LINKING MULTIPLE STRATIFICATIONS: TWO PETROLEUM SURVEYS

By Pedro J. Saavedra, Ph.D.

The literature provides extensive treatment on methods for allocating a sample where there are multiple objectives to be met. Most of these methods assume a pre-established stratification, a small number of variables for which estimates must be obtained and either few subpopulations or the assumption of similar structures in subpopulations. The purpose of this paper is to present an account of a solution that was implemented for a survey where none of the above assumptions were met, and the solution that was implemented. A secondary purpose is to present two unsolved problems associated with the current sampling design. A historical approach is taken here, since it is important to realize how some of the solutions were arrived at and why some other possible options were not exercised.

This paper discusses two surveys: the EIA-782 and the EIA-821. The design for the EIA-782 came first. This is a monthly survey of petroleum volumes and prices. It includes the EIA-782A, a census of refiners, and the EIA782B, a survey of other dealers. Its frame was then the EIA-764, now supplanted by the EIA-863. The following section discusses the EIA-782 and its frame.

Discussion Of The EIA-782 And Its Frame

The EIA-782B is a price and volume monthly survey which covers the fifty states and the District of Columbia, and includes sales of distillate fuel oil, residual fuel oil and motor gasoline to end users and resellers. The EIA-782B does not include refiners, since these are covered by the EIA-782A (which includes all of them as a census).

The EIA-782B publishes prices and volumes for residual oil and motor gasoline for all fifty states (plus D.C.), but only publishes distillate values for twenty-four states and for each of the Petroleum Allocation Defense Districts (PADD).

The frame for the EIA-782B is the EIA-863. The EIA-863 includes refiners, and these need to be used in the sample design since the EIA-782B results are published in conjunction with those of the EIA-782A. The EIA-863 includes yearly volumetric information for seven products:

- Residential No. 2 distillate
- Nonresidential No. 2 distillate
- Wholesale No. 2 distillate
- Retail residual oil
- Wholesale residual oil
- Retail motor gasoline
- Wholesale motor gasoline

In addition the EIA-863 includes:

- A question regarding whether any of six other products are sold by the dealer:
  - No. 1 distillate
  - Crude oil
  - Propane
  - Other LPG
  - No. 4 fuel oil
- Additional questions regarding operational status, subsidiary-parent relationships, and similar information.

The Original Design

Since the frame had only volumes, on an annual basis, with fewer divisions of the various products than were published in the EIA-782 report, it was decided that the best way to decide on allocations was to investigate the relationship between volume C.V.s from the frame and price C.V.s from the sample. This led to the conclusion that in order to obtain acceptable estimates at the state level, one had to obtain certain target volume C.V.s for each of the products appearing on the frame file and for each of the states (actually some states were not publication states for distillate, but the EIA-782 publishes residual and motor gasoline prices for every state).

A preliminary investigation revealed that the intercorrelation of volumes, particularly after the very large companies (which one would want to sample with certainty) were separated, was very different for different states. In other words, attempts to stratify using combinations of products or to use principal components would have to be carried out separately for each of the fifty states and the District of Columbia.

To complicate matters the frame is a somewhat dynamic frame. New information about companies which have gone out of scope, merged or sold and corrections of data result in frequent updating of the frame. Given somewhat skewed distributions, and a complex processing system, the effort of designing a different approach for each state, where that approach might have to be changed from cycle to cycle and where the system would have to be programmed differently for each state, was not considered justified.

The thought of three different surveys, one for distillate, one for residual and one for motor gasoline, was then given serious consideration. One the one hand, such an approach would cost more and would involve a greater burden (requiring probably more respondents than a single survey). On the other, there seemed to be no easy way of allocating for the multiple estimates that the survey required.

A compromise approach was then suggested of having three surveys maximizing the overlap between the surveys'respondents. The three surveys would share the same certainty strata, and each would be designed using standard methods (i.e. Dalenius-Hodges procedure for stratification, Neyman allocations, etc.). In the end, the three samples would be drawn using the same random sort of the population (in other words the same random sort would be used to fill the quotas in each sample's stratum).

This approach assumed that different weights would be used for each survey, and that only companies sampled for a particular product and state would be used to obtain estimates for that state. After experimenting with various possible forms, it was decided that it was easier to use the same questionnaire than to have three separate ones. It was thought that this was a cost-effective simplification that would create no problems. The difficulty was a non-statistical one. It was unacceptable to collect data which one will not use in obtaining estimates.

This problem led to what was perceived as a natural solution--include in the one sample any company selected for any of the three, and obtain probabilities of selection and use Horvitz-Thompson type estimators as weights. As the decision was made, no one anticipated the fact that several years later nobody has yet come up with an analytical formula for the probability of selection.

When it became obvious that probabilities of selection could not be empirically obtained, the search for an alternative began. There were three separate stratifications - distillate, residual and gasoline - with three to nine noncertainty strata for each (plus one nonrespondent stratum) in each state. Attempting to cross the strata would lead to many small cells, increasing considerably the sample size. Combining products had failed at the earlier stages of the project. Finally, the suggestion of a computer simulation of the probability of selection was made and accepted.

Thus, the design of the first cycle of the EIA-782 where the linked stratifications concept was first implemented was set up in its most basic form. It had the following features:

- A certainty stratum was defined in each state, consisting of:
  - Refiners.
  - Companies reporting in the frame doing business in more than three states.
  - Companies reporting over 5% of the volume for any product in any state.
- A nonrespondent stratum was defined in each state.
- The Dalenius-Hodges procedure was used to obtain stratum boundaries using the seven products, but replacing nonresidential retail distillate and wholesale distillate for the maximum of the two.
- A distillate stratification was made crossing residential retail with the maximum of the other two distillate variables. Nine noncertainty respondent strata were defined (zero, low and high on each variable). The strata were separately defined in each publication state and in each groups of nonpublication states found within a PADD.
- For motor gasoline seven strata were defined, since companies selling only retail gasoline were considered out of scope. However, since many companies sold no wholesale gasoline, but did sell distillate or residual, the number of strata was changed to nine at a later date.
- For residual the maximum of three levels (zero, low and high) defined by the two products was used to create three strata.
- Since the frame used annual data and since it was outdated, it was obvious that estimates of variance would be low, so after matching prior years' monthly data with a frame, inflation factors were established. The standard deviation of each cell was multiplied by an inflation factor prior to implementing the Neyman allocation program.
- A minimum of two company/state units (CSU) was sampled from each cell.
- To each cell was assigned the maximum allocation from the two or three products which defined the cell (in other words, a separate Neyman's allocation was done for retail and resale gasoline and the maximum allocation for each cell was used).
- Fifty percent of the combined sampling fraction of the noncertainty cells was used as the allocation for the corresponding nonrespondent stratum frame allocation cell.
- The sample was drawn by shuffling the frame once. The same order was used to fill the cells in each of the three stratifications. If a CSU was selected for one of the three samples it was in the sample.
- One thousand samples were drawn to obtain probabilities of selection. Probabilities were averaged for CSUs sharing the same combination of cells.
- Horvitz-Thompson type estimators were used to estimate prices and volumes from the data , but variances were calculated on only the data of respondents selected for the particular product. It was felt that this was a conservative estimate, and there did not appear to be a standard formula for the variance.

At a later date, the EIA-821, which uses the same frame, but is an annual survey and does not cover motor gasoline was designed using a similar design. Through the various cycles of each survey, a number of changes to the basic design have taken place. These have been the product of both practical necessity and continued research into the properties of the linked design. The aspects of the design related to sample rotation will be left for last, but the next section will discuss various modifications to the original design.

Modifications Of The Original Design

The earlier modifications came as evidence of bad data from the EIA-764 (the original frame) suggested several changes. In addition, concern was raised about the accuracy of weights obtained through simulation, particularly for CSUs with large weights. Finally, some needs of the EIA-821 did not match those of the EIA-782, resulting in design modifications to the EIA-821 which in many cases were later extended to the EIA-782.

The first change had to do with the stratifications. The EIA-821 required greater precision of volumes and every state was a distillate publication state. It soon became evident that crossing retail and resale would lead to too many de facto certainty cells (i.e. cells with one or two CSUs) and too large a sample.

For this reason the five relevant variables (annual volumes for residential, nonresidential retail and resale distillate fuel and for retail and resale residual fuel oil) were allowed to each form separate stratifications and a five-way linkage was conducted. This procedure was recently generalized to a seven-way linkage for the EIA-782 (though strata have been defined using more than one product for reasons which are beyond the scope of this paper).

H-T type estimators were thought unsatisfactory by some members of the research team. Two other estimators were considered:

- An estimator which merely looked at the total number sampled and the total population from each stratum (disregarding the probability of selection).
- An estimator which adjusted the sum of the weights for each product so that the number of companies in the population in a cell corresponded exactly to the sum of the weights in that cell.

A number of samples were drawn and population totals estimated from the samples and the frame using the three estimators. The H-T type estimator had a higher variance, but the second estimator, based on a sampling fraction had a bias which also made it less effective. The third estimator, which adjusted the probality of selection, constituted an improvement over the H-T and was recommended for the EIA-821. It was later adopted for the EIA-782. (In this and other instances the presence or absence of resources for system changes and the flexibility of software determined how soon a change could be implemented in one survey or the other.)

As a means of stabilizing the weights, the probability of selection of a CSU was transformed by the formula $p'=(1000p+1)/1001$ where p' is the new probability and p the old. This reduces much of the instability of the weights.

The EIA-821 also required the selection of companies rather than CSUs. Because selection of CSUs is independent for the EIA-821 this was easily done. For the EIA-782, the presence of nonpublication states made this slightly more difficult. The difficulty was in the smoothing of weights for two and three state companies. The change in sampling unit was implemented recently for the EIA-782.

The variance formula was another element which was explored further. It was clear than under most circumstances the use of only the stratified sample, disregarding the real weights or CSUs obtained from other stratifications, led to a conservative estimate of the variance. The problem was that at times it was too conservative and in certain cases it was not conservative enough. There was no formula for the joint probability of two companies, and no practical way of empirically obtaining an estimate. Eventually, a formula frequently used for PPS samples was suggested and empirically tested. It proved a better variance estimator than the previous one and was implemented.

Two unresolved issues which will not be discussed at length here are those of imputation and surprise states (a company turns out to sell in a new state). While the design creates some special issues in these areas, they also bring up problems found in other designs and which are thus unrelated to stratification linkage.

The one issue which creates a special problem is rotation of the sample. This applies to the EIA-782 which being a monthly survey presents a greater respondent burden. The EIA-821 is at this time drawn independently each cycle.

Rotation Of The Sample

One of the main difficulties of the stratified linkage design is the complexity of rotating the sample while controlling for the overlap in old and new samples. In a stratified random sample, one simply excludes half of each stratum and samples enough cases to replace them. Under linked stratification a noncertainty company can belong to as many as 21 strata (if it does business in three states).

The original approach for rotation was to rotate the random order in which the sample is selected. Thus if the sample is selected using a variable x uniformly distributed between 0 and

1, the new sample is selected using x'= (x-p) if x-p is greater than 0 and (1+x-p) otherwise.

One problem with the above approach was that if two CSUs are close to each other, the first would always be selected in preference to the second, except for the small proportion of the time when the starting point lies between the two. Given that the initial order was random, the sample will always be random, but certain companies may continue to be selected.

There are ways around this, of course. Making the degree of rotation proportional to the probability of selection tends to spread out the rotation through strata, but does not solve the problem. The difficulty was not in rotating a certain percentage, but in rotating about 50% of those in the current cycle and close to 100% of those in the previous cycle.

One approach which was at one time recommended was the creation of categories, each with 1/6 of the population. The categories would be rotated, but the order within categories would be an entirely new random variable.

Unfortunately, this approach did not take into account the fact that even without any rotation the frame itself (and thus the allocations and even stratifications) changes enough from year to year to reduce the overlap. The approach recommended above would yield far less than 50% overlap. Using more categories would tend to leave some cases in the sample repeatedly.

The rotation problem has not been satisfactorily resolved for the EIA-782 so that two conditions are met:

- Approximately 50% of the noncertainty companies sampled in one cycle remain in the sample for the next cycle.
- Close to 100% of the noncertainty companies which have been in the sample for two consecutive cycles are rotated out.

The Weight Problem Revisited

The rotation problem is one of two major ones in the EIA-782. The absence of an analytic formula for the probability of selection is the other. One of the purposes of this paper is to invite readers to try their hand at a formula. Empirical simulation is adequate to a point, but the presence of an analytic formula would enhance knowledge of the properties of a design which, after all, does reduce the necessary sample size considerably in cases like those of the two petroleum surveys.

As a simpler case, suppose it was known that the distribution of two dichotomous variables in a population was as follows:

	A	B
X	100	100
Y	100	100

Suppose one wished to sample 10 cases with value A, 10 with value B, 10 with value X and 10 with value Y. The sample is randomly ordered and cases are selected until all four conditions are met (i.e. at least 10 cases with value A on the first variable are sampled, 10 with value B, 10 with value X on the second and 10 with value Y on the second). In the process more than 10 may be sampled for one condition or another. Using this sampling approach, what is the probability of selection of any given case? Now replace the numbers with:

	A	B	Quota
X	c	d	m
Y	e	f	n
Quota	s	t	

where, of course, c+d is greater than m and so forth. Now find the probability of selection.

Finally, extend the problem to k dimensions and g(j) strata (where g depends on which dimension) and find a general formula for the probability of selection. This question stands on the way of making linked stratification far more useful as a design concept where estimators for many uncorrelated variables are involved.

Credits

The following persons contributed to some of the work reported here. Many of the concepts presented here emerged from discussions while others were developed as the effort of one individual or a team led by on individual. As happens with these kinds of projects, there are likely to have been others who were heavily involved, but were perhaps not visible to the author at the time:

Paula Weir
Mike Griffey
William Blackmore
Robert Burton
Larry Thibodeau
Robert Clickner
Glenn Galfond
and the author: Pedro J. Saavedra

The material presented in this paper is documented in a long series of reports prepared by various contractors for the Petroleum Marketing Division of the Energy Information Administration. Some of the material has also been reported in the Petroleum Marketing Monthly.

The author wishes to thank Paula Weir for reviewing this paper prior to submission.

BIBLIOGRAPHY

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

Horvitz, D.G. and Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. Jour. Amer. Stat. Assoc., 47, 663-685.

Neyman, J. (1934) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. Jour. Roy. Stat. Soc.,97 588-606.