Using Paradata to Understand Business Survey Reporting Patterns

Eric B. Fink and Joanna Fane Lineback

U.S. Census Bureau[1], Washington, D.C. 20233

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract

Paradata are increasingly used as a source of information about respondent behavior and survey outcomes. In this paper, we examine paradata from the 2010 and 2011 Annual Survey of Manufactures (ASM), including cost, follow-up and multi-unit electronic reporting instrument data. Previous research was exploratory in nature, attempting to understand basic questions such as the median length of time from downloading to uploading the survey's software. Current research will be more targeted, whereby we use auxiliary data, including paradata, to model survey response. Additionally, we begin examining ASM costs as they relate to the stages of the survey process.

Keywords: Paradata, Adaptive Design, Annual Survey of Manufactures

1. Introduction

As survey nonresponse has increased over the past two decades (de Leeuw and de Heer, 2002) and survey budgets have declined, it has become imperative to find cost savings wherever possible. One possible intervention to mitigate these trends is the implementation of an effective adaptive survey design program. Adaptive design attempts to utilize already available survey information to make decisions during data collection that balance survey costs and quality. One such source of information comes in the form of the data collected about the survey process, referred to as paradata (Couper, 1998). For instance, Rao, Glickman, and Glynn (2008) proposed data collection stopping rules that depended on comparing estimates derived from early and late responders. It is not the purpose of this paper to review the adaptive survey design approach (see Groves and Heeringa, (2006) for a better exposition on the topic). The first step in using paradata is to identify a set of variables that may provide some utility in understanding the nature of reporting throughout the data collection period.

As such, this paper will be exploratory in nature, developing a profile of respondents based off paradata variables captured in the 2011 Annual Survey of Manufactures (ASM). These data come from two sources: the Business Register (BR) and Surveyor. Surveyor is a downloadable reporting software platform used for multi-unit businesses. The BR is a centralized business database where information for enterprises, establishments, and other administrative data are stored. Additionally, we have cost data that includes costs associated with the initial survey mail-out operation, as well as mail and telephone follow-up.

For this paper, we define an establishment as a single physical location where business is conducted or where services or industrial operations are performed. Further, the terms "establishment" and "unit" are used interchangeably in this paper. Finally, we define an enterprise as a business organization consisting of one or more domestic establishments that were specified under common ownership or control. The enterprise and the establishment are the same for single-unit organizations. Each multi-unit company forms one enterprise.

In this paper, we will discuss the following: Section 2 gives relevant background information on the ASM; Section 3 will describe the methods we used to analyze the data; Section 4 will present results. Finally, Section 5 discusses the results and future research directions.

2. Background

The ASM is a mandatory response survey that provides statistics on employment, payroll, supplemental labor costs, cost of materials consumed, operating expenses, value of shipments, value added by manufacturing, detailed capital expenditures, fuels and electric energy used, and inventories for all manufacturing establishments with one or more paid employees. In this section, we provide information on the major components of the ASM program, including

[1] Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

sample design, data collection, estimation, and non-response follow-up. For information on the ASM including historical data and forms, go to http://www.census.gov/manufacturing/asm/index.html.

2.1 Sample Design

To select the ASM sample, the manufacturing population is partitioned into two groups: establishments eligible to be mailed a questionnaire, a mail stratum, and establishments not eligible to be mailed a questionnaire, a nonmail stratum. The eligible establishments consist of larger single-location, manufacturing companies and all manufacturing establishments of multi-location companies. The ineligible establishments consist of small and medium-sized, single-establishment companies based on the Economic Census. Data for these ineligible establishments are estimated using information obtained from the administrative records of the Internal Revenue Service (IRS) and Social Security Administration, and are included in the published ASM estimates.

The ASM mail sample includes approximately 50,000 establishments of which about 20,000 are selected with certainty, and about 30,000 are selected with probability proportional to a composite measure of establishment size. Although the nonmail stratum contained approximately 180,000 individual establishments in 2011, it accounted for less than 7 percent of the estimate for total value of shipments at the total manufacturing level. A new sample is selected at five-year intervals beginning the second survey year subsequent to the Economic Census. This information is supplemented with data for new companies from the IRS and the Census Bureau's Report of Organization Survey (COS).

2.2 Data Collection

Data are collected annually for the ASM except for years ending in 2 and 7 when the Economic Census is conducted. The survey is establishment-based, although for a multi-establishment business the questionnaires are mailed to the business enterprise unless another reporting arrangement has been made. Respondents can choose to report by mail or electronically using either the Census Surveyor software (for multi-unit organizations) or by the Web (for single-unit organizations). In addition, respondents can fax forms and in some cases give their responses by phone. In 2011, every enterprise in the sample received a paper form[2]. All multi-units that receive a request to complete the ASM also get a request to complete the COS in the same package. Responses are due within 30 days of receiving the form. The COS is an annual mail-out/mail-back survey of selected companies with payroll, excluding companies engaged exclusively in agricultural production. The purpose of the COS is to obtain current organization and operating information on multi-establishment firms in order to maintain the BR.

Follow-up with nonresponding businesses begins approximately two months after the initial mailout and is usually in the form of a mailed letter. After the first reminder, there are three additional reminders sent, once a month, until a case is considered a delinquent nonrespondent. For some very large establishments that are deemed important for estimation purposes, follow-up may occur via telephone. Currently, data collection continues for the ASM until the project runs out of time or money.

2.3 Estimation

Most of the ASM estimates derived for the mail stratum are computed using a difference estimator. The difference estimator takes advantage of the fact that, for manufacturing establishments, there is a strong correlation among some estimates between the current year data values and the previous Economic Census values. Because of this correlation, difference estimates are considered more reliable than comparable estimates developed from the current sample data alone. The ASM difference estimates are computed at the establishment level by adding the weighted difference (between the current data and the Economic Census data) to the Economic Census data. However, some estimates are not generated using the difference estimator because the year-to-year correlations are considerably weaker. A standard linear estimator is used for these variables. Estimates are published from the 2 – 6 digit NAICS level, and for the U.S. and by state.

3. Analysis

3.1 Analysis Variables

In analyzing the 2011 ASM Surveyor paradata, we have included additional 2010 and 2011 ASM data obtained from the BR and cost data provided by the program staff. The ASM Surveyor paradata file included the date respondents downloaded the Surveyor software and the date the respondents uploaded the Surveyor data file. From

[2] For the 2012 Economic Census if 2011 ASM responses were electronic, paper forms will not be sent.

the BR we obtained information about participation in other Census Bureau surveys, check-in dates, and the mode the respondent supplied information to the Census Bureau for the 2011 ASM. Finally, we have also obtained data indicating costs associated with initial mailing, as well as follow-up mailings and telephone costs.

3.2 Analysis Questions

Much of the research we have conducted to this point was exploratory in nature. We spent months obtaining and merging the aforementioned data and many of the initial research questions necessitate only descriptive statistics to answer. Our overarching research question is how to use paradata to plan survey contact and nonresponse follow-up strategies. We further refined our question into several more manageable parts about business-respondent behavior. We develop the following initial questions:

1. What is the time between mailout, downloading Surveyor software, and uploading data?
2. What is the cumulative response rate?
3. What is the cumulative total quantity response rate?
4. How much money are we spending on each stage of data collection relative to the achieved response rate?
5. What changes in reporting trends do we notice since the previous survey cycle?
6. What are the characteristics of early versus late responders?
7. Are there strong predictors for switching from paper to electronic reporting? From electronic to paper?

3.3 Limitations of the Analysis

There are some limitations to our analysis. Because all establishments in sample report using their parent company ID, it is not possible to distinguish between multiple establishments downloading or uploading the Surveyor software under a company, and a single individual establishment downloading or uploading the software multiple times on the paradata file. Thus, our analysis is restricted to only the initial download/upload event.

Additionally, to measure response, we calculated a check-in rate, rather than a traditional response rate, as we currently do not have the flags to indicate if a respondent had supplied sufficient information to be deemed a response. The use of a check-in rate is appropriate when used as a measure of data collection performance. It is not, however, to be interpreted as a quality indicator.

As noted above, with multi-unit businesses reporting via Surveyor and single-unit establishments reporting via Web, the analysis below in Section 4.1 is restricted only to multi-unit organizations, as we did not have access to Web-response paradata.

Finally, there are limitations with respect to the cost data presented in Section 4.2. The costs here only reflect mail form and phone call costs (direct labor, overhead, and outgoing calls). At this point, we are unable to reasonably estimate cost by survey or by survey activity such as form design, sample selection, or data processing. Additionally, it is not always possible to separate ASM and COS costs because they are conducted jointly. However, as ASM is a much more involved survey instrument in that it asks much more than does COS, a reasonable simplifying assumption for this paper is that where we are given costs for both ASM and COS, a vast majority of the resources are being utilized for ASM.

4. Results

Results are given below. Subsection 4.1 gives results on the 2,014 companies that responded using the Surveyor software package. Both subsections 4.2 and 4.3 present results on all mailed ASM establishments in NAICS 31, the manufacturing sector. Subsection 4.2 presents results on survey response and costs, and subsection 4.3 presents results on those establishments that changed their method of responding to the survey.

4.1 Surveyor Results

All 33,718 multi-unit establishments had the option of reporting via Surveyor, and approximately 13,150 actually used the Surveyor package to report their 2011 data. Of the 13,150 Surveyor records, 2,014 companies had both an initial download date and an initial upload date. All analyses in this section are restricted to only those companies.

The range in days from when the ASM forms were mailed to recipients to when recipients initially downloaded the Surveyor software package was from -5 to 281 days[3]. A negative value is reasonable,

[3] For initial mailout and the subsequent follow-up waves of mailings, we have a range of three consecutive days for each event. As such, we arbitrarily chose the second of these three days to represent the event we are analyzing.

because of some proactive establishments that are expecting the survey. There were several peak download times at roughly 45 days, 75 days, and 185 days (see Figure 1). As responses were due 30 days after receipt, and follow-ups began 30 days after the due date, and then approximately once a month for three months afterwards for nonresponsive cases, these peaks roughly correspond to those periods.

Most ASM uploads occurred within the first few days of download (see Figure 2). In fact, 81% of initial uploads occurred on the same day the Surveyor package was downloaded. There is evidence of severe decay in download/upload interval. Referring to Figure 3, this decay is still seen among those respondents taking more than ten days to upload their data, but the pattern is not as extreme.

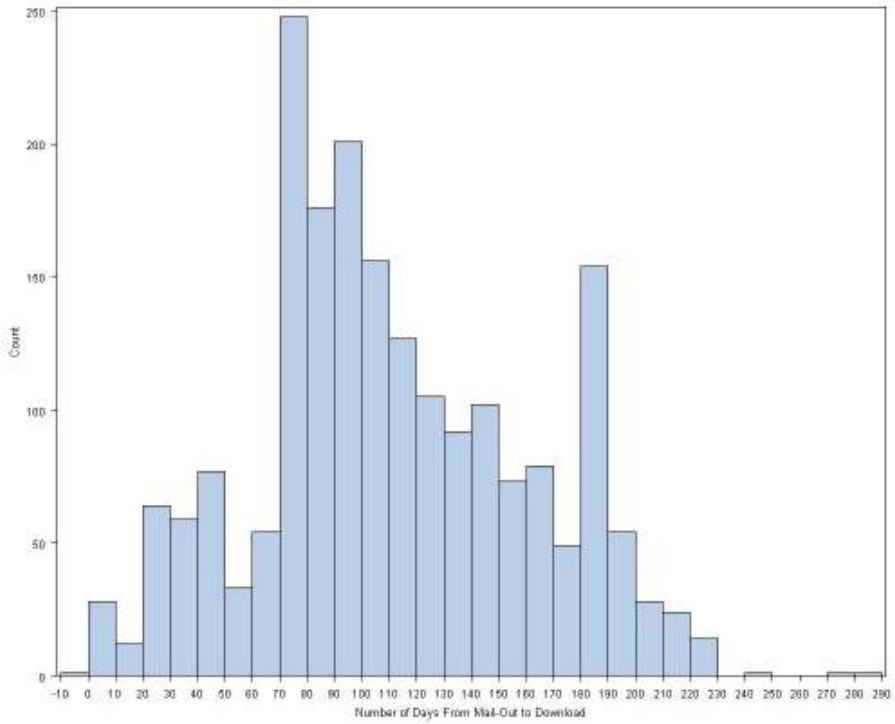

Figure 1. Duration from when the form was mailed to initially downloading the Surveyor package, in 10-day intervals.
Source: 2011 ASM Surveyor Paradata.

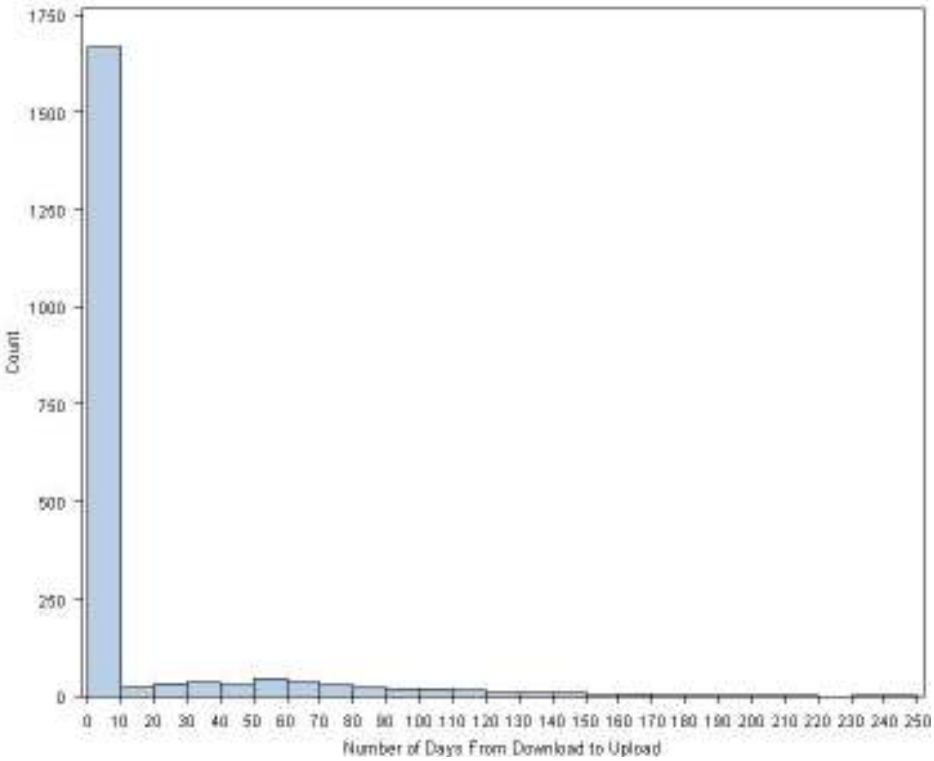

Figure 2. Time elapsed between the initial download to upload of the Surveyor software, in 10-day intervals.
Source: 2011 ASM Surveyor Paradata.

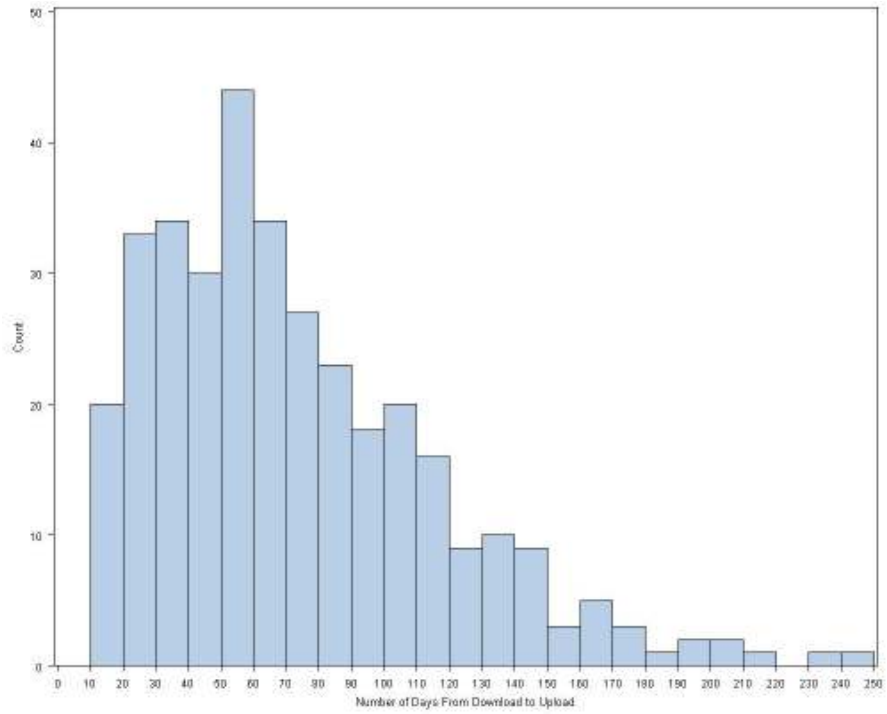

Figure 3. Time elapsed between the initial download to upload of the Surveyor software for those respondents taking longer than 10 days to upload their data, reported in 10-day intervals.
Source: 2011 ASM Surveyor Paradata.

4.2 Response Metrics and Costs

The check-in rate is the percentage of forms returned, either by paper or electronically, to those mailed. The check-in rate covers all mailed multi-unit and single-unit establishments. Again, the check-in rate serves as a measure of data collection performance. The curve in Figure 4 shows a gradual, constant increase in the check-in rate, achieving an overall rate just under 80%. Figure 5 reveals an initial spike in the number of forms checked-in occurred shortly after the due date. Another spike occurred, in the 70-80 day interval, after the first reminder was sent.

Figure 4 also shows the cumulative percentage of the mailing budget from the initial mailing through the fourth follow-up. The first percentage listed is so large because it includes the cost of printing the forms, as well as the cost of postage for the mailing plus the cost of postage on the envelope for return, in addition to early incoming and outgoing phone calls. With the monotonic increase in the check-in rate, it is difficult to assess the utility of spending on follow-up mailings. Some follow-ups have a larger percentage increase in associated costs than others, but do not appear to yield any appreciable increase in the number of check-ins. There is the possibility the follow-ups help maintain the observed monotonic increase in check-ins, but only a carefully planned experiment can help answer that question.

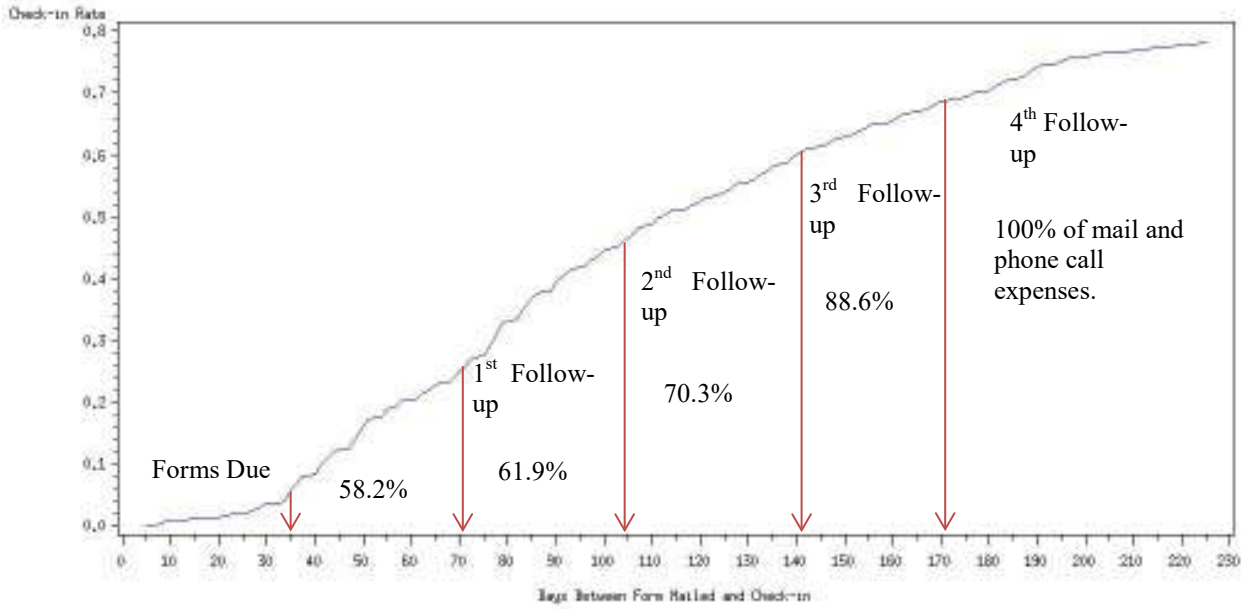

Figure 4. The check-in rate for the 2011 ASM from when forms were initially mailed to respondents. The red arrows represent mail-out dates for follow-up letters at 71, 104, 141, and 174 days for 1st, 2nd, 3rd, and 4th follow-up respectively. The percentages listed show the cumulative percent of the total mailing and telephone expenditures allocated to each stage of data collection up to, but not including, the follow-up subsequently listed.
Source: 2011 ASM.