Machine Learning and Record Linkage

Winkler, William 1/
U.S. Census Bureau, Center for Statistical Methods Research
4600 Silver Hill Road
Suitland, Maryland 20746, USA
E-mail: william.e.winkler@census.gov, william_e_winkler@msn.com

1. Introduction

Modern record linkage consist of methods for finding duplicate entities within a file or across files using non-uniquely identifying characteristics such as name, address, and date-of-birth. The idea is to have sophisticated computer algorithms that perform many of the tasks that well-trained, experienced individuals might perform. Because names, addresses, and dates-of-birth can have many conventions in terms of formatting, word order, and spelling variations, there are a variety of methods for cleaning up the files that can be rule-based (if-then-else) or, more generally and flexibly, based on hidden Markov models from machine learning and statistics. The basic hidden Markov methods (Baum-Welsh algorithm; see Bilmes (1998 Tutorial) generalize the well known EM methods.

The simplest model for record linkage is based on odds-ratios (Newcombe 1959, 1962) that was formally developed into the mathematical model of Fellegi and Sunter (1969, hereafter FS) based on hypothesis-testing ideas. Although the FS model is very general, in most situations it is applied under a conditional independence assumption which is referred to as naïve Bayes in the machine learning literature. Although naïve Bayes is often superceded by Support Vector Machines (Vapnik 20001) and Boosting (Freund and Shapire 1996; Friedman, Hastie, and Thibshirani 2000) for most applications of machine learning, naïve Bayes is still the method most widely used in record linkage.

In this paper, we describe the basic model of FS, how 'optimal' parameters are estimated when there is no training data, and how false match (false positive) rates can be estimated in a narrow but widely applicable number of situations without training data. In the FS model, based on a score for each pair of records from two files A and B and fixed upper bounds of the acceptable error rates, an upper score is chosen above which a pair is a match, a lower score below which a pair is a nonmatch, and pairs with in-between scores are held for clerical review. The 'optimal' parameters typically significantly reduce the number of clerical-review pairs by at least a factor of three. The 'optimal' parameters vary significantly between adjacent geographic regions (city versus suburb) in comparison to parameters from knowledgeable guesses. Estimating error rates is referred to as the *regression problem* (Hastie et al. 2001; Vapnik 2000) and is known to be exceedingly difficult when training data are available.

2. The Model of Fellegi and Sunter

Fellegi and Sunter (1969) provided a formal mathematical model for ideas that had been introduced by Newcombe (1959, 1962). They introduced many ways of estimating key parameters without training data. To begin, notation is needed. Two files A and B are matched. The idea is to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true matches, and U, the set of true nonmatches.

Fellegi and Sunter, making rigorous concepts introduced by Newcombe (1959), considered ratios of probabilities of the form:

$$R = P(\gamma \in \Gamma \mid M) / P(\gamma \in \Gamma \mid U) \quad (1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith", "Zabrinsky", "AAA", and "Capitol" occur. The ratio R or any monotonely increasing function of it such as the natural log is referred to as a *matching weight* (or score).

The decision rule is given by:

If $R > T_\mu$, then designate pair as a match.

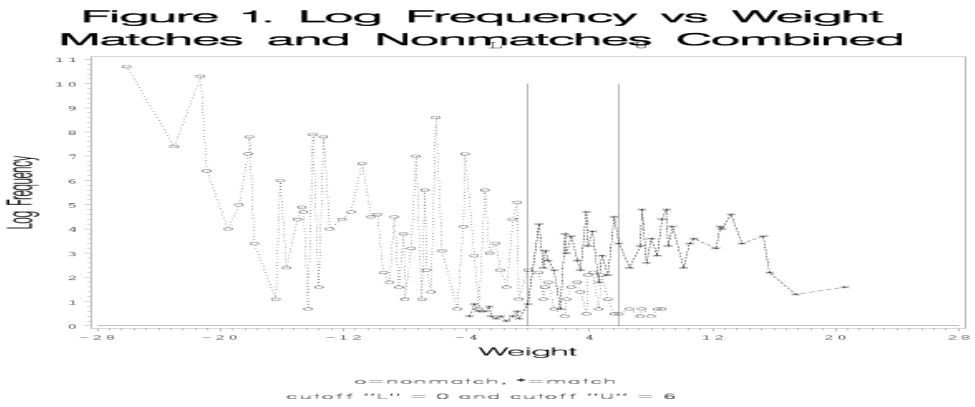
If $T_\lambda \leq R \leq T_\mu$, then designate pair as a possible match and hold for clerical review. (2)

If $R < T_\lambda$, then designate pair as a nonmatch.

The cutoff thresholds T_μ and T_λ are determined by a priori error bounds on false matches and false nonmatches. Rule (2) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than nonmatches and ratio (1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (1) would be small. Rule (2) partitions the set $\gamma \in \Gamma$ into three disjoint subregions. The region $T_\lambda \leq R \leq T_\mu$ is referred to as the *no-decision region* or *clerical review* region. In some situations, resources are available to review pairs clerically.

Fellegi and Sunter (1969, Theorem 1) proved the optimality of the classification rule given by (2). Their proof is very general in the sense in it holds for any representations $\gamma \in \Gamma$ over the set of pairs in the product space $\mathbf{A} \times \mathbf{B}$ from two files. As they observed, the quality of the results from classification rule (2) were dependent on the accuracy of the estimates of $P(\gamma \in \Gamma \mid M)$ and $P(\gamma \in \Gamma \mid U)$.

Figure 1 provides an illustration of the curves of log frequency versus log weight for matches and nonmatches, respectively. The two vertical lines represent the lower and upper cutoffs thresholds T_λ and T_μ, respectively. The x-axis is the log of the likelihood ratio R given by (1). The y-axis is the log of the frequency counts of the pairs associated with the given likelihood ratio. The plot uses pairs of records from a contiguous geographic region that was matched in the 1990 Decennial Census. The clerical review region between the two cutoffs primarily consists of pairs within the same household that are missing both first name and age.



3. Learning parameters via the methods of Fellegi and Sunter

Fellegi and Sunter (1969) were the first to give very general methods for computing the probabilities in ratio (1). As the methods are useful, we describe what they introduced and then show how the ideas led into more general methods that can be used for *unsupervised learning* (i.e., without training data) in a large number of situations.

Fellegi and Sunter observed several things. First,

$$P(A) = P(A \mid M)P(M) + P(A \mid U)P(U) \tag{3}$$

for any set A of pairs in $\mathbf{A} \times \mathbf{B}$. The probability on the left can be computed directly from the set of pairs. If sets A represent simple agreement/disagreement, under the conditional independence assumption (**CI**), we obtain

$$P(A_1^x \cap A_2^x \cap A_3^x \mid D) = P(A_1^x \mid D)P(A_2^x \mid D)P(A_3^x \mid D), \tag{4}$$

then (3) and (4) provide seven equations and seven unknowns (as x represents agree or disagree) that yield quadratic equations that they solved. Here D is either M or U. Equation (or set of equations) (4) can be expanded to K fields. Although there are eight patterns associated with the equations of the form (4), we eliminate one because the probabilities must add to one. In general, with more fields but still simple agreement/disagreement between fields, the equations can be solved via the EM algorithm in the next section. Probabilities of the form $P(A_i \mid D)$ are referred to as *m-probabilities* if D=M and *u-probabilities* if D=U.

5. Learning Parameters via the EM Algorithm

In this section, we do not go into much detail about the basic EM algorithm because the algorithm is well understood. We provide a moderate amount of detail for the record linkage application so that we can describe a number of the limitations of the EM and some of the extensions.

For each $\gamma \in \Gamma$, we consider

$$P(\gamma) = P(\gamma \mid C_1)\, P(C_1) + P(\gamma \mid C_2)\, P(C_2) \quad (5a)$$

$$P(\gamma) = P(\gamma \mid C_1)\, P(C_1) + P(\gamma \mid C_2)\, P(C_2) + P(\gamma \mid C_3)\, P(C_3) \quad (5b)$$

and note that the proportion of pairs having representation $\gamma \in \Gamma$ (i.e., left hand side of equation (5) can be computed directly from available data. In each of the variants, C_1 and C_2, or C_1, C_2, and C_3 partition $\mathbf{A \times B}$.

If the number of fields associated with γ is K > 3, then we can solve the combination of equations given by (5) and (3) using the EM algorithm. Although there are alternate methods of solving the equation such as methods of moments and least squares, the EM is greatly preferred because of its numeric stability. Under conditional independence, programming is simplified and computation is greatly reduced (from 2^k to $2k$).

Caution must be observed when applying the EM algorithm to real data. The EM algorithm that has been applied to record linkage is a *latent class algorithm* that is intended to divide $\mathbf{A \times B}$ into the desired sets of pairs M and U. The probability of a class indicator that determines whether a pair is in M or U is the missing data must be estimated along with the m- and u-probabilities. It may be necessary to apply the EM algorithm to a particular subset S of pairs in $\mathbf{A \times B}$ in which most of the matches M are concentrated, for which the fields used for matching clearly can separate M from U, and for which suitable initial probabilities can be chosen. Because the EM is a local maximization algorithm, the starting probabilities may need to be chosen with care based on experience with similar types of files. Because the EM latent-class algorithm is a general clustering algorithm, there is no assurance that the algorithm will divide $\mathbf{A \times B}$ into two classes C_1 and C_2 that almost precisely correspond to M and U.

A fast efficient means of estimating parameters in differing regions and with a variety of files were needed because Winkler (1989) had demonstrated that the optimal parameters (which are very dependent on typographical error rates) differed significantly between adjacent geographic regions. For the 1990 Census, the EM was used to get optimal parameters automatically for the 457 regions of Decennial processing. The false match rate with the computerized procedures was 0.2 percent (based on extensive field validation) and computer processing took 3-6 weeks with clerical review by 200 individuals because of extensive missing or contradictory information for a very small proportion of individuals. Prior to the development of the computerized matching procedures (which also included string comparators for dealing with typographical error – Winkler 1990, 2006a), individuals had extrapolated that the clerical review associated with primarily manual procedures would take 3000 individuals six months with an false match rate as high as 5 percent.

The EM methods showed their effectiveness when clerical review increased dramatically in three regions in one week. Upon review, we discovered that two keypunchers had managed to bypass edits on the year-of-birth and all records associated with these keypunchers could not be compared on age. The EM automatically discovered the discrepancies and the resultant decrease in optimality of parameters (name – sometimes missing – and age are the only means of distinguishing individuals in the same household) increased the size of the clerical review region.

Generalizations of the basic EM methods (Winkler 1998) are due to Ravikumar and Cohen (2003) and

Bhattacharya and Getoor (2006). These latter methods are much easier to apply than the more general methods of Winkler (1993).

6 Error rate estimation

With any matching project, we are concerned with false match rates among the set of pairs among designated matches above the cutoff score T_μ in (2) and the false nonmatch rates among designated nonmatches below the cutoff score T_λ in (2). Very few matching projects estimate these rates although valid estimates are crucial to understanding the usefulness of any files obtained via the record linkage procedures. Sometimes reasonable upper bounds for the estimated error rates can be obtained via experienced practitioners and the error rates are validated during follow-up studies (Winkler 1995). If a moderately large amount of training data is available, then it may be possible to get valid estimates of the error rates.

If a small amount of training data is available, then it may be possible to get improved record linkage and good estimates of error rates. Larsen and Rubin (2001) combined small amounts of (labeled) training data with large amounts of unlabelled data to estimate error rates using an MCMC procedure. In machine learning (Winkler 2000), the procedures are referred to as *semi-supervised learning*. In ordinary machine learning, the procedures to get parameters are "supervised" by the training data that is labeled with the true classes into which later records (or pairs) will be classified. Winkler (2002) also used semi-supervised learning with a variant of the general EM algorithm. Both the Larsen and Rubin (2001) and Winkler (2002) methods were effective because they accounted for interactions between the fields and were able to use labeled training data that was concentrated between the lower cutoff T_λ and the upper cutoff T_μ.

Belin and Rubin (1995) were the first to provide an unsupervised method for obtaining estimates of false match rates. The method proceeded by estimating Box-Cox transforms that would cause a mixture of two transformed normal distributions to closely approximate two well separated curves such as given in Figure 1. They cautioned that their methods might not be robust to matching situations. Winkler (2006a) observed that their algorithms would typically not work with business lists, agriculture lists, and low quality person lists where the curves of nonmatches were not well separated from the curves of matches. Scheuren and Winkler (1993), who had the Belin-Rubin EM-based fitting software, observed the Belin-Rubin methods did work reasonably well with a number of well-separated person lists.

Because the EM-based methods of this section serve as a template of other EM-based methods, we provide details of the unsupervised learning methods of Winkler (2006b) that are used for estimating false match rates. The basic model is that of semi-supervised learning in which we combine a small proportion of labeled (true or pseudo-true matching status) pairs of records with a very large amount of unlabeled data. The conditional independence model corresponds to the naïve Bayesian network formulization of Nigam et al. (2000). The more general formulization of Winkler (2000, 2002) allows interactions between agreements (but is not used in this paper).

Our development is similar theoretically to that of Nigam et al. (2000). The notation differs very slightly because it deals more with the representational framework of record linkage. The following equations repeat some of the ideas given in equations 1-5. Let γ_i be the agreement pattern associated with pair p_i. Classes C_j are an arbitrary partition of the set of pairs D in $\mathbf{A} \times \mathbf{B}$. Later, we will assume that some of the C_j will be subsets of M and the remaining C_j are subsets of U. For coherence and clarity

equations (6) and (7) repeat earlier equations but use slightly different notation. Unlike general text classification in which every document may have a unique agreement pattern, in record linkage, some agreement patterns γ_i may have many pairs $p_{i(l)}$ associated with them. Specifically,

$$P(\gamma_i \mid \Theta) = \sum_i^{|C|} P(\gamma_i \mid C_j; \Theta)\, P(C_j ; \Theta) \qquad (6)$$

where (i is a specific pair, Cj is a specific class, and the sum is over the set of classes. Under the Naïve Bayes or conditional independence (**CI**), we have

$$P(\gamma_i \mid C_j ; \Theta) = \prod_k P(\gamma_{i,k} \mid C_j ; \Theta) \qquad (7)$$

where the product is over the k^{th} individual field agreement γ_{ik} in pair agreement pattern γ_i. In some situations, we use a Dirichlet prior

$$P(\Theta) = \Pi_j (\Theta_{Cj})^{\alpha-1} \quad \Pi_k \quad (\Theta\, \gamma_{i,k \mid Cj})^{\alpha-1} \qquad (8)$$

where the first product is over the classes Cj and the second product is over the fields. We use Du to denote unlabeled pairs and Dl to denote labeled pairs. Given the set D of all labeled and unlabeled pairs, the log likelihood is given by

$$\begin{aligned} l_c(\Theta \mid D; z) = {} & \log (P(\Theta)) + \\ & (1-\lambda) \sum_{i \in Du} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta)\ P(C_j ; \Theta)) + \\ & \lambda \sum_{i \in Dl} \sum_j z_{ij} \log (P(\gamma_i \mid C_j ; \Theta)\ P(C_j ; \Theta)). \end{aligned} \qquad (9)$$

where $0 \le \lambda \le 1$. The first sum is over the unlabeled pairs and the second sum is over the labeled pairs. In the third terms equation (9), we sum over the observed z_{ij}. In the second term, we put in expected values for the z_{ij} based on the initial estimates $P(\gamma_i \mid C_j; \Theta)$ and $P(C_j; \Theta)$. After re-estimating the parameters $P(\gamma_i \mid C_j; \Theta)$ and $P(C_j; \Theta)$) during the M-step (that is in closed form under condition (**CI**)), we put in new expected values and repeat the M-step. The computer algorithms are easily monitored by checking that the likelihood increases after each combination of E- and M-steps and by checking that the sum of the probabilities add to 1.0. We observe that if λ is 1, then we only use training data and our methods correspond to naïve Bayes methods in which training data are available. If λ is 0, then we are in the unsupervised learning situations of Winkler (1988, 1993). Winkler (2002, 2000) provides more details of the computational algorithms.

We create 'pseudo-truth' data sets in which matches are those unlabeled pairs above a certain high cutoff and nonmatches are those unlabeled pairs below a certain low cutoff. Figure 1 illustrates the situation using actual 1990 Decennial Census data in which we plot log of the probability ratio (1) against the log of frequency. With the datasets of this paper, we choose high and low cutoffs in a similar manner so that we do not include in-between pairs in our designated 'pseudo-truth' data sets. We use these 'designated' pseudo-truth data sets in a semi-supervised learning procedure that is nearly identical to the semi-supervised

procedure where we have actual truth data. A key difference from the corresponding procedure with actual truth data is that the sample of labeled pairs is concentrated in the difficult-to-classify in-between region where, in the 'pseudo-truth' situation, we have no way to designate comparable labeled pairs. The sizes of the 'pseudo-truth' data is given in Table 1. The errors associated with the artificial 'pseudo-truth' are given in parentheses following the counts. The *Other* class gives counts of the pairs and proportions of true matches that are not included in the 'pseudo-truth' set of pairs. In the *Other* class, the proportions of matches vary somewhat and would be difficult to determine without training data.

Table 1. 'Pseudo-Truth' Data with Actual Error Rates

	Matches	Nonmatches	*Other*
A×B pairs	8817 (008)	98257 (.001)	9231 (.136)

We determine how accurately we can estimate the lower cumulative distributions of matches and the upper cumulative distribution of nonmatches. This corresponds to the overlap region of the curves of matches and nonmatches. If we can accurately estimate these two tails of distributions, then we can accurately estimate error rates at differing levels. Our comparisons consist of a set of figures in which we compare a plot of the cumulative distribution of estimates of matches versus the true cumulative distribution with the truth represented by the 45 degree line. We also do this for nonmatches. As the plots get closer to the 45 degree lines, the estimates get closer to the truth.

Figure 2a. Estimates vs Truth, File A
Cumulative Matches, Tail of Distribution
Independent EM, Lambda = 0.2

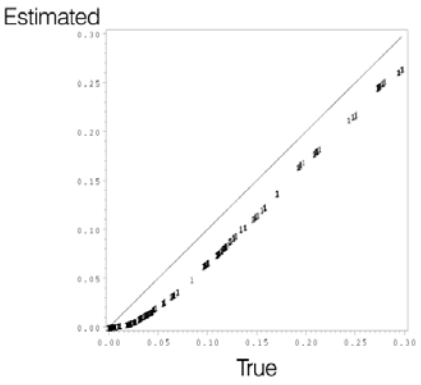
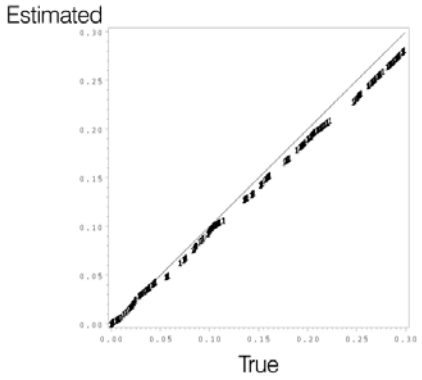


Figure 2b. Estimates vs Truth, File A
Cumulative Nonmatches, Tail of Distribution
Independent EM, Lambda = 0.2



Our primary results are from using the conditional independence model and 'semi-supervised' methods of this paper with the conditional independence model and actual semi-supervised methods of Winkler (2002). With our 'pseudo-truth' data, we obtain the best sets of estimates of the bottom 30% tails of the curve of matches and the top 5% tails of nonmatches with conditional independence and λ=0.2. Figure

2a-b illustrates the set of curves that provide quite accurate fits. The 45 degree line represents the truth whereas the curve represents the cumulative estimates of matches and nonmatches for the left and right tails, respectively. Although we looked at results for λ=0.1, 0.5, and 0.8 and various interactions models, the results under conditional independence (**CI**) were the best with λ=0.2. We also looked at several different ways of constructing the 'pseudo-truth' data. Additionally, we considered other pairs of files in which all of the error-rates estimates were better (closer to the 45 degree line) than those for the pair of files given in Figure 2. The curves with other test decks were typically three times as close to the 45 degree line as the corresponding curves of Belin and Rubin (1995).

We can use the model given in this section (essentially the same as in Winkler 2000, 2002) and the associated EM software to obtain all of the EM estimates that are used in this paper. In each situation, the inputs will vary significantly.

7. Name and Address Pre-processing and Standardization

Herzog et al. (2007) observed that the clean-up and standardization of the inputs prior to matching yields a much larger improvement in matching efficacy than improved parameter estimation due to effective use of the EM and other parameter estimation algorithms. Bilmes (1998) provides an excellent tutorial on the EM algorithm and the more general (but closely related) hidden Markov methods. Borkar et al. (2001) provide a very effective application of the hidden Markov methods to address standardization. Churches et al. (2002) provide applications of slightly different hidden Markov algorithms for both name and address standardization. They demonstrate that hidden Markov methods make it very straightforward to develop training data and that methods outperform rule-based methods (e.g., Winkler 1995) for southern and south-east Asian types of addresses but not necessarily for western style addresses as used in Australia, Western Europe, and the Americas. The rule-based methods (e.g. Winkler 1995) outperform the initial applications of hidden Markov methods to names. Table 2 provides an example of standardized and parsed names.

Table 2. Examples of name parsing

Standardized
1. DR John J Smith MD
2. Smith DRY FRM
3. Smith & Son ENTP

	Parsed						
	PRE	FIRST	MID	LAST	POST1	POST2	BUS1	BUS2
1.	DR	John	J	Smith	MD			
2.				Smith			DRY	FRM
3.				Smith		Son	ENTP	

8. Concluding Remarks

This paper provides some background on applications of machine learning methods to parameter estimation and false-match rate estimation when there is no training data. No training data is the typical situation in statistical agencies and health organizations. It provides a description of the literature on how hidden Markov models (which generalize EM methods) are effectively used in standardizing names and addresses. Winkler (2006) gives a fairly extensive overview of the machine learning methods that are used for record linkage and a set of open research problems.

1/ This report is released to inform interested parties of (ongoing) research and to encourage discussion (of work in progress). Any views expressed on (statistical, methodological, technical, or operational) issues are those of the author(s) and not necessarily those of the U.S. Census Bureau.

REFERENCES

Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False- Match Rates in Record Linkage," *Journal of the American Statistical Association*, 90, 694-707.

Bhattacharya, I., and Getoor, L. (2006), "A Latent Dirichlet Allocation Model for Entity Resolution" *SDM '06 – best paper.*

Bilmes, J. A. (1998), "A Gentle Tutorial of the EM Algorithm and its Application for Parameter Estimation for Gaussian Mixture and Hidden Markov Models," International Computer Science Institute, Berkeley, CA, available at http://ssli.ee.washington.edu/people/bilmes/mypapers/em.pdf .

Borkar, V., Deshmukh, K., and Sarawagi, S. (2001), "Automatic Segmentation of Text into Structured Records," Association of Computing Machinery SIGMOD 2001, 175-186.

Churches, T., Christen, P., Lu, J. and Zhu, J. X. (2002), "Preparation of Name and Address Data for Record Linkage Using Hidden Markov Models," *BioMed Central Medical Informatics and Decision Making*, 2 (9), available at http://www.biomedcentral.com/1472-6947/2/9/.

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

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

Freund, Y. and Schapire, R. E. (1996), "Experiments with a New Boosting Algorithm," *Machine Learning: Proceedings of the Thirteenth International Conference*, 148-156.

Friedman, J., Hastie, T.,Tibshirani, R. (2000), "Additive Logistic Regression: a Statistical View of Boosting," *Annals of Statistics*, 28, 337-407.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer: New York.

Herzog, T. N., Scheuren, F., and Winkler, W. E. (2007), *Data Quality and Record Linkage*," Springer: New York, N.Y.

Larsen, M. D., and Rubin, D. B. (2001), "Alterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 79, 32-41.

Newcombe, H. B., Kennedy, J. M. Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954-959.

Newcombe, H.B. and Kennedy, J. M. (1962) "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information" *Communications of the Association for Computing Machinery*, 5, 563-567.

Nigam, K., McCallum, A. K, Thrun, S., and Mitchell, T. (2000), "Text Classification from Labeled and Unlabelled Documents using EM, *Machine Learning*, 39, 103-134.

Ravikumar, P. and Cohen, W. W. (2004), "A Hierarchical Graphical Model for Record Linkage," *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, Banff, Calgary, CA, July 2004.

Scheuren, F., and Winkler, W. E. (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, 19, 39-58.

Vapnik, V. (2000), *The Nature of Statistical Learning Theory (2nd Edition)*, Berlin: Springer-Verlag.

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*, 667-671.

Winkler, W. E. (1989), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Fifth Census Bureau Annual Research Conference*, 145-155.

Winkler, W. E. (1990b), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359.

Winkler, W. E. (1993), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-279.

Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox *et al.* (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384 (also available at http://www.fcsm.gov/working-papers/wwinkler.pdf).

Winkler, W. E. (2000), "Machine Learning, Information Retrieval, and Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20-29. (also available at http://www.niss.org/affiliates/dqworkshop/papers/winkler.pdf).

Winkler, W. E. (2002), "Record Linkage and Bayesian Networks," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM (also at http://www.census.gov/srd/www/byyear.html).

Winkler, W. E. (2006a), "Overview of Record Linkage and Current Research Directions," available at http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf .

Winkler, W. E. (2006b), "Automatic Estimation of Record Linkage False Match Rates," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM, also at http://www.census.gov/srd/papers/pdf/rrs2007-05.pdf .

Winkler, W. E., Yancey, W. E., and Porter, E. H. (2010), "Fast Record Linkage of Very Large Files in Support of Decennial and Administrative Records Projects," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, CD-ROM.