EVALUATING NEW REPORTER SOURCES FOR A USDA PRICE SURVEY

Carol C. House, U.S. Department of Agriculture

ABSTRACT

A difference was perceived between the sampling frame used in a periodic USDA survey to estimate the prices farmers pay for production inputs and the population of establishments selling to farmers. This led to an experiment with new reporter sources. The existing frame consisted of a list of firms constructed primarily from the yellow pages, but its completeness was unknown. An additional problem existed for certain commodities when a large percentage of firms were located in metropolitan areas. A different price structure could exist there which would not reflect sales to farmers. Farm operators in a national expenditure survey were asked for the addresses of places where they purchased various commodities and for their expenditures at each place. This new frame allowed for a self-weighting sample of establishments with selection probabilities proportional to the estimated business conducted with farmers. The domains of the overlapping frames and the effect of recall errors in frame construction were examined. The technique has application to many situations where construction of a complete sampling frame is prohibitive.

INTRODUCTION

One problem frequently encountered by survey practitioners is the lack of an adequate sampling frame to cover the population of interest. Frames easily or cheaply obtained are often incomplete or quickly outdated. This paper discusses a frame building technique that can be cost efficient and easy to carry out. The frame is substantially complete and designed to be self-updating. The procedure requires a second population from which knowledge can be obtained of the first. A frame must be available to survey this second population.

Difficulties with a series of government agriculture price surveys served as the catalyst for investigating the technique. The list frame of business establishments for these price surveys was inadequate. Supplementing with a standard area frame would be inefficient at best. A study was conducted in three states during January 1980. Respondents on a national farm expenditure survey in these states were asked for names and addresses of establishments where they had purchased various production inputs. These names formed a new frame for sampling firms and were checked against the current list frame to identify and analyze the overlap domains.

We expected data collection problems. The most significant were recall errors from respondents and the difficulty in correctly identifying the overlap domain when one list source contained local or incomplete names and addresses. In practice, the problems proved manageable.

This report covers background information on the price surveys, a description of the frame building technique, properties of the frame and the estimator for average price, data collection and feasibility, and an analysis of the overlap domain with the current frame.

DEFINITION OF TERMS

Diagram 1 on the last page shows a schematic representation of the sampling frames and domains discussed in this paper. The frames are numbered from one to five and the domains are assigned letters. When a particular frame, domain or subdomain is discussed later in the paper, it will be referred to by the name assigned there.

BACKGROUND

The United States Department of Agriculture (USDA) uses a yearly series of surveys to estimate prices that farmers pay for various production inputs such as feed, fuel, and farm machinery. The Government uses these estimates, and the indexes derived from them, to make policy decisions about farm programs and compute parity prices and various income series used as national economic indicators. The desired sampling unit for the survey was a business establishment which sold at least one of a specified list of commodities to farmers. Potential reporters were identified primarily from advertisements and listings in yellow page telephone directories. Frame 5, the list built by this method, had disadvantages: 1) incompleteness, 2) inclusion of units in the frame from outside the population, 3) difficulty in obtaining information on size of establishment for proportionate sampling or weighting, and 4) the cost involved in building and maintaining the list source. A discussion of these appears below.

Incompleteness

The population consists of establishments selling any one of a specified list of farm production inputs to farmers. Many of these establishments do not advertise in the yellow pages. Of those that do advertise, it is not always clear from the advertisement whether they sell a specific commodity. The severity of the incompleteness problem is dependent on the specific item for which a price is to be estimated. For example, a yellow page produced list of firms selling new trucks should be more complete than one for firms selling fencing materials or firms selling spark plugs. These items are sold in different types of establishments in almost every town. Thus, a large discount drug store might advertise in the yellow pages but probably not in such a way as to enable one to tell if they sell spark plugs. As a further example, many farmers buy feed through a local farmer who is the distributor for a feed company. These distributors generally depend on "word-of-mouth" advertisement to become known.

Statisticians frequently use an area frame to estimate for the incompleteness of a list source.

In this situation, an establishment selling to farmers is a "rare item", and thus an area frame is inefficient for this purpose.

Inclusion of Units from Outside Population

Another standard solution to frame incompleteness is the intentional inclusion of units having a low probability of being in the population. Screening then becomes the first step in the survey procedures. Unfortunately, this method causes a decrease in efficiency as well as incompleteness as the number of screen-outs increase.

An additional problem affects the price surveys because the population of interest excludes establishments not frequented by farmers. Individual stores do not know how many of their customers are farmers, and thus this information can not be obtained in a screening interview. Establishments are not expected to differentiate in their prices between farmers and the general public. However, we expect that farmers may buy proportionately more at different types of stores serving the public. For example, lists of firms obtained from the yellow pages are dominated by metropolitan areas where farmers are less likely to do business.

Proportionate sampling or weighting

The purpose of the surveys is to estimate the average price paid by farmers for a particular production item. Because quantity of sales frequently affects the price set by an establishment, firms should be sampled proportional to sales of the commodity or the data should be weighted to reflect size. Yellow page advertisements do not give sufficient information to calculate unequal selection probabilities. The author has been disappointed in efforts to obtain "quantity sold" data during a price survey. Respondents have considered the information confidential or too much trouble to look up.

Cost

The costs of building extensive lists of firms are great in terms of staffing requirements and time constraints. Additional resources must still be expended to compensate for the inadequacies of the procedures. What is even more unfortunate, the effort at maintenance must begin as soon as the frame is constructed. The Bureau of the Census estimates that over a third of all establishments in a monthly Current Business Survey undergo a change (coming into existence, going out of business, merging, splitting) each year. (Wolter and others) Frame 5 has to be virtually reconstructed each year.

NEW FRAME - DESCRIPTION AND ESTIMATION

General Description

The most accurate method of identifying the establishments where farmers make purchases is to ask the farmers directly. This was done in a study conducted by the USDA in February 1980, as part of the 1979 Farm Production Expenditure Survey (FPES).

The FPES is a national survey used to estimate annual expenditures for all aspects of farm production. The sample was drawn from a multiple list (Frame 1) and area (Frame 2) frame to provide efficient and complete coverage of the population of farm operators. During the survey, respondents recorded expenditures for every area of farm production and then gave the names and addresses of business establishments where they purchased new machinery, supplies, feed or fertilizer. They also provided the percent of expenses associated with each establishment. In this manner we were able to construct a list frame, Frame 4, of establishments for each commodity. The expenditure data associated with each firm allowed a self-weighting sample to be drawn with probability proportional to the estimated dollar business by farmers.

This method of list construction had several clear advantages over the traditional method. First, establishments on Frame 4 have known sales to farmers. For example, the frame excludes firms in large metropolitan areas if they did not do business with farmers. Secondly, one can easily draw a self-weighting sample of firms with probabilities of selection proportional to the estimated amount of expenditures by farmers in the firm. The calculation of probabilities requires no additional "size" information from the establishments themselves. The cost of constructing the frame is minimal. The additional enumerator time during the survey to collect the names and addresses was the chief component of cost. The second component was staff-hours of editing to complete any incomplete addresses, etc. This component was only necessary for firms selected in the sample. Finally, since the FPES survey is run each year to estimate expenditures, Frame 4 can be rebuilt annually, insuring that the sample of establishments reflects current farmer buying patterns.

Completeness

The degree of completeness of the frame must be looked into more carefully. If the FPES survey contacted all farm operators for information about their purchasing practices, this list of establishments, called Frame 3 would be substantially complete. It would, however, be subject to respondent recall errors. (Establishments omitted in this way are suspected to be those with which the farmer had relatively small transactions.) Frame 4 is subset of Frame 3. The relationship is presented statistically in two ways.

One formulation defines Frame 3 as the complete sampling frame on which the survey is built. A multistage cluster sample of establishments is then selected from this frame. In the first stage of sampling, a selection of farm operators identifies the clusters of establishments. Each farm operator uniquely defines the set or cluster of establishments in which he did business. The clusters are overlapping. These first stage units are collected on the FPES and become Frame

4. The second stage of sampling involves the selection of firms from the clusters to achieve an estimate of average price.

One can view the situation statistically in another way. Each sample of farm operators is uniquely associated with a frame of establishments. Thus, considering the set of all samples of operators, we have a corresponding set or population of frames. To conduct a survey, one selects a "random frame" from the population and uses it to sample establishments. This becomes Frame 4. A single frame selected in this manner is not complete. However, the author has examined the expected "average price" estimated from these frames for a simplified case and compared it to the population parameter $\bar{Y}$ of "average price per unit sold to farmers." In the example, the estimate is biased but the bias is the type that decreases with the sample size. Sample size is the key to possible problems with the estimator. Because the expected value is taken over a population of frames, "sample size" refers to the number of "random frames" selected for repetition. The procedure calls for one repetition, Frame 4. The estimate of price depends on how representative Frame 4 is of Frame 3. (The example is worked out at the end of this section, as it can be presented more easily after the estimators are discussed.)

Selection of Establishments

Each farm operator reported his or her total expenditures for an item, the establishments where he made the purchases, and the percent of the total expended at each place. The sampling procedures select firms with probability proportional to estimated dollar expenditures by farmers. The expenditure estimates are based on a multiple frame (Frame 1 and Frame 2) of operators using the multiple frame estimator, (Hartley, 1962),

$$E = pE_1 + (1-p)\,E_2 + E_3 \qquad (1)$$

Where

E = multiple frame estimator of total expenditures

E_1 = total expenditures estimate from Frame 1

E_2 = total expenditures estimates from Frame 2 operators that overlap with Frame 1

E_3 = total expenditures estimate from Frame 2 operators that do not overlap with Frame 1

p = weight

The three groups of establishments, those generated from operators from the overlap and nonoverlap domains in Frame 2, are sampled independently. The desired sample size of establishments, n, is allocated between these groups in the following manner.

$$n_1 = n\frac{E_1}{E}\,p = \text{sample size allocated to establishments generated from frame Frame 1 (Subdomain A1 and B1)} \qquad (2)$$

$$n_2 = n\frac{E_2}{E}\,(1-p) = \text{sample size allocated to establishments generated from Frame 2 operators that are also in Frame 1} \qquad (3)$$

$$n_3 = n\frac{E_3}{E} = \text{sample size allocated to establishments generated from Frame 2 operators that are not in Frame 1} \qquad (4)$$

Once the sample is allocated between the three groups, the selection of individual establishments is proportional to the firm's estimated total sales to farmers. We define E_{ij} to be the expenditure of operator "i" in establishment "j". The selection probability of a given firm, f, becomes:

Prof(f) = Prob that f is selected from Group 1
+ Prob that f is selected from Group 2
+ Prob that f is selected from Group 3

$$= \sum_{G=1}^{3} \frac{\sum_{i=1}^{S_G}\sum_{j=1}^{m_i} (E_{ij})\,(Z_{ij}(f))}{\sum_{i=1}^{S_G}\sum_{j=1}^{m_i} E_{ij}} \qquad (5)$$

where,

$$Z_{ij}(f) = \begin{cases} 1 \text{ if f is the "j" firm frequented by operator "i"} \\ 0 \text{ otherwise} \end{cases}$$

m_i = number of firms in which operator "i" did business

S_G = number of operators in sample in "G" group

Estimator of Price

The self-weighting sample of establishments allows price data from all selected firms to be averaged together to estimate the "average price per item sold to farmers." Let $\{f_{1,1}, \ldots, f_{1,n_1}\}$, $\{f_{2,1},\ldots,f_{2,n_2}\}$, and $\{f_{3,1}, \ldots, f_{3,n_3}\}$ be the samples selected from the three groups of establishments. We define $P_{k,h}$ to be the price reported by firm $f_{k,h}$.

The FPES estimator is:

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{k=1}^{3}\sum_{h=1}^{n_k} P_{k,h} \qquad (6)$$

The estimate is biased, because we are estimating price, which is a ratio. The estimated variance is:

$$v(\hat{\bar{Y}}) = \frac{1}{n(n-1)}\sum_{k=1}^{3}\sum_{h=1}^{n_k} (P_{k,h} - \hat{\bar{Y}})^2 \qquad (7)$$

Example

In the following example, the expected value over the population of frames of the estimate of price is calculated for the simplified case where a random sample of "n" operators was drawn from

a population of size "R".

We define

R = population size of operators
n = sample size of operators
F = $\{F_1, \ldots, F_N\}$ = population of frames of establishments.

Since there is a "one-to-one/onto" correspondence between frames and "samples" of farm operators, then the number of frames is :

$$N = \frac{R!}{n!(R-n)!} \tag{8}$$

The population parameter $\bar{Y}$ to be estimated is:

$$\bar{Y} = \frac{\text{total expenditures for item by operators}}{\text{total quantity sold to operators}}$$

$$= \frac{\sum_{r=1}^{N}\sum_{i=1}^{n} E_{ri}/M}{\sum_{r=1}^{N}\sum_{i=1}^{n} Q_{ri}/M} = \frac{\sum_{r=1}^{N}\sum_{j=1}^{n} E_{ri}}{\sum_{r=1}^{N}\sum_{i=1}^{n} Q_{ri}} \tag{9}$$

where

E_{ri} = expenditures for item by operator "i" associated with frame "r"

Q_{ri} = total quantity purchase by operator "i" associated with frame "r"

$\frac{1}{M} = \frac{R}{N}$ = probability of a given operator being associated with a given frame. Since each operator is associated with more than one frame, this probability is multiplied by the operator's expenditures and quantity purchased.

For a given frame, F_r, the population parameter to be estimated is:

$$\bar{Y}_r = \frac{\sum_{i=1}^{n} E_{ri}}{\sum_{i=1}^{n} Q_{ri}} \tag{10}$$

The estimator for this, $\hat{\bar{Y}}_r$, is equation (6), described earlier in the paper. The estimate is biased since we are estimating a ratio (price), but the bias, B_r, decreases with the sample size of establishments. (Murthy, 1977). Thus,

$$E(\hat{\bar{Y}}_r) = \bar{Y}_r + B_r \tag{11}$$

Once a frame, F_r, is selected, the estimate, $\hat{\bar{Y}}$, of the population parameter (9) is defined to be

$$\hat{\bar{Y}} \equiv \hat{\bar{Y}}_r$$

The expected value of this estimate is:

$$E(\hat{\bar{Y}}) = E_{F_r \in F}[E(\hat{\bar{Y}} \mid F_r)]$$

$$= E_{F_r \in F}[E(\hat{\bar{Y}}_r \mid F_r)]$$

$$= \sum_{r=1}^{N} [E(\hat{\bar{Y}}_r \mid F_r)] \text{ Prob}(F_r)$$

$$= \frac{1}{N}\sum_{r=1}^{N} [\bar{Y}_r + B_r]$$

$$= \frac{1}{N}\sum_{r=1}^{N} \bar{Y}_r + \frac{1}{N}\sum_{r=1}^{N} B_r$$

$$= \frac{1}{N}\sum_{r=1}^{N} \bar{Y}_r + B \tag{12}$$

The bias associated with $\hat{\bar{Y}}$ is:

$$B(\hat{\bar{Y}}) = \left| \bar{Y} - E(\hat{\bar{Y}}) \right|$$

$$\leq \left| \frac{\sum_{r=1}^{N}\sum_{i=1}^{n} E_{ri}}{\sum_{r=1}^{N}\sum_{i=1}^{n} Q_{ri}} - \frac{1}{N}\sum_{r=1}^{N} \frac{\sum_{i=1}^{n} E_{ri}}{\sum_{i=1}^{n} Q_{ri}} \right| + B$$

$$= B' + B \tag{13}$$

B' is a bias associated with the estimation of a ratio, and will decrease with the sample size of frames selected from the population of frames. B is the average of B_r's, and will decrease with the sample size of establishments.

FEASIBILITY

The study was conducted during February, 1980 in three states to test the feasibility of implementing this method on a national scale. State 1 is a large midwestern state, State 2 is a southern state and State 3 is a sparsely populated western state. Fertilizers and chemicals, mixed feeds, farm supplies, and new machinery and tractors represented the various product groups associated with the price surveys. During the Farm Production Expenditrue Survey in these states, farm operators constructed a list frame (Frame 4) of establishments for each commodity group. Office personnel compared firm names on Frame 4 with firms on the current sampling frame (Frame 5) in the price surveys, and coded each as a match or nonmatch with Frame 5.

We anticipated two problems involving respondent recall errors. First, the operator might completely fail to mention a firm in which he had done business. For example, he may not recall every store where he purchased nails. We had no method of counting the number of establishments omitted in this way, but suspect that they are firms with which the farmer had relatively small transactions. Second, The respondent might have trouble giving complete firm names and addresses. For example, a firm by the name of "A & A Feed" might be known locally as "Jones Feed" (the owner's name), or respondents may not have a recollection of the street address of the firm. Also, an establishment might be located in a crossroads community sharing a postal designation with a nearby town. The respondent may provide the local community name instead of the postal designation.

To keep overlap checking procedures consistent throughout the three states, personnel followed a complicated system of instructions to identify the matches and nonmatches of names and addresses. A flowchart of the decision process is found in the appendix. Any pair of names and addresses

determined to be neither a definite match nor nonmatch on the decision chart, received further investigation. This process included searching telephone directories, contacting individuals with personal knowledge of the community involved, and/or contacting the establishment. In the end, each potential matching pair became a match or nonmatch. Personnel kept records as they reached each decision in order to evaluate the type of effort needed to perform the overlap check.

The results were encouraging and feasible. Enumerators reported that farm operators generally did not have difficulty listing the places where they made purchases, but they did have some problem with street addresses and percentages. The enumerators supplied street addresses themselves in many cases by information obtained from phone books, post offices and other respondents. In the office, the majority of the names on Frame 4 were declared a match or a nonmatch from the decision chart without further investigation. When more information was required, it was gerally available in a telephone book.

COMPARISONS OF OVERLAP AND NONOVERLAP DOMAINS

The collection of names and addresses on the FPES survey created Frame 4. Analysis of the expenditure data associated with each firm and the overlap checking between Frame 4 and Frame 5, confirmed the hypothesis that a substantial percentage of total farm-related expenditures for feed, fertilizer, new farm machinery and farm supplies were made in establishments not identifiable through yellow page advertisements. Domains A, B, C and the various subdomains are discussed below.

Domain C will receive little mention in this paper. The domain consists of two types of firms--one type that sells to farmers and the other type that does not. Firms of the first type are actually "overlap" with Frame 3. Frame 4 is a subset of Frame 3. Firms of the second type are not frequented by farmers and thus not in the population of interest. They should be removed from Frame 5. Unfortunately we have not discovered a realistic way to distinguish between the two types.

Domain B consists of firms that are on both Frame 4 and Frame 5. This domain is divided into two subdomains for analysis, B1 and B2. Firms in subdomain B1 were reported by operators sampled from Frame 1. Firms in Subdomain B2 were reported by operators sampled from Frame 2.

Establishments that appear on Frame 4 but do not on Frame 5 are in Domain A. These are firms that sell to farmers but were not identifiable through the construction techniques which produced Frame 5. Like B, this domain is divided into two parts for analysis. Subdomain A1 consists of establishments reported by operators sampled from Frame 1. Subdomain A2 consists of establishments reported by operators sampled from Frame 2.

For each of the four commodities in the three states, operators' expenditures at each firm were calculated and summed to the state level for subdomains A1, A2, B1, and B2. The mean expenditure per operator per firm was computed for each subdomain. The means between subdomains A1 and B1, and A2 and B2 were compared with t-tests. The results are tabulated in the appendix.

The results show that Frame 5 does not adequately cover the population of establishments serving farm operators. Even with farm machinery dealers (where yellow page produced lists were suspected to be the most complete), the median percent of total expenditures in Domain A firms was 22 percent. Considering all commodities, the percent of expenditures in Domain A firms ranges from a low of 2 percent to a high of 83 percent. T-statistics were computed to test the hypothesis that mean expenditures per operator in Domain A firms and Domain B firms were equal. For three of the four commodities tested, the data failed to produce a rejection of the null hypothesis in more than one state (90 percent confidence level). For farm supplies in all states, the mean expenditure in Domain A firms was significantly larger than the Domain B firms. Below are some specific observations for each commodity.

Feed

Farm operators purchased feed in both Domain A and Domain B firms, with comparable expenditures in each. In State 3, the total expenditures in Domain B firms was larger than in Domain A firms. In State 1 a difference in estimates of the percent of total expenditures existed between Subdomains A1 and A2. Firms in Subdomain A2 accounted for 72 percent of the total expenditures. An examination of the data revealed that a single large transaction, expanded by Frame 2 expansion factors, had a large influence on this. The mean expenditures per operator were significantly different at the 90 percent confidence level in State 3 for Subdomains A1 and B1. All other tests did not show a significant difference.

Fertilizer

The results of t-tests failed to show significant differences between the mean expenditures for fertilizer in Domain A and B firms in any state. The estimate of total expenditure between the domains were comparable, with expenditures estimated from Domain A slightly higher.

Farm Machinery

Domain B dominated the market for farm machinery, although there were more transactions with Domain A firms than predicted. The mean expenditure per operator was greater in Domain B firms for all states. They were statistically different in State 1 and State 3. State 3 had an estimate of 2 percent of total expenditures in Subdomain A1 firms, which was the lowest percentage for any commodity. In the other states, the percent at Domain A firms ranged from 17 percent to 30 percent.

Farm Supplies

Farm supplies are defined to be such items as shovels, nails, hammers, chain saws, and milk pails. We anticipated that Frame 5 would be the most incomplete for this commodity. As predicted, the mean expenditure estimates in Domain A firms

were judged significantly higher in all cases except between Subdomain A2 and B2 in State 3. Even here, the estimate from A2 was $646 as compared to the B2 estimate of $363. Percents of expenditures in Domain A firms were in the eighties for State 1 and 2 and somewhat lower for State 3.

CONCLUSIONS

The results showed that a large percentage of expenditures by farm operators for feed, fertitizer, farm machinery, and farm supplies are made at establishments not in Frame 5. This has not established that the estimates of average price paid by farmers have been affected by the omission of these firms. A sample of establishments has been drawn from Frame 4, and the estimated average prices will be compared with the estimates from Frame 5.

The technique of frame construction discussed in this paper provides a sampling frame that substantially covers the population of establishments selling the specified items to farmers. A PPS sample can be easily drawn to provide a self-weighting estimate of average price from the USDA price surveys.

The enumerators were able to collect names of establishments from the operators with a minimum of problems. Percentages and street addresses were more difficult to obtain, but feasible. Farm operators reported names and addresses completely enough to allow overlap checks with a second frame, and to allow sample units to be properly identified in the field. The cost of building the frame was minimal in terms of respondent burden and office time. It is a self-updating system, reflecting changes in market situations.

The sample on the FPES survey must be sufficiently large to provide an adequate list of establishments from which to sample. Otherwise, sampling with probability proportional to size will tend to clump sample units toward firms reported by a few large operators.

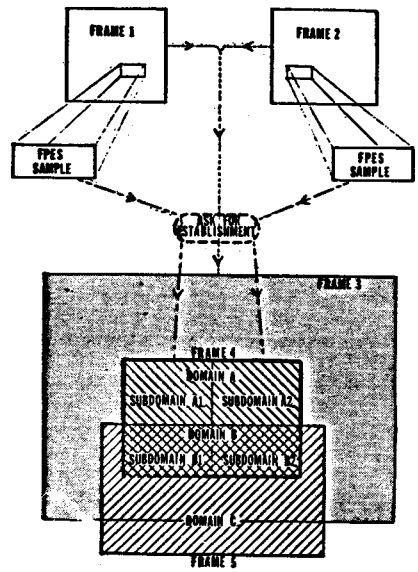
REFERENCES

Cochran, William G. Sampling Techniques 3rd Ed. New York: John Wiley & Company, 1977.

Murthy, M.N. Sampling Theory and Methods 2nd Impression. Calcutta: Statistical Publishing Society, 1977.

Wolter, Kirk M., and others. Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys. U.S. Bureau of the Census.

Diagram 1--Schematic of frames and domains



Frames

Frame 1- List of farm operators.
Frame 2- Area frame of operators.
Frame 3- Theoretical complete frame of establishments selling to farmers constructed by asking every farm operator where they made purchases.
Frame 4- Frame of establishments, constructed by asking operators on FPES sample where they made purchases.
Frame 5- Traditional list frame of establishments, constructed primarily through yellow page directories.

Domains

Domain A- Establishments on Frame 4 but not on Frame 5.
  Subdomain A1- Establishments in Domain A generated by sample of operators from Frame 1.
  Subdomain A2- Establishments in Domain A generated by sample of operators from Frame 2.
Domain B- Establishments on Frame 4 and Frame 5.
  Subdomain B1- Establishments in Domain B generated by sample of operators from Frame 1.
  Subdomain B2- Establishments in Domain B generated by sample of operators from Frame 2.
Domain C - Establishments on Frame 5 but not on Frame 4.

** The appendix tables have been omitted to observe the space restrictions for publication. They are available from the author by request.