ADMINISTRATIVE RECORD MATCHING FOR THE 1992 ECONOMIC CENSUSES

Philip M. Steel, Carl A. Konschnik, U.S. Bureau of the Census
Philip M. Steel, U.S. Bureau of the Census, Services Division, Washington, DC 20233

Key Words: Business Lists, Name and Address Matches, Record Linkage, Link Variables

1. Overview

This paper describes a matching process which improves the linkage between sole-proprietorship income tax return records from the Internal Revenue Service (IRS) and their associated payroll records on the Census Bureau's Standard Statistical Establishment List (SSEL).

The matching process supplements the linkages made previously based on a common primary identifying number on the two types of records. This number is the Employer Identification Number (EIN), issued by IRS to businesses with employees, and used by them as a principal taxpayer identification number. Unfortunately this common identifier is omitted on roughly 30% of the annual income tax returns on which it should appear. In matching, our aim was to make the linkages more complete by using other information besides the EIN--chiefly, name, city, state, ZIP code, payroll and kind-of-business activity code.

2. Context and Motivation for the Matching

Linking receipts and payroll records depends largely on associating the correct EIN with each annual income tax return. A sole-proprietorship business, when filing the required annual Form 1040, Schedule C, (or, briefly, 1040-C) tax return with the IRS, uses the owner's Social Security Number (SSN) as its taxpayer identification number. If the business has employees, it is required to have an EIN and use it for filing IRS Form 941, Employer's Quarterly Federal Tax Return. When it files its annual 1040-C tax return, the sole-proprietorship business is asked to provide its EIN if it has one. This reported EIN is the principal link between the annual business income and quarterly payroll tax returns for sole-proprietorship employers.

The IRS provides Form 941 payroll data to the Census Bureau weekly for updating the SSEL. These payroll records, along with data received monthly from the IRS Business Master File (BMF), serve to keep the SSEL current with name and address, employment, payroll, form of organization, and other key data for business. The primary identifier used for the BMF and the 941 files, is, of course, the EIN. By natural extension, for processing administrative records, the EIN is also the primary identifier on the SSEL. All employers--corporation and partnership employers, as well as sole-proprietorship employers, file their Form 941 under their EIN. Because partnership and corporation income tax returns are filed under an EIN, the linkage between receipts from annual tax returns and payroll records for these businesses is readily available. However, for sole-proprietorships, if the EIN is missing or incorrect on the 1040-C, we obviously can't rely on the EIN to update the appropriate SSEL payroll record with 1040-C receipts.

Complete updating of receipts on the SSEL is important because the Census Bureau's economic censuses use the SSEL as a frame and use IRS tax return data from the 1040-C to tabulate receipts for single-establishment (singleunit), sole-proprietorship businesses with payroll below prescribed cutoff levels. These cutoffs vary by kind of business. Singleunit businesses with payroll above the cutoffs and all multi-establishment (multiunit) businesses from the SSEL are mailed a census form. Tax return data from the 1040-Cs are also used to account for those who fail to respond to the mailing.

Incomplete linkage between 1040-C employers and the SSEL means that the file of 1040-C records from IRS for a census year such as 1992 (after removing 1040-C linked to the SSEL) still contains some employer as well as all nonemployer businesses. This causes two problems:

(1) tax-return receipts are not available on the SSEL for tabulating inscope EINs with nonzero 1992 payroll and missing receipts for the economic censuses.

(2) the 1040-C file includes an unknown number of employers with unknown total receipts. Therefore, we cannot use it directly to tabulate census year receipts for nonemployer businesses.

Both problems are alleviated by improving the linkage between the file of 1040-C records and the SSEL.

For the 1992 censuses, we obtained an EIN to SSN cross-reference (x-ref) file from IRS to aid in linking records. In addition to this, we used matching techniques to associate 1040-C records with their

associated SSEL payroll record. In the following sections, we present the technical details of this matching work and discuss the impact it had on the 1992 census estimates.

3. Description of the Files for Matching

After updating the SSEL using the reported EIN on the 1040-C and the SSN to EIN cross-reference file from the BMF, we were left with EINs on the SSEL which were still missing receipts. A file of these EINs drawn from the SSEL, formed the primary file for the matching.

The number of unlinked, potentially matchable EINs, on the file at this point was 419,494. The criteria for selecting these cases were that:

(1) The EIN be within U.S. boundaries

(2) The Legal Form of Organization (LFO) be a sole-proprietorship or form of organization unknown (as opposed to partnership or corporation form of organization).

(3) The EIN be taxable or have tax status unknown.

(4) The EIN reported nonzero payroll for the 1992 census year.

The second file for matching consisted of 1040-C sole-proprietorship tax return records. A 1040-C may have (in census processing) up to three schedules, each representing a separate business. The schedule, together with the name and address from the main 1040, form our 1040-C record. The SSN, together with one schedule number, formed the identifier for the second file. The original 1040-C file included those with EINs that linked to the SSEL, this file contained 16,540,844 schedules in all. Because this large file size exceeded the capacity of our software platform (a VAX minicomputer cluster), our matching was performed with this file split into 36 pieces.

4. Variables for Matching and Their Comparability

Records on both files contain name, address, kind-of-business and payroll fields. Each of these fields has associated problems.

• Name Fields -- The primary name field from the SSEL may be the name of a business, e.g., the American Bank Note Company, but in the case of sole-proprietors this field is usually the proprietor's name, even where the LFO has not been determined. On the other hand, the 1040-C record has the Form 1040 name, which is a personal name, often including both husband and wife for joint filing. There are several other name fields available on the SSEL, such as the census name, physical location name, and mailing name. These were examined as candidates for matching fields, but appeared to contribute very little to establishing new linkages (during testing, the census name field update was incomplete, and may yet be shown to be useful for future matching). To summarize, on the EIN file we have a name field that may or may not contain a personal name; on the 1040-C file we have a name field that may contain compound names, with either of the components a candidate for matching. To deal with this, our name parser rejects records with identifiable business names from the EIN file and generates two records for compound names on both the 1040-C file and (in a few cases) the EIN file.

• Address Fields -- Both files have address fields containing street address, city, state and ZIP code. However, the address on the SSEL is generally the business address and the 1040-C address is a personal address. Using a test file containing only known linkages between the EIN and 1040-C records, we found that the street addresses matched partially or better only 30% of the time (+/- 6%) based on a clerical review of a sample of 248 cases. This eventually led us to drop the street address as a matching variable. The same problem--that the 1040-C address and SSEL address of known linkages can be different--applies to city, state and ZIP code, but to a lesser degree. These variables were retained for matching.

• Business Classification Codes -- The EIN's business activity code from the SSEL is the Standard Industrial Classification (SIC) code, whereas the 1040-C record has a converted Primary Business Activity (PBA) code. The PBA code is an abbreviated but roughly comparable coding system. The SIC on the SSEL is generally coded from various sources. In contrast, the PBA is a self-reported code by the taxpayer. Previous studies indicate that we can expect the self-reported code to match the SIC code no more than about 67% of the time at the four-digit level and 75% at the two-digit level. See Konschnik et al. (1993) for more on the quality of self-coded PBAs.

• Annual Payroll -- We obtain a single annual payroll figure for EIN records from the SSEL. The 1040-C has two fields related to payroll--wages and cost of

labor. Technically, the wages field is supposed to correspond to payroll, and cost of labor to represent contracted labor where the employment taxes are born by the employer of the contracted laborers. Examination of the data shows this is not always the case. Although for most of the time, the SSEL payroll figure agrees with the wages figure from the 1040-C, the exact number sometimes appears in the cost of labor field or even split across both fields. Whether this is due to taxpayer reporting error, or keying error is an open question. Since legitimate (nonpayroll) data also appears in the cost of labor field, a statistical solution was required.

5. Software Used for the Matching

For the matching software, we used Winkler's mf3 matcher, with match specific modifications. We used both character-by-character comparisons and one of the native string comparators. For the numeric comparison on the payroll variable, we developed a new module, about which we will go into in some detail.

The EIN records from the SSEL were extracted and "prepped", forming a "stationary" file of 351,141 records. The 1040-C files were preprocessed and matched in 36 separated cuts of roughly 750,000 records (each).

6. Blocking Criteria

The blocking criteria, defined as the minimum characteristics necessary to consider a pair of records in the match, were the first six letters of the last name and the first letter of the first name. We originally explored the possibility of blocking by ZIP code but abandoned this when we realized the scope of the problem in business versus home addresses.

7. Matching Variables and Weights

The fit between any pair of records is determined by the sum of the weights of the match variables. We assign positive weights for agreement and negative weights for nonagreement. Below is a list of the match variables, along with their positive and negative weights. A record from the 1040-C file is considered a match to a record from the SSEL file when the pair's match score (sum of weights) exceeds 15.15.

The match variables fall into three groups: Name, Location and Business. This suggests the general strategy we employed for determining the weights. The role of the Name group was to further (beyond blocking) qualify pairs--a failure on more than one of the name variables here should disqualify a record. The other two groups were weighted to balance one another--a weak score on Location required a strong score on Business and vice versa, with Business given slight precedence over Location.

Group	Description	Positive Weight	Negative Weight
Name	last name	5.01	-8.11
	first name remainder	5.01	-7.82
	middle initial	3.00	-8.06
	middle name remainder	2.18	-0.01
Location	city	3.04	-0.00
	state	0.00	-6.31
	5 digit zip code	3.04	-0.00
	first 3 digits of zip	3.04	-0.00
	first 2 digits of zip	3.00	-2.00
Business	entire SIC	3.06	-0.00
	first 2 digits of SIC	3.00	-3.00

The annual payroll variable was handled somewhat differently when determining weights. The payroll variable looks at the ratio of 1040-C payroll (wages+cost of labor, combined in the prep phase) + 5000 to SSEL payroll + 5000. Calling this ratio R, weights were assigned based on the interval in which R fell.

Range				Weight
0.00	< R	<=	0.64	-7.00
0.64	< R	<=	0.87	0.00
0.87	< R	<=	0.93	4.50
0.93	< R	<=	1.05	7.50
1.05	< R	<=	1.13	4.50
1.13	< R	<=	2.25	0.00
2.25	< R			-7.00

The factor of 5000 keeps a small absolute difference of say 1000 (possibly a rounding error) from making R too large or too small.

8. How the Model for Relating the Payroll Variables Was Determined

We constructed two files to test competing models for the payroll comparison. A file of randomly joined payrolls from a known sole-proprietors file and a sample file of 1040-Cs was created (random set). Both payrolls were taken from an EIN linked file of sole-proprietors to create the second file (truth set).

Next, we tested three models as shown below.

model 1: $\frac{\text{wages} + X}{\text{SSEL payroll}+X}$

model 2: $\frac{\text{wages} + \text{cost of labor} + X}{\text{SSEL payroll} + X}$

model 3: $\frac{\text{cost of labor} + X}{\text{SSEL payroll} + X}$, if wages=0,

and $\frac{\text{wages} + X}{\text{SSEL payroll} + X}$, otherwise.

The discriminating power of the variable must contrast the behavior over the truth set against its behavior on the random set. The addition of a term to top and bottom of the ratio pulls the distributions toward 1. In fact, the distribution centralizes faster on the truth set than it does on the random set.

The criteria for selection was to select the model that produced the most ratios near 1 and the fewest ratios at the extremes on the truth set, and simultaneously produced the fewest ratios near 1 and the most at the extremes on the random set. Models 2 and 3 were clearly better than model 1, model 3 slightly better than model 2. Model 3 was a later invention and did not make it into production. For the selected model, value of X = 5000, and the four most critical conditions, we have:

P(strong agree|match)=53.05
P(strong agree|nonmatch)=0.4
P(disagree|match)=12.3
P(disagree|nonmatch)=88.8

9. Match Results

The Parser

The parser behaved very differently on the two files. The 1040-C name field is highly structured, generally well keyed, and contains no legitimate business names. The SSEL name field may have a sole-proprietor's personal or business name, or it may have the name of a corporation or partnership--this latter group a contribution from the unknown LFO. The personal names include more abbreviations and are less structured.

Looking at the results of the parser on the 1040-C file (excluding schedules linked to the SSEL), we see that 23,670 of 14,894,578 (0.16%) schedules failed to parse and were not included in the match. Almost all failures were due to complicated name structures. 10.6 million records with duplicate identifiers were created from joint returns, i.e. roughly 10 duplicates for every 14 schedules. With duplicates, the prepped 1040-C file had 25,429,164 records.

From a test of confirmed sole-proprietors (of about 19,000 records) we know that the parser succeeds about 97.4% of the time. The unparsed SSEL file, which included records with LFO unknown, had 419,494 records--332,441 of which parsed. Using the known rate we can deduce that the unparsed file contained about 341,315 sole-proprietors (virtually none of the non-sole-proprietors parse). Hence we have an estimated 8,874 sole-proprietor establishments whose name failed to parse, and, consequently, were not included in the match. Roughly 25% of failures were due to unrecognized name patterns, the remainder were recognized as business names. We can infer from this that sole-proprietors use a business name on the SSEL rather than their personal name about 1.9% of the time. This is computed by (8,874)(.75)/341,315. In the matter of duplication, in contrast to the 1040-C file, only about 5.6% of the parsable names on the SSEL file generate duplicates.

Unduplication

There were several varieties of duplicates among the files produced by the match. In all cases, the pair with the highest match score was designated to be the match. In the event of a tie, the first pair was taken or both discarded depending on the type of duplication. Unduplication proceeded first by EIN then by SSN/Schedule Number.

Over half the duplication was caused by duplicates created in the name parse. In effect, the matcher picks two best candidate for these records. Ties frequently occurred where husband and wife appeared in the name field of both records. For duplicates fitting this pattern, both candidates having the same EIN and the same SSN, the pair with the highest match strength. In event of a tie, the first record was taken.

When pairs were presented with the same EIN and different SSNs, the highest match strength was taken. In the event of a tie, no match was made for that EIN. Family businesses seemed to be the main cause for ties. The file of ties has been retained for further study.

After the EIN side was resolved, the file was resorted to look for instances of the same Schedule C attempting to match more than one EIN record. This occurred almost exactly 1% of the time. Again, if the matcher rated one pair higher than all others, this pair was designated a match. Otherwise, although rarely, the first instance of the tied match strength was taken. An examination of the file of duplicates and winners revealed the following common pattern. Husband and wife had distinct businesses on the SSEL, each under

their individual names. In theory each business should correspond to a distinct schedule, but for some reason one of the businesses did not fit any of the schedules. If the 1040-C had only one schedule the same error would occur.

The following table gives the unduplication by type and the final number of designated matches.

Unduplication Type	No. of Records
Match Pairs Before Unduplication	156,836
EIN Unduplication (records dropped)	
Same EIN, same SSN, same schedule no.	5,211
Same EIN, same SSN, diff. sched. no.	100
Same EIN, diff. SSN, same match strength	506
Same EIN, diff. SSN, lower match strength	2,184
SSN Unduplication (records dropped)	
Same SSN, diff. EIN, same schedule	156
Match Pairs After Unduplication	148,679

10. Match Error Rates

Modeling the Error

Our approach to error estimation is to study a population, similar to the match population, where the link between pairs has already been established. From a 1 in 50 master sample of the original 1040-C file, we identified 18,595 records that reported an EIN on the 1040-C, are known matches to a record on the SSEL, and meet the following conditions:

1. the EIN was valid
2. the EIN was reported on only 1 schedule
3. the EIN record had a sole-proprietorship LFO compatible with a 1040-C filing
4. the pair passed a mild payroll/receipts edit
5. the name field on the EIN record was parsable
6. the EIN had positive 1992 payroll on the SSEL

The count of the 1 in 50 sample that parsed, and including the 18,595 links based on a reported EIN, was 559,514 records. This set of record was used to model two situations: first, where there existed a 1040-C that ought to be linked to the SSEL record ("matchable"); and second, where no record should be linked to an SSEL record ("unmatchable"). By including or excluding the 18,595 linked 1040-C records, and always retaining the linked SSEL records, we modeled both conditions.

False Match Rate for "Matchable" Records

A match was performed with the 559,514 parsed 1040-C records against approximately 361,000 parsed SSEL records. The 1040-C file contained 18,595 linked records, the remaining 540,919 were used to represent the 25,429,164 parsed 1040-C records, giving each a weight of 47. The results of the match on the known links were as follows:

Condition	No. of Records
True matches	16,364
Type A false matches	100
Type B false matches	1
False non-matches	2,130
Total	18,595

The type A false matches involved a correct linkage between EIN and SSN, but with the incorrect schedule number. This can only happen within the sample of 559,514, since the sample was based on SSN and every (prior) linked SSN had all schedules present for the match. Thus, type A false matches represent only themselves. The type B false match involved an incorrect linkage between EIN and SSN, and the one occurrence represents approximately 47 others. Since the event is so rare, we calculated an upper bound and used it in subsequent calculations.

The apparent rate is so low, 1 in 540,919, that we are required to estimate from the binomial probabilities--the normal approximation does not apply and the Poisson approximation is poorest near the mean, where our estimation occurs. We observe that for any $p>4.6/540{,}919$ the probability of getting only 1 occurrence is less than

$$(540{,}919)\ (.0000085)\ (1 - .0000085)^{540{,}919} \approx .05$$

i.e., if the number of type B false matches were greater than 216 (i.e., 4.6 x 47) across all 25.5 million records, we would have less than a 5% chance of getting 1 occurrence in our sample. The question then arises how to distribute the additional 215 estimated false matches between converted true matches and converted false non-matches. We assume a range on the match score of a false match from 15.15 to 20.51, where 20.51 was the highest false match score observed during all testing. This range contains 1456 of the true matches--those which are in a range where it is fairly believable that they can be supplanted by a false match. Assuming an even distribution between these and the false nonmatch set, the additional false matches should be allocated by the proportion 1456:2130 or 87:128, i.e., we will take 87 from the true match count and 128 from the false nonmatch count. In the following, we estimate the results if we

were to run against the whole 1040-C file:

Condition	Frequency	% of File
True match	16,277	87.5%
Type A false match	100	.5%
Type B false match	216	1.2%
False non-match	2,002	10.8%

False Match Rate for "Unmatchable" Records

We reran the match excluding the 1040-Cs belonging to the truth set. 4 matches were produced, linking a SSEL truth record to a 1040-C when the true 1040-C was suppressed. We can only have (type B) false matches and true non-matches on this set. The highest match strength among the 4 was 20.5. Again resorting to the binomial calculation, we estimate the upper bound for the number of false matches in a hypothetical run of the whole file to be 390. That is a false match rate of 2.1% on "unmatchable" records.

Error Estimation

We are now in a position to recover the composition of the original SSEL file. Let x be the number of "matchable" records and y be the number of "unmatchable" records. Then, using upper bounds and adding the type A and B false match rates, .5% + 1.2% = 1.7%, we must solve the simultaneous equations:

$$(.875+.017)x + .021y = 148,679$$
$$x + y = 332,443$$

The solution is x = 162,684, and y = 169,759.
From this we computed the upper estimate of the false match rate to be 4.3% as indicated below.

$$(.017x162,684 + .021x169,759)/148,679 = .043$$

This is an upper bound only. A similar calculation on the point estimate gives an error rate of 2%.

11. Conclusion

We present here the impact of our matching efforts in the context of the overall task of linking SSEL payroll records of sole-proprietors with their 1040-C tax return. For the 1992 tax year, about 1.37 million sole-proprietors had their 1040-C tax return linked to their payroll records on the SSEL. The breakdown by source of linkage is given below.

Source	% of Linked Cases
An EIN reported on 1040-C	71
Use of the EIN-SSN x-ref file	15
Use of the x-ref file & matching	4
Matching on name, address, etc.	10

After all attempts at linking, we still have about 200,000 inscope sole-proprietors on the SSEL for which we could not post receipts. Of these, we estimate that about 170,000 had no 1040-C on the file we used. We believe this may be due to non-filing of the 1040-Cs because of extensions or other late filings. We estimate that about 6,000 were linked but failed to have receipts posted because they failed a payroll to receipts edit. About 9,000 are due to parse failures. We estimated that there are about 15,000 false non-matches, i.e., a 1040-C was on the file but the matching program failed to link to it.

In conclusion, the links for the 1992 censuses were much more complete than for the 1987 censuses, since in 1987 we used only the reported EIN on the 1040-C for linking purposes. The EIN-SSN x-ref file from IRS provided substantial additional links. These were nearly equaled by our matching. Overall, the matching operations were quite efficient, and added significantly to the quality of the 1992 censuses.

References

Cochran, W.G., "Sampling Techniques", Third edition, J. Wiley.

Fellegi, I. P. and Sunter, A. B., "A Theory for Record Linkage", Journal of the American Statistical Association, 1969.

Konschnik, C., Black, J., Moore, R., and Steel, P., (1993), "An Evaluation of Taxpayer-Assigned Principal Business Activity (PBA) Codes on the 1987 Internal Revenue Service (IRS) Form 1040, Schedule C," 1993, Proceedings of the International Conference on Establishment Surveys, American Statistical Association, 745-750.

Winkler, W. E., "Comparative Analysis of Record Linkage Decision Rules," Proceedings of the Section of Survey Research Methods, American Statistical Association, 1992 (to appear).

Winkler, W. E., "Matching and Record Linkage," in B. G. Cox (editor), Survey Methods for Business Farms and Institutions, New York: J. Wiley (to appear).