Discussion

*Stephen E. Fienberg*¹

1. Introduction

In countries around the world, support for traditional population censuses and large-scale sample surveys has eroded. The issues raised most often are cost, public distrust (privacy), and timeliness of data releases, and these are weighed against accuracy and alternative sources of population information (see e.g., [Fienberg and Prewitt 2010](#)). Official statisticians continue to explore alternatives and supplements to regular censal canvasses of the population and large-scale surveys for various purposes. Although various forms of population registers have replaced traditional head counts in a number of countries, new realities have led statisticians to question their accuracy and completeness. In some instances, political concerns present a problematic overlay to all of these developments.

So what has changed? First response rates in large-scale probability surveys have steadily declined, especially as individuals have become more mobile and in some senses less accessible to traditional methodologies, such as door-to-door person interviewing, telephone interviewing, and mail surveys. Second, family structures have also changed in ways that make households less apt as a unit of measurement. Third, migration, especially migration involving illegal immigrants, now poses a vexing problem in countries that have long professed not to be afflicted with such complexities, especially in the European Union. Finally, politicians have become increasingly suspicious of the costs of official statistics methodologies, believing that equivalent information could have been obtained from the internet at a fraction of the cost. Understanding the provenance and quality of the data has been the hallmark of official statistics as a field, but explaining the importance of this in the political climate of budget cuts and austerity is often difficult.

What has been the response to this situation by the official statistics community? Three mechanisms stand out:

- (1) The use of administrative records, not solely but in combination with one another and more traditional census methods.
- (2) The use of post-enumeration surveys to replace costly nonresponse follow up to such mechanisms involving telephone and personal interviewing,
- (3) The use of online census forms.

¹ Maurice Falk University Professor of Statistics and Social Science at the Department of Statistics, the Heinz College, and the Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A. Email: fienberg@stat.cmu.edu

Acknowledgment: The preparation of this discussion was partially supported by the U.S. National Science Foundation and the U.S. Census Bureau through NSF grant SES1130706 to the Department of Statistics, Carnegie Mellon University.

There are many hurdles, both methodological and practical, that need to be overcome before these approaches can yield the high-quality data that many government statistical agencies are recognized as producing. A major methodological issue is how to combine the potentially noisy information from multiple sources. In the United States, as we approach our next decennial census in 2020 the focus is on (1) and (3) with little focus on combining multiple sources for actual population estimation (see [Eddy and Fienberg 2014](#)), largely because of past political battles over post-enumeration surveys and sampling; for example, see [Anderson and Fienberg \(2001\)](#), [Anderson et al. \(2000\)](#), and [Brown et al. \(1999\)](#). In other countries the mix is different. Canada, for example, has pioneered the use of online forms, achieving a 54% online response, although set in a traditional mail-out mail-back context with mailed postcards replacing the initial mailing of census forms. But Canada, too, is exploring the use of administrative records going forward, see the overview in [Hamel and Béland \(2013\)](#). And in Israel, two separate surveys are used to capture the effects of both over- and underenumeration; see [Nirel and Glickman \(2009\)](#) for a detailed description.

2. The Articles in this Issue

The present issue of the *Journal of Official Statistics* includes contributions from statisticians in multiple countries, describing their current and recent efforts to reshape their census and population-reporting methodologies as well as more technical methodological contributions. Below I comment on some of the features of their efforts and both the novelties and the commonalities.

The primary methodological features that tie many of these articles together are (a) record-linkage methodologies, (b) the use of capture-recapture (dual systems) models and their multidimensional analogues when three or more record systems are involved, and (c) other related log-linear model methods. What I found especially gratifying in reviewing these contributions was the extent to which virtually every article linked in some way to my past and current research interests and activities on these topics. I mention a number of these links to my work as well as that of others in what follows.

2.1. Gerritse, van der Heijden, and Bakker

These authors explore the departure from independence in the dual-systems approach to two registers by modeling different values of the dependence. They do this by looking at a range of “offset” values and by considering the role of covariates, albeit with missing values. They use data from Dutch registers and introduce a stratification by country of nationality that allows for some common modeling of dependencies. They briefly discuss extensions to multiple registers based on standard multiple-recapture approaches, first laid out in [Fienberg \(1972\)](#) and [Bishop et al. \(1975\)](#), and something they refer to as the multiplier method, which presumes independence of sources as well as the availability of some joint information between them. The latter does not generalize in any useful way for most census problems I know. An alternative approach to this work worthy of note is that presented by [Kurtz \(2014\)](#) who uses a localized approach to adjust for interactions with covariates.

2.2. Zhang

The focus in this article is on adjusting census results for both underenumeration (omissions) using coverage survey data and overenumeration (erroneous enumerations). The U.S. Census Bureau approach is to do this by directly adjusting for overenumeration for a sample of census blocks prior to the use of dual-systems methods for omissions; for example, see [Hogan \(1993\)](#). There are two strong assumptions involved: (1) that all erroneous enumerations can be identified in the census data for the sample blocks, and (2) that there are no erroneous enumerations in the coverage survey for those blocks. The Israeli Central Bureau of Statistics approach uses two separate surveys, one for omissions and another to remove erroneous enumerations; for example, see [Nirel and Glickman \(2009\)](#). Through a targeted class of models and assumptions regarding the two surveys they estimate the population totals, adjusting for both kinds of errors. Zhang adopts a somewhat different strategy involving two coverage surveys, or a coverage survey and an administrative list, and he directly models the resulting three-dimensional cross classification using log-linear models, and a variant on them also involving marginal log-linear models. By assuming that the second coverage survey only has underenumerations (and no matching error), he identifies a model that allows for estimation of the relevant population totals. He approaches estimation through a form of moment estimation. I would have preferred a maximum-likelihood approach as that would have directly allowed for consistency and asymptotic variances.

The earlier literature on multiple-system estimation (with three or more lists) includes many ways to estimate population totals using log-linear and related models, which allow for both dependence among lists and individual-level heterogeneity, but they all make the unrealistic assumption of no overenumeration (or at least that overenumeration can be corrected without error). See [Darroch et al. \(1993\)](#), [Fienberg et al. \(1999\)](#), [Manrique-Vallier and Fienberg \(2008\)](#), and the [International Working Group for Disease Monitoring and Forecasting \(1995a, 1995b\)](#) for some examples. Zhang's approach, moved to a setting with at least four lists, might allow for a weakening of this assumption.

2.3. Di Consiglio and Tuoto

There are two possible ways of dealing with the uncertainty associated with record linkage for various lists, especially in a censal context. We can actually examine the error process in linkage and propagate it through into the population estimation process (see below for more details) or we can develop simplified models involving homogeneous probabilities of true and false links and then either estimate those probabilities using auxiliary data or explore the sensitivity of the assumption of no linkage error by considering a range of possible values for the parameters. [Ding and Fienberg \(1994, 1996\)](#) took the latter approach. In this article, Consiglio and Tuoto take up the Ding and Fienberg approach for two lists (1994) and extend it to allow for a slightly more complex possibility of true and false probabilities of links. Adapting their approach to multiple lists and exploring how well the approach captures the effect of heterogeneous errors of linkage associated with basic record linkage approaches are natural next steps. For related literature borrowing from methods used in the context of tag loss in animal population studies, see [Seber et al. \(2000\)](#) for the two-list case and [Lee et al. \(2001\)](#) for an extension to multiple lists.

2.4. *Yildiz and Smith*

These authors address a related but different problem involving the combination of administrative records data than those looking to create and analyze individual-level data, which are ultimately aggregated into the form of a contingency table. Rather, they consider adjusting one source of aggregate administrative data in the form of a cross classification to conform with “more accurate” values from an auxiliary dataset. The methodology is once again based on log-linear models, where the use of “offsets” allows the preservation of interaction structures from the original aggregate administrative data. They implement the approach using maximum-likelihood estimation based on iterative proportional fitting as described in [Bishop et al. \(1975\)](#).

What if there are multiple updated data sources for overlapping variables? This turns out to be a far more complex problem, and requiring consistency of overlapping margins for the updated sources is not sufficient to ensure the existence of a contingency table with the original interaction structure and adjusted margins to accord with the updated sources. For a related discussion of nonexistence in the context of privacy protection, see, for example [Yang et al. \(2012\)](#).

2.5. *Chipperfield and Chambers*

As I and others have noted (e.g., see [Bilenko et al. 2003](#)), most methods for record linkage, and especially those used in the context of official statistics, are based on variants of the now classic methodology of [Fellegi and Sunter \(1969\)](#). The book by [Herzog et al. \(2007\)](#) provides a useful update including suggested links to multiple-recapture methodology. It also describes the use of the FS methodology for duplicate detection as well as record linkage, and in practice these are done as separate activities rather than jointly.

This article is set in this FS record-linkage tradition and focuses on the use of bootstrap methodology to propagate the uncertainty from the matching process of two files into the subsequent statistical analyses. It extends the authors’ earlier work in which the results of the record linkage are used for regression modeling to the setting of linked categorical data, with a special focus on analyses based on log-linear or logistic models. They assume no duplicates within the two files to be used for linkage, and also that fields within records used for linkages either match or not. The latter is an impractical restriction for files involving alphanumerics (such as names and places) or even dates, and this can clearly be relaxed. They begin with the assumption that the files to be linked are of equal size and assume 1-1 matching, a case known to produce high probabilities of linkage, and then proceed to the more practical situation where some files will not match and the files will be of differing lengths. They demonstrate the methodology in a simulation and then in the context of the 2011 census of Australian Population and Housing, comparing it with two alternative procedures with interesting summary results. When it comes to the actual implementation of record-linkage methods, the devil is in the detail, and we clearly need more details to judge the utility of the methods presented here.

There is a hint buried in this article regarding the potential representation of linkages in terms of network structures. This is the approach adopted by [Sadinle \(2014\)](#) in the context of a duplicate detection task, and extended in [Sadinle \(2015\)](#) to joint duplicate detection and record linkage involving the possibility of multiple files. See also the related approach

in [Steorts et al. \(2014a, 2014b\)](#). Getting these methods to scale so that they work in a censal context remains a challenge.

2.6. *Blackwell, Charlesworth, and Rogers*

These authors describe in considerable detail the way in which the U.K. Office of National Statistics used administrative records and record linkage in a sample within 58 carefully selected local areas corresponding to formal administrative authorities. They describe aspects of both over- and undercoverage. There is a serious technical issue here for anyone familiar with the actual uses of Fellegi-Sunter methodology at the scale of census files, even for restricted geographical areas. As the size of lists to be matched grows, the number of possible matches increases as the product of their sizes. The only way around this is through blocking, but this rests on the assumption that the blocking variable is error free, something that is rarely the case. It will not come as a surprise that the effort was complex. At any rate, the authors do not provide any specific details on the implementation. Given my understanding of the approach used in the U.S. for coverage evaluation, the actual implementation rarely resembles what one finds in research papers and the published literature. Even without the details, their presentation here should serve as a sobering reminder that we require much more than elegant theory and methods to produce official statistics of value.

2.7. *Bryant and Graham*

These authors have as their goal replacing the traditional census in New Zealand with an administrative census, perhaps supplemented by a coverage survey, to (i) generate population estimates, and (ii) assemble individual-level socioeconomic data for different data sources. Their key methodological focus is on the use of Bayesian hierarchical models, something I approve of heartily (cf. [Fienberg 2011](#) and [Fienberg et al. 1999](#)). The unique feature of the present article, in the context of this special issue on coverage measurement, is its reliance on a demographic accounting framework. By working with multiple data sources and using the hierarchical structure to “borrow strength” across sources via the hierarchical structure, they explain how to properly account for random variation in the demographic series and in the measurement of these series and how to propagate this uncertainty into final population estimates. Accuracy of immigration and emigration numbers is crucial. However, what remains unclear from their article is how they will ultimately combine multiple sources and a coverage survey to provide individual-level data. Nor do they discuss how they propose to account for differences in the timing of data sources and target populations.

2.8. *Burger, van Delden, and Scholtus*

These authors present an approach to estimate the bias and variance resulting from classification errors in a business register. Their approach involves a homogeneous misclassification model, based on a two-parameter transition matrix, and bootstrap resampling methodology. They then explore the sensitivity of bias and variance specification of the transition matrix via a simulation study. Their study focuses on a single

data source, and what would be of special interest in the context of the problems explored in the other articles in this issue is how the kinds of misclassification errors would affect record linkage of multiple sources and statistical calculations on the resulting merged data.

3. Conclusion

Some areas that still require research to bring the theory and methods described above into actual practice in official statistics settings are as follows:

- *Record linkage methods from three or more files.* As I note above, virtually all of the record-linkage methodologies in use today deal only with pairs of files, for example, from a census and a coverage survey. But as we move to the use of administrative records, we need to think in terms of multiple files, and it does not suffice to match each separately into some common file. Our census-related project at Carnegie Mellon has produced a number of efforts in this direction, see [Sadinle \(2014, 2015\)](#), [Sadinle et al. \(2011\)](#), [Sadinle and Fienberg \(2013\)](#), [Steorts et al. \(2014a, 2014b\)](#) and [Ventura et al. \(2014\)](#). See also [Fienberg and Manrique \(2009\)](#). However, making these methods suitable for “industrial strength” applications such as those involved in census production systems will take considerable time and effort.
- *Combining duplicate detection and record linkage.* Many agencies and researchers treat duplicate detection within files and record linkage between files as separate activities. They are not and the very fact that the Fellegi-Sunter is used for both should get us thinking about ways to approach the tasks in a unified way. [Sadinle \(2014, 2015\)](#) and [Steorts et al. \(2014a, 2014b\)](#) adopt such unifying approaches, and the Bayesian framework that they use is especially suitable for this. The type of combining raises even greater issues of scalability than methods for record linkage alone.
- *Propagating duplicate detection and record linkage error into subsequent calculations.* Researchers have recognized this problem for decades, but it has received renewed attention over the past decade. The article by Chipperfield and Chambers described above and the related ones by these authors take a frequentist approach and develop bootstrap procedures to accompany them. [Sadinle \(2014, 2015\)](#) and [Steorts et al. \(2014a, 2014b\)](#), following earlier work of [Belin and Rubin \(1995\)](#), [Larsen and Rubin \(2001\)](#), and [Tancredi and Liseo \(2011\)](#), focus on Bayesian approaches. Implementing any of these methods at scale is complex.
- *Measuring both EEs and omissions.* Much of the discussion surrounding the use of coverage surveys and capture-recapture methods for population estimation focuses on the problem of omissions and census undercount. But erroneous enumerations often play as big a role, and will take on increasing importance as administrative record systems often contain many records that are out of scope. The approach of [Nirel and Glickman \(2009\)](#) suggests one way of proceeding, but with attendant assumptions, and that of Zhang suggests that simplifying assumptions and models will be useful in other contexts. Much more work needs to be done here.
- *Putting it all together with statistical models and methods.* The propagation of error issue discussed above is but one of the multiple aspects of a unified methodology for census taking. We have no such overarching framework and methodology today.

- *Moving from new methods to statistical practice.* For the new approaches to census taking to work in practice, agencies will need to be able to address essentially all of the problems listed in the bullets above, and combine them with effective use of online forms and real-time editing. In the U.S, Canada, and the U.K., as well as in other countries, censuses have combined the study of housing, households, and individuals. Most of my discussion has focused on the latter, but placing individuals in households also requires special attention, as does the linkage of households across multiple files.

While most of the articles in this special issue of the *Journal of Official Statistics* focus primarily on coverage issues in censuses and administrative record systems, many of the concerns and methodologies are relevant to the future of large-scale sample surveys as well. The changes demanded by government officials and the public are great and likely to continue increasing. The good news in the work reported on here and in other recent methodological developments is that we may be entering a new era for collaboration between statistical offices and academic statisticians.

4. References

- Anderson, M.J., B.O. Daponte, S.E. Fienberg, J.B. Kadane, B.D. Spencer, and D. Steffey. 2000. "Sample-Based Adjustment of the 2000 Census – A Balanced Perspective." *Jurimetrics* 40: 341–356.
- Anderson, M.J. and S.E. Fienberg. 2001. *Who Counts? The Politics of Census-Taking in Contemporary America*. Rev. ed. New York: Russell Sage Foundation.
- Belin, T.R. and D.B. Rubin. 1995. "A Method for Calibrating False-Match Rates in Record Linkage." *Journal of the American Statistical Association* 90: 694–707. Doi: <http://dx.doi.org/10.1080/01621459.1995.10476563>.
- Bilenko, M., R. Mooney, W. Cohen, P. Ravikumar, and S.E. Fienberg. 2003. "Adaptive Name Matching in Information Integration." *IEEE Intelligent Systems* 5: 16–23. Doi: <http://doi.ieeecomputersociety.org/10.1109/MIS.2003.1234765>.
- Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press. Reprint, New York: Springer, 2007.
- Brown, L.D., M.L. Eaton, D.A. Freedman, S.P. Klein, R.A. Olshen, K.W. Wachter, M.T. Wells, and D. Ylvisaker. 1999. "Statistical Controversies in Census 2000." *Jurimetrics* 39: 347–376.
- Darroch, J., S.E. Fienberg, G. Glonek, and B. Junker. 1993. "A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability." *Journal of the American Statistical Association* 88: 1137–1148. Doi: <http://dx.doi.org/10.1080/01621459.1993.10476387>.
- Ding, Y. and S.E. Fienberg. 1994. "Dual System Estimation of Census Undercount in the Presence of Matching Error." *Survey Methodology* 20: 149–158.
- Ding, Y. and S.E. Fienberg. 1996. "Multiple Sample Estimation of Population and Census Undercount in the Presence of Matching Errors." *Survey Methodology* 22: 55–64.
- Eddy, W.F. and S.E. Fienberg. 2014. *Envisioning the 2030 U.S. Census*. Morris Hansen Lecture, Washington Statistical Society, in preparation.

- Fellegi, I. and A. Sunter. 1969. "A Theory of Record Linkage." *Journal of the American Statistical Association* 64: 1183–2010. Doi: <http://dx.doi.org/10.1080/01621459.1969.10501049>.
- Fienberg, S.E. 1972. "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." *Biometrika* 59: 409–439. Doi: <http://dx.doi.org/10.1093/biomet/59.3.591>.
- Fienberg, S.E. 2011. "Bayesian Models and Methods in Public Policy and Government Settings." *Statistical Science* 26: 212–226. Doi: <http://dx.doi.org/10.1214/10-STS331>.
- Fienberg, S.E., M. Johnson, and B. Junker. 1999. "Classical Multilevel and Bayesian approaches to Population Size estimation Using Multiple Lists." *Journal of the Royal Statistical Society. Series A* 162: 383–406. Doi: <http://dx.doi.org/10.1111/1467-985X.00143>.
- Fienberg, S.E. and D. Manrique-Vallier. 2009. "Integrated Methodology for Multiple Systems Estimation and Record Linkage Using a Missing Data Formulation." *Advances in Statistical Analysis* 93: 49–60.
- Fienberg, S.E. and K. Prewitt. 2010. "Save Your Census." *Nature* 466: 1043, August 26, 2010. Doi: <http://dx.doi.org/10.1038/4661043a>.
- Hamel, M. and Y. Béland. 2013. "Future Developments on the Canadian Census of Population." In Proceedings of the ISI World Statistical Congress, August 25-30, 2013, Hong Kong. Available at: <http://www.statistics.gov.hk/wsc/IPS094-P3-S.pdf> (accessed August 6, 2015).
- Herzog, T., F. Scheuren, and W. Winkler. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer-Verlag.
- Hogan, H. 1993. "The 1990 Post-Enumeration Survey: Operations and Results." *Journal of the American Statistical Association* 88: 1047–1060. Doi: <http://dx.doi.org/10.1080/01621459.1993.10476374>.
- International Working Group for Disease Monitoring and Forecasting. 1995a. "Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development." *American Journal of Epidemiology* 142: 1047–1058.
- International Working Group for Disease Monitoring and Forecasting. 1995b. "Capture-Recapture and Multiple-Record Systems Estimation II: Applications in Human Diseases." *American Journal of Epidemiology* 142: 1059–1068.
- Kurtz, Z. 2014. "Smooth Post-Stratification for Multiple Capture-Recapture." Unpublished manuscript. Available at: <http://arxiv.org/abs/1302.0890>. (accessed August 6, 2015)
- Larsen, M.D. and D.B. Rubin. 2001. "Iterative Automated Record Linkage Using Mixture Models." *Journal of the American Statistical Association* 96: 32–41. Doi: <http://dx.doi.org/10.1198/016214501750332956>.
- Lee, A.J., G.A.F. Seber, J.K. Holden, and J.T. Huakau. 2001. "Capture-Recapture, Epidemiology, and List Mismatches: Several Lists." *Biometrics* 57: 707–713. Doi: <http://dx.doi.org/10.1111/j.0006-341X.2001.00707.x>.
- Manrique-Vallier, D. and S.E. Fienberg. 2008. "Population Size Estimation Using Individual Level Mixture Models." *Biometrical Journal* 50: 1051–1063. Doi: <http://dx.doi.org/10.1002/bimj.200810448>.

- Nirel, R. and H. Glickman. 2009. "Sample Surveys and Censuses." In *Sample Surveys: Design, Methods and Applications*, edited by D. Pfeffermann and C.R. Rao, 539–565.
- Sadinle, M. 2014. "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach." *Annals of Applied Statistics* 8: 2404–2434. Doi: <http://dx.doi.org/10.1214/14-AOAS779>.
- Sadinle, M. 2015. "A Bayesian Partitioning Approach to Duplicate Detection and Record Linkage." Unpublished Ph.D. Dissertation. Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, U.S.A.
- Sadinle, M. and S.E. Fienberg. 2013. "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems." *Journal of the American Statistical Association* 108: 385–397. Doi: <http://dx.doi.org/10.1080/01621459.2012.757231>.
- Sadinle, M., R. Hall, and S.E. Fienberg. 2011. "Approaches to Multiple Record Linkage." In Proceedings of the ISI World Statistical Congress, 21-26 August 2011, Dublin, 1064–1071.
- Seber, G.A.F., J.T. Huakau, and D. Simmons. 2000. "Capture-Recapture, Epidemiology and List Mismatches: Two Lists." *Biometrics* 56: 1227–1232.
- Steorts, R., R. Hall, and S.E. Fienberg. 2014a. "SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication." *Journal of Machine Learning Research* 33: 922–930. Available at: <http://jmlr.csail.mit.edu/proceedings/papers/v33/steorts14.pdf> (accessed August 6, 2015).
- Steorts, R., R. Hall, and S.E. Fienberg. 2014b. "A Bayesian Approach to Graphical Record Linkage and De-duplication." Under Submission. Available at: <http://arXiv:1312.4645>
- Tancredi, A. and B. Liseo. 2011. "A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems." *Annals of Applied Statistics* 5: 1553–1585. Doi: <http://dx.doi.org/10.1214/10-AOAS447>.
- Ventura, S. and R. Nugent. 2014. "Hierarchical Clustering with Distributions of Distances for Large-Scale Record Linkage." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer, pp. 283–298. Berlin: Springer Link. Lecture Notes in Computer Science, Vol. 8744. Doi: <http://dx.doi.org/10.1007/978-3-319-11257-222>.
- Yang, X., S.E. Fienberg, and A. Rinaldo. 2012. "Differential Privacy for Protecting Multidimensional Contingency Table Data: Extensions and Applications." *Journal of Privacy and Confidentiality* 4. Available at: <http://repository.cmu.edu/jpc/vol4/iss1/5> (accessed August 6, 2015).