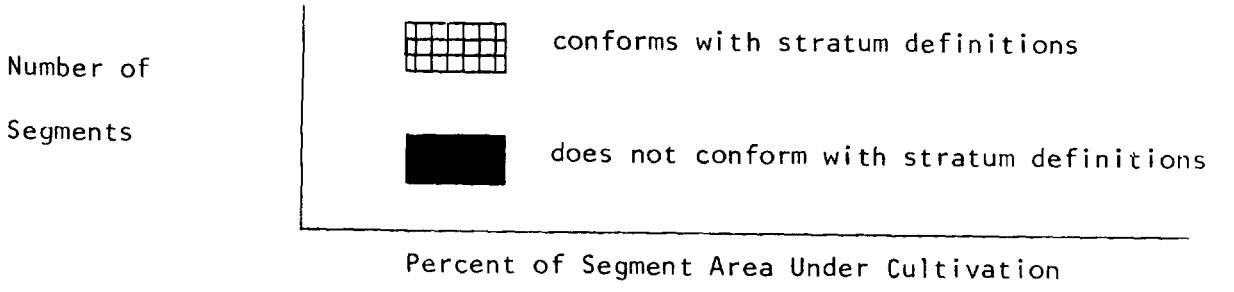
In order to pursue this last point further we look at the number and percent of segments conforming with strata definitions. For the four main agricultural strata, Table 3 presents number and percent of segments conforming to strata boundary values.

Table 3

	Ohio			Illinois -			Kansas			Minnesota		
	Total # Segments	# Conforming	%	Total # Segments	# Conforming	%	Total # Segments	# Conforming	%	Total # Segments	# Conforming	%
11	140	87	62.1	170	147	85.9	170	129	75.9	160	135	84.4
12	55	24	43.6	50	20	40.0	120	52	43.3	40	42	46.7
20	30	23	76.7	40	19	47.5	100	44	44.0	25	17	68.0
40	25	21	84.0	6	5	83.3	15	7	46.7	30	27	90.0

This table shows a little more than the previous tables. It not only picks out the same problem strata as before but also indicates other potential problem areas such as Stratum 12 in all four states. This tells us that segments within the count units do not all conform to the land use strata definitions. Segments which are outside the stratum definition results in a

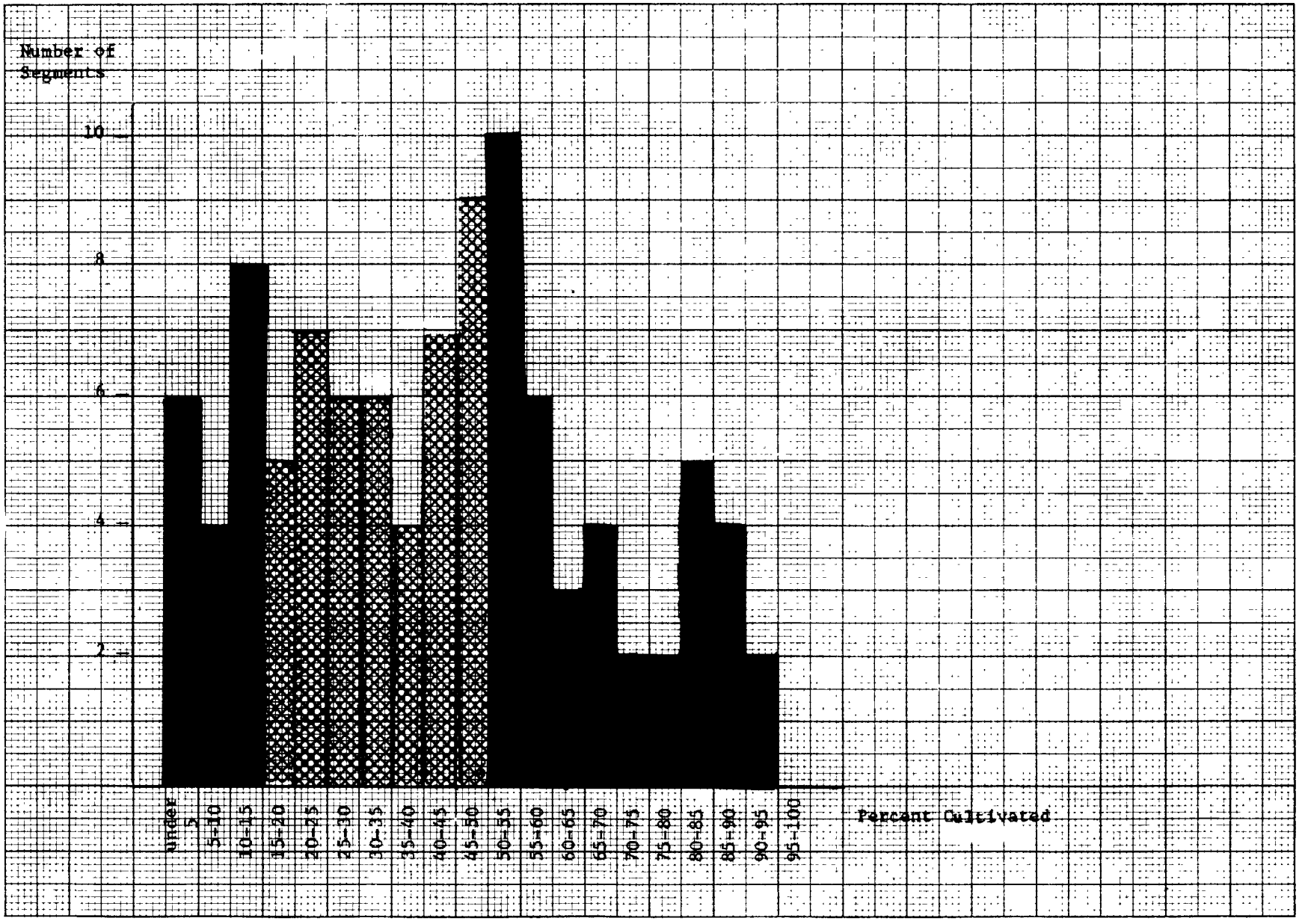
loss of precision, i.e., an increased coefficient of variation of the estimates ([13], p. 129). Of course, it is not possible to obtain perfect classification of segments because photo interpretation is imperfect, identifiable boundaries must be followed even if some land is outside the stratum definition, and not all land is uniform with a count unit. A single segment selected at random from a count unit may not reflect the land use of the majority of the count unit. Plots were constructed to pictorially show the percent of segment area under cultivation versus the number of segments. The general form of each graph is as follows:



The plots are constructed by State and by stratum. One typical plot is for stratum 20 (the remaining plots for the four main agricultural strata are located in Appendix A). The graph pictorially shows where misclassifications are occurring using the present definitions.

There are 56 segments that do not conform with strata definitions and 44 segments that do conform. Many of the other plots are not as extreme as this one but they are of a similar structure. Again, one wonders how these results affect the efficiency of the frame as presently stratified. Thus one should attempt to determine the optimum number of strata and stratum boundary values.

Kansas Stratum 20



The previous discussion indicates potential problem areas in the area frame. By keeping a historical record of descriptive statistics of this type an indication of problems in the area frame can be obtained. Therefore, frame reconstruction or updating can be initiated on a "need" basis. Known percent cultivated computations for areas inside previous sample segments might be useful as a guide to photo interpreters when reconstructing stratum boundaries. Eventually technological advances in satellite land use classification might also be used to improve the homogeneity of segments with strata.

The Stratification Variable

The important variables to be estimated were assumed to be: corn, winter wheat and soybeans platned for all purposes, total hogs and pigs and total cattle and calves. Possible stratification variables are: total agricultural acreage, permanent pasture, nonagricultural acreage, cropland, total acres, percent cultivated and percent agricultural acreage. According to Cochran [2] the best variable to stratify on is the variable you want to estimate. Obviously this cannot be done. The next best variable to use is one which is highly correlated with all five of the variables to be estimated.

Correlation coefficients between the variables to be estimated and the possible stratification variables were obtained for Ohio and Kansas over all strata and within each stratum. Since we are only concerned with agricultural data we deleted all strata that were non-agricultural. We did this because the non-agricultural starta will have to be included in any stratification of a state that will be proposed.

Summary tables of correlation coefficients between the variables to be estimated and possible stratification variables for the four main agricultural strata for Ohio and Kansas are presented in Tables 4 and 5.

Table 4

Ohio

	Total Agr. Acreage	Perm. Past.	Non-Agr. Acreage	Crop-land	Total Acres	% Cult.	% Agr. Acreage
Stratum 11							
Corn	.344	-.015	-.213	.392	.277	.325	.230
Winter Wheat	.331	-.284	-.243	.484	.237	.456	.261
Soybeans	.399	-.393	-.180	.566	.376	.468	.218
Hogs	.161	-.014	-.096	.135	.133	.087	.097
Cattle	.081	.409	-.110	-.184	.017	-.226	.110
Stratum 12							
Corn	.354	-.226	-.260	.667	.159	.543	.279
Winter Wheat	.245	-.195	-.245	.406	.039	.355	.232
Soybeans	.386	.149	-.386	.351	.062	.312	.400
Hogs	.058	-.056	-.117	.074	-.055	.093	.117
Cattle	.203	.379	-.236	.020	-.004	.006	.246
Stratum 20							
Corn	.318	-.196	-.226	.785	.154	.780	.267
Winter Wheat	.312	-.212	-.201	.656	.187	.617	.247
Soybeans	.347	-.097	-.335	.530	.021	.572	.348
Hogs	.356	.003	-.219	.316	.229	.217	.241
Cattle	.639	.543	-.509	.393	.217	.276	.537
Stratum 40							
Corn	.084	-.185	-.187	.673	-.280	.714	.177
Winter Wheat	.070	-.563	-.162	.387	-.250	.414	.135
Soybeans	*						
Hogs	.065	-.112	-.090	.302	-.080	.336	.085
Cattle	.638	.683	-.520	.277	.140	.281	.570

* No Soybeans found in this stratum

Table 5

Kansas

	Total Agr. Acreage	Perm Past.	Non-Agr. Acreage	Crop-land	Total Acres	% Cult.	% Agr. Acreage
Stratum 11							
Corn	.088	-.180	-.005	.180	.089	.162	.007
Winter Wheat	.153	-.283	-.081	.343	.134	.324	.084
Soybeans	.053	-.138	-.179	-.225	-.004	.249	-.183
Hogs	-.018	-.048	.048	.026	-.005	.034	-.050
Cattle	.251	.261	..071	-.176	.277	-.328	-.045
Stratum 12							
Corn	-.068	-0.113	.081	.096	-.010	.094	-.082
Winter Wheat	.090	-.484	-.157	.567	-.038	.057	-.156
Soybeans	-.078	-.122	.236	.025	.133	.008	-.235
Hogs	-.070	-.067	-.063	.027	-.158	.049	.062
Cattle	.140	.590	-.103	.555	.085	-.563	.106
Stratum 20							
Corn	-.348	-.120	.272	-.096	-.234	-.047	-.278
Winter Wheat	.156	-.561	-.174	.712	.013	.717	.17[illegible]
Soybeans	-.126	-.234	.131	.121	-.026	.124	-.1[illegible]
Hogs	.016	-.174	-.035	.184	-.030	.193	.0[illegible]
Cattle	.207	.435	-.219	-.326	.037	.322	.21
Stratum 40							
Corn	.058	.013	-.068	.089	-.051	.096	.067
Winter Wheat	.379	-.151	-.301	.912	.311	.895	.301
Soybeans	.058	.013	-.068	.089	-.051	.096	.067
Hogs	.092	.296	-.119	-.036	-.117	.354	.119
Cattle	.504	.513	-.522	.812	-.139	.078	.522

As can be seen from the tables no one variable clearly distinguishes itself as the best stratification variable. Of the possible seven candidates four variables approximately exhibited the same correlations with the variables to be estimated. The four possible candidates for stratification are total agricultural acreage, total acres, cropland and percent cultivated.

Since percent cultivated is the only variable of the four that is not open ended, i.e., has a definite range (0 - 100%), it is recommended that we continue to use it as the stratification variable.

Methodology

The area frame is used to obtain estimates of crops and livestock. The data used for the study was from the 1975 JES for Ohio and Kansas. The major characteristics estimated from the area frame are corn, soybeans, winter wheat, cattle and hogs. Table 6 (extracted from Table 4) gives the estimated correlation coefficients between the stratification variable and the characteristics estimated for segments by stratum for Ohio.

Table 6: Estimated Correlation Coefficients between percent cultivated and major characteristics for Ohio.

	Stratum 1	Stratum 2	Stratum 3	Stratum 4
corn	0.325	0.543	0.780	0.714
winter wheat	0.456	0.355	0.617	0.414
soybeans	0.468	0.312	0.572	*
hogs	0.087	0.093	0.217	0.336
cattle	-0.226	0.006	0.276	0.281

* no soybeans found in this stratum

It is clear from the table that the best correlations occur between percent cultivated and the crops. The same is true in Kansas although to a lesser degree. Therefore the optimatization procedures will be applied to corn, soybeans, and winter wheat. The boundary values for percent cultivated land obtained on the basis of crops will then be used to obtain livestock estimated variances. These variances will be compared to the variances obtained using the current strata to show losses and gains.

The methods used to compute the boundary values are those proposed by Dalenius and Hodges [6], [7], Durbin [8], Ekman [9], Sethi [15] and the technique known as Equal Aggregate Output. Variances for crop estimates are

compared to determine the optimum number of strata for each technique as well as to determine the best technique for stratification.

It should be noted here that this empirical study deviates from practice in one way. In practice segments are drawn from large areas preclassified into particular land use strata. As we have seen, individual segments may not conform to the stratum definition. In this study the actual amount of cultivated land contained in each segment is calculated. Based on these calculation the segments are place into the proper strata. Therefore the frame used in this study is a more exact mapping than the actual area frame and hence the estimated variances obtained here will be smaller than those obtained in practice. We now turn to a discussion of the methods used to obtain approximate optimum boundary values.

Neyman's allocation was used as the basis of comparison for the approximate stratification techniques. Let $y_o, y_1, \ldots, y_L$ be the stratum boundary values for 1, 2,..., L strata and let $\bar{y}_h$ be the sample estimate of the population mean, S_h be the population standard deviation and $W_h = N_h/N$ be the ratio of the number of sampling units in stratum h to the total number in the population. With stratified sampling the usual estimate of the population total is

$$y_{st} = N \sum_{h=1}^{L} W_h \bar{y}_h$$

with variance

$$V(y_{st}) = N^{-1} \Sigma N_h S_h^2 (1 - \frac{n_h}{N_h})/n_h.$$

Using Lagrange multipliers and ignoring the fpc, $V(y_{st})$ is minimized by

$$V_{min}(y_{st}) = n^{-1} (\Sigma N_h S_h)^2. \quad (1)$$

To compute optimum stratum boundaries for the selected stratification variables, five techniques for defining these boundaries were applied to the variable percent cultivated. It should be understood that it was not possible to consider all the different stratification techniques. While examining the literature (e.g. Anderson, et. al. [1], Cochran [2] and Kpedekpo [12]) it became clear that various methods lead to approximately equal results or that some methods were inferior. Therefore, we concentrated our effort on the five techniques Dalenius and Hodges, Durbin, Ekman, Sethi and Equal Aggregate Out-Put. A brief technical description of each technique is given in Appendix B. In addition, Appendix B contains an application of the five techniques of stratification to a numeral example. Computations are presented in Table B1. The results of applying each of the techniques to the JES data to determine boundaries for four strata are shown in Table B2. Stratum boundary values for 2, 3, 4 and 5 strata are summarized by technique in Tables B3 and B4 (Appendix B) for Ohio and Kansas, respectively.

In the following a method for determining the optimum set of stratum boundary values, i.e., picking a set of stratum boundary values in each State which gives the smallest variance is given.

Results

For each of the five techniques a variance (using equation (1)) was computed for each of the sets of strata 2, 3, 4 and 5 for Ohio and Kansas.

The first problem to resolve is the number of strata necessary. There are two methods for deciding on the number of strata.

1. A variance stabilization can be used where an increase in the number of strata does not yield a significant gain in the precision of the estimates.
2. Cochran [3] suggests running a linear regression model and comparing variances.

Cochran's technique works well if you indeed have a linear regression. If you do not have a linear regression model then one would use the variance stabilization technique. Simple linear regressions on the stratification variable were executed and did not fit the data. Therefore, we will proceed to use the variance stabilization technique.

In order to determine where a significant gain in precision occurs the ratio $V_L/V_{L-1} \doteq (L-1)^2/L^2$ (valid for the rectangular distribution, [5]) will be used. This formula gives 0.250, 0.444, 0.562 and 0.640 for L = 2, 3, 4 and 5 strata, respectively. For each strata by crop and for each approximate stratification technique the ratio V_L/V_{L-1} was computed by State. Table 7 exhibits these calculations. The criteria for selecting the "optimum" number of strata is to compare V_L/V_{L-1} to $(L-1)^2/L^2$ and count the number of times a set of strata has $V_L/V_{L-1} \geq (L-1)^2/L^2$. The strata with the 2 highest counts are selected for further consideration.

We see that in Ohio for all crops 2 strata has a significant gain in precision 2 times, 3 strata has a significant gain 3 times, 4 strata 5 times and 5 strata 6 times. In Kansas for all crops 2 strata has a significant gain 5 times, 3 strata 8 times, 4 strata 8 times and 5 strata 3 times. Therefore in Ohio we select 4 and 5 strata and in Kansas 3 and 4 strata for further study. These strata will be used to determine the optimum stratum boundary values.

Table 7

The ratio of V_L/V_{L-1} for strata 2, 3, 4 and 5 by crop, stratification technique and State.

	Strata	Dalenius Hodges	Durbin	Ekman	Sethi	Eq. Aggre.
			OHIO			
Corn	2	.206	.216	.234	.256	.256
	3	.389	.400	.350	.447	.457
	4	.586	.561	.550	.549	.534
	5	.602	.600	.663	.723	.692
Winter Wheat	2	.189	.210	.189	.238	.238
	3	.378	.378	.389	.397	.404
	4	.553	.533	.523	.574	.574
	5	.644	.624	.707	.638	.582
Soybeans	2	.170	.180	.104	.214	.214
	3	.398	.426	.671	.361	.412
	4	.554	.508	.492	.657	.569
	5	.602	.600	.666	.560	.532
			KANSAS			
Corn	2	.242	.242	.242	.190	.190
	3	.449	.419	.454	.457	.354
	4	.567	.615	.630	.555	.549
	5	.612	.587	.616	.523	.555
Winter Wheat	2	.184	.184	.184	.206	.235
	3	.444	.417	.444	.469	.424
	4	.546	.582	.606	.544	.568
	5	.642	.611	.603	.713	.693
Soybeans	2	.266	.266	.266	.287	.307
	3	.426	.420	.426	.446	.488
	4	.549	.556	.585	.576	.552
	5	.627	.596	.571	.624	.623

Now the optimum boundary values will be determined by computing the ratio of the variance for each technique to the variance using the current technique for those strata selected.

Table 8 gives the ratio of the variance computed under each approximately optimum technique to the variance using the current boundary values for each State. It is clear from Table 8 that the current boundaries and number of strata performs well but improvements can be made.

Table 8

The ratio of the variances using approximately optimum stratification technique to the variance under the current boundary values for the strata selected from Table 7 by State.

Technique	Number of Strata	Corn	Winter Wheat	Soybeans
		OHIO		
Dalenius-Hodges	4	0.781	0.676	0.699
	5	0.471	0.434	0.421
Durbin	4	0.806	0.724	0.721
	5	0.483	0.452	0.433
Ekman	4	0.753	0.656	0.638
	5	0.498	0.464	0.425
Sethi	4	1.048	0.925	0.945
	5	0.758	0.520	0.524
Eq. Agg.	4	1.045	0.942	0.934
	5	0.723	0.548	0.497
		KANSAS		
Dalenius-Hodges	3	1.133	1.316	1.282
	4	0.643	0.719	0.703
Durbin	3	1.047	1.236	1.266
	4	0.643	0.719	0.703
Ekman	3	1.133	1.316	1.282
	4	0.714	0.798	0.750
Sethi	3	0.896	1.556	1.451
	4	0.497	0.847	0.836
Eq. Agg.	3	0.818	1.604	1.698
	4	0.450	0.912	0.937

For Ohio Dalenius-Hodges and Durbin are the best stratification techniques with 5 strata. In Kansas 4 strata appears optimum but Dalenius-Hodges, Durbin, Sethi, and Equal Aggregate Output all perform at about the same level. Therefor to finally decide on the proper strata boundary values we compare livestock variances under the above stratification techniques to see if one technique results in the largest gain and/or smallest loss in precision in Table 9.

Table 9

The ratio of the variance using the approximate stratification technique to the variance under the current strata boundary values by State.

Dalenius	Cattle	Hogs
	Ohio (5 strata)	
Dalenius-Hodges	0.707	0.822
Durbin	0.718	0.861
	Kansas (4 strata)	
Dalenius-Hodges	0.342	0.461
Durbin	0.285	0.444
Sethi	0.195	0.936
Eq. Agg.	0.200	1.034

Examination of Table 9 for Ohio shows that for either technique any gains in one variable is offset by a loss in the other variable. From Appendix B, Table B3 for Ohio we would select Dalenius-Hodges with stratum boundary values 25%, 47.5%, 70% and 85%. Recall the current stratum boundary values are 15%, 50% and 75%. For ease of frame construction the boundary value 47.5% should be 50%. Table 10 gives the ratio of the variance for a boundary value of 45% and a boundary value of 50% to the variance for the optimum boundary values.

On the other hand, Table 9 shows in Kansas that Durbins technique provides the most gain for cattle and hogs. Those stratum boundary values are 35%, 62.5% and 82.5% from Appendix B, Table B4. Again to ease frame construction the possible boundary values are 35%, 60% or 65% and 80% or 85%.

From Table 10 we see that the boundary values should be 35%, 60% and 85%.

Table 10: Ratio of variance for easier frame construction to variance with optimum boundary values

Boundary Values	Corn	Winter Wheat	Soybeans
	Ohio		
25-45-70-85	1.017	0.720	1.024
25-50-70-85	0.991	0.710	0.998
	Kansas		
35-60-80	1.054	1.095	1.032
35-65-80	1.053	1.088	1.028
35-60-85	0.981	0.871	0.994
35-65-85	0.950	1.011	0.947

Summary and Recommendations

Three interrelated problems; choice of stratification variable, optimum number of strata, and optimum stratum boundary values, associated with stratified sampling were studied empirically using 1975 JES data for Ohio and Kansas. It was found that the land under cultivation, percent cultivated, should be retained as the stratification variable. In Ohio the optimum number of strata is 5, with boundary values 25%, 50%, 70% and 85%. In Kansas, four strata are optimum with boundary values 35%, 60% and 85%. The current stratum boundary values are 15%, 50% and 75%.

Motivation for this study was provided through a descriptive analysis of the properties of the second stage sampling units, segments, in Illinois, Kansas, Ohio and Minnesota. This descriptive analysis outlines a useful tool which can be repeatedly carried out in each State to indicate potential problems. Use of this type analysis would allow for the updating of a States' area frame on a "need" basis, rather than on a scheduled rotational basis. As the agriculture in a State changes the descriptive analysis would indicate when the stratification should be studied and possibly a new frame constructed.

The empirical analysis of estimated correlation coefficients indicates that in all four States percent cultivated is as good as any other variable for stratification. In fact, because percent cultivated has a definite range (0-100%) it is preferable to other stratification variables.

Recommendations based on this report. They are:

1) The descriptive analysis could be implemented in each State. Then restratification can be based on need rather than turn.
2) Percent cultivated should be retained as the stratification variable.
3) The analysis to determine the optimum number of strata and strata boundaries should be applied to each state.
4) Ohio should have stratum boundary values of 25%, 50%, 70% and 85%. Kansas should have stratum boundary values of 35%, 60% and 85%.
5) The problem of accuracy of classification into proper strata during frame construction should be studied. How difficult is it to construct a frame in Ohio with boundary values of 25%, 50%, 70% and 85%? Can it be done better, i.e., fewer misclassification, than using the current stratum boundary values? What about frame construction in Kansas where the boundary values are 35%, 60%, and 85%?
6) The techniques of optimum stratification should be applied to the list frame since these techniques assume the stratification variable is the variable to be estimated.

References

[1] Anderson, D.W., Kish, L., Cornell, R.G. "Quantifying Gains From Stratification for Optimum and Approximately Optimum Strata Using A Bivariate Normal Model", Tech. Report 4, Departmenr of Biostatistics, The Unversity of Michigan, (1975), Ann Arbor, Michigan 48104

[2] Cochran, W.G. "Comparison of Methods for Determining Stratum Boundaries", Bulletin of the International Statistical Institute, 38(2), Tokyo (1961), pp. 345-358.

[3] Cochran, W.G. Sampling Techniques, 2d. ed. John Wiley and Sons, Inc. New York, 1963

[4] Dalenius, T. "The Problem of Optimum Stratification", Skandinavisk Aktuarietidskrift, 33 (1950), pp. 203-213.

[5] Dalenius, T. Sampling in Sweden, Chapter 8, Almquist and Wiksell, Stockholm (1957)

[6] Dalenius, T. and Hodges, J.L., Jr. "The Choice of Stratification of Points", Skandinavisk Aktuarietdiskrift, (1958), pp. 198-203.

[7] Dalenius, T. and Hodges, J.L., Jr. "Minimum Variance Stratification", Journal of the American Statistical Association, 54 (1959), pp. 88-101.

[8] Durbin, J. "Review of 'Sampling in Sweden'", Journal of the Royal Statistical Society, 122 (1959), pp. 246-248.

[9] Ekman, G. "An Approximation Useful in Univariate Stratification", Annals of Mathematical Statistics, 30 (1959), pp. 219-229.

[10] Hansen, M.H., Hurwitz, W.N., Madow, W.G. Sample Survey Methods and Theory John Wiley and Sons, Inc. 1953.

[11] Hess, I., Sethi, V.K., and Balakrishnan, T.R. "Stratification: A Practical Investigation", Journal of the American Statistical Association, 61, (1966) pp. 74-90.

[12] Kpedekpo, G.M.K. "Recent Advances on Some Aspects of Stratified Sample Design. A Review of the Literature", Metrika, 20 (1973), pp. 54-64.

[13] Mahalanobis, P.C. "Some Aspects of the Design of Sample Surveys. Sankhya 12 (1952) pp. 1-7.

[14] Raj. D. The Design of Sample Surveys, McGraw Hill, New York. 1972

[15] Sethi, V.K. "A Note on Optimum Stratification of Populations for Estimating the Population Means", Australian Journal of Statistics, 5 (1963) pp. 20-33.

Appendix A

Typical plots of segments either conforming

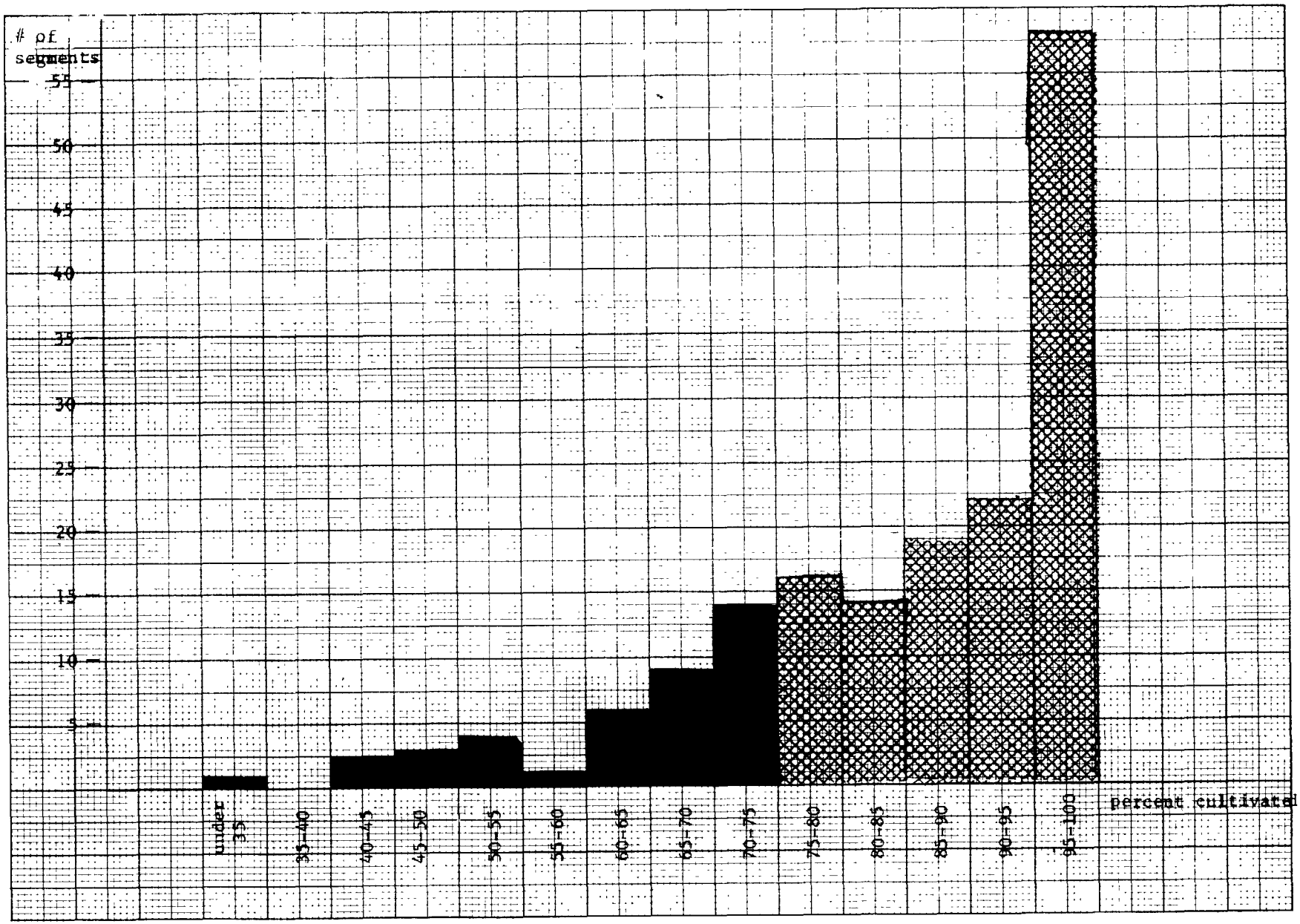
or

not conforming with stratum definitions

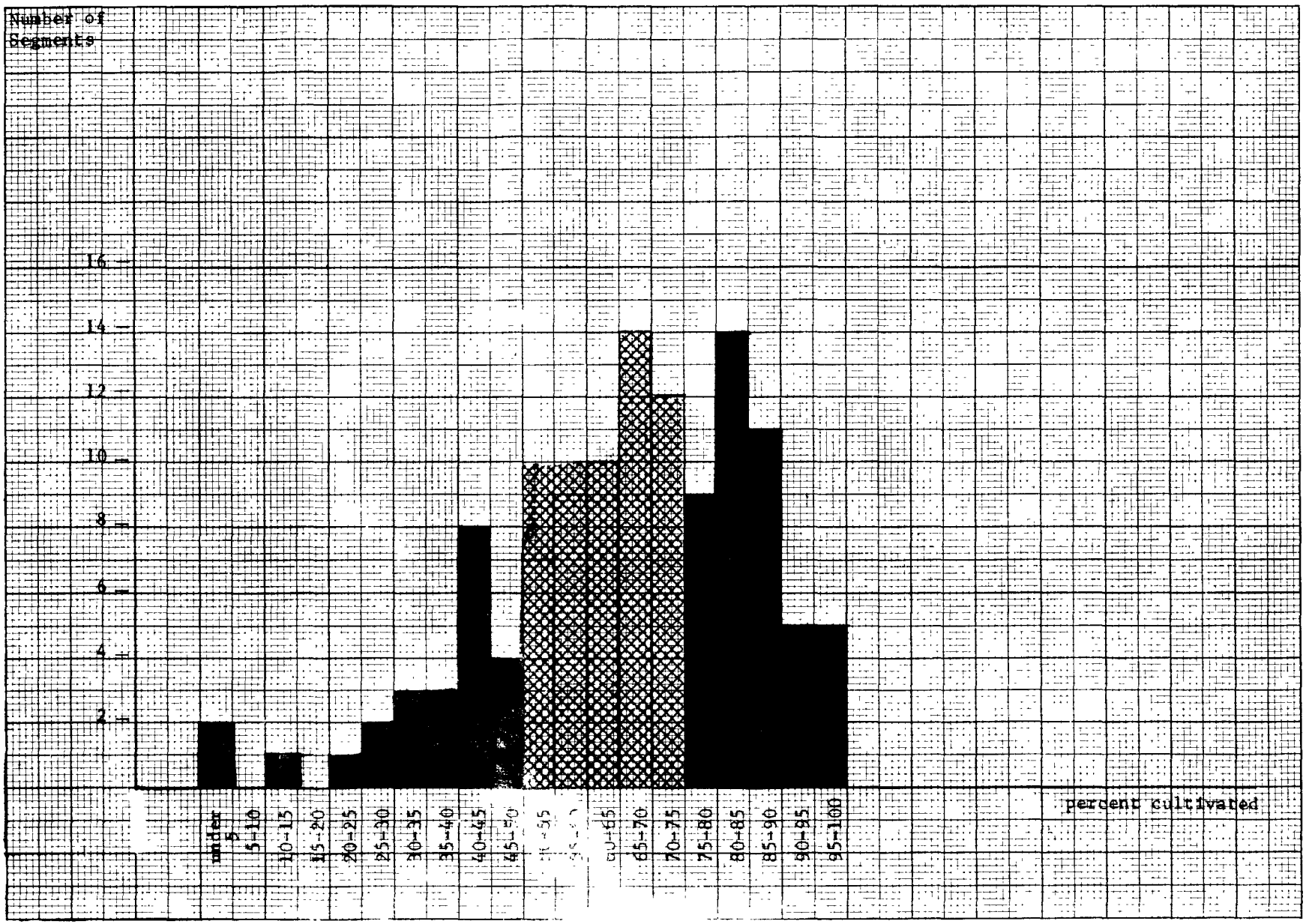
conforms with stratum definitions

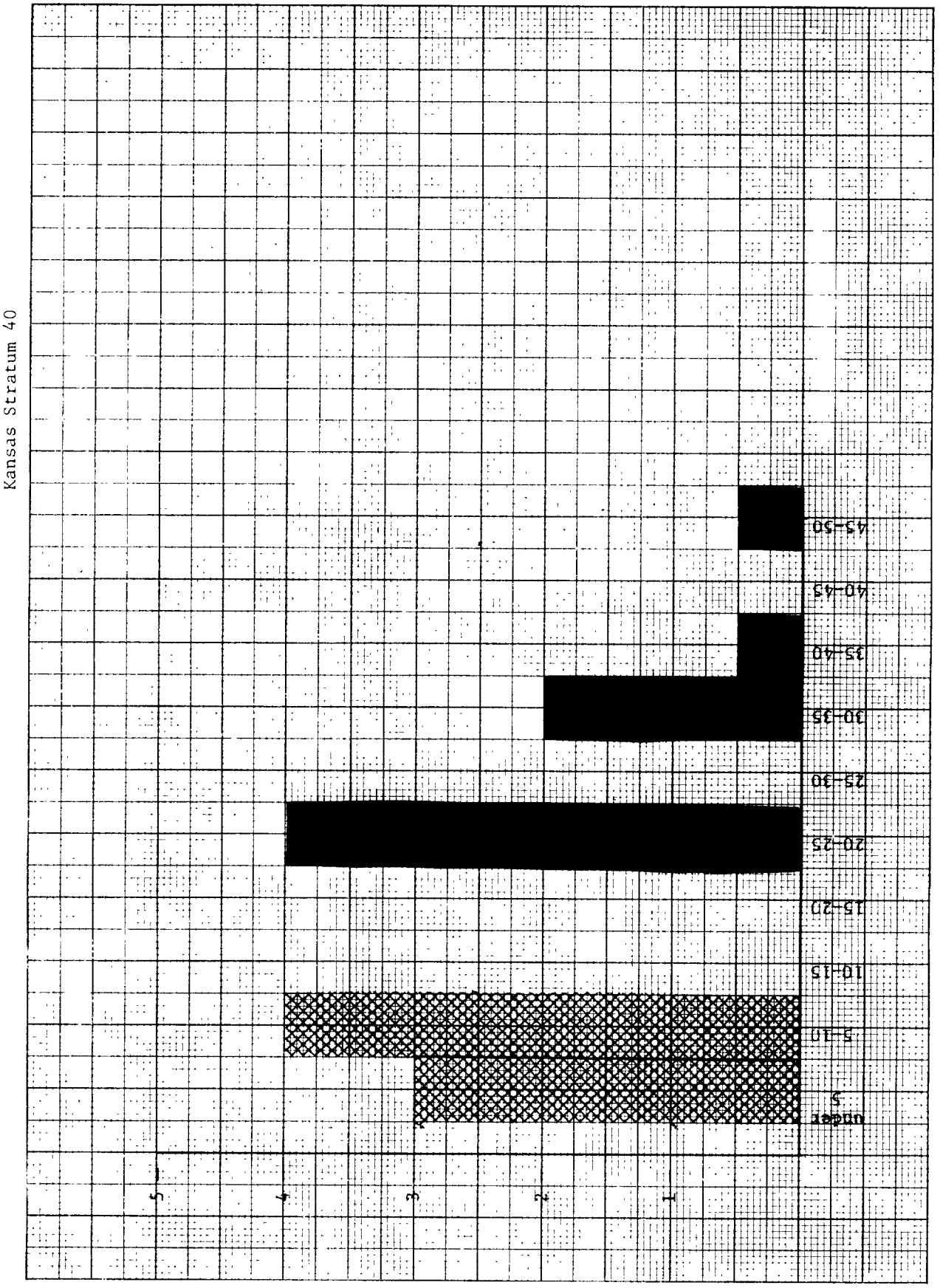
does not conform with stratum definitions

Kansas - Stratum 11

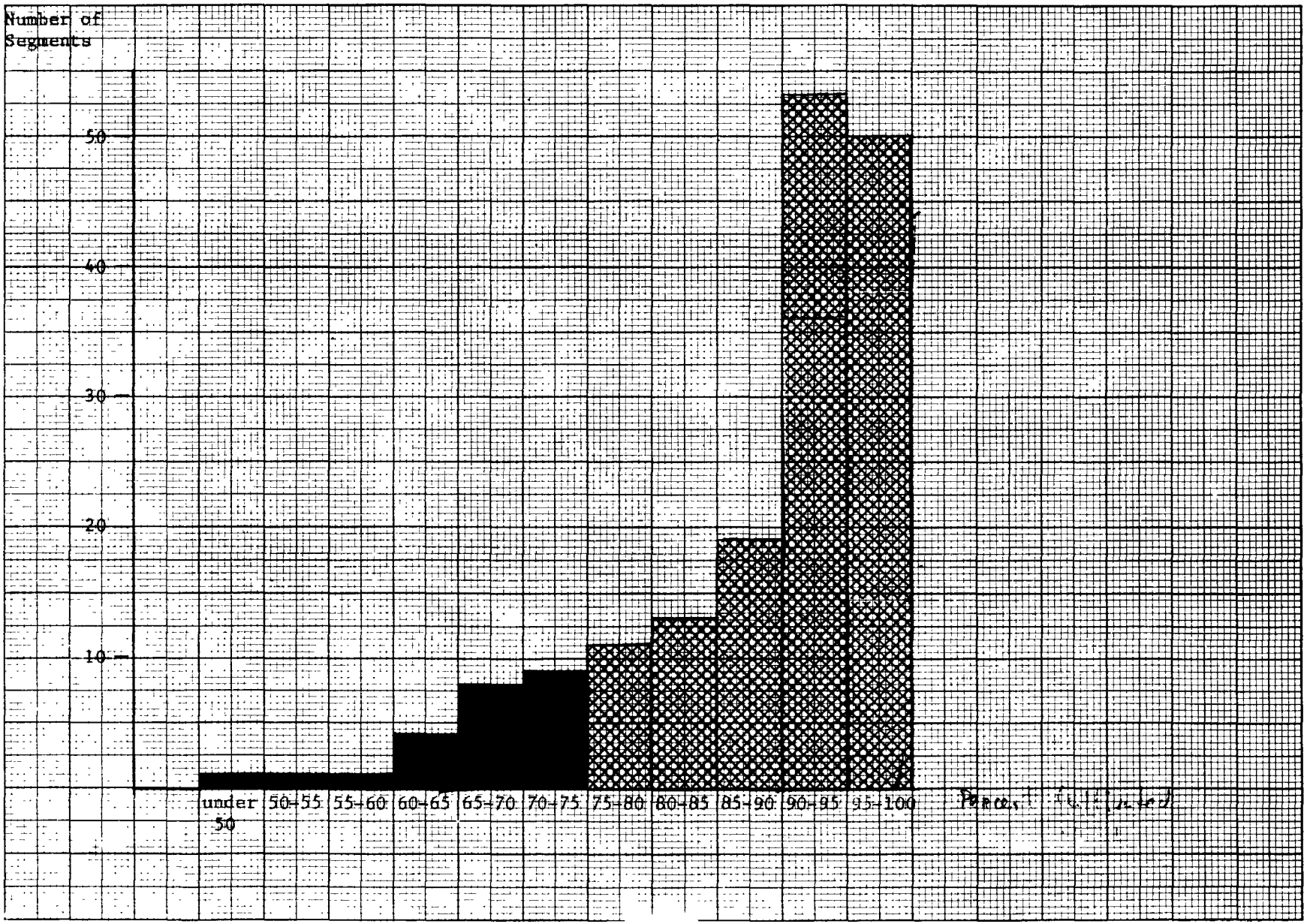


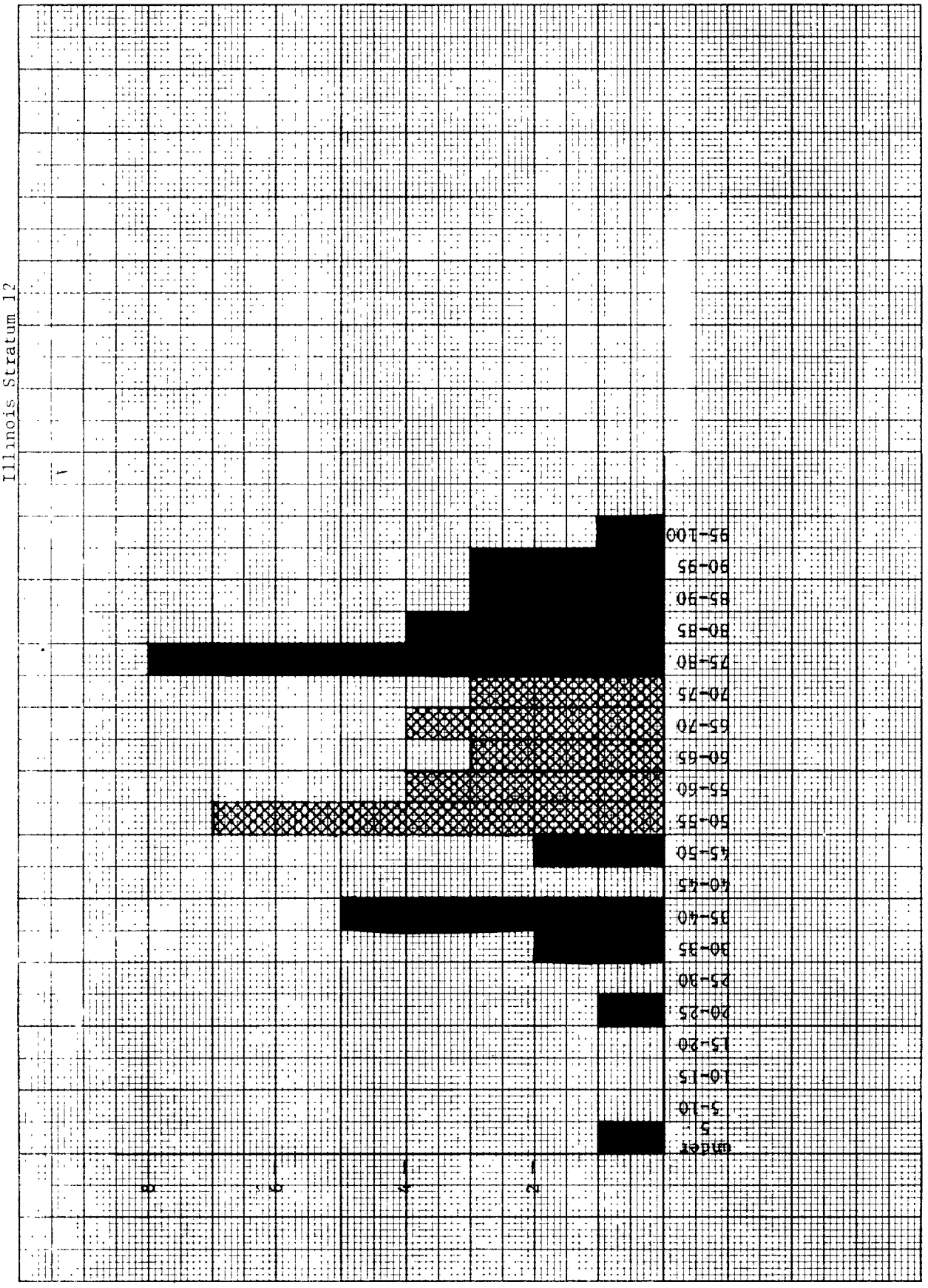
Kansas - Stratum 12



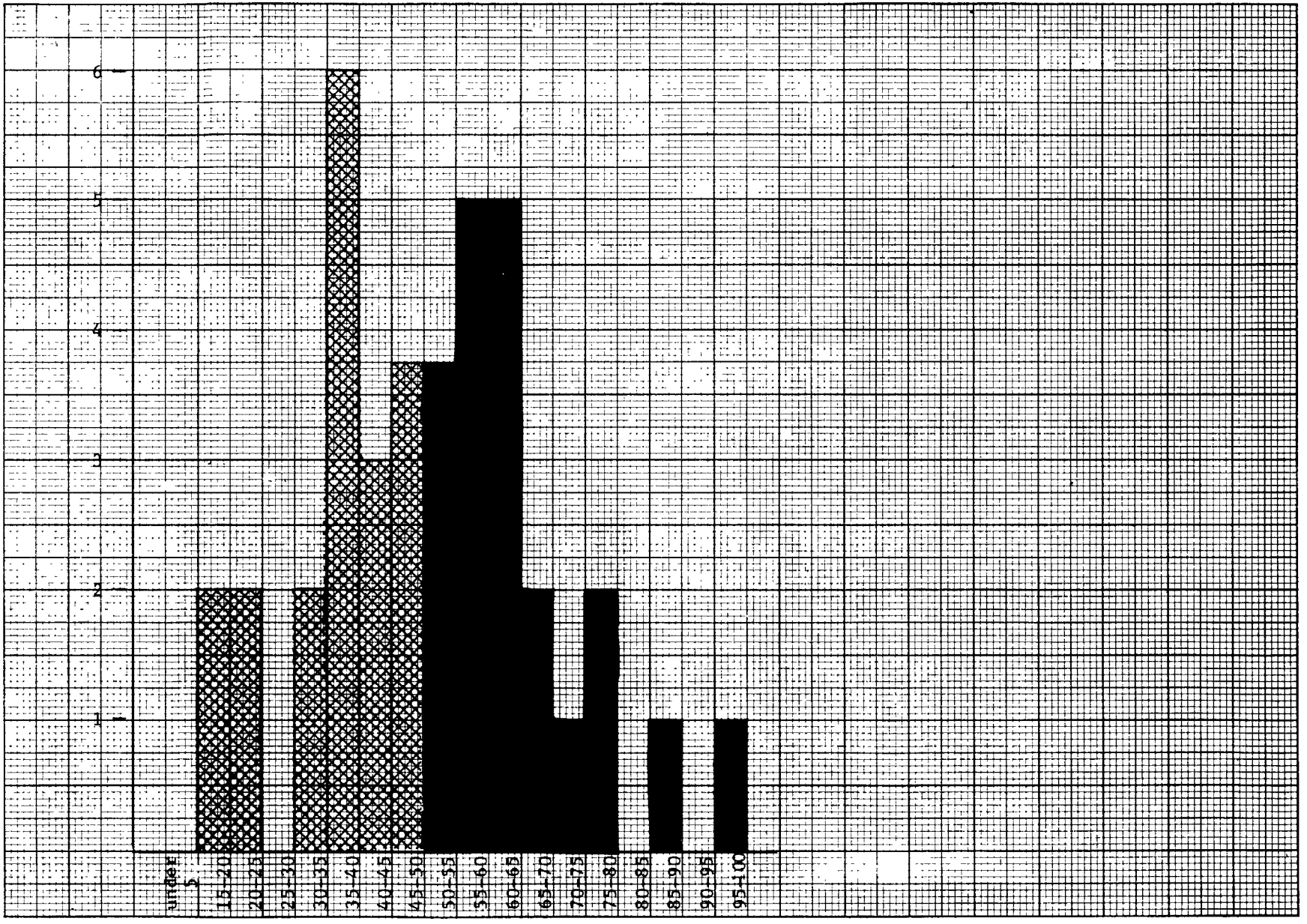

Kansas Stratum 40
45-50
40-45
35-40
30-35
25-30
20-25
15-20
10-15
5-10
under 5
5
4
3
2
1

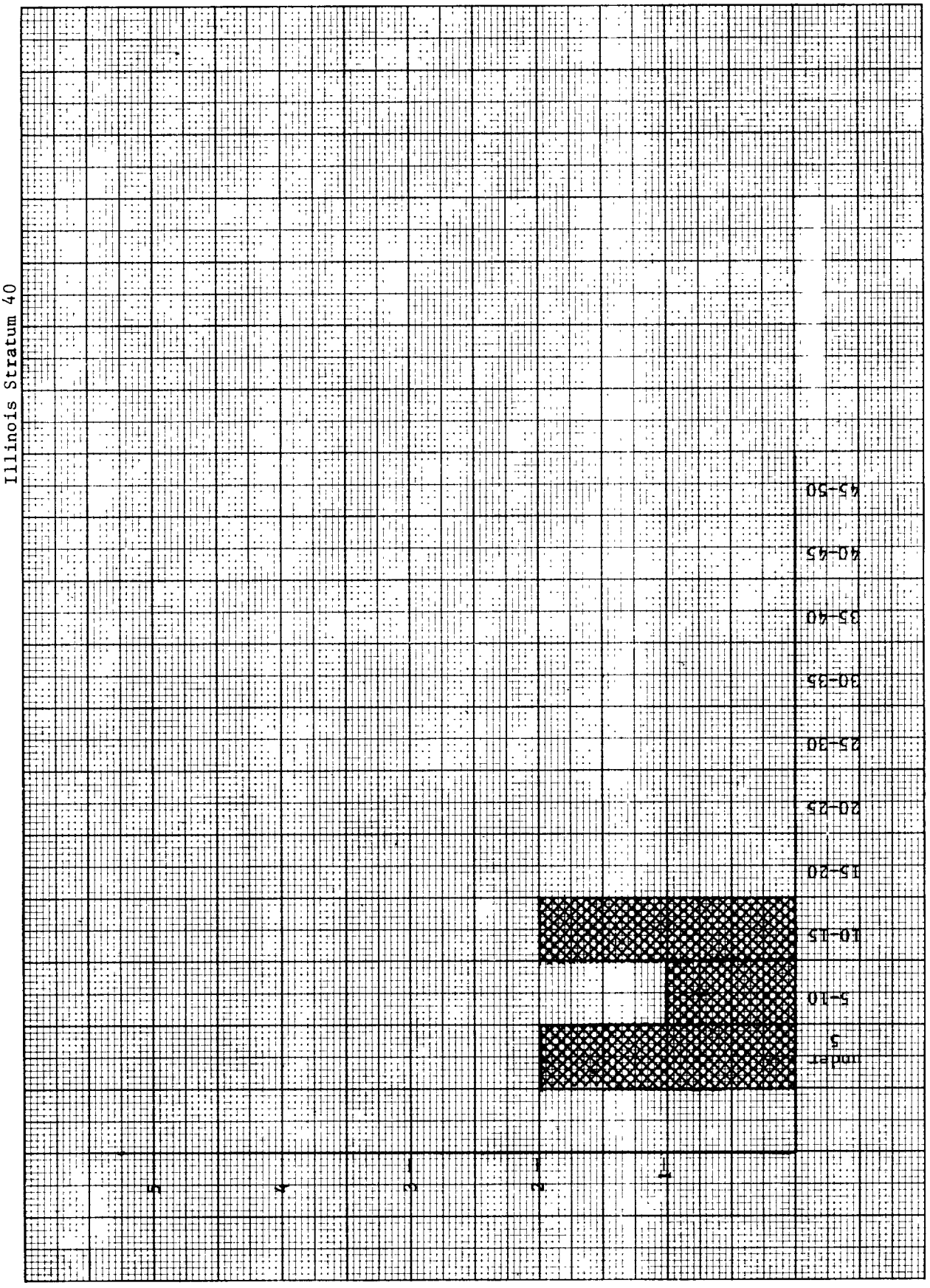
Illinois Stratum = 11

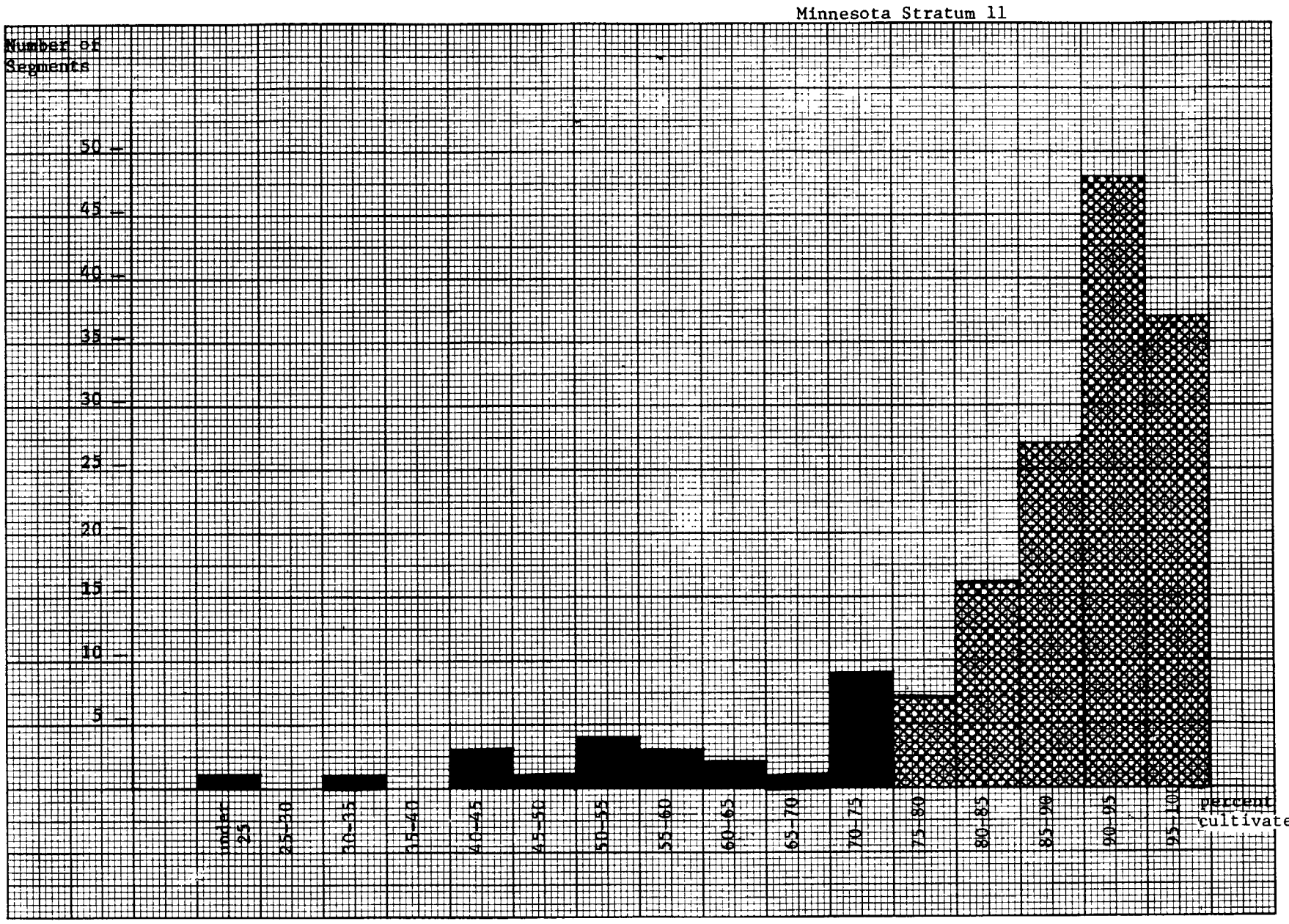


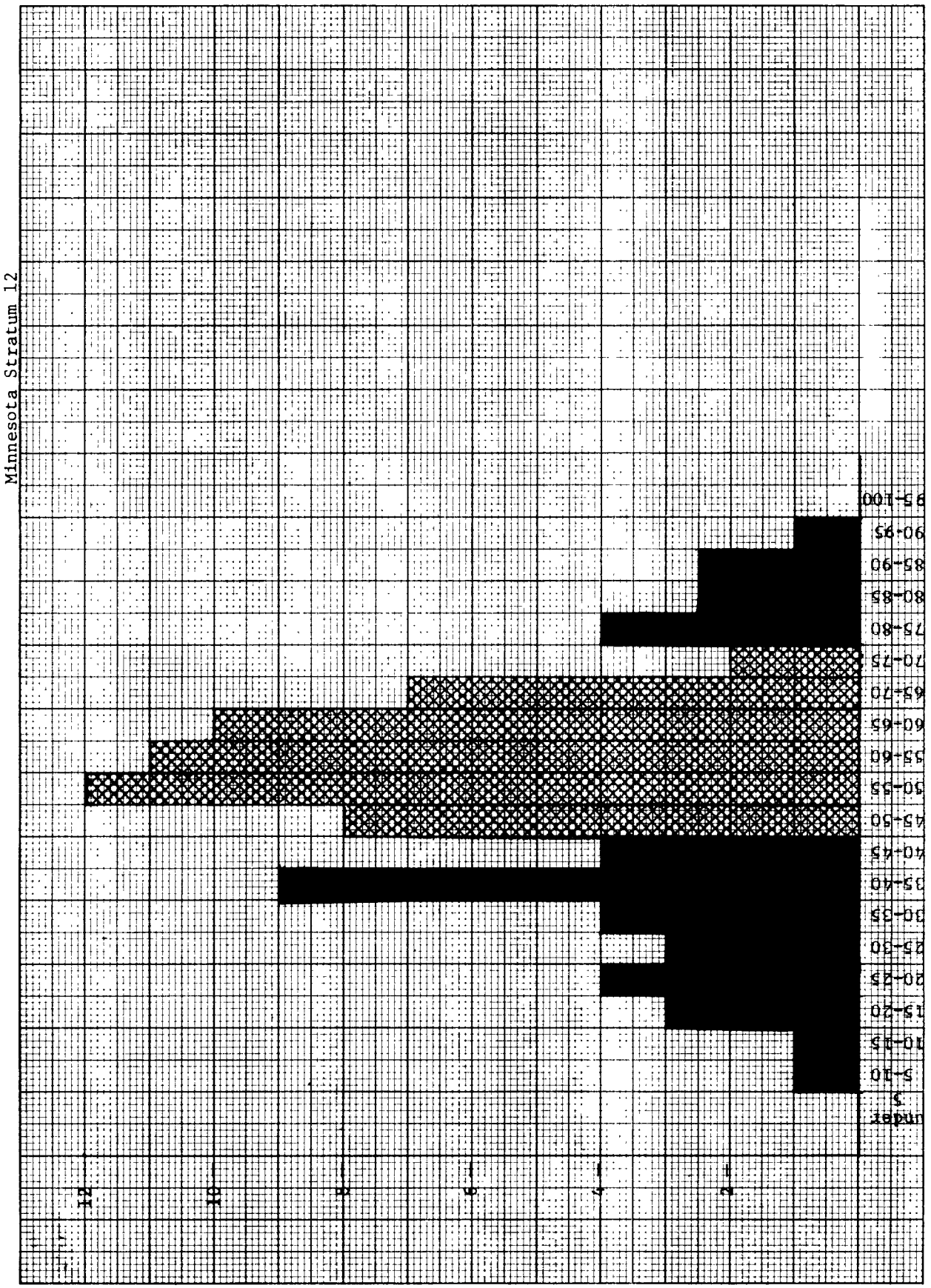

Illinois Stratum 12
95-100
90-95
85-90
80-85
75-80
70-75
65-70
60-65
55-60
50-55
45-50
40-45
35-40
30-35
25-30
20-25
15-20
10-15
5-10
under 5
8
6
4
2

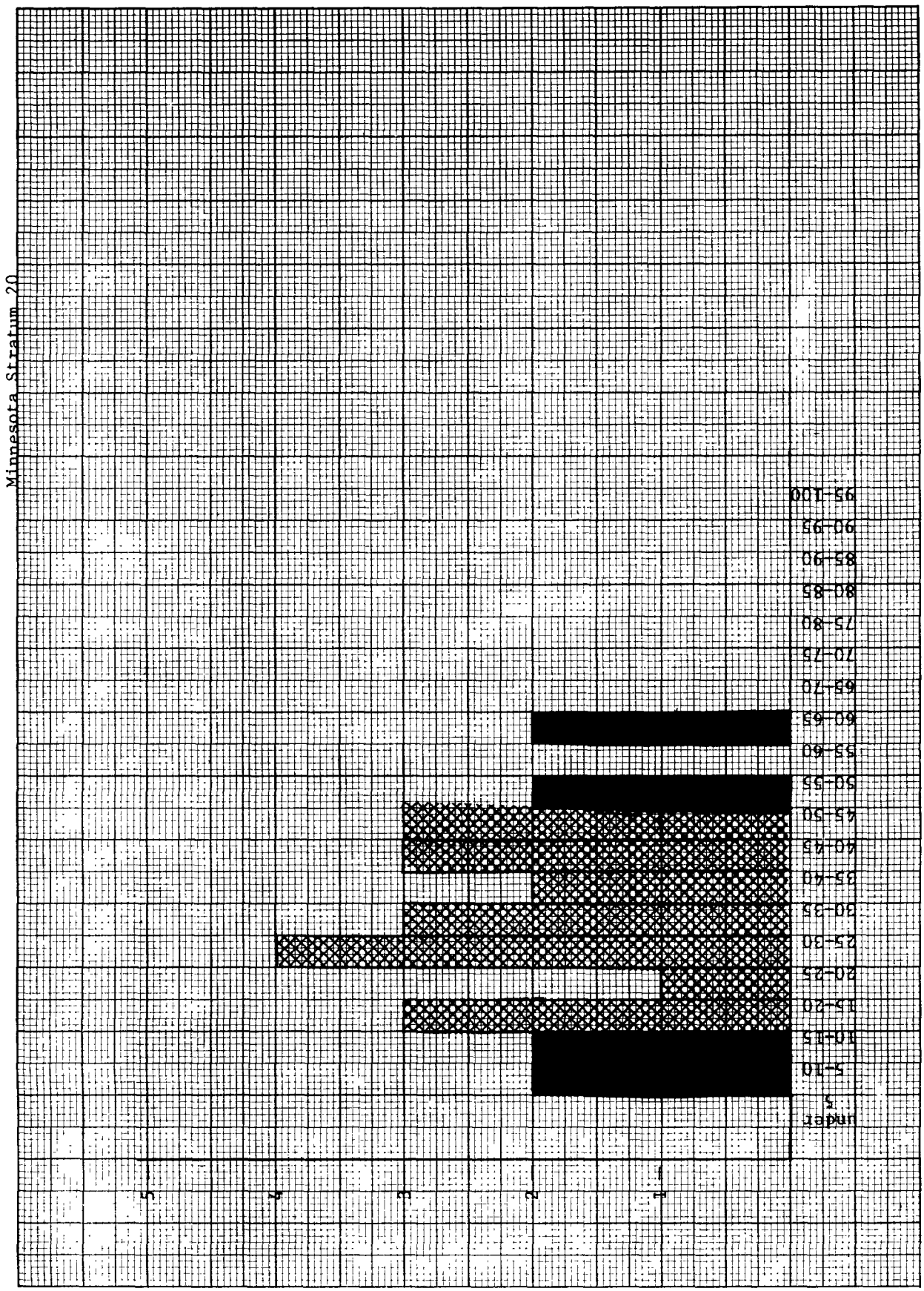
Illinois Stratum 20

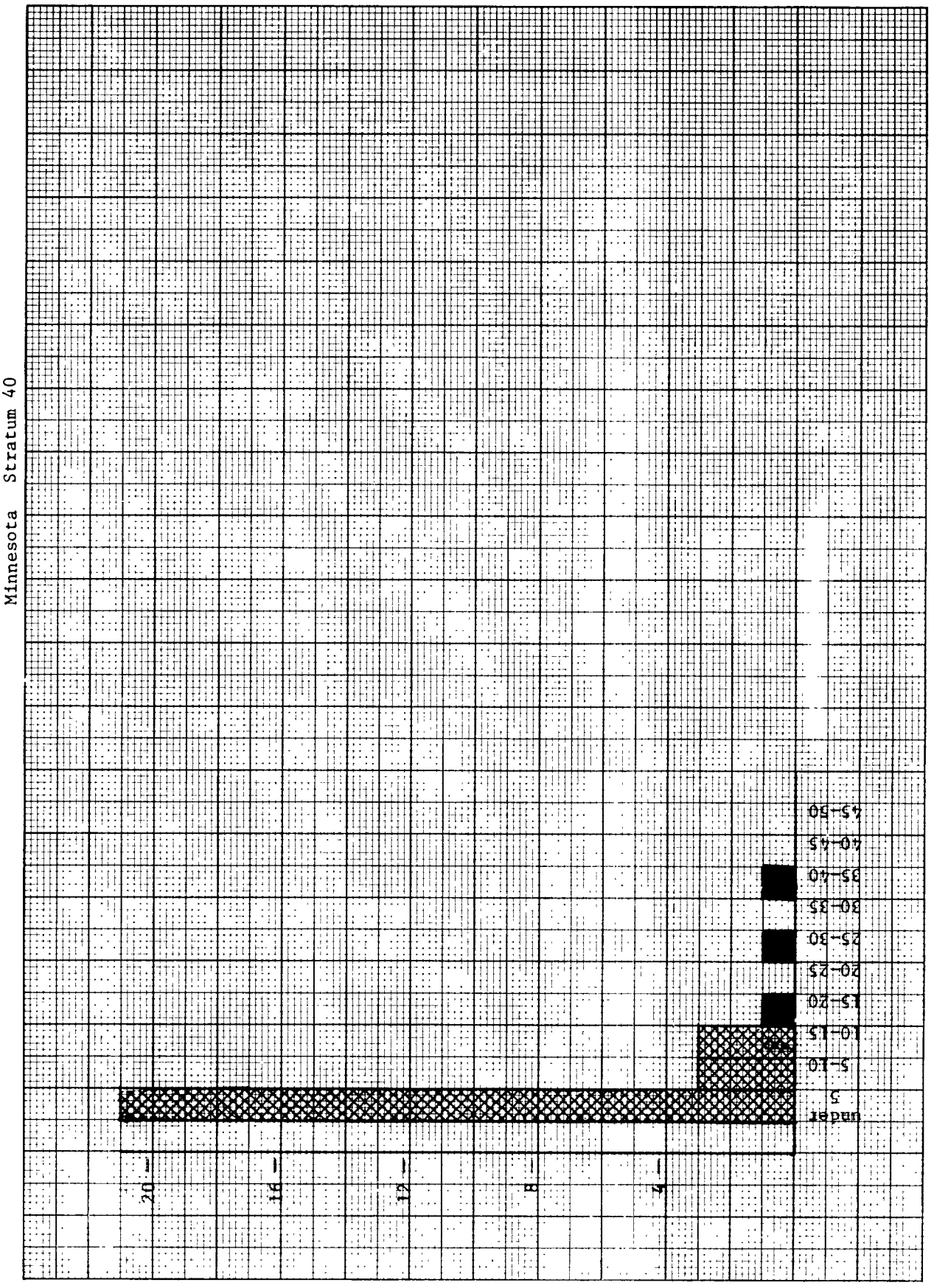


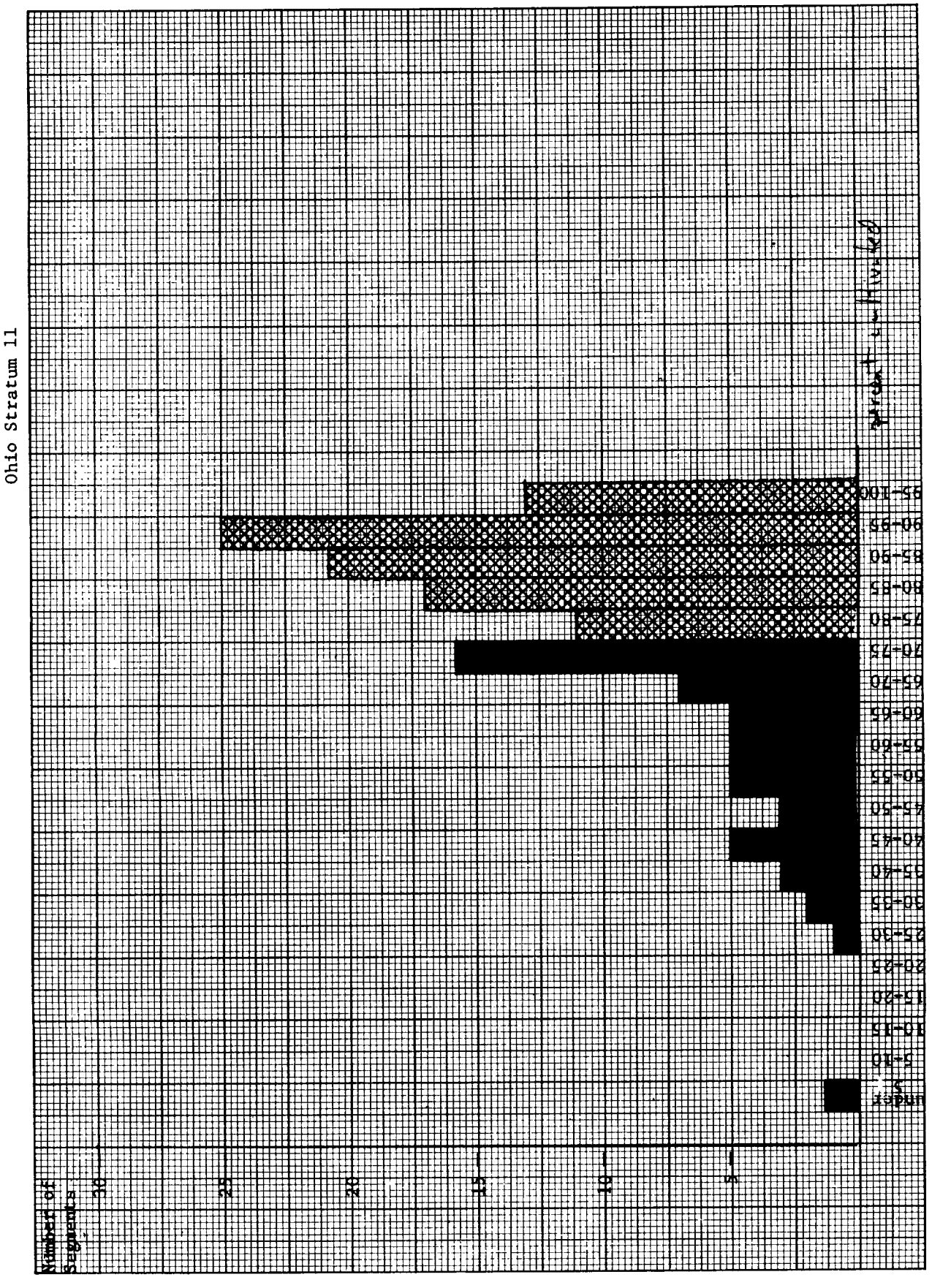

Illinois Stratum 40
45-50
40-45
35-40
30-35
25-30
20-25
15-20
10-15
5-10
Under 5
1
2
3
4
5

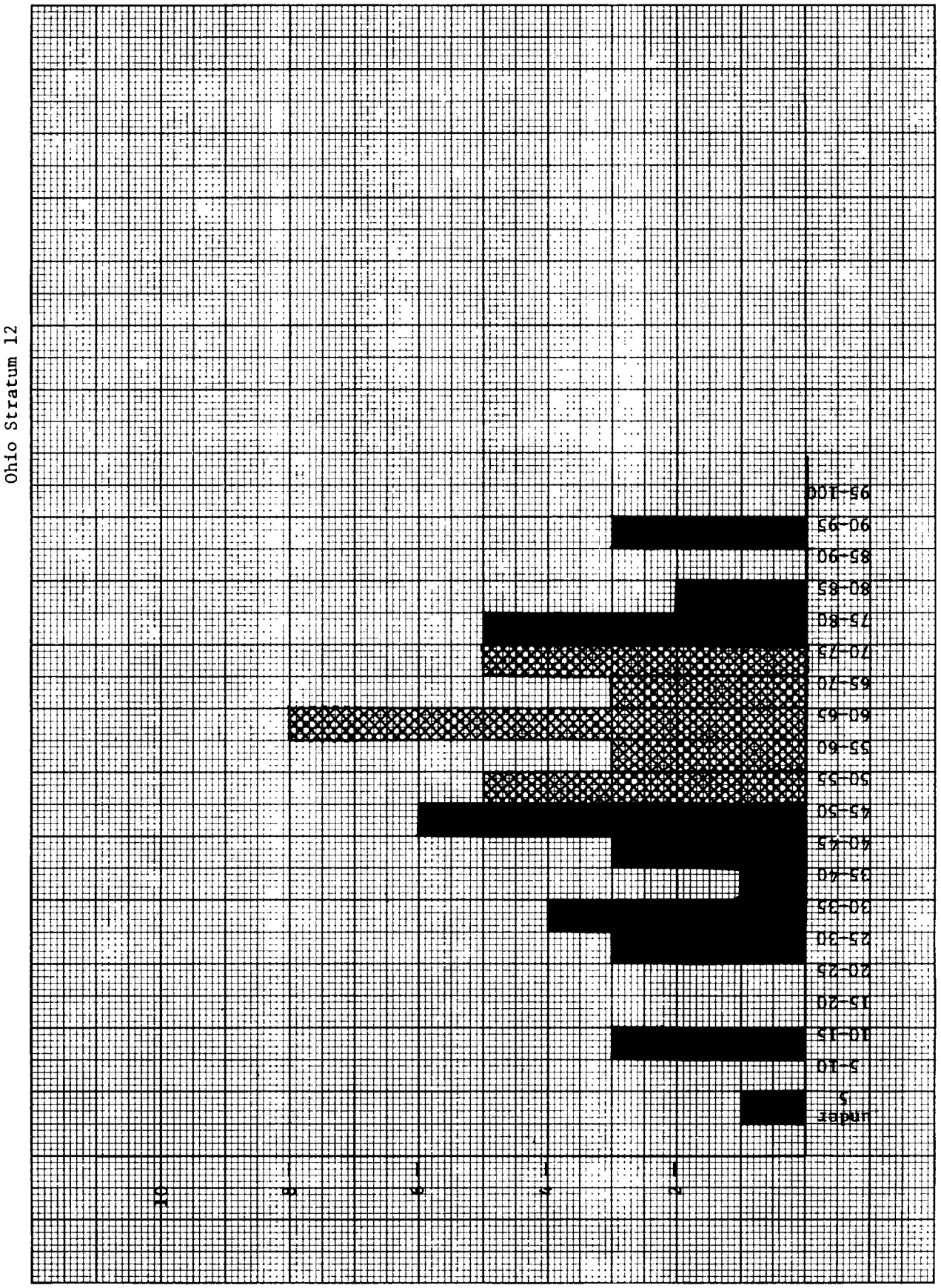

Minnesota Stratum 11
Number of Segments
50
45
40
35
30
25
20
15
10
5
under 25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
80-85
85-90
90-95
95-100
percent cultivated

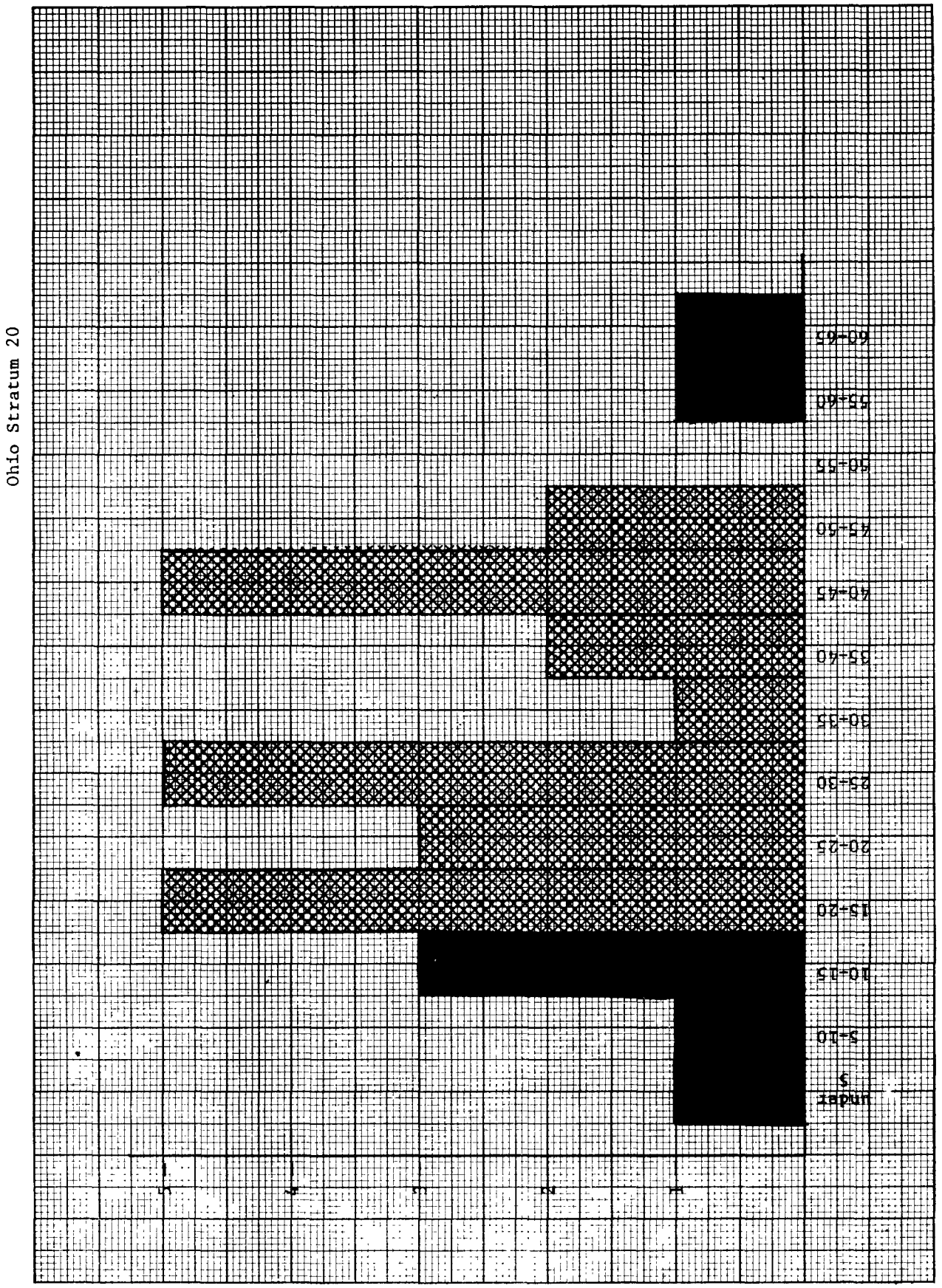

Minnesota Stratum 12
95-100
90-95
85-90
80-85
75-80
70-75
65-70
60-65
55-60
50-55
45-50
40-45
35-40
30-35
25-30
20-25
15-20
10-15
5-10
under 5
12
10
8
6
4
2

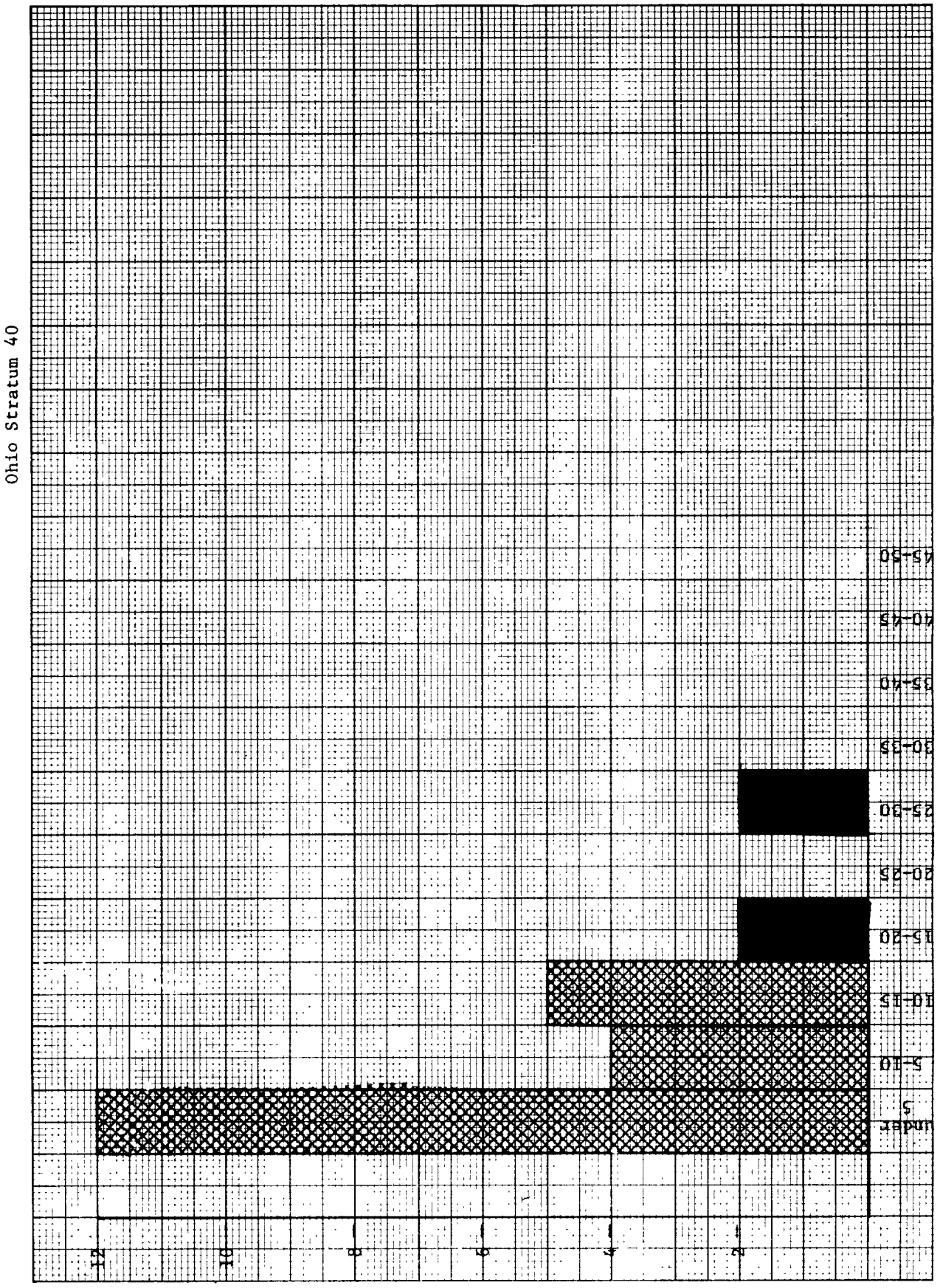

Minnesota Stratum 20
under 5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
80-85
85-90
90-95
95-100
1
2
3
4
5


Minnesota Stratum 40
under 5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
4
8
12
16
20


Ohio Stratum 11
Number of Segments
30
25
20
15
10
5
95-100
90-95
85-90
80-85
75-80
70-75
65-70
60-65
55-60
50-55
45-50
40-45
35-40
30-35
25-30
20-25
15-20
10-15
5-10
under 5


Ohio Stratum 12
95-100
90-95
85-90
80-85
75-80
70-75
65-70
60-65
55-60
50-55
45-50
40-45
35-40
30-35
25-30
20-25
15-20
10-15
5-10
under 5
10
8
6
4
2


Ohio Stratum 20
under 5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65


Ohio Stratum 40
45-50
40-45
35-40
30-35
25-30
20-25
15-20
10-15
5-10
under 5
12
10
8
6
4
2

Appendix B

A technical description of each stratification technique,

a numeral example for each technique

and

the stratum boundary values for Ohio and Kansas for each technique

1. Dalenius and Hodges - This technique is an approximate solution to the problem of minimizing the variance of the mean over all strata, i.e.,

$$\min[\mathrm{Var}(\bar{y}_{st})] = \min\left[\frac{1}{n}\left(\sum_{h=1}^{L} W_h S_h\right)^2 - \frac{1}{n}\sum_{h=1}^{L} W_h S_h^2\right],$$

where

$$S_h^2 = \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2/(N_h - 1),$$

h=1, 2,..., L and L is number of strata considered. Ignoring the fpc, the problem is to minimize $\Sigma\, W_h\, S_h$. An exact solution os obtainable by differentiating the above formula [4], however this solution is impractical because all the parameters in the solution depend on y_{hi}. Dalenius and Hodges [6] obtain approximate solutions by defining.

$$Z(y) = \int_{yo}^{y} \sqrt{f\ (t)}\ dt, = f(y)$$

where f(y) represents the frequency of occurrence of the stratification variable. If the strata are numerous and narrow, f(y) should be rectangular in each stratum, i.e., constant. Hence

$$W_h = \int_{y_{h-1}}^{y_h} f\ (t)\ dt \simeq f_h\ (y_h - y_{h-1})$$

$$S_h \simeq (y_h - y_{h-1})/\sqrt{12}$$

$$Z_h - Z_{h-1} = \int_{y_{h-1}}^{y_h} \sqrt{f\ (t)} \simeq \sqrt{f_h}\ (y_h - y_{h-1})$$

where f_h is the "constant" value of f(y) in stratum h. So

$$\sqrt{12}\ \sum_{h=1}^{L} W_h S_h \simeq \sum_{h=1}^{L} (Z_h - Z_{h-1})^2.$$

Since $Z_1 - Z_0$ is fixed, it is easy to verify that $\sum_{h=1}^{L} (Z_h - Z_{h-1})^2$ is minimized by taking $(Z_h - Z_{h-1})$ constant.

Using these approximate solutions the optimum boundaries are obtained by forming equal intervals of the cumulative of the square root of the frequencies of the stratification variable, i.e., equal intervals of cum$\sqrt{f}$. In practical applications you must divide the range of values for the stratification variable of interest into k equal intervals (for our study we set k=40). For each of k intervals, determine the frequency of occurrence of the stratification variable within each interval. Then calculate cum$\sqrt{f}$, giving you k values of the cum$\sqrt{f}$. Divide the k-th cum$\sqrt{f}$, call it $\text{cum}_k\sqrt{f}$, by the number of strata (L) you wish to use (For this study we looked at L=2,3,..,10). and the upper bounds of the intervals of the corresponding values of the cum$\sqrt{f}$ that correspond to $\frac{1}{L}\ \text{cum}_k\sqrt{f}$, $\frac{2}{L}\ \text{cum}_k\sqrt{f}$, ..., $\frac{L-1}{L}\ \text{cum}_k\sqrt{f}$, $\frac{L}{L}\ \text{cum}_k\sqrt{f}$. We have assumed k equal intervals of width x but if some intervals are of width dx then the value of $\sqrt{f}$ must be multiplied by $\sqrt{d}$ and then take the cumulative over the interval as before.

2. Durbin - This method basically amounts to forming strata by taking equal areas under a frequency distribution with density half-way between the original and rectangular distribution, i.e., form the L strata by taking equal intervals on the cumulative of (r(y) + f(y))/2, where r(y) is the rectangular distribution and f(y) is the frequency distribution. r(y) is obtained by dividing the distribution function by the range of values in the distribution i.e., r(y) = $\frac{F(y_L)}{y_L - yo}$. In practical applications you must divide the variable to be stratified into k equal intervals. Find the k cum r values by calculating $\frac{1}{k}$ cum f, $\frac{2}{k}$ cum f, ..., $\frac{k-1}{k}$ cum f, $\frac{k}{k}$ cum f. Similarly obtain the k cum f, values. We obtain the cum (r+f) by adding

corresponding values of cum r and cum f. Stratum boundaries are calculated by finding values of cum (r+f) divided by L, say cum_k (r+f). This basically amounts to finding the values of cum (r+f) closely corresponding to $\frac{1}{L}$ cum_k (r+f), $\frac{2}{L}$ cum_k (r+f),..., $\frac{L-1}{L}$ cum_k (r+f), $\frac{L}{L}$ cum_k (r+f).

3. Ekman - For this technique we try to equalize the product of the cumulative frequency within the stratum, W_h, and the width of the stratum, $y_h - y_{h-1}$, i.e. make $W_h (y_h - y_{h-1})$ a constant, where y_h, h - 1, 2, ..., L, is the upper bound of our interval corresponding to W_h for a particular stratum. This technique is a little different from the previous three in that $\Sigma_h W_h (y_h - y_{h-1})$ changes with different values of L. In practice we once again create k equal intervals of our stratification variable and obtain the corresponding frequencies, f, and the k values for the cum f. Ekman proposes that we compute Q - $(cum_k - f) (y_1 - y_0)$, where $y_1 - y_0$ is the range of values of the stratification variable. Then equalize $W_h (y_h - y_{h-1})$ with the value $\frac{Q}{L^2}$. This implies we must find the right combination of W_h and $y_h - y_{h-1}$ such that its product is $\frac{Q}{L^2}$.
4. Sethi - This technique tries to equalize the probability of falling within a stratum, i.e., 100/L = percent probability of falling within each region. Since stratification variable has tabulated frequencies we can approximate the distribution of the startification variable by constructing a histogram. The shape of the distribution is the only important aspect for stratification as far as Sethi is concerned. In practice one obtains a histogram of the stratification variable and determines what type of distribution it follows. (A close

approximation to the distribution is adequate.) Then determine the boundaries such that there are equal probabilities of falling within each stratum.

5. Equal Aggregate Output - Mahalanobis [13] and then Hansen, Hurwitz, and Madow [10] developed this technique to optimize strata boundaries. This method calls for the strata to be constructed such that $W_h \mu_h$ is constant in all strata, where W_h is as in Ekmans method and μ_h is approximated by the midpoint of the intervals of the stratification variable. The equivalent to $W_h \mu_h$ being constant in all strata is to have the cumulative of $f_h \mu_h$, cum $(f_h \mu_h)$ being constant in all strata. Practically speaking we obtain the k equal intervals of the stratification variable. Find the frequencies corresponding to the intervals and the k midpoints of these intervals also. Compute the k cum $(f\mu)$'s and then divide the k-th cum $(f\mu)$ by L, the number of strata desired, calling it $\text{cum}_k\ (f\mu)$. We obtain the upper bounds of the intervals corresponding to the values of cum $(f\mu)$ that come closest to $\frac{1}{L}\ \text{cum}_k\ (f\mu)$, $\frac{2}{L}\ \text{cum}_k\ (f\mu)$, ..., $\frac{L-1}{L}\ \text{cum}_k\ (f\mu)$, $\frac{L}{L}\ \text{cum}_k\ (f\mu)$. In Appendix B an application of the rules to a numerical example is given and the strata boundary values tabled.

The method of applying the rules to determine the stratum boundary values will be illustrated for the data as shown in Table 1. For compactness and illustration boundary values will be calculated for L=4 and intervals of k=40 for the stratification variable of percent of land cultivated in Ohio.

1. Dalenius and Hodges - For each interval the frequency of occurrence (f_h) was determined and the 40 values of the $\text{cum}\sqrt{f_h}$ were calculated

The 40-th value of the cum$\sqrt{f_h}$ is 93.8. Ideally we want the 4 strata boundaries to correspond to

$\frac{1}{4}$ cum$\sqrt{f_h}$ = 23.5

$\frac{2}{4}$ cum$\sqrt{f_h}$ = 47

$\frac{3}{4}$ cum$\sqrt{f_h}$ = 70.5 and

$\frac{4}{4}$ cum$\sqrt{f_h}$ = 93.8.

Searching in the cum$\sqrt{f_h}$ column for the values that are closest to the ones calculated we obtain the following boundaries

	STRATUM			
	0 - 30%	30 - 60%	60 - 80%	80 - 100%
cum f_h	23.4	47.5	69.2	93.8

To obtain values of cum$\sqrt{f_h}$ that correspond more closely to the calculated values a finer subdivision of the data would be required say 60 intervals or more.

2. Durbin - To obtain values for the rectangular distribution divide the total number of observations by the number of class intervals. This yields a value of $\frac{252}{40}$ = 6.3. The cumulative of the rectangular distribution is shown in column 7 of Table 8. The cumulative of the frequencies, cum f_h, is in column 6 of Table 2. The sum of these two values yields the cum $(f_h + r_h)$. The 40-th value of the cum $(f_h + r_h)$ is 504.0. Ideally the 4 strata boundaries should correspond to

$\frac{1}{4}$ cum $(f_h + r_h)$ = 126.0

$\frac{2}{4}$ cum $(f_h + r_h)$ = 252.0

TABLE B1: CALCULATION OF THE STRATUM BOUNDARIES
BY THE FIVE METHODS FOR L=4

Variable: Percentage of land cultivated in Ohio (1975)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
%	Midpoint of Class Interval μ_h (coded)	Freq, of Occurance f_h	cum $\sqrt{f}$	cum $\sqrt{f_h \mu_h}$	cum f_h	cum r_h	cum(f_h+r_h)
0- 2.5	1.25(1)	11	3.3	x11	x11	6.3	17.3
2.5- 5.0	3.75(3)	5	5.5	x26	16	12.6	28.6
5.0- 7.5	6.25(5)	6	8.0	56	22	18.9	40.9
7.5-10.0	8.75(7)	0	8.0	56	22	25.2	47.2
10.0-12.5	11.27(9)	6	10.4	110	28	31.5	59.5
12.5-15.0	13.75(11)	5	12.6	165	33	37.8	70.8
15.0-17.5	16.25(13)	3	14.4	204	36	44.1	80.1
17.5-20.0	18.75(15)	4	16.4	264	40	50.4	90.4
20.0-22.5	21.25(17)	2	17.8	298	42	56.7	98.7
22.5-25.0	23.75(19)	1	18.8	317	43	63.0	106.0
25.0-27.5	26.25(21)	7	21.4	464	50	69.3	119.3
27.5-30.0	28.75(23)	4	23.4	556	54	75.6	129.6
30.0-32.5	31.25(25)	5	25.7	681	59	81.9	140.9
32.5-35.0	33.75(27)	2	27.1	735	61	88.2	149.2
35.0-37.5	36.25(29)	0	27.1	735	61	94.5	155.5
37.5-40.0	38.75(31)	6	29.5	921	67	100.8	167.8
40.0-42.5	41.25(33)	8	32.4	1185	75	107.1	182.1
42.5-45.0	43.75(35)	5	34.6	1360	80	113.4	193.4
45.0-47.5	46.25(37)	7	37.2	1619	87	119.7	206.7
47.5-50.0	48.75(34)	4	39.2	1775	91	126.0	217.0
50.0-52.5	51.25(41)	4	41.2	1939	95	132.3	227.3
52.5-55.0	53.75(43)	6	43.7	2197	101	138.6	239.6
55.0-57.5	56.25(45)	1	44.7	2242	102	144.9	246.9
57.5-60.0	58.75(47)	8	47.5	2618	110	151.2	261.2
60.0-62.5	61.25(44)	9	50.5	3059	119	157.5	276.5
62.5-65.0	63.75(51)	5	52.8	3314	124	163.8	287.8
65.0-67.5	66.25(53)	6	55.2	3632	130	170.1	300.1
67.5-70.0	68.75(55)	4	57.2	3852	134	176.4	310.4
70.0-72.5	71.25(57)	12	60.7	4536	146	182.7	328.7
72.5-75.0	73.75(59)	9	63.7	5067	155	189.0	344.0
75.0-77.5	76.25(61)	11	67.0	5738	166	195.3	361.3
77.5-80.0	78.75(63)	5	69.2	6053	171	201.6	372.6
80.0-82.5	81.25(65)	12	72.7	6833	183	207.9	390.9
82.5-85.0	83.75(69)	7	75.3	7316	190	214.2	404.2
85.0-87.5	86.25(69)	11	78.6	8075	201	220.5	421.5
87.5-90.0	88.75(71)	10	81.8	8785	211	226.8	437.8
90.0-92.5	91.25(73)	22	86.5	10391	233	233.1	466.1
92.5-95.0	93.75(75)	6	88.9	10841	239	239.4	478.4
95.0-97.5	96.25(77)	10	92.1	11611	249	245.7	494.7
97.5-100.0	98.75(79)	3	93.8	11848	252	252.0	504.0
Total		252	93.8	11,848	252	252	504

$\frac{3}{4}$ cum $(f_h + r_h) = 378.0$ and

$\frac{4}{4}$ cum $(f_h + r_h) = 504.0$

Locating values in the cum $(f_h + r_h)$ column that correspond to these calculated values we obtain:

	STRATUM			
	1	2	3	4
cu	0 - 30%	30 - 57.5%	57.5 - 80%	80 - 100%
cum $(f_h + r_h)$	129.6	246.9	372.6	504.0

3. Ekman - As stated previously, we cannot find a single variable to partition so that the boundaries can be found rather easily. We must equalize $W_h\ (y_h - y_{h-1})$, which is the product of two variables. The approximate value that this product must come close to for each boundary is $Q/L^2 = \text{cum}_k\ (f)\ (y_1 - y_0)/L^2$

$= (252)\ (100 - 0)/4^2$

$= 1575.$

We must now find the right combination of W_h and $y_h - y_{h-1}$ such that its product is Q/L^2.

As an example of this calculation look at the boundary for stratum 1. If the upper bound is 30% we obtain $W_h = 54$, $y_h = 30$ and $y_{h-1} = 0$. This yields a product of 1620, which is as close to 1575 as you can get without finer subdivisions. For stratum 2 look at the interval from 30 - 60%. For this stratum $W_h = 110 - 54$, $y_h = 60$, $y_{h-1} = 30$ yielding a product of 1680. Similarly we obtain the next 2 strata The results are:

	STRATUM			
	1	2	3	4
	0 - 30%	30 - 60%	60 - 82.5%	82.5 - 100%
$W_h (y_h - y_{h-1})$	1620	1680	1642.5	1207.5

4. Sethi - For this method a histogram was plotted for the percent of land cultivated in Ohio (Figure 1). The plot could be approximated by a parabola (positive absicca). It should be noted that we deviate here. slightly from Sethi's original work. He used a normal, gamma or beta distribution as his standard distribution, Therefore the density function used was

$$f(x) = 3 \times 10^{-6} x^2, \; x \in [0,100].$$

For L=4 the probabilities of falling in each stratum is 0.25. Therefor the equations to be solved are

$$\int_0^{x_1} f(x)\,dx = 0.25 \qquad (1)$$

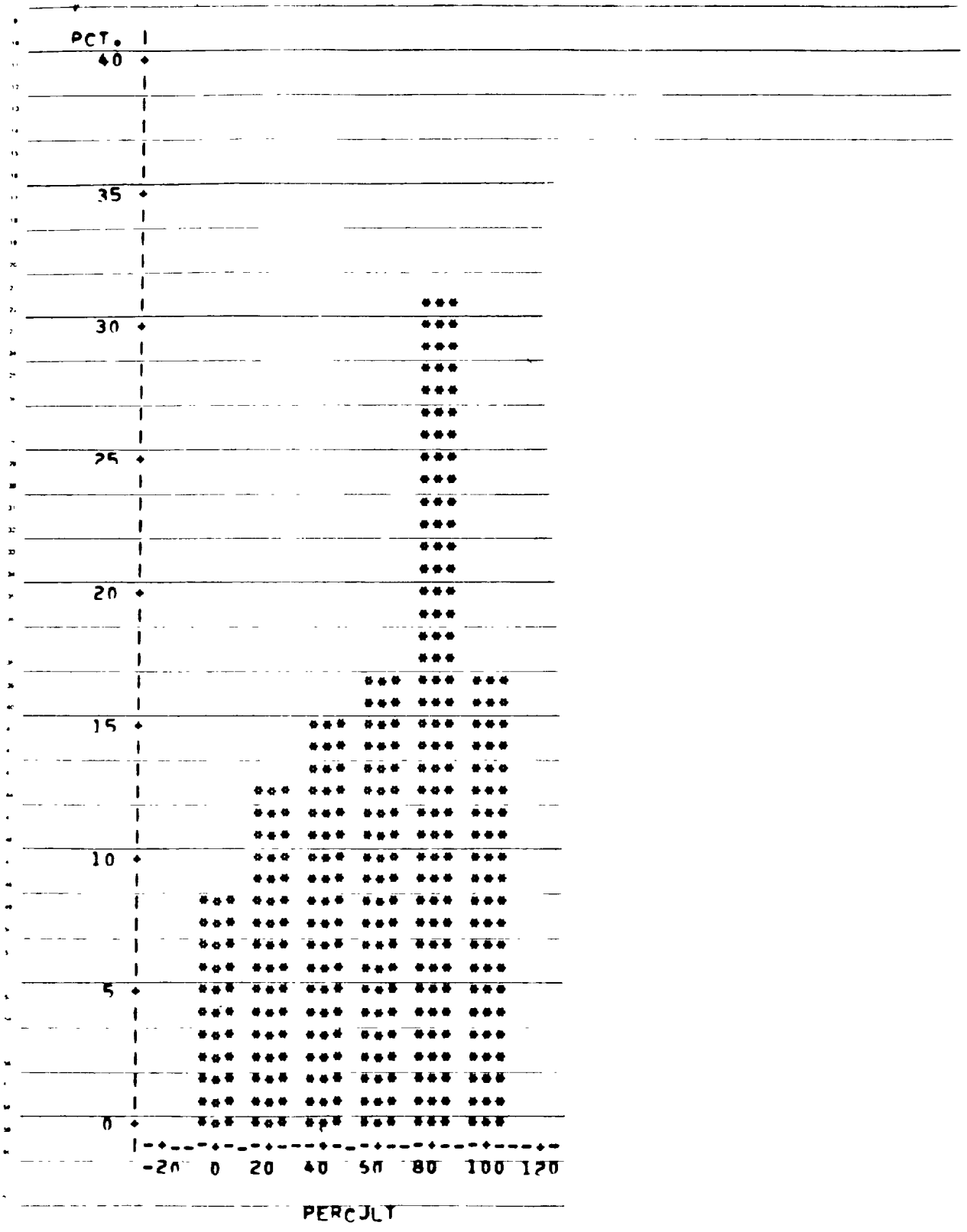
$$\int_{x_1}^{x_2} f(x)\,dx = 0.25 \qquad (2)$$

and

$$\int_{x_2}^{x_3} f(x)\,dx = 0.25 \qquad (3)$$

Solving (1) for x_1 produces a value of 63.0. Similarly we obtain $x_2 = 79.4$ and $x_3 = 90.0$.

FIGURE 1: HISTOGRAM OF PERCENT OF LAND CULTIVATED IN OHIO (1975)



Therefore the stratum boundaries are as follows:

	STRATUM			
	1	2	3	4
SETHI	0 - 63%	63 - 79.4%	79.4 - 90.0%	90 - 100%

5. Equal Aggregate Output - Since we want the cum ($f_h \mu_h$) to be constant in all strata, divide the total cum ($f_h \mu_h$) by L obtaining

$\frac{1}{4}$ cum_k $(f\mu)$ = 2,962

$\frac{2}{4}$ cum_k $(f\mu)$ = 5,924

$\frac{3}{4}$ cum_k $(f\mu)$ = 8,886

$\frac{4}{4}$ cum_k $(f\mu)$ = 11,848

The values in column 5 of Table B1 corresponding to the calculated values are:

	STRATUM			
	1	2	3	4
	0 - 62.5	62.5 - 80.0%	80.0 - 90.0%	90.0 - 100%
cum (f)	3059	6053	8785	11848

A summary of the stratum boundaries for the five methods is as follows:

Table B2: Results of stratum boundary computations in Ohio for four strata by stratification techniques.

	STRATUM			
METHOD	1	2	3	4
I	0 - 30	30 - 60	60 - 80	80 - 100
II	0 - 30	30 - 57.5	57.5 - 80	80 - 100
III	0 - 30	30 - 60	60 - 82.5	82.5 - 100
IV	0 - 63	63 - 79.4	79.4 - 90.0	90.0 - 100
V	0 - 62.5	62.5 - 80.0	80.0 - 90.0	90.0 - 100

The boundaries for the first three methods are very similar as are the boundaries for the last two methods. This result was not unexpected and further investigation will determine which boundaries are optimum. Tables B3 and B4 summarize the stratum boundary values by State for 2, 3, 4 and 5 strata. Note that the boundaries for four strata in Ohio are those from Table B2 above. The appropriate number of strata to use and the best technique and therefore the optimum strata boundaries are given in the text under Results (page 13).

Table B3

Strata boundary values by stratification technique for Ohio

	Number of Strata			
	2	3	4	5
Dalenius-Hodges	60.0	42.5	30.0	25.0
		75.0	60.0	47.5
			80.0	70.0
				85.0
Durbin	57.5	40.0	30.0	22.5
		72.5	57.5	47.5
			80.0	67.5
				85.0
Ekman	60.0	40.0	30.0	22.5
		75.0	60.0	47.5
			82.5	70.0
				87.5
Sethi	79.4	69.1	63.0	58.5
		87.1	79.4	73.7
			90.9	84.3
				92.8
Equal Aggregate Output	80.0	70.0	62.5	57.5
		87.5	80.0	72.5
			90.0	85.0
				90.0

Table B4

Strata boundary values by stratification technique for Kansas

	Number of Strata			
	2	3	4	5
Dalenius-Hodges	62.5	45.0	35.0	27.5
		75.0	62.5	52.5
			82.5	70.0
				87.5
Durbin	62.5	45.0	35.0	27.5
		77.5	62.5	52.5
			82.5	72.5
				87.5
Ekman	62.5	45.0	35.0	30.0
		75.0	60.0	52.5
			80.0	70.0
				85.0
Sethi	79.4	69.1	63.0	58.5
		87.1	79.4	73.7
			90.9	84.3
				92.8
Equal Aggregate Output	82.5	72.5	65.0	60.0
		90.0	82.5	75.0
			95.0	87.5
				97.5