Redesigning the Sample of the Company Organization Survey Using Predictive Modeling

Matthew Thompson[1], Chrishelle Lawrence[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Suitland, MD 20746

Abstract

The Company Organization Survey (COS) is conducted annually between economic census years to collect data on multi-unit companies operating in the United States. The non-probability sample consists of certainty multi-unit companies, single-unit companies suspected of being multi-unit companies, and a targeted sample of multi-unit companies. Since organizational change is likely for larger companies, multi-unit companies with 250 or more employees are selected for the sample with certainty, and single-unit companies that meet certain criteria are selected. The targeted sample aims to maximize the number of multi-unit companies with less than 250 employees with organizational changes. Organizational change occurs when at least one of a company's units is opened, closed, moved, or sold. This research focuses on revising the sample for multi-unit companies. A logistic regression model was used to predict the likelihood of an organizational change. Results suggest that an improved sample could be obtained when multi-unit companies with 500 or more employees are selected with certainty, and companies with less than 500 employees are targeted by their likelihood of organizational change.

Keywords: Non-probability sample, logistic regression, establishment survey

Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

1. Introduction

1.1 The Company Organization Survey

The purpose of the Company Organization Survey, also known as the Report of Organization Survey, is to obtain current organization and operating information on multi-establishment companies – that is, companies operating in multiple locations – in order to maintain the Census Bureau's Business Register. The United States Code, Title 13, authorizes this survey and provides for mandatory responses. The Census Bureau uses the data to maintain up-to-date company affiliation, location, and operating information for establishments of multi-establishment companies. However, this survey is taken primarily to assure full coverage and high quality of other statistical programs and does not provide data products for public use. COS provides the only direct source of information on changes in multi-establishment company organization and industry classification at the establishment level between Economic Census years, which occurs every 5 years.

The COS is an annual survey. Although most of the sample is composed of larger, multi-establishment companies, smaller companies are selected where administrative data indicate a probable organizational change. In general for COS, companies identify establishments that have been sold, closed, continued, started, and acquired; report first quarter and annual payroll, and employment during the pay period that included March 12, for each establishment; indicate any large foreign equity positions; and indicate controlling interests held by other domestic or foreign-owned organizations. However, during the Economic Census COS companies only report basic company affiliation and operations of establishments not within scope of the Economic Census.

About 42,000 multi-establishments companies and 5,000 single-establishment companies are annually chosen to complete the COS. The COS sample is comprised of three parts. The largest portion of the sample is made up of large multi-establishment companies selected with certainty. Several criteria will result in a company being selected with certainty. First any company that is selected to participate in the Annual Survey of Manufacturers (ASM) or the Business Sample Revision (BSR) is selected into the COS with certainty. In addition, any active multi-establishment with at least 50 employees and one manufacturing establishment is selected with certainty. These companies, while not ASM companies, are of interest because of the likelihood of a business birth in the manufacturing industry. And finally, any company with 250 or more employees is selected with certainty.

The selection of certainties is then supplemented with a targeted sample of non-certainty companies, as well as a selection of 5,000 single-establishment companies showing signs of operating at more than one location. The targeted sample includes smaller multi-establishment companies where administrative data indicates that an organizational change may have occurred.

1.2 Purpose of Sample Redesign

Recommendations were made to research the best indicators of organizational change and to potentially update the certainty cutoff. If a change was not warranted, justification of the current cutoff was also recommended. The current targeting methodology dates from the late 1990s. In addition, there are an increasing number of companies with over 250 employees, which has reduced the number of smaller companies available for selection as part of the targeted sample.

The goal of the COS Sample Redesign research is to establish statistical justification for the COS sample design, including the company certainty cutoff and the targeted sample. As mentioned above, the annual COS sample is set by budget to consist of approximately 42,000 multi-establishment companies, and it is not expected that this total will change. This research focuses solely on the multi-establishment sample and does not address the 5,000 single-establishment companies in sample. In addition, the ASM and BSR component of the COS will not change. This research will focus on the sample selection of non-ASM and non-BSR certainties and non-certainty (i.e. targeted) companies that compose the remainder of the overall COS sample.

The main objective of the COS is to maintain the list of establishments for multi-establishment companies, which is not available from administrative records. The most effective way to maintain this list is to capture the highest percentage possible of companies opening or closing establishments, as well as to ensure a high percentage of payroll coverage, as a measurement of size for economic activity.

It is assumed that large companies with employment over a certain employment size make a significant contribution to the economy in addition to having a greater likelihood of opening or closing establishments. This research will determine an appropriate employment size cutoff to meet the research goals of economic coverage in addition to selecting companies that are opening or closing establishments.

The research goal for non-certainty companies is to select a sample of companies to maximize the number of companies undergoing change, which is measured by establishments opening, closing, moving, or being sold. To determine which companies are likely undergoing changes, a model that best predicts organizational change was researched. The results of the modeling combined with payroll coverage will determine the certainty cutoff and selection criteria for non-certainty companies.

2. Company Characteristics

In order to better understand the data available at the time of COS sampling, we first created a file, referred to as the company characteristics file, compiling a wide variety of descriptive data. In order to be included on the file a company had to meet the same criteria necessary for inclusion in the COS. Namely, the file contains all active, multi-establishment companies with positive payroll in the year prior to COS sample selection. For modeling purposes, the 2010 COS was used as it was the most recent completed cycle of the survey at the time the research began. For testing the results of the modeling, company characteristics files were also created for the 2008 and 2009 COS samples, as well as for the 2011 COS once that survey cycle was completed.

Dozens of variables were obtained from the Business Register for each company including administrative data such as annual payroll, employment, tax filing requirements, and many other data items. Several data items were also calculated from this administrative data including the number of active establishments within each company, the number of closures as identified by tax filing requirements, and year-to-year change for employment, payroll, and receipts.

This research used COS response data to identify which sampled companies experienced some change in company organization. The variable CHG_STATUS was created from this response data and was later used for modeling the likelihood of a company undergoing change. CHG_STATUS is a binary variable set to 1 if any establishment within a company was new, closed, sold, or moved. Otherwise, CHG_STATUS was set to 0.

3. Modeling

3.1 The Models

Once all pertinent data was gathered, we could begin thinking about ways to use this data in order to assist in selecting cases for the COS sample. SAS Enterprise Miner was used to explore what information was available in our 2010 company characteristics data file. Several approaches were investigated including the use of neural networks and decision trees, but ultimately the decision was made to use logistic regression.

The goal in creating a logistic regression model was to predict for each company the likelihood of that company undergoing an organizational change, as defined by the variable CHG_STATUS. In this way, we believe we can more effectively identify companies for our targeted sample.

Recall that the targeted portion of the COS sample is not a probability sample. Instead, the companies below the certainty employment cutoff of 250 were specifically selected because they showed signs of organizational change. For the purposes of modeling the likelihood of organizational change, this means that we could not include these companies in our analysis and could only make use of those companies that were selected with certainty. Because of this, only companies with 250 to 1,000 employees were included in any modeling.

Three different regression models were tested, each with a different set of variables input into the model as potential independent variables. Certain variables were included in all three models, such as number of active establishments and filing status (which can be used as an indication of an establishment death). For the initial model employment, payroll, and receipts data for the survey year and the prior year were included. For Model 2, the percent difference and relative percent difference of employment, payroll, and receipts calculated using the survey year and prior year data were instead included in the model. Model 3 was identical to Model 2, with the addition of an indicator variable that identifies each company as small (1-4), medium (5-8), large (9-17) or very large (18+) based on its number of establishments.

To create each model, a step-wise logistic regression model was used with CHG_STATUS serving as the dependent variable. For all models the significance level for entry was set to 0.15, and the significance level to remain in the model was 0.05.

3.2 Modeling Results

Each of the three models was applied to COS data from 2008 and 2009, since 2010 COS data was used to create the models. When evaluating the quality of each of the models against one another, we looked at both the rate at which organizational change was successfully identified across the modeled probability of change (the posterior probability produced by the model) and cumulative and non-cumulative percent change captured charts[1] produced by Enterprise Miner.

When evaluating the rate at which organizational change was successfully identified, we broke the data down into groups based upon the posterior probability. We were looking for two things here. First, how well the predicted probability matched up with the actual success rate, and also the number of cases that were successfully identified as having experienced a change. The table below shows the results of this analysis for the three models for COS survey years 2008 and 2009.

[1] Note that in Enterprise Miner these charts are referred to as “response captured” charts, but here we will be looking at the amount of organizational change our model successfully captures not survey response.

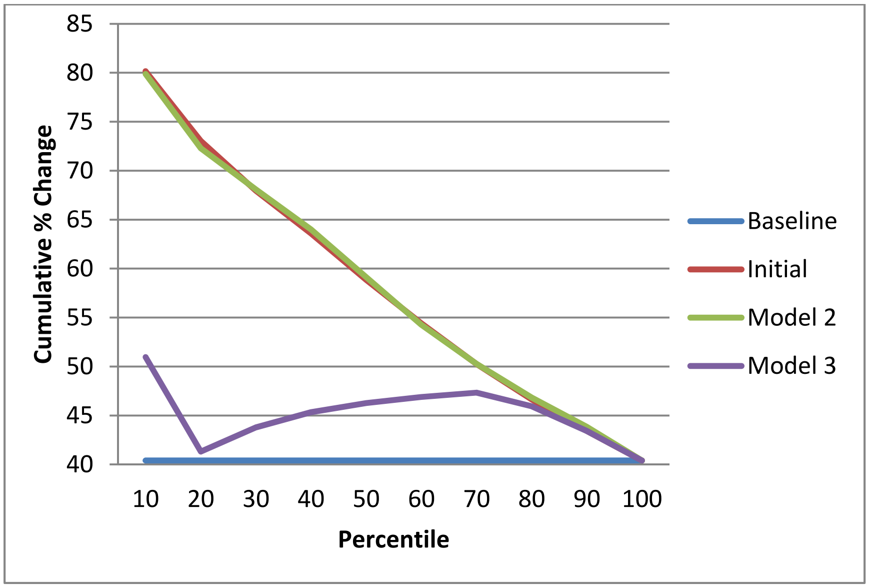
Table 1: Model Success Rate by Predicted Probability of Change

Posterior Probability	**COS 2008**			**COS 2009**		
	Initial Model	**Model 2**	**Model 3**	**Initial Model**	**Model 2**	**Model 3**
90 – 100%	100.00% (8/8)	84.85% (28/33)	100.00% (3/3)	100.00% (6/6)	81.25% (26/32)	50.00% (1/2)
80 – < 90%	84.62% (11/13)	83.67% (41/49)	90.00% (9/10)	66.67% (2/3)	80.65% (50/62)	75.00% (9/12)
70 – < 80%	68.75% (11/16)	76.54% (62/81)	75.00% (18/24)	91.67% (11/12)	72.63% (69/95)	66.67% (22/33)
60 – < 70%	79.49% (31/39)	55.88% (95/170)	69.07% (67/97)	85.71% (18/21)	66.46% (107/161)	65.35% (66/101)
50 – < 60%	62.04% (67/108)	60.13% (190/316)	55.36% (129/233)	77.05% (47/61)	62.88% (227/361)	61.48% (158/257)
40 – < 50%	56.02% (214/382)	53.11% (376/708)	50.46% (218/432)	71.43% (160/224)	59.25% (477/805)	60.44% (330/546)
30 – < 40%	28.04% (3323/11850)	39.73% (1506/3791)	38.45% (2301/5984)	31.16% (3826/12277)	47.73% (1475/3090)	44.15% (2467/5588)
20 – < 30%	30.67% (246/802)	20.87% (1576/7550)	30.38% (552/1817)	51.76% (309/597)	23.28% (1893/8131)	32.81% (650/1981)
10 – < 20%	N/A	7.12% (37/520)	13.30% (614/4618)	100.00% (1/1)	12.04% (56/465)	14.46% (677/4682)

Overall, the initial model had a higher percentage of companies with an organizational change with higher posterior probabilities, but Model 2 actually had more companies with an organizational change with higher posterior probabilities. For example in 2009, the initial model had 100.00% of companies with an organizational change with posterior probabilities greater than 90.00%, and Model 2 captured 81.25% of companies with an organizational change. The initial model had the greater percentage, but it only captured 6 companies compared to 26 captured organizational changes for Model 2. This usually occurred for companies with a posterior probability greater than 50.00%, where we would expect an organizational change. As a result, Model 2 was considered better for assigning higher posterior probabilities to more companies with an organizational change.

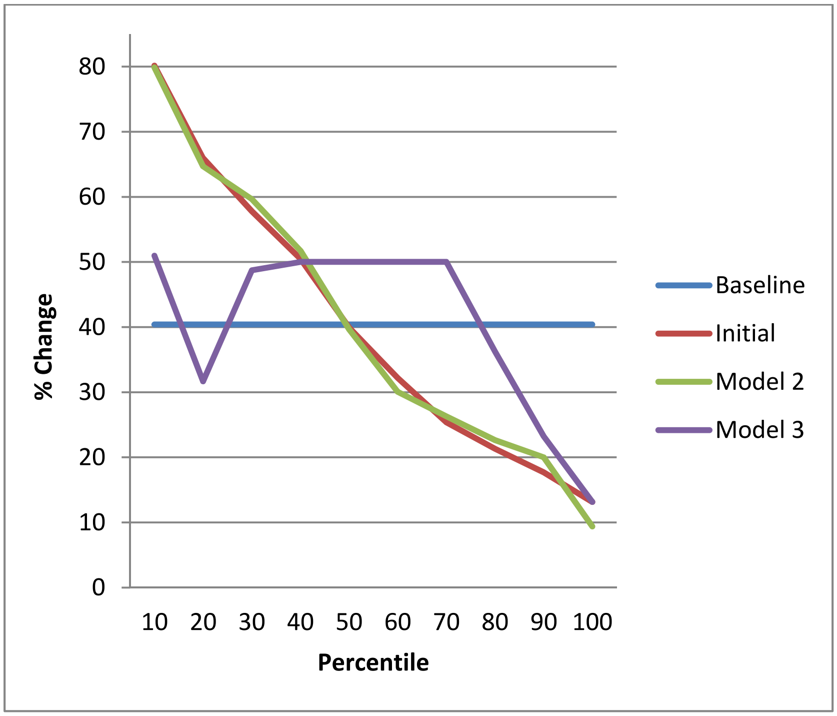
The three models were also compared graphically using cumulative and non-cumulative percent change captured charts. These charts were created using 2010 COS data from a validation data set that was set aside at the beginning of the modeling process. For these charts, as in the modeling process, companies with an organizational change are those companies for which CHG_STATUS = 1. The posterior probabilities produced by the models were sorted from largest to smallest and placed into ordered bins that contain about 10% of the data. For each bin, the response rate of those companies with an organizational change is calculated. For Chart 1, from decile to decile the response rates are summed cumulatively. A good model will have more respondents in the bins with high posterior probabilities.

Chart 1: Cumulative Change Captured – COS 2010



In the above chart, both the initial model and Model 2 are quite close to one another. Here the baseline represents the amount of companies with an organizational change that we would expect for a random sample (40.397%). The non-cumulative percent response is needed because the cumulative chart shown above hides the model's effectiveness at each decile. Chart 2 shows the proportion of companies with an organizational change at each decile (i.e., non-cumulative percent change).

Chart 2: Non-Cumulative Change Captured – COS 2010



Using the non-cumulative chart, the initial model and Model 2 are again quite similar. However, at the 100th percentile, the initial model had 13.14% of companies with an organizational change while Model 2 only has 9.39%. Model 3 does not perform very well and becomes flat between the 30th and 70th percentiles. It is clear from these charts that the decision comes down to the initial model and Model 2. These models behaved similarly in the cumulative and non-cumulative percent response charts. However, since the initial model has small parameter estimates and its odds ratio estimates are close to one for most of its variables, Model 2 was chosen as the final model to use for subsequent analyses.

4. Certainty Cutoff

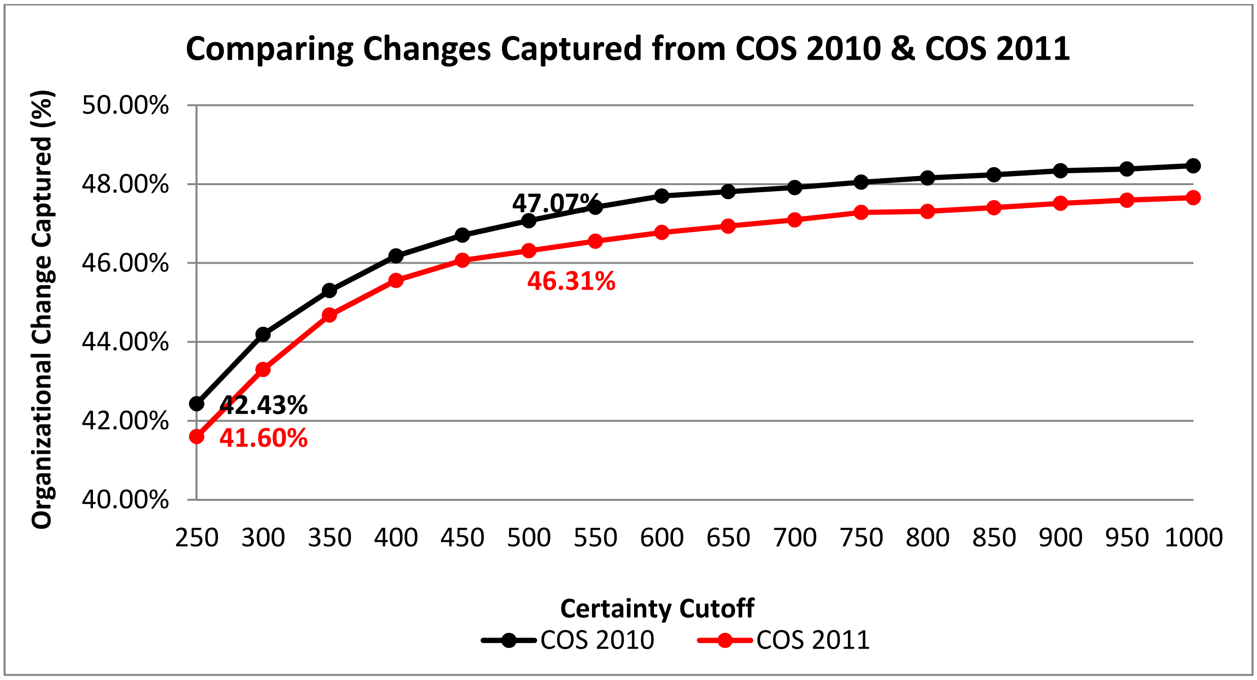
Now that Model 2 has been selected, we have addressed the identification of our targeted sample. Once certainty companies have been selected, all remaining companies will be processed through the regression model and have probabilities of organizational change assigned. These non-certainty companies will be sorted by this probability and the remaining COS sample will be composed of those companies with the highest probability of undergoing organizational change. In this way, we address the objective of capturing the highest percentage possible of companies opening or closing establishments.

The remaining question to be answered is this: “What is the appropriate employment cutoff to utilize for determining certainty companies?” The objective here is two-fold. On the one hand, we want to continue to maximize the percentage of companies in sample undergoing some organizational change. This would lead us to increase the certainty cutoff from the existing cutoff of 250 in order to have more sample available to target these companies specifically. However, recall that the second objective of this research is to ensure a high percentage of payroll coverage of companies in the survey. This objective argues for a lower certainty cutoff to ensure that large companies are included in the COS sample.

In order to balance both of these concerns, we looked primarily at the percent of change captured across various employment cutoff values but also at the percent of payroll captured once a likely new cutoff was identified. In intervals of 50 employees starting at the existing cutoff of 250 up to a possible cutoff of 1,000 employees, we calculated the estimated percent of change captured for survey years 2010 and 2011. For both of these survey years, we have complete COS results. Therefore, we know the change status of all companies that were in the sample with certainty as well as those companies that were contacted as part of the targeted sample. However, when applying our new targeting methodology to the 2010 and 2011 sample frames, there were cases selected for the targeted sample that were not previously selected under the old methodology and thus their change status is unknown. For these cases the change status was estimated by the posterior probability produced by the logistic regression model[2]. Chart 3 below, was created using the observed change status for those cases contacted in the 2010 and 2011 COS and the posteriors for those cases not in sample.

[2] Note that in reviewing the posterior probabilities it was noticed that the posterior probabilities for cases not previously targeted generally under-predicts organizational change. Therefore, using this probability in estimating the organizational change captured likely underestimates the percentage of change that will be captured under our new method.

Chart 3: Percent Change Captured at Various Certainty Cutoffs



Looking at this chart, we can see that there are substantial gains in the percent of change captured to be had by increasing the employment cutoff above its current value of 250. At about 500 employees, we begin to see diminishing returns for any further increase in the cutoff.

Based upon review of this chart, the new cutoff was tentatively set at 500 employees. However, the one remaining concern was what impact this would have on the payroll coverage of our sample as a whole.

The new design was simulated for the 2010 and 2011 COS and compared to the actual results of the COS for those years. By moving the certainty cutoff for large companies from 250 to 500 employees, the number of companies above the cutoff would be reduced from approximately 12,000 companies to 4,300 companies. In general, larger companies are more likely to be opening or closing establishments and carry a large amount of economic activity. As can be seen in Chart 3, the model improvement is more effective in identifying companies with less than 500 employees that are undergoing change than the current cutoff of 250 employees combined with the targeting methodology.

Table 2: Payroll Coverage – New Methodology vs. Old COS Methodology

Cutoff	**2010**		**2011**	
	New	**Old**	**New**	**Old**
250	89.54%	89.68%	90.37%	90.40%
500	88.18%		88.73%	

Table 2 shows that this increase in captured change due to the new cutoff is not accompanied by a large decrease in payroll coverage. In fact, the change in cutoff only reduces payroll coverage of COS-selected companies by approximately 1.5%. Therefore it was decided that using the new targeting methodology in conjunction with a new certainty cutoff of 500 employees would best meet the objectives of the COS sample redesign.

5. Ongoing and Future Work

Before the new targeting methodology and certainty cutoff can be put into production with 2013 COS, one withstanding issue needs to be resolved. As mentioned in the Section I.A., currently any multi-establishment company that is not in the Annual Survey of Manufacturers sample but has at least one manufacturing establishment and 50-249 employees is selected with certainty. A decision still needs to be made on how to revise this selection criteria. We are currently investigating whether to simply extend the employment criteria from 50-249 to 50-499 employees in keeping with the new certainty cutoff or to increase the lower bound of 50 employees in order to open up more companies to targeted selection.

We are also currently working on updating the existing sample selection programs to incorporate the new methodology for COS 2013. Once the 2013 COS is fielded and results are returned, we can evaluate the effectiveness of the changes made to the sampling methodology.