SUPERIMPOSITION OF A GEOGRAPHICAL STRATIFICATION ON A COMPLEX DESIGN

Paula Weir, Energy Information Administration
Pedro Saavedra, Macro Sytems Inc.

The purpose of this paper is to present some of the work done determining the effect of the geographic location of the survey respondent on sample rotation. Such effects could lead to discontinuity in a published data series such as, the EIA-782B, "Reseller/Retailers' Monthly Petroleum Product Sales Report". Furthermore, this paper will address incorporation of geography in the design of the seventh sample cycle of that data series. The work presented here is preliminary, and subsequent work has pursued this same issue to continue with the geographic modification.

This paper includes the following sections:

1) Background on the EIA-782B and its frame.
2) The problem.
3) Preparation of data.
4) Principal components analysis.
5) The three key states.
6) The urban-rural factor.
7) Implicit stratification.

Background on the EIA-782B and its Frame.

The EIA-782B is a price and volume monthly survey that covers the 50 states and the District of Columbia, and includes sales volumes and prices of distillate, residual fuel oil and motor gasoline to end users and resellers. The EIA-782B is filed by a sample of resellers and retailers of No. 2 distillate and residual fuel oil dealers and resellers of motor gasoline. The noncertainty portion of the sample is rotated approximately 50% each year to reduce individual company burden. The EIA-782B is complemented by the EIA-782A which is filed by a census of refiners.

Detailed sector, volume weighted price data from the EIA-782A and the EIA-782B are published for residual oil and motor gasoline for all 50 states (plus D.C.), but distillate prices are published for only 24 states, as well as the Petroleum Allocation Defense (P.A.D.) Districts.

The frame for the EIA-782B is the EIA-863, "Petroleum Products Sales Identification Survey". The EIA-863 includes 1988 annual state level 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

Price data, however, are not collected by the frame survey.

The EIA-782B sample design (referred to as a linked sample design) is described below:

A company is classified as a certainty company if it meets one of the following criteria:

1) The company is a refiner.

2) The company does business in at least five states. Some four-state companies are classified as certainties as well (based on volumes and products).

3) The company sells at least 5 percent of the volume for one of the seven products for which volumetric information exists for one of the states in which it does business.

4) The company is assigned an ad hoc certainty status because of particular small sector (end-use category) dominance before the allocation process begins.

. From this point until the final steps, company-state units (CSUs) become the unit of analysis.

. For each state and product, the frame nonrespondents, zero volume respondents and certainties are removed and a Dalenius-Hodges procedure on product volumes is used to define stratum boundaries for each product for each state. Given the fact that products are sometimes crossed in the stratification, it is not clear that the equal variance criterion recommended by Dalenius and Hodges is optimal. For this reason, different sets of boundaries are developed for each product, yielding one, two or three nonzero noncertainty strata.

. Different stratifications are developed for each product, using various combinations of boundaries. For distillate in certain states, a crossed stratification using all three distillate products has been used to compensate for known frame errors. Each stratification is carried out with different boundaries.

. Totals and standard deviations are calculated for each product and stratum. In order to account for the difference in volumes since the time of the frame, as well as the expected monthly fluctuation of the EIA-782 volumes compared to the annual frame volume, inflation factors by stratum are used. The standard deviations are multiplied by the inflation factors.

. Neyman allocations are used for each stratification and product, selecting 100% of the certainty stratum and assigning to the nonrespondent stratum half the sampling fraction used for the combined noncertainty respondent strata. The allocations are designed to obtain a target Coefficient of Variation (CV) of 15 percent for all the products.

. For residual fuel and motor gasoline, the allocation takes place independently for each state. For distillate, the allocation is carried out first for the 24 publication states and then for the rest of the P.A.D. District so as to bring the total region to the desired CV.

. For each stratification, the total number allocated for each state is calculated and the stratification yielding the smallest sample size is selected. In case of a tie, preference is given to stratifications with fewer strata. For distillate this is done first for publication states, then after fixing the boundaries for the publication states, the process is repeated for the rest of the P.A.D. District.

. A single random variable is used to draw seven samples (one for each of the seven stratifications) simultaneously. A CSU is selected for the combined sample if it is selected for one of the seven basic samples. A company is selected if one of its CSUs is selected.

. In order to control the amount of sample rotation, the random order used to draw the sample is a rotation of the previous cycle's random order. Originally a uniformly distributed variable was assigned. In subsequent cycles, a rotation by a fixed amount plus a fraction of the largest probability of selection is used. (Most recently the procedure for assuring even geographical distribution within states was implemented.)

. One thousand samples are drawn using one thousand random variables. For each combination of three strata, the number of times its members appear is averaged. This simulation (i.e. drawing 1,000 samples) is necessary since there is no theoretical formula providing probabilities of selection for linked sampling. The multiplicative inverses of each probability become the weights.

The Problem

Discontinuity has always been a potential difficulty in any ongoing survey. This is particularly true when the survey is a monthly survey and a new cycle is inaugurated on a yearly basis. Consider a hypothetical situation where an average value is sought in a survey where fifty percent of the population was in the sample. If d was the difference between the population mean and the sample, the difference between the two samples would

be 2d. In other words, if the first sample provided an overestimate, the second sample would provide an underestimate by the same amount.

It is for this reason that an overlap of adjacent samples is necessary, so that there will not be a major discontinuity between the two months that correspond to when the new cycle is instituted. In spite of such an overlap, it is possible that there be a difference through random chance on any given value being estimated. This is particularly true in a situation like that of the EIA-782B where there can be products in certain states which are sold by only a few dealers.

One possible reason for such differences is the geographical spread of the sample within a state. If prices are related to geography, and there are high intercorrelations, then a number of prices can be off by chance in any particular state. While there is no reason why the prices should be higher or lower for a particular cycle in general, it is quite possible that this would happen by chance in some cycle, in some state. In so far as geography is an important factor in the discontinuity, any procedure which insures a geographical spread of the sample across the state is likely to result in greater precision, and to lessen sample to sample discontinuity.

Thus, this investigation attempted to discern the pattern of interrelationship of the various prices, the relationships of those prices to geographical variables and to the cycles, and the more detailed patterns for three problem states: Connecticut, New York and Maine.

Preparation of the Data

The investigation began by identifying the sample data files which applied to an overlap period between the two cycles (cycle 5/cycle 6). A file was created with companies from either cycle for the overlap months of July and August 1988. In addition the month of September for Cycle 6 was also used. The price was obtained for the month of August when available. If no sales were conducted in August, then July was used; and, if not then, September was used. No volumes were used at this stage, since the focus was on price patterns.

It was desirable to conduct some analysis on a data set with no missing values. Thus a procedure was used to impute prices. This permitted the use of a full correlation matrix for analytic purposes. In some cases it was deemed appropriate to exclude imputed values, and this is specified at the appropriate places.

The Principal Components Analyses

In order to determine the relationships among the various product/end-use prices, a principal components analysis was conducted using the prices for the original EIA-782B variables with price imputations for the missing values (i.e. where the product is not sold). Even though five factors were identified with eigenvalues greater than one, it was found that a two factor solution with an oblique rotation was very interpretable. Given that a larger number of factors was impractical for the intended purposes, the analysis concentrated on the first two components and their oblique rotation, a solution supported by a somewhat ambiguous scree test.

Essentially, the first principal component was the expected general factor, with each product having a high positive loading on the component. In other words, nationally, companies with high prices in one product tended to have high prices in other products. When the two rotated components were examined they were found to correspond to motor gasoline and residual fuel oil respectively, with distillate fuel being correlated with both.

An examination of the correlations of the factors by state revealed that although there was a positive correlation at the national level, the correlation within states for in-state companies was often negative, indicating a difference between companies selling in their home state versus out of state. In order to examine this phenomenon, the prices for in-state

companies only were standardized by state, and a new principal components analysis was carried out. The results of this new analysis were quite different.

First, the criterion of eigenvalues greater than one yielded eight factors rather than five. In addition, residential fuel oil, as well as the various residual fuel oil end-use products, did not load above .10 on the first principal component. When the oblique two factor solution was obtained, Factor 2 included various types of motor gasoline and diesel sales through company-operated retail outlets, while Factor 1 included other gasoline and diesel fuel sales, plus a moderate loading for nonresidential No. 2 fuel oil.

This indicated that the between state patterns do not necessarily correspond to the within state patterns. It also indicated that the key product -- residential fuel oil -- does not share the same geographical patterns as other products, and thus, solutions that apply to some products may not apply to residential No. 2 fuel oil.

Nevertheless, the first Principal Components Analysis was used to examine differences between the cycles for all states and between the geographical regions for the three states where problems had been detected. In addition, residential No. 2 fuel oil was treated separately, both with and without imputation. Finally a composite variable averaging standardized prices (standardizing among in-state companies at the state level) and weighing residential by three (to account for its greater importance over the other reported prices) was obtained.

T-tests were done for each variable comparing Cycle 5 only companies with Cycle 6 only companies. The first unrotated and the first rotated components were significantly different at the .05 level in the same four states: Idaho, Illinois, New Jersey and North Carolina. The second rotated component was significant at .05 in Arizona, California, North Carolina, North Dakota and Wyoming. The second unrotated component was significant at .05 in Idaho. Residential fuel oil was significantly different in New Hampshire and Illinois. None of these variables were significant at the .01 level in any state.

The composite variable correlated .76 with the first unrotated component, but indicated significant differences only in the state of Illinois. These differences, however, were significant at the .01 level, indicating lower prices in Cycle 6 than in Cycle 5.

The Three Key States

This section discusses the patterns identified for the three specific states being considered.

New York

The most obvious geographical division for the state of New York lies in the separation of New York City from the rest of the state. While there were definite differences between New York City and the rest of the state in Factor 2, there was no relationship between these differences and the two cycles, nor was there a significant interaction.

Factor 2, focusing on residual fuel and some distillate products, seemed to break down along a New York City vs. upstate division. This was also the case with the composite measure that weighs residential highly. Factor 1, however, focusing on motor gasoline showed a different pattern. For this factor, there was no easily discernible pattern within New York City, but there was a trend to lower prices for other large metropolitan areas such as Buffalo, Syracuse and Albany.

Thus it seemed that New York has three major areas in terms of prices, upstate urban, upstate rural and New York City. Within NYC there were both high and low pockets for gasoline prices.

Connecticut

The results in Connecticut were also complex, but some patterns did emerge. First of all, when it came to home heating oil, northeast Connecticut did seem to have the lowest prices (though the number of

companies reporting was small). Otherwise, the higher prices were in the west and north central regions of the state, in a corridor that included zip codes 069, 068, 067 and 060, with 066 (Bridgeport) being marginally in, and the southeast corner of the state joining in.

When we looked at the two factor solution, though, the picture was far more complex. Hartford, New Haven and some of their surrounding areas seemed to be uniformly low in most products, but home heating oil seemed an exception.

Maine

In Maine there was not a discernible geographical pattern except for lower prices in general in Portland and Bangor, as well as in certain surrounding areas.

The Urban-Rural Factor

After examining all three states it seemed clear that the major differences were not between geographical regions, but, between the big urban centers and the rest of the state. To examine this, the three digit zip codes corresponding to the big urban centers were defined as urban and the remaining zip codes were defined as rural. A state by urban status ANOVA was conducted for the rotated and unrotated factors and for the composite variable. The first unrotated factor, the two rotated factors and the composite variable all yielded significant urban-rural differences, prices being in general lower in urban areas.

A subsequent study refined the geographical stratification, and obtained urban-rural differences based on seven composite products. That study is presented in a separate paper.

Implicit Stratification

As a result of the analysis, the first inclination as a way to insure greater spread of prices was to stratify future samples based on average prices for the three digit zip code. However, there are many three digit zip codes which were not represented at all in the last two sample cycles, and some zip codes where only one or two companies existed for which price information was available.

The finding that differences seem to be accounted for in a good measure by the tendency of prices to be lower in big urban areas suggests an alternative. This alternative was intended as a multi-year solution, with the first design modification being implemented for Cycle 7 and subsequent modifications in Cycle 8. This paper presents the Cycle 7 implementation.

The proposed solution first involved identifying three categories of companies. The first would be out-of-state companies. Those were easily identifiable from the combination of home state and state of sale. The second and third type of category are heavily urban and rural respectively. These categories were harder to identify, but this was done using zip codes, plus an appropriate file linking zip codes with counties and counties with SMSAs.

The process was modified slightly in some states to insure that every state had both urban and rural components. The exception was the District of Columbia, which was merely divided into urban and out-of-town. In addition the large urban centers of New York City, Los Angeles and Chicago formed a fourth category of their own.

What was implemented in Cycle 7 was the use of the same three of four categories to spread companies out through the random order used to draw the sample. Here is how the method worked:

1) The companies were classified into the three or four categories by state.

2) Companies in each category were sorted in the same random order in which they were selected for Cycle 6.

3) Each company was first assigned a number from 1 to k, where k was the number of companies in the category and state, such that their numbers followed the order of selection for Cycle 6.

4) Then the companies (indexed by j = 1 to k) were assigned the initial random number ((j-1)/k)+(e/k) where e was a random number between 0 and 1. This preserved the relative rotation order.

5) The number was then rotated as needed.

This approach of superimposition preserved the relative rotation order within each category. In fact, if this were done without frame changes between cycles, the overlap prior to any rotation would be very high.

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

Saavedra, Pedro J. (1988) Linking Multiple Stratifications: Two Petroleum Surveys. Proceedings of the American Statistical Association, Survey Section, 777-781.