MEASURING SAMPLE VARIABILITY IN THE PRODUCER PRICE INDEX

Demetra V. Collia, Bureau of Labor Statistics
600 E Street NW, Room 5217, Washington, D.C. 20212

INTRODUCTION

This paper reports on recent efforts by the Bureau of Labor Statistics (BLS) to estimate the variance of the Producer Price Index (PPI). We evaluated the applicability of series expansion and sample modification techniques to the PPI, and selected balanced half-sample replication as the most promising technique.

The first section describes the survey design of the PPI and the estimation process. The second section discusses four variance estimation techniques in relationship to the PPI data. The third section describes the PPI's variance estimation system and presents empirical results of its application. In the final section we briefly outline additional work to be done in the area of variance estimation.

SAMPLE DESIGN AND ESTIMATION

The Producer Price Index is modeled as a fixed-weight, fixed-base index as developed by Etienne Laspeyres in 1864. However, the PPI uses weights that have been updated since the original base period and base period prices are adjusted for quality changes and product substitutions. As a result, we have designated the PPI a "modified" Laspeyres index.

The work described in this paper refers to industry output price indexes. With minimal modifications the same methodology is applicable to the other indexes published by BLS such as commodity and stage-of-processing price indexes. Industry output price indexes are organized by industry according to the Standard Industrial Classification (SIC). Each industry has a publication structure defined before the industry is sampled. The publication structure defines the "cells", or lowest level indexes to be included in the industry indexes and determines how these cells are to be combined (i.e. aggregated) to higher levels.

In index estimation, prices are combined to estimate long-term cell relatives which are used to estimate cell indexes (lowest level indexes). Cell indexes are then aggregated to estimate higher level indexes benchmarked on trade values collected by the Bureau of the Census.

The survey design is based on stratified multi-stage systematic sampling with item selection probabilities being proportional to a measure of size. Ideally we would want to sample the unique products of an industry. Since a frame of products does not exist, we sample companies and within each company we subsample products. The first stage of sampling takes place at the Washington office and uses the employment of each company, as listed on the Unemployment Insurance (U.I.) file, as the first stage measure of size. Each primary sample unit is a Profit Maximizing Center (PMC). A PMC refers to the economic unit within which prices are set, records are kept, and profits are maximized. The PMC may be either a single establishment or a group or "cluster" of establishments operating within the same SIC and the same company. The second and subsequent stages take place at each selected company during initiation, using a procedure called disaggregation. Initiation of companies involves selecting unique products, and collecting baseline information on each primary sampling unit. Once a sample is "initiated", the process of "repricing" begins. On a monthly basis, each company mails in price data on all items selected during disaggregation. It is these data that are used for calculating the monthly indexes. Hill (1987) describes in detail the sample design and estimation methodology for the PPI.

VARIANCE ESTIMATION METHODOLOGIES

Most of the theoretical results presented in this section were developed by the Research Triangle Institute (RTI) under contract to the Bureau of Labor Statistics. RTI attempted to derive a theoretically exact formula for the variance of the Producer Price Index. The methodologies they considered were series expansion, and sample modification techniques.

I. Taylor Series Expansion

There are a number of references on series expansion techniques for determining the variance expression for ratios and other complex statistics, including Wolter (1985) and Koop (1968). Shah (1977), Koop (1972), and Krewski and Rao (1981) provide additional insight regarding the use of this technique in the problem of estimating regression and correlation coefficients and other statistics which are functions of stratum means.

In this section most of the theoretical development pertains to the long-term cell relative, this being the building block upon which an industry index is constructed.

The long-term relative for cell N is a weighted sum of price relatives:

$$r_N^{t,b} = \frac{\sum_{ih} E_{is} U_{ih} r_{ih}^{t,b}}{\sum_{ih} E_{is} U_{ih}}$$

and the cell index can be written as:

$$\hat{I}_N^{t,b} = \left(\frac{r_N^{t,b}}{r_N^{t-1,b}}\right)\left(\frac{r_N^{t-1,b}}{r_N^{t-2,b}}\right)\left(\frac{r_N^{t-2,b}}{r_N^{t-3,b}}\right)\left(\frac{r_N^{t-3,b}}{r_N^{t-4,b}}\right) \hat{I}_N^{t-4,b}$$

where:

i - refers to the primary sampling unit (PSU).

h - refers to a unique item arrived at by product disaggregation.

e_i - employment of primary sampling unit i.

$U_{ih} = n_{ih} \gamma_{ih} \alpha_i V_i$ - item weight

n_{ih} - refers to a multiple hit factor

γ_{ih} - refers to the relative percent. It is the relative importance of item h in PSU i.

$\alpha_i = (\sum e_i)/ne_i$ - sampling weight for PSU i.

V_i - collected value of shipments and receipts for PSU i

n - first stage sample size

$$E_{is} = \frac{\sum\limits_{i \varepsilon T_s} \alpha_i e_i}{\sum\limits_{i \varepsilon T_s} \left[\alpha_i e_i \left(\sum\limits_{h \varepsilon S(i)} n_{ih} \gamma_{ih} \right) \right]}$$ - nonresponse factor applied to stratum s

T_s - the set of all sampling units in s which are productive or refusals

S(i) - the set of all items h, in PSU i, which are being used in index calculation.

$$r_{ih}^{t,b} = \frac{P_{ih}^t}{P_{ih}^b}$$ - long term price relative for unique product h in PSU i.

Let $Y_N^{t,b} = \sum\limits_{ih} E_{is} \; U_{ih} \; r_{ih}^{t,b}$ and $X_N^{t,b} = \sum\limits_{ih} E_{is} U_{ih}$

then the cell relative becomes: $r_N^{t,b} = \dfrac{Y_N^{t,b}}{X_N^{t,b}}$ (1)
and an approximate variance is

$$V(r_N^{t,b}) = \frac{E^2[Y]}{E^2[X]} \left[\frac{V(X)}{E^2[X]} - \frac{2 \text{ Cov } (X,Y)}{E[X] \quad E[Y]} + \frac{V(Y)}{E^2[Y]} \right] \quad (2)$$

The mathematical expectation is to be determined by taking into account all relevant stages of sampling with the appropriate selection probabilities. The entire problem reduces to finding consistent estimates of the variances for X and Y and their covariance.

Attempts at deriving approximate formulae proved to be intractable. It is extremely difficult to obtain estimates of the variances and covariances that may be considered remotely close to being unbiased. There are two reasons for this: (1) due to the presence of characteristic random variables indicating whether a unit is productive or refusal, the estimating functions constituting the numerator and denominator of the cell relative are themselves nonlinear, and (2) the sample size internal to the PSU at subsequent stages is often one and this does not permit estimation of variance even if the nonlinear functions are linearized.

To eliminate this source of difficulty an attempt was made to develop variances conditionally on the set of productive or refusal establishments.

An exact form for the variance of the cell relative was proven to be unusable for the following reasons:

(1) In this method, third and higher orders were neglected each time the expansion is carried out to obtain moment functions that are tractable. The error of approximation on the true variance depends on the relative magnitude of the neglected terms.

(2) Any inferences can only relate to the set of establishments that are in Scope, in Business, and will be responsive to initiation in the base period.

(3) Because of (2), any inferences can relate only to the set of active items in all such establishments at subsequent periods of time.

II. Sample Modification Techniques

The division of a simple random sample into a number of parts to estimate the variance on the basis of means of the resulting subsample was first discussed by Mahalanobis (1946) and later in Hansen, Hurwitz, and Madow (1953). McCarthy (1966) considers the situation when two primary sampling units per stratum are available in a survey and considers the estimation of variance on the basis of the 2^L possible half samples or pseudoreplicates. For this purpose he gives attention to certain subsets of this set of 2^L that have desirable properties of orthogonality. A large number of statisticians have studied this area since then. Jackknifing, Sample Division, and Independent Replication were three sample modification techniques considered for the PPI.

JACKKNIFING, in the true sense of the word, was ruled out because it was not applicable to PPI data. This technique involves the removal of price items from the lowest level cells in self-representing units where the supply of data is not always plentiful. The cell relatives computed after Jackknifing may not be meaningful in the context of the Producer Price Index (PPI). In the case of sparse cells, the removal of price data can result in a cell index determined by products that are not representative of the cell's overall production. This index estimate does not fulfill the economic concept of a price index and can be very misleading.

An alternative method called SAMPLE DIVISION or RANDOM GROUPS was considered. In this method first stage sample units (companies) are divided up equally at random into g subsamples and cell relatives are computed from each subsample taking into account the division of the sample.

Let α_c be a random subset of k primary units (c=1,..., g), then a cell index r_c is computed computed using data from subset α_c as:

$$r_c = \frac{\sum\limits_{i \varepsilon \alpha_c} Z_i}{\sum\limits_{i \varepsilon \alpha_c} Q_i} \quad \text{and} \quad \bar{r} = \frac{1}{g} \sum_{c=1}^{g} r_c$$

where: $Z_i = \sum\limits_{h \varepsilon i} E_{is} \; U_{ih} \; r_{ih}^{t,b}$, and

$Q_i = \sum\limits_{h \varepsilon i} E_{is} \; U_{ih}$.

If w is the undivided sample, then

$V(\bar{r}) = V(E[\bar{r} \mid w]) + E[V(\bar{r} \mid w)]$

It can be shown that $V(r) \doteq V(r_N^{t,b}) + E[V(\bar{r} \mid w)]$.

Since $E[V(\bar{r} \mid w)]$ is always positive $V(\bar{r}) > V(r_N^{t,b})$ and if we use

$$\hat{V}(\bar{r}) = \frac{\sum\limits_{c=1}^{g} (r_c - \bar{r})^2}{g(g-1)}$$ as an estimator of $V(\bar{r})$

we are likely to have a conservative estimate of $V(r^{t,b}_N)$, which can be justified as partially offsetting any neglected nonsampling error.

Like Jackknifing, Sample Division was proven to be operationally unacceptable.
For certain industries, removal of first stage units could result in subsamples with no prices for certain cells, or subsamples with no prices at all. This makes cell relative and/or variance calculation impossible.

INDEPENDENT REPLICATION

In the context of classical sample survey theory the problem of independent replicated samples and its ramifications has been examined by Mahalanobis (1946), Lahiri (1954), Koop (1960, 1967), Singh and Bansal (1975) McCarthy (1969), Krewski, D. and Rao, J.N.K. (1981) and others.

Using this method we draw g samples each consisting of k_h primary units so that $g \cdot k_h = n_h$ the number allocated for stratum h. Each sample is replaced before the drawing of the next to ensure statistical independence. Common first stage units and second stage items are possible but since the multi-stage disaggregation procedure is also independent for each of the g samples, statistical independence is preserved.

Let $r^{t,b}_N(1), r^{t,b}_N(2), \ldots, r^{t,b}_N(g)$, be the g estimates of the long-term cell relative for cell N, each computed on the basis of the corresponding independent sample, and

$$r^{t,b}_N(c) = \frac{\sum_{i \epsilon \alpha_c} Z_i}{\sum_{i \epsilon \alpha_c} Q_i} \quad \text{where } \alpha_c \text{ is the } c^{th} \text{ independent sample, } c=1,\ldots,g.$$

Then the average cell relative is given by

$$r^{t,b}_N(\cdot) = \frac{1}{g} \sum_{c=1}^{g} r^{t,b}_N(c) \text{ and } V\left(r^{t,b}_N(\cdot)\right) = \frac{1}{g} \cdot V\left(r^{t,b}_N(c)\right)$$

Regardless of the functional form of $V\left(r^{t,b}_N(c)\right)$ an unbiased estimate of $V\left(r^{t,b}_N(\cdot)\right)$ is simply:

$$\hat{V}\left(r^{t,b}_N(\cdot)\right) = \frac{\sum_{c=1}^{g} \left(r(c) - r(\cdot)\right)^2}{g(g-1)}$$

As in the method of Sample Division the number of quotes available for computing a cell index will vary from sample to sample. The above formula is applicable as long as there are at least two samples each having at least one quote.

PPI's VARIANCE ESTIMATION SYSTEM

Taylor Series Expansion, Jackknifing, Sample Division, Independent Replication were all considered in great detail for the PPI by Koop (1979a, 1979b). Taylor series proved to be an interesting statistical exercise with very little practical application. In the context of the PPI's sampling design, which has as many as ten stages of sampling, higher moments can have a tremendous number of component terms. Therefore, neglecting higher order terms was perceived as a serious problem. Jackknifing, which requires the removal of data at the lowest level cell of self-representing units, was unacceptable, because in the PPI the cells can have only 3 or 4 pricing items. This left us with sample division and replication to be considered. The results of both techniques are assumed to be very similar provided that the number of groups in sample division is large. However, the optimal number of groups for the PPI was considered to be two. In this case a variance using sample division can be easily computed but it is expected to be rather unstable having only one degree of freedom. Replication on the other hand provides an estimator that is easily computed and is more stable than the one provided by sample division because it has more degrees of freedom. Therefore we decided to use a replication technique.

Once a decision was made to adopt a replication technique for variance estimation in the PPI, a computerized system was developed incorporating theoretical assumptions and program constraints. The present program is a computerized system that calculates variances based on balanced half-sample replication.

Each industry's sampling frame is divided into m variance strata (m=3,7,15,31). The number of variance strata is determined by the total sample size and the number of self-representing primary sampling units. There is at least one certainty unit or two probability units per variance stratum and (m+1) is always a multiple of 4. Within each variance stratum two independent samples are drawn (half-sample A and half-sample B). Modifications to the sample design and the disaggregation process were made to preserve as much independence between the two half-samples as possible. Items from self-representing PSU's and PSU's selected in both half-samples are divided between half-samples. With the aid of a Hadamard matrix, (m+1) orthogonal replicates are formed, McCarthy (1969).

An example of a balanced 7 (variance strata) by 8 (replicates) matrix is:

VARIANCE STRATA

1	2	3	4	5	6	7	...REPLICATES...
A	A	A	A	A	A	A	1
B	A	B	A	B	A	B	2
A	B	B	A	A	B	B	3
B	B	A	A	B	B	A	4
A	A	A	B	B	B	B	5
B	A	B	B	A	B	A	6
A	B	B	B	B	A	A	7
B	B	A	B	A	A	B	8

Each replicate is a unique combination of price items coming from half-samples A or B within each variance stratum. All prices not included in replicate j form complement j. Thus, a sample with m variance strata generates m+1 replicates and m+1 complements. So in the example above, the configuration for the third replicate is ABBAABB and the third complement is BAABBAA.

Since the number of replicates is a multiple of four and always higher than the number of variance strata, the replicates are in full orthogonal balance, Wolter (1985). An index is calculated on every level of publication from each replicate and complement sample using exactly the same methodology as for the overall sample. The variance of the index is computed using the following estimators:

$$V(\hat{I}_N^{t,b}) = \frac{\sum_{i=1}^{s} (\hat{I}_{R_i} - \hat{I}_{C_i})^2}{4\,s} \qquad (4)$$

$$V(\hat{I}_N^{t,b}) = \frac{\sum_{i=1}^{k} (\hat{I}_{R_i} - \hat{I})^2}{k} \qquad (5)$$

where: $\hat{I}_{R_i}$ is the cell index using all prices in replicate i
$\hat{I}_{C_i}$ -is the cell index using all prices in complement i
$\hat{I}$ -is the cell index using all prices
k -is the number of non-zero replicates
s -is the number of non-zero replicate/complement pairs

It has been shown that (4) and (5) are asymptotically equivalent. However, in small samples and in the case of a nonlinear estimator, such as the PPI estimator, (5) is considered to be an estimator of the MSE of the index and(4) is considered to be an estimator of the variance, Wolter, p. 119 (1985).

In addition, an estimate of the variance's stability is calculated using a Jackknifing approach. For every replicate/complement pair an estimate of the variance is calculated using:

$$V_j = \frac{\sum_{\substack{i=1\\ j\neq i}}^{s} (\hat{I}_{R_i} - \hat{I}_{C_i})^2}{4\,(s-1)}$$

and an estimate of the variance of the variance is then:

$$S_V^2 = \frac{\sum_{i=1}^{s} (V_i - \bar{V})}{s(s-1)}$$

Theoretical assumptions were compromised in the following areas:

(1) Due to cost and computational constraints the maximum number of variance strata is limited to 31. It is our belief that any possible reduction in the estimator's stability would not have a significant impact on the variances.

(2) Due to operational considerations unique items for both half-samples are selected by the same data collector during initiation. Thus the two half-samples cannot be considered truly independent.

(3) Independence is again compromised because the methods of estimating missing prices and revenue data are applied to the entire sample and not to each half-sample.

EMPIRICAL RESULTS

Variance estimation has been carried out for 114 industries. They are a representative sample of all industries in Mining and Manufacturing with respect to sample size, complexity of the industry, price movements, special sampling procedures, etc. For each industry the variance of the index, the variance of the percent change of the index, and the variance of the variance were calculated for the period of Oct. 1982 - Apr. 1984.

Figure 1 shows overlaying plots of variances over time for the two estimators of a representative industry. Both estimators displayed similar patterns with one (5) being consistently higher than the other, validating Wolter's claim that, in small samples, (5) is an approximation of the MSE. Like the selected SIC shown on Figure 1, most of the sampled SIC's show an increasing trend in the variance. This phenomenon has been studied by Valliant (1987) and it is believed to be an artifact of the index estimator. Internal BLS research on the Consumer Price Index (CPI) survey has found a similar pattern of increasing variance over time.

Similarly, for most of the sampled SIC's the coefficient of variation shows an increasing trend as illustrated on Figure 2. It indicates that the precision of the index estimator can deteriorate over time. This result agrees with Valliant's claim that the product estimator for the Laspeyres index is expected to be unstable.

While the variance of the index clearly shows an increasing trend, that pattern is not apparent in the variance of the percent change of the index. As shown in Figure 3, the variance of the monthly percent change fluctuates considerably from month to month and can be quite high for certain industries.

The variance of the variance is modelled as a function of the index variance. A sample of the results is shown on Table 1. The best model is a polynomial one of the form:

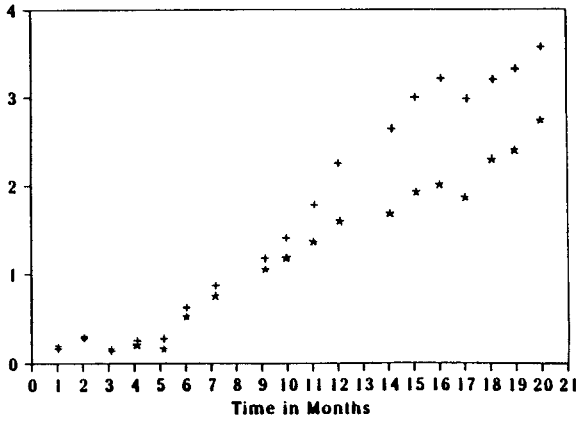
$S_V^2 = \alpha \cdot V^{\beta}$ where $1.8 < \beta < 2.1$

Regression models have been tried in an attempt to develop a generalized variance function (GVF) for the PPI. Our findings indicate that it is not feasible to develop one GVF for the entire PPI, because variances vary considerably from one industry to another. Regression models specific to each industry (SIC) predicting the variance as a function of certain covariates seem a more viable option. Preliminary regression models indicate that the monthly INDEX ESTIMATE and/or LENGTH of TIME IN VARIANCE CALCULATION emerge as the predominant covariates. These models were developed using weighted least squares with weights being the predicted values from the variance of variance models. A sample of the regression results on the variance of the index is shown on Table 2.

Figure 1

Variance Data Over Time

SIC 3533



Note: * is used to plot (4)
+ is used to plot (5)

Figure 2

Coefficient of Variation Over Time

SIC 3533

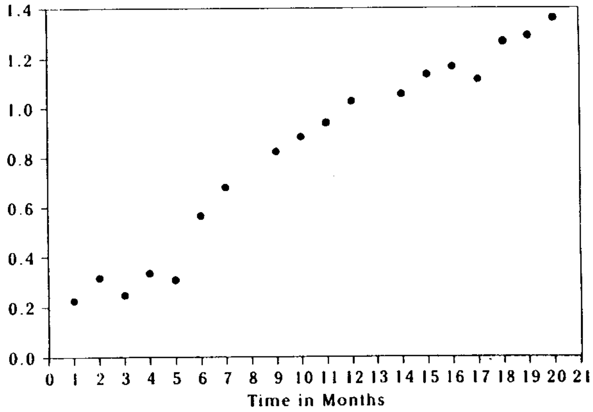


Figure 3

Variance of Monthly Change Over Time

SIC 3533

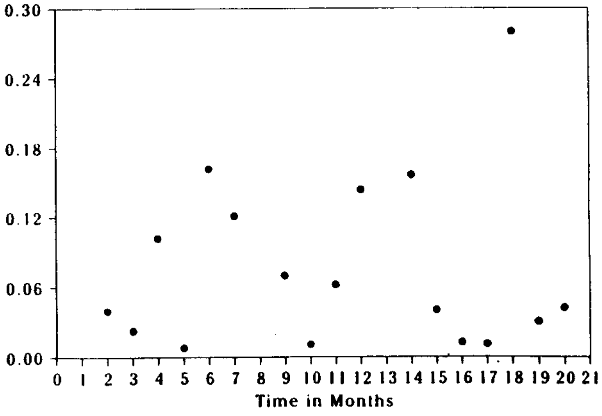


Table 1: Regression Models
Variance of Variance

Industry	R^2	t-Statistic
2061	.99	55.7
2082	.96	20.8
2221	.95	17.9
2311	.99	89.6
2653	.97	23.3

Table 2: Regression Models
Variance of Index

Industry	Predictor	R^2	t-Statistic
2079	Index	.80	19.2
3149	Index	.96	19.6
3511	Index	.97	22.9
3547	Index	.95	17.8
3554	Index	.97	21.6
2511	Time	.97	21.1
3431	Time	.96	15.1
3532	Time	.99	13.1
3554	Time	.92	14.4
3643	Time	.93	3.7

FUTURE WORK

As suggested by the preliminary empirical results, future work will be focused in two major areas; variance trends and modelling variance.

We intend to further investigate the variance trends which were first discovered in the empirical results. Our efforts will be directed toward determining whether the variance trends are an artifact of the index estimator or an indicator of sample deterioration. Quality adjustments, product substitutions, drop-out rates, and price imputations are being investigated as possible indicators of sample deterioration.

We will attempt to refine existing regression models to model the variance as a function of certain industry specific characteristics. Preliminary results show that sample size, index values, indicators of sample deterioration, and factors reflecting the complexity of an industry's aggregation structure seem promising in predicting the variance of an industry in a regression model.

The work done in the area of variance estimation is part of a larger BLS project to develop and implement a production variance estimation system.

REFERENCES

Hansen, M.H., Hurwitz, W.V., Madow, W.G. (1953). Sample Survey Methods and Theory. New York: John Wiley and Sons.

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

Koop, J.C. (1960). On theoretical questions underlying the technique of replicated or interpenetrating samples. Proceedings of the Social Statistics Section, American Statistical Association, 196-205.

——— (1967). Replicated (or interpenetrating) samples of unequal sizes. Ann. Math. Stat., 38, 1142-1147.

——— (1968). An exercise in Ratio Estimation. The American Statistician 22, 29-30.

——— (1972). Bias of the estimate of correlation for a finite universe. Symmetric Function in Statistics. University of Windsor, Ontario, Canada.

——— (1979a). Methods of Variance Estimation for the New BLS Producer Price Indexes II. Progress Report on Task 2. RTI Project No. 255U-1705.

——— (1979b). Methods of Variance Estimation for the New BLS Producer Price Index II - Addendum. Progress Report on Task 2. RTI Project No. 255U-1705.

Krewski, D. and Rao, J.N.K. (1981). Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods. Annals of Statistics, 9, 1010-1019.

Lahiri, D.B. (1954). Technical paper on some aspects of the development of the sample design: National Sample Survey Report No. 5, Sankhya, 14, 264-316.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. The Journal of the Royall Statistical Society, 109, 325-370.

McCarthy, P.J. (1966). Replication: An Approach to the Analysis of Data From Complex Surveys. National Center for Health Statistics, Series 2, No. 14, Washington, D.C.: U.S. Government Printing Office.

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

Shah, B.V. (1977). Variance estimates for complex statistics from multi-stage sample surveys. NSF Symposium on Survey Sampling and Measurement, Chapel Hill, N.C.

Singh, R. and Bansal, M.R. (1975). On the efficiency of interpenetrating subsamples in simple random sampling. Sankhya C, 37, 190-198.

Valliant, R. (1987). Laspeyres Price Index Estimation Under an Autoregressive Model. Proceedings of the Survey Research Methods, American Statistical Association.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag New York.