Optimization Targeting of the Annual Refiling Survey October 2015

Marek W. Kaminski
Bureau of Labor Statistics
2 Massachusetts Ave., NE, Suite 4985
Washington, DC 20212
kaminski.mark@bls.gov

Abstract

The *Annual Refilling Survey* (ARS) is a survey of US business establishments that is part of the Quarterly Census of Employment and Wages (QCEW) program. The purpose of the ARS survey is to ensure the correct North American Industrial Classification System (NAICS) is assigned to those establishments listed on the QCEW longitudinal data base. A complete census of the entire population of establishments takes 3 years. On an annual basis 1/3 of all QCEW establishments are surveyed, with the sample chosen from those establishments which did not participate in the prior 2 years.

One measure of the quality of the refiling survey is the estimated percent of establishments misclassified by NAICS. We present an alternative survey technique based on operations research methodology which can potentially reduce the number of ARS surveyed establishments by up to 30% while significantly improving the quality of the survey with respect to the accuracy of industrial classification.

Key Words: NAICS, Quarterly Census of Employment and Wages, Census, Annual Refiling Survey

Introduction

The *Annual Refiling Survey* is a sample survey of establishments performed as a part of the Quarterly Census of Employment and Wages program. The main purpose of the ARS is to ensure correct NAICS classification for all establishments listed on the QCEW longitudinal data base (QCEW-LDB). The accuracy of NAICS classification is important for implementation of BLS programs which use NAICS classification, which currently are: Current Employment Statistics, (CES), Job Openings and Turnover Survey (JOLTS), Local Area Unemployment Statistics (LAUS), Business Employment Dynamics, (BED), and Occupational Safety and Health Survey (OSHS). Most of these programs produce leading economic indicators on monthly basis.

Presently approximately a third of establishments on the QCEW-LDB are sampled every year. This sample rate is uniform across NAICS classification. Thus a complete survey of the whole population of establishments takes 3 years. All QCEW-LDB establishments become part of the current sample provided they did not participate in the survey in the prior two samples.

A sufficiently high sample rate in a given year is important to maintain the accuracy of NAICS classification for the following year. Establishments which are not sampled and have changed their NAICS classification are left in the QCEW-LDB with their old classification in error. At the same

time, the cost of the survey depends on the number of sampled respondents. Both accuracy and cost depend on the sample rate.

It has been observed by BLS staff that the frequency of NAICS classification changes is not uniform across all NAICS classifications; rather the frequency of NAICS classification changes varies considerably by NAICS. This observation suggests that it may be worthwhile to adjust the sampling rate for individual NAICS classifications and thereby improve the quality of the survey as well as reducing its cost. In the industries where NAICS changes more frequently, the rate of sampling can be increased, while in the industries where NAICS changes less frequently the rate of sampling can be decreased.

We recommend that the adjustment of sampling rate be done in three different modes: establishments that are sampled annually, establishments that are sampled every three years, and establishments that are sampled every six years. Thus, establishments in some industries are recommended to be sampled more frequently, in some industries at the current rate, and in some industries at a lesser rate. Establishments that are designated "new births" in the current year are to be sampled at the same rate as the existing establishments within the same NAICS code.

The method to be presented requires that every year the sample rate be recalculated for each 6 digit NAICS. The computation of sampling rates is to be performed for all NAICS separately for each state. The rates are to be computed for each state. In order to compute the rates, only four types of information are required
:

1. The total cost of the survey for a given state.
2. Total number of establishments for each NAICS for the current year on the QCEW-LDB.
3. Total number of establishments which moved from a given NAICS in the previous year to different NAICS in the current year on the QCEW-LDB.
4. The prior year recommended sampling rates for all NAICS.

Sampling rates from the previous year are known and need not to be re-computed. The frequency of NAICS changes between previous year and the current year can be directly computed from the QCEW-LDB.

To find *optimum* rates for each NAICS the method uses *NAICS classification error for establishments*. The *NAICS classification error for establishments* is defined as a percent of all establishments which are misclassified by NAICS code.

The sampling rate for individual NAICS are adjusted to minimize the projected state-level NAICS classification error for the following year. The survey is optimized in the sense that, for a given fixed cost of the survey, the sampling rates are adjusted in such a way that the projected state-level NAICS classification error for establishments is minimized.

The presented methodology can be applied whenever a census is taken in parts over an extended period of time. The generalization of this methodology is beyond the scope of this paper.

Outline of the Paper

Part One. We introduce notation, develop methodology for estimating current year NAICS classification error given previous year rates of sampling, and develop methodology for estimating projected NAICS classification error for the next year given current year rates of sampling.

Part Two. We provide analysis of the problem, develop equations which are used for finding solutions, and provide a proof that the methodology for finding rates for a current year survey is indeed *optimal*. The sample rates are optimal in the sense that, for a given fixed cost of the survey, and other constrains of the survey the projected NAICS classification error is minimized.

Part Three. We present 4 algorithms for finding an optimal solution, each achieving the desired result under given preferences which are selected by the user.

Part Fourth. We present tables with results from the first two algorithms for each state.

1. Methodology

1.1 Notation

yr_t – current year: the year and a quarter in which the ARS is taken.

yr_{t-1} – previous year: the preceding year and a quarter in which the ARS is taken.

yr_{t+1} – next year: the following year and a quarter in which the ARS is taken.

$N_t(naics)$ – true number of establishments existing in the year *t*, for a given NAICS.

$N_t^{LDB}(naics)$ – number of establishments existing in the year *t*, for a given NAICS, **as it is recorded in QCEW-LDB.**

$\Delta_{t-1,t}(naics)$ – number of establishments classified by naics in the previous year *t-1*, with a different classification in the current year *t.*

$\Delta_{t-1,t}^{LDB}(naics)$ – number of establishments classified by *naics* in the previous year *t-1*, with different classification in the current year *t*, **as it is recorded in QCEW-LDB.**

$\Delta_{t-1,t}^{\wedge\, LDB}(naics)$ – number of establishments classified by NAICS in the previous year with different classification in the current year *t* that are **not recorded on QCEW-LDB.** These are establishments which in the previous year *t-1* had classification NAICS, and changed their classification in the current year, *t*, but were not given this new classification due to the fact that they were not interviewed in ARS.

From the above definitions it immediately follows that the total number of establishments which changed NAICS in the current year from the previous year is the sum of establishments which are recorded to change their NAICS, plus the establishments which are not recorded to change their NAICS, since they are not in the survey:

$$\Delta_{t-1,t}(naics) = \Delta_{t-1,t}^{LDB}(naics) + \Delta_{t-1,t}^{\wedge\, LDB}(naics).$$

Note that the quantity $\Delta_{t-1,t}^{LDB}(naics)$ may be directly found from the QCEW-LDB, whereas the quantities $\Delta_{t-1,t}(naics)$ and $\Delta_{t-1,t}^{\wedge\, LDB}(naics)$ may only be estimated. The estimate of the second of these quantities is particularly important for this project. The value of $\Delta_{t-1,t}^{\wedge\, LDB}(naics)$ represents

a total of establishments which are misclassified by NAICS. An estimate of this total will be denoted as $\Delta_{t-1,t}^{\wedge\, \widehat{LDB}}(naics)$.

Let **ω** denote sample ratio of establishments which are included in the survey. For example: if for a given year *t*, and a given NAICS, sample selection occurs annually, then $\omega_t(naics) = 1$, if sample selection occurs every 3 years, then $\omega_t(naics) = 1/3$, if sample selection occurs every 6 years, then $\omega_t(naics) = 1/6$. Note that under the current system $\omega_t(naics) = 1/3$ for every NAICS, and every year *t*.

Similarly, define $\omega_{t-1}(naics)$ and $\omega_t(naics)$ as sample ratios of surveys in the previous and current year respectively. Note that $\omega_{t-1}(naics)$ is always known from the past while $\omega_t(naics)$ is to be determined.
Let **Cost** denote the cost of one individual survey for only one establishment in a given state. Let **Total_Cost** denote the total cost of the survey in a given state. The **Total_Cost** is a total of all funds which are available in a given state to perform the survey.

1.2 Method to Measure Accuracy of NAICS Classification
The accuracy of NAICS classification for all establishments in a given state can be measured by estimating a percent of establishments with an incorrect NAICS classification to the total of all establishments.

The percent of establishments which are **incorrectly** classified in the current year by NAICS code will be called the *NAICS classification error rate for establishments* calculated using the following formula:

$$(1) \qquad naics\_error(yr_t) = \frac{\sum_{\{NAICS\}} \Delta_{t-1,t}^{\wedge\, \widehat{LDB}}(naics)}{\sum_{\{\mathrm{NAICS}\}} N_t(naics)}.$$

Note that the denominator in (1) does not depend on whether establishments are summed with correct or incorrect NAICS, all that matters here is the total sum of establishments, calculated using the following formula:

$$\sum\nolimits_{NAICS} N_t(naics) = \sum\nolimits_{NAICS} N_t^{LDB}(naics).$$

Therefore, formula (1) can be written as follows:

$$(2) \qquad naics\_error(yr_t) = \frac{\sum_{\{NAICS\}} \Delta_{t-1,t}^{\wedge\, \widehat{LDB}}(naics)}{\sum_{\{\mathrm{NAICS}\}} N_t^{LDB}(naics)}.$$

The denominator in (2) can be found directly in QCEW-LDB, The numerator part needs to be estimated.

Similarly, *NAICS classification error for employment* can be defined as a ratio of misclassified employment to the total of all employment. The formulas for employment classification are analogous. In this work we limit our attention to the *NAICS classification error for establishments*.

1.3 Estimation of *naics_error* for the Current Year

In order to estimate the value of naics_error in current year data it is necessary to estimate $\Delta_{t-1,t}^{\wedge LDB}(naics)$ for each NAICS.

In the current survey a third of all establishments are sampled annually. Hence only one third of all changes of NAICS are being detected by the survey. Therefore, under the current system of survey the total number of establishments which are changing NAICS classification in the current year, yr_t, can be estimated (Horvitz-Thompson estimator) as follows:

$$(3) \qquad \Delta_{t,t-1}\widehat{(naics)} = 3 \times \Delta_{t-1,t}^{LDB}(naics).$$

The right hand side in (3) can be directly found from the QCEW-LDB.

In particular, in the current survey, the total number of establishments with changed NAICS which are not recorded in the QCEW-LDB can be estimated as the total of establishments which changed NAICS less the total of known establishments known from QCEW-LDB which changed NAICS. By formula:

$$(4) \qquad \widehat{\Delta_{t-1,t}^{\wedge LDB}}(naics) = \Delta_{t-1,t}\widehat{(naics)} - \Delta_{t,-1t}^{LDB}(naics) = 2 \times \Delta_{t-1,t}^{LDB}(naics).$$

Formula (4) implies that using the current sampling rate the number of establishments with incorrect NAICS can be estimated bythe known number from QCEW-LDB changes, and it is twice as high as the known (from LDB) number of changes. (Infrequently, when the total number of establishments in the current quarter is less than $2 \times \Delta_{t-1,t}^{LDB}(naics)$, this estimation needs to be reduced to the number of total number of establishments in the current quarter.)

This result can be generalized for a situation where the sampling ratio $\omega_{t-1}(naics)$ is not equal to 1/3, but it is an arbitrary number from the interval (0, 1).

Following the same arguments as above, it can be shown that for an arbitrary ratio $\omega_{t-1}(naics)$ a Horvitz - Thompson estimate will yield that the total of establishments which changed their NAICS is

$$\Delta_{t-1,y}\widehat{(naics)} = \frac{1}{\omega_{t-1}(naics)} \times \Delta_{t-1,t}^{LDB}(naics)$$

and the total number of establishments with incorrect NAICS classification in a current quarter can be estimated by the following formula:

$$(6) \qquad \widehat{\Delta_{t-1,t}^{LDB}}(naics) = \left(\frac{1}{\omega_{t-1}(naics)} - 1\right) \times \Delta_{t-1,t}^{LDB}(naics).$$

The NAICS classification error rate, $naics\_error(yr_t)$, as defined in equation (1) can now be estimated from (2) and (5) as follows:

$$(7) \qquad \widehat{naics\_error}(yr_t) = \frac{\sum_{\{NAICS\}}(\frac{1}{\omega_{t-1}(naics)}-1)\times \Delta_{t-1,t}^{LDB}(naics)}{\sum_{\{\text{NAICS}\}} N_t^{LDB}(naics)}.$$

In particular, under the current sampling design:

$$(8) \qquad \widehat{naics\_error\_current}\ (yr_t) = \frac{\sum_{\{NAICS\}} 2\times \Delta_{t-1,t}^{LDB}(naics)}{\sum_{\{\text{NAICS}\}} N_t^{LDB}(naics)}.$$

1.4 Estimation of Variance for the Current Year Estimator of *naics_error*

Having estimates of naics error without having estimates of their variance is not very useful. It can be shown that variance for naics error in (6) can be estimated by:

$$(9) \qquad \widehat{V}(\widehat{naics\_error}(yr_t)) = \left(\frac{1}{\sum_{\{\text{NAICS}\}} N_t^{LDB}(naics)}\right)^2 \sum_{\{NAICS\}} \left\{(1-\omega_{t-1}(naics)) \times (N_t^{LDB}(naics))^2 \times \frac{\hat{p}_t(naics) \times (1-\hat{p}_t(naics))}{\omega_{t-1}(naics) \times (N_t^{LDB}(naics)-1)}\right\}$$

where

$$\hat{p}_t(naics) = \left(\frac{1}{\omega_{t-1}(naics)} - 1\right) \times \frac{\Delta_{t-1,t}^{LDB}(naics)}{N_t^{LDB}(naics)}.$$

In particular, under current sam pling design with $\omega_{t-1}(naics) = 1/3$, the esti mator of variance of estimator in (8) is derived using:

$$(10) \qquad \widehat{V}(\widehat{naics\_error}(yr_t)) = \left(\frac{1}{\sum_{\{\text{NAICS}\}} N_t^{LDB}(naics)}\right)^2 \sum_{\{NAICS\}} \left\{2/3 \times (N_t^{LDB}(naics))^2 \times \frac{\hat{p}_t(naics) \times (1-\hat{p}_t(naics))}{(1/3 \times N_t^{LDB}(naics)-1)}\right\}.$$

1.5 Past NAICS error estimate and its variance

Note that the main goal of this project is NOT to find a design which minimize $\widehat{V}(\widehat{naics\_error})$, as it is commonly performed in survey sampling. The goal is to minimize $\widehat{naics\_error}$. Note that

the top priority is minimization of this error rate, with estimation of the error rate as a secondary goal. The design of the survey is to be made with the idea of achieving $\widehat{naics\_error}$ as small as possible. Also, note that $\widehat{naics\_error}$ needs to be minimized NOT for the current year, but for the subsequent year. Thus, formulas (7) and (10) can only serve to compute $\widehat{naics\_error}$ and its variance in the current year and thereby serves to evaluate how effective the survey was in the preceding year. Specifically, the value $\widehat{naics\_error}(yr_t)$ may only be used to evaluate how optimal was the choice of sampling rates ω for the preceding year.

The following section details the projection of the estimated naics error in the next year, given the selection of sampling rates ω in the current year for the survey.

1.6 Projected *naics_error*

The selection of sampling rates in the current sample year has its impact on the accuracy of NAICS classification for the following year. The projected NAICS classification error for the next year, given a specific selection of sampling rates ω in the current year is performed below. For projected NAICS classification error in the next year, using similar arguments as previously, the equations (5), (6) and (7) cited earlier can be adopted with small changes as follows:

$$\widehat{\Delta^{\wedge LDB}_{t,t+1}}(naics) = (1 - \omega_t(naics)) \times \Delta_{t,t+1}(naics) \tag{11}$$

where, $\omega_t(naics)$ is a rate for the current year.

The number of all establishments that changed NAICS classification in the next year can be estimated by the estimate obtained from the previous year, by formula

$$\Delta_{t,t+1}\widehat{(naics)} = \Delta_{t-1,t}\widehat{(naics)}. \tag{12}$$

The accuracy of estimation in (12) is outside the scope of this work and will not be performed here. From (11) it follows that:

$$\Delta^{\wedge \widehat{LDB}}_{t,t+1}(naics) = \big(1 - \omega_t(naics)\big) \times \Delta_{t-1,,t}\widehat{(naics)} \tag{13}$$

then, as previously:

$$\Delta^{\wedge \widehat{LDB}}_{t,t+1}(naics) = \big(1 - \omega_t(naics)\big) \times \frac{1}{\omega_{t-1}(naics)} \times \Delta^{LDB}_{t-1,t}(naics) \tag{14}$$

and thus the projected NAICS classification error for the next year is derived as follows:

$$\widehat{naics\_error}\ (yr_{t+1}) = \frac{\sum_{\{NAICS\}}(1 - \omega_t(naics)) \times \frac{1}{\omega_{t-1}(naics)} \times \Delta^{LDB}_{t-1,t}(naics)}{\sum_{\{\mathrm{NAICS}\}} N^{LDB}_t(naics)} \tag{15}$$

where $\omega_{t-1}(naics)$ a rate used in the previous year, and $\omega_t(naics)$ is a rate to be used in the current year. For the purpose of optimization for the next year formula (15) will be used.

2. Optimization of the Survey

2.1 Problem of Finding Sampling Rates

The initial purpose of this project was to design a survey which would produce the lowest error with the least resources. Finding a design which produces lower error means finding a set of rates $\{\omega_t(naics): naics \in NAICS\}$ which consume less resources while yielding the smallest $naics\_error(yr_{t+1})$. In this sense the process of finding the best rates at a given cost can be called survey optimization. The formula for projected NAICS error is given in (15). Next, the formula for the survey cost is derived. The total is the cost of one individual unit multiplied by the number of survey units

$$(16) \quad \text{Total Cost} = \mathrm{C} \times \sum_{\{NAICS\}} \omega_t(naics) \times N_t^{LDB}(naics)$$

where C is the cost of one individual survey, and $\omega_t(naics) \in \{1/6, 1/3, 1\}$. Therefore, we have arrived at the following optimization problem:

Find $\{\omega_t(naics): naics \in NAICS\}$ such that

MINIMIZE:

$$(\bullet) \qquad naics\_error\ (yr_{t+1}) = \frac{\sum_{\{NAICS\}}(1-\omega_t(naics)) \times \frac{1}{\omega_{t-1}(naics)} \times \Delta_{t-1,t}^{LDB}(naics)}{\sum_{\{\mathrm{NAICS}\}} N_t^{LDB}(naics)}$$

under the following constrains:

$$(\bullet\bullet) \qquad \text{Total Cost} = \mathrm{C} \times \sum_{\{NAICS\}} \omega_t(naics) \times N_t^{LDB}(naics)$$

and

$$(\bullet\bullet\bullet) \qquad \omega_t(naics) \in \{1/6,\ 1/3,\ 1\}$$

where C is the cost of one survey unit. Here the Total Cost, is a fixed given value, which is assigned to each state. Note that the first sum obtains its minimum when the sum

$$\sum_{\{NAICS\}} \omega_t(naics) \times \frac{1}{\omega_{t-1}(naics)} \times \Delta_{t-1,t}^{LDB}(naics)$$

obtains its maximum. Therefore, it is equivalent to find a maximum of the second sum in order to finding minimum of the first one. The last expression is mathematically and conceptually easier to work with than the first, and therefore the last expression is used for the further study. The above setup of finding minimum or maximum for a linear function under a set of constrains is very typical in operations research.

2.2 Analysis of the Optimization Problem

In order to find a suitable solution to the optimization problem, a new simpler notation is introduced. Assume that there are n different NAICS codes in a given state. Assume further that NAICS codes are given indices, i.e. $\{ \omega_t(naics) : \ naics \in \text{NAICS} \} = \{ \omega_i : 1 \leq i \leq n \}$.

Assume accordingly that indices are being assigned to the terms

$$\frac{1}{\omega_{t-1}(naics)} \times \Delta_{t-1,t}^{LDB}(naics)$$

and that these terms are denoted by the letter d, i.e.

$$\left\{ \frac{1}{\omega_{t-1}(naics)} \times \Delta_{t-1,t}^{LDB}(naics) : \ naics \in \text{NAICS} \right\} = \{ d_i : \ 1 \leq i \leq n \}.$$

The terms $N_t^{LDB}(naics)$ are given indices, and denoted by the letter e, i.e.

$$\{ N_t^{LDB}(naics) : \ naics \in \text{NAICS} \} = \{ e_i : 1 \leq i \leq n \}.$$

Next, let f and g be defined

$$f(\omega_1, \omega_2, \dots, \omega_n) = \sum_{i=1}^{n} (\omega_i \times d_i)$$

and

$$g(\omega_1, \omega_2, \dots, \omega_n) = C \times \sum_{i=1}^{n} (\omega_i \times e_i).$$

Thus, the optimization problem can be stated as follows:

MAXIMIZE:

$$(*) \qquad f(\omega_1, \ \omega_2, \dots, \omega_n)$$

under the constraints:

$$(**) \qquad Total\_Cost \ = g(\omega_1, \omega_2, \dots, \omega_n)$$

and

$$(***) \quad \omega_i \in \left\{ \frac{1}{6},\ \frac{1}{3},\ 1 \right\} \qquad 1 \le i \le n\,.$$

Before finding a solution to the problem stated, making some useful observations is necessary. Note that the function f is a linear function with respect to $\omega_1, \omega_2, \ldots, \omega_n$. (Function f is therefore a harmonic function, and hence it assumes both minimum and maximum values on the edges of the closed region defined by (**)). The problem of finding the maximum of this function has a solution and the solution must be located on the edge of region defined by (**).

Through a series of simple algebraic computations it can be shown that condition (**) can be incorporated into expression (*) giving the following identity:

$$(17) \qquad \sum_{i=1}^{n} (\omega_i \times d_i) = \frac{1}{n} \sum_{i=1}^{n} \omega_i \left[(n-1) \times d_i \ - \ e_i \times \sum_{j=1,\ j \ne i}^{n} \frac{d_j}{e_j} \right] + \frac{1}{n} \times \frac{Total\_Cost}{Cost} \times \sum_{i=1}^{n} \frac{d_i}{e_i}$$

for $\omega_1,\ \omega_2,\ \ldots, \omega_n$ satisfying the constraint (**).

In order to find a maximum of the function $f(\omega_1, \omega_2, \ldots, \omega_n)$ under the constraints (**) and (***), one can use the following procedure:

Order the series:

$$\{ (n-1) \times d_i \ - \ e_i \times \sum_{j=1,\ j \ne i}^{n} \frac{d_j}{e_j} \ ; \quad 1 \le i \le n \}$$

to be a decreasing sequence. For simplicity, suppose that indices i are assigned in such a way that

$$(18) \quad (n-1) \times d_1 - e_1 \times \sum_{j=1,\ j \ne 1}^{n} \frac{d_j}{e_j} \ge (n-1) \times d_2 - e_2 \times \sum_{j=1,\ j \ne 2}^{n} \frac{d_j}{e_j} \ge \ldots \ge (n-1) \times d_n - e_n \times \sum_{j=1,\ j \ne n}^{n} \frac{d_j}{e_j}.$$

It can be proven that, the first element is always positive and the last element is always negative, i.e., we have that:

$$(19) \quad (n-1) \times d_1 - e_1 \times \sum_{j=1,\ j \ne 1}^{n} \frac{d_j}{e_j} > 0 \quad \text{and} \quad (n-1) \times d_n - e_n \times \sum_{j=1,\ j \ne n}^{n} \frac{d_j}{e_j} < 0.$$

Therefore, there exists an integer $k\ such\ that\ 1 < k < n$, which divi des sequence (18) into all positive and and all negative parts.

From equation (17) it follows that the most eff ective way to maximize the value of the function $f(\omega_1, \omega_2, \dots, \omega_n)$ is to assign, under the constraints given, the maximum values for ω's with low indices, and assign the minimum values for ω's with high indices. This way the largest values of $\omega's$ are assigned to the positive part of the sequence (18) and the lowest to the negative part. That leads to relation:

(20) $\omega_1 \geq \omega_2 \geq \omega_3 \geq \dots \geq \omega_n$.

These observations provide basis on which algorithm s for finding solutions to the stated proble m can be made. The following simple algorithm can significantly reduce cost of survey and improve the quality of survey.

3. Proposed Algorithms

The algorithms below provide a solutio n to the st ated problem. Each has cert ain advantages and disadvantages depending on a particular use.

The first algorithm is the simplest one. This algorithm takes advantage of the fact that the sequence (18) can be divided into positive and negative part s by assigning 1's to the first part and 1/6' s to the second part. The first step is not needed for algorithm itself, nevertheless it is recommended to be made for a sake of being able to com pare achieved accuracy and cost to the present accuracy and cost.

Algorithm 1

Step 1

First, assume $\omega_1 = \omega_2 = \cdots = \omega_n = 1/3$, and compute:

(21) $M_0 = f(1/3,\ 1/3, \dots, 1/3)$ and $C_0 = g(1/3, 1/3, \dots, 1/3)$.

Step 2

Find an integer k such that:

$$d_k - \frac{1}{n-1} e_k \sum_{j=1,\ j\neq k}^{n} \frac{d_j}{e_j} \geq 0 \quad \text{and} \quad d_k - \frac{1}{n-1} e_{k+1} \sum_{j=1,\ j\neq k+1}^{n} \frac{d_j}{e_j} < 0.$$

From equations (19), such an integer k always exists, and $k \in \{2,\ 3,\ \dots\ ,\ n-1\}$.

Step 3

Compute $M_1 = f(\omega_1,\ \omega_2, \dots, \omega_n)$ and $C_1 = g(\omega_1,\ \omega_2, \dots, \omega_n)$ where:

$\omega_1 = 1,\ \omega_2 = 1,\ \dots, \omega_k = 1$ and $\omega_{k+1} = 1/6,\ \omega_{k+2} = 1/6,\ \dots, \omega_n = 1/6$

That is, assume for ω_i value 1 when $i \leq k$ and value 1/6 when $i > k$.
■

The next algorithm is meant to make a complete use of available resources, defined as $C_0 = g(1/3, 1/3, \dots, 1/3)$, and it is designed to slightly improve the quality of the survey in comparison to the previous algorithm. The main advantage of this algorithm versus the other one is that it is the most similar to the current system of surveying all establishments every 3 years. Thus, it is a smooth change between the current method and Algorithm 1.

Algorithm 2
Steps 1 – 3 are identical as in the previous algorithm.
Step 4
Assume $\omega_1 = \omega_2 = \cdots = \omega_n = 1/3$. Compute:

(21) $M_0 = f(1/3,\ 1/3, \dots, 1/3)$ and $C_0 = g(1/3, 1/3, \dots, 1/3)$.

Next compute the value $C_1 = g(\omega_1,\ \omega_2, \dots, \omega_n)$. Compare C_1 to C_0. There are two possibilities: $C_1 \geq C_0$ or $C_1 < C_0$.

If $C_1 \geq C_0$:
then consecutively decrease the values $\omega_1,\ \omega_2,\ \dots, \omega_k$ to 1/3 starting from the value ω_k, next ω_{k-1}, next ω_{k-2}, each time computing $g(\omega_1,\ \omega_2, \dots, \omega_n)$ until for the first time the value of $g(\omega_1,\ \omega_2, \dots, \omega_n)$ is less than C_0.

If $C_1 < C_0$:
then consecutively increase the values $\omega_{k+1},\ \omega_{k+2}, \dots, \omega_n$ to 1/3 starting from the value ω_{k+1}, next ω_{k+2}, next ω_{k+3}, each time computing $g(\omega_1,\ \omega_2, \dots, \omega_n)$ until the value of $g(\omega_1, \omega_2, \dots, \omega_n)$ is greater than C_0. Then chose the iteration which precedes the iteration when $g(\omega_1, \omega_2, \dots, \omega_n)$ is greater than C_0.

(In the second case, $C_1 < C_0$, one can also chose an option of not increasing values $\omega_{k+1}, \omega_{k+2}, \dots, \omega_n$.)

The described process can be defined more formally as follows:

If $C_1 \geq C_0$:
Define a matrix $[\omega_{i,j}]_{i \leq k, j \leq n}$ as follows:
$\omega_{1,1} = 1,\ \dots,\ \omega_{1,k-2} = 1,\quad \omega_{1,k-1} = 1,\quad \omega_{1,k} = 1$ and $\omega_{1,k+1} = 1/6,\ \dots, \omega_{1n} = 1/6$
$\omega_{2,1} = 1,\ \dots,\ \omega_{2,k-2} = 1,\quad \omega_{2,k-1} = 1,\quad \omega_{2,k} = 1/3$ and $\omega_{2,k+1,} = 1/6,\ \dots, \omega_{2,n,} = 1/6$

$\omega_{3,1} = 1,\ \ldots,\ \omega_{3,k-2} = 1,\quad \omega_{3,k-1} = 1/3,\ \omega_{3k} = 1/3$ and $\omega_{3,k+1} = 1/6,\ \ldots,\omega_{3,n} = 1/6$

. . . . . .

$\omega_{k,1} = 1/3,\ \ldots,\omega_{k,k-2} = 1/3,\ \omega_{k,k-1} = 1/3,\ \omega_{kk} = 1/3$ and $\omega_{k+1,4} = 1/6,\ \ldots,\omega_{n,4} = 1/6$

Compute g and f for each vector of ω's. Let C_i and M_i be:

$C_i = g(\omega_{1,i}, \omega_{2,i}, \ldots, \omega_{n,i})$; $\quad 2 \le i \le k$

$M_i = f(\omega_{1,i}, \omega_{2,i}, \ldots, \omega_{n,i})$; $\quad 2 \le i \le k$

Let $m = \min\{ i : \ C_i \le C_0 \}$, then $(\omega_{1,m}, \omega_{2,m}, \ldots, \omega_{n,m})$ is the vector of $\omega's$ with the maximum value for the function f, under the constraint $C_m < C_0$, i.e.

$M_m = \max\{ M_i : \ C_i \le C_0 \}$.

If $C_1 < C_0$:

Define a matrix $[\omega_{i,j}]_{i \le n-k, j \le n}$ as follows:

$\omega_{1,1} = 1,\ \ldots,\ \omega_{1,k} = 1$ and $\omega_{1,k+1} = 1/6,\ \omega_{1,k+2} = 1/6,\ \omega_{1,k+3} = 1/6, \ldots, \omega_{1n} = 1/6$

$\omega_{2,1} = 1,\ \ldots,\ \omega_{2,k} = 1$ and $\omega_{2,k+1} = 1/3,\ \omega_{2,k+2} = 1/6,\ \omega_{2,k+3} = 1/6, \ldots, \omega_{2,n} = 1/6$

$\omega_{3,1} = 1,\ \ldots,\ \omega_{3,k} = 1$ and $\omega_{3,k+1} = 1/3,\ \omega_{3,k+2} = 1/3,\ \omega_{3,k+3} = 1/6, \ldots, \omega_{3,n} = 1/6$

. . . . . .

$\omega_{n-k,1} = 1, \ldots, \omega_{n-k,k} = 1$ and $\omega_{n-k,k+1} = 1/3,\ \omega_{n-k,k+2} = 1/3,\ \omega_{n-k,n} = 1/3, \ldots, \omega_{n,3} = 1/3$

Compute:

$C_i = g(\omega_{1,i}, \omega_{2,i}, \ldots, \omega_{n,i})$; $\quad 2 \le i \le k$

$M_i = f(\omega_{1,i}, \omega_{2,i}, \ldots, \omega_{n,i})$; $\quad 2 \le i \le k$

Let $m = \max\{ i : \ C_i > C \}$, then $(\omega_{1,m-1}, \omega_{2,m-1}, \ldots, \omega_{n,m-1})$ is the vector of $\omega's$ with the maximum value for the function f, under the constraint $C_m < C_0$, i.e. $M_{m-1} = \max\{ M_i : \ C_i \le C \}$.

In both cases, the obtained vector of ω's is the final solution to the maximization of the function f, therefore it is the vector which gives the minimum NAICS classification error for the next year.

■

The third algorithm will maximize the function f, for the complete use of available resources, and it will divide all establishments into only two categories: Establishments which are surveyed every one year and establishments which are surveyed every 6 years.

Algorithm 3

Steps 1 – 3 are identical as in the previous algorithm.

Step 4

Now, Step 4 is identical to the Step 4 in the pr evious algorithm, except that instead of using 1/3 rate when $C_1 \geq C_0$ the rate 1/6 is used, and instead 1/3 when $C_1 < C_0$ the rate 1 is used.
■

In Algorithm 2 and Alg orithm 3 the value of current expense for the whole survey , C_0 , is gi ven and is matched by the process. In the fourth, a nd the last proposed algorithm the user defines any value the as the whole expense of the survey. Thus, it produces best results for any given available resources.

Algorithm 4
All steps are identical as in Algorithm 3, but now the value C_0 can be replaced by any fixed number between $g(1/6_{(1)}, 1/6_{(2)}, \ldots, 1/6_{(k)}, 1/6_{(k+1)}, 1/6_{(k+2)}, \ldots, 1/6_{(n)})$ and $g(1_{(1)}, 1_{(2)}, \ldots, 1_{(k)}, 1_{(k+1)}, 1_{(k+2)}, \ldots, 1_{(n)})$.
■

4. Results

The results are limited to the first two algorithms which by the opinion of the author are most suited to the BLS e nvironment. The presented results wer e computed from QCEW-LDB, first quarters between previous year of 2011 and the current year of 2012.

Table – Results for Algorithm 1 and 2. Previous year 2012, Current year 2013.

state	Total Under Number Currently Performed Survey	Projected Total Number Surveyed	Projected Total Number Surveyed Under Algorithm 2 (incudes 3 years mode)	Total Num. of Estab. from the Previous Quarter with diff. naics Code in the Curr. Quar. As Recorded in LDB	Estimated Total m. of Estab. from the Previous Quarter with Incorrect naics Code in the Curr. Quarter	Estimated Current naics Error	Estimated naics Standard Error of Current naics Error	Projected naics Error Under Algorithm 1 (no 3 year mode)	Projected naics Error Under Algorithm 2 (includes 3 years mode)
AL	3616 4	21137	36132	1334	2604	0.024	0.00038	0.017674	0.014429
AK	6517 6016		6514 462		831	0.0425	0.00121	0.012786	0.012096
AZ	4693 3	29739	46825	4587	8318	0.05908	0.00047	0.043395	0.03668
AR	2714 2	18930	27085	1071	2125	0.0261	0.00051	0.0203861	0.017248
CA	4185 38	274415	410997	1071	2125	0.0261	0.00051	0.033621	0.026915
CO	5567 9	33398	55413	28744	56465	0.04497	0.00016	0.023363	0.019103
CT	3402 8	20000	33926	2730	4995	0.0299	0.00035	0.021551	0.017686

DE	8939		5674	8845	1259	2445	0.02395	0.00043	0.02144	0.018121
DC	9780	7651	9733	394	756	0.02819	0.00133	0.037661	0.033741	
FL	2007	35	125485	200503	1429	2830	0.09645	0.00179	0.043831	0.036153
GA	8633	1	60021	86149	14383	27696	0.04599	0.00036	0.020435	0.017591
HI	1216	6	11547	12164	7106	14141	0.0546	0.00049	0.014932	0.014576
ID	1659	0	13181	16567	570	1137	0.03115	0.00122	0.02833	0.025336
IL	1239	06	62874	122761	1151	2172	0.04364	0.0012	0.028341	0.022807
IN	5079	2	31676	50768	4468	8814	0.02371	0.00035	0.023462	0.019357
IA	2941	4	19847	29340	2043	4010	0.02632	0.00056	0.024903	0.020886
KS	2586	8	21225	25839	1436	2810	0.03184	0.0008	0.026352	0.023807
KY	3426	9	24464	34251	1504	2980	0.0384	0.00094	0.015077	0.01299
LA	3946	7	22567	39118	1127	2242	0.02181	0.00062	0.027998	0.022761
ME	1506	0	13481	15059	2024	3986	0.03367	0.00069	0.015604	0.01484
MD	5258	9	31824	52447	680	1341	0.02968	0.00107	0.039963	0.032985
MA	7028	9	39981	69925	3362	6629	0.04202	0.00069	0.02453	0.019631
MI	7466	8	67590	74645	2624	5169	0.02451	0.00046	0.016026	0.015109
MN	5280	3	31445	52560	7263	14525	0.06484	0.00057	0.028802	0.023635
MS	2155	4	19947	21509	2536	4889	0.03086	0.00059	0.008119	0.007756
MO	5587	3	35764	55654	535	1068	0.01652	0.0007	0.01569	0.012961
MT	1349	1	10071	13464	1638	3197	0.01907	0.00045	0.021804	0.018975
NE	2138	4	13983	21356	679	1337	0.03303	0.00119	0.02455	0.020116
NV	2400	1	12894	23840	984	1914	0.02983	0.00091	0.023957	0.019429
NH	1527	9	12260	15252	1557	2655	0.03687	0.00077	0.017235	0.01585
NJ	8384	8	52832	83571	631	1243	0.02712	0.00103	0.031485	0.026119
NM	1710	2	10606	17100	8195	16267	0.06467	0.00052	0.029235	0.024187
NY	1934	46	143025	193019	849	1656	0.03228	0.00107	0.02421	0.020953
NC	8035	6	42502	79684	22520	44628	0.0769	0.00033	0.044831	0.03635
ND	9279		8287	9278	4904	9635	0.03997	0.00055	0.005119	0.004832
OH	8932	0	47160	88378	201	394	0.01415	0.00095	0.039782	0.032354
OK	3281	9	21089	32664	4873	9649	0.03601	0.0005	0.031536	0.026316
OR	4097	6	30011	40900	1804	3519	0.03574	0.00081	0.022228	0.019503
PA	1103	04	59417	108165	1669	3300	0.02685	0.00064	0.038311	0.031206
RI	11367	7404	11323	5940	11748	0.0355	0.00044	0.010703	0.009134	
SC	3547	9	24155	35362	303	573	0.0168	0.00091	0.032202	0.027434
SD	9604		7449	9595	1998	3955	0.03716	0.00079	0.010932	0.009943
TN	4502	8	26234	44968	316	607	0.02107	0.00113	0.027002	0.022434
TX	1880	90	96875	187138	1854	3681	0.02725	0.00061	0.040163	0.032237
UT	2708	7	15857	27000	10011	19673	0.03487	0.00033	0.025535	0.020668

VT	7428	5013	7417	1086	2140	0.02634	0.00077	0.01178	0.010254
VA	7519	8 46811	75111	214	413	0.01853	0.00122	0.022906	0.01875
WA	7689	4 52065	75987	3787	7519	0.03333	0.00047	0.018879	0.015484
WV	1456	5 10641	14520	2502	4985	0.02161	0.00042	0.008696	0.007712
WI		48561 37476	48524	353	695	0.01591	0.0008	0.02073	0.018159
WY	7899	8011	7566	5064	9219	0.06328	0.00058	0.004642	0.005401
	2,884,899	1, 852,007	2,865,911						

As it is seen in the table the Algorithm 1 cut over 1,000,000 surveyed units (thus, can save hundreds of thousands of dollars!) while delivering for almost all states projected NAICS error about half of present estimated error. (Note that Algorithm 1 does not necessarily lead to decrease in the number of surveyed units for every state, for example, WY.) Algorithm 2 used all available resources while still improving NAICS error over Algorithm 1. The results for other years are similar.

Recommendations:

1. Perform more simulations
2. Examine bias and balance of flows in-and-out of industries
3. Consider cost and efficiency
4. Recommend an algorithm to replace the current methodology

Acknowledgement:
Many thanks to my coworker Mark Crankshaw for helping me with the final edition of this article.

Disclaimer:
Views express in this document are those of the author and do not necessarily reflect the views or policies of the Bureau of Labor Statistics.

References:

Michel A. Searson, Redesigning Data Collection Strategies for Cost Reduction in Two Bureau of Labor Statistics Surveys, U.S. Bureau of Labor Statistics, Proceedings of International Conference on Establishment Surveys - III, June 18-21, 2007, Montreal, Quebec, Canada