# THE EFFECTS OF SAMPLE SIZE ON VARIANCES OF THE PRODUCER PRICE INDEX

Judith Hellerstein, Bureau of Labor Statistics
600 E Street, N.W., Washington, D.C. 20212

## Introduction

This paper reports on a simulation study that was conducted at the Bureau of Labor Statistics (BLS) to examine the effects of sample size on the variance levels of price change in the Producer Price Index (PPI). These effects must be considered in order to more appropriately determine PPI sample allocations.

## Background

The PPI is a modified Laspeyres index which estimates a fixed input output price index model. Since the 1978 revision of the PPI, the industry output indexes have been based on the Standard Industrial Classification (SIC). Samples are drawn separately for four digit SIC industries, where each industry is assigned a publication structure prior to sampling. The publication structure partitions the SIC into detailed cells which generally make up the different homogeneous (with respect to price-determining characteristics) product categories within that SIC. These detailed cells are combined into higher level cells which eventually aggregate to the four digit SIC.

The sample design is a two-stage systematic sample with item probabilities assigned proportionally to measures of size. In the first stage, the primary sample unit (PSU) is a profit maximizing center. This profit maximizing center is one or more establishments within which prices are set and records kept. The first stage sample is drawn in the Washington Office using a frame derived from the Unemployment Insurance file. Employment, as a proxy for company revenue, is used as the measure of size. The second stage of the sample is performed at each selected company using a sampling procedure known as disaggregation. Unique products are selected and information concerning each PSU is obtained. Subsequently, monthly price data on the selected unique products are obtained and used to measure price change in the PPI. These are published as the monthly price indexes.

The current sample design process permits the computation of variances using balanced half-sample replication as described in Collia (1988). The sampling frame is divided into n variance strata, where n is dependent on the sample size and the number of self-representing PSUs and where (n+1) is a multiple of 4. The first-stage sample is then drawn by taking two independent half-samples within each variance stratum. Each of these strata must contain at least one certainty unit or two probability units. During the second-stage sampling process, items from self-representing PSU's and PSU's selected in both first-stage half samples are divided between half-samples. Using a Hadamard matrix, (n+1) orthogonal replicates are formed from unique combinations of price quotes. The remainder of the quotes are formed into corresponding complements. Indexes are then calculated for each replicate and its complement and the variance is computed as a function of these indexes. McCarthy (1969) provides a clear discussion of the use of the complement in this type of variance estimation.

The first stage sample allocation for a four digit SIC is presently determined by examining such industry characteristics as industry concentration and product diversity and by considering the importance to PPI users of specific lower level indexes. The second stage allocation is usually determined using a formula based on the expected diversity of production within each first stage sample unit.

The final two criteria for determining sample allocations are user need and budgetary constraints. Both first and second stage allocations may be supplemented in order to ensure publication of cell indexes which are of particular interest to PPI users and whose monthly publication might otherwise be in jeopardy. However, these enhancements must be balanced against cost considerations. The PPI operates on a fixed overall budget; when sample allocation enhancements are made in one industry, compromises in sample size may be necessary in others. This tradeoff between quality and quantity gives rise to a crucial question: What are the effects of sample size on variance levels and index quality in the PPI?

## The Study

Analytical variance formulas based on sample size are difficult to derive for PPI data due to the nonlinearity of the indexes. In order to examine the effects of sample size on PPI variances, a simulation study was conducted on price data from a lowest level cell in six different four-digit SIC industries. Using simple random subsampling, subsamples of varying sizes were drawn from the full samples of price quotes in each of the six cells. Since these subsamples were drawn from the probability proportional to size (PPS) samples of the PPI, the subsamples were also PPS samples from the total universe of quotes in these industries. Index and variance levels of these subsamples were then compared to the indexes and variances for the original full samples from each cell.

## Selection and Preparation of Data

The six cells that were selected for the study contained large numbers of price quotes which facilitated subsampling at various sample sizes. Lowest level cells were specifically chosen in order to eliminate complications arising from the weighted aggregation structure of the PPI. Thirteen consecutive months (January, 1987 - January, 1988) of collected price data for these cells were extracted from the PPI estimation system and the data were modified to exclude quotes which did not appear in all thirteen months. Because a quote is deleted from repricing only when the reporter permanently ceases repricing that item, these full samples for each of the six cells contained some estimated data. The typical estimation method is to estimate missing price data by using the average price change for the lowest level cell in which that particular item falls.

The six cells and their full sample sizes are listed below:

| Cell Code | Name | Number of Price Quotes |
|---|---|---|
| 208630111 | Cola, excluding diet, returnable bottles | 59 |
| 2095116 | Ground, roasted coffee | 73 |
| 207230301 | Tufted broadloom - nylon | 108 |
| 2411911 | Contract logging | 219 |
| 2711722 | Newspaper publishing; local, regional, and other advertising (hereafter referred to as Regional Newspaper Advertising) | 170 |
| 2752697 | Commercial printing, lithographic: all other general commercial printing, n.e.c.,sheet-fed | 109 |

### Index and Variance Estimation

Each item that is selected at initiation for monthly repricing in the PPI is assigned an item weight which is derived from the PSU and item selection processes. The following simplified formula is used to determine the item weight of each selected item (for a full description see Hill 1987):

$$\omega_{ij} = \gamma_i \rho_{ij} V_{ij}$$

where

$\omega_{ij}$ = item weight of item j in PSU i
$\gamma_i$ = (probability of selection of PSU i)$^{-1}$
$\rho_{ij}$ = (probability of selection of item j within PSU i)$^{-1}$
$V_{ij}$ = total value of shipments and receipts for item j within PSU i

Since the full samples of this study were prepared to contain a constant set of items each month, the monthly cell index of each universe was easily calculated with the following formula:

$$\hat{I}^t = \frac{\sum\sum_{ij} \omega_{ij} (P^t_{ij}/P^b_{ij})}{\sum\sum_{ij} \omega_{ij}} \times 100$$

where

$\hat{I}^t$ = index value in month t
$\omega_{ij}$ = item weight of item j in PSU i
$P^t_{ij}$ = price of item j from PSU i in the current month t
$P^b_{ij}$ = price of item j from PSU i in the base period b

The price index of each cell in the PPI is set to 100 in the base period. For the purposes of this study, the base period was considered to have been December 1986.

In order to calculate the variance for the full sample of a given cell, separate index levels for each replicate and its associated complement were computed using the same methodology as for the overall sample. Utilizing balanced half-sample replication, the variance estimate of the cell was then computed as a function of the indexes of the replicates and complements:

$$V(\hat{I}^t) = \frac{\sum_{\alpha=1}^{k} (\hat{I}_\alpha - \hat{I}^c_\alpha)^2 \Phi_\alpha}{4 \times \sum_{\alpha=1}^{k} \Phi_\alpha}$$

where

$V(\hat{I}^t)$ = variance of index $\hat{I}$ in month t
$k$ = total number of replicate/ complement pairs
$\hat{I}_\alpha$ = index calculated from data in the $\alpha$th replicate
$\hat{I}^c_\alpha$ = index calculated from data in the $\alpha$th complement
$\Phi_\alpha$ = 1 if $\alpha$th replicate and $\alpha$th complement are both non-empty
= 0 if either $\alpha$th replicate or $\alpha$th complement is empty

( The variance is not computed if $\sum_{\alpha=1}^{k} \Phi_\alpha = 0$)

These six full samples of price quotes, though large in absolute number, were clearly considerably smaller than the actual universe frames from which the original PPI samples were drawn. Because variance estimation was also to be done for subsamples of small sizes from these cells, the probability of encountering empty replicate/complement pairs in subsamples was reduced by reducing the number of variance strata n to 7 in all of the cells.

Variances, as functions of the indexes from which they are derived, are only directly comparable when the index values themselves are equal. Since the price indexes in this study were all set to 100 in December, 1986 and indexes were computed for only thirteen subsequent months, these index values would not be expected to vary as widely as they might in the published PPI. However, in order to eliminate any possible effects of differing indexes, coefficients of variation were computed. This facilitated direct comparisons of price variation across industries and months. The coefficient of variation for a given cell is:

$$CV(\hat{I}^t) = \frac{(V(\hat{I}^t))^{1/2}}{\hat{I}(t)}$$

where

$\hat{I}^t$ = index value in month t

$CV(\hat{I}^t)$ = coefficient of variation of index value $\hat{I}$ in month t

$V(\hat{I}^t)$ = variance of index value $\hat{I}$ in month t

**Drawing Subsamples**

In order to examine the effects of sample size on these cells, subsamples of varying sizes were selected from the full sample of each cell. These subsamples, like the full samples from which they were taken, had to be drawn with probabilities proportional to size with respect to the entire PPI industry universes. Since PPS sampling had already been utilized in the selection of the full samples, simple unweighted subsampling without replacement was conducted in order to obtain proper PPS subsamples of the six cells.

Sets of one thousand subsamples of varying sizes were drawn from each cell. As the size of the subsamples approached the size of the full sample for that cell, the variances of the subsamples differed only slightly from the variance of the full sample and therefore the results were not particularly interesting. However, since the size of the full samples varied greatly, the point at which the size of the subsamples became too big to obtain interesting results was different for each cell. Therefore, in this study, five sets of one thousand subsamples from each of the six cells will be examined, with the sets representing 5%, 10%, 15%, 20%, and 25% of the absolute number of quotes in that cell.

**Indexes, Variances, and Coefficients of Variation for Subsamples**

Index and variance levels for each subsample were calculated over the thirteen month period using the same methodology as for the full samples. Average index and variance levels were computed from the values obtained for the one thousand subsamples of each sample size. These average values were then used to compute coefficients of variation. The average indexes and coefficients of variation were then compared to the values calculated previously for the full sample. The results can be seen in Figures 1 and 2.

Figure 1 illustrates that in each industry the average index values for each sample size closely matched the population index values for that cell. It is evident that the subsamples and the full samples from which they were drawn were indeed estimating the same index value.

In Figure 2, coefficients of variation, as calculated from average index and variance levels for the subsamples of each sample size, are depicted in relation to the full sample coefficients of variation. The results in each industry indicate that, in all cases, reductions in sample size did lead to significant increases in variance. However, there was no constant proportional relationship between sample sizes and variance levels. Additionally, as the graphs illustrate, the magnitude of the effects of sample size reductions are quite industry specific. The reasons for this can be tied to the underlying universe characteristics of each cell.

Contract Logging, SIC 2411911, provides a clear illustration of the effects of these characteristics. Contract Logging, though the largest cell in the study, was the only one which had wide fluctuations in both index values and coefficients of variation. The underlying reasons for these fluctuations are due to economic characteristics which are extremely specific to this industry. Contract Logging is a highly seasonal industry, with production heaviest in the spring and fall months. Production and prices are also determined by geographical characteristics which affect weather and the type of wood available. When a logging contract is made, the amount and type of wood to be felled is determined and the fixed price for the contract is established. This price always remains constant through the duration of that contract, no matter what market or environmental forces prevail. Therefore, the index and variance levels in this SIC fluctuate as new contracts are negotiated with prices quite different than those of previous and existing contracts. The sample size for this SIC must be quite large in order to reduce the effects of isolated large price changes on PPI variance levels.

In contrast, SIC 2711722, Regional Newspaper Advertising, experienced little variance, even for small subsamples. Again, the reasons for this are quite industry specific. This cell represents the advertisements that newspapers sell to advertisers for distribution in non-national markets. The major cause of price changes in newspaper advertising is price changes in the input cost of newsprint. All companies that produce newsprint change their prices at the same rate, and these price changes are often made public up to three months prior to the date on which they take effect. Since all newspaper publishers rely on this newsprint, they are equally affected by newsprint price changes, and generally all pass along these cost increases to their advertising customers. Therefore, the variance in SIC 2711722 is quite low, and the price index for Regional Newspaper Advertising closely follows the index for Newsprint. Additionally, newspapers have begun to fear that they are losing advertising revenue to other media sources and are trying to continue to attract advertising customers by standardizing advertising rate schedules across the industry (Dunaway 1988). This ongoing implementation of standardized rates has also contributed to the low variance levels in SIC 2711722. It is apparent from the coefficients of variation depicted in Figure 2 that the quality of the index for this cell would not be hurt significantly if the sample size were somewhat reduced.

As previously described, missing price data was estimated by the average price data for the entire cell rather than the half-sample to which it was assigned. The computed variances were therefore lower than they would have been had separate imputations for missing data been made in each replicate and complement.

## Conclusions

The results of this study illustrate the importance of sample size to PPI data. The number of price quotes used in the estimation of price change in an industry can have dramatic effects on variances and the quality of published indexes. However, since variance levels differ so greatly across industries, it would be difficult to develop general standards for determining PPI sample allocations.

In the absence of universally similar effects of sample size on PPI data, it is necessary to develop industry specific standards with respect to adequate numbers of price quotes in particular cells. This is dependent upon the formulation of regression models which predict industry specific variance levels. Some research has been conducted on a generalized variance function in the PPI and, as expected, it seems that a single function cannot feasibly be developed for the entire PPI. Initial regression modeling specific to particular SICs has been attempted. These models were developed using variables such as sample size, industry value of shipments, number of companies in the industry, and complexity of PPI aggregation structures. However, the variables used in these regressions still do not account for significant aspects of PPI variances.

## Future Research

Future research will focus on isolating the causes of PPI variances. One of the main tasks of this research will be to determine whether the variance within a particular cell is a product of underlying industry variance or whether it is more a function of errors due to sampling. This will require close examination of actual price data and significant input from the PPI economists assigned to monitoring specific industries.

Work on regression modeling will continue. This research will focus on ways of incorporating into these models variables which are more industry specific than the ones listed above. For instance, for Contract Logging, SIC 2411911, any model of variances will have to take into account the way that contracts are set, the inherent seasonality in the industry, and the geographic distribution of logging contracts. Many of these characteristics, of course, are not applicable to regression modeling in other industries.

Once industry specific variance regression models are developed, they can be utilized in determinations of PPI sample allocations. Various sample allocations can be considered with respect to their effect on variance levels over time in specific industries. Sample allocations which fall within the budgetary constraints of the PPI can then be more appropriately distributed among the various SICs which are due to be resampled. For example, the results described above indicate that a redistribution of some sample units from Regional Newspaper Advertising to Contract Logging would be beneficial.

Similarly, these models can be utilized during the life of a given price index to monitor the effects of the deterioration of the sample on variance levels. In months when high variance levels in a particular cell are attributable to low number of price quotes (rather than to inherent industry characteristics), publication for that cell can be suppressed. Ultimately, this information can be utilized together with nonresponse data to predict the gradual deterioration of a sample so that timely resampling of that industry can occur.

## REFERENCES

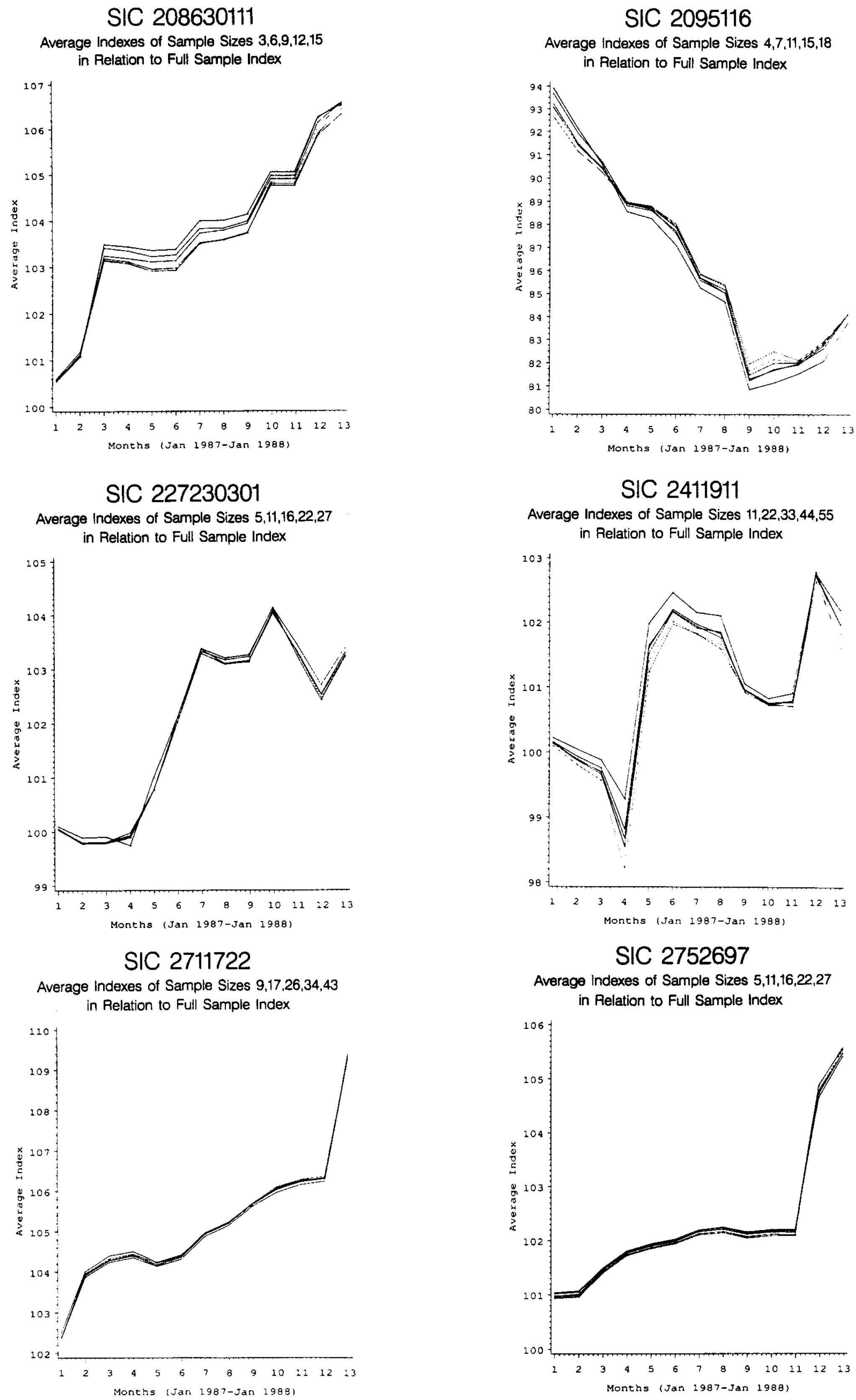
Collia, D (1988), *"Measuring Sample Variability in the Producer Price Index,"* Proceedings of Survey Research Methodology Section, American Statistical Association.

Dunaway, J. (1988), *"Nab Winds Up 'Future Advertising Project' --Continues Activities In Four Key Areas,"* Newspaper Advertising Bureau, Inc, Press Release.

Gousen, S. and Monk, K. (1986), Producer Price Measurement. Concepts and Methods, U.S. Department of Labor, Bureau of Labor Statistics.

Hill, K (1987), *"Survey Design in the Producer Price Index,"* Proceedings of Survey Research Methodology Section, American Statistical Association.

McCarthy, P. (1969), *"Pseudoreplication, Further Evaluation and Application of the Balanced Half-Sample Techinique,"* National Center for Health Statistics, Series 2, No. 31, Washington, D.C.: U.S. Government Printing Office.



SIC 208630111
Average Indexes of Sample Sizes 3,6,9,12,15
in Relation to Full Sample Index

SIC 2095116
Average Indexes of Sample Sizes 4,7,11,15,18
in Relation to Full Sample Index

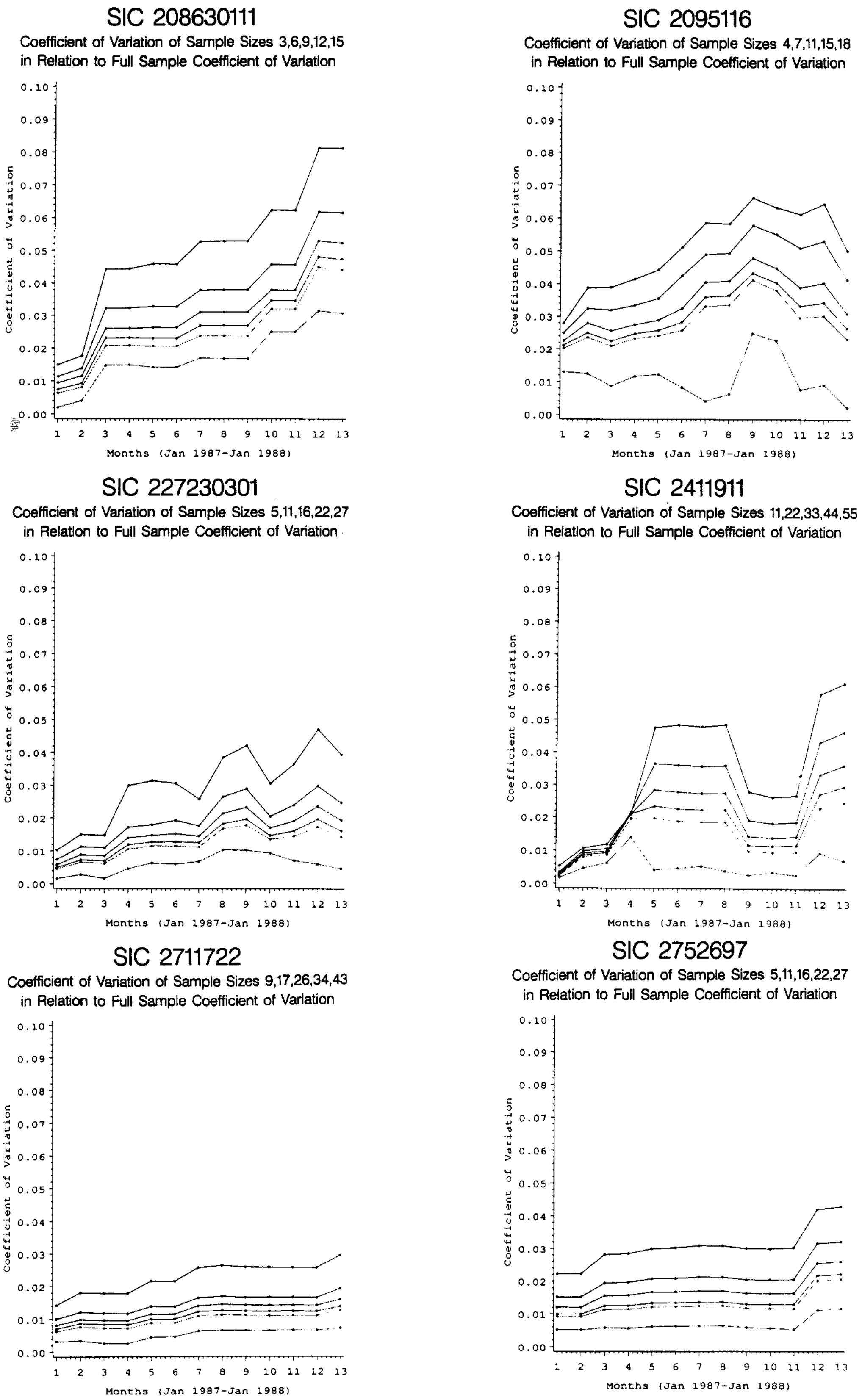
SIC 227230301
Average Indexes of Sample Sizes 5,11,16,22,27
in Relation to Full Sample Index

SIC 2411911
Average Indexes of Sample Sizes 11,22,33,44,55
in Relation to Full Sample Index

SIC 2711722
Average Indexes of Sample Sizes 9,17,26,34,43
in Relation to Full Sample Index

SIC 2752697
Average Indexes of Sample Sizes 5,11,16,22,27
in Relation to Full Sample Index

Figure 1. Average indexes and full sample indexes plotted over time for each of the six cells in the study.



SIC 208630111
Coefficient of Variation of Sample Sizes 3,6,9,12,15
in Relation to Full Sample Coefficient of Variation

SIC 2095116
Coefficient of Variation of Sample Sizes 4,7,11,15,18
in Relation to Full Sample Coefficient of Variation

SIC 227230301
Coefficient of Variation of Sample Sizes 5,11,16,22,27
in Relation to Full Sample Coefficient of Variation

SIC 2411911
Coefficient of Variation of Sample Sizes 11,22,33,44,55
in Relation to Full Sample Coefficient of Variation

SIC 2711722
Coefficient of Variation of Sample Sizes 9,17,26,34,43
in Relation to Full Sample Coefficient of Variation

SIC 2752697
Coefficient of Variation of Sample Sizes 5,11,16,22,27
in Relation to Full Sample Coefficient of Variation

**Figure 2. The top line on each graph depicts the coefficient of variation for the 5% samples, with successive lines representing the coefficients of variation for successively larger sample sizes. The bottom line on each graph depicts the full sample coefficient of variation for that industry.**