Assessing the Usefulness of Census Bureau Multi-Establishment Data to Facilitate Linking Firms with Establishments in BLS Microdata October 2014

Elizabeth Weber Handwerker[1], Lowell G. Mason[2]
[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Room 4945, Washington, DC 20212; Handwerker.Elizabeth@bls.gov.
[2]U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Room 4945, Washington, DC 20212; Mason.Lowell@bls.gov.

Abstract

For researchers interested in linking firm-level datasets with establishment-level Bureau of Labor Statistics (BLS) microdata, the fundamental challenge is to find all the EINs that these firms use in their reports to the Unemployment Insurance programs of the 50 states. In this paper, we assess the extent to which data compiled by the Census Bureau for multiple-establishment firms can aid this task. We find that for a convenience sample of firms, the Census Bureau data contain most of the Employer Identification Numbers (EINs)—and add additional EINs—to those found through labor-intensive searches of the BLS Longitudinal Database (LDB). However, not every apparently valid EIN for these firms in the BLS LDB appears in the Census Bureau's lists of EINs for these firms. Furthermore, some of the EINs that appear in the Census Bureau's lists do not appear to be valid EINs for these firms in the BLS LDB in the relevant year or quarter. We conclude that using Census Bureau data on multi-establishment firms can reduce (but not replace) the labor-intensive work of finding all the establishments for particular firms in BLS microdata.

Key Words: Firm, establishment, linking

1. Introduction

The Bureau of Labor Statistics (BLS) collects data from firms at the establishment level[1]. Most firms have only one establishment, but larger firms can be comprised of multiple establishments. For certain applications, it is desired to match firm-level data with the establishment data maintained by the BLS. Finding all of the establishments that comprise a firm is a fundamental challenge.

An example of matching together firm-level and establishment data is given in Handwerker, Kim, and Mason (2011). They attempt to find all the establishments associated with the 500 largest multinational manufacturing employers identified (at the firm level) in surveys conducted by the Bureau of Economic Analysis (BEA). Outside researchers interested in using the establishment-level microdata collected by the Bureau of Labor Statistics often suggest merging these microdata with firm-level datasets, such as

[1] An establishment is an economic unit, such as a farm, mine, factory, or store that produces goods or provides services. It is typically at a single physical location and engaged in one, or predominantly one, type of economic activity for which a single industrial classification may be applied.

corporate datasets compiled from firms' mandatory filings with the Securities and Exchange Commission (SEC) [2].

As part of Handwerker and Mason (2013), we outlined the efforts involved in linking establishment data into firms. This article is an extension of that work in that it examines the capacity of data on multi-establishment firms assembled by the Census Bureau to facilitate linking establishments in BLS data with their parent firms. In section 2 of this paper, we discuss the BLS establishment-level data and Census multi-establishment data. We discuss how the Census data might be expected to facilitate the process of linking establishments into firms and the methods used to evaluate the usefulness of the Census data in section 3. In section 4, we discuss the results of the evaluation. We conclude in section 5.

2. BLS Establishments and Census Multi-Establishment Data

The Quarterly Census of Employment and Wages (QCEW) is one of the primary sources of employer microdata at BLS. The QCEW contains quarterly records of all U.S. establishments subject to state Unemployment Insurance (UI) laws. The records in the QCEW are based on the quarterly contribution reports to the state agencies responsible for administering UI programs. As noted in Handwerker and Mason (2013), the QCEW covers approximately 9.2 million establishments and 98% of U.S. employment as of the 4th quarter of 2009[3]. QCEW records include monthly employment and quarterly total payroll data as well as establishment industry classification. Additionally, QCEW records include establishment names (legal and/or trade names), addresses (physical, mailing, and/or headquarters), and identifiers (UI account number and EIN). These data fields, most notably the identifiers, facilitate linking establishment data into firms.

While these identifiers facilitate linking establishment into firms, they are not sufficient to do so. UI accounts, issued by state unemployment agencies to identify employers for unemployment insurance purposes, are state-specific and do not distinguish firms that have establishments in more than one state. Handwerker and Mason found that 4.0% of employers (as identified by a distinct EIN number) have establishments in more than one state. Even within a state, a firm might have multiple UI accounts. Mason and Handwerker found that 4.1% of employers (as identified by a distinct EIN number) have more than one UI account. As such, without knowing every UI account number for a firm, UI account numbers do not uniquely identify firms. EIN numbers are issued by the Internal Revenue Service (IRS) to identify firms for tax purposes. While EIN numbers identify firms more accurately than UI accounts do—EINs are not state-specific, for instance, and as shown by Elvery, Foster, Krizan, and Talan (2006), most employers have only one EIN—EINs also do not uniquely identify firms[4]. Firms can have multiple EINs. For instance, firms may use

[2] On a limited basis, BLS allows eligible researchers to access confidential data for purposes of conducting valid statistical analyses (see http://www.bls.gov/bls/blsresda.htm for more information).

[3] Proprietors, domestic workers, unpaid family members, the self-employed, members of the armed forces, and railroad workers covered by the railroad unemployment insurance system are excluded.

[4] It should be noted that BLS does produce various estimates by firm size, using unique EINs to identify firms; however, this is different from being able to use an EIN to find all of the establishments of a firm.

different EINs in different states. Additionally, establishments involved in mergers and acquisitions may retain the old EINs for Unemployment Insurance reporting but use their parent company's EINs for tax purposes. Even if a firm does use only one EIN, the EIN they use in Unemployment Insurance reports (the source of the establishment data in the QCEW) may be different (yet equally valid) than the EINs they use for other purposes. These scenarios explain why, as Handwerker and Mason show, not all EINs used in SEC filings always appear in the QCEW.

The Census Bureau's multi-establishment data are derived from the Census Bureau's Report of Organization Survey, whose purpose is to obtain current organization and operating information on large multi-establishment firms. Other than the 5-year Economic Census, this annual survey provides the only direct source of information on changes in multi-establishment firm organization at the establishment level. For multi-establishment firms, the survey identifies establishments that have been sold, closed, continued, started, or acquired during the reference year. Annual payroll, first quarter payroll, and employment as of March 12 are also collected for each establishment. In addition, large foreign equity positions, and controlling interests held by other domestic and foreign-owned organizations are collected at the firm level. Additionally, identifiers such as EINs are associated with each establishment. While the survey asks for every EIN used by a firm for payroll tax filing, it is possible that the respondents who answer this survey in large firms are different staff than those who file Unemployment Insurance reports. For example, in some firms, the accounting or legal department may fill out Census forms, while the human resources department may fill out unemployment insurance reports. Thus, if a different EIN is used for unemployment insurance than for other purposes, it may be omitted from the list of every EIN used by a firm provided to the survey.

The BLS was granted access to the set of microdata from the Census Bureau's Report of Organization Survey for employers with multiple establishments, as part of a data sharing agreement between BLS and Census. This data sharing agreement is authorized by the 2002 Confidential Information Protection and Statistical Efficiency Act.

3. Identifying the Establishments of a Firm

As noted above, EINs define businesses for tax purposes and each establishment in the QCEW is associated with an EIN. Thus, linking establishments with EINs belonging to the same parent firm provides a means of identifying all the establishments belonging to that firm. Therefore, the fundamental challenge of combining firm-level data to the establishments in the QCEW that comprise the firm can be restated as finding all the EINs for the firm. We examine whether the Census multi-establishment data, which include a list of EINs for each multi-establishment firm, can help us in this work.

We examine the capacity of the Census data to facilitate finding all establishments for two sets of firms. For both sets, we can compare the number of EINs, establishments, and total employment for firms identified independently from the Census multi-establishment data and those we could find using the Census multi-establishment data. The first set of firms we examine are a random sample of 100 firms drawn from a list of firms identified in the 2007 Benchmark Survey of Foreign Direct Investment in the United States as affiliates of foreign multinational companies. The second is a subset of 20 of the 500 largest multinational manufacturing firms from the 2004 Benchmark Surveys of U.S. Direct Investment Abroad, which we had previously attempted to link with QCEW establishment

data, as described in in Handwerker, Kim, and Mason (2011). These data were provided to BLS under a data sharing agreement between BLS and the Bureau of Economic Analysis (BEA).

There are two reasons why these two sets of firms are convenient samples for a study of linking firms with establishments in BLS data. First, these firms report total firm employment in the BEA surveys. Reported total employment provides a target to determine the quality of establishment linking—the sum of establishment employment should be close to the target employment as reported by the two sets of firms. Second, we have some idea about firm composition. These firms report at least one (and often more) EIN, as well as addresses and firm names, which gives us a starting point for finding their establishments in the QCEW data. In addition, for the firms in the second set, we had already determined, as best we could without using the Census multi-unit data, all of the establishments that compose the firm.

We use two methods to find the list of EINs for these firms, independently of the Census multi-establishment data. First, tools already developed for similar projects are used to automatically search for establishments (and the EINs associated with them). These "automated matching procedures" are a set of SAS programs, macros, and linked excel spreadsheets that take a company name, address, and/or lists of EINs, and search for all establishments (and their associated EINs) in the QCEW with the same name, the same address, the same EIN, or which appear in lists of related companies from various BLS programs. Second, since these procedures can erroneously find "matches" with unrelated companies or miss the EINs of subsidiaries with different names, we use additional "hand-matching" procedures to remove erroneously matched EINs and add additional names, addresses, and their associated EINs. Handwerker and Mason (2013) discuss the difficulties of linking establishments and firms in more detail.

This iterative process of automated and manual matching procedures can be very time-consuming. In general, it requires more time to search for additional EINs than to remove erroneously matched EINs. When searching for additional EINs, it is often necessary to consult various sources, such as firm websites, SEC 10-k filings, or corporate databases, to compile lists of firm and subsidiary names, establishment addresses, and EINs. Removing those that are erroneously matched takes much less time. While still needing to consult the same sources, it is a directed search to determine if a particular name, address, or EIN belongs to the firm in question. Table 1 summarizes the time spent in hand-matching activities, both in finding additional EINs for firms and for removing EINs that were erroneously linked with these firms.

Table 1: Summary of Minutes Spent Hand-Matching Sample of 100 Firms

Type of hand-matching	Mean	SD	Range		Percentiles		
			Min	Max	Q1	Median	Q3
Finding establishments	8	9.87	0	65	2	5	11.25
Removing erroneously matched establishments	1.3	4.03	0	30	0	0	0

We summarize the total employment reported to BEA as well as the total employment and firm structure (the number of EINs that comprise the firm as well as their associated establishments) as determined by the iterative process of automated and hand-matching work in Table 2 for the sample of 100 firms from BEA's 2007 Benchmark Survey of Foreign Direct Investment in the United States. Also shown is a summary of percent

differences between reported employment and total employment found in the QCEW for the establishments of these firms through this matching process.

Table 2: Summary of Reported and Matched Employment and Firm Structure

Variable	Mean	SD	Percentiles				
			P5	P25	P50	P75	P95
Reported Total Employment							
Firm employment	307.4	753.7	1	7	57.5	179.5	1394.5
Matched Employment and Firm Structure (without using the Census Multi-unit data)							
Employment	303.3	801.9	0.5	5	46.5	171	1324.5
Number of EINs	1.7	1.8	0.5	1	1	1	5.5
Number of establishments	10.3	27.0	0.5	1	2	7	48.5
Comparison of Reported and Matched Total Firm Employment							
Percent difference	-8.3%	34.0%	-98%	-8%	0%	3%	26%

Table 2 shows that the majority of these firms are small (<50 employees) and medium (50-250 employees) sized, as reported to BEA. Based on the results of the matching procedures, most appear to have simple structures. 43 of the firms have a single establishment and only 24 firms have multiple EINs. For these 24 firms, the number of EINs ranges from 2 to 12 with a mean of 3.9 EINs (and standard deviation of 2.5).

There is also a large difference between reported and matched employment for the sampled firms. We could not find EINs for 5 of the sampled firms. On average we found less employment than reported, as is seen in the total employment for all 100 sample firms; reported total employment is 30,735 employees but we have only 30,332 employees in the establishments matched to these 100 sampled firms. Using the convention from Handwerker, Kim, and Mason (2011), matches are considered "good enough" when the absolute percent difference between reported and matched employment is 20% or less. 76 of the 100 sampled firms are good matches. Of the good matches, 44 are firms that have multi-establishments (out of 52 such firms in total).

4. Evaluating the Census Multi-Establishment Data to Aid Matching

We are interested in learning whether we can replace the time-intensive hand-matching work by using the Census Bureau's multi-unit data. These data should greatly facilitate linking establishments in the QCEW to firms in two ways. First, they provide a simple and direct procedure for linking establishments to firms: establishments are linked using the collections of EINs as given in the Census multi-establishment data. We first determine which firms in our sample have multiple establishments by seeing if the EIN(s) listed for the sample of firms appear in the Census multi-establishment data. If at least one EIN does appear in the Census data, we link all of the QCEW establishments in the firm using the whole collection of EINs provided by Census. The second way in which the multi-establishment data facilitates linking is by noting if the EIN(s) listed for each firm in our sample do not appear in the Census multi-establishment data. Ideally, we could then infer that it is not a firm with multiple establishments. As such, the EIN(s) listed for each firm

can be directly linked to the establishments in the QCEW. All data (BEA, Census' multi-establishment data, and QCEW) are for reference year 2007.

Using the reported total employment of the sampled firms, as well as their composition as determined through automated and hand-matching procedures, provides a comparison to evaluate the total employment and firm composition that we find from using the Census multi-establishment data to match these firms instead.

Table 3 shows the results of matching the EINs derived from the Census multi-establishment data with the EINs associated with the establishments within the QCEW.

Table 3: Results of Census EINs matched to QCEW microdata

Census EINs matched QCEW	Number of firms	Type of firms as determined during auto/hand matching			
		Unknown, No matches found	Single establishment	Multiple establishments	
				Single EIN	Multiple EINs
Yes	40	1	7	14	18
No	60	4	36	15	5

Only 40 of the 100 firms in this sample have EINs in the Census data. Of the 60 that did not, 36 were identified as single establishment firms in the automated/hand-matching procedures. The Census Bureau's Report of Organization Survey only samples large firms with multiple establishments or firms that are believed to have multiple establishments, based on administrative records. It can be expected that few single establishment firms are in the survey. Further, under the data sharing agreement, BLS does not have access to single establishment microdata. An additional 4 firms could not be found during the automated/hand-matching procedures. However, 20 firms that have multiple establishments (of which 5 also have multiple EINs) were not in the Census data.

Table 4 shows summary statistics for the 40 sampled firms for which we found EINs in the Census multi-establishment data.

Table 4: Summary of Reported and Census Employment and Firm Structure

Variable	Mean	SD	Percentiles				
			P5	P25	P50	P75	P95
Reported Total Employment							
Firm employment	307.4	753.7	1	7	57.5	179.5	1394.5
Matched Employment and Firm Structure							
Employment	890.3	1,812.3	12.5	47	185.5	838	4833
Number of EINs	4.1	4.1	1	2	2	5	12.5
Number of establishments	24	39.5	1	2	7	25	123
Comparison of Reported and Matched Total Employment							
Percent difference	204%	1,135%	-83%	-8%	1%	9%	621%

The summary statistics for the Census-derived EINs for the sample in Table 4 are much different from those in Table 2. Instead of finding less total employment, the table indicates

we are finding more, though the additional employment is concentrated in just a few of the firms. In fact, of the total Census matched employment of 35,612 employees, 19,235 are associated with 3 large firms. The reported employment is 27,243 for the 40 firms and 9,156 for these 3 large firms.

We further examine the differences in Tables 2 and 4 by comparing the EINs found during automated/hand-matching procedures and the EINs that are given in the Census multi-establishment data. Of the 163 EINs found by the matching procedures and the 117 Census EINs, 77 overlap. We consult firm websites, SEC 10-k filings, and corporate databases to determine why the remaining do not overlap, and whether the EIN (as identified through automated/hand-matching procedures or the Census multi-establishment data) belong to the firm or not. Table 5 show the results.

Table 5: Reconciling EINs Matched to Sampled Firms

Source of EINs	Number of EINS			Employment		
	Total	Correct	Incorrect	Total	Correct	Incorrect
In both	77	77	0	22,689	22,689	0
Census	40	15	25	12,923	1,402	11,521
Auto/hand-matching	86	86	0	7,643	7,643	0

Noting that the hand-matching procedures also make use of firm websites, SEC 10-k filings, and corporate databases, it is not surprising that the 163 EINs found during the automated/hand-matching procedures are deemed correct. It is also not surprising that they do not give full coverage—it is much easier and time efficient to remove incorrect matches then it is find missing matches. As such, we note that the additional 15 correct EINs found in the Census multi-establishment data are quite beneficial.

We did not expect to find EINs that are not for the sampled firms, however. Examining the incorrect EINs further, we find that the 25 incorrect EINs are associated with 6 firms. The firms are on average large (mean employment is 1,040.7) and complex (mean number of EINs is 4.5 and number of establishments is 33.2). Further examination shows that in all but one firm, the EINs correctly identify establishments belonging to these firms in other periods. The firms either were acquired by or acquired other establishments after the reference date for the firms in our sample. It is not that the Census multi-establishment data incorrectly identify a firm's establishments, but rather that they do so at the wrong time.

We find similar results when we use the Census multi-establishment data to replicate portions of the matching effort described in Handwerker, Kim, and Mason (2011). In this work we attempted to find all EINs, establishments, and employment for 20 of the 500 largest multi-national manufacturers in 2004. Without the benefit of the Census multi-unit data, our matching efforts came within 20% of the employment totals reported to BEA for 454 of these firms (the "well-matched"). We assessed how many of the EINs found for 2004 in the earlier project could have been found using the Census Bureau's 2007 links, and how many additional EINs could be accurately identified in the Census Bureau data. We examined 10 firms that were "well-matched" in our earlier efforts and 10 firms that were not "well-matched," including the largest firms that were not "well-matched" in our earlier efforts. Overall, there is a great-deal of overlap between the Census lists of EINs for 2007 and the EINs we found for 2004 without the benefit of this Census data. Furthermore, the additional EINs located by using the Census data generally appear to be

correctly identified for the relevant firms. The additional EINs found with the Census data appear to be particularly useful in accurately finding establishments for firms that were not "well-matched" in our earlier efforts. There were some additional EINs in the Census data that do not appear to be correct, and these appear to come from businesses acquired by these firms after 2004.

5. Conclusion

We find that the Census Bureau multi-establishment data cannot replace our automated matching and hand-matching procedures for linking establishments to firms. However, these data can augment our existing procedures. First, they can add additional EINs to our work. The Census multi-establishment data include additional correct EINs that we had not been able to find with our automated/hand-matching procedures. In total, for a random sample of 100 firms, the additional matches found using the Census Bureau multi-establishment data added 1,402 employees towards the target of 30,735 employees reported by these firms.

More significantly, the Census multi-establishment data can reduce the amount of time spent searching for relevant EINs in our hand-matching work. As shown in Table 1, the amount of time spent searching for EINs is much greater than that spent removing erroneously matched EINs. If we add matching using the Census multi-establishment data to our automated matching procedures, we will have fewer EINs to find during hand-matching procedures. Though there will also be some erroneously matched EINs, the time required to remove these is a fraction of that to add missing EINs.

That the Census data will be a good complement to our existing procedures but not replace them is not due to inadequacies in the Census data. Rather, it is not designed for this purpose. The QCEW contains quarterly data, and the Census Bureau's focus is on data quality for their Economic Census, which takes place every five years, and on the sample frame for their economic surveys, for which they want the best possible current data, not necessarily the best representation of particular reference dates in the past.

Acknowledgements

We thank Dan Yorganson, Brandy Yarbrough, Justin McIllece, Mark Loewenstein, and Anne Polivka for helpful discussions. The views expressed in this paper are solely those of the authors and do not reflect the views of the U.S. Bureau of Labor Statistics.

References

Elvery, Joel, Lucia Foster, C. J. Krizan, and David Talan, "Preliminary micro data results from the Business List Comparison Project," *Proceedings of the 2006 American Statistical Association Annual Meeting* (Alexandria, VA, American Statistical Association, 2006).

Handwerker, Elizabeth Weber, and Lowell G. Mason, "Linking firms with establishments in BLS microdata," Monthly Labor Review, June 2013, Vol. 135, No. 6.

Handwerker, Elizabeth Weber, Mina M. Kim, and Lowell G. Mason, "Domestic employment in U.S.-based multinational companies," Monthly Labor Review, October 2011, Vol. 134, No. 10.