Partial Synthesis of the Longitudinal Employer-Household Dynamics (LEHD) Database[*]

Simon D. Woodcock[†]
Simon Fraser University
simon_woodcock@sfu.ca

Gary Benedetto
US Census Bureau
gary.linus.benedetto@census.gov

November 2008

Preliminary and incomplete
Please do not cite without permission

Abstract

We describe ongoing efforts to develop public-use partially-synthetic microdata based on the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) database. These are longitudinal linked data on employers and employees, constructed from a variety of administrative and survey data sources. Confidential characteristics of workers, firms, and jobs are replaced with synthetic values. The graph of employment relationships is preserved, to the extent that doing so does not compromise privacy of workers and firms. We describe methods used to generate the synthetic values, and provide a preliminary assessment of utility and disclosure risk in the partially synthetic data.

[*] This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. This document is released to inform interested parties of research and to encourage discussion. This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grants SES-9978093 and SES-0427889 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant R01 AG018854, and the Alfred P. Sloan Foundation. The views expressed on statistical, methodological, or technical issues are those of the authors and not necessarily those of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau supports external researchers' use of these data through the Research Data Centers (see www.ces.census.gov). For other questions regarding the data, please contact Jeremy S. Wu, Program Manager, U.S. Census Bureau, LEHD Program, Demographic Surveys Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA. (Jeremy.S.Wu@census.gov http://lehd.dsd.census.gov).

[†] Correspondence to: Simon Woodcock, Department of Economics, Simon Fraser University, 8888 University Dr., Burnaby, BC V5A 1S6, Canada.

1 Introduction

Statistical agencies face two competing objectives when preparing data for public release. On the one hand, they seek to provide their users with high quality data. On the other hand, they must maintain the privacy of respondents. There is a critical trade-off between these objectives, because protecting privacy usually entails information loss (Duncan et al., 2001). Unless care is taken, measures to protect privacy can invalidate statistical inferences.

In this paper, we describe ongoing work to develop multiply-imputed, partially synthetic data based on the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) database. These are longitudinal linked data on employers and employees, constructed from a variety of administrative and survey data sources.

Synthetic data are an alternative to traditional disclosure limitation methods that permit valid statistical inferences using standard software and methods. They are comprised of synthetic values sampled from an estimate of the joint distribution of the confidential database. Rubin (1993) proposes multiple imputation to generate the synthetic values;[1] Fienberg (1994) suggests bootstrap methods. Under either approach, the released data pose little or no disclosure risk because they are completely synthetic, i.e., contain no actual data on actual respondents. However, this approach requires knowledge, or a good estimate, of the joint distribution of the data. This is impractical in many instances, including the one considered here. Specifically, a fully synthetic approach would require modeling which individuals are employed at which firms – a problem that remains intractable.

Thus we adopt an alternative approach: partial synthesis. These are data on actual respondents, with confidential characteristics replaced with synthetic values sampled from an estimate of the joint distribution of the confidential data conditional on disclosable data. As described in Reiter (2003), partially synthetic data allow valid statistical inferences about population quantities. In our application, we replace confidential characteristics of workers, firms, and jobs with multiple synthetic values sampled from the posterior predictive distribution of an imputation model.

The current version of the partially synthetic LEHD data preserves the graph of employment relationships. That is, we preserve the history of which individuals were ever employed at which firms. This replicates a number of simple summaries of individuals' and firms' employment histories in the partially synthetic data. This is unlikely to directly increase the risk of identity disclosure, since an intruder is unlikely to have reliable secondary measures of these summaries. However, they do uniquely identify a large proportion of units in

[1] This proposal is developed more fully in Raghunathan et al. (2003). Reiter (2002) provides a simulation study, Reiter (2005b) discusses inference, and Reiter (2005a) provides an application.

the partially synthetic data. Consequently, in many cases an intruder will be able to link records corresponding to the same unit across synthetic implicates. This increases the risk of attribute disclosure, because it increases the information available to an intruder who seeks to predict the actual values of confidential characteristics. We consequently describe some possible strategies to limit an intruder's ability to link records across implicates, without invalidating inferences based on the partially synthetic data.

The remainder of the paper is organized as follows. We describe the LEHD data in Section 2, and our synthesis approach in Section 3. Section 4 provides a preliminary assessment of data quality, and we discuss disclosure risk in Section 5. Section 6 concludes.

2 The LEHD Data

The LEHD data are administrative, constructed from quarterly Unemployment Insurance (UI) system wage reports. Every state, through its Employment Security agency, collects quarterly earnings and employment information to manage its unemployment compensation program. The characteristics of the UI wage data vary slightly from state to state. However the Bureau of Labor Statistics (1997, p. 42) claims that UI coverage is "broad and basically comparable from state to state" and that "over 96 percent of total wage and salary civilian jobs" were covered in 1994. Further details regarding UI wage records and the LEHD data can be found in Abowd et. al. (forthcoming). With the UI wage records as its frame, the LEHD data comprise the universe of employers required to file UI system wage reports — that is, all employment potentially covered by the UI system in participating states.

Individuals are uniquely identified in the data by a Protected Identity Key (PIK). Employers are identified by an unemployment insurance account number (SEIN). Some SEINs operate in more than one location; each location is separately identified by an account number (SEINUNIT). The UI wage records associate each PIK with an SEIN in each quarter that the individual was employed. When the individual is employed at an SEIN that operates in multiple locations, the specific SEINUNIT at which the individual was employed is only known in some states. In other states, the identity of the SEINUNIT is multiply-imputed. See Abowd et. al. (forthcoming) for details on the imputation model.

In addition to employer and employee identifiers, the UI wage records contain information about employment earnings in each quarter. Earnings are a measure of total compensation that includes gross wages and salary, bonuses, stock options, tips and gratuities, and the value of meals and lodging when these are supplied (Bureau of Labor Statistics (1997, p. 44)).

The LEHD database integrates the UI wage records with internal Census Bureau data to obtain additional demographic characteristics of individuals (e.g., sex, race, date of birth, residential geography) and their employers (e.g., industry, geography, size, payroll). Given this structure, it is conceptually useful to consider the LEHD data as comprising data from three separate sampling frames (individuals, employers, and jobs), linked via the employment relation as given by the UI wage record.

Our application is based on data from one state (whose identity is confidential), over the period 1993-2004. The sample comprises about 1 million individuals employed in the state at some time during the sample period. We observe these individuals in a total of about 3.5 million distinct employment relationships at one of approximately 75,000 distinct employers (SEINs) who operate a total of about 85,000 SEINUNITs.

We select a subset of variables for partial synthesis. These include characteristics of individuals, their jobs, and employers. Specifically, we synthesize individuals' sex, race, date of birth, and county of residence; employers' industry (NAICS sector) and county; the complete history of quarters in which each job was active, and the history of employment earnings in each of those quarters. The only un-synthesized information is the employment graph: the history of which individuals were ever employed at each firm.[2] The extent to which this information can be used to identify individuals and employers, and possible approaches to further reduce disclosure risk by perturbing the graph, are discussed in Section 5. We note, however, that synthesis of the above-mentioned characteristics implicitly perturbs a number of additional characteristics of employers and employees that can be constructed via the employment graph, e.g., an employer's total employment and payroll in each quarter, the dates in which employers were active in the labor market, the history of industries in which an individual was employed, etc.

Various data items are missing, generally with quite low frequency, in the LEHD data. Missing data items have been multiply-imputed by Census Bureau staff for other purposes. Our application is based on four completed data implicates of the LEHD data.

3 Synthesis Approach

We generate synthetic values sequentially, one or more variables at a time, in a procedure described below. In each case, we sample synthetic values from the posterior predictive distribution of an imputation model appropriate for the variable(s) in question, conditional on available information from all three sampling frames. We generate four synthetic values

[2] Here and throughout, we use the terms "firm" and "employer" synonymously with "establishment": an SEIN, when the SEIN operates in a single location only, or an SEINUNIT when the corresponding SEIN operates in multiple locations.

of each confidential variable for each completed data implicate, for a total of 16 partially-synthetic implicates.

As we proceed through the imputation sequence, we condition draws from the predictive distribution on synthetic values of all variables earlier in the sequence, and on actual values of all variables later in the sequence. This procedure approximates sampling from an estimate of the joint distribution of the confidential variables, conditional on disclosable information (the employment graph). This is formalized as follows.

Let Y denote the complete set of confidential variables. We denote a collection of elements (one or several confidential variables) in Y by Y_k, where $k = 1, ..., K$ indexes the order in the imputation sequence. Disclosable information (in our case, the employment graph) is denoted X. Synthetic values of Y_k, denoted $\tilde{Y}_k$, are sampled from the posterior predictive distribution:

$$p_k\left(\tilde{Y}_k|X, Y\right) = \int p_k\left(\tilde{Y}_k|X, \tilde{Y}_1, ..., \tilde{Y}_{k-1}, Y_{k+1}, ..., Y_K, \theta_k\right) p_k\left(\theta_k|X,Y\right) d\theta_k. \tag{1}$$

where $p_k\left(\tilde{Y}_k|X, \tilde{Y}_1, ..., \tilde{Y}_{k-1}, Y_{k+1}, ..., Y_K, \theta_k\right)$ is the likelihood of an appropriate imputation model for Y_k, and $p_k\left(\theta_k|X,Y\right)$ is the corresponding prior. We discuss the specifics of the imputation models for each Y_k below. In each case, we estimate the imputation model on each completed data implicate, and sample four multiply-imputed synthetic values from the corresponding predictive distribution.

The first collection of variables in the imputation sequence (Y_1) is the set of discrete individual characteristics: sex, race, and county of residence. We synthesize these variables jointly using a multinomial likelihood for cells defined by their cross-classification. The likelihood conditions on the individual's modal county and industry of employment, and the individual's quartile in the marginal distributions of average employment earnings, number of employers, number of quarters worked, date of birth, average employer size, and average earnings per worker at their employers. The prior is an equally-weighted mixture of an uninformative prior and two informative Dirichlet priors, given a total weight equivalent to 100,000 individuals. The first informative prior is based on marginal counts of individuals in each sex×race×county cell; the second further conditions on quartiles of the distribution of average employment earnings.

The next collection of variables in the synthesis sequence (Y_2) is the set of discrete firm characteristics: industry (NAICS sector) and county. These are synthesized jointly using a multinomial likelihood, conditional on employees' modal county of residence and the firm's quartile in the marginal distributions of average employment, mean earnings per employee, first and last year of operation, number of establishments (locations) operated by the SEIN, average employee age, proportion of employees that are male, and proportions black, hispanic, or another minority race. The prior is an equally-weighted mixture of an uninformative prior and three informative Dirichlet priors, given a total weight equivalent to 10,000 firms. The first informative prior is based on marginal counts of firms in each industry×county cell; the second and third condition on quartiles of the distribution average employment and average earnings per employee, respectively.

We synthesize date of birth (Y_3) next. This variable is measured with daily detail. The imputation model is a linear regression, coupled with the density-based transformation defined in Woodcock and Benedetto (2007). The density-based transformation replicates the distribution of birth date in the synthetic data, up to sampling error in our estimate of its distribution, on subdomains defined by the cross-classification of sex, race, number of quarters worked (3 categories), number of distinct employers (3 categories), and ten-year age ranges.[3] The linear regression model further conditions on various functions of the individual's employment and earnings history (e.g., the first three moments of the individual's distribution of earnings and quarters worked), the number of distinct jobs held, the average birthdate of employing firms, and the first three moments of earnings and quarters of employment in the firms, counties, and industries where the individual worked, and in her county of residence.

Finally, we synthesize the employment and wage histories (Y_4). The exact imputation models, and their application, are relatively complex. We give a detailed account in the Appendix, and a brief description here. For each job, we first synthesize the sequence of quarters in which the job was active, and then impute earnings in each quarter. In each case, the imputation model conditions on various functions of previously synthesized variables from both the individual and firm sampling frames, functions of the employment graph, and various functions of the individual's employment and wage history. The main imputation model for the employment history is a binary logistic regression for each quarter in the sample period, which we use to impute whether the job was active in that quarter. The imputation model for earnings is a linear regression model, again paired with the Woodcock and Benedetto (2007) density-based transformation, which preserves the distribution of earnings on subdomains defined by the cross-classification of sex, race, age category, and various indicators for earnings and employment volatility. We assign uninformative priors in both cases.

4 Preliminary Assessment of Data Utility

This section provides a preliminary assessment of utility in the partially synthetic data. Throughout, completed data estimates are based on Rubin's (1976) combining rules, and

[3] The regression model and density transformation are estimated separately on each of this subdomains.

synthetic data estimates are based on Reiter's (2004) combining rules. The former accounts for uncertainty arising from the multiple imputation of missing data. The latter accounts for uncertainty arising from the double use of multiple imputation to complete missing data and generate the synthetic values.

Table 1 presents the first four moments of the univariate distributions of continuous variables in the completed and partially synthetic data. Synthetic moments of birthdate, in-sample job duration, and earnings, are very close to the corresponding completed data moments. We synthesize these variables directly, so this is not too surprising. In contrast, there are larger discrepancies between completed and synthetic moments of derived firm-level variables (the number of quarters the firm has positive employment, quarterly employment, and quarterly payroll). Our synthesis procedure only targets these variables indirectly. That is, they are derived by aggregating job histories to the firm level. In general, the completed and synthetic moments do not differ wildly, which is encouraging.

Figure 1 presents a simple comparison of univariate margins of categorical variables in the completed and partially synthetic data. For each variable (county of residence, employer county, industry, and race) we plot completed and synthetic proportions in each category.[4] In each case, values lie very close the 45 degree line, i.e., synthetic proportions are very close to completed proportions.

Figure 2 summarizes the relationship between completed and synthetic means of earnings and in-sample job duration on a large number of subdomains. On each subdomain, we calculate the absolute value of the percentage deviation of the synthetic mean from the completed data mean, and the proportion of the completed data 95% confidence interval (CI) covered by the corresponding synthetic confidence interval. The figure summarizes the distribution of these percent deviations across cells, distinguishing between cells with good CI overlap (over 50 percent) and cells with poor CI overlap (50 percent or less). In the vast majority of cases where CI overlap exceeds 50 percent, the completed data CI is completely covered by the synthetic CI. In contrast, in most cases where CI overlap is 50 percent or less there is no overlap at all. In many of the latter cases, the point estimates are actually quite close, but very precisely estimated.

In the left panels of Figure 2, cells are given equal weight. In the right panels, cells are weighted by size. It is clear, comparing the left and right panels, that cells with large percent deviations between completed and synthetic data means, and cells with poor CI overlap, are typically small cells. When weighted by cell size, percent deviations are small

[4] We omit two large cells to control the scale of the plot. This facilitates comparisons between completed and synthetic proportions among the large number of small cells. Omitted cells are: County of Residence = Out of State (completed proportion = 0.436, synthetic proportion = 0.432); and Race = White (completed proportion = 0.855, synthetic proportion = 0.850).

and CI overlap is good in the vast majority of cells. Weighted by cell size, the synthetic and completed means of average quarterly earnings differ by 10 percent or less in 63 percent of cells, and CI overlap is good in over 90 percent of cells. Similarly, the synthetic and completed means of in-sample job duration differ by 10 percent or less in 90 percent of cells, and CI overlap is good in over 97 percent of cells.

Figure 3 is a similar plot for two derived firm-level variables (quarterly employment and payroll). Again, percent deviations are largest, and CI overlap is lowest, in small cells. Weighted by cell size, CI overlap remains quite good for these variables, but percent deviations are somewhat larger than for variables we synthesized directly. In particular, the synthetic and completed means of employment differ by 10 percent or less in 36 percent of cells, and CI overlap is good in over 93 percent of cells. For payroll, the percent deviation is 10 percent or less in 21 percent of cells, and CI overlap is good in 95 percent of cells.

The Quarterly Workforce Indicators (QWI) are a public use data product based on the LEHD microdata. They are a series of aggregate measures of employment and wage dynamics, similar to those defined by Davis and Haltiwanger (1992), released for detailed geographic, industrial, and demographic cells; see Abowd et. al. (forthcoming) for details. In Figures 4-6, we summarize the relationship between some key QWI estimands in the completed and synthetic data, again for a large number of subdomains. Figure 4 presents two quarterly measures of employment (beginning of quarter employment, and full-quarter employment). CI overlap approaches 100 percent in every cell. There are a large number of small cells where percent deviations are quite large (50 percent or more). Weighted by cell size, however, only about 15 percent of cells exhibit such large differences between completed and synthetic data estimands, and about one third of cells exhibit deviations of 10 percent or less. Results are very similar in Figure 5, where we present two quarterly measures of employment dynamics (accessions and separations) and in Figure 6, where we present two wage measures.

Table 2 presents estimates of a simple log earnings regression of the type typically estimated on the LEHD data. For the most part, coefficient estimates are similar in the completed and partially synthetic data. The main discrepancies are the coefficients on industry and year dummies. We note that the imputation model for earnings did not include main effects for year or industry, which explains this result.[5] To assess whether the completed and synthetic data yield similar inferences about the regression coefficients, we investigate their repeated sampling properties. We treat the LEHD data as a population, and take 1,000

[5] The earnings imputation model includes conditions on moments of the distribution of earnings in the current industry, but not industry main effects. The model is estimated separately for each quarter, but does not include year main effects. See Appendix 7 for further details.

random samples of 10,000 jobs from the completed and synthetic data. We estimate the regression model in Table 2 on each of the completed and synthetic data samples, combining estimates across implicates using the Rubin (1976) and Reiter (2004) combining rules. We then calculate the RMSE of each estimated coefficient, treating the completed data estimates in Table 2 as population values. We take the ratio of the synthetic data RMSE to the completed data RMSE in each sample, and average the relative RMSE ($RRMSE$) over samples. Its value for each coefficient is reported in the third column of Table 2. Values are all in the neighborhood of one. The median value over coefficients is 1.01, indicating that RMSE is only about one percent larger in the partially synthetic data than in the completed data for the typical coefficient. Thus, both data sources yield very similar inferences about the regression coefficients in Table 2.

5 Preliminary Assessment of Disclosure Risk

We present a very preliminary analysis of disclosure risk in this Section. For now, the analysis focuses on attribute disclosure risk only. An assessment of identity disclosure risk is forthcoming.

Our measures of attribute disclosure risk presume that an intruder is able to link records across synthetic implicates. In most instances, this would be a conservative assumption. In the current context, however, it is probably realistic. Recall that our synthesis procedure does not perturb the employment graph, i.e., the history of which individuals were ever employed at each firm. Consequently, a number of simples summaries of employment histories are replicated across implicates in the partially synthetic data: the number of distinct firms at which each individual was ever employed (R), and the number of distinct individuals ever employed at each firm (E). About 80 percent of individuals are uniquely identified by the combination of: R, the value of E at each of their employers, the value of R for each of individual ever employed at one of their employers (their coworkers), and the value of E for each of their coworkers' employers. Iterating further in this fashion is likely to uniquely identify even more individuals. It is also likely that a similar procedure would uniquely identify a large fraction of employers. Thus it seems likely that an intruder could link an individual or firm's records across implicates using simple summaries of the employment graph.

Consequently, we assume an intruder estimates unit i's value of the k^{th} confidential variable, $y_{k,i}$, by averaging the unit's synthetic values across all partially synthetic implicates: $\bar{y}_{k,i} = \sum_{m=1}^{M} \tilde{y}_{k,i}^{m}$, where $M = 16$ is the number of partially synthetic implicates. Our main measure of attribute disclosure risk is based on the $RRMSE$ of this estimator of $y_{k,i}$ for each unit:

$$RRMSE_{k,i} = \left(\sqrt{ \left(y_{k,i} - \bar{y}_{k,i} \right)^2 + M^{-1} \left(M - 1 \right)^{-1} \sum_{m=1}^{M} \left(\tilde{y}_{k,i}^{m} - \bar{y}_{k,i} \right)^2 } \right) / y_{k,i}.$$

The distribution of $RRMSE$ in the synthetic data provides a measure of variability in the imputations.

The upper panel of Table 3 reports quantiles of the distribution of the $RRMSE$ of prediction for average quarterly earnings and in-sample job duration. For both variables, $RRMSE$ exceeds 30 percent for the median observation. Thus there is considerable uncertainty about actual (completed data) values of earnings and job duration, even in the case where an intruder can link records across synthetic implicates.

Our second measure of attribute disclosure risk continues to assume that an intruder attempts to predict the actual values of synthetic variables by combining information across implicates. We assume the intruder estimates $\bar{y}_{k,i}$ for each unit, as above, and its variance based on the Reiter (2004) combining rules, and uses these to construct a 95 percent confidence interval for $y_{k,i}$. We then calculate the proportion of the empirical density of y_k that lies within the interval. The idea here is that predictions are more informative when the interval contains a small proportion of the empirical density. This arises when the prediction is very precise (i.e., the confidence interval is very narrow), or when the predicted value lies in a low-density region of the distribution. By definition, low-density regions correspond to uncommon values of y_k. Hence in these cases the synthetic data provides information about the value of $y_{k,i}$ that the intruder would be unlikely to infer without the additional information. Thus the risk of attribute disclosure is largest when the confidence interval contains a small proportion of the empirical density.

The lower panel of Table 3, summarizes the ability of an intruder to predict values of average quarterly earnings and in-sample job duration. Cases at greatest risk of attribute disclosure are those where the 95 percent CI covers the actual (completed data) value, and contains 10 percent of the empirical density or less. For earnings, about 11 percent of cases satisfy this definition. The corresponding value for job duration is 7 percent. It is clear, however, that intruders are also likely to make incorrect inferences when the CI covers a small fraction of the empirical density: for earnings, in about one third of cases where the CI covers 10 percent of the empirical density or less, the CI does not cover the actual value. For job duration, this holds in about one quarter of cases. Elsewhere, the CI covers the actual value with high probability, but also covers a substantial proportion of the empirical density. Thus, we have further evidence that the risk of attribute disclosure is quite low for most individuals.

Two strategies to further reduce disclosure risk are apparent. Both are aimed at reducing the ability of an intruder to combine information across synthetic data implicates. One possibility is to release a sample of observations, rather than the entire LEHD population. If different synthetic implicates are based on different samples, combining information across implicates is difficult: unique summaries of the employment graph in a sample do not guarantee uniqueness in the population, and hence the intruder must assign probabilities that records with identical summaries correspond to the same unit. Furthermore, since most units will not appear in all samples, an intruder will have fewer implicates on which to base predictions about any unit's actual data values, and hence obtain less precise predictions. Another possibility is to slightly perturb the employment graph. Multiply-imputing the identity of a fraction of individuals' employers, possibly restricted to candidate employers in the same industry and geography, would perturb the employment graph while still allowing users to obtain valid inferences. We expect that a fairly small number of such imputations would be sufficient to render non-unique summaries of the employment graph that otherwise uniquely identify individuals.

We anticipate these strategies will also be useful for reducing the risk of identity disclosure. When an intruder is unable to link records across implicates, they obtain less precise predictions about actual values. This limits their ability to match these predictions to a secondary data source with a high degree of confidence.

6 Conclusion and Next Steps

We have described our work in progress to develop multiply-imputed, partially synthetic data based on the LEHD database. Overall, our results thus far suggest quite good data utility and quite low risk of attribute disclosure.

There remains much to do. Further work to assess data utility and disclosure risk is clearly required. Our results thus far already indicate ways in which the synthesis procedures described above can be improved, e.g., estimates in Table 2 suggest the imputation model for earnings should include main effects for year and industry. More fundamentally, we anticipate taking steps to reduce the ability of an intruder to combine information across synthetic data implicates, either by sampling or multiply-imputing elements of the employment graph.

7 Appendix

In this appendix, we provide further detail on the imputation model(s) for employment and wage histories. The unit of observation in this case is a job – an employment relation between a worker and firm.

First, we impute a vector of indicators for earnings volatility. A volatile job is defined as one that persists for 8 or more quarters, and where the difference between the maximum and minimum of quarterly earnings exceeds one quarter of average quarterly earnings. We impute an indicator for such volatility, and indicators for whether earnings have a maximum in quarter one, two, three, or four, via Bayesian bootstrap. The indicators are resampled within cells by sex, race, age category, indicators for the first and last year that the employer was active, industry, and quartiles of the distribution of the proportion of the firm's employees whose earnings meet the volatility definition.

Second, we impute indicators for whether the job's start date was before the first period of the sample, and whether it was before the last period. Collectively, these indicators measure whether the job began in an "interior quarter." Unsurprisingly, the distribution of observed start dates exhibits a large spike in the first quarter of the sample. We impute both indicators using a binary logistic regression model with uninformative prior. The regression models condition on sex, race, age, the volatility indicators, number of jobs held, total number of quarters worked, average earnings at this job, the proportion of quarters worked at this job between the job's first and last quarter (we call this "employment persistence" below), the first and last year that the firm was active in the sample, firm-average age and earnings; the first three moments of the distributions of age, earnings, and quarters of employment in the firm, industry, county of work, and county of residence; and the first three moments of the distribution of the numeric quarters in which the individual worked, weighted by quarterly earnings.

Third, for all cases where the job is imputed to start in an interior quarter, we impute the start quarter of the job via a linear regression with uninformative prior. The regression conditions on sex, race, age, the volatility indicators, number of jobs held, total number of quarters worked, average earnings at this job, the employment persistence measure, the firm's first and last active quarter, the mean job start and end quarters in the industry, quartile of the distribution of start quarter by industry, county of work, and county of residence; the first three moments of the distributions of age, earnings, and quarters of employment in the firm, industry, county of work, and county of residence; and the first three moments of the distribution of the numeric quarters in which the individual worked, weighted by quarterly earnings.

Fourth, we impute an indicator for whether the job lasts more than one quarter. The distribution of job duration exhibits a large spike at one quarter. The imputation model is a logistic regression with uninformative prior. It conditions on sex, race, age, the volatility indicators, number of jobs held, total number of quarters worked, average earnings at this job, job start quarter, the firm's first and last active quarter; and the first three moments of the distributions of age, earnings, and quarters of employment in the firm, industry, county of work, and county of residence.

Fifth, we impute an indicator for whether the job's end date is before the final quarter of the sample. Again, the distribution of end dates exhibits a large spike in the final quarter. The imputation model is a logistic regression with uninformative prior. It conditions on sex, race, age, the volatility indicators, number of jobs held, total number of quarters worked, average earnings at this job, the employment persistence measure, job start quarter, the firm's first and last active quarter; the first three moments of the distributions of age, earnings, and quarters of employment in the firm, industry, county of work, and county of residence; and the first three moments of the distribution of the numeric quarters in which the individual worked, weighted by quarterly earnings.

Sixth, for all cases where the job is imputed to end before the final quarter, we impute the end quarter via a linear regression with uninformative prior. The regression conditions on sex, race, age, the volatility indicators, number of jobs held, total number of quarters worked, average earnings at this job, the employment persistence measure, the firm's first and last active quarter, the mean job start and end quarters in the industry, quartile of the distribution of end quarter by industry, county of work, and county of residence; the first three moments of the distributions of age, earnings, and quarters of employment in the firm, industry, county of work, and county of residence; and the first three moments of the distribution of the numeric quarters in which the individual worked, weighted by quarterly earnings.

Having imputed the start and end quarters of each job by the above procedure, we impute a binary indicator for whether the job was active in each quarter between the job start and end. These indicators are imputed sequentially, moving forward through time. The imputation model is a binary logistic regression with uninformative prior. It conditions on sex, race, age, the volatility indicators, number of jobs held, total number of quarters worked, average earnings at this job, the employment persistence measure, the firm's first and last active quarter; the first three moments of the distributions of age, earnings, and quarters of employment in the firm, industry, county of work, and county of residence; the first three moments of the distribution of the numeric quarters in which the individual worked, weighted by quarterly earnings; up to four quarterly lags of the indicator for whether the job was active, firm employment, the ratio of the firm's employment in the current quarter to its maximum over the sample period, time until the final quarter, and several weighted measures of distance between the current quarter and the center of the job spell.

Finally, we impute earnings in each quarter that the job is active. Again, the imputation is sequential, moving forward through time. The imputation model is a linear regression, paired with the Woodcock and Benedetto (2007) density-based transformation. The transformation and regression model are fitted separately on subdomains defined by the cross-classification of sex, race, age categories, the volatility indicators, and indicators for whether the job was active in the previous and subsequent quarters and in the previous year. The regression model further conditions on earnings in the previous quarter and four quarters (one year) ago (also subject to the density-based transformation), the individual's average quarterly earnings up to this date, up to 4 quarterly leads and lags of the employment indicators, number of jobs held, total number of quarters worked, average earnings at this job, the employment persistence measure, the firm's first and last active quarter, the average first and last quarter of jobs in this industry; the first three moments of the distributions of age, earnings, and quarters of employment in the firm, industry, county of work, and county of residence; the first three moments of the distribution of the numeric quarters in which the individual worked, weighted by quarterly earnings; firm employment, the ratio of the firm's employment in the current quarter to its maximum over the sample period, time until the job's final quarter, several weighted measures of distance between the current quarter and the center of the job spell; quartile of the distribution of start quarter by industry, county of work, and county of residence.

References

Abowd, J. M., B. E. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. D. Woodcock (forthcoming). The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. In T. Dunne, J. B. Jensen, and M. J. Roberts (Eds.), *Producer Dynamics: New Evidence from Micro Data*. Cambridge, MA: National Bureau of Economic Research.

Bureau of Labor Statistics (1997). *BLS Handbook of Methods*. U.S. Department of Labor.

Davis, S. J. and J. Haltiwanger (1992, August). Gross job creation, gross job destruction, and employment reallocation. *The Quarterly Journal of Economics 107*(3), 819–863.

Duncan, G. T., S. A. Keller-McNulty, and S. L. Stokes (2001). Disclosure risk vs. data utility: The r-u confidentiality map. National Institute of Statistical Sciences Technical Report No. 121.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Carnegie Mellon University Department of Statistics Technical Report No. 611.

Raghunathan, T., J. Reiter, and D. Rubin (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics 19*(1), 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics 18*(4), 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology 29*, 181–188.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology 30*, 235 – 242.

Reiter, J. P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 185–205.

Reiter, J. P. (2005b). Significance test for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference 131*, 365–377.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1993). Discussion of statistical disclosure limitation. *Journal of Official Statistics 9*(2), 461–468.

Woodcock, S. D. and G. Benedetto (2007, September). Distribution-preserving statistical disclosure limitation. Mimeo.

Table 1

Univariate Moments of Continuous Variables

		Value in	Value in
Variable	Statistic	Completed Data	Synthetic Data
	Person- and Job-Level Variables		
Birthdate	Mean	1,213	1,214
	Standard deviation	5,743	5,738
	Skewness	-0.516	-0.519
	Kurtosis	-0.166	-0.177
Quarterly Earnings	Mean	4,653	4,649
	Standard deviation	9,563	7,286
	Skewness	357	281
	Kurtosis	301,809	249,558
In-sample Job Duration (Quarters)	Mean	5.34	5.43
	Standard deviation	7.87	7.84
	Skewness	3.01	2.97
	Kurtosis	9.74	9.44
	Derived Firm-Level Variables		
Number of Quarters with Positive Employment	Mean	17.2	13.7
	Standard deviation	14.8	14.7
	Skewness	0.728	1.079
	Kurtosis	-0.851	-0.246
Quarterly Employment	Mean	15.6	11.5
	Standard deviation	75.5	62.2
	Skewness	23.8	26.4
	Kurtosis	839	1025
Quarterly Payroll	Mean	72,519	53,562
	Standard deviation	490,288	381,557
	Skewness	31.3	34.9
	Kurtosis	1,420	1,832

Table 2
Coefficients in Log(Earnings) Regression

	Completed Data Estimate	Synthetic Data Estimate	Relative RMSE in Repeated Samples
Male	0.357	0.406	1.00
White	0.118	0.135	1.00
Black	-0.056	-0.025	1.00
Hispanic	-0.012	-0.004	1.00
Age	0.142	0.141	0.98
0.1*(Age Squared)	-0.201	-0.211	1.01
0.01*(Age Cubed)	0.006	0.008	0.94
Job Tenure	0.062	0.061	1.03
Log(Firm Employment)	0.048	0.043	0.97
Industry Main Effects (NAICS Sector)			
21	0.653	0.159	0.96
22	0.703	0.390	1.07
23	0.227	0.113	1.02
31-33	0.258	0.049	1.04
42	0.243	0.148	1.02
44-45	-0.119	-0.099	1.01
48-49	-0.157	0.003	1.00
51	0.175	0.034	1.00
52	0.590	0.171	1.00
53	-0.156	-0.104	1.00
54	0.259	0.118	1.02
55	-0.040	0.015	1.02
56	-0.696	-0.227	1.01
61	-0.359	-0.158	1.01
62	0.070	0.025	1.01
71	-0.509	-0.243	1.01
72	-0.402	-0.352	1.01
81	-0.207	-0.054	1.01
92	-0.205	0.009	1.00
Year Dummies			
1993	0.142	0.131	1.09
1994	0.071	-0.025	0.99
1995	0.017	-0.070	1.03
1996	-0.032	-0.096	1.00
1997	-0.029	-0.093	0.92
1998	-0.034	-0.104	0.97
1999	-0.014	-0.068	1.11
2000	-0.014	-0.055	1.01
2001	0.002	-0.031	0.96
2002	0.008	-0.018	0.99
Intercept	5.370	5.564	1.18

Table 3
Attribute Disclosure Risk

	Percentiles of RRMSE of Prediction				
	1st	5th	10th	25th	50th
Avg Quarterly Earnings	0.035	0.064	0.087	0.151	0.309
In-sample Job Duration	0.014	0.088	0.122	0.187	0.347
	Percent of Empirical Distribution Covered by Synthetic 95% CI				
	≤ 10%	10-20%	20-30%	30-40%	> 40%
Avg Quarterly Earnings					
Synthetic 95% CI **Does Not** Contain Completed Value	5.22	3.54	2.15	1.18	0.85
Synthetic 95% CI **Does** Contain Completed Value	10.9	13.7	13.2	11.4	37.8
In-sample Job Duration					
Synthetic 95% CI **Does Not** Contain Completed Value	2.29	1.49	4.5	2.09	1.12
Synthetic 95% CI **Does** Contain Completed Value	7.02	5.32	5.74	8.29	62.1

Figure 1: Sample Proportions in Race, County, and Industry Cells

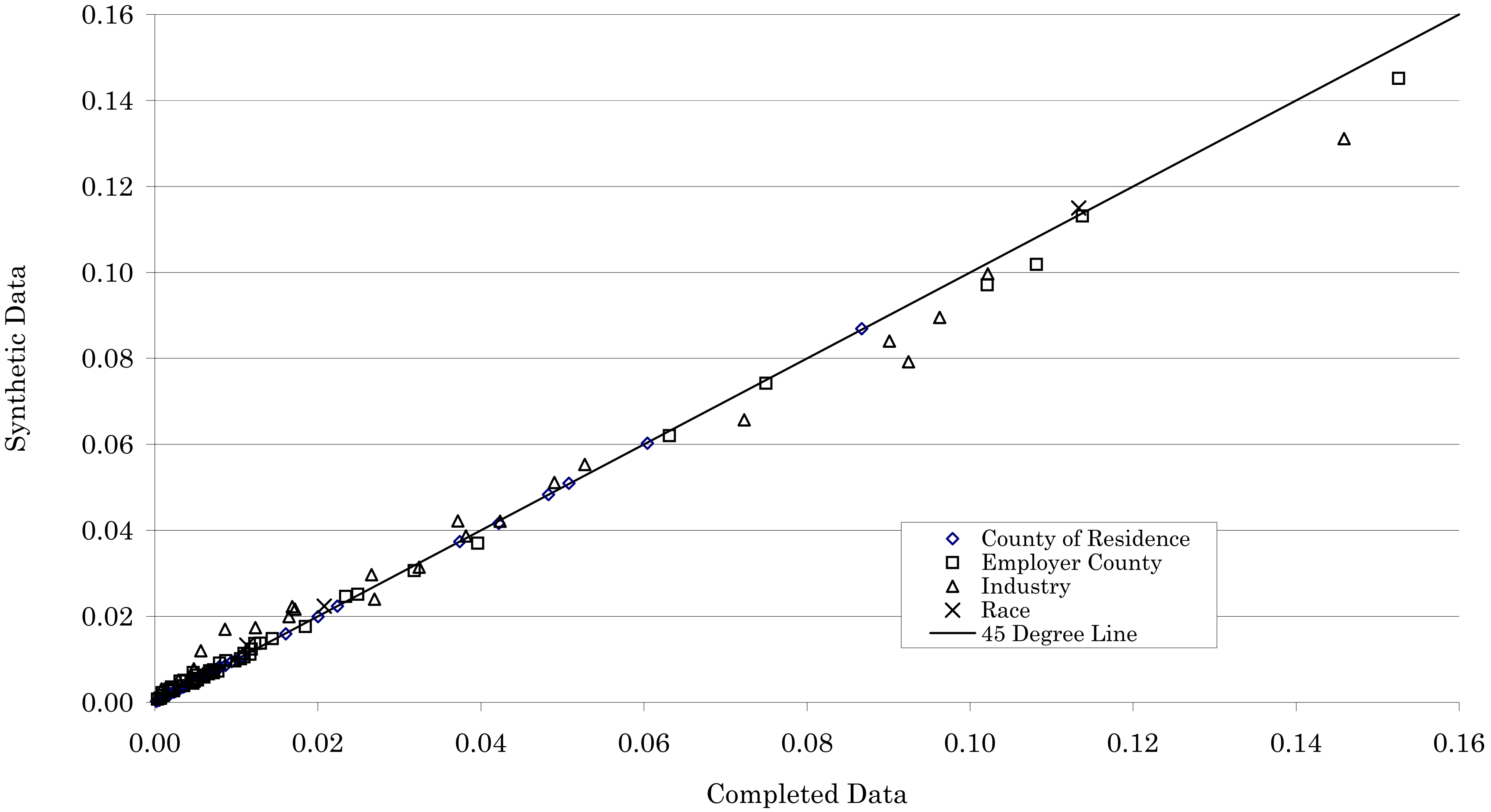
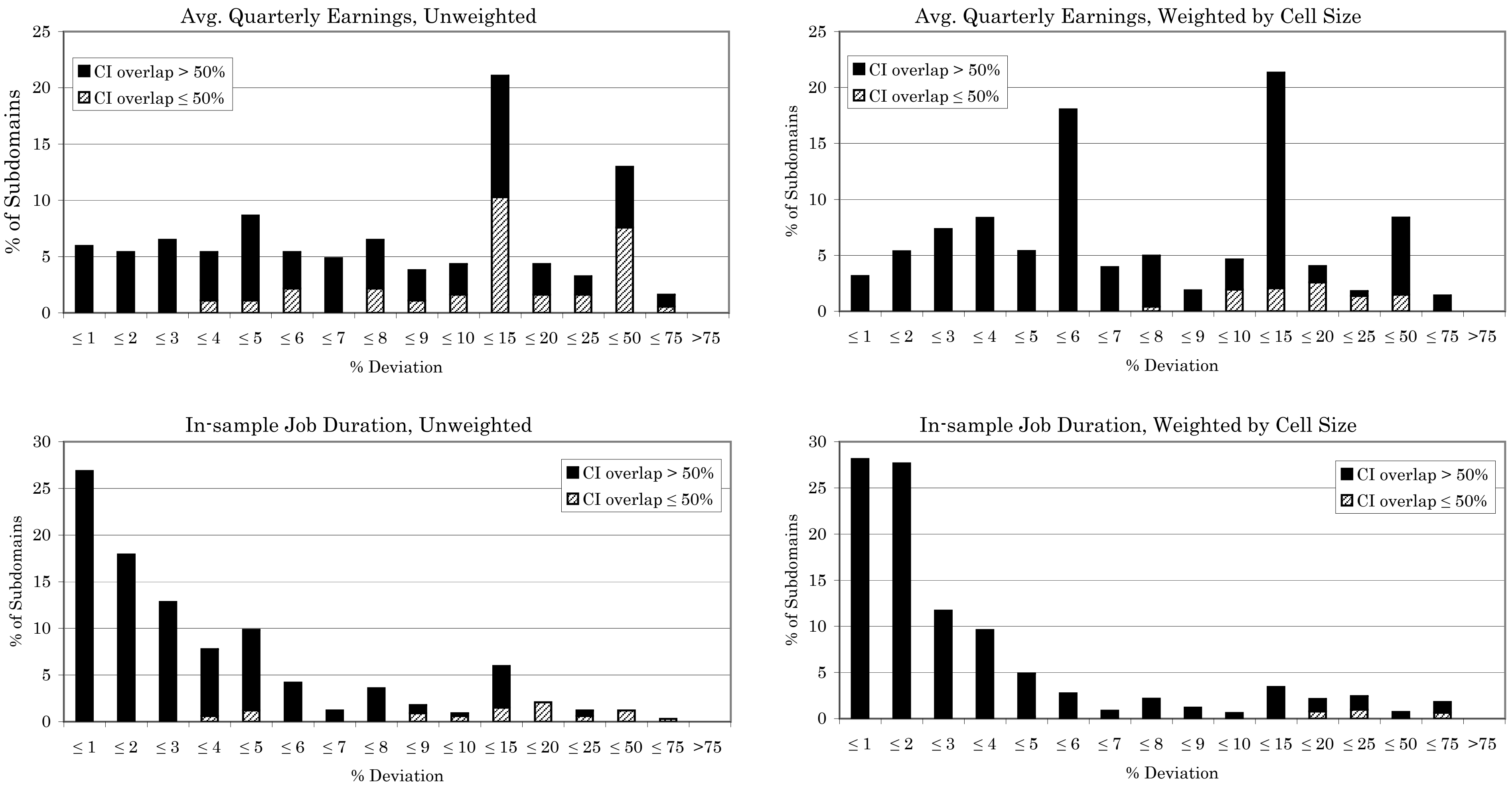


Figure 2: Percent Deviation of Synthetic Data Means from Completed Data Means on Subdomains, Job-Level Variables



Avg. Quarterly Earnings, Unweighted

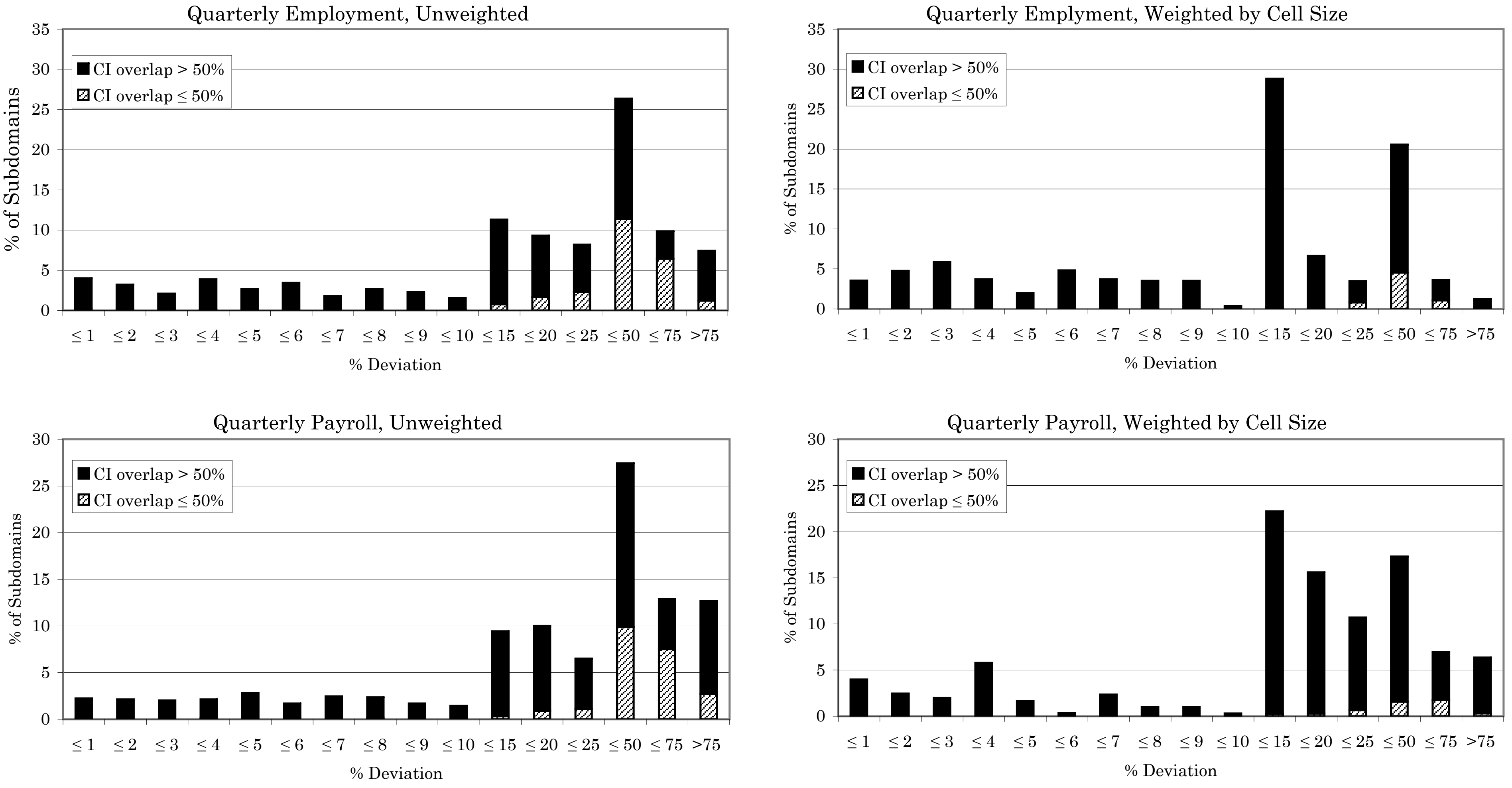
Avg. Quarterly Earnings, Weighted by Cell Size

In-sample Job Duration, Unweighted

In-sample Job Duration, Weighted by Cell Size

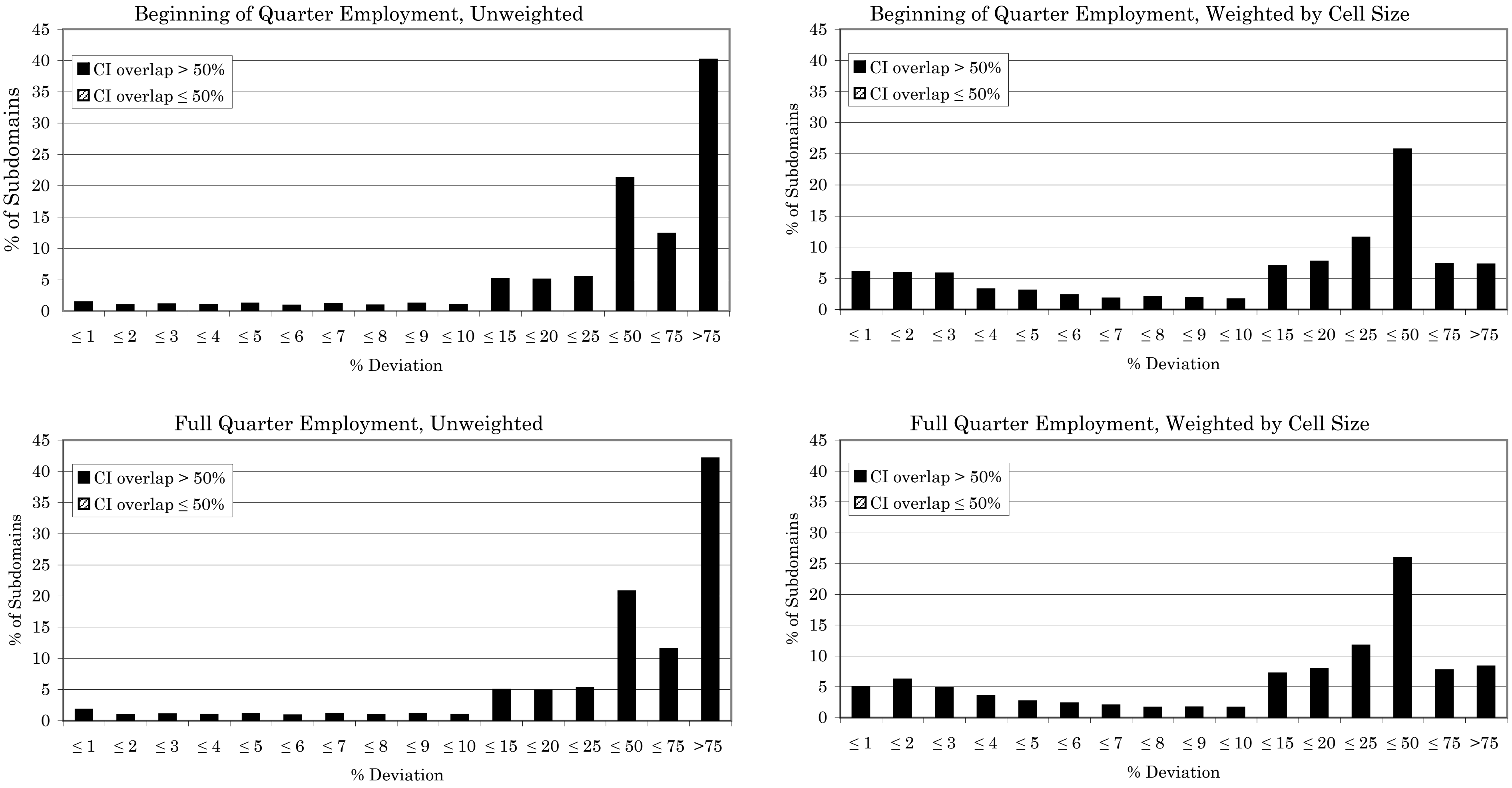
Note: Subdomains are: sex x race x age category; sex x employer county; sex x county of residence; sex x industry; age category; race; industry; sex; county of residence; and employer county.

Figure 3: Percent Deviation of Synthetic Data Means from Completed Data Means on Subdomains, Quarterly Firm-Level Variables



Note: Subdomains are: industry x county; industry; and county.

Figure 4: Percent Deviation of Synthetic Data Means from Completed Data Means on Subdomains, QWI Employment Variables



Beginning of Quarter Employment, Unweighted

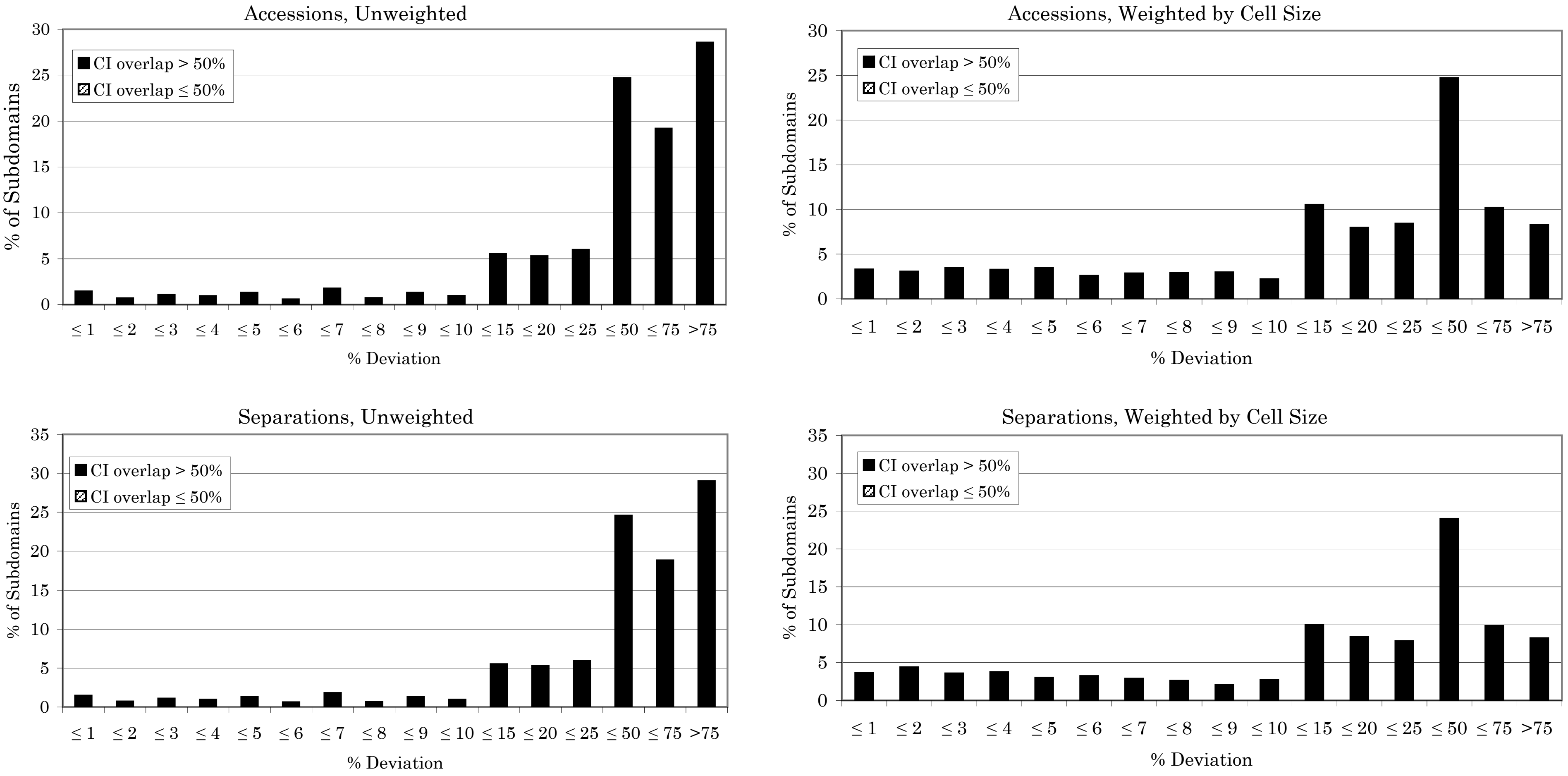
Beginning of Quarter Employment, Weighted by Cell Size

Full Quarter Employment, Unweighted

Full Quarter Employment, Weighted by Cell Size

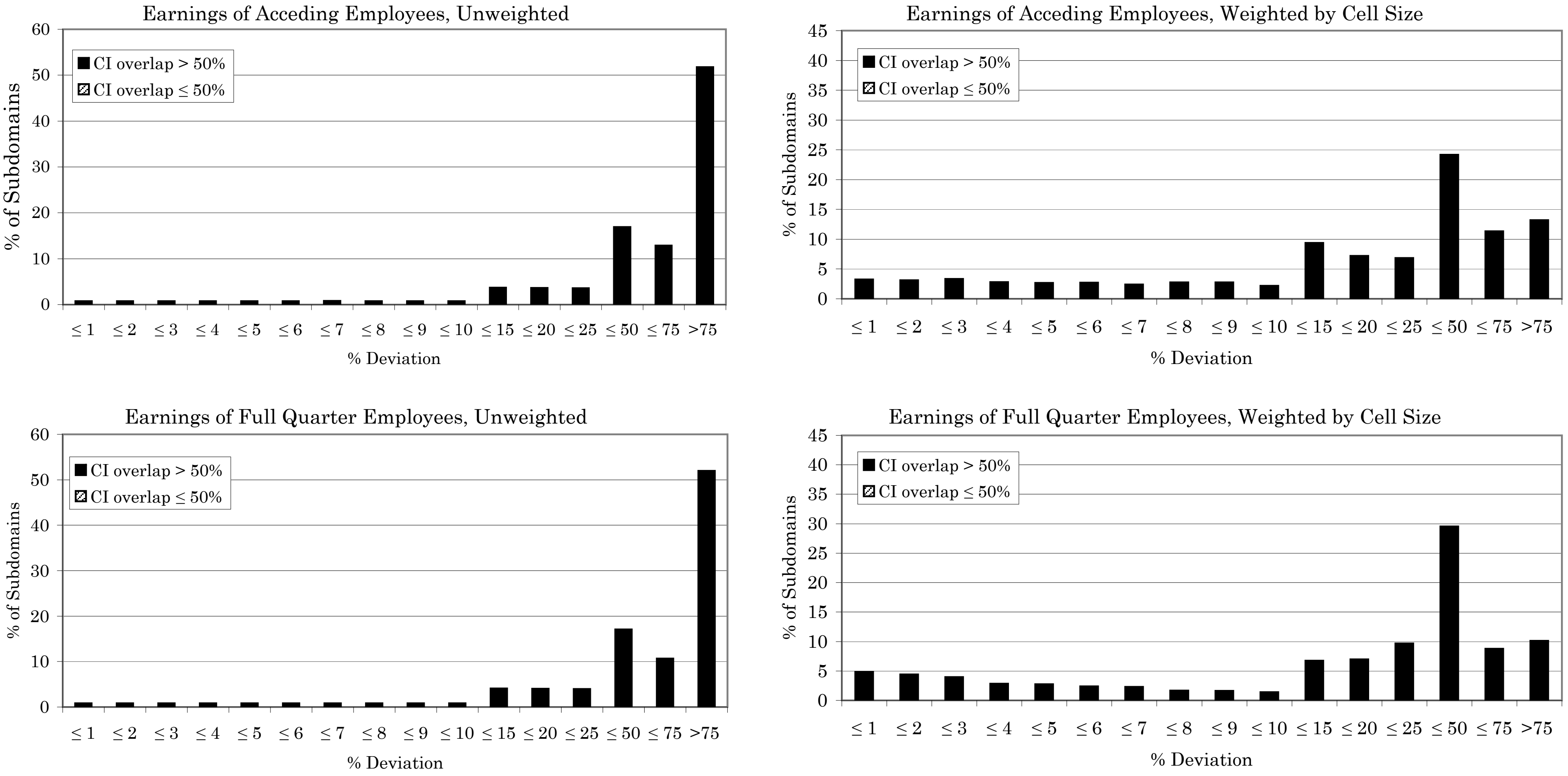
Note: Subdomains are: year x quarter x sex x age category x industry x employer county; year x quarter x employer county x county of residence; year x quarter x sex x race x age category; and year x quarter x industry x employer county.

Figure 5: Percent Deviation of Synthetic Data Means from Completed Data Means on Subdomains, QWI Employment Dynamics



Note: Subdomains are: year x quarter x sex x age category x industry x employer county; year x quarter x employer county x county of residence; year x quarter x sex x race x age category; and year x quarter x industry x employer county.

Figure 6: Percent Deviation of Synthetic Data Means from Completed Data Means on Subdomains, QWI Wage Variables



Note: Subdomains are: year x quarter x sex x age category x industry x employer county; year x quarter x employer county x county of residence; year x quarter x sex x race x age category; and year x quarter x industry x employer county.