

ADDITIONAL REFERENCES

- Abrevaya, J., Hausman, J., and Scott-Morton, F. (1998), "Identification and Estimation of Polynomial Errors-in-Variables Models," *Journal of Econometrics*, 87, 239–269.
- Bound, J., Brown, C., Duncan, G., and Rodgers, W. (1994), "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics*, 12, 345–368.
- Bound, J., Brown, C., and Mathiowetz, N. (2001), "Measurement Error in Survey Data," in *Handbook of Econometrics*, Vol. 5, eds. J. J. Heckman and E. Leamer, Amsterdam: North-Holland, Chap. 59.
- Bound, J., and Krueger, A. (1991), "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right," *Journal of Labor Economics*, 12, 1–24.
- Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, New York: Chapman & Hall.
- Carroll, R., and Wand, M. (1991), "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society*, 53, 573–585.
- Chen, X., Hong, H., and Tamer, E. (2003), "Measurement Error Models With Auxiliary Data," working paper; forthcoming in *Review of Economic Studies*.
- Chesher, A. (1991), "The Effect of Measurement Error," *Biometrika*, 78, 451–462.
- Frish, R. (1934), *Statistical Confluence Study*, Oslo: University Institute of Economics.
- Fuller, W. (1987), *Measurement Error Models*, New York: Wiley.
- Hausman, J., Ichimura, H., Newey, W., and Powell, J. (1991), "Measurement Errors in Polynomial Regression Models," *Journal of Econometrics*, 50, 271–295.
- Hong, H., and Tamer, E. (2003), "A Simple Estimator for Nonlinear Errors-in-Variables Models," *Journal of Econometrics*, 117, 1–19.
- Horowitz, J., and Manski, C. (1995), "Identification and Robustness With Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.
- Hsiao, C., and Wang, L. (1995), "A Simulation Based Semi-Parametric Estimation of Nonlinear Errors-in-Variables Models," working paper, University of Southern California.
- Lee, L. F., and Sepanski, J. H. (1995), "Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data," *Journal of the American Statistical Association*, 90, 130–140.
- Li, T. (2002), "Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models," *Journal of Econometrics*, 110, 1–26.
- Mahajan, A. (2002), "Identification and Estimation of Single Index Models With Misclassified Regressors," working paper, Stanford University.
- Newey, W. (2001), "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *Review of Economics and Statistics*, 83, 616–627.
- Schennach, S. M. (2004), "Estimation of Nonlinear Models With Measurement Error," *Econometrica*, 75, 33–75.
- Sepanski, J., and Carroll, R. (1993), "Semiparametric Quasi-Likelihood and Variance Estimation in Measurement Error Models," *Journal of Econometrics*, 58, 223–256.
- Taupin, M. L. (2001), "Semiparametric Estimation in the Nonlinear Structural Errors-in-Variables Model," *The Annals of Statistics*, 29, 66–93.

Comments

William W. COHEN

Center for Automated Learning & Discovery, Carnegie Mellon University, Pittsburgh, PA 15213
(wcohen@cs.cmu.edu)

Stephen E. FIENBERG

Department of Statistics, Center for Automated Learning & Discovery, Carnegie Mellon University, Pittsburgh, PA 15213 (fienberg@stat.cmu.edu)

Pradeep RAVIKUMAR

Center for Automated Learning & Discovery, Carnegie Mellon University, Pittsburgh, PA 15213 (pradeepr@cs.cmu.edu)

We congratulate the authors on an interesting and technically innovative article. The article illustrates the sophistication required to deal with statistical issues of data integration involving diverse government databases, especially in the economics domain; furthermore, it fits within a broader program of work on the creation of longitudinal economic datasets for secondary analysis, for which Abowd in particular has provided leadership (see, e.g., Abowd and Lane 2004; Abowd and Woodcock 2001, 2004). The article has also stimulated us to think about more general issues regarding probabilistic model-based methods for linkage—the primary topic of our joint research.

1. RECONSIDERING THE PROBLEM

Abowd and Vilhuber show that certain "flow" statistics connected with longitudinal studies are surprisingly sensitive to linkage errors. To illustrate their main point, consider the following simplified problem. Suppose that we are presented with a set of N tuples (a_i, b_i, c_i) , where a_i and c_i describe the employee i filling some position x in the first and third quarters of

2003, b_i describes the employee i' filling position x in the second quarter, and i may or may not be identical to i' . To resolve these ambiguities, we clean the data by applying a probabilistic linkage method (Fellegi and Sunter 1969; Winkler 2002) to the collection of (i, i') pairs, which links together some fraction p of the most-similar pairs.

Now consider using the linked data to count the number of "recalls," occasions in which an employee left her job and then returned after one quarter. The best estimate for this from the linked data will be $R = N(1 - p)$. Because true recalls are likely to be rare, however, a small number of linkage errors could easily lead to an estimate quite different from the true recall rate (proportionally speaking). A statistic that is even more sensitive to linkage errors (in absolute terms) is the number of "job changes," which would be estimated as $C = 2Np$. Many other

natural statistics not specifically measuring changes in employment status (e.g., average starting salary) will also be biased by linkage errors.

In the terminology of Abowd and Vilhuber, (a_i, c_i) is a “hole”—a possible gap in i 's employment history—and $b_{i'}$ is a “plug.” The main technical contribution of their article is a data-cleaning method that corrects longitudinally linked employment histories by looking for “plugs” to match an incorrect “hole.” The technique substantially improves the accuracy of estimates of flow variables (like “job changes” in our foregoing example). Another contribution is a detailed analysis of their data, which shows that errors in flow variables can be substantial (up to 15% in one case) even given relatively accurate initial longitudinal links. The analysis also suggests that some errors are exacerbated, rather than ameliorated, by aggregation.

The problem considered by Abowd and Vilhuber is much more complex than the simple case just outlined, and raises a number of technical issues regarding the linkage of “holes” to “plugs.” This problem requires formulation (either implicitly or explicitly) of two models, one model for employment histories in which the hole is erroneous and hence should be filled with a plug; and another for histories in which the hole is a true alternation between two different employees, i and i' .

The error model for holes is complicated by the fact that in Abowd and Vilhuber's data, job histories can be of any length. This raises the possibility of sequentially correlated errors, for instance, data-entry mistakes that are copied over several quarters. This issue is finessed by considering holes and plugs that span a single quarter. This restriction is justified by an observation that longer gaps are less common; also, long gaps due to systematic errors, such as a wrong social security number (SSNs) in the employer database, cannot be easily corrected by a matching method. However, although this is a reasonable restriction to consider, a limitation of the study is that there is no analysis of how longer holes might affect flow statistics.

The model proposed for true alternations between two employees is quite simple. It is assumed that the start and end times of the one-quarter jobs are uniformly distributed, which implies that the expected time worked in a real one-quarter job is only one month. Reflection suggests that this assumption is not suitable for at least some jobs (e.g., jobs that are undesirable to leave unfilled), so ideally this model should be grounded empirically, perhaps by survey data for a similar population. It is unclear how sensitive the model is to this assumption, however; indeed, other evidence suggests that the proposed edits are still quite conservative.

We have a few other comments and suggestions regarding this specific application:

- Some data sources are thought to be more accurate than others. How do we build that into the matching and correction process?
- Many record linkage methods assume a one-to-one mapping between the objects to be linked (in this case, holes and plugs). Errors resulting from fraudulent use of a single SSN by multiple people violate this assumption and this problem suggests the need to develop a more elaborate model, for example, a Bayesian mixture of models of different complexities, as we will describe in the next section.

- The discussion in section 4.4 suggests that Abowd and Vilhuber are actually missing many edits. Their approach seems to follow the “do no harm” rule, which may or may not be appropriate.
- The related literature on statistical disclosure limitation tries to see how much an intruder can infer from masked data (cf. Fienberg, Makov, and Sanil 1998; Labert 1993). It is interesting to ask the extent to which the corruption of the data in the present context provides “protection” against such an intruder.

In summary, Abowd and Vilhuber provide an excellent case study of how linkage and statistical analysis interact, and of how understanding that interaction can be used to drive development of new and better linkage methods.

2. A MORE GENERAL LESSON

The interaction between data cleaning and linkage explored in this article suggests to us that there may be other situations in which data cleaning cannot be performed well without considering what sort of statistical analyses will be applied to the cleaned data. The most general lesson of Abowd and Vilhuber's article may be that data cleaning and analysis can be, and should be, more tightly coupled than is usually done in practice.

One interesting coupling is suggested by the work of Winkler and Scheuren (1996), who described a method called “analytic linkage” for jointly finding links between two sets of records and a regression model on the linked data. In their work the goal was to use the results of regression to improve the quality of linkage; however, one can imagine variants of the method that optimize the linkage decisions to improve the regression model.

To carry the coupling one step further, we note that in almost all uses of data cleaning, the end result is a *single* “clean” database, against which *all* subsequent analysis is performed. From a Bayesian standpoint, data cleaning would ideally result in a posterior *distribution* over possible “clean” databases that are consistent with the observed raw data, rather than a single database. Subsequent statistical analysis would then be performed via queries to this posterior distribution.

Let us return to our very simple example. Suppose that the result of linkage was a distribution D over the fraction p of pairs (i, i') that should be linked. This distribution could be used to compute probability intervals for estimates of the number of recalls, R , or job changes, C , rather than simply point estimates.

For more realistic problems, this sort of integration of data cleaning and analysis would require more general methods for representing and reasoning about uncertainty over the data cleaning process. This is a difficult task, particularly for datasets describing complex relationships between objects, as is the case in the longitudinal data considered by Abowd and Vilhuber. However, there are many reasons to believe that the approach could be practical in the not-too-distant future. Although a “clean” relational database is still by far the most efficient way to store large amounts of information, there has been steady and continual progress in representing and reasoning with uncertainty about data (e.g., Buntine 1994; Domingos and Richardson 2004; Friedman, Getoor, Koller, and Pfeffer 1999; Heckerman, Chickering, Meek, Rounthwaite, and Kadie 2000).

Representing uncertainty about how records should be linked introduces additional complications, because this implies uncertainty about the number of objects in the world. However, Pasula, Marthi, Milch, Russell, and Shpitser (2002) recently described a Markov chain Monte Carlo (MCMC) technique that allows sampling from a posterior over RPN structures associated with different numbers of objects. The method that they describe has not been applied to databases containing more than a few hundred objects; however, in earlier work, Cohen, Kantz, and McAllester (2000) described an $O(N \log N)$ search technique for finding a “clean” database that is locally optimal with respect to posterior probability, and the data structures that they described could be adapted to MCMC sampling as well. In fact, the Abowd–Vilhuber approach of providing a single clean database might be viewed as a maximum a posteriori (MAP) approximation to the full Bayesian approach. What we might want to do is have multiple draws from the posterior distribution, in the spirit of “multiple imputation,” which again relates to issues concerning protection against statistical disclosure (cf. Raghunathan, Reiter, and Rubin 2003).

To be practical for large-scale problems, these sorts of representations of data cleaning uncertainty will have to be combined with conventional schemes for storing data whose attributes and identity are not in question. We note that in Abowd and Vilhuber’s dataset, the vast majority of the longitudinal links are based on apparently uncorrupted SSNs, and only 800,000 of 96,000,000 records are modified by their data cleaning method.

This Bayesian framework for representing the output of data cleaning clearly requires refinement and the imposition of many approximations to make it practical for large datasets. Because of this, it is still essential to obtain a good understanding of which data cleaning decisions most effect the analyses that will be performed on the data and the effect of those decisions on the bias and accuracy of derived statistics. Abowd and Vilhuber’s article is an excellent step toward achieving this understanding.

Rejoinder

John M. ABOWD

Edmund Ezra Day Professor of Industrial and Labor Relations, School of Industrial and Labor Relations, Cornell University, Ithaca, NY 14850 (john.abowd@cornell.edu)

Lars VILHUBER

Senior Research associate, CISER (Cornell Institute for Social and Economic Research), Cornell University, Ithaca, NY 14850 (lars.vilhuber@cornell.edu)

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

ADDITIONAL REFERENCES

- Abowd, J. M., and Lane, J. (2004), “New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers,” in *Privacy in Statistical Databases*, eds. J. Domingo-Ferrer and V. Torra, New York: Springer-Verlag, pp. 282–289.
- Abowd, J. M., and Woodcock, S. D. (2001), “Disclosure Limitation in Longitudinal Linked Data,” in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, Amsterdam: North-Holland, pp. 215–277.
- (2004), “Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data,” in *Privacy in Statistical Databases*, eds. J. Domingo-Ferrer and V. Torra, New York: Springer-Verlag, pp. 290–297.
- Buntine, W. (1994), “Operations for Learning With Graphical Models,” *Journal of Artificial Intelligence Research*, 2, 159–225.
- Cohen, W. W., Kautz, H., and McAllester, D. (2000), “Hardening Soft Information Sources,” in *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining*, pp. 255–259.
- Domingos, P., and Richardson, M. (2004), “Markov Logic: A Unifying Framework for Statistical Relational Learning,” in *Proceedings of SRL2004: Statistical Relational Learning and Its Connections to Other Fields*, <http://www.cs.umd.edu/projects/srl2004/>.
- Fellegi, I., and Sunter, A. (1969), “A Theory for Record Linkage,” *Journal of the American Statistical Society*, 64, 1183–1210.
- Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1998), “A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data,” *Journal of Official Statistics*, 13, 75–89.
- Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. (1999), “Learning Probabilistic Relational Models,” in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, pp. 1300–1309.
- Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. (2000), “Dependency Networks for Inference, Collaborative Filtering, and Data Visualization,” *Journal of Artificial Intelligence Research*, 1, 49–75.
- Labert, D. (1993), “Measures of Disclosure Risk and Harm,” *Journal of Official Statistics*, 9, 313–331.
- Pasula, H., Marthi, B., Milch, B., Russell, S., and Shpitser, I. (2002), “Identity Uncertainty and Citation Matching,” in *Advances in Neural Processing Systems 15*, Vancouver, British Columbia: MIT Press, pp. 1425–1432.
- Raghunathan, T. E., Reiter, J., and Rubin, D. B. (2003), “Multiple Imputation for Statistical Disclosure Limitation,” *Journal of Official Statistics*, 19, 1–16.
- Winkler, W. E. (2002), “Methods for Record Linkage and Bayesian Networks,” research report, Statistical Research Division, U.S. Bureau of the Census.
- Winkler, W., and Scheuren, F. (1996), “Recursive Analysis of Linked Data Files,” in *Proceedings of the 1996 Census Bureau Annual Research Conference*, pp. 920–935.

contain many specific questions regarding the data-correction procedures and analyses presented in our article. All commen-