Improving EM Algorithm Estimates for Record Linkage Parameters[1]

William E. Yancey
U.S. Bureau of the Census
`william.e.yancey@census.gov`

Abstract

The EM algorithm can be used to estimate conditional probabilities for matching field patterns for the Fellegi-Sunter model for record linkage. The algorithm is based on a latent class model for the record pairs where one of the classes is the set of true matches. If the number of true match pairs in the data set is too small, then the EM algorithm cannot detect the correct latent class. We consider methods for enriching the density of matches in the set of examined record pairs in order to obtain improved EM algorithm estimates for the record linkage conditional probability parameters.
KEY WORDS record linkage, EM algorithm

1. Record Linkage Background

Record linkage is a procedure to find pairs of records in two files that represent the same entity. When the two files are the same file, record linkage can be used to find duplicate records within a file. If we let A and B be two files, the set of record pairs $A \times B$ can be partitioned into two sets M and U, where M represents the set of record pairs where both records represent the same underlying entity (*e.g.* person), and U is the set of record pairs where the records represent different entities. The object of record linkage is to infer which pairs of records belong to M. To do this, we make a set of k comparisons between the records, which we may express as a comparison function

$$f : A \times B \to \Gamma$$

where each element γ of the comparison space Γ is a k-tuple called an *agreement pattern*. For the sake of simplifying the discussion, we assume that all of our comparisons are binary, either agreement or disagreement, so that there are 2^k possible agreement patterns in Γ. The Fellegi-Sunter theory of record linkage (Fellegi, 1969) says that if we know for each agreement pattern γ, the conditional probabilities $\Pr(\gamma|M)$ and $\Pr(\gamma|U)$, the probability that a pair of records produces the agreement pattern γ given that the pair is a match (resp. non-match), then the optimal record linkage rule is formed by ordering the patterns by their weights

$$w_\gamma = \log \left(\frac{\Pr(\gamma|M)}{\Pr(\gamma|U)} \right),$$

choosing cutoff values $w_H > w_L$, and designating as links (inferred matches) all pairs with agreement pattern γ with

$w_\gamma > w_H$, designating as non-links (inferred non-matches) all pairs with agreement pattern γ with $w_\gamma < w_L$, and consigning to the clerical review region all record pairs with agreement pattern γ with $w_L \le w_\gamma \le w_H$. The record linkage rule is optimal in the sense that of all the linkage rules that have the same or better false match and false non-match error rates, this rule produces the minimal clerical review region.

1.1 Conditional Independence

This ideal record linkage rule can only be approximated since the fundamental conditional probabilities are unknown and can only be approximated. The task of approximating these conditional probabilities is usually simplified by making the conditional independence assumption, that for each k-tuple agreement pattern $\gamma = (\gamma_l)$, the conditional probabilities can be computed by

$$\Pr(\gamma|M) = \prod_{l=1}^{k} \Pr(\gamma_l|M) \qquad (1)$$

$$\Pr(\gamma|U) = \prod_{l=1}^{k} \Pr(\gamma_l|U)$$

where $\Pr(\gamma_l|M), \Pr(\gamma_l|U)$ are the marginal conditional probabilities of the lth component of γ.

For census person data, we have achieved better conditional independence results when we use three classes (Winkler, 1995). Since the individual information is in categories that either distinguish individuals or household groups, it is reasonable to partition the non-match class U into

$$U = U' \cup U''$$

where U' indicates the class of record pairs reflecting different individuals in the same household and U'' indicates the class of record pairs from different individuals in different households. We can recover our marginal probability by

$$\Pr(\gamma|U) = \Pr(\gamma|U' \cup U'') \qquad (2)$$
$$= \frac{\Pr(\gamma|U')\Pr(U') + \Pr(\gamma|U'')\Pr(U'')}{\Pr(U') + \Pr(U'')}$$

1.2 The EM Algorithm

This still leaves the problem of estimating the marginal conditional probabilities. We have found that the EM algorithm can produce effective marginal probability estimates specific to a given subset $S \subset A \times B$ directly from the comparison data of the record pairs in S without the use of training data or historical parameter values from other data record

sets (Winkler, 1988). The input data is the agreement pattern counts

$$n_j = \# \left\{ (a,b) \in S \mid f(a,b) = \gamma^j \right\},$$

the number of times a record pair in S produces the agreement pattern γ^j for each $1 \leq j \leq |\Gamma|$. Under the binary comparison assumption, the number of agreement patterns $|\Gamma| = 2^k$. When we collect this count data, we may express the maximum likelihood function as

$$L(S) = \prod_{(a,b) \in S} \Pr(\gamma(a,b)) = \prod_{j=1}^{2^k} \Pr(\gamma^j)^{n_j},$$

where the probabilities $\Pr(\gamma^j)$ are functions of the marginal conditional probabilities and the latent class proportions. Details of the adaptation of the EM algorithm to this application are given in (Yancey, 2002).

The EM algorithm is an iterative numerical procedure to compute values of the parameters that maximize the likelihood function. Dempster, Laird, and Rubin (Dempster, 1977) prove that this algorithm will converge, and in this conditional independence context, we have found that the convergence is efficient and fairly insensitive to initial conditions. However, there is no *a priori* guarantee that the latent classes C_1, C_2, C_3 determined by the EM algorithm will correspond to the classes M, U', U'' that we had in mind. Failure to converge to the desired class parameters happens when the proportion of one of the classes M, U', U'' is too small to be detected by the algorithm. In practice, it is the class M of matches that is generally smallest, and when the proportion of this class drops below 0.05 or so, the EM algorithm can converge to parameter values that are not relevant to record linkage calculation.

1.3 Match Enriched Sampling

Our purpose here is to experiment to see whether, in a case where the matches M are of so small proportion of the total set S of pairs that the EM algorithm fails to converge to classes representing M, U', U'', we can select a sample $S' \subset S$ where the agreement pattern counts $\{n_j\}$ from the subset S' produce meaningful EM parameter estimates that can in turn be used for record linkage on the full set S. The procedure is to modify the agreement pattern counting procedure so that a record pair is added to the count based on a calculation of its preliminary matching weight. We choose some initial marginal probability parameters to compute a preliminary matching weight. We count all pairs whose weight is high; we select a random sample of lower weight pairs. In this way we select a subset S' of record pairs in which the proportion of matches is relatively high.

1.4 One-to-One Matching

An important feature of the Census Bureau record linkage program is that it performs one-to-one matching using

Test Deck	# A	# B	# Pairs	Max Match Prop
STL	15048	12072	18,514,495	0.00065
2021	4539	4859	2,225,132	0.00204
3031	5022	5212	2,766,086	0.00182

Table 1: File Sizes and Blocked Pairs

a linear sum assignment algorithm (Winkler, 1994). Thus if we are comparing m records from one file with n records from another file, where say $m \leq n$, we compute a matching weight array (w_{ij}), $1 \leq i \leq m, 1 \leq j \leq n$, and the linear sum assignment algorithm is used to find an optimal permutation $\hat{\sigma} \in P_n$ such that

$$\sum_{i=1}^{m} w_{i,\hat{\sigma}(i)} = \max_{\sigma \in P_n} \left\{ \sum_{i=1}^{m} w_{i,\sigma(i)} \right\}.$$

The output of the one-to-one matching is only the m record pairs $\left(a_i, b_{\hat{\sigma}(i)} \right)$. In our empirical study, we examined the true and false match distributions of the output of the one-to-one matcher under different matching parameter inputs, not the distributions of all of the pairs (a_i, b_j). Consequently, whenever we use valid marginal probability parameters with $\Pr(\gamma_l|M) > \Pr(\gamma_l|U)$, we generally see a high proportion of true matches in the output. Better parameter choices are evidenced by more effective separation between the true and false matches.

2. Empirical Results

2.1 Census Test Decks

We considered for our test decks three pairs of Census files that have been extensively clerically reviewed to determine match status. We chose a light blocking criterion that would result in the subclass of true match pairs being a small proportion of all of the pairs considered. Specifically, we blocked on the last digit of the cluster number, a fairly random blocking scheme that divides each set into ten subsets, and thus the resulting set of pairs should be around 10% of all possible pairs. The maximum possible proportion of the size of the match class in the set of all pairs is given by

$$\frac{\min\{\#A, \#B\}}{\#\text{Pairs}}$$

which corresponds to every record in the smaller file having matched a record in the larger file. In Table 1, for each test deck, we show the number of records in each file, the number of record pairs brought together under our loose blocking criterion, and the maximum possible proportion of match pairs in this blocking set. We see in Table 1 that the match class in each of the three test decks is well below 1% of the pairs considered.

For each test deck, we use ten matching fields for binary comparison, making a total of 1024 possible matching patterns. The matching fields are: last name, first name, house

Test Deck	$\Pr(C_1)$	$\Pr(C_2)$	$\Pr(C_3)$
STL	0.0542	0.4931	0.4527
2021	0.0318	0.5391	0.4291
3031	0.0262	0.4968	0.4771

Table 2: All Pairs Class Proportion Estimates

Field	$\Pr(\gamma|C_1)$	$\Pr(\gamma|C_2)$	$\Pr(\gamma|C_3)$	$\Pr(\gamma|C_2 \cup C_3)$
last	0.6570	0.0641	0.0286	0.0471
first	0.2296	0.0861	0.0534	0.0705
hsnm	0.9975	0.3190	0.0383	0.1846
strt	0.9841	0.3465	0.0035	0.1823
phone	0.6103	0.0160	0.0215	0.0186
age	0.3674	0.3184	0.0577	0.1936
rel	0.2807	0.4321	0.0001	0.2254
marit	0.5426	0.6113	0.0027	0.3200
sex	0.5968	0.5109	0.5458	0.5276
race	0.9767	0.6382	0.5295	0.5862

Table 3: Marginal Probabilities for Deck STL

Field	$\Pr(\gamma|C_1)$	$\Pr(\gamma|C_2)$	$\Pr(\gamma|C_3)$	$\Pr(\gamma|C_2 \cup C_3)$
last	0.8685	0.0151	0.0194	0.0170
first	0.3643	0.0061	0.0600	0.0300
hsnm	0.9826	0.6683	0.0005	0.3723
strt	0.9824	0.6317	0.1506	0.4185
phone	0.7272	0.0174	0.0088	0.0136
age	0.4646	0.1114	0.4594	0.2656
rel	0.3395	0.1830	0.4260	0.2907
marit	0.6053	0.4175	0.4741	0.4426
sex	0.6417	0.4945	0.5122	0.5023
race	0.9232	0.6915	0.4104	0.5669

Table 4: Marginal Probabilities for Deck 2021

Field	$\Pr(\gamma|C_1)$	$\Pr(\gamma|C_2)$	$\Pr(\gamma|C_3)$	$\Pr(\gamma|C_2 \cup C_3)$
last	0.9332	0.0254	0.0142	0.0199
first	0.4108	0.0098	0.0650	0.0369
hsnm	0.9658	0.5140	0.0248	0.2743
strt	0.7121	0.5968	0.0667	0.3371
phone	0.7885	0.0249	0.0037	0.0145
age	0.4904	0.0974	0.4199	0.2554
rel	0.3208	0.1080	0.5086	0.3042
marit	0.5608	0.3206	0.5811	0.4482
sex	0.6419	0.4670	0.5225	0.4942
race	0.9371	0.7082	0.5144	0.6133

Table 5: Marginal Probabilities for Deck 3031

number, street, 4 digit phone number, age, relationship to head of household, marital status, sex, race.

2.2 Raw EM Estimates

When we run the three class EM algorithm on pattern counts from this full blocking criterion, we get the estimated class proportions shown in Table 2. We see that in all cases the smallest estimated latent class proportion, namely that of class C_1, is at least an order of magnitude larger than the actual proportion of matches. Actually the latent class proportion estimates are fairly consistent from one test deck to the next, but it is difficult to interpret what might characterize these classes. Likewise, the marginal conditional agreement probabilities estimated by the EM algorithm in Tables 3, 4, and 5 do not conform to the values we would expect if the latent classes represent matches, non-matches within household, and non-matches outside household. While the resulting parameters do have the necessary property that

$$\Pr(\gamma|C_1) > \Pr(\gamma|C_2 \cup C_3)$$

so that agreement weights will all be positive and disagreement weights will all be negative, we see that the fields, such as first name and age, that primarily identify individuals have weak distinguishing power.

If we proceed by identifying the first class with the class of matches

$$M = C_1$$

and the other two classes with the class of non-matches

$$U = C_2 \cup C_3$$

then we can compute for each matching field the agreement weight

$$A = \log \frac{\Pr(\gamma|M)}{\Pr(\gamma|U)}$$

the disagreement weight

$$D = \log \frac{1 - \Pr(\gamma|M)}{1 - \Pr(\gamma|U)}$$

and the discriminating power of the field

$$P = A - D.$$

The results for these test decks are summarized in Tables 6, 7, and 8.

2.3 Match Enriched Samples

We now want to recompute our EM parameter estimates using pattern counts from a subset of the files that should contain a higher proportion of the set of matches. In order to do this, we can select to count more of the pairs with higher weights and fewer of the pairs with lower weights, where the matching weights of the pairs has been determined from an *a priori* set of marginal conditional weights. For the record, the initial marginal weights are given in Table 9. We tried counting all pairs with matching weight above 0, and randomly selecting 1% of all pairs with matching weight below 0 and above -6, and 0.1% of all pairs with matching weight

Field	$\Pr(\gamma\mid M)$	$\Pr(\gamma\mid U)$	A	D	P
last	0.6570	0.0471	2.6354	-1.0218	3.6572
first	0.2296	0.0705	1.1807	-0.1877	1.3685
hsnm	0.9975	0.1846	1.6871	-5.7874	7.4745
strt	0.9894	0.1823	1.6915	-4.3456	6.0371
phone	0.6103	0.0186	3.4908	-0.9236	4.4144
age	0.3674	0.1936	0.6407	-0.2427	0.8834
rel	0.2807	0.2254	0.2194	-0.0741	0.2935
marit	0.5426	0.3200	0.5281	-0.3965	0.9246
sex	0.5968	0.5276	0.1232	-0.1584	0.2816
race	0.9767	0.5862	0.5105	-2.8769	3.3875

Table 6: Matching Weights for Test Deck STL

Field	$\Pr(\gamma\mid M)$	$\Pr(\gamma\mid U)$	A	D	P
last	0.8685	0.0170	3.9336	-2.0116	5.9452
first	0.3643	0.0300	2.4968	-0.4226	2.9194
hsnm	0.9826	0.3723	0.9705	-3.5856	4.5561
strt	0.9824	0.4185	0.8533	-3.4977	4.3510
phone	0.7272	0.0136	3.9791	-1.2853	5.2645
age	0.4646	0.2656	0.5592	-0.3160	0.8752
rel	0.3395	0.2907	0.1552	-0.0713	0.2265
marit	0.6053	0.4426	0.3131	-0.3452	0.6582
sex	0.6417	0.5023	0.2449	-0.3286	0.5736
race	0.9232	0.5669	0.4877	-1.7298	2.2174

Table 7: Matching Weights for Test Deck 2021

Field	$\Pr(\gamma\mid M)$	$\Pr(\gamma\mid U)$	A	D	P
last	0.9332	0.0199	3.8479	-2.6860	6.5339
first	0.4108	0.0369	2.4099	-0.4914	2.9013
hsnm	0.9658	0.2743	1.2587	-3.0549	4.3137
strt	0.7121	0.3371	0.7478	-0.8340	1.5819
phone	0.7885	0.0145	3.9960	-1.5389	5.5349
age	0.4904	0.2554	0.6524	-0.3792	1.0316
rel	0.3208	0.3042	0.0531	-0.0242	0.0773
marit	0.5608	0.4482	0.2241	-0.2282	0.4524
sex	0.6419	0.4942	0.2615	-0.3453	0.6068
race	0.9371	0.6133	0.4239	-1.8161	2.2400

Table 8: Matching Weights for Test Deck 3031

Field	$\Pr(\gamma\mid M)$	$\Pr(\gamma\mid U)$
last	0.97	0.10
first	0.99	0.01
hsnm	0.95	0.03
strt	0.96	0.16
phone	0.94	0.05
age	0.79	0.02
rel	0.53	0.29
marit	0.89	0.34
sex	0.91	0.45
race	0.97	0.84

Table 9: A Priori Marginal Weights

Test Deck	# Pairs	Max Match Prop
STL	71,729	0.1683
2021	16,117	0.2816
3031	13,665	0.3675

Table 10: Reduced Counts of Blocked Pairs

Test Deck	$\Pr(C_1)$	$\Pr(C_2)$	$\Pr(C_3)$
STL	0.1597	0.3981	0.4423
2021	0.2131	0.4951	0.2918
3031	0.2744	0.4746	0.2510

Table 11: Reduced Pairs Class Proportion Estimates

below -6. We based this sifting method on a count of pairs from the STL deck. When we binned the pairs by integer weight values, the number of pairs tended to increase in order of magnitude at around 0 and -6, hence we chose to discard proportionally more of them while retaining a sample of the lower weight pairs. This seemed to work better than our previous approach of simply discarding all pairs below a cutoff weight. Using such an absolute cutoff may retain all match pairs, but the included non-match pairs tend to agree on at least a few fields. This tends to raise some of the marginal probability estimates conditioned on non-matching and thus degrades the distinguishing power of some fields. The reduced number of pairs counted in the test decks and the corresponding maximum proportion of matches is given in Table 10 (*cf.* Table 1).

2.4 Enriched EM Estimates

When we run the three class EM algorithm using these reduced counts for data, we get the estimated class shown in Table 11 (*cf.* Table 2). Here we see that the estimated size of class C_1 more reasonably corresponds to a possible proportion of the match class M. The corresponding marginal probability estimates from the EM algorithm are given in Tables 12, 13, and 14 (*cf.* Tables 3, 4, and 5).

Since we want to extend the conditional probability estimates from sample S' to the full set S, in order to compute

Field	$\Pr(\gamma\mid C_1)$	$\Pr(\gamma\mid C_2)$	$\Pr(\gamma\mid C_3)$	$\Pr(\gamma\mid C_2\cup C_3)$
last	0.9341	0.7571	0.0038	0.0050
first	0.9885	0.0490	0.0267	0.0268
hsnm	0.9506	0.9874	0.1640	0.1653
strt	0.9650	0.9874	0.2587	0.2598
phone	0.6749	0.6312	0.0001	0.0011
age	0.9024	0.2659	0.1091	0.1093
rel	0.4847	0.2985	0.1373	0.1375
marit	0.8642	0.6044	0.3074	0.3078
sex	0.9836	0.5341	0.4346	0.4347
race	0.9746	0.9649	0.5063	0.5070

Table 12: Marginal Probabilities for Deck STL

Field	$\Pr(\gamma|C_1)$	$\Pr(\gamma|C_2)$	$\Pr(\gamma|C_3)$	$\Pr(\gamma|C_2 \cup C_3)$
last	0.9479	0.7074	0.0135	0.0160
first	0.9769	0.0522	0.0450	0.0451
hsnm	0.9332	0.9888	0.0386	0.0420
strt	0.9507	0.9948	0.4203	0.4224
phone	0.6721	0.6216	0.0015	0.0037
age	0.8847	0.2608	0.2758	0.2758
rel	0.4862	0.2727	0.2255	0.2257
marit	0.8513	0.5158	0.4886	0.4887
sex	0.9753	0.5190	0.6040	0.6037
race	0.9226	0.8990	0.6755	0.6764

Table 13: Marginal Probabilities for Deck 2021

Field	$\Pr(\gamma|C_1)$	$\Pr(\gamma|C_2)$	$\Pr(\gamma|C_3)$	$\Pr(\gamma|C_2 \cup C_3)$
last	0.9327	0.6783	0.0071	0.0087
first	0.9616	0.0645	0.0284	0.0284
hsnm	0.9280	0.9890	0.0218	0.0241
strt	0.6422	0.8693	0.1133	0.1150
phone	0.6786	0.5996	0.0000	0.0014
age	0.8645	0.3029	0.1757	0.1760
rel	0.4521	0.2142	0.1282	0.1284
marit	0.7686	0.4941	0.3540	0.3543
sex	0.9827	0.5076	0.5721	0.5720
race	0.9295	0.8961	0.5297	0.5306

Table 14: Marginal Probabilities for Deck 3031

$\Pr(\gamma|C_2 \cup C_3)$ using (2), we need to re-estimate the class proportions $\Pr(C_2)$ and $\Pr(C_3)$. Essentially we assume that all of the uncounted pairs from S are in class C_3, which probably is not quite the case, but the class C_3 in S should be by far the largest. Specifically, we estimate the class proportions in S by

$$\Pr(C_2)_S = \frac{|S'|\Pr(C_2)}{|S|}$$

$$\Pr(C_3)_S = \frac{|S'|\Pr(C_3) + |S - S'|}{|S|}$$

2.5 Comparison of Match Results

We see in Tables 15, 16, and 17 that the discriminating power of the variables, especially the individual matching variables like first name and age has been increased over the values derived from the full file parameter estimates as shown in Tables 6, 7, and 8.

The sampled pair counts have produced significantly different marginal probability estimates for some of the fields, but we want to see what effect the new parameter estimates have on the matching output. One effect is that by generally increasing the discriminating power, the range of possible pair weights from total agreement to total disagreement widens. Thus in order to compare the outputs of the matcher using different marginal probability parameters, we rescale the out-

Field	$\Pr(\gamma|M)$	$\Pr(\gamma|U)$	A	D	P
last	0.9341	0.0050	5.2302	−2.7146	7.9448
first	0.9885	0.0268	3.6078	−4.4382	8.0460
hsnm	0.9506	0.1653	1.7493	−2.8271	4.5765
strt	0.9650	0.2598	1.3122	−3.0516	4.3638
phone	0.6749	0.0011	6.4193	−1.1225	7.5418
age	0.9024	0.1093	2.1110	−2.2111	4.3221
rel	0.4847	0.1375	1.2599	−0.5151	1.7750
marit	0.8642	0.3078	1.0324	−1.6287	2.6611
sex	0.9836	0.4347	0.8166	−3.5401	4.3566
race	0.9746	0.5070	0.6535	−2.9658	3.6193

Table 15: Matching Weights for Test Deck STL

Field	$\Pr(\gamma|M)$	$\Pr(\gamma|U)$	A	D	P
last	0.9479	0.0160	4.0820	−2.9388	7.0208
first	0.9769	0.0451	3.0761	−3.7204	6.7965
hsnm	0.9332	0.0420	3.1001	−2.6625	5.7626
strt	0.9507	0.4224	0.8112	−2.4602	3.2715
phone	0.6721	0.0037	5.2014	−1.1112	6.3126
age	0.8847	0.2758	1.1657	−1.8373	3.0030
rel	0.4862	0.2257	0.7674	−0.4101	1.1775
marit	0.8513	0.4887	0.5550	−1.2348	1.7897
sex	0.9753	0.6037	0.4796	−2.7753	3.2549
race	0.9226	0.6764	0.3105	−1.4307	1.7412

Table 16: Matching Weights for Test Deck 2021

Field	$\Pr(\gamma|M)$	$\Pr(\gamma|U)$	A	D	P
last	0.9327	0.0087	4.6809	−2.6899	7.3708
first	0.9616	0.0284	3.5208	−3.2319	6.7527
hsnm	0.9280	0.0241	3.6516	−2.6073	6.2589
strt	0.6422	0.1150	1.7198	−0.9057	2.6255
phone	0.6786	0.0014	6.1781	−1.1335	7.3116
age	0.8645	0.1760	1.5918	−1.8052	3.3971
rel	0.4521	0.1284	1.2588	−0.4642	1.7230
marit	0.7686	0.3543	0.7745	−1.0264	1.8009
sex	0.9827	0.5720	0.5412	−3.2066	3.7478
race	0.9295	0.5306	0.5607	−1.8956	2.4563

Table 17: Matching Weights for Test Deck 3031

Score Region		Raw EM	Enrich EM
$s > 0.55$			
	Total Pairs	9914	9525
	#True Matches	9425	9452
	False Match Rate	4.93%	0.77%
$0.35 < s < 0.55$			
	Total Pairs	544	589
	#True Matches	127	419
	False Match Rate	76.65%	28.86%

Table 18: STL Sample Linkage Rule Comparison

Score Region		Raw EM	Enrich EM
$s > 0.6$			
	Total Pairs	3602	3584
	#True Matches	3509	3522
	False Match Rate	2.58%	1.73%
$0.4 < s < 0.6$			
	Total Pairs	313	259
	#True Matches	55	62
	False Match Rate	82.42%	76.06%

Table 19: 2021 Sample Linkage Rule Comparison

Score Region		Raw EM	Enrich EM
$s > 0.6$			
	Total Pairs	3583	3552
	#True Matches	3488	3503
	False Match Rate	2.65%	1.38%
$0.4 < s < 0.6$			
	Total Pairs	375	263
	#True Matches	75	82
	False Match Rate	80%	68.8%

Table 20: 3031 Sample Linkage Rule Comparison

put weight to a score s, where

$$s = \frac{w - w_{\min}}{w_{\max} - w_{\min}}$$

where w is the pair's output weight by the matcher and $w_{\max}$ and $w_{\min}$ are the maximum and minimum weights output by the matcher respectively. Thus we can compare the outputs of all matching runs on a scale of 0 to 1. Graphs of the output distributions of the matches and non-matches can be seen in (Yancey, 2002).

To get an indication of the improved separation between matches and non-matches, we can consider drawing comparable cutoff level scores for links and clerical regions for the deck. For example, we can see in Table 18 that if we designate as links pairs with a score above 0.55 (actually, since the bar graphs are centered at midpoints, this cutoff score is 0.525), the enriched EM parameters produce a set of links with some more true matches and a lot lower false match rate. If we designate the clerical region by the pairs included in bars with scores between 0.35 and 0.55, while the enriched EM distribution contains somewhat more pairs to consider, the clerical region has a much higher ratio of matches to non-matches. For the other two test decks, we set the cutoff levels slightly higher. For test decks 2021 and 3031, if we take the designated links to be contained in the frequency bars with score 0.6 and above and the clerical region to be in the bars with scores from 0.4 up to 0.6, then for both cases, the enriched EM matching provides more true matches with a lower false match rate in the designated links and a clerical region with fewer pairs, more matches, and a lower false match rate, as seen in Tables 19 and 20 respectively.

3. Conclusions

The EM algorithm can be an effective unsupervised learning technique for estimating parameters for record linkage. When the class of match pairs is too sparse for the algorithm to conform to the appropriate classes, it is possible to sample the pairs to obtain a match-enriched subset, obtain reasonable EM parameter estimates, and extend these estimates for use in record linkage for the entire file. The sample can be taken without clerical input, using default parameters, and examining the default matching weight statistics of the output. The use of these sample-based EM parameters can result in improved separation of matches and non-matches in a record linkage program.

References

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society,* Series B, **39**, pp. 1–38.

Fellegi, I. P. and A. B. Sunter (1969). "A Theory for Record Linkage." *Journal of the American Statistical Association.* **64**, pp. 1183–1210.

Winkler, W. E. (1988). "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage." *Proceedings of the Section on Survey Research Methods.* American Statistical Association. pp. 667–671. (longer version report rr00/05 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1994). "Advanced Methods for Record Linkage." *Proceedings of the Section on Survey Research Methods.* American Statistical Association. pp. 467–472. (longer version report rr94/05 available at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (1995). "Matching and Record Linkage." in B. G. Cox *et. al.* (ed.) *Business Survey Methods.* New York: J. Wiley, pp. 355–384

Yancey, W.E. (2002) "Improving EM Algorithm Estimates for Record Linkage Parameters," http://www.census.gov/srd/www/byname.html