Recursive Merging and Analysis of Administrative Lists and Data

Fritz Scheuren, George Washington University, and William E. Winkler, Bureau of the Census

Use of multiple administrative lists for statistical purposes has widespread appeal due to the cost savings from not collecting data and to possible increased accuracy because analyses are not based on relatively small samples. Producing accurate analyses when quantitative data reside in multiple files has previously been virtually impossible if unique common identifiers are not present. This paper demonstrates a methodology for analyzing two or more files when the only common information is name and address that are subject to significant error and each source file contains quantitative data. Such a situation might arise with lists of businesses. We assume that a small proportion of records can be accurately matched using name and address information. The matched pairs are used to build an edit/imputation model that is then used to add predicted quantitative values to each file. Matching is then repeated with common quantitative data and with name and address information. If necessary, the edit/impute and matching steps can be repeated in a recursive fashion.

Introduction

To model the energy economy properly, an economist might need company-specific microdata on the fuel and feedstocks used by companies that are only available from Agency A and corresponding microdata on the goods produced for companies that are only available from Agency B. To model the health of individuals in society, a demographer or health sciences policy worker might need individual-specific information on those receiving social benefits from Agencies B1, B2, and B3, corresponding income information from Agency I, and information on health services from Agencies H1 and H2. Such modeling is possible if analysts have access to the microdata and if unique, common identifiers are available (e.g., Oh and Scheuren, 1975; Jabine and Scheuren, 1986). If the only common identifiers are error-prone, nonunique name and address information, then probabilistic matching techniques (e.g., Newcombe et al., 1959; Fellegi and Sunter, 1969) are needed.

In earlier work (Scheuren and Winkler, 1993), we provided theory showing that elementary regression analyses could be accurately adjusted for matching error. For applications where name and address information was of sufficiently high quality, we applied an error-rate estimation procedure of Belin and Rubin (1995). In later work (Winkler and Scheuren, 1995, 1996), we showed that we could actually use noncommon quantitative data from the two files to improve matching and adjust statistical analyses for matching error. The main requirement--even in heretofore seemingly impossible situations--was that there exists a very small subset of pairs that could be accurately matched using name and address information only and that the noncommon quantitative data be highly or moderately correlated.

The intuitive underpinnings of our methods are based on record linkage (**RL**) and edit/imputation (**EI**). The ideas of modern **RL** were introduced by Newcombe (Newcombe et al., 1959) and mathematically formalized by Fellegi and Sunter (1969). Recent methods are described in Winkler (1994, 1995). **EI** has traditionally been used to clean up erroneous data in files. The most pertinent methods are based on the **EI** model of Fellegi and Holt (1976).

To adjust a statistical analysis for matching error, we employ a four-step recursive approach that is very powerful. We begin with an enhanced **RL** approach (e.g., Winkler, 1994; Belin and Rubin, 1995) to delineate a subset of pairs of records in which the matching error rate is estimated to be very low. We perform a regression analysis, **RA,** on the low-error-rate linked records and partially adjust the regression model on the remainder of the pairs by applying previous methods (Scheuren and Winkler, 1993). Then, we refine the **EI** model using traditional outlier-detection methods to edit and impute outliers in the remainder of the linked pairs. Another regression analysis (**RA**) is done, and, this time, the results are fed back into the linkage step so that the **RL** step can be improved (and so on). The cycle continues until the analytic results desired cease to change. Schematically, we have

↗RA↘

RL⇦ RA⇦ EI

Beginning with the introduction, this paper is divided into five sections. In the second section, we provide background on edit/imputation and record linkage. The third section describes the empirical data files constructed and the regression analyses undertaken. In the fourth section, we present results. The final section consists of some conclusions and areas for future study.

■ EI and RL Methods Reviewed

In this section, we undertake a short review of Edit/Imputation (**EI**) and Record Linkage (**RL**) methods. Our purpose is not to describe them in detail but simply to set the stage for the present application. Because Regression Analysis (**RA**) is so well known, our treatment of it is covered only in the particular application (the third section).

Edit/Imputation

Methods of **editing** microdata have traditionally dealt with logical inconsistencies in data bases. Software consisted of **if-then-else** rules that were data-base-specific and very difficult to maintain or modify. Imputation methods were part of the set of **if-then-else** rules and could yield revised records that still failed edits. In a major theoretical advance that broke with prior statistical methods, Fellegi and Holt (1976) introduced operations-research-based methods that both provided a means of checking the logical consistency of an edit system and assured that an edit-failing record could always be updated with imputed values so that the revised record satisfies all edits. An additional advantage of Fellegi-Holt systems is that their edit methods tie directly with current methods of **imputing** microdata (e.g., Little and Rubin, 1987).

Although we will only consider continuous data in this paper, **EI** techniques also hold for discrete data and combinations of discrete and continuous data. In any event, suppose we have continuous data. In this case, a collection of edits might consist of rules for each record of the form

$$\mathbf{c_1X < Y > c_2X}$$

In words,

> **If Y less than c_1X and greater than c_2X, then the data record should be reviewed.**

Here, Y may be total wages, X the number of employees, and c_1 and c_2 constants such that $c_1 < c_2$.

While Fellegi-Holt systems have theoretical advantages, implementation has been very slow because of the difficulty in developing general set covering routines needed for implicit-edit generation and integer programming routines for error localization (i.e., determining the minimum number of fields to impute).

Record Linkage

A record linkage process attempts to classify pairs in a product space **A × B** from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (e.g., Newcombe et al., 1959; Newcombe et al., 1992), Fellegi and Sunter (1969) considered ratios **R** of probabilities of the form

$$\mathbf{R = Pr(\gamma \in \Gamma \mid M) / Pr(\gamma \in \Gamma \mid U)}$$

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 surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Scheuren or Winkler, occur. The fields compared (surname, first name, age) are called <u>matching variables</u>.

The decision rule is given by

If R > *Upper*, then designate pair as a link.

If *Lower* ≤ R ≤ *Upper*, then designate pair as a possible link and hold for clerical review.

If R < *Lower*, then designate pair as a nonlink.

Fellegi and Sunter (1969) showed that this decision rule is optimal in the sense that, for any pair of fixed bounds on **R**, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds, *Upper* and *Lower*, are determined by the error bounds. We call the ratio **R** or any monotonely increasing transformation of it (typically a logarithm) a matching weight or total agreement weight.

With the availability of inexpensive computing power, there has been an outpouring of new work on record linkage techniques (e.g., Jaro, 1989; Newcombe, Fair, Lalonde, 1992; Winkler 1994, 1995). The new computer-intensive methods reduce, or even sometimes eliminate, the need for clerical review.

Simulation Setting

The intent of our simulations is to use matching scenarios that are worse than what some users will encounter and to use quantitative data that are both easy to understand and difficult to use in matching.

Matching Scenarios

For our simulations, we considered one matching scenario in which matches are virtually indistinquishable from nonmatches and three levels of file overlap. In our earlier work (Scheuren and Winkler, 1993), we considered three matching scenarios in which matches are more easily distinguished from nonmatches than in the scenario of this paper and only the high-file-overlap situation of this paper. The basic idea was to generate data having known distributional properties, adjoin the data to two files that would be matched, and then evaluate the effect of increasing amounts of matching error on analyses. Because the methods of this paper work better, we only consider a matching scenario that we label 2nd poor because it is more difficult than the poor (most difficult) scenario we considered previously.

We started with two files (sizes 12,000 and 15,000) having good matching information and for which true match status was known. In the high overlap situation, about 10,000 of these were true matches (before introducing errors)--for a rate on the smaller or base file of about 83 percent. In the medium overlap situation, we took a sample of one file so that the overlap of the two files being matched was approximately 25 percent. In the low overlap situation, we took samples of both files so that the overlap of the files being matched was approximately 5 percent.

We then generated quantitative data with known distributional properties and adjoined the data to the files. These variations are described below and shown in figure 1 where we show the poor scenario (labelled 1st poor) of the previous paper and the 2nd poor scenario of this paper. In the figure, the match weight, the logarithm of **R**, is plotted on the horizontal axis with the frequency, also expressed in logs, plotted on the vertical axis. Matches (or true links) appear as asterisks (*), while nonmatches (or true nonlinks) appear as small circles (o):

- **1st Poor Scenario (figure 1a).**--The 1st poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names. Moderately severe typographical errors were made in one fourth of the addresses. Matching probabilities were chosen that deviated substantially from optimal. The intent was for them to be selected in a manner that a practitioner might choose after gaining only a little experience. The true mismatch rate here was 10.1 percent.

- **2nd Poor Scenario (figure 1b).**--The 2nd poor matching scenario consisted of using last name, first name, and one address variation. Minor typographical errors were introduced independently into one third of the last names and one third of the first names. Severe typographical errors were made in one fourth of the addresses. Matching probabilities were chosen that deviated substantially from optimal. The intent was to represent situations that often occur with lists of businesses in which the linker has little control over the quality of the lists. The true mismatch rate here was 14.6 percent.

With the various scenarios, our ability to distinguish between true links and true nonlinks differs significantly. With the 1st poor scenario, the overlap is substantial (figure 1a); and, with the second poor scheme, the overlap is almost total (figure 1b). In the earlier work, we showed that our theoretical adjustment procedure worked well using the known true match rates in our data sets. For situations where the curves of true links and true nonlinks were reasonably well separated, we accurately estimated error rates via a procedure of Belin and Rubin (1995), and our procedure could be used in practice. In the poor matching scenario of that paper (1st poor scenario of this paper), the Belin-Rubin procedure was unable to provide accurate estimates of error rates, but our theoretical adjustment procedure still worked well. This indicated that we either had to find an enhancement to the Belin-Rubin procedure or develop methods that used more of the available data.

A crucial practical assumption for the work of this paper is that the analyst be able to separate out a low-error-rate set of pairs on which to do matching. Although neither the Belin-Rubin procedure (1995) nor an alternative procedure of Winkler (1994) that requires an ad hoc intervention could be used to estimate error rates, we believe it is possible for an experienced matcher to pick out a low-error-rate set of pairs even in the 2nd poor scenario. A naive matcher might not easily do so. Until now, an analysis based on the 2nd poor scenario would not have seemed even remotely sensible. As we will see in the fourth section, something of value can be done.

Quantitative Scenarios

Having specified the above linkage situations, we used SAS to generate ordinary least squares data under the model $\mathbf{Y} = \mathbf{6}\ \mathbf{X} + \epsilon$. The $\mathbf{X}$ values were chosen to be uniformly distributed between 1 and 101, and the error terms ϵ are normal and homoscedastic with variances 13,000, 36,000, and 125,000, respectively--all such that the regressions of $\mathbf{Y}$ on $\mathbf{X}$ have an R^2 value in the true matched population of 70 percent, 47 percent, and 20 percent, respectively. Matching with quantitative data is difficult because, for each record in one file, there are hundreds of records having quantitative values that are close to the record that is a true match. Additionally, to make modelling and analysis much more difficult in the high overlap scenario, we used all false matches and only 5 percent of the true matches; in the medium overlap scenario, we used all false matches and only 25 percent of true matches.

See figure 2a for the actual true regression relationship and related scatter plot, as they would appear if there were no matching errors. Note all of the mismatches are plotted, but only 5 percent of the true matches are used. This has been done to keep the true matches from dominating the results so much that no movement can be seen. Second, in this figure and the remaining ones, the true regression line is always given for reference. Finally, the true population slope or **beta** coefficient (at 5.85) and the R^2 value (at 43 percent) are provided for the data being displayed.

■ Simulation Results

We begin by presenting graphs and results of the recursive process for the 2nd poor scenario, R^2 value of 47 percent, and the high overlap situation. These results best illustrate the procedures of this paper. Later in the paper, we summarize results over all R^2-situations and all overlaps. The regression results for two cycles are given in the first two subsections. In the third section, we present results that help explain why such a dramatic improvement can occur.

First Cycle Results

- **Regression after Initial RL ⇒RA Step.**--In figure 2b, we are looking at the regression on the actual observed links--not what should have happened in a perfect world but what did happen in a very imperfect one. Unsurprisingly, we see only a weak regression relationship between **Y** and **X**. The observed slope or **beta** coefficient differs greatly from its true value (2.47 v. 5.85). The fit measure is similarly affected, falling to 7 percent from 43 percent.

- **Regression after Combined RL⇒RA⇒EI ⇒RA Step.**--Figure 2c completes our display of the first cycle of our recursive process. Here, we have edited the data in the plot displayed as follows. First, using just the 99 cases with a match weight of 3.00+, an attempt was made to improve the poor results given in figure 2b. Using this provisional fit, predicted values were obtained for all the matched cases; then, outliers with residuals of 460 or more were removed and the regression refit on the remaining pairs. This new equation was essentially **Y** = 4.5**X** + ϵ with a variance of 40,000. Using our earlier approach (Scheuren and Winkler, 1993), a further adjustment was made in the **beta** coefficient from 4.5 to 5.4. If a pair of matched records yielded an outlier, then, predicted values using the equation **Y** = 5.4**X** were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value.

Second Cycle Results

- **True Regression (for reference).**--Figure 3a displays a scatter plot of **X** and **Y** as they would appear if they could be true matches based on a second **RL** step. The second **RL** step employed the predicted **Y** values as determined above; hence, it had more information on which to base a linkage. This meant that a different group of linked records was available after the second **RL** step. Since a considerably better link was obtained, there were fewer false matches; hence, our sample of all false matches and 5 percent of the true matches dropped from 1,104 in figures 2a through 2c to 650 for figures 3a through 3c. In this second iteration, the true slope or **beta** coefficient and the R^2 values remained though, virtually identical for the slope (5.85 v. 5.91) and fit (43 percent v. 48 percent).

- **Regression after Second RL ⇒RA Step.**--In figure 3b, we see a considerable improvement in the relationship between **Y** and **X** using the actual observed links after the second **RL** step. The slope has risen from 2.47 initially to 4.75 here, still too small but much improved. The fit has been similarly affected, rising from 7 percent to 33 percent.

- **Regression after Combined RL⇒RA⇒EI⇒RA Step.**--Figure 3c completes the display of the second cycle of our recursive process. Here, we have edited the data as follows. Using this fit, another set of predicted values was obtained for all the matched cases. This new equation was essentially **Y** = 5.5**X** + ϵ with a variance of about 35,000. If a pair of matched records yields an outlier, then predicted values using the equation **Y** = 5.5**X** were imputed. If a pair does not yield an outlier, then the observed value was used as the predicted value. The plot in figure 3c gives the adjusted values, which have slope 5.26 and fit 47 percent, which improves over first cycle results.

Further Results

We do not show results for the medium- and low-overlap situations because the matching was somewhat easier. The reason it was easier is that there were significantly fewer false-match candidates, and we could more easily separate true matches from false matches. For the high R^2 scenarios, the modelling and matching were more straightforward than they were for the medium R^2 scenario in the section with the second cycle results. For the low R^2 scenario, we were unable to distinguish true matches from false matches. This is understandable because there are so many outliers associated with the true matches. We can no longer assume that a moderately higher percentage of outliers in the regression modelling is due to false matches.

Future Study

In principle, the recursive process of matching and modelling could have continued. Indeed, while we did not show it in this paper, the **beta** coefficient of our example did not change much during a third matching pass.

At first, it would seem that we should be happy with the results. They take a seemingly hopeless situation and give us a fairly sensible answer. A closer examination, though, shows a number of places where the approach taken is weaker than it needs to be or simply unfinished.

We have looked at a simple regression of one variable from one file with another variable from another. What happens when this is generalized to the multiple regression case? We are working on this now, and sensible results are starting to emerge, which have given us insight into where further research is required. There is also the case of multivariate regression. Here, the problem is harder and will be more of a challenge.

First, to make use of multivariate data, we need to have better ways of modelling them than the simple method of this paper. The likely best methods will be variants and extensions of Little and Rubin (1987, Chapters 6 and 8) in which predicted multivariate data have important correlations accounted for. If we take two variables from one file and two from another, then we can make use of the fact that the two variables taken from one file have the correct two-variable distribution but may be falsely matched.

Second, we have not yet developed effective ways of utilizing the predicted and unpredicted quantitative data. Simple multivariate extensions of the univariate comparison of **Y** values in this paper do not seem to work.

■ 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**, pp. 694-707.

Fellegi, I. and Holt, T. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, **71**, pp. 17-35.

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

Jabine, T.B. and Scheuren, F. (1986), "Record Linkages for Statistical Purposes: Methodological Issues," *Journal of Official Statistics*, **2**, pp. 255-277.

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **89**, pp. 414-420.

Little, R.J.A. and Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, J. Wiley: New York.

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

Newcombe, H.; Fair, M.; and Lalonde, P., (1992), "The Use of Names for Linking Personal Records," *Journal of the American Statistical Association*, **87**, pp. 1193-1208.

Oh, H. L. and Scheuren, F. (1975), "Fiddling Around with Mismatches and Nonmatches," *American Statistical Association Proceedings, Social Statistics Section.*

Scheuren, F. and Winkler, W. E. (1993), "Regression Analysis of Data Files that are Computer Matched," *Survey Methodology*, **19**, pp. 39-58.

Winkler, W. E. (1994), "Advanced Methods of Record Linkage," *American Statistical Association Proceedings, Survey Research Methods Section*, pp. 467-472.

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.

Winkler, W. E. and Scheuren, F. (1995), "Linking Data to Create Information," *Proceedings of Statistics Canada Symposium 95.*

Winkler, W. E. and Scheuren, F. (1996), "Recursive Analysis of Linked Data Files," *Proceedings of the 1996 Census Bureau Annual Research Conference.*

Figure 1a. 1st Poor Matching Scenario

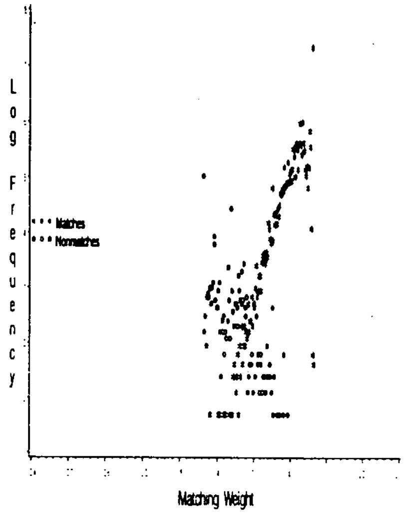

Figure 1b. 2nd Poor Matching Scenario

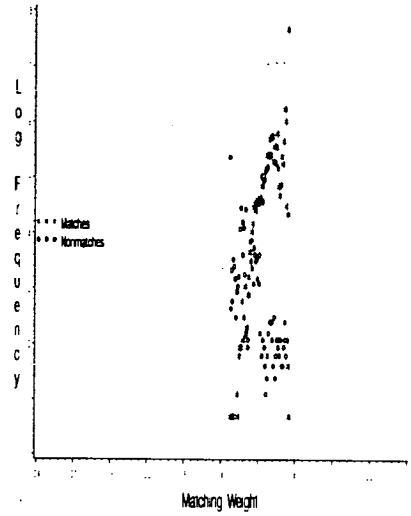

Figure 2a. 2nd Poor Scenario, 1st Pass
All False & 5% True Matches, True Data, HighOverlap
1104 Points, beta=5.85, R-square=0.43

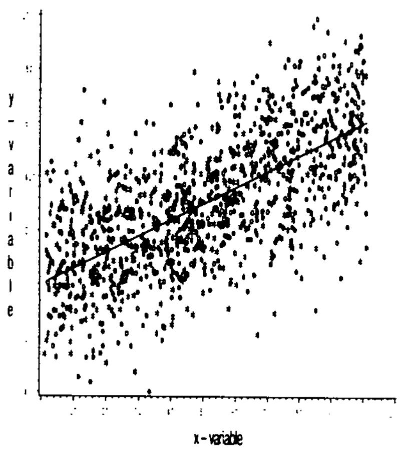

Figure 2b. 2nd Poor Scenario, 1st Pass
All False & 5% True Matches, Observed Data, HighOverlap
1104 Points, beta=2.47, R-square=0.07

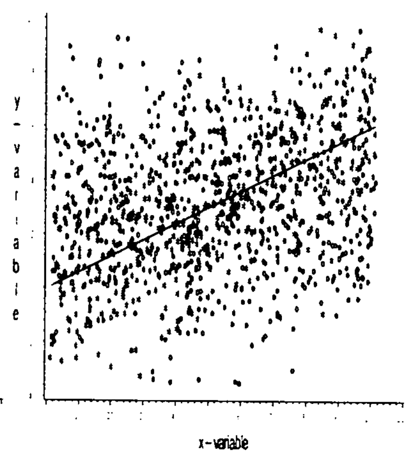

Figure 2c. 2nd Poor Scenario, 1st Pass
All False & 5% True Matches, Outlier-Adjusted Data
1104 Points, beta=4.78, R-square=0.40

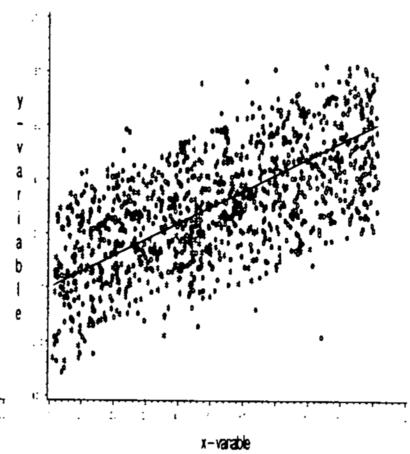

Figure 3a. 2nd Poor Scenario, 2nd Pass
All False & 5% True Matches, True Data, High Overlap
650 Points, beta=5.91, R-square=0.48

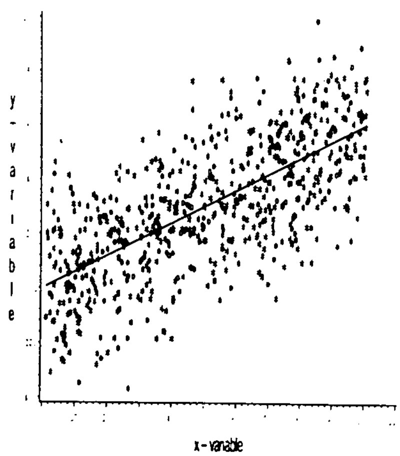

Figure 3b. 2nd Poor Scenario, 2nd Pass
All False & 5% True Matches, Observed Data, High Overlap
650 Points, beta=4.75, R-square=0.33

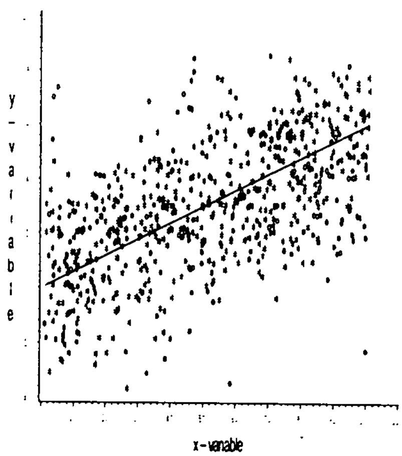

Figure 3c. 2nd Poor Scenario, 2nd Pass
All False & 5% True Matches, Outlier-Adjusted Data
650 Points, beta=5.26, R-square=0.47