Identifying and Comparing Characteristics of Nonrespondents throughout the Data Collection Process

Morgan Earp[1], Daniell Toth[1], Polly Phipps[1], & Charlotte Oslund[1]

[1]Bureau of Labor Statistics, PSB Suite 1950, 2 Massachusetts Avenue NE, Washington, DC 20212

Abstract

Establishment surveys usually go through a process of verifying address and contact information, as well as attempting to gain participation before data collection begins. At each phase in the data collection process, there is potential for nonresponse. Characteristics of establishments that are difficult to locate and contact may not be the same as those that refuse to participate or respond to the survey; therefore, it is important that we assess nonresponse at each phase of data collection. The Bureau of Labor Statistics Job Openings and Labor Turnover Survey (JOLTS) uses a three-fold data collection process: address refinement to verify address and contact information, enrollment to recruit participants into the survey, and data collection when the survey is administered. Using auxiliary data related to key JOLTS estimates, we identify and compare characteristics associated with nonresponse at each of the three phases of nonresponse, as well as both unit and item nonresponse during the data collection phase. The results of this study can be used to better allocate resources when attempting to reduce nonresponse.

Key Words: Nonresponse; Regression Trees; Establishment Surveys

1. Introduction

Throughout the survey data collection process, there are several opportunities for nonresponse. Most establishment surveys go through a process of validating addresses and contact information and gaining participation before data collection even begins. Very little work to date has focused on identifying and comparing the characteristics of establishment nonrespondents prior to data collection. At the time of data collection, respondents may choose to decline survey participation, or participate selectively by responding only to certain data items and not others. This paper focuses on understanding how characteristics of nonrespondents shift throughout the data collection process.

The Bureau of Labor Statistics' (BLS) Job Openings and Labor Turnover Survey (JOLTS) is a panel survey with three documented phases of data collection: address refinement, enrollment, and data collection. Nonresponse can happen during any one of these phases, suggesting several potential questions. Are some establishments more likely to be nonrespondents during one phase and not at others? Which establishments are least likely to have their address verified or to be successfully enrolled in the survey? Are characteristics of nonrespondents different during actual data collection? For example, what if private versus public sector ownership is more important during address refinement and enrollment, while the type of industry is of greater significance during

data collection? If the characteristics do indeed shift, it may be important to model each phase separately to better manage the risk of nonresponse at each phase.

JOLTS attempts to maintain high response rates at each phase of data collection since low response rates carry the threat of nonresponse bias, loss of stakeholder confidence, and the potential to inflate variance in survey estimates. Maintaining high response rates requires substantial effort and resources. Traditionally, survey methodologists use several approaches for dealing with nonresponse, such as increasing participation through incentives, notification letters, or providing alternative data collection modes (Dillman, 1978, Dillman, Smyth, and Christian 2009; Groves et al., 2002), or after data collection using adjustment (Kalton and Flores-Cervantes 2003) or imputation. This paper attempts to determine which establishments are least likely to respond during each phase of data collection so that BLS can make the best use of resources throughout the data collection process. Using regression trees, we identify subgroups of establishments least likely to respond at each phase. The results of our regression tree models can be used to develop strategies for increasing participation, including adaptive design and weighting methods (Phipps and Toth 2012).

2. Methodology

JOLTS collects data every month from establishments to provide national estimates of job openings, hires, and total separations in the United States. JOLTS samples approximately 16,000 establishments per month from all 50 states and includes both the government and private sectors. The JOLTS sample is stratified by ownership (private or public), region, North American Industry Classification System (NAICS) industry sector, and employment size class. Once selected, establishments remain in the survey for 24 months until their panel is rotated out. Panels are rotated in and out every month.

Our study sample consisted of 16,598 establishments sampled for JOLTS during July of 2012. We excluded establishments that were out of business (n = 51), post offices (n = 682), and those that had not yet been contacted (n = 71). Post offices were excluded as the postal service provides data to JOLTS as a census by state. We excluded a small number of establishments with no record of any contact or collection attempt since we are interested in classifying establishments that do not respond given the opportunity. After removing these records, our final dataset used for analysis consisted of 15,794 establishments.

JOLTS data collection takes place in three phases: 1) address refinement, 2) enrollment, and 3) data collection (see Figure 1). At each of the three phases, there is potential for nonresponse. For example, for the July 2012 sample panel, 1.5 percent of establishments did not make it through the address refinement phase, while 9.1 percent of establishments that made it through the address refinement phase, but did not agree to participate in the survey, resulting in "enrollment nonresponse."

During address refinement, BLS locates and verifies the contact information of sampled establishments by telephone. Establishment contact information is provided by each State and is included as part of the sample frame. By the time the frame is used to draw the sample for JOLTS this contact information is at least 12 months old. Most sampled establishments have some known contact information, but there are a few with little or no

contact information available. Even in the case where contact information is provided, the quality and extent varies. A street address is provided for most establishments and in some cases a telephone number, but for the majority there is not a contact name. Even when contact information is available, it may be out of date, given the 12-month lag time. If the contact information for an establishment cannot be verified by the BLS these establishments are considered here as "address nonrespondents" (BLS, 2013a).

Once an establishment's address is verified, its data collection status is updated to "address refinement complete," and it is moved to enrollment. The goal of the enrollment phase is to gain compliance from the potential respondent to participate in the JOLTS program, which involves providing monthly employment and turnover data. During the enrollment phase, each establishment is mailed an "introductory packet" explaining the survey and the importance of their participation; these packets include a customized cover letter, JOLTS Brochure, Business Information Guide, Fact Sheet explaining how the data are used, and JOLTS Survey Form. About three to five days after the introductory packet is mailed out, interviewers follow-up by calling the establishment to solicit participation (BLS, 2013b).

After an establishment is successfully enrolled in the survey, the interviewer schedules an appointment and moves the unit into the data collection phase, at which point, the interviewer attempts to collect the requested data. For the first five months, most establishments complete the survey via computer-assisted telephone interviewing. After that time, an establishment may be transitioned to other data collection modes like Web, Email, or fax. Offering a variety of collection methods helps accommodate respondent preferences, which is important since JOLTS is a voluntary survey program. (BLS, 2012)

Response status at each of the phases of data collection is constructed using data collection status codes, which are available in the JOLTS survey management system. These codes specify the last phase at which the establishment was contacted. If an establishment is counted as a nonrespondent in an earlier phase, we have excluded them from the analysis in subsequent phases. In order to be counted as a respondent in the data collection phase, the establishment must have provided data that was used for the survey estimates. Establishments that did not report one or more items (total separations, hires, and/or job openings) are considered item nonrespondents.

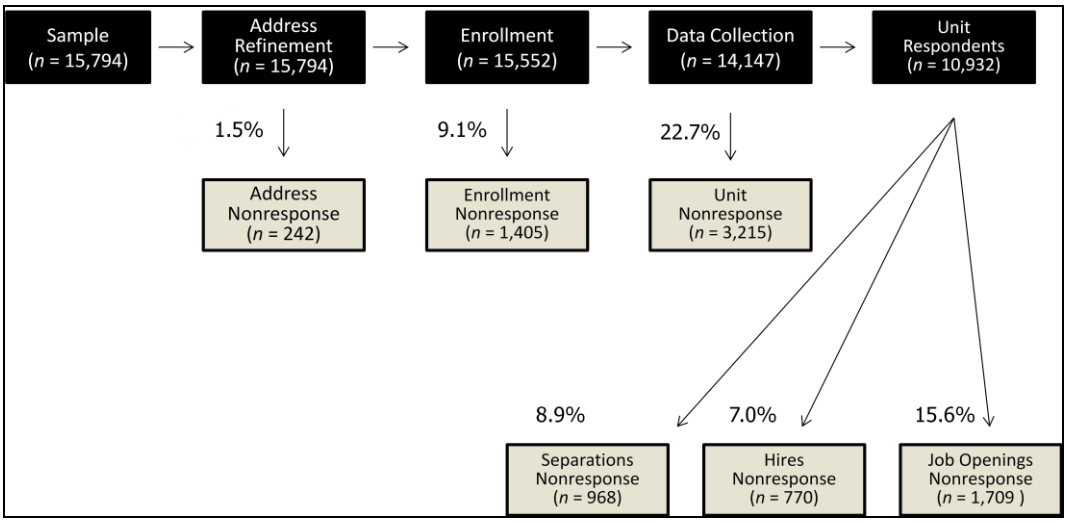

Figure 1: Jolts Data Collection Phases

The most common way to model response propensity is through logistic regression. Logistic regression requires the analyst to hypothesize variables thought to be associated with nonresponse and then uses the observed data to fit the model parameters. The predicted response propensities obtained from the model are then used to form groups. Usually groups are formed using quantiles of the predicted response rates (Eltinge, J. and Yansaneh 1997). Both the logistic regression models and the groups formed from their predicted propensity are often difficult to interpret because of interactions between the characteristic variables. In contrast, regression trees are designed to give interpretable models using interaction effects. By recursively splitting the establishments into two groups based on their characteristics and propensity to respond, the resulting tree model provides a partitioning of the data that is easy to interpret based on the characteristic variables.

Our goal is to identify interpretable classes based on establishment characteristics that help identify likely nonrespondents. Using regression trees, we identify characteristics of nonrespondents at each phase of data collection. A regression tree model is constructed by recursively splitting the data based on characteristic variables and response propensity. Recursive partitioning is applied on each group of sampled establishments until a minimum threshold of sample size is reached. At each iteration the variable and breakpoint are chosen to maximize the heterogeneity across subgroups and the homogeneity within groups with regard to nonresponse.

The regression tree models were built using the CRT method in SPSS. Final subgroups were required to have at least 100 cases; also the depth of trees was limited to three to keep explanations simple and predictions stable. After reviewing the initial results, the trees were simplified (pruned) to provide clean and easily interpretable results.

Separate trees were built to identify characteristics associated with each phase of nonresponse: one tree to model address nonresponse, another tree to model enrollment nonresponse, and four separate trees to model data collection nonresponse – one for unit nonresponse and three for item nonresponse (total separations, hires, and job openings).

Nonresponse was modeled at each phase of data collection using auxiliary data from the BLS Quarterly Census of Employment and Wages sample frame, including: average employment size of establishment in 2011; industry super sectors, (1) mining and logging, 2) construction, 3) manufacturing, 4) trade/transportation/utilities, 5) information, 6) financial activities, 7) professional and business services, 8) education and health services, 9) leisure and hospitality, 10) other services, and 11) government; white collar services (including industry super sectors 5, 6, and 7 above); type of ownership (public versus private); population size of metropolitan area; region; and whether the establishment was part of a state multi-establishment firm.

3. Results

As discussed earlier, some sample members are not located, contacted, or verified at the first phase of possible contact—address refinement. At first glance it seems that nonresponse is small at this phase, 1.5 percent, and therefore may not be of much concern. However, the tree model, as shown in Figure 2, identifies two groups for which this type of nonresponse is a concern. The first is federal government – approximately 12.1 percent of federal government establishments are nonrespondents at this phase. Second, higher nonresponse is observed in large establishments (>182 employees) in the trade, transportation, and utilities industries, with address refinement nonresponse rate of 14.8 percent.

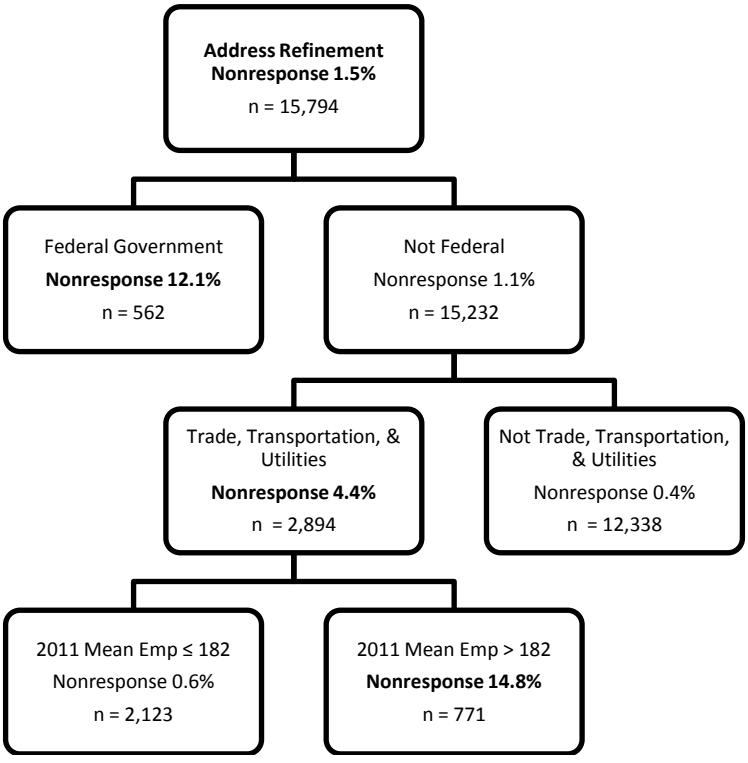

Figure 2: Address Refinement Nonresponse Tree Model

Enrollment is the next phase of the data collection, where interviewers contact sample members and solicit their participation in the survey. Nonresponse at enrollment (9%) is higher than at the time of address refinement, as seen in Figure 3. Unlike the address

refinement phase, nonrespondents in the enrollment phase tend to be privately owned: nonresponse is five percentage points higher in private as opposed to publicly-owned establishments. In addition, employment size plays a role in nonresponse for privately-owned establishments, as those with greater than 75 employees on average during 2011 have a 14.8 percent nonresponse rate compared to 6.4 percent for those with 75 or fewer employees.

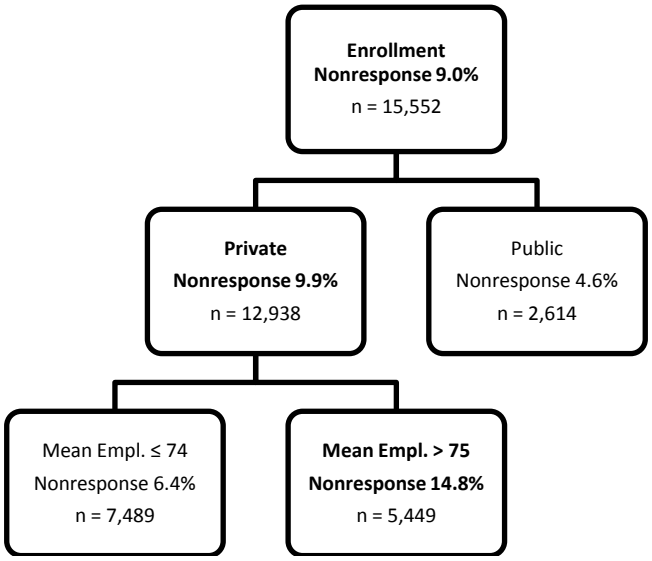

Figure 3: Enrollment Nonresponse Tree Model

Once an establishment has agreed to participate in the survey, they are contacted each month. In July 2012, approximately 22.7 percent of those in the data collection phase did not respond to the survey request (Figure 4). The data collection tree models are not driven by whether the establishment is private or publicly owned, but instead by the type of industry and employment size. The first split in the tree model is between what we define as white-collar service sectors (professional and business services, information, and financial activities super sectors) and non-white collar industries (all other industry super sectors), with subsequent splits on employment size. Establishments in white-collar services with greater than 180 employees on average in 2011 have a 40.8 percent nonresponse rate compared to 20.2 percent for those with 180 employees or less. Establishments in non-white collar industries with greater than 23 employees have a 25.4 percent nonresponse rate compared to 14.9 percent for those with 23 or fewer employees.

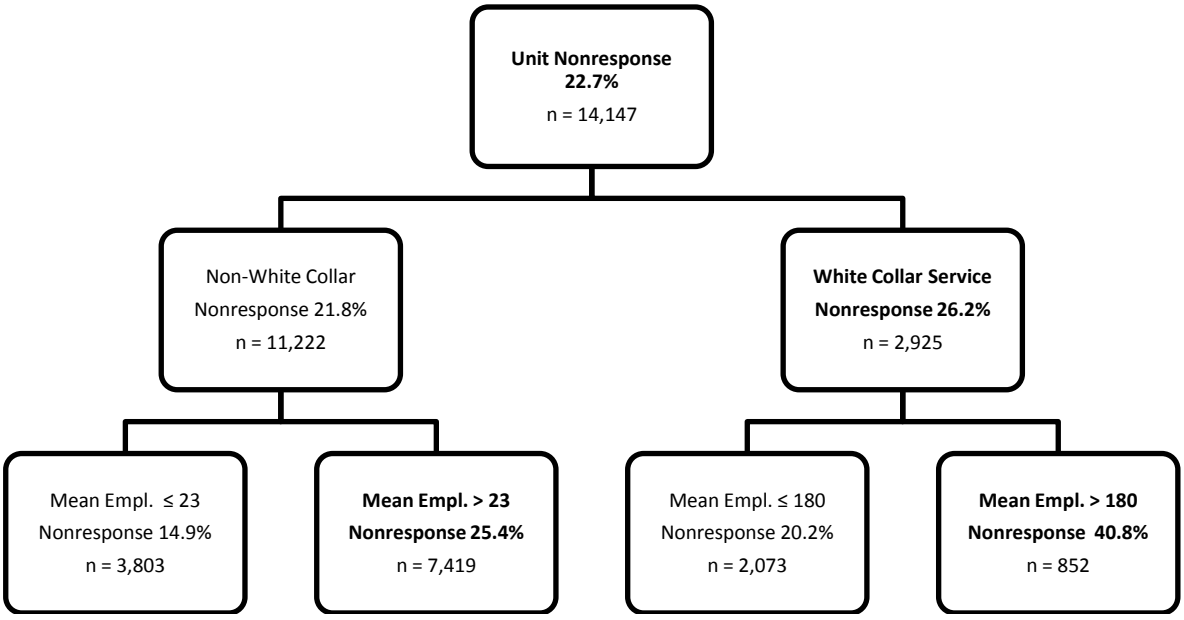

Figure 4: Unit Nonresponse Tree Model

Item nonrespondents during the data collection phase are modeled using three key JOLTS data items, including responses to total separations, hires, and job openings. Item nonresponse is modeled separately for each item; we do not assess combinations of item missingness. Overall, total separations have an item nonresponse rate of almost nine percent (Figure 5) versus seven percent for hires (Figure 6). Similar to unit nonresponse, the first tree split for both total separations and hires is the white-collar service sectors. For both total separations and hires, establishments in white-collar service sectors have about a three percentage point higher nonresponse rate than those in non-white collar sectors (Figure 5 & 6). Non-white collar establishments with greater than 74 employees on average in 2011 have an eight to nine percentage point higher item nonresponse rate for total separations and hires compared to those with 74 or fewer employees (Figures 5 & 6). Employment size plays an even larger role in the white-collar services. Establishments with greater than 72 employees have a 17 percentage point higher item nonresponse rate for total separations than those with 72 or fewer employees, with an overall item nonresponse rate of 23.1 percent (Figure 5). White-collar service sector establishments with greater than 45 employees have a 13 percentage point higher item nonresponse for hires than those with 45 employees or less, with an overall item nonresponse for hires at 17.2 percent (Figure 6).

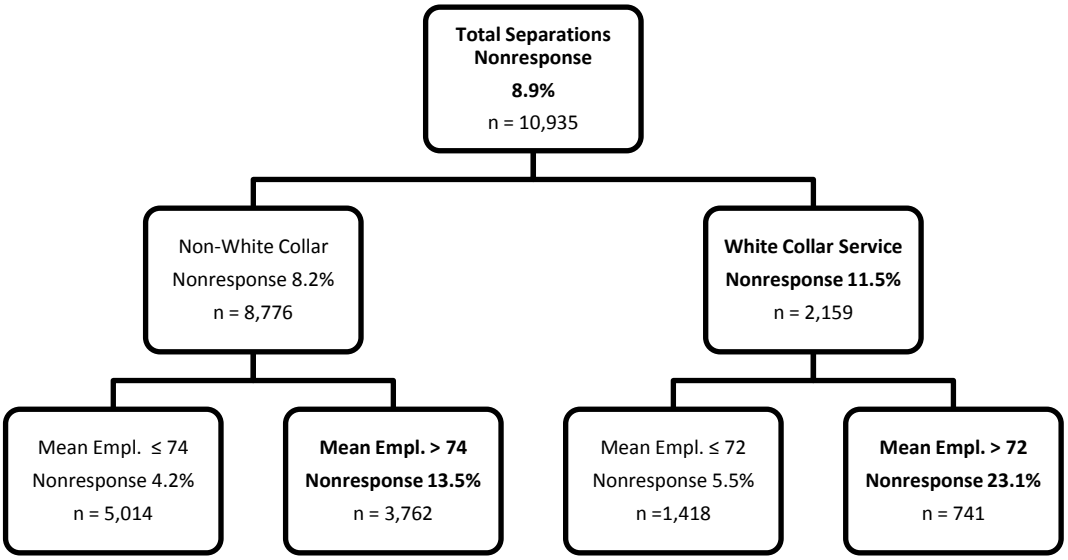

Figure 5: Total Separations Nonresponse Tree Model

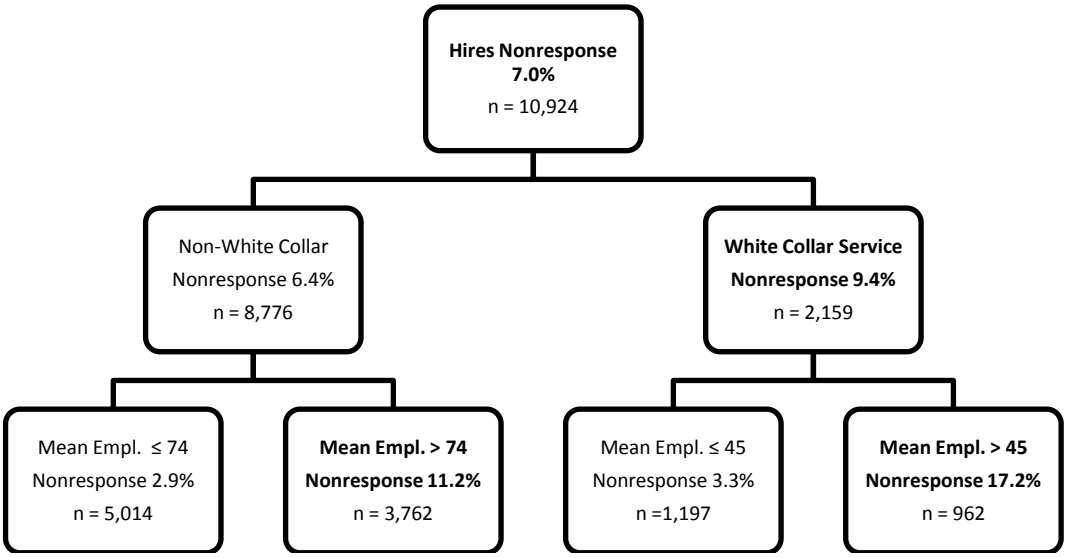

Figure 6: Hires Nonresponse Tree Model

The third data item is job openings, which has the highest rate of item nonresponse at 15.6 percent (Figure 7). Unlike total separations and hires, the first distinguishing split for job openings is whether an establishment is part of a multi-establishment firm. These establishments have a 23.2 percent nonresponse rate for job openings compared to 10.7 percent for single establishments (Figure 7). Size again, is important in determining nonresponse. Single establishments with greater than 79 employees on average in 2011 have a job openings nonresponse rate that is 15 percentage points higher than those with 79 or fewer employees. Establishments that are part of a multi-establishment firm with greater than 72 employees have a nonresponse rate almost 20 percentage points higher than those with 72 or fewer employees, with an overall item nonresponse of 32.2 percent for job openings.

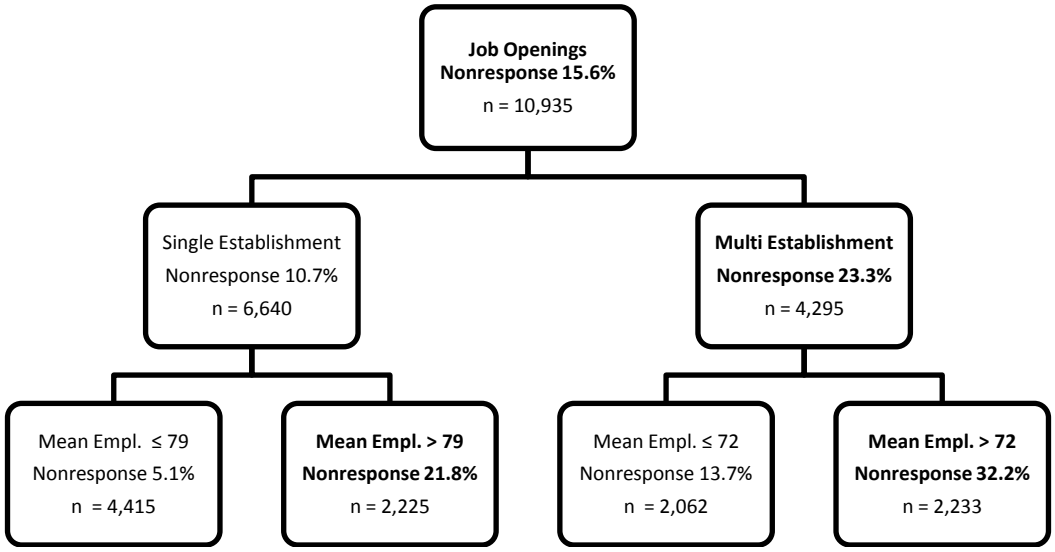

Figure 7: Job Openings Nonresponse Tree Model

4. Discussion

This study compares the characteristics of nonresponding establishments across the various phases of data collection – both before and during data collection, and for the latter, unit as well as item nonresponse. At all phases, we find that groups with higher average employment size for the previous year have higher rates of nonresponse. Federal government establishments have the highest rates of nonresponse during the address refinement phase, while privately owned establishments have the highest rates during the enrollment phase. Establishments in white-collar service sectors have the highest nonresponse rates during the data collection phase for both unit nonresponse and total separations and hires item nonresponse. Lastly, establishments that are part of a multi-establishment firm have the highest nonresponse rates for job openings. Our findings on higher nonresponse rates for larger employment size, white-collar services, and multi-establishment firms at the data collection phase are similar to those observed in another BLS survey, the Occupational Employment Statistics survey (Phipps and Toth, 2012).

By looking at each phase of nonresponse separately, we can see that the characteristics of nonrespondents vary at each phase, which helps us to better understand when and for whom nonresponse is an issue. For example, we now know that directing efforts toward federal government establishments during data collection would not be nearly as effective as doing so during address refinement. Also, targeting white-collar service sector establishments during data collection is a potential strategy, since there is less difficulty locating and verifying their addresses and contact information compared to getting a response after survey enrollment.

In future work, we would like to explore the relationship between employment size and nonresponse. Size appears to be a significant variable in all of the nonresponse models; however, it is unclear whether this is a linear relationship, or if the risk of nonresponse potentially goes back down for the largest establishments.

We plan to use paradata on the establishment contact and interviewer to further explore and characterize data collection nonresponse. This will allow us to determine if there are establishment or interviewer characteristics that are associated with successful data collection. Also, we are interested in utilizing focus groups with interviewers to understand why certain subgroups are more prone to nonresponse during various phases of the data collection process. Focus groups could provide us with the insight needed to potentially remedy or at least reduce nonresponse at each phase of data collection, specifically for problematic groups.

In this paper we focus on a single period of data collection. We are interested in expanding our study of nonresponse to include longitudinal aspects, to potentially determine if there are patterns as to when establishments become nonrespondents in this 24 month panel survey.

5. References

Bureau of Labor Statistics. (2012). *Job Openings and Labor Turnover Survey Data Collection Training Manual.*

Bureau of Labor Statistics. (2013a). *Job Openings and Labor Turnover Survey Address Refinement Training Manual*.

Bureau of Labor Statistics. (2013b). *Job Openings and Labor Turnover Survey Enrollment Training Manual.*

Dillman, D. 1978. *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley & Sons.

Dillman, D, J. Smyth, and L. Christian. 2009. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* New Jersey: John Wiley & Sons.

Eltinge, J. and Yansaneh, I. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology* 23**,** 33–40.

Groves, R.M., D. Dillman, J.L. Eltinge, and R. Little (Eds.). 2002. Survey Nonresponse. New York: Wiley.

Kalton, G. & Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics,* 19 (2), 81-97.

Phipps, P. & Toth, D. (2012). Analyzing Establishment Nonresponse Using and Interpretable Regression Tree Model with Linked Administrative Data. *Annals of Applied Statistics, 6* (2), 772-794.