

## A COMPARISON OF A MODIFIED TILLÉ SAMPLING PROCEDURE TO POISSON SAMPLING

John Slanta<sup>1</sup> and Gary Kusch<sup>2</sup>

### ABSTRACT

Since the late 1950's, the probability surveys in the manufacturing sector within the Manufacturing and Construction Division (MCD) were almost exclusively selected by using Poisson sampling with unit probabilities assigned proportionate to some measure of size. Poisson sampling has the advantage of simplistic variance calculations. Its disadvantage is that the sample size is a random variable, thus adding an additional (and usually positive) component of variance to the survey estimates. In the 1998 survey year, MCD initiated the use of the modified Tillé sampling procedure in some of its surveys. This sampling procedure is used when there is unequal probability of selection and the sample size is fixed. In this paper, we briefly describe this modified procedure and some of its features, and for a variety of dissimilar surveys we contrast variance results obtained using the Tillé procedure to those resulting from the earlier Poisson procedure.

KEY WORDS: Poisson Sampling; Tillé Sampling; Horvitz-Thompson Estimator.

### 1. INTRODUCTION

Prior to 1998, the Manufacturing and Construction Division (MCD) of the U.S. Bureau of the Census primarily used Poisson sampling in the selection of its manufacturing surveys where the sampling units were assigned unequal probabilities of selection. Poisson sampling implies that sampling units are selected or rejected from the sample independently of any other sampling unit. This simplifies variance calculations. One drawback to Poisson sampling is that the sample size is variable. There exist scenarios where an increase in variance will occur because of this variability of the sample size.

Fixed sample size selection schemes involving unequal probabilities were well known (Hanif and Brewer, 1983), but they were difficult to implement in large scale sampling operations, and more importantly, the unbiased sample estimator of the Yates-Grundy-Sen variance (Cochran, 1977) took on negative values. In 1997, MCD modified an existing sampling procedure by Yves Tillé (Tillé, 1996) that was practical to implement and that avoided the problem of negative estimates of variance (Slanta and Fagan, 1997). MCD implemented this sampling procedure for the 1998 Manufacturing Energy Consumption Survey (MECS), the 1998, 1999, and 2000 Industrial Research and Development (R&D) Surveys, the 1999 Survey of Plant Capacity (MQ-C1), and the 1999 Pollution Abatement Capital Expenditures (PACE) survey.

---

<sup>1</sup> John Slanta, U.S. Bureau of the Census, Manufacturing and Construction Division, Room 2225-4, Washington, D.C. 20233, USA

<sup>2</sup> Gary Kusch, U.S. Bureau of the Census, Manufacturing and Construction Division, Room 2225-4, Washington, D.C. 20233, USA

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

This paper reports on a study that compares the coefficients of variation (CVs) of estimates from these surveys using the fixed sample size methodology to what they would have been if Poisson sampling had been used. Select years for three of the four surveys (final data were not available for the PACE survey) were examined. Each of these surveys has unique characteristics that will make the results differ from survey to survey.

## 2. METHODOLOGY OF COMPARING VARIANCES

Our first thought was to compare the variances for consecutive periods for each survey: the first period being the last use of the Poisson based panel and the second being the first use of the fixed sample size based panel. For example, we compared CV's from 1994 MECS to those from the 1998 survey and they showed significant reductions in variance. These results were very promising but we were concerned, especially for this survey because of the four-year difference between panels, that there could be factors affecting the reduction in CVs other than the change to a fixed sample size. We wanted to compare apples with apples and not apples with oranges. We finally concluded that period to period differences were likely subject to additional influences other than the fixed sample size for each of these surveys and that these influences made such comparisons tenuous.

Since the true variances are mathematical functions, we decided to compare estimates of these functions. True variances depend on the data values (mostly unknown), what estimator is being used, and the probabilities and joint probabilities of the sampling units.

Most of these surveys use the Horvitz-Thompson estimator (Cochran, 1977). The Survey of Plant Capacity uses a ratio estimator, where the numerator and denominator are both Horvitz-Thompson estimators. The variance of the Horvitz-Thompson estimator is:

$$\sigma^2(\hat{Y}_{HT}) = \sum_{i=1}^N \left( \frac{1}{\pi_i} - 1 \right) y_i^2 + 2 \sum_{i=2}^N \sum_{j=1}^{i-1} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \quad (1)$$

where,  $y_i$  is the value of interest for the  $i^{\text{th}}$  sampling unit  
 $B_i$  is the probability of selecting the  $i^{\text{th}}$  sampling unit  
 $B_{ij}$  is the probability of selecting both the  $i^{\text{th}}$  and  $j^{\text{th}}$  sampling units  
 $N$  is the population size

There are two relations that should be noted when  $i \dots j$ . The first relation is that when Poisson sampling is used,  $B_{ij} = B_i B_j$ , and the double sum portion of the variance becomes zero. The second relation is that when Tillé sampling is used,  $B_{ij} \neq B_i B_j$ . Therefore, if all the data (values of interest) are nonnegative, then the true variance under Tillé sampling will always be less than or equal to the true variance under Poisson sampling.

Yates, Grundy, and Sen showed that, for fixed sample sizes, the above variance, can also be expressed as:

$$\begin{aligned} \sigma^2(\hat{Y}_{HT, \text{fixed sample size}}) &= \sum_{i=2}^N \sum_{j=1}^{i-1} (\pi_i \pi_j - \pi_{ij}) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \\ &= \sum_{i=2}^N \sum_{j=1}^{i-1} (\pi_i \pi_j - \pi_{ij}) \left( \frac{y_i^2}{\pi_i^2} - 2 \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} + \frac{y_j^2}{\pi_j^2} \right) \end{aligned} \quad (2)$$

Notice that the cross term in (2) equals the double sum term in (1). Therefore, when the sample size is fixed:

$$\sum_{i=1}^N \left( \frac{1}{\pi_i} - 1 \right) y_i^2 = \sum_{i=2}^N \sum_{j=1}^{i-1} (\pi_i \pi_j - \pi_{ij}) \left( \frac{y_i^2}{\pi_i^2} + \frac{y_j^2}{\pi_j^2} \right) \quad (3)$$

Since we had fixed sample sizes for the surveys in question, we will use the form of the Poisson variance that is on the right-hand side of equation (3). Estimates of the mathematical functions (2) and (3), respectively, are:

$$\sum_{i=2}^n \sum_{j=1}^{i-1} (\beta_{ij} - 1) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad \text{and} \quad \sum_{i=2}^n \sum_{j=1}^{i-1} (\beta_{ij} - 1) \left( \frac{y_i^2}{\pi_i^2} + \frac{y_j^2}{\pi_j^2} \right) \quad (4)$$

where  $n$  is the fixed sample size and  $\beta_{ij} = \frac{\pi_i \pi_j}{\pi_{ij}}$ .

If  $\beta_{ij} \neq \beta_j$  for every  $i$  and  $j$  in the population where  $i \dots j$  then the estimates are nonnegative; and if  $\beta_{ij} > 0$  for every  $i$  and  $j$  in the population where  $i \dots j$  then the estimates are unbiased. Note again that if all the data (values of interest) are nonnegative, then the estimate of (2) will always be less than or equal to the estimate of (3), a desirable property.

We made two other changes to (4); the first was to simplify and reduce the number of mathematical operations, and the second was to ratio adjust these functions to reduce the excessive variability of these numbers if more than one sample was taken.

The first adjustment stems from a property of Tillé sampling whereby many values of  $\$_{ij}$  equal  $\$_{i1}$ , for  $j < i$ . The first adjustment is, therefore, a substitution replacing  $\$_{ij}$  with  $\$_{i1}$ .

The second adjustment is a ratio adjustment that compensates somewhat for the instability of the Yates, Grundy, and Sen sample variance. When the sample size is fixed and the probabilities of selection are skewed, some joint probabilities,  $\beta_{ij}$ , could be extremely small. This could lead to an excessively overstated sample variance. The reverse could also be true. Some samples may be picked where the sample variance will be less than a known lower bound (Biyani, 1980). Therefore we multiply the sample variance by:

$$\frac{n^2}{\sum_{i=1}^n \pi_i + 2 \sum_{i=2}^n \sum_{j=1}^{i-1} \beta_{ij}} \quad (5)$$

Note that the expected value of the denominator equals the numerator. For ease in calculation, we replaced  $\$_{ij}$  with  $\$_{i1}$ . This may not fully address the problem presented by Biyani, but this adjustment does add stability to the sample variance.

The estimator for the Survey of Plant Capacity is a ratio estimator, so we will need the covariance terms as well. The covariance between  $\hat{Y}$  and  $\hat{X}$ , where these are Horvitz-Thompson estimators, is:

$$\sigma^2(\hat{Y}_{HT}) = \sum_{i=1}^N \left( \frac{1}{\pi_i} - 1 \right) y_i x_i + 2 \sum_{i=2}^N \sum_{j=1}^{i-1} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{x_j}{\pi_j} \quad (6)$$

If the sample size is fixed, then one estimate of the covariance of  $\hat{Y}$  and  $\hat{X}$  is:

$$\sum_{i=2}^n \sum_{j=1}^{i-1} (\beta_{ij} - 1) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right) \left( \frac{x_i}{\pi_i} - \frac{x_j}{\pi_j} \right), \quad (7)$$

and the estimate of the function  $\sum_{i=1}^N \left( \frac{1}{\pi_i} - 1 \right) y_i x_i$ , i.e. the covariance of  $\hat{Y}$  and  $\hat{X}$  when Poisson sampling is used, is:

$$\sum_{i=2}^n \sum_{j=1}^{i-1} (\beta_{ij} - 1) \left( \frac{y_i x_i}{\pi_i^2} + \frac{y_j x_j}{\pi_j^2} \right). \quad (8)$$

The same two adjustments will be made to these estimates as well. It is necessary to use these covariance estimates as opposed to others so that the variance of the ratio estimate will always be nonnegative.

### 3. SURVEY RESULTS

#### 3.1 R&D Survey

The R&D survey is a company survey conducted annually and collects, among other things, information on R&D expenditures by types of R&D, by industry groupings, and by state. An important feature of the R&D survey is that only a small proportion of the companies in the survey actually have R&D activity, so most of the companies report no expenditures at all. Our study dealt with one of the main publication tables, A-2, which displayed estimates of domestic employment, domestic net sales, number of scientists and engineers, federal R&D, company R&D and total R&D. These six items are broken out by industry groupings and by employment categories. The sampling constraints were univariate and based on total R&D by industry groupings. Our study dealt with the 1999 survey year and excluded aggregated groupings of industries other than the grand total.

The first part of this study dealt exclusively with total R&D by detailed industry groupings. There were originally 48 publication cells that were controlled for sampling. If the estimate in this cell was zero or if the Poisson estimate of variance was zero then that cell was excluded from the study. This left 46 cells to be studied. We then calculated the estimates of the Tillé variance and the Poisson variance. We next derived the “percent reduction” defined as:

$$100 \left( \frac{\hat{\sigma}_{Poisson} - \hat{\sigma}_{Tille}}{\hat{\sigma}_{Poisson}} \right) \quad (9)$$

Histograms were then produced showing the “percent reduction” on the horizontal axis and the counts of the number of publication cells associated with the “percent reduction” on the vertical axis. In the following charts, if the “percent reduction” was within one-half width band of a value in the horizontal axis, then it was grouped with that value. Chart 1 is listed below.

Descriptive Statistics for Chart 1:

Number of Total Cells	46	
Number of Zero Cells	6	(Number of cells where “percent reduction” equaled zero exactly)
Number of Positive Cells	40	

Lowest Percent Reduction            0  
 Highest Percent Reduction        21  
 Median Percent Reduction         3

Most of the cases had marginal improvement, but a few cases showed considerable improvement, especially the one cell that showed a “percent reduction” of 21%. Of note is the fact that the Tillé sampling methodology never resulted in an increased variance relative to Poisson sampling.

Chart 1  
 Total R&D by Industry Groupings

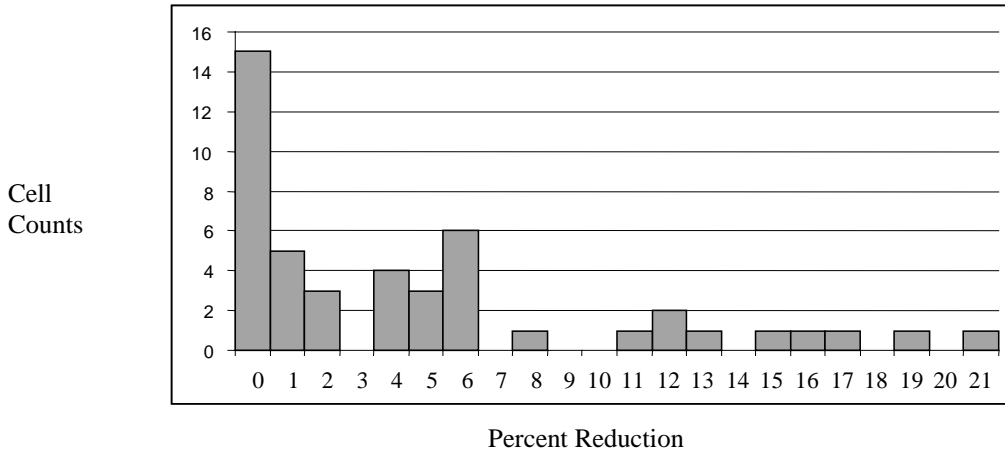
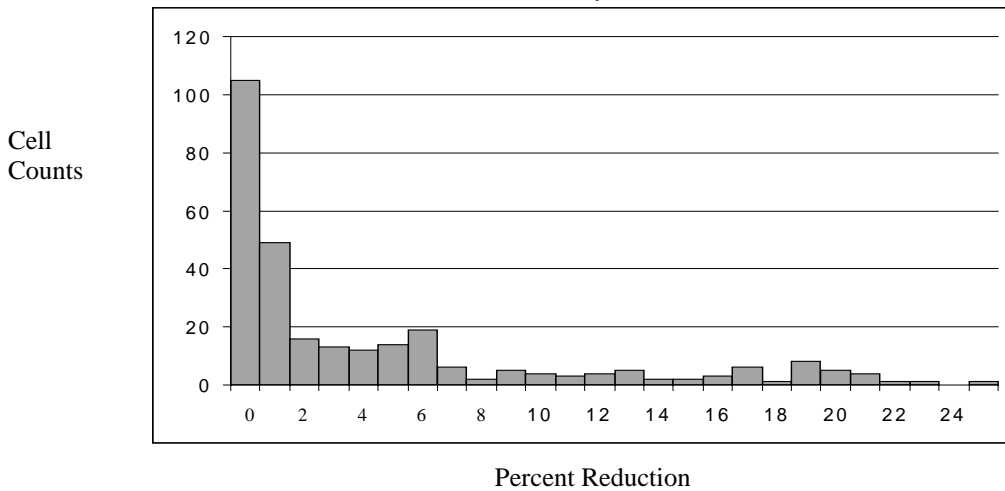


Chart 2 summarizes the results on all cells in publication table A-2, excluding aggregated groupings of industries other than the grand total.

Chart 2  
 R&D Survey Table A-2



Descriptive Statistics for Chart 2:

Number of Total Cells	291	Lowest Percent Reduction	0.0
Number of Zero Cells	50	Highest Percent Reduction	24.7
Number of Positive Cells	241	Median Percent Reduction	1.3

The results are similar, although many more zero values occur. A zero value, which implies the estimate of the Tillé variance equals the estimate of the Poisson variance, could occur when only one noncertainty

sampling unit has a value other than zero. With only a small proportion of noncertainty companies reporting non-zero R&D, this is not an unusual occurrence.

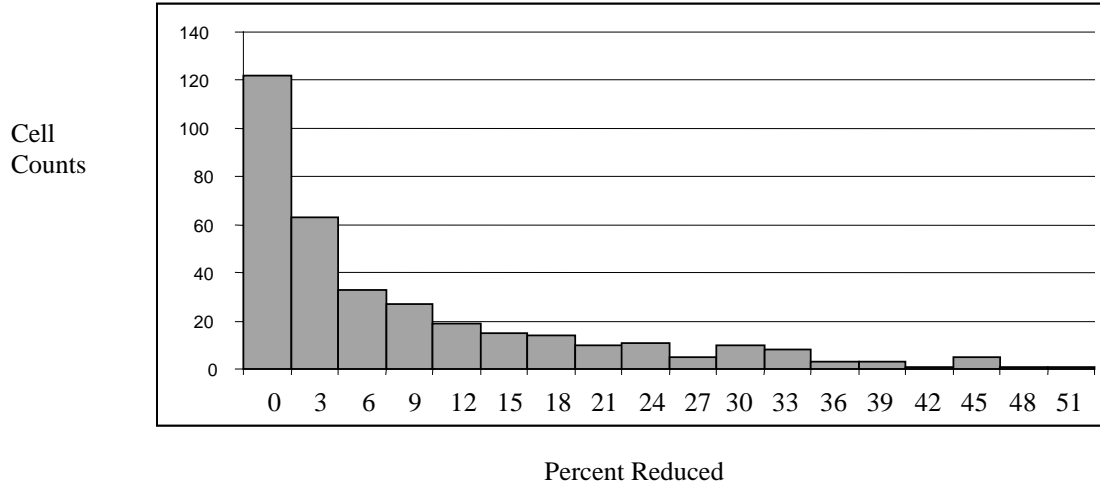
### 3.2 MECS

MECS is an establishment survey and gathers information on energy consumed using different types of fuel. In recent years the survey has been conducted in four-year cycles with the most recent occurring in 1998. We concentrated our study on table N3.1 for that year. This table can be found on the website

<http://www.eia.doe.gov/emeu/mecs/mecs98/datatables/contents.html>

This table shows energy consumption by various types of fuel and by geographic region. We looked exclusively at the total U.S. level and disregarded the geographic regions. The fuel types published in this table were net electricity, residual fuel oil, distillate fuel oil, natural gas, liquefied petroleum gas and natural gas liquids, coal, coke and breeze, other, and total. Chart 3 is listed below and the width of each bar in the histogram is 3.

Chart 3  
MECS Table N3.1



Descriptive Statistics for Chart 3:

Number of Total Cells	351	Lowest Percent Reduction	! 0.05
Number of Negative Cells	3	Highest Percent Reduction	51.4
Number of Zero Cells	33	Median Percent Reduction	3.3
Number of Positive Cells	315		

We investigated the 3 cells where the variance increased using the modified Tillé sampling method. We found a mixture of positive and negative microdata that were used in the calculation of the estimates. The analysts working on the survey confirmed that there could be legitimate situations where the data could be negative. Therefore, one can no longer guarantee that there will be a reduction of variance with a fixed sample size selection. Even without this guarantee, MECS did appear to fare better than the R&D Survey and this is because the R&D Survey had a high percentage of zero-value microdata, which caused the two variance estimates to be more nearly comparable.

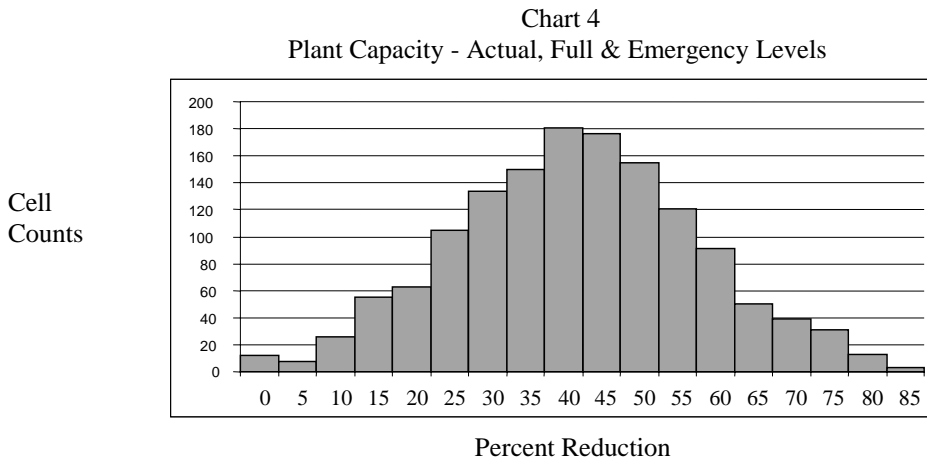
### 3.3 Survey of Plant Capacity

The Survey of Plant Capacity is also an establishment survey and collects production level information. We concentrated our efforts on the 1999 survey year. Three of the items that we looked at were as follows:

1. Market value of actual production for the fourth quarter (we will label this item Actual)
2. Estimated market value of production if plant had been operating at full capacity (we will label this item Full).
3. Estimated value of production if plant had been operating under national emergency (we will label this item Emergency).

Utilization rates are derived from these items. The first utilization rate was the full capacity utilization rate (Actual over Full), and the second utilization rate was the emergency capacity utilization rate (Actual over Emergency). The survey publication includes only the utilization rates; the level estimates are not published but are sometimes made available as special tabulations. Our analysis considered both level estimates and utilization rate estimates, but only for NAICS-6 industries.

Chart 4 summarizes the “percent reduction” for the NAICS-6 level estimates.

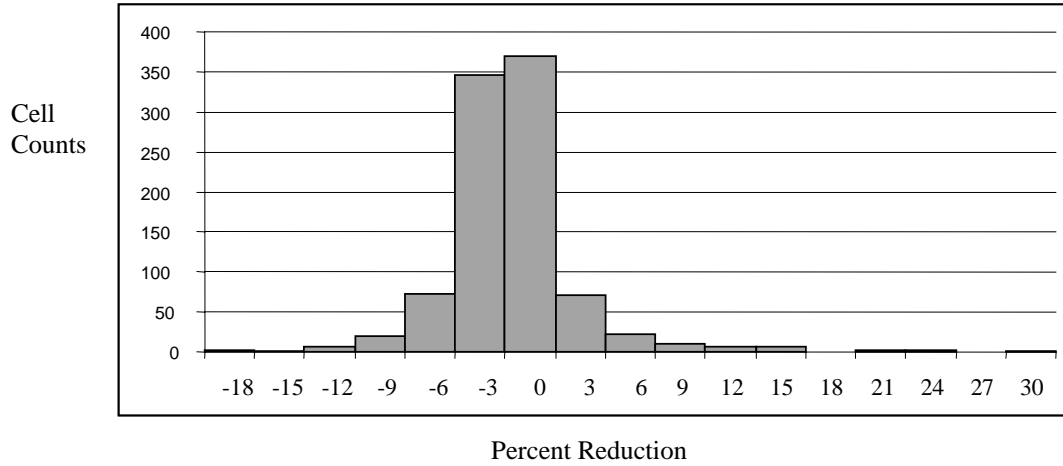


Descriptive Statistics for Chart 4:

Number of Total Cells	1413	Lowest Percent Reduction	0.0
Number of Zero Cells	9	Highest Percent Reduction	85.1
Number of Positive Cells	1404	Median Percent Reduction	42.0

Chart 5 considers “percent reduction” for utilization rates.

Chart 5  
Plant Capacity - Full & Emergency Utilization rates



Descriptive Statistics for Chart 5:

Number of Total Cells	942	Lowest Percent Reduction	! 18.0
Number of Negative Cells	725	Highest Percent Reduction	29.7
Number of Zero Cells	18	Median Percent Reduction	! 1.4
Number of Positive Cells	199		

The “percent reduction” in CVs for level estimates was very good, the median reduction being 42% with no increase in variance. On the other hand, the “percent reduction” in CVs for the utilization rates was poor. In one cell, there was an 18% increase in the CV and there was also an increase in CVs for 77% of the cells. Fortunately, most of the increases in variance were small in magnitude, as seen in Chart 5. The reason for an increase in variance for a high percentage of cells is that the correlation coefficients between level estimates is very high for Poisson sampling. Since the sample size is variable in Poisson sampling, all level estimates tend to be small when the sample size is small and large when the sample size is large. This leads to a high correlation. In a ratio estimator, the higher the correlation the lower the variance.

#### 4. CONCLUSION

In the R&D survey, we saw consistent improvement, even though much of it was marginal. MECS had similar results, but generally showed more improvement than the R&D survey. MECS also demonstrated, however, that a mixture of negative and positive data could cause an increase in variance for the modified Tillé sampling method, but in MECS it was an insignificant increase. The Survey of Plant Capacity showed tremendous reduction in variance for the level estimates that make up the utilization rates. Unfortunately, about three fourths of the utilization rates showed an increase in variance by switching to the modified Tillé sampling. This was due to the estimator for the utilization rates being a ratio estimator and not a Horvitz-Thompson estimator. The high correlation of level estimates when Poisson sampling is used helped reduce the variance of the ratio estimator.

It appears that if the Horvitz-Thompson estimator is used and the data are nonnegative, then there is no penalty switching to Tillé sampling, and the benefits could be large if the data are all or nearly all nonzero for several noncertainty sampling units. If there are only a few sampling units with negative data, then it appears the fixed sample size method is beneficial as well. If the estimators are not Horvitz-Thompson, then there may be no benefit at all to switching to a fixed sample size method.

## REFERENCES

Biyani, Shriram H. (1980), "On Inadmissibility of the Yates-Grundy Variance Estimator in Unequal Probability Sampling", *Journal of the American Statistical Association*, Volume 75, Number 371, pp. 709-712.

Cochran, William G. (1977), *Sampling Techniques*, New York: John Wiley & Sons.

Hanif, M., and Brewer, K. R. W. (1983), *Sampling with Unequal Probabilities*, New York: Springer-Verlag.

Slanta, John G., and Fagan, James T. (1997), "A Modified Approach to Sample Selection and Variance Estimation with Probability Proportional to Size and Fixed Sample Size", unpublished report, MCD Working Paper Number: Census/MCD/WP-97/02.

Tillé, Yves (1996), "An Elimination Procedure for Unequal Probability Sampling Without Replacement", *Biometrika*, **83**, 1, pp. 238-241.