

MULTIPLE IMPUTATION ANALYSIS OF RECORDS LINKED USING MIXTURE MODELS

Michael Larsen¹

ABSTRACT

Computerized record linkage is used to join together two files that contain information on the same individuals, but lack unique personal identification codes. Mixture models, when fit to measurements of agreement between pairs of records on the two files, produce estimated probabilities that a pair is a match and can be used to impute links between record pairs. When estimating relationships between variables on the two linked files, it is necessary to use estimators that adjust for potential linking error. Multiple sets of imputed links can be generated under the mixture models, and, for each set of imputed links, the relationships between variables can be estimated. The difference in estimates can be used to express uncertainty in the true relationship due to unknown match status of the pairs. Through simulation we compare various procedures for using mixture model results together with estimators of linear regression coefficients. Our work here is preliminary and intended to generate interest and motivate further work.

KEY WORDS: Data augmentation; EM algorithm; Fellegi and Sunter; Measurement error; Scheuren and Winkler.

RÉSUMÉ

Le couplage automatisé d'enregistrements est employé pour joindre ensemble deux fichiers qui contiennent de l'information sur les mêmes individus, mais qui ne contiennent pas de codes d'identification personnelle uniques. Les modèles mixtes, lorsqu'ajustés pour mesurer l'accord entre les paires d'enregistrements des deux fichiers, produisent des probabilités estimées qu'une paire soit un couple et peuvent être utilisés pour imputer des liens entre les paires d'enregistrements. Lors de l'estimation des relations entre les variables des deux fichiers appariés, il est nécessaire d'utiliser des estimateurs qui s'ajustent selon les erreurs potentielles de couplage. Plusieurs ensembles de liens imputés peuvent être générés à l'aide des modèles mixtes, et, pour chaque ensemble de liens imputés, les relations entre les variables peuvent être estimées. Les différences entre les estimations peuvent être utilisées pour exprimer de l'incertitude envers la vraie relation due à l'état inconnu du couplage des paires. À l'aide des simulations, nous comparons diverses procédures pour l'utilisation conjointe des résultats des modèles mixtes et des estimateurs des coefficients de régression linéaire. Notre travail présenté ici est préliminaire et vise à créer de l'intérêt et à encourager la poursuite du travail.

MOTS CLÉS : Augmentation de données; algorithme EM; Fellegi et Sunter; erreur de mesure; Scheuren et Winkler.

1. INTRODUCTION

The goal of record linkage is identifying pairs of records (a,b), a from file A and b from file B, that correspond to the same person or entity. If there are not unique codes that identify the matching pairs of records, then links can be designated by comparing information contained in the two files. Clerks can review pairs of records, but that is time consuming and costly. Computers can quickly score the level of

agreement between records in the two files, and, given a decision rule, designate pairs as links or nonlinks. Often the designated links correspond closely to matching pairs.

Mixture models can be used to estimate the probability that a pair of records is a match before clerical review determines actual status. Estimates can be calculated using maximum likelihood or Bayesian techniques. Maximum likelihood estimates of mixture model

¹ Michael D. Larsen, University of Chicago, Department of Statistics, 5734 University Ave, Chicago Illinois 60637 USA, larsen@galton.uchicago.edu.

parameters can be used to rank pairs in terms of their probability of being matches. Bayesian simulations from the posterior distributions of parameters can be used to multiply impute, or multiply assign, links between records in the two files.

Here we consider the problem of estimating a linear regression relationship between two variables, one recorded in each of file A and B, and a measurement of uncertainty in our estimate after the files have been linked using mixture models. Errors in linking pairs create a measurement error problem, which attenuates the apparent regression relationship. Estimation methods of Scheuren and Winkler (1993) are among those we consider. When links are multiply imputed, estimation techniques can be applied to each version of the linked data set. The collection of estimates and measurements of uncertainty can be combined according to methods of Rubin (1987) to reflect uncertainty in linking. The goal of integrating the use of mixture models in computerized record linkage and model-based adjustments for analysis of linked files is to produce methods that are quick, flexible without requiring training data or clerical review, and accurately estimate relationships.

In section 2 we review record-linkage theory of Fellegi and Sunter (1969) and describe simulation data sets. Section 3 discuss mixture models and their estimation. Section 4 presents regression estimators and our use of multiple imputation. Sections 5 and 6 contain results and discussion, respectively.

2. RECORD LINKAGE

Assume the files A and B do not contain duplicates and that K comparisons can be made between information in pair of records (a,b) . Let $\{v_k(a), k=1, \dots, K\}$ be K pieces of information in file A on person a and $\{w_k(b), k=1, \dots, K\}$ be the information in file B on person b . The comparison of information on pair (a,b) yields a comparison vector $I(a,b) = \{I_k(a,b), k=1, \dots, K\}$ where $I_k(a,b)=1$ if $v_k(a)=w_k(b)$ and 0 otherwise. If a from A and b from B were produced by the same person, then it would be expected that $I(a,b)$ would contain more '1' entries than if they were produced by different people.

Not all pairs are compared since such a procedure would create many pairs and most of them would not be plausible as matches (e.g., (a,b) live in different geographical areas, have different first letters of last name, etc.). The records in the two files are divided into S 'blocks' and comparisons are made between records within blocks. We will refer to the records in

block s in file A as $A(s)$ and in file B as $B(s)$. It is assumed here that blocking does not create errors in matching, i.e., that records involved in matches are blocked together.

2.1. Fellegi-Sunter Theory

Fellegi and Sunter (1969) proposed decomposing the space $A \times B$ into sets M , the matches, and U , the unmatched or nonmatching pairs. They showed that, given probabilities $P(I|M)$ and $P(I|U)$, if pairs with $P(I|M)/P(I|U)$ above a cutoff are designated links and pairs with the ratio below a second cutoff are designated nonlinks, then the number of undeclared pairs is minimized at the error levels corresponding to the two cutoffs. In practice, there have been several methods of estimating the probabilities involved in the ratio (see, e.g., Belin and Rubin 1995, Winkler 1995, Larsen and Rubin 1999).

2.2. Simulated Data Sets

Here we simulate one hundred data sets to study the use of mixture models for estimating the probabilities used in the Fellegi-Sunter ratio and the proposed estimators of the regression relationship. It is assumed that all records a from A and b from B have a match in the other file. For an individual data set, the size n of each file was generated uniformly from 500 to 4000. A predictor variable X was generated as standard normal for each $i=1, \dots, n$. Correlation r was generated uniformly from 0.3 to 0.9. A response variable Y was generated to be standard normal with correlation r with X . If match status were known for every pair, the variables $\{(X_i, Y_i), i=1, \dots, n\}$ would be recorded and we would estimate the regression relationship using standard least-squares formulas.

Variables $\{v_k(i), k=1, \dots, K\}$ and $\{w_k(i), k=1, \dots, K\}$ are now generated for $i=1, \dots, n$. The number of comparisons K was uniformly chosen between 5 and 10, yielding 32 to 1024 possible comparison vectors. For each $k=1, \dots, K$, the numbers of values v_k and w_k can assume are generated uniformly from two to ten (so $v_k(i)$ will have probability 0.1 to 0.5 of matching $w_k(i')$ for i different from i'). The probability that $v_k(i) = w_k(i)$ is generated uniformly from 0.6 to 1.0. One can consider file A to contain variables v and X and B to have variables w and Y . For pair (a,b) , the comparison vector $I(a,b)$ has k^{th} element 1 if $v_k(a)=w_k(b)$ and 0 otherwise.

Blocking information also is generated for each data set. Block sizes are generated uniformly from 5 to 20 such that blocking does not create errors in matching,

i.e., that records involved in matches are blocked together. If $n=4000$ in both files, there are 16 million possible pairs, but with the blocking there are only between 20000 and 80000 pairs considered.

In our simulations, the agreements and disagreements across fields of information are independent for matches and nonmatches. That is, the conditional independence (Fellegi and Sunter 1969, Winkler 1988) or latent-class (Haberman 1974 and 1979, Goodman 1974) model is true for the simulated data. In reality, there is evidence that this model is not accurate (Belin and Rubin 1995, Larsen and Rubin 1999).

3. MIXTURE MODELS

Consider the observed data to be the patterns of agreements for the pairs of records considered, $\{I(a,b), a \text{ from } A(s), b \text{ from } B(s), s=1, \dots, S\}$. Before clerical review determines match status, the data can be considered to arise from a mixture of comparison vectors from matches and nonmatches. The probability of observing pattern $I(a,b)$ is

$$P(I(a,b)) = P(I(a,b) | M) P(M) + P(I(a,b) | U) P(U), \quad (1)$$

where $P(I(a,b)|M)$ and $P(I(a,b)|U)$ are the probabilities of the pattern among the matches and nonmatches, respectively, $P(M)$ is the probability that a pair is a match, and $P(U)=1-P(M)$. The observed-data likelihood is a product of (1) over pairs (a,b) .

The conditional-independence model specifies that the conditional probability of pattern $I(a,b)$ is the product of the probabilities for agreeing or disagreeing on the K fields of information is

$$P(I(a,b) | C) = \text{product}_{k=1, \dots, K} (P(I_k(a,b) | C) / (I_k(a,b) + (1 - P(I_k(a,b) | C)) (1 - I_k(a,b)))),$$

where C represents M or U . More complex models have been used to model discrete mixtures (e.g., Becker and Yang 1998) and in record linkage (e.g., Winkler 1989, 1993, Armstrong and Mayda 1993, Thibaudeau 1993).

Let $z(a,b)=1$ if pair (a,b) is a match and 0 otherwise. The vector of indicator variables z is unobserved before clerical review. The probability that $z(a,b)$ equals 1 given the agreement pattern $I(a,b)$ is

$$P(z(a,b) = 1 | I(a,b)) = P(M | I(a,b)) = P(M) P(I(a,b) | M) / P(I(a,b)).$$

Fellegi and Sunter's (1969) ratio $P(I|M)/P(I|U)$ can be computed if values are given for the parameters of the mixture model, which in the case of the conditional independence model are $\{P(C), P(I_k(a,b) | C), k=1, \dots, K, C=M,U\}$. If z were observed, the complete-data likelihood would be a product over pairs (a,b) of

$$P(I(a,b) | M) P(M)^{z(a,b)} + P(I(a,b) | U) P(U)^{(1-z(a,b))}.$$

3.1. Maximum Likelihood Estimation

A convenient method of finding maximum likelihood estimates of mixture model parameters is to treat the unobserved indicators z as missing data and use the EM (Dempster, Laird, Rubin 1977) or ECM (Meng and Rubin 1993) algorithms. The algorithms have been presented for record linkage using mixture models by the sources noted previously in this section.

In the case of the conditional-independence model, the EM algorithm may be used. Given current estimates of parameter values, the E-step is completed by computing $P(z(a,b)=1 | I(a,b))$, the expected value of $z(a,b)$, for all pairs (a,b) , which is the same for each unique comparison vector I . At iteration $t+1$ call the values $z(a,b)^{t+1}$. The M-step is completed by calculating maximum likelihood estimates of parameters with entries in z held at their current expectations. The estimate of $P(M)$ is the sum of $z(a,b)^{t+1}$ over pairs (a,b) divided by the total number of pairs. The estimate of $P(I_k(a,b)=1|M)$ is the sum of $z(a,b)^{t+1}$ over pairs (a,b) for which $I_k(a,b)$ equals 1 divided by the sum of $z(a,b)^{t+1}$ over all pairs. The estimate of $P(I_k(a,b)=1|U)$ is the sum of $1-z(a,b)^{t+1}$ over pairs (a,b) for which $I_k(a,b)$ equals 1 divided by the sum of $1-z(a,b)^{t+1}$ over all pairs. The algorithm iterates between the E- and M-steps until the observed-data likelihood converges to a maximum.

3.2. Bayesian Estimation

Bayesian analysis requires prior distributions on parameters in addition to the likelihood. Prior distributions could reflect knowledge about record linkage and information about the data set. In this application, we consider only the conditional-independence model and simple prior distributions with little "weight" relative to the size of the data set. The prior distribution on $P(M)$ is $\text{Beta}(\delta, \delta)$, where $\delta = (.005)2^K$. The prior distributions on $P(I_k(a,b)=1|M)$ and $P(I_k(a,b)=1|U)$, $k=1, \dots, K$, are $\text{Beta}(\delta/2, \delta/2)$. All prior distributions are mutually independent.

Posterior distributions can be simulated by sampling from alternating conditional distributions in steps analogous to those of EM. Given initial values for parameters, each variable $z(a,b)$ is drawn independently from a Bernoulli distribution with parameter equal to $P(z(a,b)=1 | I(a,b))$. At iteration $t+1$ call the values $z(a,b)^{t+1}$. Given the drawn values of z , the parameters are drawn independently from their current conditional distributions. $P(M)$ is Beta with parameters $(\delta + \sum z(a,b)^{t+1})$ and $(\delta + \sum (1 - z(a,b)^{t+1}))$, where \sum indicates summation over all pairs. $P(I_k(a,b)=1|M)$ is $\text{Beta}(\delta/2 + \sum z(a,b)^{t+1} I_k(a,b), \delta/2 + \sum z(a,b)^{t+1} (1 - I_k(a,b)))$ for $k=1, \dots, K$. $P(I_k(a,b)=1|U)$ is $\text{Beta}(\delta/2 + \sum (1 - z(a,b)^{t+1}) I_k(a,b), \delta/2 + \sum (1 - z(a,b)^{t+1}) (1 - I_k(a,b)))$ for $k=1, \dots, K$. The algorithm cycles between drawing values of parameters and values of missing indicator variables until drawn values are being sampled from the posterior distribution (see, e.g., Gelman and Rubin 1992, Larsen 1994).

The Bayesian sampling algorithm produces several sets of values for mixture model parameters from the posterior distribution and several sets of imputed links. Each set of parameter values can be used to calculate probabilities that pairs are matches, $P(z(a,b)=1 | I(a,b))$. A set of imputed links is a set of drawn values for z that identify which pairs (a,b) are designated as links at a particular iteration of the algorithm.

Gelman et al. (1995) and Schafer (1997), and references therein have given more complex Bayesian models for discrete data that allow interactions between fields.

3.3. One-to-one Assignment and Blocking

The algorithms presented here do not explicitly model the blocking structure and do not enforce in the model the assumption that matches occur only within blocks. Methods used by Larsen and Rubin (1999) and Winkler (1995) also do not use blocking structure explicitly. Future work will have to investigate models and algorithms for using blocking.

The algorithms also do not enforce one-to-one linking. That is, for an individual a in file A, there is no constraint in the model that forces the probability that a has a match to be 1. Several records in file B could agree closely with person a and have high model-based probabilities of matching person a . Future work also will have to incorporate appropriate constraints.

Despite the two limitations mentioned above, mixture models such as those used in this paper have produced

good results in applications. The models identify patterns that are typical of matches and do not introduce much error by ignoring one-to-one assignment and blocking issues. For examples and discussion, see Winkler (1994, 1995), Larsen (1996), Larsen and Rubin (1999), and Alvey and Jamerson (1997), and references in these sources.

4. REGRESSION OF MATCHED FILES

In this section we consider methods for estimating a linear regression relationship and the uncertainty in our estimates on a data set created through record linkage. We assume the model is $Y=X\beta + \epsilon$ where error terms are independent with mean 0 and variance σ^2 . If clerks had found the matches without error, then we would simply compute the linear least-squares regression estimate of β and its standard error.

Suppose according to the mixture model pair (a,b) is a designated link. For person a , variable X_i is observed on file A. Let the response variable observed on file B and linked to record a be variable Z_i . The “naive” estimator, which simply computes the linear least-squares regression of Z on X and the usual formula for its standard error, tends to underestimate the regression relationship when there is linking error, because inaccuracies in linking error introduce a type of measurement error (Fuller 1987).

4.1. Scheuren-Winkler

Scheuren and Winkler (1993) extend work of Neter, Maynes, and Ramanathan (1965) to adjust for matching error. Let Z_i equal Y_j , the response that should be matched to X_i , with probability p_i , and equal Y_j with probability q_{ij} , j not equal to i . Scheuren and Winkler (1993) show that the naive estimator should be adjusted for bias. If b_{ZX} is the naive estimator, then their estimator is $b_{YX}^{SW} = b_{ZX} - B_{YX}/\sigma_X^2$, where B_{YX} is given in Scheuren and Winkler (1993) and σ_X^2 is the observed variance of the variable X . They also give indications of how to estimate the variance of their estimator.

Scheuren and Winkler (1993) advocate using only the best link and the second best link in calculations. The method can be extended to use all pairs considered as possible links. In our application, the probabilities used to estimate the probabilities p_i and q_{ij} are taken from the mixture models. Maximum likelihood or Bayesian estimates can be used. Scheuren and Winkler (1993) used the method of Belin and Rubin (1995) to estimate probabilities. In an extension of

their work, Scheuren and Winkler (1997) iteratively compute regression estimates and initially use just the best links. Future work will consider incorporating their extensions with the methods presented here.

4.2. Multiple Imputation and Truncation

In the Bayesian context, each draw from the posterior distribution of mixture model parameters can be used to estimate probabilities and compute either the naïve or Scheuren and Winkler's (1993) regression estimates and estimates of variance. Multiple imputation formulas (Rubin 1987) can be used to combine results from the separate sets of results into a single estimate and estimate of variance. The multiple imputation estimate of variance reflects both the estimated variance of the estimates as well as the variability among estimates.

Scheuren and Winkler (1993) report results by matching weight or log of the ratio used in the Fellegi-Sunter (1969) algorithm. In general, regression estimates computed using the pairs with higher matching weights have lower bias than those computed with all pairs. The naïve estimates and two versions of b_{yx}^{sw} , using all and using only two links per record, could be calculated for pairs with estimated match probability above a cutoff value, say 0.5. Whether or not estimates should be formed using only the best links will be explored in future research.

5. RESULTS

One hundred data sets were generated as described in section 2.2. For each data set, we calculate naïve estimates and Scheuren and Winkler's (1993) estimates using two and using all pairs with probabilities from the mixture models. Maximum likelihood estimates were used to calculate probabilities for one set of estimates. Bayesian estimates were used for estimates based on multiple imputation methodology. Results are presented that describe how well the methods

reproduce the true correlation between variables X and Y and the degree to which 95% confidence intervals cover the true correlation. Note that the correlation is the same as the regression slope parameter in our simulation.

Table 1 contains measures of how well the estimation methods reproduce the correlation and how often intervals based on the methods cover the true correlation for the 100 data sets. The naïve estimator has the largest values of sum of squared errors (sum over 100 data sets of $(r - \text{estimate})^2$), sum of squared relative error (sum of $(r - \text{estimate})^2/r$), sum of absolute error (sum of $|r - \text{estimate}|$), sum of absolute relative error (sum of $|r - \text{estimate}|/r$), and bias (an average of -0.079 per data set). The 95% nominal coverage of intervals made using the naïve estimator and twice its standard error cover the true values less than half the time. When Bayesian estimates of probabilities are used to produce ten estimates for multiple imputation, coverage is increased only slightly.

When Scheuren and Winkler's (1993) estimator is used, measurements of performance are reduced by 40 to 60%. Average bias reduces to -0.029 per data set. Improvement in coverage occurs both with and without multiple imputation. When all pairs are used in Scheuren and Winkler's method, slight additional improvement is found. Average bias is reduced to about -0.01 and coverage is increased. Again multiple imputation increases coverage.

6. DISCUSSION AND FUTURE WORK

Scheuren and Winkler's (1993) estimates are much better than the naïve estimates. When their method is used with all pairs rather than just two pairs, the results are a little better still. When multiple imputation is used to reflect uncertainty about the linkages between the files, coverage of the true correlation is again improved. Additionally, the mixture model estimates do not involve clerical review or training data from

Table 1. Measurements of performance of three estimators.

	Naïve Estimator	Scheuren-Winkler (2 links)	Scheuren-Winkler (All pairs)
Probabilities based on MLE			
Sum of squared error	1.031	0.396	0.302
Sum of squared relative error	1.621	0.655	0.481
Sum of absolute error	8.146	4.832	4.190
Sum of absolute relative error	13.693	8.209	7.082
Sum of bias	-7.916	-2.916	-0.955
95% nominal coverage	44%	55%	64%
Probabilities from Bayesian estimation			
95% nominal coverage incorporating multiple imputation	46%	66%	72%

another record-linkage operation. Although the methods used here are much better than the naive estimators are, there still is room for improvement. It could be that the differences between the mixture model probability estimates and the actual probabilities are responsible for the failure to achieve 95% coverage.

One possible area for improvement is to use ideas of Scheuren and Winkler (1997) to iteratively improve regression and linking results. Larsen and Rubin (1999) have investigated the combination of mixture models with a small amount clerically reviewed data to improve mixture model estimates of probabilities. In general, these iterative methods seem to have a lot of promise for improving model-based record linkage results.

Another area for research is using blocking information directly in the mixture models and in models for adjustment of regression estimates. A related concern is formulating the models so that one-to-one linking is enforced. It is possible that the additional modeling effort could yield better estimates of probabilities for use in current estimators as well as new estimators.

Furthermore, although gains were realized when all pairs rather than just the best links were used in adjustment in the simulations, it is not known whether this is a general result or was a finding particular to this simulation. Studies could be conducted to choose cutoffs below which pairs with low probabilities of being matches are not used in estimation.

ACKNOWLEDGEMENTS

The author would like to thank William Winkler and Donald Rubin for introducing him to the problem of record linkage, Partha Lahiri for discussions, and the organizers of the conference in Regina, Saskatchewan.

REFERENCES

- Alvey, W., and Jamerson, B. (Eds.) (1997). Record Linkage Techniques -- 1997. *Proceedings of an International Record Linkage Workshop and Exposition*, March 20-21, 1997, Arlington, VA.
- Armstrong, J.B., and Mayda, J.E. (1993). Estimation of record linkage models using dependent data. *Survey Methodology*, 19, 137-147.
- Becker, M.P., and Yang, I. (1998). Latent Class Marginal Models for Cross-Classifications of Counts. *Sociological Methodology*, 28, 293-325.
- Belin, T.R., Rubin, D.B. (1995). A method for calibrating false match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.
- 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, Series B*, 39, 1-22.
- Fellegi, I.P., and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- Fuller, W.A. (1987). *Measurement error models*. New York: John Wiley.
- Gelman, A. Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian data analysis*. London: Chapman & Hall.
- Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.
- Haberman, S.J. (1974). Log-linear models for frequency tables derived by indirect observation: maximum likelihood equations. *Annals of Statistics*, 2, 911-924.
- Haberman, S.J. (1979). Product models for frequency tables involving indirect observation. *Annals of Statistics*, 5, 1124-1127.
- Larsen, M.D. (1994). Data augmentation with Bayesian iterative proportional fitting applied to a Census Bureau latent-class problem. *Proceedings of Government Statistics Section, American Statistical Association*, 116-121.
- Larsen, M.D. (1996). *Bayesian approaches to finite mixture models*. Ph.D. Thesis. Harvard University.
- Larsen, M.D., and Rubin, D.B. (1999). Iterative automatic record linkage using mixture models. *In preparation*.

- Meng, X-L., and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267- 278.
- Neter, J., Maynes, E.S, and Ramanathan, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.
- Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley.
- Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Scheuren, F., and Winkler, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39- 58.
- Scheuren, F., and Winkler, W.E. (1997), Regression analysis of data files that are computer matched -- Part II. *Survey Methodology*, 23, 157- 165.
- Thibaudeau, Y. (1993). The discriminating power of dependency structures in record linkage. *Survey Methodology*, 19, 31-38.
- Winkler, W.E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 667-671.
- Winkler, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Bureau of the Census Annual Research Conference*, 5, 145-155.
- Winkler, W.E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 274-279.
- Winkler, W.E. (1994). Advanced methods of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472.
- Winkler, W.E. (1995). Matching and record linkage. *Business Survey Methods*, (Eds. B.G. Cox et al.). New York: John Wiley, 355-384.