A Comparison of Record Linkage Techniques

Lowell G. Mason[1]
[1]Bureau of Labor Statistics, 2 Massachusetts Avenue, Suite 4925, Washington, DC 20212

Abstract
It has become increasingly common to create new statistical products by integrating existing data rather than engaging in new data collection; using existing data sources is less expensive and does not increase respondent burden. However, it is usually not possible to satisfactorily integrate the multiple data sources without manual intervention. An example is the integration of the Bureau of Economic Analysis (BEA) enterprise-level data on Foreign Direct Investment (FDI) with establishment data from the Bureau of Labor Statistic's Quarterly Census of Wages and Employment (QCEW). In this particular case, the initial error rate was 87.7%. After manual review and correction, the error rate was reduced to 19.0%. The labor cost, however, was considerable: almost 1,510.5 hours. To reduce linkage error and labor costs, we implement several record linkage techniques. We consider supervised learning techniques, such as Support Vector Machines (SVM) and Random Forests. Finally, as a baseline comparison, we implement the methods developed by Fellegi and Sunter (1969).

Key Words: Record linkage, integrated data, machine learning

1. Introduction

There have been several instances recently in which enterprise-level from the Bureau of Economic Analysis (BEA) data have been integrated with establishment-level data from the Bureau of Labor Statistic (BLS). For instance, in a pilot study Handwerker, Kim, and Mason (2011) linked BEA data on the top 500 U.S.-based, multinational manufacturing firms to establishment data from the BLS. More recently, BEA data on foreign multinational firms with ownership in U.S. affiliates were merged with BLS establishment data. This work is described in Talan, Mason, and Clayton (2015), and an official news release is expected soon. While there are many benefits of integrating these data sources, there are challenges as well. Primarily, a significant amount of labor is required to manually integrate the data sources. This paper describes the preliminary exploration of automated record linkage in the context of integrating BEA and BLS data sources. Several record linkage classification methods are compared.

The paper is organized as follows: Section 2 provides a detailed examination of the two data sources, and describes the advantages of integrating the data sources. Additionally, the challenges inherent in linking the data sources are described. Section 3 gives a brief overview of the record linkage process. For each step in the process, Section 3 also describes the associated implementation details particular to linking BEA and BLS data. An evaluation of the results is given in Section 4. Finally, concluding remarks and future directions for this work are detailed in Section 5.

2. Data Sources

This section describes the two data sources: BEA inward Foreign Direct Investment (FDI) data and BLS establishment data from the Quarterly Census of Employment and Wages (QCEW). The two data sources complement and augment one another, and integrating the data provides many benefits. However, there are many challenges associated with integrating the data as well.

2.1 Inward FDI data

The BEA data provide a variety of measures on inward Foreign Direct Investment (FDI). BEA collects this data through benchmark surveys every 5 years and samples in the intervening years.

BEA collects data at the enterprise level—the unit of measure is a U.S. firm (commonly referred to as an affiliate) that has at least 10% foreign-ownership. In terms of U.S.-based employer structure, this is the highest level, as seen in Figure 1.

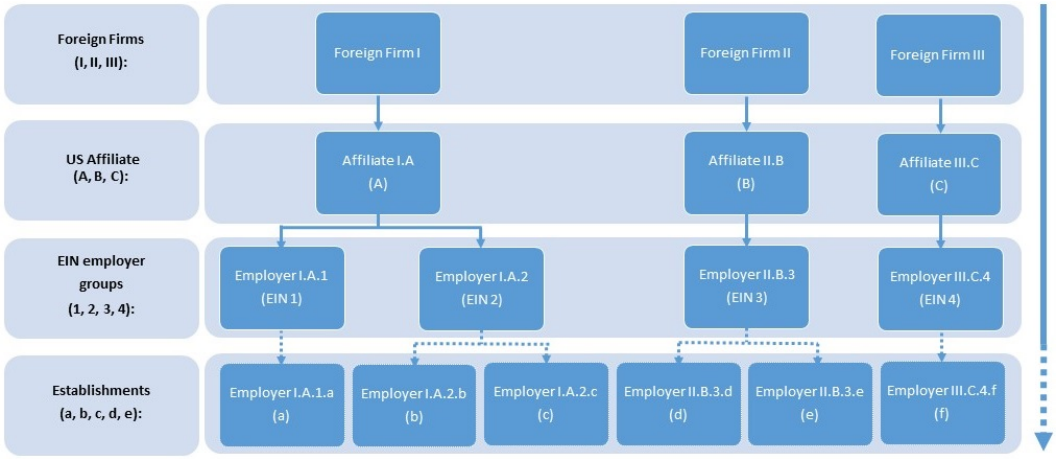

Figure 1: Inward FDI structure

For each U.S. affiliate, data on balance sheets, income statements, goods and services supplied, and employment (broken out by state) and compensation are collected. Additionally, the industrial classification of the affiliate and its foreign parent are provided as well as the affiliate address and contact information. Finally, as affiliates vary in size and complexity, all names and EINs associated with the affiliate and any subsidiaries of the affiliate are collected.

The varying complexity of the affiliates are seen in Figure 1. There are three foreign firms (I, II, and III), each with one U.S. affiliate (I.A, II.B, and III.C). Affiliate I.A has a more complex structure than the others, however. It has two subsidiaries (I.A.1 and I.A.2) as represented by the presence of two EINs. Overall, Affiliate I.A would then have three names: an overall affiliate name, and a name for each of the subsidiaries. Affiliates II.B and III.C, however, would only have one name each (the overall affiliate name).

The arrow on the right-hand side of Figure 1 indicates that knowledge of the affiliate structure flows from the top, down. It also indicates the extent of knowledge of affiliate structure. The dashed lines in Figure 1 denote that the knowledge of establishment structure is unknown. While it is possible to determine if an affiliate has subsidiaries, it is not

possible to determine the establishments associated with these subsidiaries. In fact, that is the primary objective of integrating the BEA and BLS data sources.

The establishments of Affiliate II.B hint at the difficulty in integrating the two data sources. What looks like a simple structure in terms of the data available from BEA is in fact not as simple as Affiliate III.C. Affiliate II.B has multiple establishments as opposed to a single-establishment for Affiliate III.C.

The inward FDI data used for this project are from the BEA 2012 Benchmark Survey. There are 5,684 affiliates in the data, with an associated 33,914 subsidiaries (BEA, 2015).

2.2 QCEW data

The BLS data are from the Quarterly Census of Wages and Employment (QCEW). The QCEW provides a comprehensive measure of U.S. establishments, covering over 95% of U.S. employment. The QCEW also serves as the frame for most BLS surveys, including the Occupational Employment Survey (OES), which provides a measure of occupational distribution.

The QCEW collects establishment employment and total compensation, as well as industrial classification and geographic location. For each establishment, trade and legal names are available, as well as up to three address (physical, mailing, and/or headquarters). Finally, the EIN for the establishment is provided.

The unit of measurement in the QCEW is an establishment, a single physical location where one predominant industrial activity occurs. This is the lowest level in terms of employer structure, as seen in Figure 2.

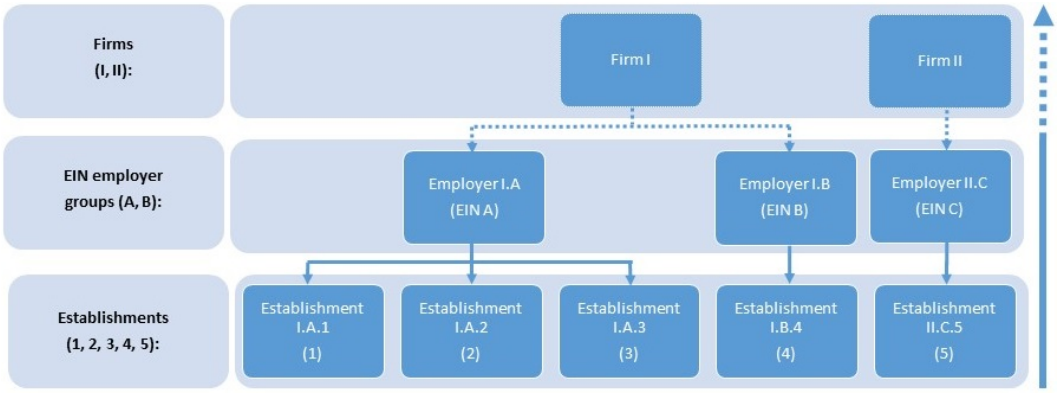

Figure 2: QCEW structure

As opposed to inward FDI, in the QCEW, knowledge of the employer structure flows from the bottom, up. Establishments are given, but can be aggregated to more complex employer structures using the provided EIN. If establishments share an EIN, then these establishments form a multiple-establishment employer, as in Employer I.A. These are opposed to single-establishment employers I.B and II.C.

Similar to inward FDI, the knowledge of the complete firm structure is unknown in the QCEW. The dashed lines in Figure 2 indicate that the knowledge of firm structure is unknown. Firms can be composed of multiple EINs. Employer I.B hints at the difficulty in integrating the two data sources in a similar manner as before. While employer I.B is a

single-establishment employer at the EIN level, it is in fact part of a multi-establishment firm. This is in contrast to Employer II.C, which is both a single-establishment employer (at the EIN level) and a single-establishment firm.

2012 QCEW data were used to link to the BEA data depending on the fiscal date reported by the affiliate (affiliates whose fiscal date was in January, February, or March where matched to QCEW 2012Q1, etc.). Only private establishments in the QCEW were linked to the BEA data. By definition, at Federal, State, or Local government cannot have foreign ownership. For 2012, there were 8,826,016 private establishments in the QCEW (BLS, 2016).

2.3 Advantages of integrating the two data sources

By integrating the two data sources, the BLS establishment data augment the enterprise-level BEA FDI data, allowing a detailed look at the distribution of establishments (across industry or geographic location, for instance) that comprise foreign-owned enterprises, as well as the ability to compare those to domestically-owned establishments. Further, as the QCEW is the frame for the OES, it is also possible to track the distribution of occupations within foreign-owned enterprises.

While integrating the two data sources is not without cost, compared to directly collecting the data in an enhanced survey, the cost savings are significant. For instance, there is no increase in respondent burden when integrating the data.

2.4 Challenges in integrating the two data sources

However, integrating the two data sources is not a straight-forward activity. While common identifiers between the datasets do exist—Employer Identification Numbers (EINs) assigned by the Internal Revenue Service (IRS) for tax purposes—they are extremely noisy. Merging the data sets using just these identifiers results in an error rate—defined as the absolute difference in reported employment between the two data sources—of 87.7%. Handwerker and Mason (2013) propose a number of reasons why EIN numbers prove so ineffective at linking data sources.

To reduce the error rate, analysts reviewed the initial linkages. Invalid EIN linkages were removed. Additionally, any missing linkages were manually added. This involved extensive research and time. After nearly 1,510.5 hours of analyst review, the error rate was reduced to 19.0%.

There are two fundamental reasons why integrating the two data sources is so time-consuming. The first reason stems from the fact that the data sources are at fundamentally different levels in terms of structure. In the case of multiple establishments, it is very inefficient to try and determine if one of the establishment should be linked to a particular affiliate without considering all of the other establishments as well. The second reason is because the type of linkage is not one-to-one in the sense that one affiliate can have multiple subsidiaries (represented by EINs).

To counter this, prior to linking, multi-establishment data is aggregated to the EIN level. For numeric variables, this is a straightforward. For instance, establishment employment is simply summed to give total employment for the EIN, which can be directly compared to the affiliate employment. For categorical variables such as establishment industrial classification, when aggregating to the EIN it is necessary to look at the distribution of each establishment in the EIN across all of the values of the categorical variable. Establishments

can be simply counted or weighted by their employment. This gives a vector (of length equal to the number of classes in the categorical variable) that can be compared to the affiliate categorical variable (also stated as a vector, either with all of the weight in given class or distributed). Finally, for names and addresses, it is necessary to maintain a list of all unique names or addresses for the EIN and each of these can be compared to the names and address for the affiliate.

While aggregating to the EIN level in the QCEW does address the inefficiencies related to linking multiple-establishment EINs, it does not address the fact that an affiliate can be composed of multiple EINs. This can be addressed in record linkage by using a classification method that is collective. In this preliminary study, this issue was not addressed.

3. Record Linkage Process

Record linkage is the process of joining the observations in one or more datasets in the absence of reliable unique identifiers. Record linkage is also commonly referred to as data matching, entity resolution, co-reference resolution, or deduplication.

The goals of record linkage are two-fold. First, it is necessary to accurately and reliably model linkages that fit the data while accounting for uncertainty, and second, to do this efficiently. These goals are contradictory. Accuracy and reliability imply examining all possible pairs of records in the two datasets. However, this is often computationally infeasible for even moderate sized data sources. Furthermore, it is difficult to assess quality and completeness of linked pairs without knowing the true linkage status, and obtaining labeled data to evaluate linkage quality and completeness is expensive. The ideal is "Gold-standard" labels that are double-blind coded, with disagreements adjudicated by a third coder. As an example of the expense, almost 1,510.5 hours where needed to integrate the BEA inward FDI dataset to the QCEW.

The record linkage process consists of a number of steps, as visualized by the following diagram:

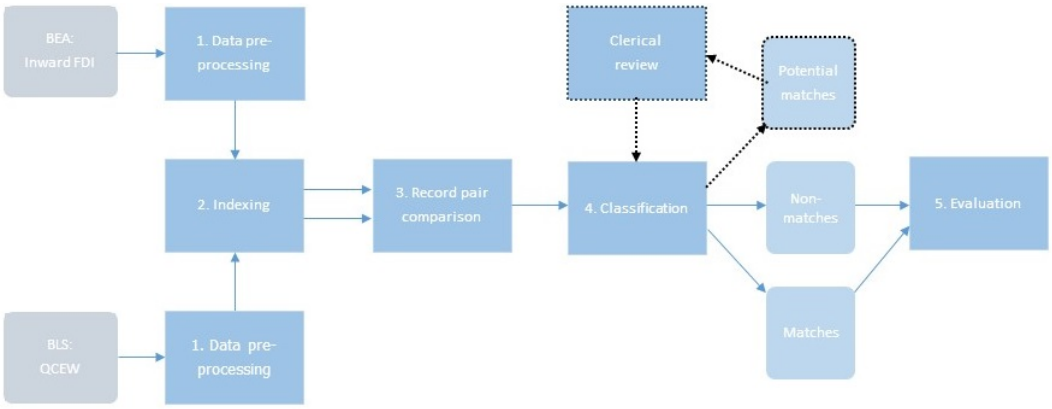

Figure 3: Record linkage process (Christen, 2012)

The steps are briefly described below. Additionally, implementation details in the context of linking the two datasets are given for each step.

3.1 Data Preprocessing

Data preprocessing is concerned with data quality and consistency between sources. This includes imputation of missing data and the standardization of data elements. In particular, preprocessing is important for names and addresses. It is necessary to remove unwanted characters and tokens, to standardize and tokenize, and also to parse into multiple output fields.

3.1.1 Data preprocessing implementation

Given the data are at different levels of employer structure, the majority of pre-processing involved transforming the data sources such that features could directly be compared. This meant aggregating QCEW establishments to the EIN level. For each EIN:

- Employment was summed across states to create a vector of state-by-state employment.
- Employment was summed across industry sectors and normalized to create a vector giving the share of employment in each sector.

Correspondingly for BEA affiliates, average distributions across sectors calculated from QCEW sectors are used.

For both inward FDI data and QCEW establishments:

- Physical addresses were parsed into street address, city, state, and zip code.
- All strings (street address, city, state, contact names and phone numbers, and trade and legal names) were standardized and converted to lowercase.

3.2 Indexing

The indexing step is concerned with reducing the search space for the remaining record linkage steps, particularly classification. Deterministic indexing techniques partition the search space into blocks by requiring subsets of the features of each dataset to match according to some function. Probabilistic techniques, such as Locality Sensitive Hashing (LSH) as described by Steorts (2014), compresses the search space such that similar pairs are mapped to the same sub-space with high probability. The compressed search space is referred to as the set of all candidate pairs.

3.2.1 Indexing implementation

Indexing used a hybrid approach of deterministic techniques and LSH. This is due to the fact that both affiliates and EIN employer aggregations vary greatly in terms of size, geographic distribution, and employer structure. Given a large affiliate with multiple subsidiaries will have many names associated with it, it is better to only compare these to correspondingly large EIN employer aggregations that also have many names. As such, deterministic techniques based on employer size, geographic distribution, and employer structure were first used to partition the search space. For each partition, LSH using Cosine similarity of 3-grams of the BEA and BLS employer names, normalized by TF-IDF was then used.

3.3 Record Pair Comparison

Record pair comparison takes all of the candidate pairs and compares like-features. This is mainly done using normalized similarity measures, where values close to 1 are highly similar, and values close to 0 are highly dissimilar. For numeric variables, similarity can be defined using 1 – the normalized Euclidean distance, for example. Categorical variables can be compared using set-based similarity measures, such as Jaccard Similarity. Finally,

text features such as names or addresses are first transformed to numeric vectors using a bag-of-words or bag-of-n-grams strategy. Using this strategy, the words or n-grams of the name or address are converted to tokens, the tokens are counted, and finally the counts are normalized and weighted to account for the importance of specific tokens. Once transformed to feature vectors, names and addresses can be compared using similarity measures. Normalized Cosine similarity is often used in this context.

3.3.1 Record pair comparison implementation

Similarity measures were applied to the candidate pairs. Euclidean distance was used for comparing employment distributions across states and industry sectors. Cosine similarity was used on the vectors corresponding to names, street addresses, city, state, and contact name. For names, both the affiliates and the EIN employer aggregations can have multiple names. For all of the other vectors, the EIN employer aggregations can have multiple vectors. As such, the maximum similarity measure of the Cartesian cross product of affiliate vectors and EIN employer aggregation were taken for each candidate pair. Lastly, Jaccard similarity measures were calculated for contact phone and zip codes, taking the maximum value for the similarity measures when the EIN employer aggregation have more than one contact phone or zip code.

3.4 Classification

Using the similarity measures for the candidate pairs as well as features that are unique to the two datasets, the classification step is concerned with predicting the linkage type (match, non-match) for the candidate pairs. There are several classification techniques, including probabilistic and supervised machine learning techniques.

3.4.1 Classification implementation

A subset of the 5,684 affiliates for which it was believed the analyst reviewed BEA-BLS linkage was of good quality was used as training data. The following classification methods were employed:

- Probabilistic:
  - Fellegi-Sunter
- Supervised machine learning:
  - Logistic regression
  - Support Vector Machines (SVM)
  - Random Forests
  - Gradient Boosting

The same features were used for all of the machine learning techniques. In addition to all of the similarity measures, the supervised learning techniques allow the use of features that are specific to one data source but not the other. BEA specific features include the number of subsidiaries and the geographic location of the parent foreign enterprise. BLS specific features include the number of establishments and features describing the type of EIN employer aggregation.

Additionally, to ensure a balanced training set, the number of “non-match” labels were down-sampled randomly to match the number of “match” labels for each affiliate.

3.5 Evaluation

After the candidate pairs are classified as matches or non-matches, it is necessary to evaluate the performance of the classification step. With labeled data, there are a number

of evaluation metrics that summarize the concordance between the actual to predicted linkages.

These include:

- Accuracy: the proportion of predicted labels ("match", "non-match") that are correct.
- Recall: the ability of the classifier to find all the "match" samples.
- Precision: the ability of the classifier not to label as "match" a sample that is a "non-match".
- F1 Score: harmonic mean of precision and recall.

Additionally, it is desirable to compare the classification techniques in terms of time and complexity.

3.5.1 Evaluation implementation

To ensure that the evaluation of classification methods were not measured on data used in training, a 5-fold cross-validation strategy was employed. Within each fold, a 75- 25% split was randomly selected for the training and testing data, respectively.

Accuracy, recall, precision, F1 scores, and AUC-ROC scores were tracked. ROC curves were plotted additionally. Processing and implementation time were recorded as well as the number and sensitivity of hyper-parameters that need to be tuned for each classification method.

4. Results

Table 1 gives evaluation measures for the classification methods employed. The probabilistic method of Fellegi-Sunter performs the worst amongst all other classification techniques for all measures except for recall, in which logistic regression was lower. This is not surprising given the supervised learning methods have more features to work with, particularly, those that are specific to one data source but not the other. The Fellegi-Sunter method can only look at directly comparable features. Logistic regression performs better than Fellegi-Sunter in the other evaluation measures, but worse than all of the other classification methods. SVM and gradient boosting perform roughly similarly, while random forests outperform all of the techniques.

Table 1: Comparison of the Classification Methods Using Evaluation Measures

Measure	*Classification Method*				
	Fellegi-Sunter	*Logistic*	*SVM*	*Random Forest*	*Gradient Boosting*
Accuracy	0.7564	0.8018	0.8641	0.9087	0.8752
Recall	0.8661	0.8463	0.8875	0.9437	0.8980
Precision	0.6080	0.7407	0.8360	0.8706	0.8483
F1	0.7145	0.7900	0.8610	0.9057	0.8724

These observations are apparent graphically in Figure 4, which provides ROC curves for all of the classification methods as well as summary AUC-ROC scores.

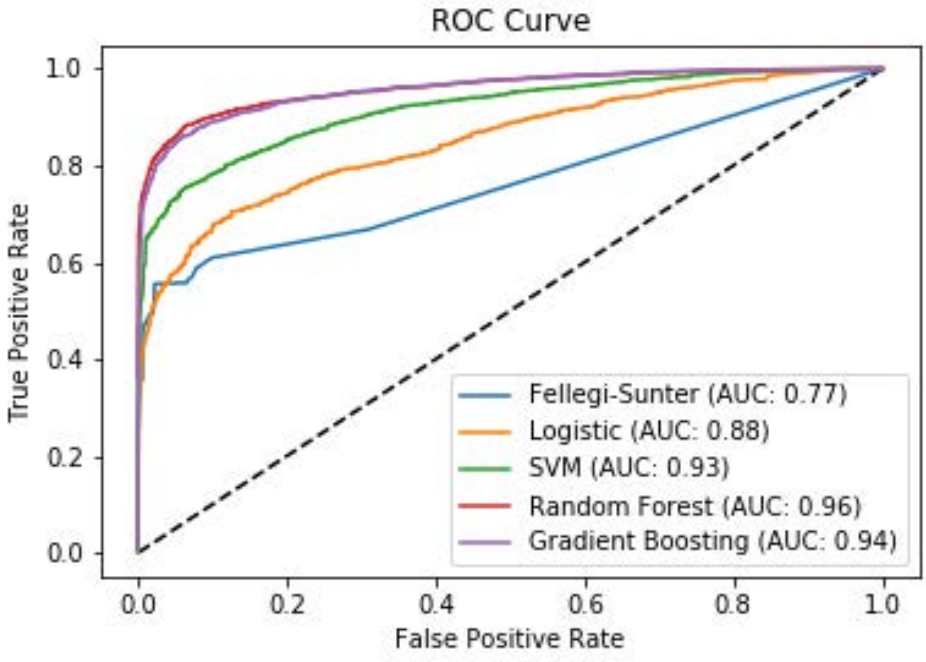

Figure 4: ROC curves for classification methods

It is important to consider other aspects of the classification methods as well. For example, while Fellegi-Sunter scores poorly in terms of its prediction power, it is much easier to implement and runs much quicker that the supervised learning methods. The supervised learning methods in contrast are more difficult to implement (although, having being implemented, revising the programs to run in production in place or in addition to an analyst is not nearly as long as the initial cost) and take longer to run. Furthermore, they have more parameters to tune. Some of these are highly sensitive and the values must be chosen carefully.

Table 2: Comparison of the Classification Methods Using Additional Evaluation Measures

Measure	*Classification Method*				
	Fellegi-Sunter	*Logistic*	*SVM*	*Random Forest*	*Gradient Boosting*
Processing time (in seconds)	1.52	41.84	1,360.13	310.82	16.32
Implementation time	Low	Medium	Medium	High	High
Number of hyper-parameters to tune	1	2-3	6-8	14	16
Sensitivity to hyper-parameters	Medium	Low	Low	High	High

5. Conclusion

This paper described a preliminary implementation of an automated record linkage process to integrate data that otherwise is quite labor intensive. Several different classification methods were compared.

The results are promising. In general, supervised learning techniques are more involved than probabilistic techniques, but have several advantages, including better performance and the ability to account for features particular to each data source. Of these techniques, random forests performed the best, correctly predicting the linkage type ("match", "non-match") approximately 91% of the time. Assuming this holds on subsequent data (for instance, integrating the 2013 inward FDI data with the QCEW), labor costs in terms of processing time for automated record-linkage is $1/130^{th}$ of manual analyst review time with only a 9% degradation in quality.

Many additional improvements are necessary. For instance, in data pre-processing, company names and street addresses could be further parsed into multiple outputs, allowing for a more refined record pair comparison. The indexing step needs refinement to better match affiliates to EIN employer aggregations. Further, a better comparison of the classification techniques would have the features selected independently for each method. Additionally, there are better ways to account for class imbalance than through the use of down-sampling.

More fundamentally, the classification methods implemented in this study do not collectively account for the fact that firms can be comprised of multiple EINs. Rather, this must be accounted for in a post-classification step. Ideally the classification step would account for these collectively. One promising method might be the newer Bayesian record linkage techniques, such as those by Stoerts (2015).

References

Bureau of Economic Analysis (2015). "Foreign Direct Investment in the United States: Final Results from the 2012 Benchmark Survey".

Bureau of Labor Statistics (2016). "Employment and Wages Online Annual Averages, 2012".

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer.

Handwerker, E. W., and Mason, L. (2013). "Linking firms with establishment in BLS microdata," Monthly Labor Review, Vol. 135, No. 6.

Handwerker, E. W., Kim, M., and Mason, L. (2011). "Domestic employment in U.S.-based multinational companies," Monthly Labor Review, Vol. 134, No. 10.

Steorts, R. C., Ventura, S., Sadinle, M., and Feinberg, S. (2014). "A Comparison of Blocking Methods for Record Linkage." In Privacy in Statistical Databases, 253-268. Springer.

Steorts, R. C. (2015). "Entity Resolution with Empirically Motivated Priors." Bayesian Analysis, 10, Number 4, pp. 849-875.

Talan, D., Mason, L., and Clayton, R. (2015). "Linking Foreign Direct Investment Data (FDI) to the BLS Business Register: Developing Employment Measures Related to Foreign Direct Investment." Paper presented at Joint UNECE, Eurostat, OECD Expert Group Meeting on Business Registers.