Record Linkage Methodology in Longitudinal Database of Quarterly Census of Employment and Wages

Marek W. Kaminski, Vinod Kapani
Bureau of Labor Statistics
2 Massachusetts Ave. NE, Suite 4985,
Washington, DC 20212

Abstract

The Longitudinal Database of Quarterly Census of Employment and Wages (QCEW) links quarterly reports of employment and total wages of all nonfarm establishments covered by State Unemployment Insurance. QCEW uses two types of linkage: deterministic and probabilistic. Deterministic linkage is performed by SESA IDs – a combination of state FIPS code, Unemployment Insurance Number, and Reported Unit Number. The remaining establishments, those not linked by SESA IDs, are linked by a probabilistic method (weighted matching), based on the Fellegi and Sunter theory. This paper performs and evaluates both deterministic and weighted linkage. Recommendations for weighted linkage are given.

Keywords: Deterministic Record Linkage, Probabilistic Record Linkage, EM algorithm, Accuracy of Linkage, Missing Data

Disclaimer

Views express in this document are those of the authors and do not necessarily reflect the views or policies of the Bureau of Labor Statistics.

1. Introduction

The Quarterly Census of Employment and Wages (QCEW) is a quarterly census of all US establishments subject to state unemployment insurance taxes. QCEW data are collected by each state on a quarterly basis through the production of the Enhanced Quarterly Unemployment Insurance (EQUI) file. The states produce the EQUI file approximately five months after the end of the reference quarter. This file contains records for all business establishments for the applicable reference quarter and updates transactions for prior quarters. EQUI records include the following information:

1. Unemployment Insurance (UI) Account Number, Reporting Unit Number (RUN)
2. Predecessor's UI and RUN
3. Successor's UI and RUN
4. EIN (Employer ID Number)
5. Trade Name
6. Legal Name
7. Addresses – two lines of address are used, Address Line 1 and Address Line 2
8. Phone number
9. North American Industry Classification System (NAICS)
10. Geo-coding Information, geo latitude, geo longitude
11. Monthly Employment and Quarterly Wages

A part of a typical (EQUI) records might look as follows:

		predecessor		successor				
ui acct	rpt unit	ui acct	rpt unit	ui acct	rpt unit	ein	legal name	trade name
0006999999	00028					949999999	Johns Roofing	John Smith
0091299999	00000	000229999	00001			459999999		Zero Buffet
0009199999	00000			000338888	00012	348888888	8 in 1 Pat	Xpat Group

Fig. 1 Example of 3 records shows some of the fields related to this article

Each state's EQUI file is processed through the BLS –Washington edit system, which produces sets of errors reports. These reports are then sent to states which "clean" their data. States produce updated EQUI files which are sent to Washington. Updated files are linked by the National Office to create the QCEW longitudinal data base. Two consecutive quarters are linked together; the first quarter in this paper is referred to as the *previous quarter* and the second quarter is referred to as *current quarter*.

LDB processing is intended to link continuous establishments over time, regardless of any transfer of ownership, location, changes of employment, or changes of type of business. Linkage is a critical part of LDB creation. This linkage process is designed to accurately link the maximum number of establishments.

Linkage is performed in 9 steps:

1. Identification of SESA ID Links
2. Identification of inter-quarter predecessor/successor code links
3. Identification of inter-quarter breakouts/consolidations
4. Identification of inter-quarter UI breakouts/consolidations
5. Identification of inter-quarter weighted matches
6. Identification of intra-quarter predecessor/successor links
7. Identification of intra-quarter breakouts/consolidations
8. Analyst review matching process at state, regional, national level
9. Fully-imputed single units links

Previous and current quarter records which were not linked in one step are passed to the next step for linkage. Records from the current quarter, except the *fully imputed units*, participate in all 8 steps of linkage. The definition of "fully imputed units" is given in reference (1). Fully imputed units are linked to the records from the previous quarter in the last 9-th step.

A primary identifier for linkage is concatenation of three numbers (state-fips, ui account, rpt unit) called a SESA ID. More than 95% of all linkages are one-to-one linkages with the use of the SESA ID. This first step of linkage is called a SESA ID Link. While most establishments maintain the same identifying information over time, establishments that are sold from one owner to another typically experience a

change to their primary identifier, the SESA ID. The primary identifier can also be changed due to a breakout or consolidations of establishments.

Steps 2 through 8 are designed with the purpose of linking establishments when the SESA ID is changed. Records from the previous and the current quarter which are not linked in the first step are put through the linkage process by use of predecessor and successor SESA ID fields in steps 2, 3, 4, 6, 7. Predecessor and successor fields can be used to define one-to-many (that is, breakout) or many-to- one (that is, consolidation) linkages. Linkage via predecessor and successor fields can be performed between prior and current quarter (inter), or within current quarter (intra), depending on the individual state's record of when the transfer of ownership occurred. The detailed description of each of these steps and the process for performing them is given in reference (1).

The fifth step is a probabilistic linkage (weighted matches) that is about 0.2% of total linkage. Probabilistic linkage is the only step in which none of SESA identification numbers are used. The purpose of this step is to perform links which cannot be established via SESA variables. To determine a probabilistic linkage, the proposed method needs to incorporate the use of other fields such as trade_name, ein, naics, etc.. Currently, the QCEW uses Automatch/Vality software.

2. Goal

The purpose of this research is to examine all steps of linkage and to improve the current probabilistic linkage and reduce the internal costs.

3. Methodology and Results

3.1 Linkage by SESA ID Fields – Deterministic Linkage

Linkage by SESA ID fields is a type of deterministic linkage. A deterministic linkage is performed in steps 1 through 7 as detailed in the introduction. Currently, once linkage is performed it is then final. In our research, a deterministic linkage was performed on LDB using nationwide data for the first quarter of 2007 to the second quarter of 2007. The count of all links found for each step was tabulated. The count of unlinked records from both quarters was likewise tabulated. Table 1 below summarizes the tabulated results. The first two columns detail the number of units in the previous and current quarter for each matching process. The third and the fourth column detail number of links obtained in previous and current quarter. If the links are one-to-one matches then the number of links in the previous quarter will be equal to the number of links in the current quarter. If the links are one-to-many, or many-to-one, then number of links in the previous quarter will be different than the number of links in the current quarter. Thus the total number of links may not be equal to the total number of establishments which are to be linked. In some (very rare) cases an establishment can be linked between prior and current quarters (inter linkage) as well as in the current quarter (intra linkage).

Table 1

	Number of Units		Number of Links		Percent of Links	
Linkage_type	previous quarter	current quarter	previous quarter	current quarter	previous quarter	current quarter
SESA ID Linkage	8,209,982	8,209,982	8,209,982	8,209,982	95.9595	95.8893
Inter-quarter Predecessor Linkage	24,578	24,578	24,578	24,578	0.2873	0.2871
Inter-quarter Successor Linkage	334	334	334	334	0.0039	0.0039
Inter-quarter Breakout Identification	961	4,906	976	4,906	0.0114	0.0573
Inter-quarter Consolidation Identification	162	63	162	63	0.0019	0.0007
UI Breakouts	697	4,296	728	4,296	0.0085	0.0502
UI Consolidations	2,003	532	2,003	553	0.0234	0.0065
Weighted Linkage or Analyst Review(*)	14,297	14,297	14,297	14,297	0.1671	0.167
Intra-quarter Predecessor Identification	14,159	14,159	14,159	14,159	0.1655	0.1654
Intra-quarter Successor Identification	3,873	3,873	3,873	3,873	0.0453	0.0452
Intra-quarter Breakout Identification	125	548	129	548	0.0015	0.0064
Intra-quarter Consolidation Identification	149	49	149	53	0.0017	0.0006
Fully Imputed Single Units	284,300	284,300	284,300	284,300	3.3229	3.3205
Unmatched Units	272,140	253,867	0	0	0	0
Total Number of Units	8,809,156	8,808,183	8,555,670	8,561,942	100	100

(*) In this case, the number of Weighted Links or Analyst Review is estimated as a difference between the total number of links performed by LDB number and the total number of links performed by SESA ID fields in all 8 steps of deterministic linkage.

Looking at the last column of table 1, the most important observation to be made is that the one-to-one deterministic linkage performed through SESA ID fields constitutes approximately 99.2% of all performed links. Conversely all breakouts and consolidations, that is, the one-to-many and many-to- one linkages, that are performed via SESA ID fields account for only about 0.67% of all performed linkages, and weighted links account for only 0.17%.

3.2 Evaluation of Accuracy of the first step of linkage - SESA ID Linkage

SESA ID Linkage is the first and most important step of linkage in the current linkage methodology. As mentioned in the introduction, linking in this first step is performed with a linking key defined as the concatenated vector of three fields (state-fip, ui account, rpt unit). All units from the current quarter can participate in this step, with the exception of *fully imputed units*. The linking process is a one to one match. As seen in Table 1 above, this initial step accounted for the overwhelming majority (95.89%) of all linked matches in the simulation of linkage performed on QCEW-Longitudinal Data Base. Thus, the evaluation of the accuracy of this initial linkage step would be a vital part for any investigation and research on the current linkage methodology.

The evaluation of the accuracy of the SESA ID Link step EQUI data was performed using data from one small and one large state between the third and the fourth quarter of 2009. In order to evaluate the accuracy of the SESA ID linkage, an estimation of the number of mismatched records was performed by

attempting to identify *possible mismatches*. A possible mismatch is identified by comparison with fields other than the SESA ID. For example, if two records linked by SESA ID have different values in fields such as *legal_name*, *trade_name*, *ein*, *naics*, *phone_num* etc. then it is possible that the records may have been mismatched. Although having one or two fields different might not strongly indicate the occurrence of a mismatch, having six or more fields different is definitely a stronger indication of a possible mismatch. For the purpose of our investigation of possible mismatches the following definition of strong disagreement in a field was used:

Definition of Strong Disagreement in Field

For a given linked pair of records we say that the linked pair strongly disagrees in this field if the following two conditions are satisfied for a given field:

1. The values are not missing in this field in either element of pair
2. The values are not identical

For example, if *trade_name* in the previous quarter is *David Lee* and in the current quarter it is *Lee* then there is strong disagreement in trade_name. If in the previous quarter it is *David Lee* and in the current quarter the value is missing, then strong disagreement cannot be detected. Thus, a strong disagreement in a given field can be detected only if the values are not missing and are not identical.

Note that having no strong disagreement in a field does not mean having agreement. The definition of *agreement in the field* is given in the next section and it is NOT exactly logically opposite to strong disagreement in the field.

Table 2 presents statistics for some of combinations of strong disagreements, denoted by 0, and not strongly disagreements denoted by 1. The following fields were used: legal_name, trade_name, phone_num, addr_line1, geo_longitude, and geo_latitude. Records were linked by SESA ID Link between third quarter of 2009 and fourth quarter of 2009 for the same two states.

Table 2 Some combinations of strong disagreement and not strong disagreements
strong disagreement (= 0) and not strong disagreement (=1)

state fip	prior quarter	SESA ID linked pairs	legal_name=0 trade_name=0 phone_num=0 addr_line1=0 geo_long=1 geo_lati = 1	legal_name=0 trade_name=1 phone_num=1 addr_line1=1 geo_long=0 geo_lati = 0	legal_name=0 trade_name=1 phone_num=1 addr_line1=1 geo_long=n/a geo_lati = n/a	legal_name=0 trade_name=0 phone_num=0 addr_line1=0 geo_long=0 geo_lati = 0	legal_name=0 trade_name=n/a* phone_num=n/a* addr_line1=1 geo_long=1 geo_lati=1	legal_name=0 trade_name=1 phone_num=1 addr_line1=1 geo_long=1 geo_lati=1
1	20093	35,624	0	0	3	0	5	3
2	20093	265,573	4	197	437	6	330	237

(*) Here the n/a in a given field means that this field was used for comparison.

In Table 3, columns 1 and 4 show that there were only 4 pairs such that strong disagreement was detected in legal_name, trade_name, phone_num, addr_line1. There were only 6 pairs such that legal_name, trade_name, phone_num, addr_line1, geo_long, geo_lati strongly disagree. These 10 pairs were selected

for inspection. None of the pairs was determined to be a certain mismatch. The strong disagreements were due to misspelling of names, use of different abbreviations, use of abbreviations with no abbreviations, use and no use of commas. Thus, in over 300,000 cases of linked pairs by the SESA ID link, the method identified not even one pair that was determined to be a mismatch.

Therefore, it is reasonable to estimate that SESA ID Linkage produces 0 mismatches. The following estimation can be made:

(1) P(match / SESA ID = 1) = 1 and P(mismatch / SESA ID = 1) = 0

The conditional probability of SESA ID = 1 given match can be estimated from the Bayes' theorem:

(2) $$P(\,SESA\,ID = 1 \mid match) = \frac{P(match \mid SESA\ ID=1)P(SESA\ ID=1)}{P(match|SESA\ ID=1)P(SESA\ ID=1)+P(match|SESA\ ID=0)P(SESA\ ID=0)}$$

From Table 1:

1. P(SESA ID =1) = 8,209,982/8,808,183 = 0.932086, as 8,209,982 units in the current quarter were matched by SESA ID Linkage, and 8,808,183 units existed in the current quarter.
2. P(SESA ID = 0) = 1 – 0.932086 = 0.067914
3. P(match / SESA ID =0) = P(links performed by Weighted Method and Analyst review Matches) =0.00167

(3) $$P(\,SESA\,ID = 1 \mid match) = \frac{1*0.932086}{1*0.932086+0.00167*0.067914} = 0.999878 \approx 1$$

Again using Bayes' theorem and P(unmatch / SESA ID = 1) = 0 it follows that

$$P(\,SESA\,ID = 1 \mid mismatch) = 0$$

Thus, the first step of linkage, SESA ID Linkage, can be considered as error free. This is very fortunate and useful result.

4. Probabilistic linkage

Probabilistic linkage can be done in several steps; the first step is finding initial *m* and *u* probabilities.

4.1 Finding initial *m* and *u* probabilities

Probabilistic Linkage is performed in several steps. The first step is finding initial *m* and *u* probabilities. To this end, following definitions are given:

Let $\gamma^s = (\,\gamma_1^s,\ \gamma_2^s\ ,\dots,\gamma_n^s\,)$ denote a *vector of strong agreements* for a given pair of records with *n* fields. Assume $\gamma_i^s = 1$ if there is strong agreement in field *i* , and assume $\gamma_i = 0$ if there is no strong agreement in field *i* . Let M denote a set of pairs which are true matches, and let U denote a set of pairs which are all possible mismatches. Let *m* and *u* denote conditional probabilities defined as follows:

(4) $m(\gamma_i^s) = P(\gamma_i^s|M)$, $u(\gamma_i^s) = P(\gamma_i^s|U)$ $i = 1, 2, \dots, n$

$m(\gamma^s) = P(\gamma^s|M)$, $u(\gamma^s) = P(\gamma^s|U)$

Let p denote probability of match, i.e. $p = P(M)$.

The result from the previous section that the first step of linkage, SESA ID Linkage, is error free has strong implications in probabilistic linkage. Because of this result SESA ID linkage is used as a way to determine the true match and mismatch status for all pairs which can be linked via SESA ID. The first step of QCEW linkage performs over 95% of all links; hence there is an abundance of available links which can be used for estimation of initial *m* and *u*. That is a very desirable situation, since it will produce estimates of *m* and *u* that are accurate. The *m* and *u* probabilities are estimated using frequencies of strong agreements on a set of pairs linked by the SESA ID. The following estimates of *m* and *u* are called *initial*, since later on they will be refined through the Estimation and Maximization (EM) algorithm. Let the initial *m* probability be defined as:

(5) $$m\_initial(\gamma_i^s = 1) = \frac{\#\{\gamma_i^s = 1 \ and \ SESA \ ID \ link\}}{\#\{SESA \ ID \ link\}} \qquad i = 1, 2, \dots, n$$

similarly, initial *u* probabilities are defined:

(6)

$$u\_initial(\gamma_i^s = 1) = \frac{\#\{\gamma_i^s = 1 \ and \ no \ SESA \ ID \ link\}}{\#\{records \ in \ previous \ quarter\} \times \#\{records \ in \ current \ quarter\} - \#\{SESA \ ID \ link\}}$$

and initial *p* probability is defined:

(7) $$p_{initial} = \frac{\#\{SESA \ ID \ link\}}{\#\{records \ in \ previous \ quarter\} \times \#\{records \ in \ current \ quarter\}}$$

where:

$\#\{SESA \ ID \ link\}$ is the total number of all SESA ID links

$\#\{records \ in \ previous \ quarter\} \times \#\{records \ in \ current \ quarter\}$ is the total number of all possible links

$\#\{\gamma_i = 1 \ and \ SESA \ ID \ link\}$ is the total number of pairs which are linked by SESA ID and and agree on *i*, field,

$\#\{\gamma_i = 1 \ and \ SESA \ ID \ no \ link\}$ is the total number of pairs which strongly agree on *i* field, and can't be linked via SESA ID,

$\#\{records \ in \ previous \ quarter\} \times \#\{records \ in \ current \ quarter\} - \#\{\gamma_i = 1 \ and \ SESA \ ID \ link\}$

is total number of possible pairs which cannot be linked via SESA ID.

Table 3 details the initial *m* and *u* probabilities as computed for individual fields. Linkage was performed on EQUI data for one of the largest states between the fourth quarters of 2009 to the first

quarter of 2010. The total number of units in the previous quarter, which were subjected to linkage was 1,544,121, and in the current quarter 1,566,386. The total number of SESA ID links which were made was 1,527,361.

Table 3 initial *m* and *u* probabilities

	Number of nonmissing entries in 20094	Number of nonmissing entries in 20101	Number of pairs linked by SESA ID that strongly agree in the field	m prob	u prob	weight - log(m/u)
legal_name	1,544,078	1,565,626	1,520,990	0.9958	3.3833E-08	17.20
trade_name	245,147	251,377	236,398	0.1548	1.8484E-08	15.94
addr_line1	1,544,029	1,565,072	1,432,422	0.9378	2.6009E-07	15.10
addr_line2	1,894	1,868	1,849	0.0012	5.0027E-10	14.70
Town	1,543,268	1,564,881	1,500,700	0.9825	6.0768E-07	14.30
State	1,543,454	1,564,560	1,523,549	0.9975	6.4480E-07	14.25
Zip	1,543,984	1,565,106	1,495,345	0.9790	5.8764E-07	14.33
zip and zip_ext (*)	937,864	1,035,303	910,983	0.5964	2.3771E-07	14.74
Cnty	1,544,121	1,566,386	1,514,470	0.9916	6.4396E-07	14.25
phone_num	992,470	1,003,635	951,646	0.6231	3.4139E-08	16.72
geo_location	793,752	763,874	763,458	0.4999	3.2778E-07	14.24
geo coordinates (**)	1,402,226	1,373,850	1,236,461	0.8095	5.5472E-08	16.50
Naics	1,544,121	1,566,386	1,507,177	0.9868	6.2830E-07	14.27
Ein	1,518,653	1,538,819	1,489,048	0.9749	2.1545E-07	15.33

(*) here zip and zip ext is a vector of two numbers (zip, zip_ext)
(**) here geo coordinates is a vector of two numbers (geo_latitide, geo_longitude)

There is an obvious relationship between the number of non-missing entries and the number of strong agreements. Since strong agreement can be determined only if both compared values are non-missing, the number of strong agreements is always less than the minimum of the number of non-missing entries in previous and current quarter, as indicated by the following formula:

(8) $$\#\{strong\ agreements\} \leq \min\left(\#\{nonmissing\ entries\ prev.quarter\},\ \{nonmissing\ entries\ curr.quarter\}\right)$$

Since *m* probabilities, which describe a power to link by a given field, depend on the number of strong agreements and strong agreements depend on non-missing data, the probabilities *m* depends on non-missing data. Once initial *m* and *u* probabilities are determined for every field they are used to obtained

refined *m* and *u* probabilities through the EM algorithm. Having accurate initial estimates for vales *m*, *u* and *p* speeds up EM algorithm.

A common technique used in probabilistic linkage is blocking. Blocking entails linking by a predefined set of fields before probabilistic process takes place. Blocking produces a set of pairs which are linked by a *blocking field* or a *blocking vector of fields*. The following definition of blocking is used:

Definition of Blocking by Field
A set of pairs which **strongly agree** in a given field, is called blocked by this field. A process of selecting pairs which are blocked in field is called blocking.

Similarly, blocking by a vector of fields is defined. For blocking by a vector of fields the requirement is that records strongly agree for all fields in a vector.

The blocking technique for probabilistic linkage is well described in literature. See references (2) and (3).

We use blocking for two reasons:
1. In order to speed up computation performed by EM algorithm
2. In order to predefine types of links which are acceptable and allowing only these type of links

First reason is well explained in the literature (look refrence 3). The second reason comes from the fact that some probabilistic links may not be acceptable. Probabilistic linkage might link an ice cream shop to a shoe store based on the same phone number, and location. These situations can happen since the data used is not without errors. In order to prevent some of the worst mistakes, blocking may be of help. Blocking predefines which links are to be considered for matches, and which links are not considered. Well defined blocking can improve the accuracy of the matching process without affecting the total number of matches. For this particular application, the LDB Record Linkage, the accuracy of linkages is by far the most important aspect of the blocking technique.

4.1. Finding Fields Best Suitable for Blocking

An important observation is that weights, as they are computed for each field, cannot be used by themselves to determine which fields should be used for blocking. For example, let's consider the weight computed from initial probabilities (such as those in table 3) for the field phone_num is 16.72. This is one of the highest weights, but it is non-missing, only for 992470 units in the previous quarter, which is 64.27% of total number of units. As a result the *m* value is only 0.6231. If only one block was used phone_num, it could have missed approximately 35% of possible links. On the other hand, using a field with high *m* value but low weight can result in too many links. Too many links may create a heightened opportunity for many mismatches. For example, county (cnty) has very high value of *m* (0.9916), but its weight is relatively low (14.25). Hence, if cnty was used only by itself, it may allow too many mismatches. Thus, choosing appropriate fields for blocking has to be done with careful consideration for all parameters. In the above example for the large state, the following fields (legal_name, ein, addr_line1, zip, and phone_num) seem to be most suitable for blocking. These fields can be considered as *identifiers* of a valid link. Whereas, fields such as: naics, cnty, geo_location, geo_coordinates, town, state (zip,

zip_ext) can serve as *verifiers*. Looking at the results from the small state one more field can be used as a good identifier: trade_name.

Fields which are strong identifiers are used for blocking whereas fields which are verifiers are used for probabilistic verification. In order for a match to be determined as "true" it has to pass through both identification and verification. The Identification process is manually predefined and performed in blocking steps, whereas probabilistic verification is done through a program.

The proposed QCEW linkage procedure is a matching routine where blocking is performed 7 times using different blocking fields. Previous and current quarter records which are not linked in one block are then passed to the next block for linkage. Blocking is performed always by at least 2 fields, that is, two fields are used as one single blocking vector. At least two of the fields must have the property of being a strong identifier. A set of pairs of records which are linked in each block are subjected to probabilistic linkage. The proposed system is made of the following blocks:

1. (ein, naics, phone_num)
2. (ein, naics, trade_name)
3. (ein, naics, legal_name)
4. (phone_num, trade_name)
5. (phone_num, legal_name)
6. (addr_line1, naics, trade_name)
7. (addr_line1, naics, legal_name)

Blocks of pairs produced in each step of blocking are subsequently put through probabilistic linkage.

4.3. Execution of Probabilistic Linkage

Sets of linked pairs produced in each block are subjected to the probabilistic linkage. In this step all links are divided into 3 sets: 1) a set of pairs which are verified to be matches, 2) a set of pairs which are possible matches (these links are to be examined by a human), and 3) a set of pairs which are verified to be mismatches. A set of pairs which is verified to be mismatches is later to be divided into a set of single previous quarter records, and a set of single current quarter records.

There are 3 steps in the execution of probabilistic verification.

- EM algorithm – performed on blocks of pairs to find refined *m* and *u* probabilities
- Computation of score for each linked pair
- Dividing records into 4 sets, as described above

4.3.1. EM algorithm

EM algorithm is used to find the Maximum Likelihood Estimator (MLE) for *m* and *u* probabilities. The general discussion of the EM algorithm is given in references (2) and (3). For the purpose of performing EM algorithm, the following definition of agreement and disagreement in a field is used.

Definition of Agreement and Disagreement in Field

For any pair of records, if there are identical values in a given field in each of these records then we say that these records agree in this given field. Otherwise we say that the pair of records disagrees in this given field.

Note that in this definition there is no requirement that the data are non-missing. Therefore, if data are missing in both records in some field, then these records agree in this field. Similarly, as in the definition of strong agreement, the following definitions and terms are used: Let $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_n)$ denote a vector of agreement and disagreement values for a given pair of records with *n* fields. Here assume $\gamma_i = 0$ if there is disagreement in field *i*, and assume $\gamma_i = 1$ if there is agreement in field *i* .

After blocking is performed the EM algorithm is run on a block of pairs. Through the EM algorithm the refined *m* and *u* probabilities are then found. Since the EM algorithm is performed only on blocks, the *m* and *u* values obtained are actually conditional probabilities with conditions: {*match* and *blocking*} and {*mismatch* and *blocking*}. Let m^b and u^b be conditional probabilities defined as follows:

$$m^b(\gamma_i) = P(\gamma_i | M \text{ and } block) \qquad u^b(\gamma_i) = P(\gamma_i | U \text{ and } block)$$

If field *i* is used for blocking, then the obvious equations hold:

$$m^b(\gamma_i = 1) = 1 \quad \text{and} \quad u^b(\gamma_i = 1) = 1 \qquad \gamma_i \in block$$

In this case, the EM algorithm is using initial *m* and *u* probabilities obtained from the previous step with the sole difference that if field *i* is used for blocking then $m^b(\gamma_i) = 1$ and $u^b(\gamma_i) = 1$ as noted before. For example, the initial probabilities for block defined by fields (ein, naics, phone_num) for one of the largest states from the table 3 are:

Table 4 Example of initial *m* and *u* probabilities for the block (ein, naics, phone_num)

	legal_name	trade_name	addr_line1	addr_line2	town	state	zip
m - prob	0.9958	0.1548	0.9378	0.0012	0.9825	0.9975	0.979
u - prob	3.38E-08	1.85E-08	2.60E-07	5.00E-10	6.08E-07	6.45E-07	5.88E-07

	zip and zip_ext	cnty	phone_num	geo_location	geo coordinates	naics	ein
m - prob	0.5964	0.9916	1	0.4999	0.8095	1	1
u - prob	2.38E-07	6.44E-07	1	3.28E-07	5.55E-08	1	1

Convergence of EM algorithm

EM algorithm performs iterations until convergence is achieved. Convergence is achieved when

(8) $\sum_{i=1}^{n}\left[m_j^b(\gamma_i = 1) - m_{j-1}^b(\gamma_i = 1)\right]^2 + \sum_{i=1}^{n}\left[u_j^b(\gamma_i = 1) - u_{j-1}^b(\gamma_i = 1)\right]^2 + (p_j^b - p_{j-1}^b)^2 \leq 10^{-4}$

where subscript *j* denotes the number of iterations. With the use of initial probabilities, as defined above, the algorithm converges very rapidly. In the test data from a 3 different states the convergence was achieved in less than 40 iterations.

After *m*, *u* and *p* values are found through the EM algorithm, for the next step weights for each field are then calculated.

4.3.2. Computation of weights

Weights for individual fields are computed according to the following formulas:

1. In a case of agreement in field i **:**
   If $m_i, u_i > 0$ then $w_i = log\left(\frac{m_i}{u_i}\right)$
   If $m_i = u_i = 0$ then $w_i = 0$
   If $m_i > 0$ *and* $u_i = 0$ then $w_i = 20$
   If $m_i = 0$ *and* $u_i > 0$ then $w_i = -20$
2. In a case of disagreement in field i :
   If $m_i, u_i < 0.99999999$ then $w_i = log\left(\frac{1-m_i}{1-u_i}\right)$
   If $m_i > 0.99999999$ *and* $u_i > 0.99999999$ then $w_i = 0$
   If $m_i < 1$ *and* $u_i > 0.9999999999$ then $w_i = 20$
   If $m_i > 0.99999999$ *and* $u_i < 1$ then $w_i = -20$

The score is a sum of all w_i's for *n* fields, as in the following formula

(9) $score = \sum_{i=1}^{n} w_i$

The score is to be computed for every paired linked in a block.

4.3.3. Finding Matches, Possible Matches, and Mismatches – Errors μ and λ

The block of pairs is divided into 3 subsets: 1) Matches 2) possible matches and 3) mismatches. These are denoted as subsets A_1, A_2, and A_3, respectively. Possible matches are to be reviewed manually. The determination of matches or mismatches is to be the final outcome of the process. If we define M to be a set of pairs which are true matches, and U to be a set of pairs which are mismatches then two types of errors can likewise be defined:

1. Not matching records which are in fact actual matches – that is, A_3 while M
2. Matching records which are not in fact actual matches – that is, A_1 while U

If we assume the null hypothesis H_0 that a given pair is a true match, then using the language of hypothesis testing:

A_3 while M is error type I, and

A_1 while U is error type II.

Let $\mu = P(A_1|U)$, and $\lambda = P(A_3|M)$ be probabilities for these two types of errors. The selection process between sets A_1, A_2, and A_3 depends on μ and λ values. The values μ and λ are chosen before the selection process takes place. For the purpose of this research we have defined $\mu = 10^{-7}$ and $\lambda = 10^{-5}$.

In order to find sets A_1, A_2, and A_3 the following procedure is undertaken on the set of blocked pairs:

1. From the set of linked records a subset of records which contain only unique combination of γ is selected. In this case, there are N_γ unique combinations of γ.
2. Records are sorted by the values of their weights. The record with the greatest value of weight is the first record and the record with smallest value of weight is the last record.
3. For each pair of records *j*, values $m_j(\gamma) = P(\gamma|M)$ and $u_j(\gamma) = P(\gamma|U)$ are estimated by product of *m* and *u* probabilities

(10) $\quad m_j(\gamma) = \prod_{i=1}^{n} m_j(\gamma_i) \qquad u_j(\gamma) = \prod_{i=1}^{n} u_j(\gamma_i)$

where $m_j(\gamma_i)$ is *m* probability corresponding to *j* pair of records and *i* field, $u_j(\gamma_i)$ is *u* probability corresponding to *j* pair of records and *i* field.

4. For each pair of records, two sums are computed s_k^1, s_k^2. The sum of all *u* values down to the given record and the sum of all *m* values up to the given record. If a record number is *k* then

(11) $\quad s_k^1 = \sum_{j=1}^{k} u_j, \qquad s_k^2 = \sum_{j=k}^{N_\gamma} m_j$ where N_γ is the total number of records.

5. The file with computed s_k^1, s_k^2 values for each unique γ is merged with file of pairs in a block by γ vector. Thus associating to each pair in a block two sums s_k^1 and s_k^2.
6. Let μ and λ be admissible error levels, $\mu < \lambda$. We then divide all records in a block according to the following formulas:

$$A_1 = \{records:\ s_k^1 < \mu \ \ and \ \ s_k^2 > \lambda\}$$
$$A_2 = \{records:\ s_k^1 \geq \mu \ \ and \ \ s_k^2 \leq \lambda\}$$
$$A_3 = \{records:\ s_k^1 > \mu\}$$

By the fundamental linkage theorem, reference (2), we have that the set A_2 is the smallest possible for a given μ and λ. The set A_2 as defined above is simplified in comparison to the definition given by Fellegi and Sunter. A_2 as defined here can contain one extra record however, for all practical purposes, having one extra record in A_2 does not present any problem.

Table 5 present the outcomes of linkage performed on 2 states. The second and the third column are number of units which participate in linkage in the previous and current quarter. Note that the number of units in the current quarter is far smaller than in the previous quarter. It is due to the fact that fully imputed units were removed from the current quarter even before SESA ID Link was performed. The fourth column is the number of pairs which are in a given block. These pairs satisfy strong agreement criteria for all listed fields.

Observe that, in comparison to the number of units which participate in linkage, only very few are linked in a block. One reason that this may happen is due to the problem of missing data. Blocks are formed from pairs linked by strong agreements, which may entail non-missing data for linking variables. The fifth column is the number of linked pairs which are determined by probability linkage as true matches, that is, the A_1 set. The sixed column is the number of possible matches, that is, the A_2 set.

Notice that the total number of pairs linked in a block is here the sum of true matches plus possible matches. This is due only to the very strict linkage criteria used for blocking in this design. If blocking criteria is relaxed then outcomes may be dramatically different. Consider the first listed state linkage, if instead of the blocking criteria used to create the first block (ein, phone_num, naics) there are only two fields used (phone_num, naics), then there is going to be as many as 156 links in this block (instead of 1).

Probabilistic linkage does not have to be one-to-one. Links one-to-many and many-to-one are indeed common. For example, in linkage for the third statein a particular block (addr_line1, trade_name, naics), there were 31 records in the previous quarter which were matched to 3 records in the current quarter, with the total number of pairs being 32. If a one-to- many or many- to- one match occurs, then manual inspection is recommended.

Summary

Examination of all steps of deterministic part of linkage done by SESA ID fields did not resulted in recommendation for a change in the current procedures; proposed in house developed probabilistic linkage method will improve the weighted linkage and minimize the cost of linkage process to BLS.

Table 5 linkage for two states

block	state	num of previous quarter records which entered blocking	num of current quarter records which entered blocking	num of pairs linked in a block	num of matches verified by probabilistic matching	num of possible matches
(ein, phone_num, naics)	1	4877	997	1	1	0
(ein, trade_name, naics)	1	4876	996	2	0	2
(ein, legal_name, naics)	1	4874	994	1	1	0
(phone_num, trade_name)	1	4873	993	1	1	0
(phone_num, legal_name)	1	4872	992	0	0	0
(addr_line1, trade_name, naics)	1	4872	992	0	0	0
(addr_line1, legal_name, naics)	1	4872	992	0	0	0
(ein, phone_num, naics)	2	158729	35583	461	450	11
(ein, trade_name, naics)	2	158475	35212	305	291	14
(ein, legal_name, naics)	2	158314	34972	1206	1201	5
(phone_num, trade_name)	2	157170	34369	90	87	3
(phone_num, legal_name)	2	157083	34328	99	98	1
(addr_line1, trade_name, naics)	2	157024	34245	32	32	0
(addr_line1, legal_name, naics)	2	156993	34242	40	40	0

References

1. Bureau of Labor Statistics, *Summary of Initial LDB Record linkage System*, August 8, 2007
2. Ivan P. Fellegi and Alan B. Sunter, *Theory of Record Linkage*, Journal of the American Statistical Association, Vol. 64, No. 328 Dec., 1969, pp. 1183-1210
3. Thomas N. Herzog, Fritz J. Scheuren, William E. Winkler, *Data Quality and Record Linkage Techniques*, Springer, 2007