A Simplified Approach to Administrative Record Linkage in the Quarterly Census of Employment and Wages October 2014

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

Abstract
In the Quarterly Census of Employment and Wages program of the U.S. Bureau of Labor Statistics, quarterly establishment records are linked longitudinally by a software-based logistic regression system that assigns linkage probabilities to potential establishment matches. Each field in the administrative record is given a weight according to its non-missing rate, and records are linked if their estimated linkage probability exceeds a minimum threshold. In this paper, we consider a simplified matching alternative that relies on the principal data fields related to the administrative definitions of establishment linkage. A linkage system based on critical matching criteria, such as location and administrative business information, is constructed. The results of the alternative record linkage approach are compared to those generated from the logistic regression software.

Key Words: Record linkage, QCEW

1. Introduction

The Quarterly Census of Employment and Wages (QCEW) program of the U.S. Bureau of Labor Statistics (BLS) maintains business establishment registers for all states, which include such information as employment, total paid wages, industry codes and physical location. Compiled from data obtained by state-level Unemployment Insurance (UI) programs, the QCEW registers cover nearly the entire universe of U.S. business establishments. On a quarterly cycle, these lists are updated with the most recently available administrative and economic data. Continuous establishments—businesses that continue operations under the same ownership from one quarter to the next—are linked between quarterly files by a hierarchical record linkage system.

Since the QCEW files are used to construct sampling frames for BLS establishment surveys, such as those conducted by the Current Employment Statistics (CES) and Occupational Employment Statistics (OES) programs, improper linkage affects their sample selection procedures. For example, in the CES survey, a supplemental birth sample of new businesses is annually selected, as described at the CES website. Improper linkage, or incorrectly identifying businesses as new or continuous, would lead to new businesses being ineligible for birth sampling and continuous businesses to be eligible. Business

[1] Views expressed are those of the authors and do not necessarily reflect the views or policies of the Bureau of Labor Statistics.

Employment Dynamics, discussed in Section 4.3, would also be negatively impacted by improper linkage.

In the first step of the record linkage system, most continuous establishments are linked via a unique combination of state code, UI Number, and Reporting Unit Number. This combined field, called SESA ID, links the vast majority of the establishments. The subsequent three steps use such information as predecessor/successor (P/S) codes, which identify change of ownership of continuing businesses, and repeats of SESA IDs in either the prior or current quarter to identify business breakouts and consolidations. These four administrative steps constitute the bulk of the record linkage system.

The fifth step of the record linkage system applies the licensed software package AutoMatch, a probabilistic record linkage system, to the remaining subsets of records unlinked in the previous steps. Establishment pairs that exceed a minimum probability threshold are linked between quarters. Matching variables, blocks (minimum criteria for establishment pairs to be further evaluated in the matching process), and probability thresholds are specified by the user, but the internal methodology and computation, including how variables are compared and scored, is wholly contained within the system. A more detailed description of key AutoMatch features is given by Thomas (1999). Comparisons of linkage systems made in this paper are based solely on the production settings QCEW deploys when using AutoMatch.

In this paper, an alternative linkage system developed at BLS, called the Weighted Match (WM), is presented. Target links are described according to the QCEW definition of the P/S establishment relationship, and critical variables are identified. Variable scoring methodology, with special emphasis on text strings, is presented. Overlapping and non-overlapping establishment links generated by the two systems are compared for several test states. Performance of linked records is evaluated by estimating accuracy and power based on data expert review of results. Impact on statewide employment of business opening and closings is measured. Feasibility of replacing the AutoMatch software by the WM system is discussed.

2. Predecessor/Successor Establishments

The QCEW program defines the P/S relationship as follows:

“A predecessor/successor relationship is defined as one where the successor (the new owner of an establishment) performs similar operations to the predecessor (the previous owner of an establishment) using some or all of the predecessor’s employees. These operations are frequently, but not necessarily, performed at the same location as the predecessor” (QCEW Operating Manual).

A few key concepts from this definition are: performing similar operations; using some or all of the same employees; and frequently performing operations at the same location. Relevant QCEW data elements are mapped to these concepts. Some elements offer information of specific interest to the P/S definition, while others serve as proxies in the absence of more detailed information.

The basic functionality of the WM system is to create a list of potential P/S relationships, or links, and evaluate them based on certain criteria to select the subset of good links from the larger list. Establishment pairs are formed by matching a business record from the prior

quarter file (File 1) to a record from the current quarter file (File 2). These records will be referred to as A and B, respectively, while AB represents the linkage of these two establishments.

Table 1: QCEW Variables, Similar Operations

Data Element	*Variable*	*Definition*
6-Digit North American Industrial Classification System Code	NAICS	Classifies type of business activity performed by establishment
Trade Name	TRADE	Identifies business name and/or type of business activity performed
Legal Name	LEGAL	Identifies business name and/or type of business activity performed
Employment Identification Number	EIN	Federal Tax Identification Number
Reporting Unit Description	RUD	Additional operational or locational information for some multi-establishment firms

Not all variables fall solely into one category. For example, RUD may be useful for identifying the specific operations of an establishment or for identifying its location, depending on the reporting structure of the firm.

Table 2: QCEW Variables, Retained Employees

Data Element	*Variable*	*Definition*
Average Monthly Employment	EMP	Count of establishment's average quarterly employment
Total Wages	WAGE	Count of total quarterly wages paid to employees

The variables EMP and WAGE do not specify that the same employees are used. Instead, it is assumed that increasing similarity of employment and wage data between A and B indicates increasing likelihood that AB is a good link.

Table 3: QCEW Variables, Same Location

Data Element	*Variable*	*Definition*
Address Line 1	ADDR1	Physical street address of an establishment
City, Zip Code	ADDR2	Physical city and zip code of an establishment
County Code	CNTY	Physical county code of an establishment
Phone Number	PHONE	Establishment phone number

The address variables in Table 3 are not necessarily related to physical location. A hierarchy is followed that uses physical location data when available. UI address information is used if physical location data is unavailable. Sometimes the information contained in these variables represents corporate headquarters, rather than the physical location of the establishment.

Since no single variable among those listed in Tables 1 through Table 3 is solely indicative of a good link, they are used collectively in the matching process. The target links for the WM system are establishment pairs with enough similarity among these variables to suggest that the P/S relationship definition is satisfied.

3. Methodology

The WM system is composed of three primary steps: blocking, scoring, and selection. The blocking step constructs an initial match file based on record pairs that meet a baseline matching requirement. The scoring step assigns a numeric value to all record pairs assembled in the blocking phase. The selection step determines final one-to-one links based on the results of the scoring phase. All results are based on linking establishments between the 4th quarter of 2012 and the 1st quarter of 2013.

The street address field, ADDR1, is standardized prior to blocking. Data entries are converted to standard postal abbreviations to improve matching capabilities. For example:

- Street → ST
- Avenue → AVE

The details of the standardization process are not addressed in this paper.

3.1 Blocking

The quarterly state data files (File 1 and File 2) that serve as inputs into the WM system tend to have, at minimum, several thousand records. Input files for some large states contain over 100,000 establishments. A Cartesian product of File 1 and File 2 for these large states would result in match files consisting of billions of potential establishment pairs. Since the QCEW linkage system is run on a quarterly basis for all states, this is computationally impractical.

Blocking minimizes computational burden by screening File 1 and File 2 for record pairs that match on at least one prescribed matching criterion, or block. Blocks include single and concatenated variables, compressed (blanks removed) and truncated variables, cross-matching variables, etc. The WM system blocks cast a wide net, endeavoring to catch as many good links as it can, with file size a secondary concern. Blocking constraints, described at the end of this section, are implemented to protect against the match file becoming too large for efficient computation.

The WM system constructs the match file using the blocks listed in Table 4. Missing values are not considered matches. Other restrictions are applied depending upon the properties of the variables, such as EIN equal to all zeroes or all nines. For brevity, the specific restrictions are not listed. Variables combined by a plus sign (+) indicate concatenated fields, all of which must be non-missing to be eligible for blocking. Variables combined by an asterisk (*) indicate cross-matching fields, where a variable from File 1 is matched to a different variable from File 2.

Table 4: Weighted Match Blocks

Block	*Data Elements*	*Compressed?*	*Truncation*
1	EIN		

2	LEGAL	✓	First 10 characters
3	TRADE	✓	First 10 characters
4	PHONE + CNTY + NAICS		First 3 characters (NAICS only)
5	LEGAL * TRADE	✓	First 10 characters
6	TRADE * LEGAL	✓	First 10 characters
7	RUD		
8	ADDR1 + CITY + ZIP		

The truncation of LEGAL, TRADE, and NAICS is applied to capture more potential links in the blocking phase. Some good links inconsistently report business names, while others may have a slight NAICS reclassification.

Without safeguards, it is possible that a block could generate a computationally cumbersome number of record pairs. This occurs if there are many repeated entries in both input files for a particular block. For example, a corporation could have a very large number of establishments in a state, all with the same EIN. Since the blocks act as Cartesian products of subsets of the input files, this may exponentially increase the size of the match file, increasing run time to an unacceptable level.

A constraint of 50 repeated entries is implemented to eliminate the possibility of generating an overly large match file. The maximum Cartesian product, or number of record pairs generated by a specific data entry, is 2500 (50^2), which is not large enough to cause noticeable slowdown of the WM system. It is assumed that data entries repeated more than 50 times are not particularly useful for identifying good links that are otherwise unidentified by other blocks.

3.2 Scoring

Once the match file is constructed, the similarity of all record pairs is quantified by comparing the eleven variables listed in Table 1. Each variable receives a matching score from zero to one. How the score is generated depends upon the category of the variable. A fundamental characteristic of the WM system is the ability to quantify variables, particularly text strings, as something other than a perfect match or non-match. This is critical since the same business establishment may report the same data differently in successive quarters.

3.2.1 Categorical match scores

Categorical variables include EIN, CNTY, PHONE, and NAICS. The scoring rules for these four variables are presented below. The rules always refer to the match level of the specified variables between records A and B of the potential link AB.

EIN, CNTY, and PHONE are binary because even slight differences are enough to be considered a non-match. For example, if A and B possess similar phone numbers, it may be reasonable to believe that they operate in the same general neighborhood, but that is not useful for identifying specific matches.

Note that NAICS, the only categorical element that does not follow a binary scoring system, is scored based on its highest match level only. Unlike phone numbers or county codes, similarity of NAICS industry codes has value from a matching perspective, since the products or services offered by a business may evolve over time.

$$EIN, CNTY, PHONE = \begin{cases} 1, & exact\ match \\ 0, & otherwise \end{cases}$$

$$NAICS = \begin{cases} 1, & 6\ digit\ match \\ .67, & 4\ digit\ match \\ .33, & 2\ digit\ match \\ 0, & otherwise \end{cases}$$

While there are many cases where similar but inexact phone numbers may be representative of the same establishment, particularly for large businesses, it was found that relaxing the phone number criterion had a negative impact on overall quality of the WM links.

3.2.2 Numeric match scores

Numeric variables include EMP and WAGE. The numeric scoring formula is given below.

Let
e_a = Average quarterly employment of A
e_b = Average quarterly employment of B
w_a = Total quarterly wages of A
w_b = Total quarterly wages of B

$$EMP = \begin{cases} 1 - \dfrac{|e_a - e_b|}{MAX(e_a, e_b)}, & if\ e_a > 0, e_b > 0 \\ 0, & otherwise \end{cases}$$

$$WAGE = \begin{cases} 1 - \dfrac{|w_a - w_b|}{MAX(w_a, w_b)}, & if\ w_a > 0, w_b > 0 \\ 0, & otherwise \end{cases}$$

Positive values are required for both establishments in AB so that missing or zero employment or wages are not given any matching value. This formulation gives preference to continuous matches (those with positive employment and wages in both files) over zero-to-positive and positive-to-zero matches. For records with positive e_a and e_b (or w_a and w_b, analogously), using the maximum of the two numeric variables in the denominator stabilizes the scores and restricts them to the (0,1] range.

3.2.3 Text match scores

Text variables include TRADE, LEGAL, RUD, ADDR1, and ADDR2. While ADDR2 (which includes city and 5-digit zip code) could be standardized and considered categorical, it is currently treated as a text element for scoring purposes.

Of all variable types, accurately comparing and scoring two text strings is typically the most difficult. While it is not always obvious among categorical and numeric variables what match score should be attributed to a pair of variables in AB, there are significantly fewer situations that must be considered. Text strings are complicated. How to quantify the similarity of any given pair of strings is not necessarily clear.

Among the issues comparing text strings is the inconsistency with which the data are reported. Specific to the QCEW linkage system, important matching variables such as TRADE and LEGAL may be entered differently in successive quarters, though they

represent the same business. While not all-inclusive, most differences that occur in legitimate P/S relationships fall into one or more of the situations listed below. In Table 5, mock TRADE examples, motivated by real data, illustrate each situation.

Table 5: Reporting Inconsistencies, Trade Name

Reporting Inconsistency	*File 1*	*File 2*
Misspelling	John Doe Tree Service	John Do Tree Service
Transposition	John Doe Tree Service	Doe John Tree Service
Change in Business Name (ownership transfer)	John Doe Tree Service	Jane Doe Tree Service
Word Substitution	Jane Doe Tree Service	Jane Doe Tree Removal
Word Addition/Subtraction	Jane Doe Tree Removal Service	Jane Doe Tree Service
Abbreviation	Jane Doe Tree Service	Jane Doe Tree Srvc
Variable Crossing	Jane Doe Tree Service (TRADE)	Jane Doe Tree Service (LEGAL)

Any of these pairs may be considered a match, but the text strings being compared are different in each case. None are exact matches, which could be problematic in a binary scoring system. A string comparison method is needed that allows a linkage process to recognize some similarity between the text pairs above and score them accordingly.

The WM system addresses this need with a simple algorithm dubbed SCOPE (String Comparison, Ordered Pair Enumeration) for short. Let T1 and T2 represent two text strings being compared, where T1 has at least as many total characters as T2. SCOPE looks through T1, storing every ordered character pair until it comes to the end of the string. Any string including a blank is dropped, such that word order does not negatively impact the comparison. SCOPE then searches T2 for each character pair. The final match score for the text strings is the proportion of ordered character pairs from T1 that are located in T2. Therefore, the match score for text data has the range [0,1]. SCOPE is a variation of the Jaccard distance, which is available in the *stringdist* R package (van der Loo, 2014).

Example of SCOPE algorithm:

T1 = “John Doe Tree Service”
T2 = “John Do Tree Service”

Ordered Character Pairs from T1
Since T1 contains 21 characters, there 20 ordered character pairs:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
JO	OH	HN	N_	_D	DO	OE	E_	_T	TR	RE	EE	E_	_S	SE	ER	RV	VI	IC	CE

Pairs 4, 5, 8, 9, 13, and 14 all contain a blank and are dropped from the array. The others are renumbered to represent the total number of eligible character pairs. A character pair receives a value of one if it is located in the T2 string or zero otherwise.

1	2	3			4	5			6	7	8			9	10	11	12	13	14
J	O	H			D	O			T	R	E			S	E	R	V	I	C
O	H	N			O	E			R	E	E			E	R	V	I	C	E
1	1	1			1	0			1	1	1			1	1	1	1	1	1

Of the fourteen character pairs, thirteen are located in T2. Therefore, the match score that results from comparing these two text strings is 13/14, or approximately 0.93.

Applying the SCOPE algorithm to all the examples given above yields the following match scores:

Text String 1 (T1)	*Text String 2 (T2)*	*Match Score*
John Doe Tree Service	John Do Tree Service	0.93
John Doe Tree Service	Doe John Tree Service	1.00
John Doe Tree Service	Jane Doe Tree Service	0.79
Jane Doe Tree Service	Jane Doe Tree Removal	0.57
Jane Doe Tree Removal Service	Jane Doe Tree Service	0.70
Jane Doe Tree Service	Jane Doe Tree Srvc	0.64

The variable crossing example, where the business name is entered for TRADE in one file and entered for LEGAL in the next, would receive a match score of 1.00 in the WM system. LEGAL and TRADE are cross-checked for matching entries.

To prevent inflation of match scores due to incidental character pair matches, minimum thresholds are set for every text variable. If the match score is below the threshold, the match score for that variable is reset to zero.

Table 6: Variable Scoring Thresholds

Matching Variable	*Threshold*
ADDR1	0.50
ADDR2	0.70
LEGAL	0.25
TRADE	0.25
RUD	0.70

The SCOPE algorithm is a flexible tool that generally adapts well to the many situations that arise with text string data. Non-egregious misspellings and transposed words are handled cleanly without any prior specification. Small changes in business names can be accounted for and still given a nonzero match score. SCOPE tolerates a limited number of word additions, subtractions, substitutions, and abbreviations, allowing some boost to potential links containing partial matches of the ordered character pairs. Alternatively, SCOPE could be applied without any thresholds for a less-parameterized system, since the overall scoring calculation uses squared terms that drives low-percentage matches close to zero.

3.2.4 Linkage scores

Once all variable matches are scored, a linkage score D is calculated for the record pair AB. The linkage score is a weighted Euclidean distance calculated from the eleven variable

match scores generated between A and B. Higher weights are given to "unique" variables that have greater specificity for individual establishments: EIN, LEGAL, TRADE, ADDR1, and RUD. (Note that despite the heavier weighting, any variable may contain non-unique information, such as multiple businesses sharing an EIN or street address.) Lesser weights are ascribed to the remaining six "general" variables, which often do not offer information specific to an individual establishment: PHONE, NAICS, ADDR2, CNTY, EMP, and WAGE. Initially, PHONE was included among the unique variables, but too many instances of new businesses assuming the phone number of closed businesses (in the same location) led to its reclassification as a general variable.

The distinction between unique and general variables is especially relevant in dense business areas, such as shopping malls or business districts. For example, different clothing stores in the same mall or business district will likely have the same data for NAICS, ADDR2, and CNTY, and similar data for EMP and WAGE. While useful for confirming area and type of business, these five variables do not provide enough specific information to identify a link. Some unique identifying information is required before the WM system will declare the record AB as a link.

To penalize variable match scores based on r repeated data entries, such as multiple establishments containing the same values for EIN or ADDR1, a downweight factor u is calculated for each of the 11 variables used in scoring:

$$u_i = \frac{1}{\sqrt{r_i}}$$

Letting w represent the variable weights and x represent the variable match scores, as defined in Section 3.2, the general formula for the linkage score D is given by:

$$D = \sqrt{\frac{\sum_{i=1}^{11} w_i u_i x_i^2}{\sum_{i=1}^{11} w_i}}$$

The range of D is [0,1]. The values of the weights that generated the results in this paper were 1.00 for general variables and 1.75 for unique variables.

3.3 Selection

A record pair AB is accepted as a link if it meets either of the following conditions:

1. $D \geq k$, where k is the cutoff score, and AB is the record pair with the highest score of all potential links involving A and B
2. $D < k$, but AB sufficiently matches on a combination of critical variables

Regarding the first condition, the value of k used to generate the results in this paper is approximately 0.58, which is a measure of the proximity of two weighted vectors, not a link probability as would be generated from a logistic regression procedure. This choice of k is based on observational review of the results, as it seems to generate a large number of good links without introducing an unacceptable number of bad links.

Regarding the second condition, links are accepted if, prior to applying the penalty factor *u*, the variable match scores for ADDR1 and either LEGAL or TRADE are all very close to one. Data experts identified many good links with scores below *k* matching on these specific variables. The low scores resulted from heavy penalty factors, associated with firms with a large number of establishments, and from missing data in other fields, lowering the linkage score *D*. Essentially, this condition considers a record pair AB a good link if the establishments match on business name and street address.

4. Results

The objective of the WM system is to serve as an adequate replacement for the AutoMatch software currently used by the QCEW program. Thus, the primary results are presented as a series of comparisons of how these two systems perform at linking records that could not be linked during the administrative steps:

1. Overlap rates of WM and AutoMatch links
2. Expert review of links generated by each system
3. Business employment dynamics measures

Additionally, as a validity test, the WM system was run on a handful of full state data files and compared to links generated by the four administrative linkage steps discussed in Section 1. Approximately 99% of links determined administratively are also identified by the WM system. Administrative links missed by the WM system sometimes suggest an alternative link that may be superior, due to errors in administrative codes. However, identification of errors in administrative linkage has not been an objective of this system development. Any evaluation of that property of the WM system would require extensive research.

4.1 Overlap Rates

Overlapping links are those identified by both the WM and AutoMatch systems. Results were compared in seven test states: Alabama, California, Delaware, Florida, Georgia, New York, and Texas. Overlap rates, presented in Table 7, were calculated with respect to the total links generate by the WM system. Therefore, these rates represent the percentage of WM links that were also identified by AutoMatch.

Table 7: Total Links and Overlap Rates

State	*WM System*	*AutoMatch*	*Overlap*	*Overlap Rate*
Alabama	157	126	98	62.4%
California	752	543	215	28.6%
Delaware	17	9	5	29.4%
Florida	1,838	1,924	1,038	56.5%
Georgia	380	1,635	238	62.6%
New York	798	2,084	232	29.1%
Texas	667	725	216	32.4%
Total	4,609	7,046	2,042	44.3%

The number of links and overlap rates vary significantly by state. Across the seven test states, the overlap rate is not as high as might be expected, indicating that the composition of the linkage results between the two systems is quite different.

4.2 Expert Review

Given the fairly low overlap rates from Table 7, an evaluation of the links generated by each system was required. QCEW data experts were enlisted to review some of the results and qualify them as good links, bad links, or indeterminable. Since the manual review process requires multiple analysts and a significant amount of time, only the links from Alabama and a representative sample of 161 links from New York were graded.

To score the results, a link received one point if the data experts identified it as a good link, zero if they identified it as a bad link, and a half-point if there was not enough information to confidently judge the link either way. For the New York review, the sample weights were incorporated into the accuracy estimates.

Results of the expert review are provided in Table 8. The figures in these tables apply only to Alabama and New York during the reference period and cannot be generalized.

Table 8: Estimated Linkage Accuracy

State	*System*	*Estimated Good Links*	*Estimated Bad Links*	*Total Links*	*Accuracy Rate*
Alabama	WM	144.5	12.5	157	92.0%
Alabama	AutoMatch	102	24	126	81.0%
New York	WM	699	99	798	87.6%
New York	AutoMatch	396	1,688	2,084	19.0%

According to expert review, the WM system shows improvement over the AutoMatch system in both Alabama and New York. The difference in accuracy level is particularly striking in New York, with the new system scoring nearly 70% better in the accuracy estimates.

Besides increased accuracy, the WM system also identifies a larger number of good links, despite generating a smaller number of total links in these two states. Considering these factors, the low overlap rates observed in Table 7 are more reassuring than alarming. Particularly in New York, about 80% of the links generated by AutoMatch are false positives. It is a desirable result that the WM system ignores most of them.

Analyzing subgroups provides further detail about the behavior of the two linkage systems. Table 9 presents estimated accuracy rates for links generated by both systems and links generated by only one system.

Table 9: Estimated Linkage Accuracy, Subgroups

State	*System*	*Estimated Good Links*	*Estimated Bad Links*	*Total Links*	*Accuracy Rate*
Alabama	Both	96	2	98	98.0%
Alabama	WM Only	48.5	10.5	59	82.2%
Alabama	AutoMatch Only	6	22	28	21.4%
New York	Both	224	8	232	96.6%
New York	WM Only	474	92	566	83.7%

New York	AutoMatch Only	151	1701	1852	8.1%

An additional component of the success of a linkage system is its power, or percent of good links properly identified. In each state, there is an unknown number of good links missed by each system. Therefore, power is presented in Table 10 as an inequality relative to the combined good links identified by the two systems. Quantifying the difference between each system's true power and the maximum power listed would require further research. The accuracy rates from Table 8 are included to complete the expert review-based profile of the two systems.

Table 10: Maximum Linkage Power

State	*System*	*Links Identified*	*Total Links*	*Power ≤*	*Accuracy Rate*
Alabama	WM	144.5	150.5	96.0%	92.0%
Alabama	AutoMatch	102	150.5	67.8%	81.0%
New York	WM	698	849	82.2%	87.6%
New York	AutoMatch	625	849	73.6%	19.0%

4.3 Business Employment Dynamics

Of interest to the QCEW program is Business Employment Dynamics (BED) data, which "consist of gross job gains and gross job losses" and "help to provide a picture of the dynamic state of the labor market," according to descriptions given at the BED website. Openings and closings would be impacted by the difference in the number of links generated by the two systems for a given state. Further, total employment and wages associated with continuous, opening, and closing establishments would be affected. The effects of the new linkage system on BED data for seven test states is given in Table 11. Since employment-level differences are consistent among months within a quarter, the impact on BED employment is listed for only the third month (March 2013) of the reference quarter used in this research.

Table 11: Impact of New Linkage System on BED Data, Statewide Level

State	*Change in Continuous Establishment Count*	*Change in Continuous Employment*	*Change in Continuous Quarterly Wages*
Alabama	31	225	2,770,627
California	209	6,580	219,282,643
Delaware	8	303	3,042,430
Florida	-86	2,729	28,999,312
Georgia	-1,255	-685	133,667,869
New York	-1,286	-499	28,592,013
Texas	-58	6,710	56,305,671
Total	-2,437	15,363	472,660,565

For the test states in the reference quarter, the WM system would generate fewer links, which would result in more openings and closings. However, the total employment and wages associated with those openings and closings would be decreased compared to AutoMatch, since the WM links have higher average employment and wages.

More research will be conducted to thoroughly assess the BED impact of changing linkage procedures. Considering the improvements in accuracy and power of the WM system observed thus far, the BED data should also be improved.

5. Summary and Conclusions

According to initial results, the WM linkage system developed at BLS offers many advantages to the AutoMatch software for use in the QCEW program. Besides cost savings, the WM system outperforms AutoMatch when linking QCEW business establishments between quarters, showing improvement in both accuracy and power. It can be fully controlled and adjusted to meet the specific needs of QCEW data.

To thoroughly evaluate the impact of replacing AutoMatch, the WM system will be run on all states over several years. Linkage quality and the full effects on BED data will be analyzed. Adjustments to the scoring cutoffs and variable thresholds will be evaluated with respect to linkage accuracy and power. Provided that the results of this expanded research are as similarly positive as for the test states, the WM system could significantly improve record linkage in the QCEW program, while offering substantial cost savings to BLS.

References

Current Employment Statistics website. *www.bls.gov/ces*. Bureau of Labor Statistics. Regularly revised.

Thomas, B. 1999. Probabilistic record linkage software: a Statistics Canada evaluation of GRLS and AutoMatch. In *SSC Proceedings*, Survey Methods Section. 187-192.

QCEW Operating Manual. Bureau of Labor Statistics. Regularly revised.

van der Loo, M.P.J. 2014. The stringdist package for approximate string matching. *The R Journal*, Volume 6/1. 111-122.

Business Employment Dynamics website. *www.bls.gov/bdm*. Bureau of Labor Statistics. Regularly revised.