

Rejoinder

John M. Abowd and Lars Vilhuber

November 23, 2004

We appreciate the time and effort that the discussants have spent providing us with concise and useful comments and suggestions. We are also grateful to both Alastair Hall and Torben Andersen, successive *JBES* editors, for inviting us to this forum and for supporting us in our endeavor. The discussants each came to the table with a different background, and we appreciate the wide range of concerns that they raised. The comments contain many specific questions regarding the data correction procedures and analyses presented in our article. All commentators offer suggestions concerning the scope, usefulness, and direction that research on the statistical use of administrative data should address. We would like to use this rejoinder to reply to these questions and to elaborate our own vision of useful directions

for this research.

1 Current status of implementation and research

To begin, we want to point out that the methods proposed in our article have already been implemented within the U.S. Census Bureau as part of the production system for the Quarterly Workforce Indicators (see <http://lehd.dsd.census.gov>). Since prototyping and analyzing the methods with California data, a number of states have been processed in much the same way. Each state's historical data provided a new challenge, since each had a different way of capturing the original UI wage record data in the first place, of archiving and extracting it, and of post-processing. The actual employment-generating process in each state also differs, and so does the end result. Although we have not repeated the quite extensive and costly analyses at all levels of aggregation that we discussed in our article, we have observed widely varying data characteristics. Some states historically captured only a small portion of the names on the wage records, others have specific ways of dealing with apparently "invalid" name-SSN combinations. All factors affect the match rates, and thus the

number of holes plugged, which vary widely. One key statistic, the reduction in single-quarter job spell interruptions, varies between 2 and 15 percent. California, with 11 percent, is above-average in this respect.

The great variety in match rates brings us directly to a comment by Van der Klauw. He points out that states differ in fundamental employment characteristics. Our editing process highlighted some of these characteristics, and Van der Klauw cites some other items, such as the actual definition of an employer within the administrative data. Our ongoing application of the methods described in our article to other states has clearly shown this to be a legitimate worry.

2 Eliminating bias, adding bias, or changing bias

Van der Klauw is also worried that our procedure potentially introduces new biases. His example is the differential impact of the procedure on wage earners with high and low earnings volatility, as well as the differences in data transmission methods used by firms. For instance, smaller firms are more likely to have submitted paper forms, and are thus more likely to be subject

to a data-keying entry error than are larger firms. Our correction procedure would find a higher match rate among smaller firms than large firms, and as a result, the relation between exit rates in smaller firms relative to exit rates in larger firms will have been altered. When expanding the analysis to multiple states, this differential data error is likely to be exacerbated. Not only will the conditional ratios within any single state be changed, but the change will differ by state, depending on the mix of small and large firms, the (unobserved) mix of submission methods, and other factors.

We should not forget that the original incidence of SSN coding errors is correlated with the same factors that influence match rates. In the example above, we find more matches among smaller firms, leading to larger adjustment of exit rates observed in smaller firms, because there are more coding errors in smaller firms to begin with. Nevertheless, since we lack a true validation dataset, it is very difficult to assess how much bias due to miscoding error is still present in the data.

Statistical methods that can take into account knowledge about matching errors, surveyed in the forthcoming chapter by (Moffitt & Ridder 2005) suggested by Van der Klauw, will become increasingly important. How to actually transmit the detailed knowledge, for instance from the matching pro-

cedures regularly used within the Census Bureau and the Bureau of Labor Statistics, is a different problem—one that is still in search of a solution.

3 Earnings model and the administrative data generating process

Several commentators discussed the earnings model we used to augment the matching procedure with economic pattern data. We model the earnings pattern generated by two alternate processes: continuous employment of a single person by the same employer in the presence of an identifier coding error, and the interrupted employment of a person with the same employer combined with the short-term employment of a different person at the same employer. We derive the earnings patterns associated with each process, and use the difference in earnings patterns as an additional element in the matching process.

Cohen, Fienberg, and Ravikumar (henceforth CFR) point out that a third earnings-pattern generating process could confound this binary choice. This third pattern would be the result of a model of “costly vacancies,” where employers faced with the absence of a single long-term employee for

a prolonged period of time immediately hire a replacement worker for the duration of the absence. As we will explain shortly, the model they propose is accounted for by our scheme, but their proposition highlights the need to distinguish the true earnings and employment generating process from the data generating process of the administrative agency assembling the data. The latter, in effect, filters and coarsens the true employment patterns.

To illustrate, consider a simplified version of the “costly vacancy” model, in which no lag arises between the separation of the long-term employee i and the hire of the replacement worker j (seamless employment of some worker filling the job). To facilitate the exposition, assume that separations occur on the last day of a month, and hiring occurs on the first day. Next, generate employment patterns such that the long-term worker i is not observed working in the second quarter of the year, but is observed working in the first and third quarter of the year, *i.e.*, a “hole” in quarter II, with $e(I) > 0$, $e(II) = 0$ and $e(III) > 0$ in the notation of Section 4.2. Nine possible employment patterns can be generated by processes that fit this pattern: worker i could have left on the last day of January, February, or March, and returned on any one of July 1, August 1, and September 1. Note however that each of the 9 true employment patterns has a different earnings pattern, which also differs

slightly from our section 4.2 since the separation process is now assumed to be discrete rather than continuous. In the notation of section 4.2,

$$E[e(I)] = 2/3, e(II) = 0, E[e(III)] = 2/3 \quad (1)$$

In the “costly vacancy” model, each of the nine true employment patterns corresponds to an employment pattern of a replacement worker. If worker i had quit in January and returned in August, then worker j would have worked from February to July. But these are the true employment patterns—not the data representation of those employment patterns. Because of the way the data are collected—the data generating process—only one of the nine possible replacement worker employment histories generates a single-quarter data record. For instance, if worker j worked from February to July, he would have generated a three-quarter sequence of wage records in the data. Only if he worked from April to June would he have generated a potential “plug.”

To further elaborate, recall that the expected earnings for “plugs” under H_1 are not computed over the population of replacement workers, but over the population of work histories that generate one-quarter sequences of wage

records at the same employer. Thus, although the replacement worker described above will be among the candidates, so will the worker k who worked only the month of July. Thus, with the assumptions above concerning hiring and separations, seven possible true employment patterns arise, and the expected earnings of all plugs, including the replacement worker’s history, are

$$E[e(II)|e(I) = 0 \text{ and } e(III) = 0] = 4/7 \quad (2)$$

which is smaller than the expected earnings of the “hole.”

This simple example serves to illustrate several points. First, if one is willing to make certain assumptions about the employment process, implications that are different from the ones we derive in Section 4.2 may arise. However, most alternate models will still generate the basic implication, that *on average*, plugs will have lower earnings than holes. Second, and more important, the data generating process that filters the true employment patterns has implications about the candidate records. In the simple example above, one-ninth of all replacement workers are filtered out and simply not eligible, because their employment patterns do not generate the same data pattern that a true miscode would generate. Finally, in the actual matching we do not use the exact (predicted) earnings. Rather, we use the popula-

tion decile into which earnings from hole and plug fall. We allow hole and plug earnings to be in adjacent deciles although if earnings fall into the same decile, the match weight is higher. Interestingly, whereas Winkler suggests increasing the precision of the earnings comparison, *i.e.*, moving away from using deciles, in order to increase the match rate, Van der Klauw suggests the opposite, fearing that a comparison that allows only for earnings within a narrow band would increase the non-match rate by penalizing workers with volatile earnings patterns. We will continue to investigate both avenues.

4 Improving knowledge on matching methods

The application of the methods proposed in our article is not unique to administrative records data. In fact, we used a commercial record linking product that is widely used in business environments. While we could easily publish the parameters used for the California data (and we will gladly provide these to readers with an interest in using them), our own experience in expanding the methods to other states has shown that weights and probability cutoffs need to be adjusted to the characteristics of the specific data

in order to obtain satisfactory results. The parameters used for California do not yield satisfactory results in other states. Methods do exist that use computing power rather than operator experience to find both appropriate weights and cutoffs, in particular using the EM algorithm. A number of our discussants are active in those fields (Winkler 2000b, Winkler 2000a, Yancey 2000, Yancey 2004)

A second factor related to improving our knowledge of matching methods relates to the choice of ancillary information to use in the linking exercise. As we noted in our article, legal constraints prevented expanding the information set used in the edit beyond the full history of the businesses and individuals present in the California data. In particular, the statute under which the data were received and edited (Title 15 U.S. Code) prohibited commingling these data with other Title 13 data at the Census Bureau. Such a commingling would have prevented returning the edited file to its original custodian, the California employment security office. There would have been substantial additional information had we used these off-limits Title 13 data. In particular, additional name and geography information could have been used both to improve the match rates and to assess the false match rate. Some research along these lines is clearly warranted; however, for statutory reasons it must

proceed outside the context of the Quarterly Workforce Indicator production system.

5 Moving away from one-to-one matches

CFR suggest an interesting extension or modification; namely, moving away from one-to-one matching towards explicit Bayes matching. The typical approach in matching applications, and which we have adopted here as well, is to use one-to-one matching. CFR propose a modification in the same spirit that imputation models were expanded in the last decades, by assigning multiple probable links to a single candidate employment history. In the case of missing data imputation, multiple imputation is achieved (in parametric models) by successive draws from the posterior predictive distribution of the missing values. In the case of record linking the actual application is non-trivial, since not only must the posterior distribution cover a draw from a distribution of candidate pairs, but must also account for the changing likelihood of no match being admissible. At present, we are not aware of an actual implementation of a multiple-match or Bayesian matching algorithm. Some authors, as pointed out to us by William Winkler in private correspondence,

have relaxed some aspects of the one-to-one match decision rule, by tracking some indicator of match quality from the matching process to the regression analysis (Lahiri & Larsen 2004, Scheuren & Winkler 1993, Scheuren & Winkler 1997). The effect of different possible matches is considered when the variable z_i used in the regression analysis $x_i = \beta z_i$ and stemming from a matching procedure is defined as (Lahiri & Larsen 2004)

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij} \end{cases}$$

The true but unobserved match is $z_i = y_i$, but the record linkage provides for each possible $y = j, j = 1, \dots, n$ a corresponding probability q_{ij} , $\sum_j q_{ij} = 1$. In this analysis, the one-to-one match is a special case, with $q_{ii} = 1$ for all i . As a result, the effect of mismatches, in particular the potential bias of using only one-to-one matches, can be evaluated in a regression framework.

In the Scheuren-Winkler framework, all possible matches for a given set of matching parameters are used in the analysis. This may not always be feasible, if the set of possible matches is large. A variation on this theme could be the introduction of a step into the matching procedure that perturbs the weights upon which the one-to-one rule is applied. N repetitions of this step

would result in as many sets of possibly different one-to-one matches, which could be analyzed using standard multiple imputation techniques (Rubin 1987). Such a method still remains speculation at this time, but we find this a most promising path to follow in future research.

The processing associated with a multiply-linked analysis file remains challenging. When the resulting data meet Rubin's (1987) conditions, the Bayesian statistics for any regular estimand are known. These techniques permit a direct implementation of CFR's suggestion and an assessment of the variability due to the resolution of the missing data via probabilistic record linking. Complicated statistical analyses, such as the ones performed at the Census Bureau to form the Quarterly Workforce Indicators (QWI), can be programmed to recognize multiply-imputed missing data. The QWIs, for example, recognize ten implicates for every data item that can be imputed. However, even the simplest of states takes 24 hours of parallel computing time (10 thread maximum) to complete with the assumption that the input employment histories, the items we edit in our article, are fixed. States like California take several weeks. Feasible multiple imputation of the record linkage in the wage record edit would have to be combined with the other missing data models in the QWI so that the processing of all implicates could

proceed in parallel.

6 Access to the data

One issue that multiple commentators have mentioned is access to the data. The data we used for this article are not public-use. In particular, the correction methods proposed in this paper required access to highly sensitive confidential information (SSN-laden employment histories). At the Census Bureau, the edited data files (and all other files used by the LEHD Program in producing the QWIs) are purged of all name and SSN information before being processed further within the statistical system, where they are still considered confidential. Only a few analysts have access to the SSN and name-laden files, even at the Census Bureau. Nevertheless, our access is not exclusive. Access can be granted based on approved research projects. The secondary analysis proposed by most commentators, studying the effect of the correction procedure on micro economic models of behavior and economic activity, can be done using the anonymized but still confidential micro data, since processing information from the correction is carried forward with the data. We refer interested researchers to the U.S. Census Research Data

Centers and the Center for Economic Studies web site (www.ces.census.gov), which contains information on applying to use confidential Census Bureau data.

7 Final remarks

We agree with all of the discussants that efforts to improve the quality of administrative record data for use as an input to general statistical analyses must go hand-in-hand with efforts to elaborate conventional analysis models. In particular, we think that the several discussions of the differences between the statistical processes generating the actual missing data and the assumptions of classical missing data models are instructive. We welcome the efforts to incorporate more appropriate underlying statistical assumptions into mainstream analyses. These efforts are clearly in their infancy and will become widespread as the benefits of exponentially improving computer speeds continue to manifest themselves. We are particularly optimistic about the integration of large scale data base management techniques with the statistical modeling of record linkage and linked data files. Once the software to facilitate using these data structures as inputs to statistical models is widely

disseminated, analyses like the one we have proposed here should become commonplace.

References

Lahiri, P. & Larsen, M. D. (2004). Regression analysis with linked data, *Preprint 04-9*, Iowa State University, Department of Statistics.

Moffitt, R. & Ridder, G. (2005). The econometrics of data combination, *in* J. J. Heckman & E. E. Leamer (eds), *Handbook of Econometrics*, Vol. 6, Elsevier.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons.

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

Scheuren, F. & Winkler, W. E. (1997). Regression analysis of data files that are computer matched - part II, *Survey Methodology* **23**: 157–165.

Winkler, W. E. (2000a). Frequency-based matching in fellegi-sunter model of record linkage, *Research Report Series RR00/06*, U.S. Census Bureau.

Winkler, W. E. (2000b). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage, *Research Report Series RR00/05*, U.S. Census Bureau.

Yancey, W. E. (2000). Frequency-dependent probability measures for record linkage, *Research Report Series RR00/07*, U.S. Census Bureau.

Yancey, W. E. (2004). Improving em algorithm estimates for record linkage parameters, *Research Report Series RRS2004/01*, U.S. Census Bureau.