Constructing a Bridge between the LexisNexis Directory of Corporate Affiliations and the U.S. Business Register

Technical Note

Aaron Flaaen

January 8, 2013

The purpose of this matching exercise is to construct a concordance between the international corporate ownership structure information available in the LexisNexis Directory of Corporate Affiliations (DCA), and the Business Register (BR) of the U.S. Census Bureau. As there exists no identifier link between the two datasets, the principal merging procedure will utilize name and address information from each dataset to match at the establishment-level. This document will summarize each dataset, describe the principal features and issues surrounding the matching exercise, outline in detail each step of the procedure, and finally provide some summary statistics of the match.

1 Descriptions of the Data

1.1 The Directory of Corporate Affiliations

The Directory of Corporate Affiliations from LexisNexis describes the organization and hierarchy of public and private firms. The directory consists of three separate databases: U.S. Public Companies, U.S. Private Companies, and International – those parent companies with headquarters located outside the United States. The U.S. Public database contains all firms traded on the major U.S. exchanges, as well as major firms traded on smaller U.S. exchanges. To be included in the U.S. Private database, a firm must demonstrate revenues in excess of $1 million, 300 or more employees, or substantial assets. Those firms included in the International database, which includes both public and private companies, generally have revenues greater than $10 million. Each database contains information on all parent company subsidiaries, regardless of the location of the subsidiary in relation to the parent. The source of the data is a combination of public filings and independent research undertaken by LexisNexis. Specifically, they report using a "multi-faceted data maintenance strategy" that combines "the latest technology gathering techniques with teams of highly-skilled analysts based in the U. S. and India" to update their database on a daily basis. The version of the dataset used is an annual directory that spans the years 1993 to 2009.

1.2 The Census Bureau's SSEL/ Business Register

The U.S. Census Bureau has maintained a list of U.S. business establishments and companies since 1972. Originally known as the Standard Statistical Establishment List (SSEL), this register of information forms the backbone of many firm and establishment -level reporting to statistical and other federal agencies. In 2002 the SSEL was renamed the Business Register after a through redesign in order to improve coverage and quality control. There are two primary sources of information: First the IRS compiles information on single establishments and the administrative units of multi-establishment firms from payroll tax records.

I would like to thank Maggie Zhou for help with the LexisNexis data, and Kristin McCue for valuable comments on an earlier draft. Any opinions and conclusions expressed herein are those of the author and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

The Census Bureau's annual Company Organization Survey (COS) provides information on multi-unit establishments. The content of the Business Register includes business name and address, industry classification, and selected operating data (such as sales and employment). The frequency for updating individual data items varies from every quarter to every five years. The available sample period is 1972-2009.

2 Constructing the DCA to BR Bridge

2.1 Background on Name and Address Matching

Matching two data records based on name and address information is necessarily an imperfect exercise. Issues such as abbreviations, misspellings, alternate spellings, and alternate name conventions rule out an exact merging procedure, leaving the researcher with probabilistic string matching algorithms that evaluate the "closeness" of match – given by a score or rank – between the two character strings in question. These programs are notoriously time and computer-intensive, and because of this it is common to use so-called "blocker" variables to restrict the search samples within each dataset. A "blocker" variable must match exactly, and as a result this implies the need for a high degree of conformity between these variables in the two datasets. In the context of name and address matching, the most common "blocker" variables are the state and city of the establishment.

The current exercise uses the Stata program *reclink* created by Michael Blasnik, although there are other (and perhaps more sophisticated) options available. This program uses a bigram string comparator algorithm on multiple variables with differing user-specified weights.[1] This way the researcher can apply, for example, a larger weight on a near *name* match than on a perfect *zip code* match. Hence, the "match score" for this program can be interpreted as a weighted average of each variable's percentage of bigram character matches.

The danger associated with probabilistic name and address procedures is the potential for false-positive matches. For example, the two hypothetical names *Alpha Systems Inc.* and *Aloha Systems Inc.* will generate a high match score, but designating the two records as a match would be incorrect.[2] Thus, there is an inherent tension for the researcher between a broad search criteria that seeks to maximize the number of true matches and a narrow and exacting criteria that eliminates false-positive matches. Section 2.4 provides further details on the approach taken in this DCA-BR bridge with respect to this tradeoff.

2.2 The Unit of Matching

The primary unit of observation in both the DCA and BR datasets is the business establishment. Hence, the primary unit of matching for this bridge will be the establishment, and not the firm. However, there are a number of important challenges with an establishment-to-establishment link. First, the DCA and BR may occasionally have differing definitions of the establishment. One dataset may separate out several operating groups within the same firm address (i.e. JP Morgan – Derivatives, and JP Morgan - Emerging

[1] the term bigram refers to two consecutive characters within a string (the word *bigram* contains 5 possible bigrams: "bi", "ig", "gr", "ra", and "am"). Thus, for each variable, the program assigns a score based on the percentage of matching bigrams between the two datasets.

[2] In fact, this example demonstrates the benefits of matching based on multiple variables in addition to the establishment name.

Markets), while another may group these activities together by their common address. Second, the name associated with a particular establishment can at times reflect the subsidiary name, location, or activity (i.e. Alabama plant, processing division, etc), and at times reflect the parent company name. Recognizing these challenges, the primary goal of the bridge will be to assign each DCA establishment to the most appropriate business location of the parent firm identified in the BR. As such, the primary matching variables will be the establishment name, along with geographic indicators of street, city, zip code, and state.

This bridge seeks to match the establishments between these two datasets not merely at a point in time, but over a span of many years. Thus, the researcher is presented with a choice between matching the datasets year by year, or pooling the records for the entire sample, finding matches in any given year, and then relying on the Longitudinal Business Database (LBD) to track the matches across the sample.[3] While the former option is more repetitive and time-intensive, the latter presents nearly insurmountable data requirements. The Business Register contains roughly 7 million establishments per year, and pooling across even 10 years would create a dataset much too large to handle, much less apply a name/address matching algorithm. More importantly, because the DCA information on ownership structure can change from year to year without an address (or even name) change, the "pooling" approach suffers more seriously from errors when applying matches across time. Thus, this bridge matched the two datasets year by year, and uses only a limited degree of longitudinal information.

2.3 The Matching Process: An Overview

There is a generally conservative approach taken with constructing the bridge. The methodology will favor criteria that limit the potential for false positives at the potential expense of slightly higher match rates. As such, the procedure generally will require a match score exceeding 95 percent, except in those cases where ancillary evidence provides increased confidence in the match.[4]

The outline below summarizes the steps taken to merge the DCA dataset to the BR for a given year. This is an iterative process, in which a series of matching procedures are applied with decreasingly restrictive sets of matching requirements. In other words, the initial matching attempt uses the most stringent standards possible, after which the non-matching records proceed to a further matching iteration, often with less stringent standards. In each iteration, the matching records are assigned a flag that indicates the standard associated with the match. The possible match flags are identified in table 3, with further details provided below.

2.4 Steps for DCA-BR Matching

1. Match DCA to Compustat (and then to Compustat-Bridge) for those DCA observations with a Compustat Identifier (see Zhou (2011) for more details.)

2. Implement Tier 1 changes to name and address variables of DCA (see table 1). Separate out DCA observations that have matched via Compustat.

[3] See Jarmin and Miranda (2002) for a description of the LBD.

[4] The primary sources of such ancillary evidence are manual (ocular) review of the matches, and additional parent identifier matching evidence.

3. Tier 1 Matching

   (a) Restrict BR to LBD observations (save non-matching observations for Tier 2)

   (b) Implement Tier 1 changes to Name/Address variables of BR.

   (c) Apply *reclink* of Compustat-linked DCA observations to BR (using name, street address, zip code, and requiring city, state, and firmid to match exactly)

   (d) Apply *reclink* of non Compustat-linked DCA observations to BR (using name, street address, zip code, and requiring city and state to match exactly)

   (e) Apply *reclink* of non-matching DCA to BR (using name, street address, zip code, and city, but now only requiring state to match exactly)

   (f) Evaluate matches

      - if "match score" is above 0.95, classify as a match[5].
      - if "match score" is between 0.80 and 0.95, evaluate manually [6]
      - if "match score" is below 0.80, classify as a non-match

   (g) Append evaluated-as-match DCA observations to the other matched observations, and send non-matching DCA observations to Tier 2 matching

4. Tier 2 Matching

   (a) For the non-matching DCA observations, try to find an existing match with the same (DCA) parent identifier. Take the corresponding BR firm identifier (alpha or ein) for this match, and search for match over BR observations with identical alpha/ein

      - apply *reclink* of DCA to BR (using name, street address, and city, and requiring state and alpha/ein to match exactly)
      - if "match score" is above 0.70 classify as as match – spot checks have shown no false positives when requiring the alpha/ein to match

   (b) Implement Tier 2 changes to name variable of DCA (see table 2)

   (c) Take non-LBD matched BR observations and implement Tier 1 and Tier 2 changes to name and address variables.

   (d) Apply *reclink* of DCA to non-LBD-matched BR observations (using name, street address, and city, and requiring state to match exactly)

      - if "match score" is above 0.95, classify as a match

[5] I've looked at a several thousand of these potential matches and see that the false-positive rate for these is VERY small (i.e. less than 0.5 percent)

[6] The manual evaluation of matches is the one step in which I utilize some longitudinal information. (Without this, the set of potential matches to evaluate was too large – in the range of 5-6 thousand per year.) Rather than continue to manually review common matches (and non-matches) from year to year, I keep the pool of manually evaluated matches from previous years and automatically accept as a match any potential match that exactly aligns with a match evaluated in a previous year. The same is true for previously-evaluated non-matches.

2.5 Cleaning Methods

The final step in the matching process involved a number of checks and reviews of the matched observations. A common issue that arises in name and address matching to the Business Register involves the presence of food service contractors (NAICS code 722310) and fitness facility management services (NAICS code 713940). In the cases when these services are performed on-site for large business establishments, the address reported for these vendors will be identical to that for the actual business establishment. More problematic is when the vendor reports its name using the client name rather than its own business name (ostensibly to avoid confusion with its other locations) in its reporting to the Census Bureau. These cases are particularly problematic for any name/address matching algorithm, unless one is able to incorporate industry characteristics in the match. A large component of the cleaning was to remove these false matches from the data.

3 Results and Discussion

Panel A of Table 4 details the overall year-by-year establishment match rates from the DCA to the BR. The first column indicates the establishment coverage of the DCA data, which ranges from roughly 60,000 establishments in the 1990s to over 110,000 establishments in the late 2000s. It is clear that the sample increased significantly during years 2001 to 2003. This is chiefly a result of increased industry coverage, as well as a slightly lower sales/employment threshold for establishment inclusion by LexisNexis. The match rates are quite high by the standards of name/address matching, ranging from a low of 0.65 in 1996 to a high of over 0.73 in the 2000s.

Panels B and C of Table 4 describe additional information on two important subsets of the DCA data, and their subsequent match rates to the BR. Panel B details statistics on those DCA establishments whose firm owns affiliates outside the United States (U.S. multinationals), while Panel C describes those DCA establishments whose parent firms are headquartered in foreign countries (foreign multinationals). These are useful subsets of the DCA data for a variety of reasons. First, because multinational firms tend to be among the largest (in terms of the number of distinct locations, or metrics such as sales and employment) in the economy, these distinctions can serve as a proxy for whether the match rates cover the most significant firms in the economy. Second, one might expect these establishments to present some particular challenges for a name/address matching procedure. Because of the complex ownership patterns of these firms, they may suffer disproportionately from the subsidiary-vs-parent and location-vs-activity issues in the name variable identified above. Moreover, the non-English based names common to the U.S. affiliates of foreign multinationals may be more prone to misspellings and/or alternate naming conventions. Finally, the Census Bureau data infrastructure contains surprisingly little information on the locations and operations of multinational firms. Hence the data provided via the DCA-BR bridge may be of particular value for this subset of firms.

The first column of Panels B and C detail the number of U.S. establishments of U.S. and foreign multinationals respectively. The magnitude of these numbers make clear that U.S. multinationals comprise a substantial share of the establishments covered by the DCA data. The U.S.-based affiliates of foreign multinationals are a smaller, yet still significant, subset of the DCA data. It is also clear that the significant rise in establishments during the 2001 to 2003 period is less pronounced among the multinational samples

– hence confirming that the expansion in coverage was primarily due to the inclusion of smaller firms into the directory. The second and third columns of each panel show that the match rates of these two subsets are broadly consistent with the overall match rates shown in Panel A. The match rates for U.S. multinationals range from 0.62 in 1996 to 0.73 in the late 2000s while the foreign multinational match rates are generally a little higher. Figure 1 provides a graphical representation of the match rates found in Table 4.

Information on the match rates based on the firm-level, rather than the establishment level, is not provided here due to disclosure concerns. In general these match rates are considerably higher than those at the establishment level. At first glance this be relatively unsurprising since a match of only one establishment within a firm will provide a firm-level match, even though many other establishments within the firm may remain unmatched. However, the establishment-level match rate does not constitute a lower bound of the firm-level match rate, and so in general it could be either higher or lower.

Perhaps as important as the magnitude of the matches between these two datasets is the degree of reliability of those matches. Figures 2, 3, and 4 decompose the establishment match rates based on the type of match identified in the matching process. The figures detail a slightly condensed form of the classification scheme outlined in table 3. The figures are constructed in such a way that the highest-confidence match types begin at the bottom, with subsequent layers representing the stock of additional matches with slightly lower standards. To some degree this "ranking" of match types is somewhat arbitrary – one could legitimately argue that a match that has been visually identified in the *Evaluated* bin should exhibit a lower probability of a false-positive match than in the *City/State 0.95 < score < 0.99* bin.

4 Conclusion

This note has documented the data sources, methodology, and results of an annual concordance between the LexisNexis Directory of Corporate Affiliations and the Census Bureau's Business Register for a period spanning 1993-2009. There are surely numerous avenues of research for which this bridge will prove to be a valuable resource, and the Census Bureau in particular will benefit from the corporate ownership structure maintained by LexisNexis. Annual updates to the bridge, and further improvements to the current matched period would be beneficial exercises for continued work into the future.

Specifically, it remains possible to improve on the match rates identified in this document if one were to look for non-matched establishments/firms in a given year using an identified match of that DCA establishment/firm in a different year. The concern with this process involves the transfer of firm identifiers across time, for which ownership changes (and, indeed, the differences in DCA vs BR definitions of the firm highlighted above) can introduce significant errors into the matches. However, a careful implementation of this approach could improve on the matches. This will be left for future work.

References

Blasnik, Michael, (2010), RECLINK: Stata module to probabilistically match records.

Jarmin, Ron and Javier Miranda. 2002. "The Longitudinal Business Database", mimeo, available at http://www.vrdc.cornell.edu/info7470/2007/Readings/jarmin-miranda-2002.pdf

Zhou, YM. 2007. "Structural Complexity and Diversification," Working Paper

Table 1. Tier 1 String Variable Modifications

Changes to *Name* Variable

All characters changed to lowercase
Remove all commas and single quotes from string
Remove leading, trailing, and doubles spaces
If first word of string is "the ", remove
Each of the following treated as identical[1]

" incorporated" — " inc." — "inc"	" corporation" — " corp." — " corp"
" company" — " co." — " co"	" limited" — " ltd." — " ltd"
" association" — " assn." — " assn"	" manufacturing" — " mfg." — " mfg"
" international" — " intl." — " intl"	" division" — " div." — " div"
" &" — " +" — " and"	

Changes to *Street* Variable

All characters changed to lowercase
Remove all commas and single quotes from string
Remove leading, trailing, and doubles spaces
Each of the following treated as identical[1]

" street" — " st." — " st"	" drive" — " dr." — " dr"
" road" — " rd." — " rd"	" boulevard" — " blvd." — " blvd"
" avenue" — " ave." — " ave"	" court" — " ct." — " ct"
" circle" — " cir." — " cir"	" lane" — " ln." — " ln"
" place" — " pl." — " pl"	" parkway" — " pkwy." — " pkwy"
" expressway" — " expwy." — " expwy"	" highway" — " hwy." — " hwy"
" freeway" — " fwy." — " fwy"	" center" — " ctr." — " ctr"
" building" — " bldg." — " bldg"	" suite" — " ste." — " ste"
" floor" — " fl." — " fl"	
" n." — " n"	" w." — " w"
" s." — " s"	" e." — " e"
" n.w." — " nw." — " nw"'	" s.w." — " sw." — " sw"'
" n.e." — " ne." — " ne"'	" s.e." — " se." — " se"'
" first" — " 1st"	" second" — " 2nd"
" third" — " 3rd"	" fourth" — " 4th"
" fifth" — " 5th"	" sixth" — " 6th"
" seventh" — " 7th"	" eighth" — " 8th"
" ninth" — " 9th"	" tenth" — " 10th"
"p.o." — "po"	

Changes to *City* Variable

All characters changed to lowercase
Each of the following treated as identical

"saint " — "st. " — "st "	"fort " — "ft. " — "ft "
"north " — "n. " — "n "	"south " — "s. " — "s "
"east " — "e. " — "e "	"west " — "w. " — "w "

[1] Note that the use of spaces before each character string reduces the chance that altering an abbreviation may result in changing a non-abbreviated (but identically denoted) string. Any unintended changes that may still result are not necessarily a problem, however, as they are implemented on both datasets. Thus in principle the match should be unaffected.

Table 2. Tier 2 String Variable Modifications

Changes to *Name* Variable

Remove the characters "-" and "/"

Remove each of the following from the string[1]

"national"	"systems"
"industries"	"securities"
"management"	"insurance"
"association"	"america"
"american"	"north america"
"north american"	"intl"
"ltd"	"corp"
"inc"	

[1] Note here the general lack of spaces before each character string. One must be careful that the string to be removed is not embedded as part of a larger string that should be maintained in the variable. The chances of this appear to be very low, and once again any unintended changes would be implemented on both datasets.

Table 3. Classification of Matches

Tier 1 Match Categories

Flag Value	Variables Used	Variables Required	Score
COMP	name, street, zip	city, state, firmid	> 0.95
A1	name, street, zip	city, state	1
A2	name, street, zip	city, state	> 0.99 & < 1
A3	name, street, zip	city, state	> 0.95 & ≤ 0.99
A2-A	name, street, city, zip	state	> 0.99
A3-A	name, street, city, zip	state	> 0.95 & ≤ 0.99
E	name, street, zip	city, state	> 0.80 & ≤ 0.95 & Evaluated

Tier 2 Match Categories

Flag Value	Variables Used	Variables Required	Score
T2-A1	name, street, city	alpha/ein, state	> 0.99
T2-A2	name, street, city	alpha/ein, state	> 0.95 & ≤ 0.99
T2-B1[1]	name, street, city	alpha/ein, state	> 0.90 & ≤ 0.95
T2-B2[1]	name, street, city	alpha/ein, state	> 0.80 & ≤ 0.90
T2-B3[1]	name, street, city	alpha/ein, state	> 0.70 & ≤ 0.80
T2-C1[2]	name, street, city	state	1
T2-C2[2]	name, street, city	state	> 0.99 & < 1
T2-C3[2]	name, street, city	state	> 0.95 & ≤ 0.99

[1] A lower threshold value for a match is used here because of the "alpha"/"ein" blockers. Manual checks have shown no false-positives with a score above 0.70, however in principal these could occur.

[2] Matching based on the non-LBD sample of the SSEL. See the outline above for further explanation.

Table 4. DCA Establishments and Match Rates, by Firm Type

	Panel A: Total DCA			Panel B: U.S. Multinationals			Panel C: Foreign Multinationals		
	DCA (Total)	Matched to BR	Match Rate	DCA (Total)	Matched to BR	Match Rate	DCA (Total)	Matched to BR	Match Rate
1993	61,646	43,190	0.70	21,482	14,387	0.67	8,270	5,810	0.70
1994	64,090	44,904	0.70	22,396	15,110	0.67	9,326	6,437	0.69
1995	65,223	45,743	0.70	22,952	15,448	0.67	9,365	6,414	0.68
1996	64,152	41,713	0.65	22,353	13,806	0.62	10,057	6,331	0.63
1997	60,884	41,290	0.68	20,962	13,583	0.65	9,556	6,328	0.66
1998	59,043	40,854	0.69	20,012	13,218	0.66	9,416	6,282	0.67
1999	58,509	40,697	0.70	20,157	13,408	0.67	9,218	6,054	0.66
2000	68,672	48,875	0.71	18,728	12,631	0.67	9,900	6,755	0.68
2001	70,522	50,105	0.71	18,516	12,477	0.67	10,089	6,864	0.68
2002	97,551	66,665	0.68	31,260	21,004	0.67	13,168	8,483	0.64
2003	123,553	86,838	0.70	25,905	17,465	0.67	11,101	7,398	0.67
2004	117,639	84,450	0.72	24,028	16,923	0.70	10,152	7,156	0.70
2005	110,106	80,245	0.73	20,870	15,191	0.73	9,409	6,865	0.73
2006	110,826	79,275	0.72	21,335	15,539	0.73	9,981	7,243	0.73
2007	112,346	81,656	0.73	22,500	16,396	0.73	10,331	7,555	0.73
2008	111,935	81,535	0.73	23,090	16,910	0.73	9,351	6,880	0.74
2009	111,953	81,112	0.72	22,076	16,085	0.73	11,142	8,193	0.74

FIGURE 1. DCA-BR ESTABLISHMENT MATCH TYPES: ALL DCA ESTABLISHMENTS

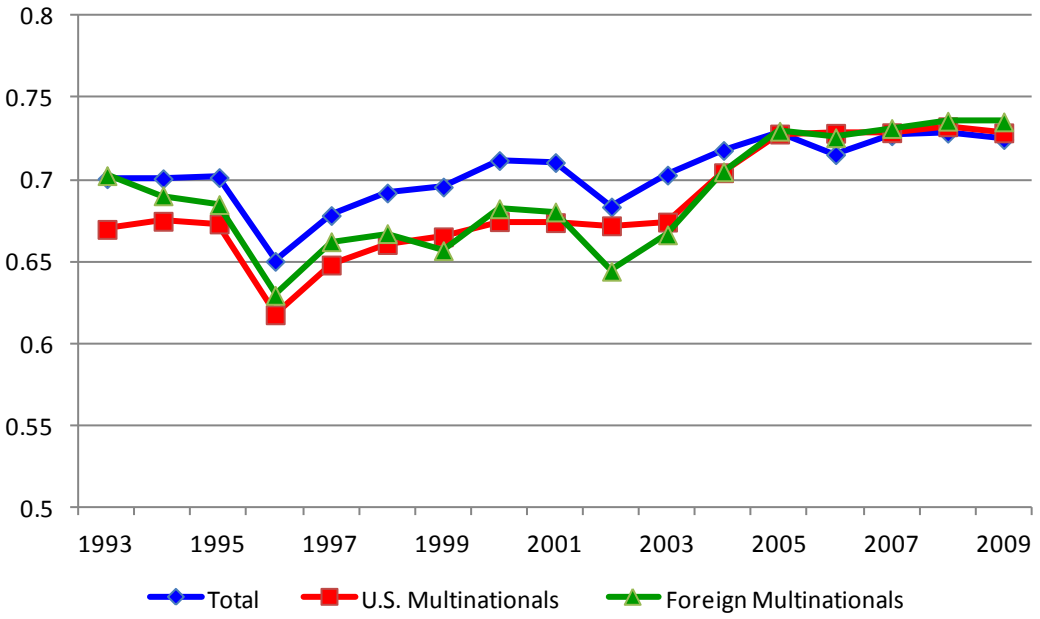


FIGURE 2. DCA-BR ESTABLISHMENT MATCH TYPES: DCA ESTABLISHMENTS

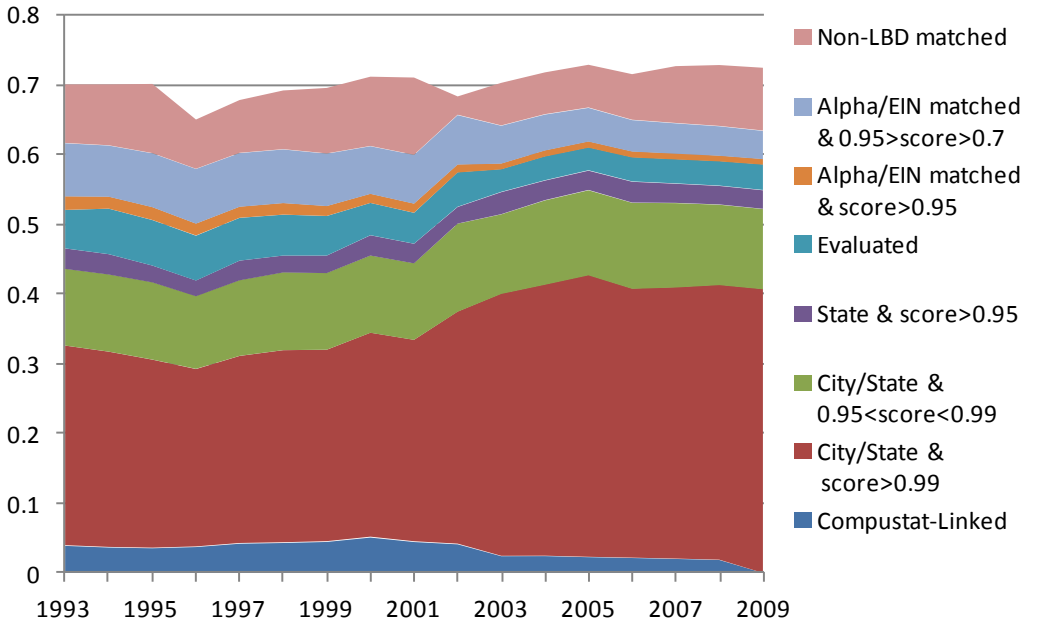


Figure 3. DCA-BR Establishment Match Types:
U.S. Multinational Establishments

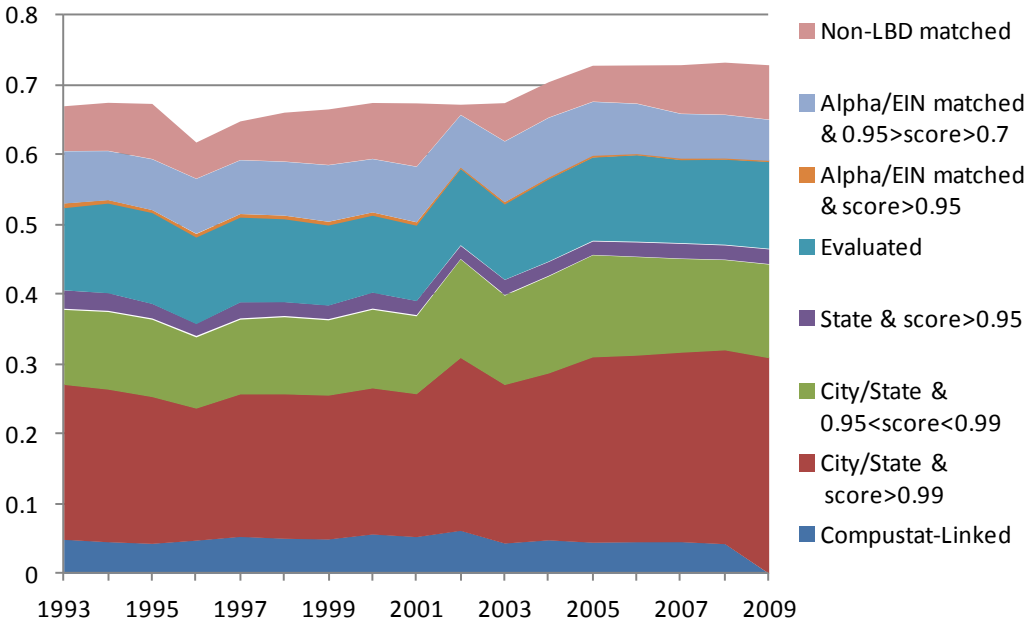
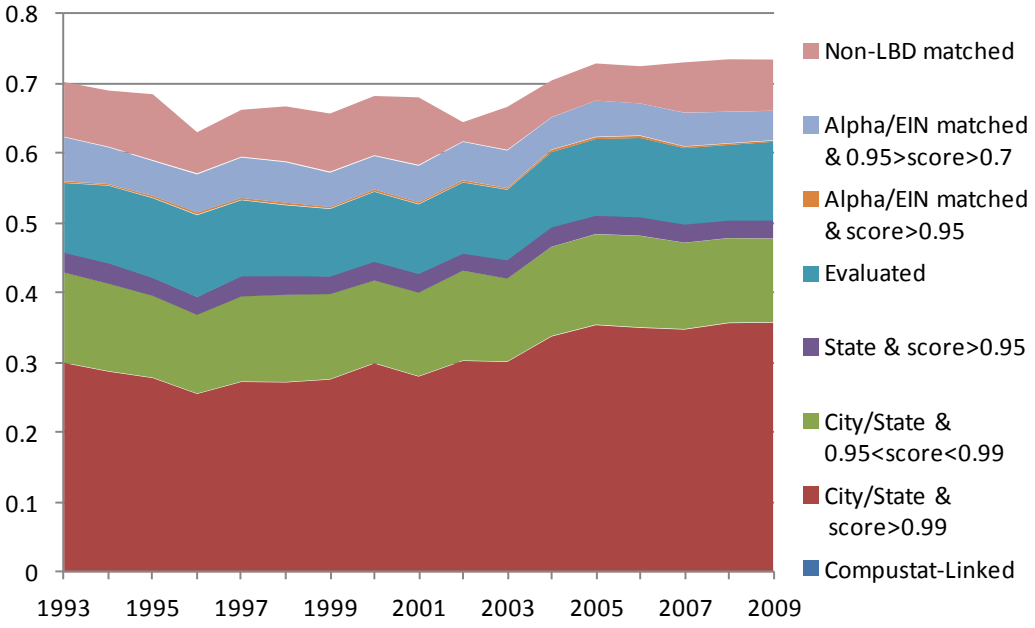


Figure 4. DCA-BR Establishment Match Types:
Foreign Multinational Establishments



Appendix: Weighted Bigram Matching Using Stata's Reclink[7]

The fuzzy matching procedure *reclink* for Stata, written by Michael Blasnik, assigns a match score between each observation i in a master dataset and observation j in the using dataset. This appendix first describes how the program defines a match score, and subsequently details the matching algorithm in full.

A.1 Defining the observational match score

The *reclink* program allows matching across multiple variables in each dataset, with differing weights applied to each variable. Moreover, the program distinguishes between matching and non-matching weights: a match weight reflects the relative likelihood that a variable match identifies a matching record, whereas a non-match weight reflects the relative likelihood that a variable non-match rules out a potential matching record. This distinction is important, as Blasnik states "A variable such as a telephone number may have a large match weight but a small non-match weight because matches are unlikely to occur randomly, but mismatches may be fairly common due to changes in phone numbers over time or multiple phone numbers owned by the same person/entity."

Let B_k be the bigram score (defined as the percent of bigrams in common) between variable k in the master and using datasets. Let $B_{\min}$ be defined as a minimum raw bigram score that serves correct for short strings. Further let w_k be the weight assigned to the matching score of variable k, and then let w_k^n be the weight assigned to the non-matching score of variable k. Then the matching score for variable k is defined as:

$$M_k = \begin{cases} w_k * \frac{1}{2}B_k^2 + \frac{1}{2}\left(\frac{B_k - B_{\min}}{1 - B_{\min}}\right)^{\frac{1}{3}} & : \text{if} B_k > B_{\min} \\ w_k * \frac{1}{2}B_k^2 & : \text{otherwise} \end{cases}$$

and the non-matching score of variable k is defined as:

$$NM_k = \begin{cases} w_k^n & : \text{if } B_k < B_{\min} - 0.2 \\ 1 - B_k^2 & : \text{if } B_{\min} - 0.2 <\leq B_k < B_{\min} \\ (1 - B_k)^2 & : \text{if } B_k \geq B_{\min} \end{cases}$$

Thus, the match score between observation i and j is the weighted average of these scores across the set of variables K used in the matching. Specifically, for each j $\ni$ J, the match score is :

$$\text{score}_j = \frac{\sum_{k=1}^{K} M_k}{\sum_{k=1}^{K} M_k + \sum_{k=1}^{K} NM_k} \tag{1}$$

A.2 Steps in the Reclink Algorithm

1. Identify exact matches, set aside, and remove from master and using datasets
2. For each observation i in master data, load the set J of observations in the using dataset that satisfy the sampling requirements (i.e. the block and require statements)

[7]I would like to thank Nada Wasi for her work identifying some of the match score definitions described in this appendix.

3. Then, for each j in J calculate the score, defined in equation (1)

4. Rank scores j = 1,...,J which are above the user-specified "minimum score"

   (a) Keep the observation with the maximum score as "matched" for observation i.

   (b) If no observation in J obtains a score above the "minimum score", mark as "non-matched".

5. Repeat steps 2-4 for all observations in the master dataset

6. Merge back the exact matches from step 1.