THE USE OF ROTATING SAMPLES IN THE CENSUS BUREAU'S MONTHLY SURVEYS

By: Ralph S. Woodruff, Bureau of the Census

Rotating panels are used on several of the monthly surveys of the Bureau of the Census. Examples are the Current Population Survey, the Monthly Retail Trade Survey, the Monthly Accounts Receivable Survey, and the Monthly Wholesale Survey. Rotation is used in monthly (and other repetitive) surveys because of one or more of the following advantages:

1. Rotation spreads the burden of reporting among more respondents.

2. Rotation permits the use of data from past samples to improve the current monthly estimate. This is done by means of the composite estimation procedure.

3. Rotation may afford an unbiased solution to the problem of large observations which occur in the sample.

The advantages of rotation may be so great, that I believe that the possibility of rotation should be considered for every monthly survey. This is especially true if the data being surveyed are of such a nature that they are expected to have a high month-to-month correlation.

The first advantage, that of spreading the burden of reporting on a sample survey among more respondents, may be very important from the standpoint of maintaining the rate of response. However, this advantage will not be discussed further in this paper. The nature of gains from the composite estimation and large observation procedures will be discussed, however, and their effect on variances of estimates obtained from the survey will be approximated. Examples of the use of these principles will be given in terms of the Monthly Retail Trade Report since this is the survey with which I am most familiar and since this is probably the survey where the greatest gains from rotation have been realized. The principles illustrated in these examples, however, can be applied to any repetitive survey concerned with any subject matter.

I shall describe briefly the sample for the Monthly Retail Trade Survey to serve as a background for the illustrations. This sample provides information on retail sales for individual kinds of business and all kinds of business combined. The sample can be divided into two main categories--the list sample and the area sample. The list sample consists of multiunit organizations and individual establishments which have been identified from previous Censuses and which are large enough to justify their inclusion in a nonrotating sample to be surveyed each month. We shall not further concern ourselves with this portion of the sample since it does not involve rotation. All remaining establishments are represented by the area sample. This area sample consists of a 2-stage sample, the first stage of which is the selection of 230 primary sampling units (counties or groups of counties) from 230 strata which account for the entire United States. In effect, each of these 230 primary sampling units is completely subdivided into area sample segments with definable boundaries and containing on the average about four retail stores each but varying considerably around this average. A sample of these segments equivalent to an over-all rate of 6 percent was drawn. This sample was divided into 12 equal panels each representing a $\frac{1}{2}$ percent sample of all the retail stores in the United States. Each of the 12 panels is assigned to a particular month and this panel is enumerated for that month each year. Two months of data (current and previous) are obtained from each respondent during each enumeration which is made by personal visit of the enumerator each year. This is the rotating sample which is used for illustration of principles in the remainder of the paper.

A. Use of the Composite Estimation Procedure With Rotating Panels

We shall first discuss the use of the composite estimation procedure with rotating panels. I should like to go back to the fundamental principle behind this use of rotating panels. This principle is that in order to develop the most efficient estimate possible, a search for correlated data should always be made. In order to be useful, these correlated data must be either universe data or based on a sample different from that used for the estimate. Then a means of linking these correlated data to the desired estimate must be found. This linking is usually done through a sample survey where data on both the estimate and the correlated item are obtained for an identical sample. There are many ways of using correlated data which may already be available or developing such data when they are not available.

The use of rotating sample is a direct application of this principle. If we consider the estimates which can be made from a rotating sample for the month of June, we can of course obtain the simple estimate for the month of June from the June panel. However, we can also obtain an estimate for the month of June from the May panel by applying the ratio of the June-May results from the June panel to the estimate for May obtained from the May panel. Progressively less reliable estimates for the month of June can be produced from the April, March, etc., panels by using products of the month-to-month ratios which can be developed from the sample. Now, instead of a single estimate for the month of June, we have a number of estimates for June at practically no additional cost and by proper weighting of these estimates can produce a much more reliable composite estimate than the single simple unbiased estimate.

At this point, it should be noted that there are two different systems of rotation which can be used for developing the data necessary for producing these estimates. The system used in the retail trade report is to completely rotate the sample from one month to the next and to obtain from the entire panel two months of data. Another system often used, for example in the Current Population Survey, is to obtain only one month's data at each

enumeration and to rotate only part of the panel, retaining part of the panel to provide information from identicals.

The rotation scheme used in the retail trade survey is more efficient in terms of variance per report because an entirely new panel is available in successive months and because the entire panel is used for the identical links. Therefore, it is to be preferred over the alternative method if it is possible to obtain two months of data (current and previous) with usable accuracy in a single enumeration. This is often not the case. Even in the case of retail sales, which are largely a matter of record, we have had some difficulty with the "previous" sales being regularly reported below the level of the "current" sales. Investigation showed that this was due to a general tendency on the part of enumerators to ignore stores in business the "previous" but not the "current" month. When they were urged to take special care in accounting for such establishments, the differences between the current and previous reports dropped to a much lower level. In many surveys based on on-the-spot observation or upon memory, it may be impractical to attempt to obtain data for two periods in one enumeration. In this case, the less efficient form of rotation must be used. While this will result in different optimum constants and different percentages of gain over the nonrotating system than those presented for the retail system of rotation, the principle is the same and the gains will be well worthwhile if the month-to-month correlations are high.

In the Retail Trade Survey, we use a composite estimate (equation 1) which uses information available from all panels through the ith panel to make an estimate for the ith month. This estimate was used as the only estimate in the Monthly Retail Survey through 1959 and since then as a preliminary estimate. It has the following form:

$$(1)\quad \chi'''_i = (1-W)\,\chi'_i + W\left(\frac{\chi'_i}{\chi''_{i-1}}\right)\left(\chi'''_{i-1}\right)$$

In the above equation:

χ'''_i = the composite estimate for the current (ith) month (note that the composite estimate for the previous month, χ'''_{i-1} is used in the estimate for the current month).

χ'_i = the simple unbiased estimate for the current month.

χ''_{i-1} = the simple unbiased estimate for the previous (i-1) month.

Note that χ'_i and χ''_{i-1} arefrom the same panel and that $\frac{\chi'_i}{\chi''_{i-1}}$ is therefore the month-to-month ratio for an identical panel.

W = a constant having a value less than one (.8 is the value of W now used in the Monthly Retail Trade Survey).

If this composite estimate is used for an indefinite number of months, it can be reduced to a series of the following form:

$$(2)\quad \chi'''_i = (1-W)\,\chi'_i + W(1-W)R'_i\chi'_{i-1} + W^2(1-W)R'_iR'_{i-1}\chi'_{i-2} + \ldots W^m(1-W)R'R'_{i-1}\ldots\ldots R'_{i-m+1}\chi'_{i-m}$$

In the above equation:

R'_i = the ratio of the current to the previous estimate for the ith panel $\left(\frac{\chi'_i}{\chi''_{i-1}}\right)$

This form clearly shows that the estimate uses all estimates available from past panels. A weight of $W^n(1-W)$ is placed on the estimate derived from the panel n months prior to the month being estimated. It can be shown using the type of reasoning used by Patterson in his comprehensive article on this subject (2) that the weights used on each term of the series yields optimum results provided the following conditions are met:

1. The relvariances are equal for all months.
2. The month-to-month correlations are equal for all months.
3. There are no correlations between the results from different panels.
4. The constant W is chosen in optimum fashion (equation 5).

These conditions differ somewhat from those used by Patterson because a different plan of rotation and a different form of estimate is used. It is believed that the above conditions are approximately met in the Retail Trade Survey.

It has been stated that the preliminary composite estimates shown in equations (1) and (2) make approximate optimum use of all data available from all panels through the ith panel. However, data become available from the i plus one panel which can be used to improve the estimate for the ith month. This of course requires a revision of the preliminary estimate but we have recently decided that this revision is worthwhile because it results in some striking variance gains particularly in the ratio between the estimates for the two most recent months. The form of this final estimate (issued one month after the publication of the preliminary estimate) is:

$$(3)\quad \chi''''_i = K\left(\chi'''_i\right) + (1-K)\,\chi''_i$$

In the above equation:

χ''''_i = final composite estimate.

K = a constant having a value less than one. (.83 is the value of K proposed to be used in the Monthly Retail Trade Survey.)

Note that the new piece of information available from the i plus one panel namely the simple estimate for the previous month ($\bar{x}''_i$) is averaged with the preliminary composite estimate ($\bar{x}'''_i$) with weights (1-K) and K.

Two remarks should be made about these composite estimates before we proceed to determining the optimum constants and examining the gains from the estimate. The first is that the estimates are of the ratio form. Regression or difference estimates could have been used instead. The regression estimate would result in a gain in reliability if regression coefficients were properly computed. However, computation of the regression coefficients involve considerable labor and where correlations are very high and relvariances are roughly the same from month-to-month (as in the case of retail trade) regression and ratio estimates yield very similar results. Actually, although the estimates used are ratio estimates at the United States level they are put in a linear form similar to that of the regression or difference estimate at the primary sampling unit level in order to facilitate the computation of variances.

The second remark which should be made is that there is additional information which theoretically could have been used in the composite estimate. In the first place, since the rotation scheme provides identical panels for the same month each year a ratio to the year-ago composite estimate could be used. However, the composite estimate a year earlier is closely correlated with existing terms in the estimate so that it yields very little additional information. The weight indicated for this term and the resulting variance gain appeared to be too small to justify the considerable complications which result from its use.

Another source of information which could be used to improve the estimate is the information available from panels succeeding the month being estimated. We have already indicated that we have decided to revise the estimate on the basis of information available from the panel following the month being estimated. Theoretically all following panels could also be used but this would require successive revisions and it is our opinion that the resulting variance gains do not justify the cost and confusion resulting from this procedure.

If we accept the preliminary and final composite estimates (equations 1 and 3) as the ones we are to use, the next problem is to optimize the constants K and W which determine the weights on the various parts of the composite estimate. This is done by expressing the variance of the composite estimates and then minimizing these variances with respect to the constants. The variance of the preliminary composite estimate (equation 1) may be expressed as:

$$(4) \quad V^2_{x'''} = V^2_{x'} \left\{ \frac{1+W^2-2W\rho}{1-W^2} \right\}$$

In the above formula:

$V^2_{x'''}$ = the relvariance of the preliminary composite estimate.

$V^2_{x'}$ = the relvariance of the simple unbiased estimate from a single panel.

ρ = the month-to-month correlation between the current and previous estimate from the same panel.

If the above variance is minimized with respect to W then:

$$(5) \quad W = \frac{1-\sqrt{1-\rho^2}}{\rho}$$

The variance of the final composite estimate (equation 3) may be expressed as:

$$(6) \quad V^2_{x''''} = K^2 V^2_{x'''} + (1-K)^2 V^2_{x'}$$

In the above formula:

$V^2_{x''''}$ = the relvariance of the final composite estimate.

If the above variance is minimized with respect to K then:

$$(7) \quad K = \frac{V^2_{x'}}{V^2_{x'''} + V^2_{x'}}$$

or if optimum W is used in $\bar{x}''' = \frac{1-\sqrt{1-\rho^2}}{\rho^2}$

The above relationships are subject to the conditions previously mentioned (i.e., that the relvariances ($V^2_{x'}$) and month-to-month correlations (ρ) be equal for all months and that there be no correlation between the results for different panels). As previously mentioned these conditions are approximated by the Monthly Retail Trade Survey. The relvariances and month-to-month correlations are roughly equal although not exactly so. There are slight correlations among the various panels due to the fact that the sampling was done without replacement. Also there are year-to-year correlations since the same panel is used each year for a given month. These latter correlations are associated with powers of K and W of 12 or greater and should not affect the variance significantly. The fact that the conditions are not exactly met does not bias the results but means that the constants may not be precisely optimum and that variances may be slightly greater than those indicated by theoretical computations based on the assumptions.

The constants actually used in the Monthly Retail Trade Survey are W = .8 and K = .83. The variance results obtained from the use of these constants is compared with the results obtained from optimum constants in table 1. The results are obtained for month-to-month correlation (ρ) = .98, .99 and .95. The correlation of .98 is roughly that obtained for all kinds of business combined and is approximately the median of the correlations for individual kinds of business. The correlations of .99 and .95 are relatively high and low respectively among those obtained for individual kinds of business. The loss over optimum constants is not great even for the high and low correlations. All the relvariances are expressed as multiples of the relvariance of the simple estimate, or in other words the variance which would be obtained from a nonrotating sample. Note that all relvariances whether or not optimum constants are used are a third or less of the variance from a nonrotating sample.

Table 1: OPTIMUM CONSTANTS UNDER HIGH, MEDIUM AND LOW CORRELATION ASSUMPTIONS AND COMPARISON OF VARIANCES OBTAINED USING THESE CONSTANTS WITH VARIANCES OBTAINED WITH SPECIFIED CONSTANTS

Correlation assumption	Optimum constants		Relvariances of composite estimates of level for a single month				Ratio of variance obtained with specified constants to variances obtained with optimum constants	
			With optimum constants		With specified constants[1]			
ρ =	W	K	Preliminary	Final	Preliminary	Final	Preliminary	Final
			(As multiples of $V_{\bar{x}'}$)[2]					
.99 (high)	.868	.876	.141	.124	.156	.136	1.11	1.10
.98 (med.)	.817	.834	.199	.166	.200	.167	1.01	1.01
.95 (low)	.724	.762	.312	.238	.333	.258	1.07	1.08

[1] W = .8, K = .83, (the constants used).

[2] Since $V_{\bar{x}'}$ is equal to the relvariance of the simple estimate obtained from a nonrotating panel the relvariances shown are in the form of ratios of variances of composite estimates to variances of estimates from nonrotating panels of the same size.

Computations are theoretical-based on stated correlations and other stated assumptions.

The optimum values of the constants K and W shown in table 1 are those needed to produce the most efficient estimate of level for a single month. Of course, the statistics may be used in many other relationships, for example, to obtain month-to-month trends, month-to-year ago trends or annual totals to name a few of the most common relationships. These relationships may be more important to the user than the level itself. The optimum constants for any one of these relationships will in general not be the optimum constants for a single month's level. However, we adopted the approximate optimum constants for a single month's level on the philosophy that the estimates can and will be used in a very large number of relationships and the only way to insure that all these relationships will have a reasonably low variance is to produce a good level estimate.

It should be noted at this point that there is one very important exception to the general rule that constants which are optimum for level are not optimum for other relationships. If constants necessary to obtain optimum results for the level of the preliminary composite estimate and the level of the final composite estimate are used, optimum results will also be obtained for the ratio between the two most recent months.

Table 2 shows the variance results obtained from composite estimates for the month-to-month ratio, the month-to-year ago ratio and the annual totals. These are compared with the results obtained from a nonrotating sample. All results in this table are theoretical based on certain correlation assumptions and other assumptions. All computations for composite estimates have been made with the constant W = .8 and constant K = .83 which are the constants used rather than the optimum constants. On the month-to-month and month-to-year ago relationships, two results are shown, one for the ratio between a preliminary and final estimate which would be the ratio available when the month in the numerator is first published. The ratio between two final composite estimates which would not be available until one month later is also shown.

Some generalizations can be made from this table. The variance of the preliminary month-to-month ratio from composite estimates is equal to or slightly smaller than the variance of the month-to-month ratio which can be obtained from a nonrotating (identical) sample of the same size. The variance of the month-to-month ratio between two final composite estimates is substantially higher. However, primary interest is probably centered on this relationship when it first appears. On the month-to-year ago relationship the variance of the final ratio is somewhat smaller than the variance of the preliminary ratio. In this case, however, the variance of both the preliminary and final ratios are smaller than those obtained from an identical nonrotating sample. On annual totals the variances of the sum of 12 final composite estimates is far below the result which would be obtained from a nonrotating sample. For this particular statistic, however, an even lower variance would be obtained from a sum of the 12 simple results from the rotating sample.

Table 2: RELVARIANCES OF MONTH-TO-MONTH RATIOS, MONTH-TO-YEAR-AGO RATIOS AND ANNUAL TOTALS FROM COMPOSITE ESTIMATES COMPARED WITH RELVARIANCES OF ESTIMATES FROM A NONROTATING SAMPLE OF THE SAME SIZE

Correlation assumptions	Month-to-month ratios			Month-to-year-ago ratios			Annual totals		
	Non-rotating sample	Preliminary composite[4]	Final composite[5]	Non-rotating sample	Preliminary composite[4]	Final composite[5]	Nonrotating sample[6]	Sum of final composite estimates	Sum of simple estimates from rotating sample
	(All relvariances expressed as multiples of V_X^2)								
High[1].....	.020	.020	.047	.200	.126	.082	.908 - .991	.093	.083
Medium[2].....	.040	.040	.062	.300	.210	.154	.863 - .982	.105	.083
Low[3].....	.100	.098	.107	.500	.454	.359	.771 - .954	.143	.083

[1] Month-to-month correlation = .99, Year-to-year correlation = .90.
[2] Month-to-month correlation = .98, Year-to-year correlation = .85.
[3] Month-to-month correlation = .95, Year-to-year correlation = .75.
[4] Variance of ratio between preliminary composite estimate for most recent month and final composite estimate for earlier month. This is the ratio available when the current month is first published.
[5] Variance of ratio between two final composite estimates (note that this ratio is not available until one month after the data are first published).
[6] The relvariance of the annual estimate from a nonrotating sample is dependent on the average correlation between estimates one to twelve months apart. We have assumptions for only the two extremes. The computations are made assuming the average correlation is at these two extremes. The actual variances are within the ranges shown.

All computations on a theoretical basis. See text for assumptions made. In composite estimates the constants used were W = .8, K = .83.

B. Use of Rotation to Establish a Panel of Large Observations Which can be Sampled at Heavier Rates

The occurrence of large observations is one of the principal problems in sampling. If the sample is nonrotating one is usually confronted with the unhappy choice of accepting the considerable increase in variance they create or of taking a bias by arbitrarily reducing their weight. In the rotating system these large observations can be placed in a special panel which can be sampled at heavier than normal rates, thus permitting the weights to be reduced without biasing the results.

The principle of this procedure is simple. Identify in n-1 previous panels these large observations and survey them in the current panel. The weight of these observations (including all like them in the current panel) is then divided by n which may drastically reduce their effect on the estimate and the variance of the estimate. While the principle is simple, it is sometimes difficult to put into effect because it is required if the estimate is to remain unbiased, that the definition of "large" that is used be applied equally to all of the n panels. This is difficult because data obtained in each of the rotating panels is usually for different months so that there is no common statistic which is readily available for all the n panels. However, this difficulty can often be overcome as will be illustrated in the Monthly Retail Trade Survey.

While the illustrations which will be given concern large retail establishments which appear in the area sample of the retail survey, the principle can be applied to any survey using rotating panels. For example, a similar feature termed the "rare event universe" has been established in the Current Population Survey for those area sample segments which contain an unusually large number of households.

A number of large establishments appear in the area sample of the Monthly Retail Trade Survey in spite of the existence of the list sample of large establishments and firms taken from the most recent Census. These large establishments in the area sample may be establishments which were born or have become large since the latest Census or they may be establishments too small to put on the certainty list but large enough to cause considerable variance in the area sample. The problem is to create what we call the large area sample panel for such establishments.

In the case of the Retail Trade Survey, we faced the previously mentioned difficulty of having no common statistic available in the 12 panels to use for the definition of "large" since we had obtained only two months of data from each respondent. A criterion suggested by Max Bershad was used in this case. This criterion was that each large area sample panel member equal or exceed a certain sales cutoff for each month of the year. By looking at the particular months we had, we could determine if it was possible for the establishment to meet the criterion. Where the establishment equalled or exceeded the criteria in the months we had, it was placed on a "potential" large area sample list. At the end of 12 months, all members of the "potential" large area sample were surveyed for their sales in each of the 12 months. Those that qualified

(about half of the potential list) were placed on the "permanent" large area sample list.

The above procedure requires some time after the end of the criterion period to determine the members of the permanent large area sample panel. For this reason, it is probably practical only in cases where the rotation is periodically repeated (as in the monthly retail trade sample). However, with other systems of rotation other procedures can be used, some completely unbiased and others with biases much smaller than those resulting from an arbitrary reduction in weight.

While the establishment of the permanent large area sample panel made an important reduction in variances, we found that large observations were still appearing in the area sample. These establishments were those which had appeared since the permanent large area sample panel was last brought up to date or which had failed the most recent large area sample test because some months were low. To reduce the variance from such cases, we set up what we call a temporary large area sample panel. Each adjacent pair of panels contains a common month of data (due to the fact that two months of data are obtained from each panel). If, for example, we consider the adjacent panels of May and June--we have information from both panels for the month of May. We then set a criterion for the month of May. Any establishments which equal or exceed this criterion for the current month for the May panel are also surveyed for June and placed in this tabulation at half weight. At the same time, the weight of all establishments in the June panel whose previous month's sales exceed the criterion are halved. In this fashion, the weight of all "large" establishments (except those "large" in the current month only) are halved. The "temporary" and "permanent" large area sample procedures are integrated by using the same cutoffs for both. Those used in the "temporary" large area sample procedure therefore constitute the "potential" large area sample which is surveyed for the permanent large area sample panel after the end of 12 months.

The optimization problem involved in the use of a large area sample panel is to decide that large area sample cutoff which will produce the most efficient results. All establishments with sales equal to or greater than this cutoff are placed on the large area sample list while those with sales smaller than this cutoff are left in the regular area sample.

The large area sample cutoff was determined empirically by fixing a cost and then approximating the variance from the combined large area sample and regular area sample strata that would be obtained with various cutoffs.

The formula for the variance obtained from the combined strata for a fixed cost and stated cutoff may be expressed as follows:

$$(8)\quad \frac{\frac{M^2\sigma_y^2}{12} + M^2\sigma_x^2\left\{(.83)^2\left[\frac{1+.8^2-2\rho_x(.8)}{1-.8^2}\right]+(.17)^2\right\}}{\frac{C}{12C_yN_y + C_xN_x}}$$

The numerator of the above fraction represents the combined variance from the area sample and large sample strata per area sample segment drawn. A final-type composite estimate is assumed with the values of W and K which are proposed to be used.

The denominator of the fraction represents the number of area sample segments which can be drawn for the fixed cost and designated cutoff.

The meaning of the individual symbols in the numerator as follows:

M = the number of area sample segments in the universe.

σ_y^2 = the variance per area sample segment for the large area sample stratum.

$= \frac{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}{m}$ where Y_i is the segment total of those defined to be in the large area sample universe by the cutoff.

σ_x^2 = the variance per area sample segment for the area sample stratum (of similar form as the variance shown above but with X, or area sample values, substituted for Y values).

ρ_x = the month-to-month correlation for the segment totals of the X values.

Note that the variance of the large area sample universe is divided by 12 (since 12 area sample panels are included) while the area sample portion is multiplied by the reduction factor achieved through the use of the composite estimate. Note also that the values of σ_y^2, σ_x^2 and ρ_x are dependent on the cutoff selected while the other values in the numerator are constants. This variance form assumes that there is no correlation between the X and Y values. (In a small scale study made in the New York primary sampling unit these covariances had a negligible effect on the results).

In the denominator the individual symbols have the following meaning:

C = total resources available for expenditure on the combined area sample and large area sample strata.

C_y = the cost per unit per month of obtaining and processing a large area sample report. (Note that this constant is multiplied by 12 because large area sample establishments from all 12 panels are surveyed each month.)

$\bar{N}_y$ = the average number of large sample establishments per segment.

$C_x, \bar{N}_x$ are similar values relating to the area sample universe.

Note that $\bar{N}_x$ and $\bar{N}_y$ are dependent on the cutoff selected but their sum is constant ($\frac{N}{M}$). All other values in the denominator are constants.

In the above formula, the variance per segment draw declines as does the number of segments which can be afforded for a given cost. The problem is to find that design which gives the lowest value for the fraction. This was done empirically by designating various large area sample cutoffs and then computing the variances for these cutoffs using the above formula. The variances in table 3 are for an estimate of sales of all kinds of business in the New York Metropolitan District while the variances in table 4 are for an estimate of sales of proprietary stores in the United States. Two sets of per unit cost assumptions are used. In the first set (those shown in columns 2 and 4 of each table) it is assumed that the per report cost of an area sample case is four times that of a large area sample case. This approximates the conditions of the retail survey where the area sample reports are collected by personal enumeration and the large area sample reports by mail. In columns 3 and 5 it is assumed that the method of collection and the per unit costs are the same for both strata. The optimum large area sample cutoff in this case is somewhat higher.

Only limited empirical evidence relating to optimum large area sample cutoffs is available because of the labor of computation. However, those data available point to the same general conclusion as column 4 of tables 3 and 4 namely, that there is a broad range for the cutoff centered around 3 to 5 times the average sales value per establishment where losses over the optimum cutoff are apparently small. This is true both for optimums computed from the standpoint of estimates for individual kinds of business and for those computed from the standpoint of the total estimate for all kinds of business combined. The large area sample cutoff for the Retail Trade Survey has been set at about three times the average sales value per establishment. Since only one cutoff could be used for each kind of business, arbitrary compromises were made where the average sales for a particular kind of business differed significantly from the average sales for all kinds of business combined.

Formula (8) and the computations in tables 3 and 4 are designed to produce optimum results for the final composite estimate. Optimum cutoffs could have been computed also for the preliminary composite estimate, the month-to-month change or other relationships.

Table 3: EMPIRICAL COMPUTATION OF APPROXIMATE OPTIMUM CUTOFF BETWEEN AREA SAMPLE AND LARGE AREA SAMPLE STRATA--ALL KINDS OF BUSINESS IN THE NEW YORK METROPOLITAN DISTRICT

Cutoff (monthly sales)	Variance per area sample segment drawn ($\times 10^{12}$)	Number segments drawn with fixed cost[3]		Variance obtained with stated cutoff	
		With $C_x = 4C_y$	With $C_x = C_y$	With $C_x = 4C_y$ (1) + (2) ($\times 10^{12}$)	With $C_x = C_y$ (1) + (3) ($\times 10^{12}$)
	(1)	(2)	(3)	(4)	(5)
(00) [1]	187,678	199.6	278.5	940	674
\$100,000 ($20\bar{x}$) [2]	136,391	198.5	271.0	687	503
40,000 ($8\bar{x}$) [2]	116,531	196.1	253.7	594	459
25,000 ($5\bar{x}$) [2]	93,126	191.3	225.1	487	414
15,000 ($3\bar{x}$) [2]	89,078	182.0	182.0	489	489
10,000 ($2\bar{x}$) [2]	92,689	169.8	141.9	546	653
5,000 ($\bar{x}$) [2]	94,145	129.1	69.6	729	1,353

[1] In other words, all cases in area sample stratum.

[2] $\bar{x}$ is the average sales per establishment in the combined strata ($5,025).

[3] Cost is fixed at level needed for 182 segments with cutoff = 3 $\bar{x}$ (approximate present design) but note that the optimum is independent of the size of the fixed cost. As indicated in Equation 8, Cy is the per unit cost of collecting and tabulating a large area sample report while Cx is the same cost for an area sample report. The first cost assumption (Cx = 4 Cy) is approximately the relationship in the Monthly Retail Trade Report because area sample reports are collected personally while the large area sample reports are collected by mail. The other assumption is used to show the variance relationships if the two types of reports were collected by the same method.

Table 4: EMPIRICAL COMPUTATION OF APPROXIMATE OPTIMUM CUTOFF BETWEEN AREA SAMPLE AND LARGE AREA SAMPLE UNIVERSES--PROPRIETARY STORES

Cutoff (monthly sales)	Variance per area sample segment drawn ($\times 10^{12}$) (1)	Number segments drawn with fixed cost[3]		Variance obtained with stated cutoff	
		With $C_x = 4C_y$ (2)	With $C_x = C_y$ (3)	With $C_x = 4C_y$ (1) + (2) ($\times 10^9$) (4)	With $C_x = C_y$ (1) + (3) ($\times 10^9$) (5)
(00) [1]	13,495	2,061	2,785	6,548	4,846
\$33,048 ($8\bar{x}$) [2]	12,476	2,043	2,656	6,107	4,697
20,655 ($5\bar{x}$) [2]	11,348	1,923	2,000	5,901	5,674
12,393 ($3\bar{x}$) [2]	11,211	1,900	1,900	5,901	5,901
8,262 ($2\bar{x}$) [2]	10,951	1,797	1,540	6,094	7,111
4,131 ($\bar{x}$) [2]	10,297	1,284	643	8,019	16,014

[1] In other words all cases in area sample stratum.
[2] $\bar{x}$ is the average sales per establishment in the combined strata (\$4,131).
[3] Cost is fixed at level needed for 1,900 segments with cutoff = $3\bar{x}$ (approximate present design). See note [3], table 3 for per unit cost assumptions used.

C. Summary of Effect of Composite Estimation Procedure, and Large Observation Procedure (both temporary and permanent) on Variances of the Monthly Retail Trade Survey

The actual combined effect of the composite estimation procedure and the large observation procedures (both permanent and temporary) is shown by comparing the variances of the estimates obtained using these procedures with the variance of a nonrotating sample. These variances have been computed for all kinds of business combined in the United States for May and June 1959 in table 5. The comparisons shown are the variances of the preliminary and final composite estimates versus the variance of the simple estimate and the variance of the preliminary-to-final month-to-month ratio versus the variance of the ratio of the simple estimates from a nonrotating panel.

These are total variances and a between primary sampling unit contribution from the entire sample is included in the variances of both the simple estimates and the composite estimates. This contribution has not been affected by any of the devices discussed in this paper, therefore percentage of gain is not as large as for the within-primary sampling unit variance alone which is the component affected by the procedures discussed. In spite of this, it appears that the variance of level has been reduced to about 30% of that which would have been obtained from a nonrotating sample while the variance of the ratio between the two most recent months has been reduced to about 45% of that obtained from a nonrotating sample. These results are only approximate because actual conditions are different for each estimate considered and because these variance results are themselves subject to variance. The cost of the rotating sample used may be between 10 and 15% larger than the cost of a nonrotating sample due principally to the cost of the large area sample procedure.

Table 5: COMPARISON OF VARIANCES OF ESTIMATES OBTAINED USING COMPOSITE ESTIMATE AND TEMPORARY AND PERMANENT LARGE AREA SAMPLE PROCEDURES WITH VARIANCES OF THE SIMPLE ESTIMATES FROM NONROTATING PANEL; ALL KINDS OF BUSINESS: MAY AND JUNE 1959

Estimate	Relvariances of simple estimate	Relvariances of composite estimate	Ratio of variances of composite estimate to variances of simple estimate
May 1959 level.....	.000688	[2].000197	.29
June 1959 level.....	.000685	[3].000202	.29
June-May 1959 ratio..	[1].000033	[4].000015	.45

[1] Same panel used for both May and June.
[2] Final composite estimate (relvariance of preliminary composite estimate for May was .000230).
[3] Preliminary composite estimate.
[4] Ratio preliminary to final composite estimate.

Bibliography:

(1) Hansen, M.H., Hurwitz, W.N., Madow, W. G., (1953). Sample survey methods and theory. John Wiley and Sons, Inc., New York.

(2) Patterson, H.D., (1950). Sampling on successive occasions with partial replacement of units. Jour. Roy. Stat. Soc. Series B, 12, 241-255.

(3) Jessen, R.J., (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa Agr. Exp. Sta. Res. Bull. 304

(4) Yates, F. (1949). Sampling methods for censuses and surveys. Charles Griffin and Co., London

(5) Cochran, W.G., (1953). Sampling techniques. John Wiley and Sons, Inc., New York