A MULTIPLE FRAME DESIGN TO ESTIMATE ECONOMIC DISTRIBUTIONAL EFFECTS

by

Douglas G. Kleweno

Statistical Research Division

Economics and Statistics Service

U.S. Department of Agriculture

Washington, D.C. 20250

December 1980

ESS NO. AGESS801126

A Multiple Frame Design to Estimate Economic Distributional Effects. By Douglas G. Kleweno; Economics and Statistics Service; U.S. Department of Agriculture; Washington D.C. 20250; December 1980, ESS Staff Report No. AGESS801126.

ABSTRACT

A multiple frame design was used to study the economic distributional effects in a rural Kentucky area. The objectives, frame construction, sample design and data collection procedures were described as a preface to variable estimation. A combined list and area frame estimator was used to estimate subpopulation means and totals. This estimation technique proved to be a feasible approach and is recommended for future use in surveys of households and establishment traits.

Key words: area frame, list frame, subpopulation, domain, frame, unduplication, composite estimate, domain estimate, nonresponse.

CONTENTS

SUMMARY

An economic distributional effects study of household and establishment subpopulations in a rural area of Kentucky was conducted in 1979 using a multiple frame design. The study relied on two independent frames - a list frame of establishments (private and public inclusive) and an area frame of households and establishments. Both frame sources were stratified. The list was stratified by firm function (nine standard industrial classes) and substratified into three sizes: 1-19 employees, 20-99 employees, and 100 or more employees. The area frame was stratified into three geographical areas - urban, suburban, and rural. Sampled list establishments provided a subsample of their employees for enumeration. The list firms and their sampled workers formed the overlap domain of the population. Sampled households and firms in the area frame not represented on the list formed the nonoverlap domain of the population. This determination was necessary under the assumptions of multiple frame sampling and occurred during the screening process. Data collection was confined in the area frame to sampling units in the nonoverlap (NOL) domain because of time and monetary constraints.

Three questionnaires were used to gather data - a household version, a private establishment questionnaire, and a government version. Field work extended over a three month period with data collection divided into three stages. (At this date questionnaires have been edited and summarization has started).

All subpopulation estimates of means and totals were constructed at the strata level. Domain estimates of totals were made for each subpopulation group. The independent domain estimates were then combined into composite total estimates for each subpopulation. A combined ratio estimator was used to compute household mean estimates.

INTRODUCTION

A survey of households and establishments in a nine-county area of southeastern Kentucky was conducted by the Economics, and Statistics Service (ESS), U.S. Department of Agriculture in late 1979. The study was designed to provide data to assess: (1) the effects of recent rapid population and employment growth in a typical rural area and (2) the impact federal economic development programs have on employment and income changes. Specific objectives of the survey were: (1) to determine how employment and income growth had been distributed among various population subgroups, and (2) to determine the effect of government programs on growth and distribution of employment and income.

The survey design was an application of two-stage multiple frame sampling. The first stage of sampling involved selecting area segments from a stratified area sampling frame, and establishments from within a stratified list sampling frame. Second stage sampling involved the selection of households and employees from area frame segments and list frame establishments respectively. Multiple frame estimates and their sampling errors were computed for employee (households) and establishment subpopulations. This paper was written to provide a summary of the sample design and to give a closer look at estimation procedures. A much broader view of the project is presented in another report [4].

SAMPLE DESIGN

To study the target population there was a need to identify and distinguish several subpopulation groups. Households were of interest for individuals employed, unemployed, and out of the labor force. Establishment characteristics were desired by size of employer and standard industrial classification (SIC). There was also a need to cross-reference or link data between employers and their employees. Because of bias and cost problems associated with a single frame design, a multiple frame design was chosen for the nine-county Kentucky study.

A complete list frame of establishments or employed people could not be constructed. Overall firm estimates based on only a list sample would have been biased. The list, however, even with some incompleteness, was efficient as a sampling frame because it could be stratified/sub-stratified with the substrata sampled at different rates. An area frame cluster sample was ruled out because of the high costs in terms of both time and resources required for defining and enumerating the reporting units for the population of interest. The large number of segments necessary to provide a reliable estimate was prohibitive particularly for rare items such as households with unemployed persons which accounted for less than 10 percent of the population. Joint use of the list and area frame insured complete coverage of the population. The list frame provided satisfactory estimates of the population by strata, and the area frame provided an estimate of the incompleteness of the list. The theory of this multiple frame sampling approach was developed by Hartley [3] and Cochran [1] and has been used by ESCS for several operational surveys.

Table 1 shows the establishment list of private and public firms stratified by SIC code and number of workers. There were nine strata based on SIC code with each divided into three substrata based on employment size. Private firms were grouped into eight SIC code strata and government units formed a ninth stratum. Reasons for stratifying by function and size were: (1) list information was available to classify firms into homogeneous groups for reducing variance estimates, (2) a major study objective was to compute estimates for this breakdown, and (3) the function and size classification insured representation across all firms of interest in the subpopulation.

Table 1: Stratification of Establishment List

Strata			Substrata	
Stratum Code	Industry	SIC	Substratum Code	Firm Size
1	Mining	10-14	0	1-19 employees
2	Construction	15-17	1	20-99 employees
3	Manufacturing	20-39	2	100 plus employees
4	Transportation	40-49		
5	Wholesale	50-51		
6	Retail	52-59		
7	Finance	60-67		
8	Service	07-09		
		70-89		
9	Government	91-97		

The list of establishments was constructed using the primary name, SIC Code, and address to identify each potential sampling unit. The firm list was constructed by combining several lists which included telephone directories, a private economic information service list, and a state employment security list. Considerable effort was expended to identify and remove list duplication. Firms operating at different locations and/or carrying out different functions (SIC) were listed separately. This distinction was not always an easy one particularly in the public sector. To determine all potential sample units with each being independent and mutually exclusive required frequent contact with local officials living in the study area. A sampling unit was associated frequently with several secondary names because all the information was available from one primary source. For example, the city government was identified generally through the mayor's or city clerk's office. Secondary bureaus or agencies associated with the office included the fire department, police department, and water works. Because all units were not correctly classified, a proration factor (P_{hi}) was necessary to adjust reported data if duplication occurred for the i^{th} firm of stratum h. This factor can be found in the estimator shown later.

The list frame units were randomly ordered within strata before a systematic sample of units was selected for each stratum. A sample of 458 firms was selected from the population of 3641 sampling units. Sample size was conditioned on budget constraints and the desire to have estimates for major characteristics within 10 percent of the true value with 95 percent confidence. The sampling rates by size of establishments (substrata 0, 1, 2,) were 1/10, 1/4, and 1/1 respectively. This proportional allocation was used for all nine strata because the smaller firms (substratum zero) accounted for over 90 percent of the subpopulation.

To obtain household characteristics for employees of the sampled firms, a subsample was selected from a list of all employees. A systematic sample of employees was selected from firms proportionate to the size (substrata) of the firm. The employee sampling rate was 1/4 for establishments in substratum zero, 1/10 in substratum one and 1/40 in substratum two. This procedure made the data self weighting and permitted employer traits

to be linked to employee traits by industry and size.

Because the establishment list was incomplete, an area frame was used so all households and firms would have a chance for selection. Area frame development involved review of several options. Consideration of a totally new frame was ruled out by time and cost limitations. Existing area frames from the Census Bureau and ESS-- Statistics were compared. The land-use area frame constructed and maintained by ESS-Statistics was selected. This frame for Kentucky was developed in 1976. It was modified to meet the study objectives by redefining the land-use strata based on agricultural intensity to be compatible with the economic study based on density of population.

A two stage stratified cluster design was used for the area frame. The population was classified into three strata for sampling - urban, suburban and rural. This primary break provided homogenous groupings and made data collection more manageable. The primary sampling unit (PSU) was an area segment and the secondary sampling unit the household or establishment. Size of sampling unit varied depending on the particular stratum. A segment was one city block in the urban stratum which contained densely populated areas. Because block sizes varied, adjustments were made to equalize as close as possible the number of dwellings per block. Sampling units in the suburban stratum were defined as one eighth of a square mile and in the rural stratum as one square mile in area.

A total of 9011 primary sampling units were identified for possible selection. The units were replicated to permit selection of additional segments if necessary after the pretest. A sample with 318 PSU's was selected. The urban, suburban, and rural strata consisted of 69, 183, and 66 segments respectively. A larger sample of segments was selected from the urban stratum (6 percent) and less from the rural stratum (2 percent) since emphasis was placed on identifying establishments and households. Households were sampled within the sample segments at an overall population rate of 1 percent. All establishments within the sample segments were enumerated.

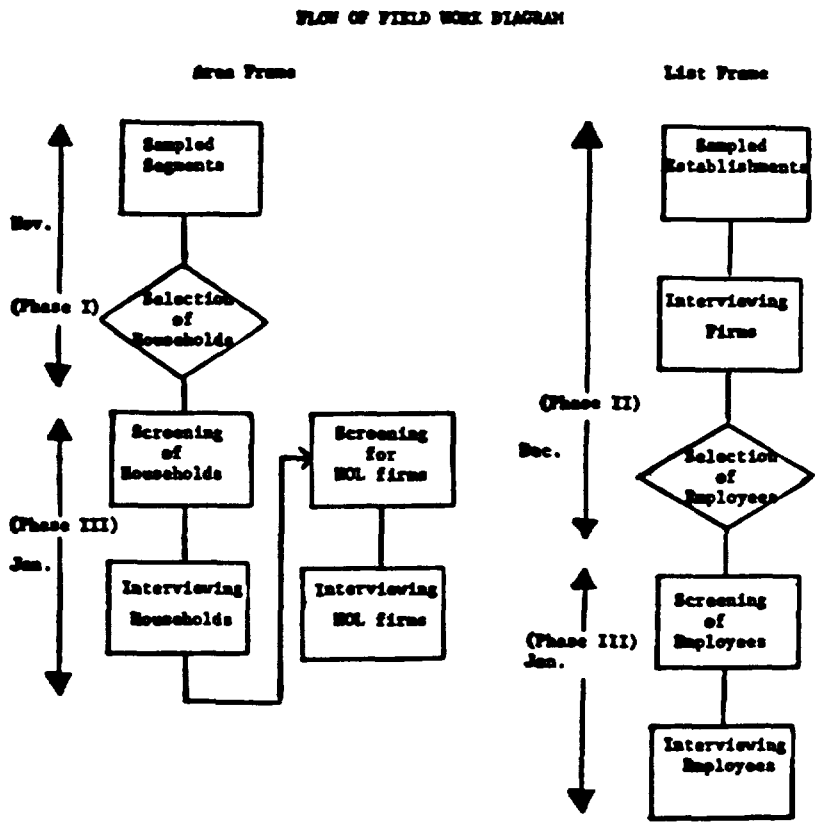
SURVEY PROCEDURES

Detailed discussion of survey procedures and data collection activities would be too massive for this report so only highlights have been given for cohesiveness.

Three questionnaire versions were used to gather information on the selected sample units. Each version asked for longitudinal data for 1974 and 1979 to determine historical trend. The household questionnaire obtained demographic, work and resident history, and income information. The public and private establishment versions collected information on firm size, type of industry, employment characteristics, capital resources, payroll, employee work hours, and sales (private firms only).

Pretest results from five area segments (PSU's) and twenty-five list frame establishments were used to refine the survey questionnaires. The questionnaires were initially too long, certain items proved too difficult, and the flow of questions was poor. A key test of the design was whether employers would provide a list of their workers for subsampling by the interviewer. The results showed over 95 percent of the employers interviewed agreed to this procedure. The pretest nonresponse rate and sampling errors were reasonable.

Data collection was completed in three phases for the actual survey. This was necessary because of limited time and staff. Dividing the project into stages maximized use of resources. The multiple frame design did not conflict with this plan. Field activity in phase I consisted of locating the selected PSU's in the area frame and then listing all secondary sampling units (occupied households and establishments) for screening and interviewing in phase III. Phase II activities involved working with the list frame. Sampled list establishments were interviewed and employee lists were used to subsample workers. Phase III fieldwork identified area frame establishments not on the list frame (nonoverlap). All area frame establishments and households classified nonoverlap during the screening step were interviewed. All list frame households selected in phase II were also interviewed at this time subject to any screenout. A diagram of the work flow is given below.



Two domains within each subpopulation were identified because a complete area frame was used in the two frame sampling procedure. The nonoverlap (NOL) domain was formed by the area frame units not on the list. The overlap domain was formed by the area frame units on the list. The frame unduplication process was necessary to meet the assumptions inherent to multiple (two) frame sampling. (Each element of the population must belong to either of the two domains and each element must be classified into the domain which it represents). The first assumption was met because the area frame was inclusive of the entire subpopulation under study. The second assumption required the frame unduplication process to identify units of the subpopulation (households or firms) contained in both the list and area frame.

Operationally the area frame unduplication process used the list frame of establishments. An area firm was classified overlap if the firm was on the list. An area household was classified overlap if any household member was employed by a firm on the establishment list. No area frame questionnaire (household or firm) was completed for overlap units. Only NOL area households and firms completed a questionnaire. This unduplication was made during field screening.

ESTIMATION

Analysis of the survey data was developed around estimators for totals and an occasional mean for households (employees) and establishments. A general, unbiased domain estimator for subpopulation totals was constructed at the strata level for the area and list. This estimator with few modifications was functional for either frame. The independent domain estimates for the area and lists were then combined for composite total estimates of the subpopulations. A combined ratio estimator was used to estimate household (employee) means because the total households in the subpopulation was unknown.

Notation used to develop the domain estimators can be found in Table 2 below. Each estimate $\hat{X}_{h(g)}$ was identified by the subpopulation group and frame source (g) and stratum (h) for the variable (X) of interest. For example, the domain total estimate $\hat{X}_{h(1)}$ was summed over nonoverlap establishments in stratum h of the area frame to estimate for a characteristic X.

All estimators were constructed at the stratum level. The finite population correction factor was ignored since the sampling fractions were generally less than 5 percent.

Table 2: Domain Estimators by Stratum $h = \begin{cases} 1, 2, 3 & \text{Area frame} \\ 1, \ldots, 9 & \text{List frame} \end{cases}$

Notation $\hat{X}_{h(g)}$, g=1,2,3,4	Subpopulation Group	Domain	Frame Source
$X_{h(1)}$	Establishment	Nonoverlap	Area Frame
$X_{h(2)}$	Establishment	Overlap	List Frame
$X_{h(3)}$	Household (employee)	Nonoverlap	Area Frame
$X_{h(4)}$	Household (employee)	Overlap	List Frame

Domain Total Estimates

The general unbiased domain estimator for household (employee) or establishment totals, referenced from Cochran [2], for stratum h was

$$(g)\hat{X}_h = \sum_{i}^{n_h} \sum_{j}^{m_{hi}} \left\{ f_{oh} \cdot P_h \left[\phi_{hi} \cdot X_{hij} \right] \right\} \qquad (1)$$

where,

f_{oh} = original expansion factor for h^{th} stratum,

$$P_h = \begin{cases} P_{hij} & \text{List Frame Households (g=4): probability of selecting a household} \\ P_{hi} & \text{List Frame Establishments (g=2): proration factor for firm duplication} \\ 1.0 & \text{Area Frame NOL Households and Establishments (g=1,3)} \end{cases}$$

$0 < P_h \leq 1.0$,

$$\phi_{hi} = \begin{cases} 1 & \text{If attribute exists for sampling unit, in the area frame the sampling unit must also be classified NOL} \\ 0 & \text{otherwise,} \end{cases}$$

x_{hij} = variable to be estimated (for domain estimator $(2)\hat{X}_h$ variable only summed across i^{th} PSU in stratum h) in sampling unit,

g = 1,2,3,4 frame and subpopulation (Table 2),

h = stratum level,

i = primary sampling unit (PSU) level: (firm or segment),

m_{hi} = sampled number of subunits (list frame establishments do not sum over m_{hi} subunits), and

n_h = sample size of PSU in stratum h.

Domain total estimates for each stratum of a frame were added to get the household (employee) domain total estimates $(3)\hat{X}$ and $(4)\hat{X}$. The same procedure was used to estimate establishment domain totals $(1)\hat{X}$ and $(2)\hat{X}$.

If a firm with multiple operations could not report data separately for the selected sampling unit or was duplicated on the establishment list, an adjustment was required using the proration factor (P_{hi}). The factor was always one unless an adjustment was needed.

The unbiased sample estimate of the variance for household (employee) and establishment domain totals $_{(g)}\hat{X}_h$ in stratum h was

$$v(_{(g)}\hat{X}_h) = \frac{n_h}{n_h - 1}\left\{\sum_{i}^{n_h}\left[(f_{oh})\ _{(g)}\hat{X}_{hi} - {}_{(g)}\bar{\hat{X}}_h\right]^2\right\} \qquad (2)$$

where, $_{(g)}\hat{X}_{hi}$ = sample total in the i^{th} PSU of the h^{th} stratum:

$$_{(1)}\hat{X}_{hi} = \sum_{j}^{m_{hi}} (X_{hij})$$

$$_{(2)}\hat{X}_{hi} = (P_{hi})(X_{hi}) \ ,$$

$$_{(3)}\hat{X}_{hi} = \sum_{j}^{m_{hi}} \left[(P_{hij})(X_{hij})\right] ,$$

$$_{(4)}\hat{X}_{hi} = \sum_{j}^{m_{hi}} \left[(P_{hij})(X_{hij})\right] , \text{ and}$$

$_{(g)}\bar{\hat{X}}_h = \frac{_{(g)}\hat{X}_h}{n_h}$ = overall sample mean per stratum,

The variance estimate was computed based on only the first stage of sampling because the subunits were self weighting.

Composite Total Estimates

To obtain an estimated subpopulation total for an attribute, strata estimates were first summed to domain totals for a frame. Composite estimates were then generated by summing the domains (NOL area estimate and the list estimate). This procedure, shown in Table 3 below, illustrates the firm and employee (household) composite total estimates $_{T_1}\hat{X}$ and $_{T_2}\hat{X}$ respectively. The corresponding variance estimates for the firm $v(_{T_1}\hat{X})$ and employee (household) $v(_{T_2}\hat{X})$ were also obtained by summing domain total variances.

Table 3: Combining Domain Estimates

Composite Total Estimates	Composite Variance Estimates	Subpopulation Estimate Type
$\hat{T}_1 = \hat{(1)} + \hat{(2)}$	$v(\hat{T}_1) = v(\hat{(1)}) + v(\hat{(2)})$	Composite Firm Estimate
$\hat{T}_2 = \hat{(3)} + \hat{(4)}$	$v(\hat{T}_2) = v(\hat{(3)}) + v(\hat{(4)})$	Composite Employee Estimate

For example, to estimate total employment in the study area the composite estimate $(\hat{T}_2)$ was composed of two parts. The first component $(\hat{(3)})$ was the NOL domain estimate from the household area frame. The second domain estimate $(\hat{(4)})$ was the total employee estimate from the list frame. The total number of establishments was estimated in a similar manner with the composite estimate $(\hat{T}_1)$ the sum of the area nonoverlap estimate $(\hat{(1)})$ and list overlap estimate $(\hat{(2)})$.

Estimating Household (Employee) Means

A combined ratio estimator $(\hat{R})$ was used to estimate means for household (employee) characteristics. The estimator is a ratio of two (dependent) random variables X and Y which both vary from unit to unit. The denominator was a random variable because total units in the household subpopulation were not known.

To compute $\hat{R} = (\hat{T}_2 / \hat{T}_2)$ the estimated composite totals for Y and X were required. Computation of $\hat{T}_2$ for the variable of interest was shown in Table 3. The composite estimate $\hat{T}_2$ for the new variable Y was computed using the same approach with domain totals first estimated. The domain totals $\hat{(3)}$ and $\hat{(4)}$ were then combined to a composite total $(\hat{T}_2)$.

The combined ratio estimator for estimating household (employee) averages was

$$\hat{R} = \frac{\hat{T}_2}{\hat{T}_2} \qquad (3)$$

Average income of retail employees was an example using the combined ratio estimate. The total estimated income for retail employees formed the numerator and the total estimated employees in the retail trade was the denominator.

The estimated variance for employee averages using the combined ratio estimator was

$$v(\hat{R}) = \frac{1}{{}_{T_2}\hat{X}^2}\left[v({}_{T_2}\hat{Y}) - 2\hat{R}\,\mathrm{Cov}({}_{T_2}\hat{X},\ {}_{T_2}\hat{Y}) + \hat{R}^2 v({}_{T_2}\hat{X})\right] \tag{4}$$

where,

$v({}_{T_2}\hat{Y})$ = composite variance estimate for Y computed in like manner as shown in Table 3,

$v({}_{T_1}\hat{X})$ = composite variance estimate for X computed in Table 3, and

$$\mathrm{Cov}({}_{T_2}\hat{X},\ {}_{T_2}\hat{Y}) = \sum_g \left[\mathrm{Cov}({}_{(g)}\hat{X},\ {}_{(g)}\hat{Y})\right] \tag{5}$$

was the sample covariance between ${}_{T_2}\hat{X}$ and ${}_{T_2}\hat{Y}$ where,

$$\mathrm{Cov}({}_{(g)}\hat{X},\ {}_{(g)}\hat{Y}) = \sum_h \left\{ \frac{n_h}{n_h - 1} (f_{eh})^2 \left[\sum_i^{n_h} X_{hi}\, Y_{hi} - \frac{\sum_i^{n_h} X_{hi} \sum_i^{n_h} Y_{hi}}{n_h} \right] \right\} \tag{6}$$

Estimating Establishment Means

The firm mean estimate (${}_{T_1}\hat{\bar{X}}$) used the composite total estimate ${}_{T_1}\hat{X}$ in Table 3 with corresponding variance estimate $v({}_{T_1}\hat{X})$. The combined ratio estimator (3) was not used since the denominator of the estimator (N') was not a random variable. A typical characteristic of interest was average establishment employment. The generalized formula for the establishment mean estimate was

$${}_{T_1}\hat{\bar{X}} = ({}_{T_1}\hat{X}/N') \tag{7}$$

where,

$N' = \sum_g {}_{(g)}N$ and,

g = 1,2; Area and list frame establishments, and

${}_{(g)}N$ = total firms in frame source of interest.

The corresponding variance estimate for firm means was

$$v(_{T_1}\hat{\bar{X}}) = \frac{1}{(N')^2} \; v(_{T_1}\hat{X}). \tag{8}$$

Nonresponse Adjustment

To summarize the data it was necessary to deal with the problem of nonresponse. For this study nonresponse was defined as a refusal, inaccessible, or nonusable questionnaire. The general formula to adjust the expansion factor for nonresponse was

$$f_{ah} = \frac{C}{(C-B)} \; (f_{oh}) \tag{9}$$

where,

f_{ah} = adjusted expansion factor in stratum h,

C = total number of units sampled in stratum/substratum,

B = total number of nonresponse units in stratum/substratum, and

f_{oh} = original expansion factor.

The original expansion factor (f_{oh}) was only adjusted in a stratum or substratum for nonresponse. The adjusted expansion factor then replaced the original factor in equations (1), (2), and (6).

References

[1] Cochran, Robert S. "Theory and Applications of Multiple Frame Surveys." Ph.D. dissertation. Iowa State University, 1965.

[2] Cochran, Willian G., Sampling Technique, third edition, New York: John Wiley and Sons, 1977.

[3] Hartley, H. O. "Multiple Frame Surveys," paper presented at American Statistical Association meeting, Minneapolis, Minnesota, September, 1961.

[4] Kleweno, Douglas G., Applications of the Multiple Frame in an Economic Distributional Effects Study. U.S. Department of Agriculture, Economics and Statistics Service, paper to be published in 1980.