Implementation and Results of a New Administrative Record Linkage Methodology in the Quarterly Census of Employment and Wages November 2016

Jessica Helfand[1], Justin McIllece[1]
[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, DC 20212

Abstract
The Quarterly Census of Employment and Wages (QCEW) program of the U.S. Bureau of Labor Statistics (BLS) uses a statistical matching component for the longitudinal linking of quarterly establishment records. The original proprietary statistical matching element, implemented in 1999, was recently replaced by a BLS created administrative record linking methodology, specifically designed for QCEW data. This paper describes the implementation and result of the new methodology.

Key Words: Record linkage, QCEW, Business Register

1. Introduction

The U.S. Bureau of Labor Statistics (BLS) Business Register can trace its roots back to 1990. Since its inception, the quarterly linkage of records on the Business Register has seen many developments and improvements. The most recent enhancement included the replacement of AutoMatch statistical software with an internally developed Weighted Match linkage methodology.

2. Quarterly Census of Employment and Wages

The Quarterly Census of Employment and Wages (QCEW) is the source of the BLS Business Register, covering roughly 97 percent of U.S. businesses.[2] Over nine million records are collected each quarter by state Unemployment Insurance (UI) tax filings, funneled through State Employment Security Agencies (SESAs).[3] In addition to meeting their tax liability, businesses provide data on their monthly employment, total wages, industry, geography, and administrative characteristics, such as name and predecessor or successor relationships with other businesses. This file of economic and administrative data is known as Enhanced Quarterly Unemployment Insurance (EQUI) data.

The BLS Business Register is used not only as a sample frame for high profile economic indicators such as the Current Employment Statistics (CES) Employment Situation, but also as the source of BLS Business Demography, known as Business Employment Dynamics (BED). BED publishes data on job churn in the economy, measured by gross

---

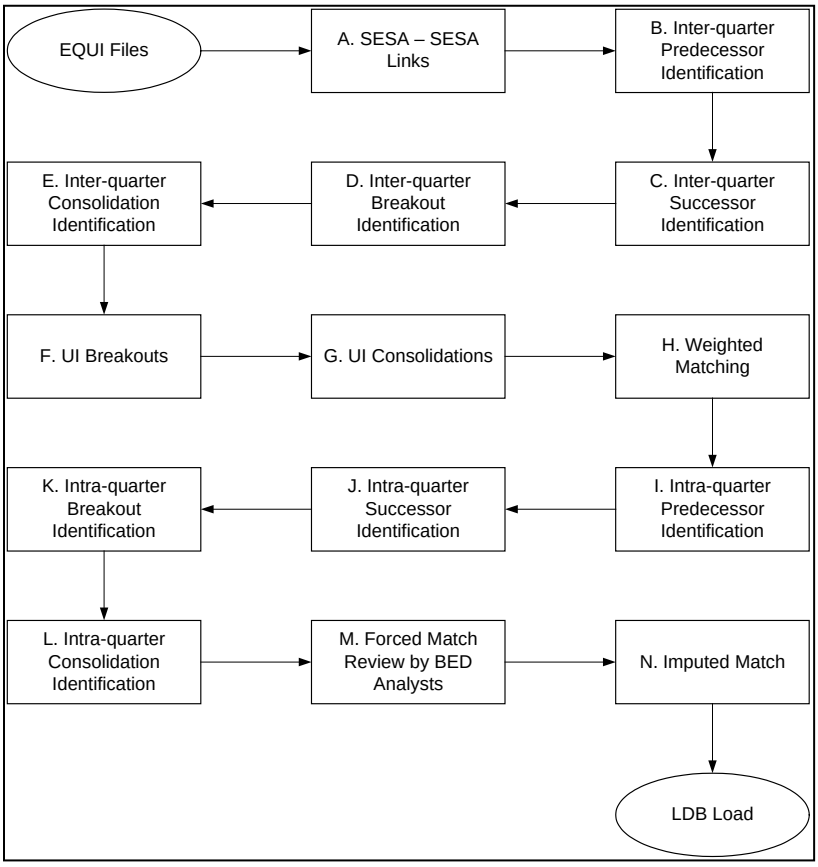
[1] Views expressed are those of the authors and do not necessarily reflect the views or policies of Bureau of Labor Statistics.
[2]http://www.bls.gov/cew/
[3]Ibid.; State Employment Security Agencies (SESA) are also known as State Workforce Agencies (SWA)

job gains and gross job losses.[4] These categories are further disaggregated into establishment births and deaths, among other categories.[5] For all applications of the Business Register, the proper treatment of business births and deaths is paramount; overstating births and deaths leads to inaccurate measures of entrepreneurship, inflated job churn, and faulty birth/death modeling for sample frame users. In order to mitigate false openings and closings on the Business Register, much attention is given to the quarterly linkage methods.

3. Quarterly Linkage Process

Each of the millions of records sent to BLS have a unique SESA-ID number. These microdata are then linked quarter-to-quarter by their SESA-ID number within a Longitudinal Database (LDB) of records and assigned an LDB number.

Chart 1: LDB Linkage Flow



[4] http://www.bls.gov/bdm/

[5] For the purpose of BED statistics and the Longitudinal Database, births are defined as units with positive third month employment for the first time in the current quarter, with no links to the prior quarter; or units with positive third month employment in the current quarter and zero employment in the third month of the previous four quarters. Births are a subset of openings, not including re-openings of seasonal businesses. Similarly, deaths are defined as units with no employment or zero employment reported in the third month of four consecutive quarters following the last quarter with positive employment. Deaths are a subset of closings, not including temporary shutdowns of seasonal businesses.

While roughly 96 percent of all records are linked by SESA-ID, the remainder are subject to economic or administrative changes. These changes reflect business openings, closings, and reorganizations such as business consolidation, new multi-establishment reporting, and changes in ownership structure. Whereas business births and deaths reflect new and ending entries in the Business Register, economic and administrative reorganizations are tracked with record linkages. Linkages include one-to-one, as well as one-to-many (breakout) and many-to-one (consolidations) relationships. By default, all records that are not linked during this process are considered births (new records) and deaths (records which end).

Chart 1 shows a simplified version of the QCEW linkage process. The automated process flows through SESA-ID matches; inter-quarter linkages based on state-populated predecessor and successor fields; a statistical match program to address any unlinked potential inter-quarter matched; and intra-quarter linkages based on state-populated predecessor and successor fields. Quarterly linking concludes with a final review conducted by a group of experienced BED analysts, where all large unmatched units are manually reviewed. This hands-on assessment is the final line of defense against spurious business openings and closings on the Business Register.

Since the beginning of the BLS Business Register, a Weighted Match component has been included in the overall linkage scheme (see Step H in Chart 1). This process creates links based on administrative characteristics in order to match records which experience SESA-ID reporting changes over the quarter. The original matching algorithm was replaced in 1999, and again recently in 2015. The Weighted Match process has historically linked between 0.09 and 0.25 percent of records.

Table 1: SESA-ID and Weighted Matches Over-Time
Matches in First Quarter 2003 – First Quarter 2015
Remaining records on the Business Register experience analyst matches, breakouts, consolidations, or are births and deaths.

Year	SESA-ID Matches	SESA % of total records	Weighted Matches	WM % of total records
2003	7,908,388	95.30%	18,535	0.22%
2004	8,012,705	95.15%	18,788	0.22%
2005	8,217,148	95.52%	13,951	0.16%
2006	8,448,492	95.64%	14,113	0.16%
2007	8,626,215	95.77%	15,066	0.17%
2008	8,786,490	95.78%	13,740	0.15%
2009	8,841,109	96.40%	12,980	0.14%
2010	8,768,027	96.39%	13,904	0.15%
2011	8,813,611	96.55%	11,535	0.13%
2012	8,915,483	96.24%	12,431	0.13%
2013	8,920,110	96.48%	12,532	0.14%
2014	9,026,489	96.61%	17,098*	0.18%
2015	9,147,636	96.51%	8,920**	0.09%

* The spike in 2014 Q1 is due to a reclassification of a number of establishments from private households (NAICS 814110) to services for the elderly and persons with disabilities (NAICS 624120). Private households are not within the scope of BED and, as a result, those establishments impacted by this industry reclassification are now within scope, causing more records to be subject to the Weighted Match process.
** 2015 Q1 reflects the new more efficient Weighted Match program.

This component's contribution to total matches has been trending down over time, even as the total number of records on the Business Register has grown (see Table 1); the decline in weighted matches can be attributed to a number of factors, including improved state identification of predecessor and successor relationships. While the number of records linked by Weighted Match may seem trivial, BLS research has shown that the weighted match component has a significant impact on the net number of births and deaths within the Business Register.[6]

4. Evolution of Statistical Linkage in the BLS Business Register

4.1 Original Weighted Match

The precursor to the modern BLS Longitudinal Database was the Universe Database (UDB), established in the early 1990s. This limited linkage system included a Weighted Match process that identified matches based on three blocks of shared criteria: [7]

- Block 1: Trade Name match, based on the first 7 consonants of the field
- Block 2: Physical Location Address match, based on the first 15 positions of the field
- Block 3: Phone Number match, identical fields

If two records matched within one or more blocks with a sufficiently high weight, the match was considered valid and the records linked. A significant shortcoming of this incarnation of the Weighted Match process was that in order for records to be eligible, they had to first match on location (county – or township for New England) and four digit Standard Industry Classification (SIC) code. For example, if a potential predecessor and potential successor were both classified *Alcohol Wholesale* (SIC 518), but one was under *Beer and Ale* (SIC 5181) and the other under *Wine and Distilled Alcohol* (SIC 5182) a match could not even be considered.

4.2 AutoMatch Software

In fiscal year 1995, BLS was provided with Congressional funding to create a database that would allow for the longitudinal analysis of business. This project came to fruition in 1999, when the Longitudinal Database (LDB) was released to internal users. It included significant revisions of the entire linkage process, in conjunction with the overhaul of data collection and editing procedures within the QCEW program. At this time new data elements, along with database upgrades, allowed for new and improved longitudinal analysis. Included in these enhancements was the replacement of the simple three-block Weighted Match with a proprietary linkage software called AutoMatch.[8]

---

[6]Kenneth Robertson, Larry Huff, Gordon Mikkelson, Timothy Pivetz, and Alice Winkler, "Improvements in Record Linkage Process for the Bureau of Labor Statistics' Business Establishment List" *1997 Record Linkage Workshop and Exposition Proceedings*, pp. 212-221.

[7] Ibid., p. 214.

[8] AutoMatch software was purchased from Matchware Technologies Incorporated, later to be serviced by IBM Websphere.

The AutoMatch software uses a probabilistic-based Weighted Match process, and allows for customization of blocking and weights for unique datasets. Weights are determined by using *m-probability*, defined as the probability of the variable agreeing in a matched pair, and *u-probability*, defined as the probability that a field agrees at random. Each variable or field contributes some information that improves the classification; the amount of improvement is the weight.

After much research and deliberation, BLS chose to use three sets of criteria, employing a total of 21 blocks. Variables included standardized business name,[9] physical location address, 3 and 6 digit NAICS,[10] county code, phone number, Federal Employer Identification Number (EIN), and zip code. Table 2 shows the grouping of these variables into three sets, where STD indicates address or name standardization and KEY indicates matching on the non-standardized field.

Table 2: Blocking Variables Used under AutoMatch

Match	Block	Blocking Variables
SET 1	Block 1	Trade Name (STD), PL address (STD), NAICS6, County
	Block 2	Trade Name (STD), PL address (STD), County
	Block 3	Trade Name (STD), PL address (KEY), NAICS6, County
	Block 4	Trade Name (KEY), PL Address (STD), NAICS6, County
	Block 5	Trade Name (STD), PL Address (KEY), NAICS3, County
	Block 6	Trade Name (KEY), PL Address (STD), NAICS, County
	Block 7	Trade Name (STD), PL Address (KEY), NAICS6, ZIP
	Block 8	Trade Name (KEY), PL Address (STD), NAICS6, ZIP
SET 2	Block 1	Trade Name (KEY), PL Address (KEY), NAICS6, Phone
	Block 2	Trade Name (KEY), PL Address (KEY), Phone
	Block 3	Trade Name (KEY), NAICS3, County, Phone
	Block 4	Trade Name (STD), Phone
	Block 5	Trade Name (KEY), County, Phone
	Block 6	PL Address (KEY), NAICS3, County, Phone
SET 3	Block 1	Trade Name (STD), NAICS3, County
	Block 2	Trade Name (STD), NAICS3, ZIP
	Block 3	Trade Name (KEY), PL Address (KEY), NAICS6, County
	Block 4	Trade Name (KEY), PL Address (KEY), County
	Block 5	EIN, PL address (KEY), County
	Block 6	EIN, ZIP
	Block 7	Trade Name (KEY), ZIP, County

Probability weights were adjusted based on characteristics of the data. For example, if records had similar street addresses, but different suite numbers, the weights were

[9] Trade Name, or "Doing Business As" name was primarily used. If the Trade Name was not populated, the Legal Name was used instead.

[10] The North American Industry Classification System (NAICS) is organized in a hierarchical structure, where a 6 digit classification is a detailed subsector of a 3 digit classification.

adjusted downward; conversely, blank fields were treated as potential matches and weights were adjusted upward.[11]

Testing was conducted using data from California, West Virginia, Georgia, and Florida, where cutoff values for each block were fine-tuned in order to maximize good matches, while minimizing incorrect matches. The results of the AutoMatch linkages did not differ dramatically from the original Weighted Match process, but did provide marginal improvements. Moreover, the Weighted Match process continued to have a notable impact on the assignment of births and deaths within the database.[12]

4.3 2015 Replacement

In 2013 development began to replace this proprietary software with an in-house BLS Weighted Match system. This replacement was motivated by several factors: a need to improve weighted match record linkages; limited technical support from the vendor; and significant annual cost savings for the QCEW program. Over the course of two years, a program was written and refined by BLS Mathematical Statisticians Justin McIllece and Vinod Kapani. As described in their 2014 publication, the replacement Weighted Match system measures the similarity of two records by calculating a weighted Euclidean distance between them.[13] This is a departure from the classical method of probabilistic record linkage, originally developed by Fellegi and Sunter (1969) and utilized by the AutoMatch software.[14]

Based on a relatively small number of QCEW variables, given in Table 3, this distance is scaled to the [0-1] range and is constructed such that higher numbers, or scores, represent greater similarity between records. Thus, a score of one would constitute a perfect record match, while a score of zero would suggest that there is no measurable similarity between the two records at all; i.e. they are perfectly dissimilar in the context of the 2015 matching system. Unique variables, including EIN, names, Reporting Unit Description, and address are given higher weights. Conceptually, it is desirable that record pairs with high scores are flagged as links, while those with low scores are discarded. Informed by empirical review, the value applied as a cutoff was about 0.58, which was selected as a satisfactory compromise between missing too many good links (by setting the cutoff too high) and flagging too many bad links (by setting the cutoff too low). Additionally, a second criterion was implemented: if the record pair sufficiently matches on a combination of critical variables, despite having a low score (typically due to missing data), a link between the records is established.

The 2015 Weighted Match replacement includes significant improvements from the AutoMatch linkages. For example, both prior Weighted Match systems compared Trade Name to Trade Name only, and would not consider a match where the Legal Name of one reporter matched the Trade Name of another. As employers are not necessarily

---

[11] Robertson et al. “Improvements in Record Linkage Process for the Bureau of Labor Statistics’ Business Establishment List” p. 216.
[12] Ibid., p. 218.
[13] Justin McIllece, Vinod Kapani (2014) “A Simplified Approach to Administrative Record Linkage in the Quarterly Census of Employment and Wages” JSM Proceedings, 2014, pp. 4392-4404.
[14] Ivan P. Fellegi, Alan B. Sunter (1969) “A Theory for Record Linkage,” Journal of the American Statistical Association, Vol 64, No. 328, pp. 1183-1210.

Table 3: New Weighted Match Components

Linkage Variable	Type	Weight
EIN	Categorical	1.75
County	Categorical	1.00
Phone Number	Categorical	1.00
NAICS	Categorical	1.00
Average Quarterly Employment	Numeric	1.00
Total Quarterly Wages	Numeric	1.00
Standardized Trade Name	Text	1.75
Standardized Legal Name	Text	1.75
Reporting Unit Description	Text	1.75
Physical Location Address	Text	1.75
City and Zip Code	Text	1.00

required to report both their Legal Name and Trade Name,[15] employing this cross-checking in the new methodology has the capacity to capture more matches.

Another substantial change is the treatment of non-unique values. If any value of a blocking variable occurs over 50 times, the value is down-weighted. This becomes important particularly with multi-establishment employers, where administrative values within their reports, such as Legal and Trade Names, Phone Number, or EIN, will repeat. These repeated values are down-weighted in order to minimize false linkages of similar but unrelated accounts.

For a more complete description of the general linkage methodology, its motivation and limited examples, see the 2014 Joint Statistical Meeting article by McIllece and Kapani: *A Simplified Approach to Administrative Record Linkage in the Quarterly Census of Employment and Wages.*

5. Implementation and Results

The final rounds of testing were conducted using second quarter 2014 QCEW data. One primary goal of the new software was to improve the Weighted Match process, but in doing so, to not create a break in series. As mentioned above, the weights assigned to the AutoMatch program in the late 1990s had not been updated in nearly two decades. During this time data collection and editing techniques had evolved and improved; it had become apparent during internal review that while many of the AutoMatch linkages were correct, the program was also creating an unacceptable number of mismatches. The new BLS linkage software needed to mimic the net result of the AutoMatch system, while improving upon the linkages themselves.

As noted by McIllece and Kapani, the overlap rate, where AutoMatch and the new Weighted Match system identified the same links, was highly varied across test states.[16] Of the seven states they tested, Georgia showed the highest rate (62.6 percent overlap), and California the lowest (28.6 percent overlap). However, despite the low levels of

[15] Requirements for employer reports differ based on State Unemployment Insurance Tax laws.
[16] McIllece and Kapani, "A Simplified Approach to Administrative Record Linkage in the Quarterly Census of Employment and Wages" p. 4403.

overlap, it became clear during analyst review of five quarters of matches that the new Weighted Match system captured more valid links and significantly fewer false matches than the AutoMatch.

Further, when reviewing summary data of all states with second quarter 2014 data, BLS became confident that despite the differences in the programs, a break in series is not expected. Despite capturing fewer links, the new Weighted Match program produced higher quality links and accounted for more employment. Table 4 shows employment by openings and closings at the national level, both by total establishments and firm size class, when employing AutoMatch and under the new Weighted Match program. The employment in these categories is little changed.

Table 4: Private Sector Employment at Openings and Closings, 2014 Q2
In Thousands, Not Seasonally Adjusted

	Openings			Closings		
	AutoMatch	New WM	Difference	AutoMatch	New WM	Difference
Establishments	1,525	1,527	2	1,052	1,053	1

	Openings			Closings		
Firm Size Class	AutoMatch	New WM	Difference	AutoMatch	New WM	Difference
1 to 4	651	652	1	414	414	-
5 to 9	191	191	-	111	111	-
10 to 19	131	131	-	70	69	(1)
20 to 49	95	95	-	48	48	-
50 to 99	30	30	-	16	16	-
100 to 249	12	13	1	8	8	-
250 to 499	3	3	-	3	3	-
500 to 999	1	1	-	2	2	-
1,000 or more	0	0	-	1	1	-

Tabulations of employment at establishment birth were evaluated by industry and by state. Table 5 shows employment at establishment births for the private sector and industry super-sectors under the AutoMatch and new Weighted Match. Again, little was changed, with a 1.2 percent increase in employment across all industries. This includes a 5.6 percent increase in the Education and Health Services sector. This large change was acceptable, as during the first several quarters of 2014 a large number of number of establishments were reclassified from private households (NAICS 814110) to services for the elderly and persons with disabilities (NAICS 624120). Private households are not within the scope of Business Employment Dynamics tabulations, and a result, those establishments are counted as births when moving into the Education and Health Services sector.

On average, states experienced an increase of 1.1 percent in employment at establishment births, not seasonally adjusted. At this time, more detailed test tabulations at the state level cannot be released due to non-disclosure restrictions.

The new BLS Weighted Match program was implemented with the fourth quarter 2014 linkage process, conducted in May 2015. These data were published July 29, 2015. As of July 27, 2016, the new Weighted Match program has been used to publish data through fourth quarter 2015and continues to produce satisfactory results.

Table 5: Private Sector Employment at Establishment Births, 2014 Q2
Not Seasonally Adjusted

Industry	AutoMatch	New WM	Difference	% Difference
Total Private	798,621	808,288	9,667	**1.2%**
Natural Resources and Mining	14,856	14,734	(122)	-0.8%
Construction	64,681	64,762	81	0.1%
Manufacturing	19,912	20,070	158	0.8%
Wholesale Trade	25,557	25,492	(65)	-0.3%
Retail Trade	87,785	88,999	1,214	1.4%
Transportation and Warehousing	20,950	20,969	19	0.1%
Utilities	720	726	6	0.8%
Information	14,489	14,044	(445)	-3.1%
Financial Activities	39,647	40,028	381	1.0%
Professional and Business Services	132,060	132,629	569	0.4%
Education and Health Services	85,355	90,154	4,799	**5.6%**
Leisure and Hospitality	175,166	177,631	2,465	1.4%
Other Services	35,810	35,664	(146)	-0.4%
Unclassified	81,633	82,386	753	0.9%

6. Conclusions

The BLS Business Register has evolved over time, as has the statistical matching linkage component used within. Building on past experience, BLS Mathematical Statisticians and Economists were able to design a new, in-house Weighted Match program capable of maximizing record linkages. Thus far, the results of the new Weighted Match program have been more than acceptable. The goals of the project were met. Weighted Match linkages have improved, with comparable employment captured and with fewer false linkages; and QCEW is no longer paying for software with limited technical support from the vendor. Additionally, many linkage opportunities await the QCEW and other BLS programs in the world of "Big Data", and future applications of the Weighted Match program will be explored.

References

Fellegi, Ivan P., and Alan B. Sunter (1969) "A Theory for Record Linkage," Journal of the American Statistical Association, Vol 64, No. 328, pp. 1183-1210.

McIllece, Justin, and Vinod Kapani (2014) "A Simplified Approach to Administrative Record Linkage in the Quarterly Census of Employment and Wages" JSM Proceedings, 2014, pp. 4392-4404.

Robertson, Kenneth, Larry Huff, Gordon Mikkelson, Timothy Pivetz, and Alice Winkler, "Improvements in Record Linkage Process for the Bureau of Labor Statistics' Business Establishment List" *1997 Record Linkage Workshop and Exposition Proceedings*, pp. 212-221.

U.S. Department of Labor, Business Employment Dynamics website, accessed frequently. www.bls.gov/bdm

U.S. Department of Labor, Current Employment Statistics website, accessed frequently. www.bls.gov/cew