DETERMINING RECORD LINKAGE PARAMETERS USING AN ITERATIVE LOGISTIC REGRESSION APPROACH

David S. Harville and Richard A. Moore[1], US Bureau of the Census
Richard A. Moore, US Bureau of the Census, Room 1176- FB3, Washington, DC 20233

Key Words: Record Linkage, Logistic Regression, Iteration

1. Abstract. Fellegi and Sunter (1969) developed an algorithm for linking records. Each pair of records is scored on a field by field basis. If the records agree on a given field, a bonus is added to the score. If they disagree, a penalty is subtracted from the score. The records are designated as a match provided the score exceeds a certain threshold. In order to employ this method, one has to determine which fields to compare, bonuses and penalties for each field, an appropriate threshold for designating matches, and the accuracy of a particular linkage, based on its score. Parameters for many previous record linkage applications have been set using an iterative approach involving manual review and re-calibration of the matching fields, bonuses, penalties, and the threshold. The results of such ad hoc procedures can be difficult to replicate, justify, and interpret. This paper describes a more rigorous methodology for setting these parameters based on an iterative logistic regression approach. The method also facilitates the evaluation of the accuracy of the resulting linkages.

2. Introduction. When using administrative lists, it is always convenient and often necessary to combine all the information about each entity onto a single record. To do so requires some type of record linkage procedure, either an exact, statistical, or probabilistic match. An *exact match* is a linkage performed by combining information on records with a common, unique identifier (such as a Social Security number (SSN) or an Employer Identification Number (EIN)). These matches are rarely incorrect. A *statistical match* links pairs of records having a similar set of characteristics, such as name and address. For statistical matches, linked records need not correspond to the same entity. If the lists are created independently, one can expect some disagreement between the fields. For example, one file may contain nicknames, abbreviations, and/or outdated addresses. Hence, pairs do not need to agree on all fields in order to be linked as the "same" entity. A *probabilistic match* (also referred to as *an exact match using probabilistic methods* in the literature*)* is a statistical match where one is able to estimate the likelihood of being correct for each resulting match.

The *scoring procedure* provides a way for linking records. Records from File A are first paired with records from File B. Initially the pair is assigned a score of zero. Each corresponding pair is then compared (on a field by field basis) for a select set of *matching fields.* When values in a given field agree, a bonus is added to the score. When they disagree, a penalty is subtracted from the score. Bonuses and penalties vary from field to field. Their magnitudes are determined by how well each variable determines a correct or incorrect linkage. Once all fields have been compared, the pair's score is recorded. A record from File A is linked to a record from File B, provided the pair has the largest score and this score exceeds a certain *acceptance threshold.* Even though points can be assigned so that all resulting linkages seem reasonable, the scoring procedure does not define a statistical match unless the values are assigned to the bonuses and penalties in a manner that one can use the final score to estimate the likelihood that each resulting linkage is correct.

Ivan Fellegi and Alan Sunter (1969) developed an algorithm to determine a bonus and penalty for each field. Their algorithm involves fairly complicated mathematics, concepts of conditional probability, odds ratios, and assumptions of independence between the matching fields. Under the Fellegi-Sunter algorithm, one can use the final score to predict the likelihood that each pair was correctly linked. Although difficult to understand and program, the algorithm transforms the scoring procedure from a statistical to a probabilistic match.

William Winkler (1995) of the U.S. Bureau of the Census developed record linkage software that employs the scoring procedure above. Steel and Konschnik (1994)

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

used this software to successfully link files of administrative tax records. Their method assigned points using the following trial-and-error approach: (1) use automatic parameter software to determine which fields to compare, agreement bonuses, disagreement penalties, and the acceptance threshold, (2) simulate the linkage on a truth deck (i.e., a set of two files which contain the fields used to link as well as a unique identifier (e.g, an EIN)), (3) evaluate the accuracy of the linkages, (4) adjust the parameters and repeat Steps (2) through (4), and (5) continue with this iterative procedure until a satisfactory result is obtained.

Although this procedure gave satisfactory results when the parameters were applied to files on 1992 tax data, it has several drawbacks. First, it is difficult to replicate. Second, it is difficult to justify why it works well. Third, it is difficult to interpret the results.

This paper describes a more straight-forward approach for setting these parameters. The technique is based on multiple logistic regressions. It is, therefore, easy to use, to replicate, to justify, and to interpret the results. In addition, we believe the approach to be consistent with that of Fellegi-Sunter. Section 3 of this paper discusses the scoring procedure. Section 4 briefly examines the Fellegi-Sunter algorithm for determining bonuses and penalties. Section 5 compares logistic regression with the Fellegi-Sunter algorithm and the scoring procedure. Section 5 evaluates a large application. Section 6 is the conclusion.

3. The Scoring Procedure. Let B_i ($B_i > 0$) be the bonus assigned for agreement of values of the i-th field. Let P_i ($P_i < 0$) be the corresponding penalty for disagreement. Let A = set of i's where the values of the i-th fields agree. We can write the final score as

$$SCORE = \sum_{i \in A} B_i + \sum_{j \notin A} P_j. \qquad (3.1)$$

Theorem 3.1. Let $X_i = +1$, when the values of the i-th field agree, and $X_i = -1$, otherwise. Then there exist unique coefficients, $\{a_i\}$, and translations, $\{t_i\}$, such that

$$SCORE = \sum_{i=1}^{n} a_i * (X_i + t_i). \qquad (3.2)$$

Proof. In order for Eq. (3.1) and (3.2) to be consistent, we need

$$\textit{With Agreement:} \quad a_i * (+1 + t_i) = B_i$$

$$\textit{With Disagreement:} \quad a_i * (-1 + t_i) = P_i$$

Solving this system simultaneously, we find

$$a_i = \frac{B_i - P_i}{2} \quad \textbf{and} \quad t_i = \frac{B_i + P_i}{B_i - P_i} \qquad (3.3)$$

Corollary 3.2 Under the conditions in Theorem 3.1, Eq. (3.2) can be written as

$$SCORE = a_0 + \sum_{i=1}^{n} a_i X_i, \qquad (3.4)$$

with the a_i unique.

Proof. Theorem 3.1 guarantees $a_1, a_2, ..., a_n$ to be unique. Expanding (3.2) and comparing coefficients, we find that

$$a_0 = \sum_{i=1}^{n} a_i t_i. \qquad (3.5)$$

Theorem 3.1 also guarantees the uniqueness of the t_i's, therefore a_0 is forced to be unique.

Example 3.1. You want to statistically match two files by comparing name and city fields. The table below gives the corresponding bonuses and penalties.

Field	Agreement Bonus	Disagreement Penalty
NAME	+11	-7
CITY	+3	-9

Use Eq. (3.3) to find that $a_1 = 9$, $a_2 = 6$, $t_1 = 4/18$, and $t_2 = -6/18$. Substitute these values in Eq(3.2) and simplify to get Eq. (3.4), namely,

$$SCORE = -1 + 9X_1 + 6X_2 . \qquad (3.6)$$

We have finished our motivation of the scoring procedure. We will now concentrate on the Fellegi-Sunter algorithm for determining the bonuses and penalties.

4. Fellegi-Sunter Algorithm for Determining Bonuses and Penalties. Suppose one is given two files, File A and File B. Further suppose that some but not all of the entities on each file contain unique identifying numbers (e.g., SSN's for a file of individuals). One wishes to compare the agreement patterns of another set of matching fields (e.g, name and addresses) to identify supplementary linkages for which the entity's record does not contain the unique identifier on at least one of the files. Fellegi and Sunter (1969) realized that one could use the records from each file which contain unique identifiers to model the agreement pattern of the set of matching variables against the likelihood that the resulting linkage was correct. They proposed the following methodology for determining bonuses and penalties.

Step 1. Assume that the set of matching fields contains a single field. Pair the records. For a given pair, let $X_1 = +1$, if this field agrees , and let $X_1 = -1$, if we have disagreement. Then the probability that a linkage is correct given a value of X_1 is given by the formula:

$$PROB(COR \mid X_1) = \frac{PROB(COR \cap X_1)}{P(X_1)} \quad . \qquad (4.1)$$

Similarly, the probability of an incorrect linkage is

$$PROB(INC \mid X_1) = \frac{PROB(INC \cap X_1)}{P(X_1)} \quad . \qquad (4.2)$$

Given a value of X_1, the odds that the linkage is correct is given by the ratio of Eq (4.1) over Eq. (4.2), namely,

$$(ODDS \mid X_1) = \frac{PROB(COR \cap X_1)}{PROB(INC \cap X_1)} \quad . \qquad (4.3)$$

We are assuming some degree of correlation between the agreement matching variable and the correctness of the resulting linkage, otherwise the variable should not have been chosen as a matching field. Therefore, we expect the odds ratio (4.3) to be greater than 1, if there is agreement ($X_1 = +1$); and the odds ratio to be less than 1, if there is disagreement ($X_1 = -1$).

Step 2. Suppose we now have **n** matching fields. Calculate odds ratios in the form of Eq. (4.3) for each. Assuming the agreement of each matching field is independent of agreement on the other fields, we can then multiply the individual odds ratios together to get an overall odds ratio for the agreement pattern. Namely,

$$(ODDS \mid (X_1,X_2,\ldots,X_n)) = \pi_{i=1}^{n} \frac{PROB(COR \cap X_i)}{PROB(INC \cap X_i)} \quad . \qquad (4.4)$$

Step 3. Take the log transform of Eq. (4.4). For a given set $(X_1, X_2, \ldots , X_n)$, we have

$$\ln(ODDS) = \sum_{i=1}^{n} \ln \frac{PROB(COR \cap X_i)}{PROB(INC \cap X_i)} \quad . \qquad (4.5)$$

Refer to the note at the end of Step 1. If $X_i = +1$, then the i-th term in the summation is positive. If $X_i = -1$, then the i-th term in the summation is negative. Consequently, Fellegi and Sunter make the following definitions:

$$B_i = \ln \frac{PROB(COR \cap X_i = +1)}{PROB(INC \cap X_i = +1)}$$

$$P_i = \ln \frac{PROB(COR \cap X_i = -1)}{PROB(INC \cap X_i = -1)} \qquad (4.6)$$

$$SCORE = \ln \; (ODDS \mid (X_1, X_2, \ldots, X_n) \;)$$

Substituting the identities in (4.6) into Eq. (4.5), we get Eq. (3.1).

Step 4. For a given agreement pattern, the Fellegi-Sunter bonus-penalty system yields a final score which is the natural logarithm of the odds that the match is correct. This yields the following identities:

$$SCORE = \ln \frac{LIKE}{(1 - LIKE)}$$

$$LIKE = \frac{e^{SCORE}}{1 + e^{SCORE}} \quad . \qquad (4.7)$$

Since the likelihood that the linkage is correct can be expressed as a function of the final score, using the Fellegi-Sunter approach to define the bonuses and penalties makes the scoring procedure a probabilistic match. This alleviates much of the subjectiveness of any statistical match. The natural question is, "Do we have to perform all the messy tabulations in Eq. (4.6) to obtain bonuses and penalties which make the scoring procedure a probabilistic match?" We will now show that the answer is "No."

5. A Logistic Regression Approach for Determining Bonuses and Penalties. A closer inspection of the Fellegi-Sunter technique reveals that it is very strongly related, if not equivalent, to logistic regression. Logistic regression software is a common component in any of today's multivariate analysis packages. We will now use this software to derive bonuses and penalties.

The requirements for this routine are similar to those of the Fellegi-Sunter. We are given two files, File A and File B with unique identifying numbers for some but not all of the entities on each file. We wish to use the records with the unique identifiers to model likelihood of a correct match against the agreement patterns of a set of matching fields.

Step 1. Pair all potential linkages and develop agreement patterns in terms of the X_i ($X_i = \pm 1$); then define $Y = +1$, if the unique identifiers agree, and $Y = -1$, otherwise. For each pair, create the vector $(X_1, X_2, \ldots, X_n, Y)$.

Step 2. Use logistic regression, to model the likelihood of Y as a function of the X_i. This will result in an equation similar to Eq. (3.4):

$$SCORE = a_0 + \sum_{i=1}^{n} a_i X_i. \qquad (5.1)$$

Problem. How do we derive bonuses and penalties from Eq. (5.1) that are consistent with the Fellegi-Sunter approach? From Corollary 3.2, we know that

$$SCORE = a_0 + \sum_{i=1}^{n} a_i X_i, \qquad (5.2)$$

with the a_i unique for the set of Fellegi-Sunter bonuses and penalties. The same corollary guarantees that there exist unique t_i's, such that

$$a_0 = \sum_{i=1}^{n} a_i t_i. \qquad (5.3)$$

If we can determine the t_i, we can combine Eq. (5.2) and (5.3) and then use Theorem 3.1 to obtain the bonuses and penalties

$$B_i = a_i * (+1 - t_i) \quad \textbf{and}$$
$$P_i = a_i * (-1 + t_i)\ , \ \textit{for } i = 1, 2, \ldots, n. \qquad (5.4)$$

Step 3. Suppose that we select a proper subset of the X_i and model the logistic regression. We would get

$$SCORE^* = a_0^* + \sum_{i=1}^{n} a_i^* X_i\ , \qquad (5.5)$$

where $a_i^* = 0$ when the i-th matching variable has been dropped from the model. Analogous to Eq. (5.3), we can assume that there exist unique t_i^*, such that

$$a_0^* = \sum_{i=1}^{n} a_i^* t_i^*. \qquad (5.6)$$

Note that when $a_i^* \neq 0$ in Eq. (5.5), it will probably differ from that of a_i in Eq. (5.2). Note also that the t_i^* in Eq. (5.6) will probably differ from the t_i in Eq. (5.3). Assume that we carefully choose a subset such that $a_i^* \approx a_i$, for all values i where the i-th matching variable appears in the subset. Under this condition, we will assume $t_i^* = t_{i.}$. This allows us to eliminate the "*" on the t_i's in Eq. (5.6), so

$$a_0^* = \sum_{i=1}^{n} a_i^* t_i. \qquad (5.7)$$

Eq. (5.3) and (5.7), now define a system of 2 linear equations in n unknowns (the t_i's). If we select a different proper subset, which generate coefficients $a_i^* \approx a_{i.}$, we will be able to create a third linear equation for this system. We can continue to iteratively select proper subset, model, and add equations to the system until we have generated **n** linearly independent equations.

Step 4. Solve the system of equations to find the unique values for each t_i. Then substitute into Eq. (5.4) to obtain the desired bonuses and penalties.

Example 5.1. Let's assume the values in the hypothetical Example 3.1 were determined using the Fellegi-Sunter approach. If we used a logistic regression modeling the likelihood of agreement against name (Variable 1) and city (Variable 2), we would get

$$SCORE = -1.0 + 9.0X_1 + 6.0X_2\ . \qquad (5.8)$$

Theorem 3.1 tells us that $a_1 = 9$ and $a_2 = 6$.

Suppose when we eliminate the name variable (X_1) and remodel, we get

$$SCORE = -2.9 + 0.0X_1 + 5.8X_2\ . \qquad (5.9)$$

Since $a_2 \approx a_2^*$, we will assume the translations associated with Eq. (5.8) and (5.9) are approximately equal. We can use Corollary 3.2 to generate a system of equations in unknowns t_1 and t_2. Solving this system, we find that $t_1 = 2/9$ and $t_2 = -1/2$.

We can now use Eq. (5.4) to generate the following table of bonuses and penalties for our hypothetical example.

Field	Agreement Bonus	Disagreement Penalty
NAME	+11	-7
CITY	+3	-9

6. Application. We now set out to apply the principles above to the following linkage problem. We have two lists of individual tax returns with similar information. We seek to determine parameters (i.e., bonuses and penalties) which will statistically link the two files. The files are described in more detail below.

Quarterly tax return file. List 1 is the list of all quarterly tax returns. Every business is required to submit this return for each quarter in which it compensates individual(s) with wages for work . The file contains the following information: (1) the operation's unique Employer Identification Number (EIN), (2) the proprietor's name, (3) a mailing address, (4) the operation's Standard Industrial Code (SIC), and (5) the operation's quarterly payroll.

The EIN is the control number used by the Internal Revenue Service (IRS) to record wages. All quarterly tax returns contain this number. There are no restrictions placed on the proprietor's name. Some forms contain nicknames, others contain initials. Use of a middle name, middle initial, or a generational suffix (Jr., Sr., III,...) is left to the discretion of the filer. Likewise, the filer is given latitude for the mailing address. He can use either the address of his home residence or that of the physical location of the operation. The SIC code is supplied only for statistical purposes. A large percentage of returns contain no SIC. For returns which contain an SIC, there is no guarantee that the code assigned is accurate.

Annual sole proprietorship income tax return (IRS Form 1040, Schedule C) file. List 2 is the file of all sole proprietorship tax returns. This file contains (1) the individual's Social Security Number (SSN), (2) the EIN of his operation, (3) the proprietor's name, (4) his mailing address, (5) his operation's SIC, and (6) his annual gross receipts.

The SSN is the control number that the IRS uses to record gross receipts and taxes paid. It appears on all records in this file. There is space on the IRS Form 1040, Schedule C for the filer to provide an EIN, if he chooses to do so. Most forms do not contain an EIN. If supplied, there is no guarantee that the individual has transcribed the number correctly. The quality of the information in the name, address, and SIC fields on this form also suffers from the same problems found in the quarterly payroll file.

Problem. There are about 400,000 sole proprietorships (EINs), which exhibit payroll in at least one quarter of 1997, for which no corresponding EIN could be found in the IRS Form 1040, Schedule C file. We want to probabilistically link these 400,000 records to a record in the Form 1040, Schedule C file. When finished, we will have a file of all sole proprietorships with the quarterly payroll and receipts of each operation on a single record.

Solution. As previously stated, Steel and Konschnik (1994), solved an analogous problem for tax year 1992, using parameters determined by an iterative trial-and-error approach. Using the logistic regression method described earlier in this paper, we revisited this problem and determined the parameters in the table below.

Variable	Agreement Bonus	Disagreement Penalty
First Name	0.01	-3.00
Middle Initial	0.01	-3.89
Middle Name	0.01	-1.21
Last Name	0.01	-2.47
Jr/Sr/...	1.96	-1.98
Street	5.00	-0.01
City	4.24	-0.02
State	0.03	-5.37
2-digit SIC	1.51	-0.01
4-digit SIC	0.55	-0.25

Notice that the state fields and all name fields, except Jr/Sr/..., have a negligible bonus and a large penalty. When matching tax files, you expect the name fields and the state field to agree. Pairs that do match on these fields receive no bonus points, but pairs that don't receive extremely large penalties. Notice also that the city and street fields have large bonuses and negligible penalties. Recall that each filer has the latitude to specify the address (residence or office) to which he wants each form mailed. These points indicate that we don't expect agreement (no penalty for disagreement) for each pair. However, if a pair has the same address, we almost certainly have a correct linkage (indicated by the large bonus). Finally, notice that the Jr/Sr/... and industry fields do not provide a bonus of sufficiently large size to indicate a definite match or a penalty sufficient to disqualify a pair for linking. These fields are primarily used as tie-breakers. They only come into play when the name and address information does not provide a single clear linkage.

Under this algorithm, a payroll record for "Robert A Smith .... 123 East Pratt Street, Baltimore, MD...SIC 5932," would first eliminate all records where (1) the name was not "Robert A Smith" and (2) the state was not "MD". It would then restrict itself to the remaining records whose city was "Baltimore." If there were still ties, it would look for a business on "123 East Pratt Street." As a last resort, it would eliminate records based on Jr/Sr/... or industry information.

Results. We selected a 20 percent systematic sample of sole proprietorships from the quarterly tax returns file. We then selected a 1 percent sample of those returns from the IRS Form 1040, Schedule C file which contained an EIN. These files contained 9,333 records with a common EIN. They were then subjected to two different matching algorithms --- one using the ad hoc parameters developed by Steel and Konschnik; the other using those derived from the logistic regression model. The table below compares the results.

Linkages	Ad Hoc Parameters	Log Reg Parameters
Total	7,110	6,923
Def. Correct*	6,897 (97.0%)	6,853 (99.0%)
Def Incorrect	71 (1.0%)	15 (0.2%)
Can't Tell	142 (2.0%)	55 (0.8%)

* There were a total of 9,333 definitely correct linkages.

After each match was executed, the resulting set of linkages were analyzed for correctness. Linkages were separated into one of three classes: (1) Definitely Correct — linkages where the record had the same EIN, (2) Definitely Incorrect — linkages where the EINs and names differed, but the software treated the difference as a typo (e.g., John Smith to Joan Smith), (3) Can't Tell — linkages where the EINs differed, but the names were similar.

Even though 99 percent of all linkages were definitely correct, we were concerned that the logistic regression parameters were too restrictive, since they identified only 6,853 (or 73%) of the 9,333 definitely correct linkages. We then compared the results with those using Steel and Konschnik's ad hoc parameters. As the table shows the ad hoc parameters generated 187 more linkages, however only 44 of these were definitely correct. We suspect it would be very difficult to identify significantly more definitely correct matches without also picking up a large number of definitely incorrect and can't tell linkages.

8. Conclusion. In conclusion, we believe that the logistic regression model can be used effectively to set parameters for probabilistic record linkage between the file of IRS quarterly payroll return records and the IRS Form 1040, Schedule C tax return records. The method is easy to use, gives good results, and the resulting parameters are consistent with the principles outlined by Fellegi and Sunter.

9. References

- Cochran, W.G. (1977). *Sampling Techniques, Third edition*, New York: J. Wiley.
- Fellegi, I. P. and Sunter, A.B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Society, 64*, pp 1183-1210.
- Harville, D. S. and Moore, R.A. (1998). User's Guide for SRD's Name Standardizing and Record Linkage Software. *U. S. Bureau of the Census*. (internal document).
- Jaro, M. A. (1989). Advances in Record-Linkage methodology as Applied to the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association, 84*, pp 414-420.
- Newcombe, H. B., Fair, M. E., and Lalonde, P. (1992). The Use of Names for Linking Personal Records, *Journal of the American Statistical Association, 87*, pp 1193-1204.
- Sharma, S. (1996), *Applied Multivariate Techniques*, New York: J. Wiley.
- Steel, P. M. and Konschnik, C. A. (1994). Post-Matching Administrative Record Linkage Between Sole Proprietorship Tax Returns and the Standard Statistical Establishment List, *Proceedings of the Section on Survey Research Methodology. American Statistical Society*. pp 473-478.
- Winkler, W. E.,(1995). Matching and Record Linkage. *Business Survey Methods*. (B.G. Cox, et al, editors), New York, J. Wiley, pp 355-384.