The Estimation of Match Validity under the Fellegi-Sunter Paradigm without Assuming Identifier-Agreement Independence

Dean M. Resnick[1]
[1]National Opinion Research Center at the University of Chicago, 4350 East-West
Highway, Suite 800, Bethesda MD 20814

Abstract
The Fellegi-Sunter record linkage paradigm specifies the functional relationship between agreement probabilities and match weights for each identifier. This paradigm assumes conditional independence of identifier agreements. However, many identifier agreements are, in fact, dependent. For example, within the set of non-matched pairs, if we know the first names agree, then it is more likely that last names also agree since name distributions vary by ethnicity. In this paper we present an approach to specify and estimate agreement probabilities, the relationship between them, and the total number of links without the use of training data. This, in turn, yields estimates of match rates (i.e., the proportion of matches among a set of pairs) for a given identifier agreement pattern.

Key Words: Record Linkage, Fellegi-Sunter, Machine Learning, Optimization

1. Record Linkage Overview

Record Linkage describes the process of comparing records (those being compared called a pair) on a single file or multiple files to infer which represent the same entity, be it a person, business, address, or something else. Comparisons are made by looking at the similarity of the values for identifiers that are on the two records being compared. For example, if records represent persons, we can look at whether or not values ***agree*** (that is, they are exactly or more-or-less the same[1]) for first name, last name, year-of-birth, sex, race, etc. What makes record linkage different from a standard data join is that it allows for some discrepancies between the fields used for comparison. For this reason, sometimes it is called fuzzy matching.

There are many ways we can imagine performing such an analysis, like developing ad-hoc rules to infer which pairs are matches (i.e., they represent the same entity) and which pairs are non-matches. However, the standard approach is based on a methodology whose application was statistically formalized by Ivan Fellegi and Alan Sunter.[2] Here, for each pair being scrutinized, we look to see if they agree on identifiers in common on the files being linked. In the simplest development of this method, each identifier comparison will be classified as an agreement (when the values are exactly or nearly the same) or a non-agreement (when they are substantially different).

[1] The exact demarcation between agreement and non-agreement does not affect the theoretical treatment.
[2] See Fellegi and Sunter (1969).

When the values for the set of identifiers selected for the analysis (e.g., first name or day-of-birth, etc.) agree, the pair weight (total score) is incremented by the agreement weight specific to that identifier. However, if these identifiers do not agree, the pair weight is decremented by the non-agreement weight[3] specific to that identifier. The sum of all these agreement and non-agreement weights is the pair weight. At this point, we defer the discussion of how these weights should be set, and stipulate that the pair weight is compared to two cutoffs for disposition (inference if the pair is a match or a non-match):

- If the pair weight is greater than the upper cutoff, the pair will be linked, meaning we have inferred that the paired records do represent the same entity (i.e., they are a match).
- If the pair weight is less than the lower cutoff, the pair will be left unlinked, meaning we have inferred that the paired records do not represent the same entity (i.e., they are not a match).
- If the pair weight is in between the lower and upper cutoffs, then we will further scrutinize it (usually with human judgment) to determine if it should be linked.

Fellegi and Sunter demonstrated that identifier weights that are set in accordance with certain algebraic formulas, as shown below, have the property of minimizing the expected Type I error for a given expected level of Type II error—i.e., they are uniformly most powerful.

$$Agreement\ Weight_i = \frac{\ln(M_i/U_i)}{\ln(2)}$$

and

$$Non\text{-}Agreement\ Weight_i = \frac{\ln((1 - M_{i)}/(1 - U_i))}{\ln(2)}$$

The computation then depends on knowing the true values for the (M) probability that values agree for a specific identifier, when they come from matches (records representing the same entity) and the (U) probability that values agree for a specific identifier when they come from non-matches (records representing different entities). For example, we can assume that there is about a 1 in 12 chance (i.e., the U-probability for month of birth is $\frac{1}{12} \approx .0833$) that two records representing different persons will have the same value for month-of-birth (i.e., by chance) and a much higher (M) probability (say at least above 90%) that they have the same value when they do represent the same person (the rare cases of non-agreement occur in the case of a transcription error).

To actually compute weights from the above formulas, we need to know the M- and U-probabilities for each identifier used in the linkage analysis. In actuality, however, these are usually unknown but need to be estimated, such as by using training data. Here, the training data would be records from the same files where we knew which pairs were actually matches and which were not. Then, computing the probabilities would only require simple frequency tabulations.

More often, training data are not available. In these cases, these values can be estimated using machine learning or a statistical fitting procedure (Winkler 2011). For this, Monte

[3] Both agreement and non-agreement weights are summed, but whereas agreement weights are positive, non-agreement weights are negative and so their addition decrements the pair weight.

Carlo Markov Chain (MCMC) is one of several methods that can be used. Even if we can accurately estimate the M- and U- probabilities, the cutoffs for assigning pairs as matches or non-matches depend on how the linked data will be used for further analysis. In some cases, the analysis will require the use of only links that have a very high probability of being matches (i.e., we are focusing on minimizing Type I error). In other cases, we can tolerate a higher probability of matching error but want to try to identify as many links as possible (i.e., we are focusing on minimizing Type II error).

Even if the weights can be estimated with good precision, it is unclear of how to set the cutoffs from a theoretical perspective. From a practical view, the objective of the exercise is as follows:

- The upper cutoff is set to be the pair score above which pairs can *almost certainly* be inferred to be matches.
- The lower cutoff is set to be the pair score below which pairs can *almost certainly* be inferred to be non-links.

Of course, operationalizing "almost certainly" is quite indefinite, and might well be considered to depend on how the linked records will be used subsequently for analysis. If Type I errors are to be minimized, this would suggest using higher cutoff scores, and if Type II errors are to be minimized, this would suggest using lower cutoff scores. Again, the actual selection of these values is left to human review and judgment. Additionally, unless the clerical review for pairs between the lower and upper cutoffs brings new data to bear on a specific pair, it is unclear that human judgment can better ascertain whether a given pair is a match or a non-match that using the pair weight.

2. Fitting Approach

For many reasons, developing formulaic guidance for setting cutoffs, rather than needing to rely on human judgment, would be desirable. Here, being able assign each pair an estimate of the match rate (by which we mean, the proportion of pairs having the same identifier agreement pattern, such as, first names agree, middle initials do not agree, and last names agree, which are matches) would enable users of the linked data to precisely set the linkage cutoffs. For example, if the cutoff is set at a point where the probability of a Type I error is less than 1%, we would assign all pairs with an estimated link probability of 99% or more to the link set.

Generally, the Fellegi-Sunter paradigm is well suited to estimating the match probabilities. Assuming that among matches and among non-matches, agreement for each identifier is a Bernoulli random variable, and agreement for each identifier is independent of the remaining identifiers (i.e., by naïve Bayes), then the probability that matches will have a given agreement pattern is shown below:

$$P(A_M) = \prod_{i=1}^{n_a} {M_i}^{a_i} \cdot (1 - M_i)^{(1-a_i)} \qquad \text{(Eq. 1)}$$

where:

A_i is the agreement status, 1 – Agree, 0 – Not Agree, for identifier i
M_i is probability of match for identifier i

and the expected number of matched pairs having this pattern is

$$E(N_M) = \ Matches_{Total} \cdot P(A_M) \qquad \text{(Eq. 2)}$$

Similarly, the probability that a non-matched pairs will have a given agreement pattern is given by:

$$P(A_U) = \prod_{i=1}^{n_a} {U_i}^{a_i} \cdot (1 - U_i)^{(1-a_i)} \qquad \text{(Eq. 3)}$$

and the expected number of non-matched pairs having this pattern is

$$E(N_U) = \ Non\text{-}Matches_{Total} \cdot P(A_U) \quad \text{(Eq. 4)}$$

One form the estimation can take is to compare the estimated number of pairs with a given identifier agreement pattern (i.e., a specific agreement vector) to the actual number of pairs (see Table 1):

Table 1: Example Agreement Vector and Associated Statistics

Agreement Vector					*Associated Statistics*				
***A**₁*	***A**₂*	***A**₃*	***A**₄*	***A**₅*					
First Name	*Last Name*	*Year of Birth*	*Month of Birth*	*Day of Birth*	$E(N_M)$	$E(N_U)$	*Expected Total Pairs*	*Actual Pairs*	*Chi-Square*
1	0	1	1	1	500	50	550	520	$\frac{(550\text{-}520)^2}{550}$

Then, to evaluate the overall agreement level, we can compute a chi-square goodness-of-fit statistic. We would expect that the set of parameters that minimizes this chi-square value (summed over all agreement vectors) should closely agree with the actual, and presumably unknown, parameter values. Then, our fitting problem reduces to an optimization problem. Note that in addition to seeking the values for the agreement probabilities that minimize the chi-square statistic, we need also seek (as another parameter to be estimated) the value of the number of matches among the full set of pairs.

So, if the M- probabilities, U- probabilities, and the number matched pairs among all of the pairs are known, we can estimate the total number of matched pairs and unmatched pairs that will have a given agreement pattern. Then the match rate (probability of a randomly selected pair being a match) for the agreement pattern, *A*, can be estimated as follows:

$$\widehat{P(M_A)} = \frac{E\left(N_{M_A}\right)}{E\left(N_{M_A}\right) + \ E\left(N_{U_A}\right)}$$

where

$E\left(N_{M_A}\right)$ is the expected number of matched pairs with agreement pattern *A*
$E\left(N_{U_A}\right)$ is the expected number of unmatched pairs with agreement pattern *A*

In the absence of any information about the true values of these parameters, it is natural to seek a set of parameter estimates that comes closest to matching the counts associated with each agreement vector.

This leads us to use the chi-square goodness-of-fit statistic as the objective function that we seek to minimize:

$$\mathrm{X}^2 = \sum_A \frac{(N_A - \widehat{N_A})^2}{N_A} \qquad \text{(Eq. 5)}$$

where,

N_A is the number of pairs with agreement pattern A, and
$\widehat{N_A} = E(N_{M_A}) + E(N_{U_A})$ is the estimated number of pairs with agreement pattern A

To find the parameters that do yield this best fit, i.e., $\min(\mathrm{X}^2)$, we use a Newton-Raphson search methodology.[4] We start with a guess of parameter values. For our guesses, we (arbitrarily) use the value 0.9 for M-probabilities and the value 0.1 for U-probabilities. Also, our guess for the number of matched pairs is equal to half of the total number of pairs being analyzed. Next we seek to determine the direction (in the multi-dimensional parameter space) of maximum improvement (reduction) of the objective function. We determine this direction by estimating the partial derivatives of the objective function with respect to each of the parameters. Each partial derivative is estimated by increasing the parameter associated with it by a small amount (ΔP) and observing how this changes the objective function. We estimate the partial derivative for each parameter *i* as

$$\widehat{\partial'_\iota} = \Delta\mathrm{O} / \Delta\mathrm{P}.$$

The vector of the full set of partial derivatives $(\widehat{\partial'_\iota}, \widehat{\partial'_2}, \ldots, \widehat{\partial'_n})$ is the gradient and estimates the direction of maximum increase of the objective function. Since we seek to minimize the objective function, we actually want to more in the directly opposite direction.

Now that we have an estimate of the direction of maximum improvement, we need to estimate the optimal interval to move in this direction. This is done by estimating the first and second derivatives along the vector of improvement. For the first derivative, this is done by moving a small amount on this vector, but here we are simultaneously adjusting all of the parameter values by the magnitude of its component (i.e., the partial derivative for it) and seeing how this changes the objective function. To estimate the second derivative, we take a second step in the same direction and estimate

$$\widehat{f''} = (\Delta\mathrm{O}_1 - \Delta\mathrm{O}_2) / \Delta\mathrm{P},$$

where,

$\Delta\mathrm{O}_1$ is the change in objective function from P_0 to P_1, $\mathrm{O}(\mathrm{P}_1) - \mathrm{O}(\mathrm{P}_0)$
$\Delta\mathrm{O}_2$ is the change in objective function from P_1 to P_2, $\mathrm{O}(\mathrm{P}_2) - \mathrm{O}(\mathrm{P}_1)$
O(P) is the value of the objective function with parameter set *P*

With this, we estimate the optimal step size as

Optimal Step Size = $-\widehat{f'}$ (objective function) / $\widehat{f''}$ (objective function).

[4] See Adler 2003.

Based on these estimates of the direction and magnitude of the step for best improvement, we are able to generate a substantially improved set of parameters with respect to minimizing the objective function. At this point, we have completed the first iteration of the fitting, resulting in an updated set of parameter estimates.

We continue to iterate this approach until we are no longer making improvement to the fit.

2.1 Testing of Fitting Methodology

To test the basic functioning of the approach, we ran it on a simulated set of pairs. For this simulation, we used an agreement identifier vector with five components. We generated 50,000 matched pairs and 1.2 million unmatched pairs. For the matched pairs, for each or the five identifiers, we assigned an M probability specific to it. Thus for the first identifier, the M probability was set (based on generation from a pseudo random number generator) to be 0.9393. This means that for the generated matched pairs, approximately 93.9% were shown with agreement for the first identifier. For the unmatched pairs, for each of the five identifiers, we assigned a U probability specific to it.

Having run the optimization algorithm on the full set of pairs obtained these estimates of the M- and U- probabilities:

Table 2: M- and U- Probability Estimates Compared to Actual Values

	Matched (M-Probabilities)			*Unmatched (U-Probabilities)*		
Identifier	Simulation Parameter (Target)	Actual Prop.	Identifier	Simulation Parameter (Target)	Actual Prop.	Simulation Parameter (Target)
$\mathbf{A}_1$	93.93%	94.06%	94.09%	6.11%	6.13%	6.13%
$\mathbf{A}_2$	94.42%	94.59%	94.57%	9.81%	9.80%	9.80%
$\mathbf{A}_3$	99.12%	99.08%	99.11%	9.12%	9.12%	9.12%
$\mathbf{A}_4$	87.02%	86.95%	86.95%	6.51%	6.50%	6.51%
$\mathbf{A}_5$	98.59%	98.59%	98.61%	8.73%	8.72%	8.73%

In terms of the predicted and actual match rates we obtained these results:

Table 3: Selected Results*, by Agreement Vector Values

	Agreement Vector				*Actual Count of Recs.*	*Estim. Count of Recs.*	*Actual Valid Match Rate*	*Estim. Valid Match Rate*	*Goodness of Fit Statistic*
A_1	A_2	A_3	A_4	A_5					
No	No	No	Yes	No	55,012	54,857	0.0%	0.0%	0.44
No	Yes	No	No	No	85,542	85,624	0.0%	0.0%	0.08
No	Yes	No	No	Yes	8,327	8,189	0.1%	0.0%	2.34
Yes	No	No	No	Yes	4,971	4,923	0.1%	0.1%	0.47
Yes	No	Yes	No	No	5,156	5,169	0.1%	0.1%	0.03
Yes	Yes	Yes	No	Yes	5,694	5,700	99.2%	99.1%	0.01
Yes	Yes	Yes	Yes	No	574	571	93.6%	93.2%	0.01
					165,276	*165,032*			*3.38*

*To reduce table size, selection of agreement vectors shown on this table was random.

3. Extension to Include Missing Agreements

In record linkage analyses, it is usually the case that for some of the identification variables, the value for a given record is missing. For example, if the linkage was for individual persons, in some cases one or more of the date-of-birth fields, middle initial fields, or other identifier fields may be missing. To account for these occurrences, the fitting model requires some adaptation. In terms of computing the probability that the matched or unmatched pair may have a specific agreement pattern *A*, it is simple enough to exclude missing identifiers from the computation. That is we exclude each *i* with missing agreement status from the products (see Eq. 1 and Eq. 3, above).

However, instead of then multiplying this probability by the *total number of matches* (or total number of non-matches, as in Eq. 2 and Eq. 4, it is instead multiplied by the estimated number of matches (or the estimated number of non-matches) *with agreement pattern P*:

$$E(N_M) = \ M\widehat{atche}s_P \cdot P(A_M)$$
$$E(N_U) = \ \ Non\text{-}\widehat{Matche}s_P \cdot P(A_U)$$

The estimated number of matches and non matches with agreement pattern P are computed as

$$M\widehat{atche}s_P = \ Matches_{Total} \ \cdot N_P \,/ N_{Total}$$
$$Non\text{-}\widehat{Matche}s_P = \ Non\text{-}Matches_{Total} \ \cdot N_P \,/ N_{Total}$$

where,

N_P is the number of pairs with missing pattern *P*

By missing pattern, we mean the vector of agreement identifiers' missing statuses. For example, assume that the linkage process uses First Name, Middle Initial, Last Name, Year of Birth, Month of Birth, and Day of Birth and the agreement status is missing for Middle Initial, Year of Birth, and Month of Birth (see Table 4).

Table 4: Missing Pattern

Agreement Value Missing Status

A₁	***A₂***	***A₃***	***A₄***	***A₅***	***A₆***
First Name	*Middle Initial*	*Last Name*	*Year of Birth*	*Month of Birth*	*Day of Birth*
0=No	**1=Yes**	**0=No**	**1=Yes**	**1=Yes**	**0=No**
(*Present*)	(*Missing*)	(*Present*)	(*Missing*)	(Missing)	(*Present*)

Then the agreement vectors falling under this missing pattern would be as follows (See Table 5):

Table 5: Agreement Vectors Falling Under (See Table 4, Above) Missing Pattern

Identifier Agreement Status*:

A₁	***A₂***	***A₃***	***A₄***	***A₅***	***A₆***
First Name	*Middle Initial*	*Last Name*	*Year of Birth*	*Month of Birth*	*Day of Birth*
No	-	No	-	-	No
No	-	No	-	-	Yes
No	-	Yes	-	-	No
No	-	Yes	-	-	Yes
Yes	-	No	-	-	No
Yes	-	No	-	-	Yes
Yes	-	Yes	-	-	No
Yes	-	Yes	-	-	Yes

*("-" indicates agreement status is missing)

Then, if among the 1.5 Million pairs under analysis, 20,000 had this missing pattern, then the estimated number of matches, with this missing pattern would be $\widehat{Matches}_P = \widehat{Matches}_{Total} \cdot 20{,}000/1{,}500{,}000$. Each identifier agreement pattern A falling under this missing pattern P would apply this ratio (20,000 / 1,500,000) to the initial estimated number of pairs (see Eqs. 2 and 4) of having the identifier pattern.. This adjusted estimate would then be compared to the actual counts to compute the goodness of fit statistic (see Eq. 5 above).

4. Relaxation of Assumption of Agreement Non-Independence

Real linkage analysis diverges from the naïve Bayes probability that assumes independence of identifier agreements among matches and non-matches. For example, if we are linking business locations using address information, then two different businesses within the same ZIP code are much more likely to have a common street name than if they were located in different ZIP codes. For this extension, we focus our attention on non-matches (rather than on matches). Also, we only model the simplest level of interaction among the identifiers, that being pairwise. This is not to say that interactions will not also occur among matched pairs, or there will not be more than two-way interactions. On the other hand, the goal of this research was to investigate whether the optimization routine could be modified in a way that allows some of interactions to be estimated as part of the process of generating estimates of the first-order M- and U- probabilities. So here we were looking for a straightforward yet robust way to include interactions as part of the estimation process. The way we have proceeded is to consider there to be an odds-ratio adjustment associated with each pair of identifiers. So, if there are four identifiers, for example (linking addresses):

1. Address Number
2. Street Name
3. City
4. State

then the ${}_4C_3$ = 6 odds-ratio adjusters to be estimated would be

Table 6: Odds Adjusters with 4 Agreements (4-dimensional agreement vector)

Odds Adjuster	*Identifier A*		*Identifier B*
1	**A_1** (Address Number)	↔	**A_2** (Street Name)
2	**A_1** (Address Number)	↔	**A_3** (City)
3	**A_1** (Address Number)	↔	**A_4** (State)
4	**A_2** (Street Name)	↔	**A_3** (City)
5	**A_2** (Street Name)	↔	**A_4** (State)
6	**A_3** (City)	↔	**A_4** (State)

Then we determine the operative odds adjusters as those for which both identifiers are in agreement. The remaining odds adjusters (i.e., those for which either or both of the components are in non-agreement) are excluded in the odds adjustment computation. For example, if the identifier agreement pattern is:

Table 7: Agreement Vector

A₁	***A₂***	***A₃***	***A₄***
Address Number	*Street Name*	*City*	*State*
1	0	1	1
(Agreement)	(Non-Agreement)	(Agreement)	(Agreement)

Then the operative odds adjusters for this pattern would be Odds Adjuster 2 ($A_1 \leftrightarrow A_3$), Odds Adjuster 3 ($A_1 \leftrightarrow A_4$), and Odds Adjuster 6 ($A_3 \leftrightarrow A_4$). The full odds-adjustment is simply the product of all the operative odds adjusters. To compute the adjusted probability that a matched pair will have a given agreement pattern, we first convert the probability computing under the naïve-Bayes assumption into odds

$$\text{Odds}_{NB} = P_{NB} / (1 - P_{NB})$$

The odds is then multiplied by the product of the operative odds adjusters

$$\text{Odds}_{Adj} = \text{Odds}_{NB} \text{ x } \{\text{Odds Adjuster 2}\} \text{ x } \{\text{Odds Adjuster 3}\} \text{ x } \{\text{Odds Adjuster 6}\}$$

and then this adjusted odds value is then converted back into an adjusted probability

$$P_{Adj} = \text{Odds}_{Adj} / (1 + \text{Odds}_{Adj}).$$

To test the efficacy of this approach in accounting for agreement interactions, we need to generate a set of simulated pairs which have these interactions. This is done by generating each pair's agreements sequentially, but making the probability used for simulating some of the later agreements to be dependent on the simulated agreement status of previous agreements. So for matches, we might say that the first agreement has a Bernoulli distribution of

$$P(A_1{=}1) = U_1$$
$$P(A_1{=}0) = 1 - U_1$$

For the second agreement, the distribution is dependent on the value of the first agreement status:

$$P(A_2{=}1 \mid A_1 = 0) = U_{2,1}$$
$$P(A_2{=}1 \mid A_1 = 1) = U_{2,2}$$
$$\text{where } U_{2,2} \neq U_{2,\,1}$$

Using randomly generated pairs with interactions for the agreement components for unmatched pairs (see Appendix for fuller specification of the simulated pairs), we obtained these results presented as regression fits. Here we are fitting the actual match proportion to the estimated match proportion (in Figure 1 we show the fitting with no interaction parameters and in Figure 2 we show the fitting using all two-way interaction parameters, i.e., odds-adjusters). The individual records being analyzed are all the possible values of the agreement vector:

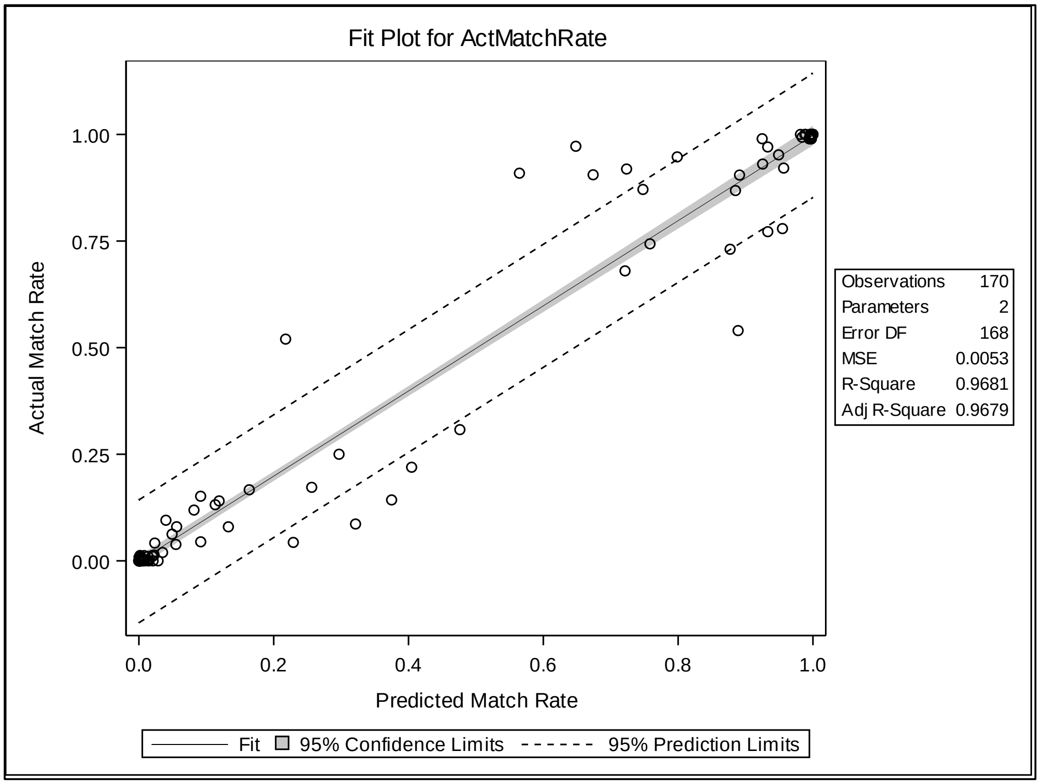

Figure 1: Actual vs. Modeled Proportion Matches by Agreement Vector – No Interaction Parameters

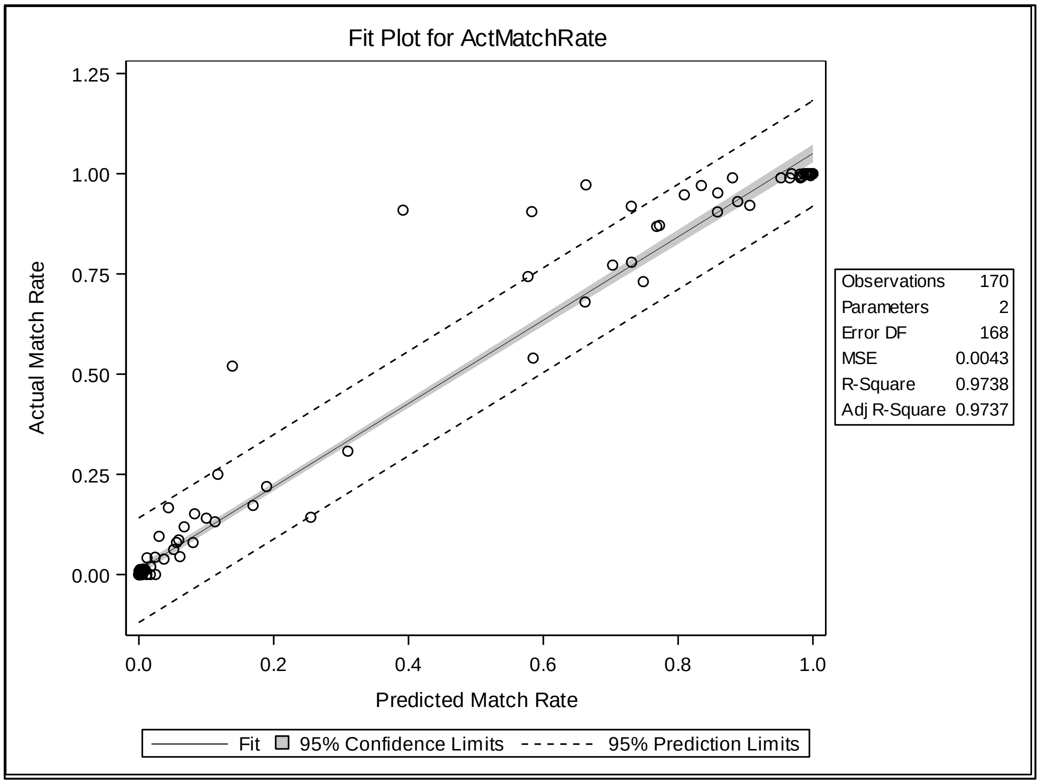

Figure 2. Actual vs. Modeled Proportion Matches by Agreement Vector – With Interaction Parameters.

We see that indeed the use of the odds adjusters to account for interactions does, in fact, substantially improve the fit from $R^2 = 0.9681$ to $R^2 = .9738$.

Conclusion

The basic approach laid out in this paper does produce quite good estimates of the M- and U- probabilities, the total number of matched pairs, and the proportion of matches among pairs with the same agreement pattern using simulated pairs with independent agreements. Additionally, modifications made to handle missing agreements likewise function well and enable this method to be applied to actual record linkage problems. The work presented in this paper yields a proof of concept that the extension to enable adjustment for interactions can be included in the model fitting with improvement over the basic model in the face of agreement interactions. It is likely that the parameterization of the interactions can be improved with work. At a minimum, it could include agreement interactions for matches as well as non-matches (which is the extent of the current modeling work). Also, the use of more than two-way interactions could be explored. In this case, care must be taken so that the optimization problem is not over-parameterized. Perhaps this concern can be addressed by making a basic fit using the naïve Bayes assumption and then using the fitted pairs to evaluate which interactions seems most significant within them. Then, the optimization can be set up to evaluate parameters that enable fit specifically to these interactions and not insignificant ones.

References

Adler, Andrew. July 16, 2003. *The Newton-Raphson Method*. Retrieved (September 30, 2017), from https://www.math.ubc.ca/~anstee/math104/104newtonmethod.pdf.

Fellegi, Ivan P., and Alan B. Sunter (1969). "A theory for record linkage." *Journal of the American Statistical Association* 64.328: 1183-1210.

Winkler, William (2011). "Machine Learning and Record Linkage." *58th World Statistics Congress ISI*.

Appendix: Development of Simulated Pairs for Testing Inclusion of Interaction Terms

These were the parameters used for the simulation created to test impact of inclusion of odds-adjusters in fitting model:

Matched pairs had the M-probabilities

$M_1=0.862$, $M_2=0.995$, $M_3=0.944$, $M_4=0.980$, $M_5=0.954$

Unmatched pairs were randomly assigned (with equal probability) to two pools with different U- probabilities:

Pool 1: $U_{1,1}=0.0043$, $U_{1,2}=0.0491$, $U_{1,3}=0.0014$, $U_{1,4}=0.0218$, $U_{1,5}=0.0150$

Pool 2: $U_{2,1}=0.0794$, $U_{2,2}=0.0017$, $U_{2,3}=0.0593$, $U_{2,4}=0.0764$, $U_{2,5}=0.0625$

There were no interactions for matched pairs, but the interactions for unmatched pairs were…

- $P(A_2=1 \mid A_1 = 0) = U_{2,i}$
- $P(A_2=1 \mid A_1 = 1) = 1.75 \text{ x } U_{2,i}$

- $P(A_3=1 \mid A_1 = 0 \text{ or } A_2 = 0) = U_{3,i}$
- $P(A_3=1 \mid A_1 = 1 \text{ and } A_2 = 1) = 2.35 \text{ x } U_{3,i}$

- $P(A_2=1 \mid A_3 = 0) = U_{4,i}$
- $P(A_2=1 \mid A_3 = 1) = 0.5 \text{ x } U_{4,i}$

In addition to these interactions, we further complicated the data set by creating duplicate records that simulate the relationships shown between records for different members of the same household. The idea here is that generally household members will have some characteristics in common, such as address and last name and others not in common, such as first name and date-of-birth. For a randomly selected set of 25% of the originally generated match records, we generated a single duplicate record (See Table A-1 below) represented to be a non-match that had all of the same identifier agreements statuses, except for two of the randomly selected agreeing identifiers they were re-set to non-agreement.

Table A-1: Example Creation of Duplicate Records

Record	*Match*	*Agreement Status Vector*				
Type	*Status*	*A_1*	*A_2*	*A_3*	*A_4*	*A_5*
Original	Match	1	1	1	0	1
↓	↓	↓	↓*	↓*	↓	↓
Duplicate	Non-Match	1	0	0	0	1

*Note: A2 and A3 changed from Agreement to Non-Agreement