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

Simon Woodcock

Simon Fraser University

November 2008

Introduction

The well-known trade-off(s) in disclosure avoidance:

- Statistical agencies seek to provide their users with high quality data. But they must maintain the privacy of respondents.
- 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

An alternative to traditional disclosure limitation methods that permits valid statistical inferences using standard software and methods is to release data sets comprised of synthetic values sampled from an estimate of the joint distribution of the confidential database.

- Rubin (1993), Raghunathan, Reiter, Rubin (2003): multiple imputation
- Fienberg (1994): bootstrap methods.

Under either approach, the released data pose little disclosure risk: they contain no actual data on actual respondents.

However, this requires knowledge, or a good estimate, of the joint distribution of the data. This is impractical in our case.

- Would require modeling which individuals are employed at which firms – this remains intractable.

Partial Synthesis

We adopt an alternative approach: partial synthesis.

Partially synthetic data are data on actual respondents. Confidential characteristics are replaced with synthetic values sampled from an estimate of the joint distribution of the confidential data conditional on disclosable data.

Reiter (2003): multiply-imputed partially synthetic data allow valid statistical inferences about population quantities.

Estimates on each implicate are combined using simple formulae. Variance estimates reflect uncertainty due to imputation (for synthesis, possibly also to complete missing data).

Our Basic Approach

We replace confidential characteristics of workers, firms, and jobs with multiple synthetic values sampled from the posterior predictive distribution of an imputation model.

We do not synthesize the employment graph: the history of which individuals were ever employed at which firms.

This solves the tricky problem of modeling who works where.

But it has implications for disclosure risk: some summaries of individuals' and firms' employment history are preserved, and this may allow an intruder to link records across partially synthetic implicates.

Remainder of Talk

1. Description of the LEHD data
2. Details on synthesis
3. Preliminary assessment of data utility
4. Preliminary assessment of disclosure risk, and discussion
5. Conclusions and next steps

The LEHD Data

The LEHD data are administrative, constructed from quarterly Unemployment Insurance (UI) system wage reports.

The Bureau of Labor Statistics (1997) 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.

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.

Nearly all states now participate in the US Census Bureau's LEHD partnership. Our application is based on one state, whose identity is confidential.

Structure of the LEHD Data

Convenient to represent the LEHD data as derived from three sampling frames:

- Individuals
- Firms
- Jobs

Employment relationships ("jobs") link the individual and firm frames.

"Firms" correspond approximately to establishments (i.e., a business location). They are based on unemployment insurance account numbers. When businesses operate at multiple locations, the specific location at which the individual is employed is known in some states, and multiply-imputed in others.

Structure of the LEHD Data, continued

The UI wage records associate each individual with an employing firm in each quarter that the individual was employed. Also includes a measure of employment earnings.

The LEHD project adds demographic characteristics of individuals (sex, race, date of birth, county of residence), and characteristics of firms (industry, county), to the UI wage records. These characteristics are based on internal Census Bureau sources.

Linkage defines some additional derived characteristics of firms (size, payroll).

The LEHD data: final details

Sample comprises approx. 1 million individuals employed in this state between 1993 and 2004, at approx. 85,000 firms. About 3.5 million employment relationships total.

Some missing data, but not much. These have been multiply-imputed by Census Bureau staff for other purposes.

Our application is based on four completed data implicates. For each completed data implicate, we generate four partially-synthetic implicates $\Longrightarrow$ total of 16 partially-synthetic implicates.

Synthesis: Basic approach

We generate synthetic values sequentially, one or several variables at a time.

In each case, we sample multiple 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.

As we proceed through the imputation sequence, we condition draws from the predictive distribution on synthetic values of variables earlier in the sequence, and on actual values of all variables later in the sequence.

This approximates sampling from an estimate of their joint distribution.

Formalization

Y is the complete set of confidential variables.

$Y_k \in Y$ is a collection of elements (one or several confidential variables), where $k = 1, ..., K$ indexes order in the imputation sequence.

X is disclosable information (the employment graph).

Synthetic values of Y_k, denoted $\tilde{Y}_k$, are sampled from the posterior predictive distribution:

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

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 imputation model for Y_k, and $p_k\left(\theta_k|X, Y\right)$ is the corresponding prior.

For each Y_k, we estimate the imputation model on each completed data implicate, and sample four multiply-imputed synthetic values from the corresponding predictive distribution.

Synthesis order and details

1. Y_1 is all discrete individual characteristics: sex, race, and county of residence.
   - Multinomial likelihood for cells defined by their cross-classification. Likelihood conditions discrete representations of information from all three frames.
   - Prior: equally-weighted mixture of an uninformative prior and two informative Dirichlet priors (based on marginal counts in each sex $\times$ race $\times$ county cell; and these plus quartiles of employment earnings).
2. Y_2 is all discrete firm characteristics: industry (NAICS sector) and county.
   - Also a multinomial likelihood.
   - Prior: equally-weighted mixture of an uninformative prior and three informative Dirichlet priors (based on marginal counts in each industry $\times$ county cell; and these plus quartiles of average employment, and average earnings per employee).

Synthesis order and details, continued

3. Y_3 is date of birth (daily).
   - The imputation model is a linear regression, coupled with the Woodcock and Benedetto (2007) density-based transformation.
   - Transformation procedure replicates the distribution of birth date in the synthetic data, up to sampling error in our estimate of its distribution, on a collection of subdomains (sex $\times$ race $\times$ functions of employment graph).
   - Regression model conditions on additional functions of the individual's employment and earnings history, industry, county, etc.
   - Uninformative prior.

Synthesis order and details, continued

4. Y_4 is the employment history.
   - For each job, we synthesize the sequence of quarters in which the job was active.
   - A hierarchy:
     - First, a logit model to impute whether the job was active in the first quarter of the sample.
     - If job not active in the first quarter, impute the start quarter via linear regression.
     - Then a logit model to impute whether the job lasts more than one quarter.
     - If the job lasts more than one quarter, impute whether the job was still active in final quarter via a logit model.
     - If the job lasts more than one quarter and was not active in the last quarter, impute the end quarter via linear regression.
     - Then impute whether job was active in each quarter between the job's first and last. This is sequential, moving forward through time, conditional on the past (plus other information), via a logit model in each quarter.
   - Uninformative priors throughout.

Synthesis order and details, continued

5. Y_5 is the earnings history.
    - Impute earnings in each quarter that the job is active.
    - Sequential, moving forward through time.
    - The imputation model is a linear regression, paired with the Woodcock and Benedetto (2007) density-based transformation.
    - Transformation preserves distribution of earnings on subdomains of sex $\times$ race $\times$ age categories $\times$ functions of the employment history.
    - Regression model conditions on past earnings history, future employment history, and other information from all three frames.
    - Uninformative prior.

Table 1
Univariate Moments of Continuous Variables

Variable	Statistic	Value in Completed Data	Value in 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

Densities of True and Synthetic Age and Quarterly Earnings on Selected Subdomains

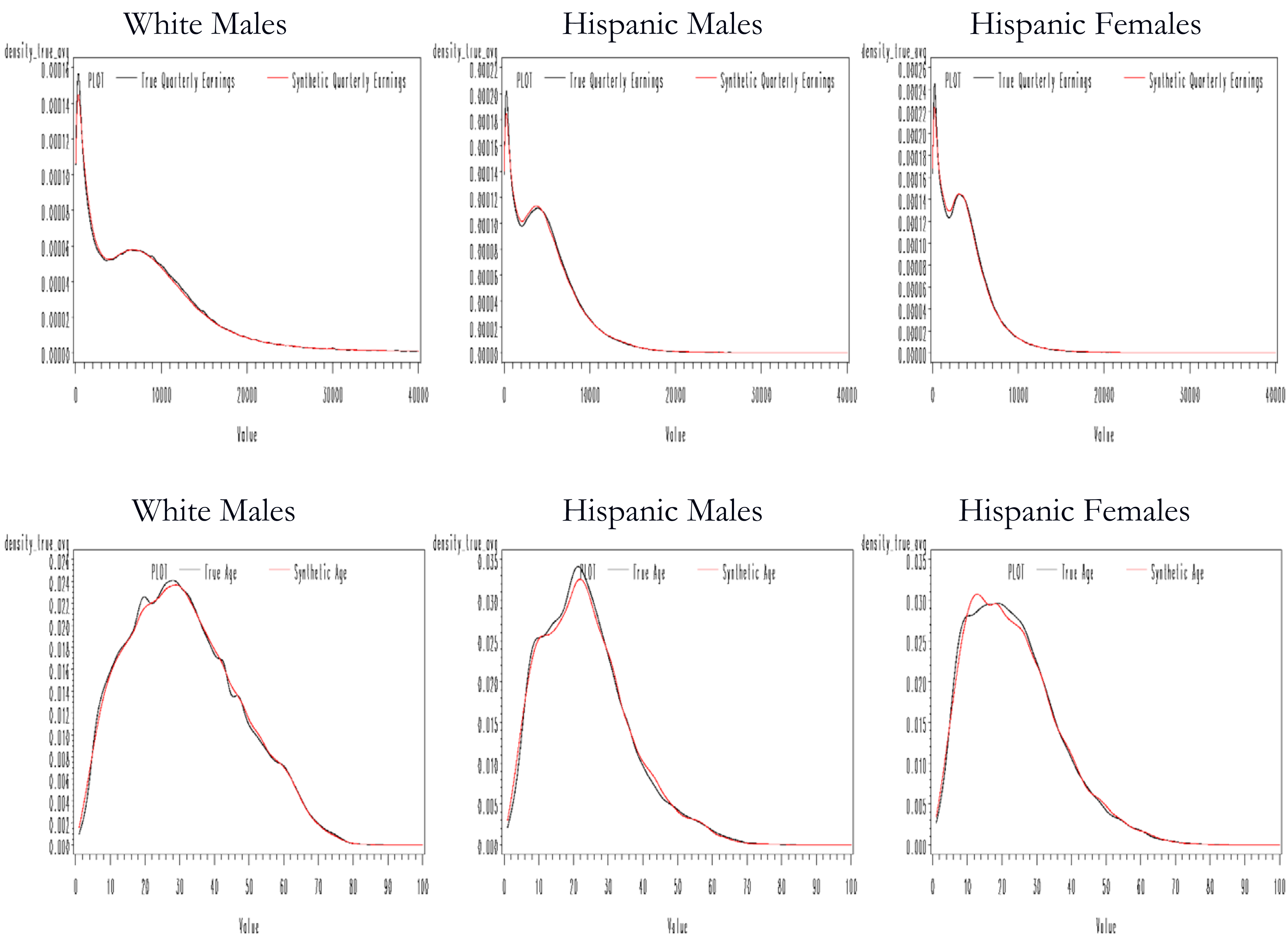


Figure 1: Sample Proportions in Race, County, and Industry Cells

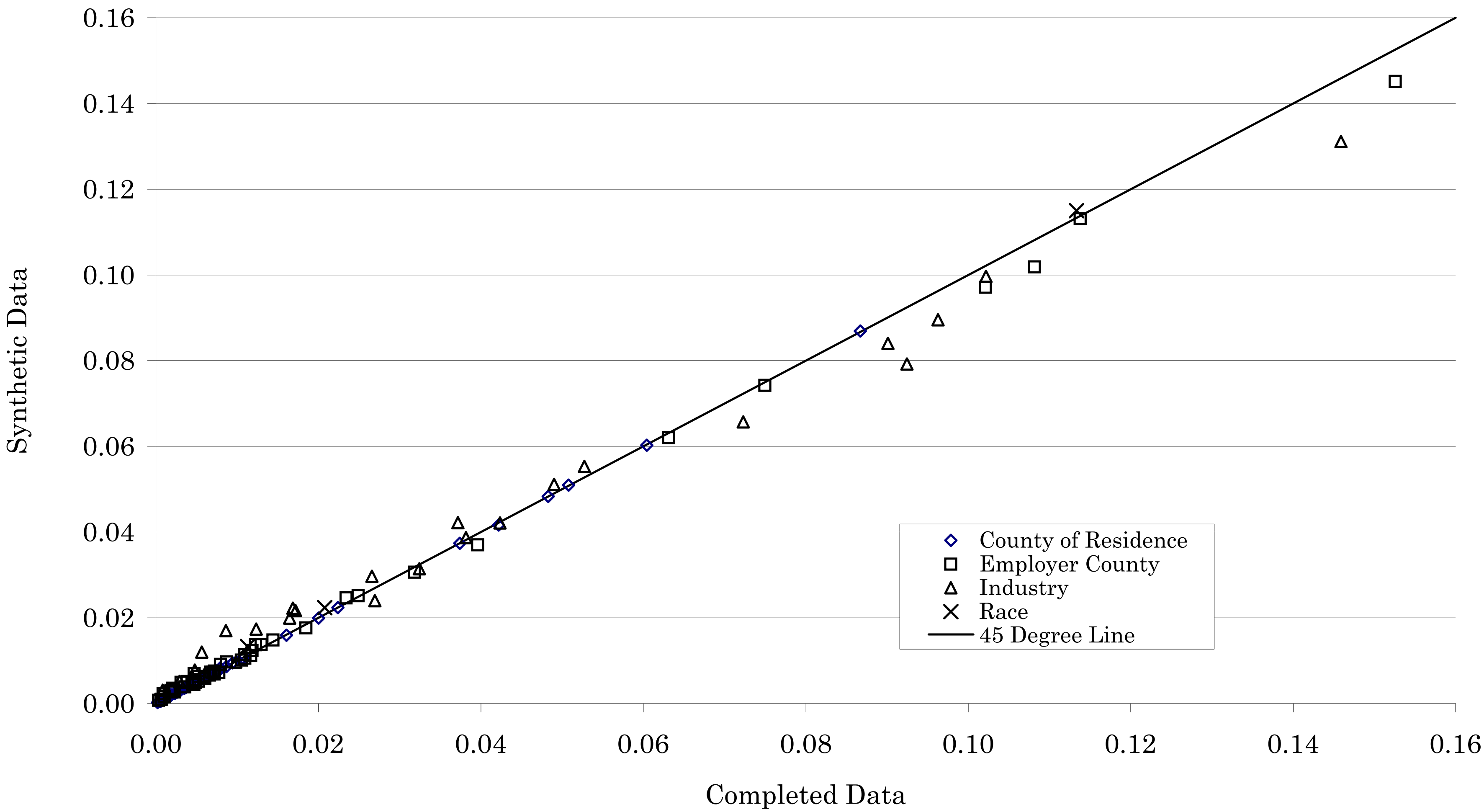
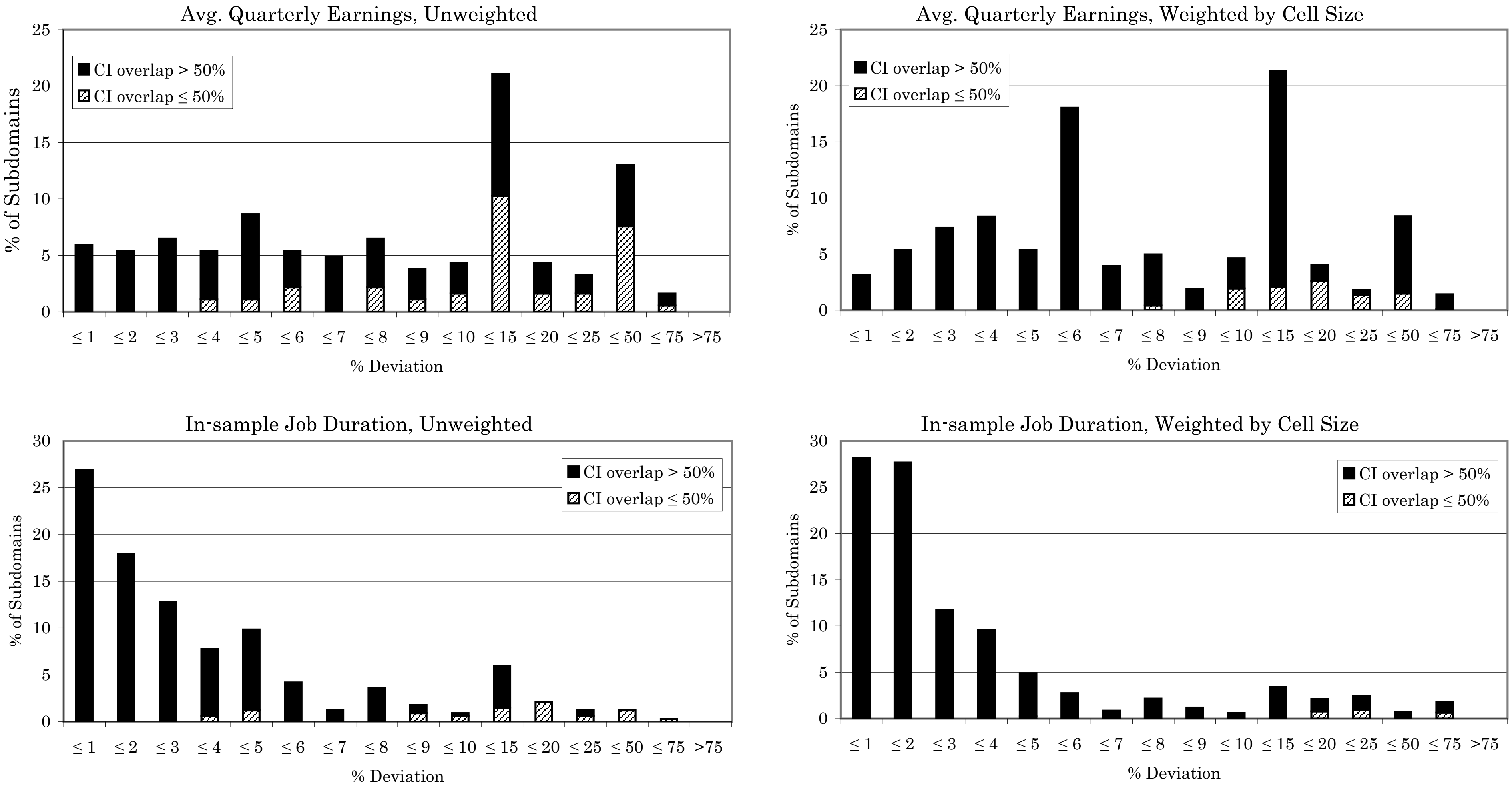
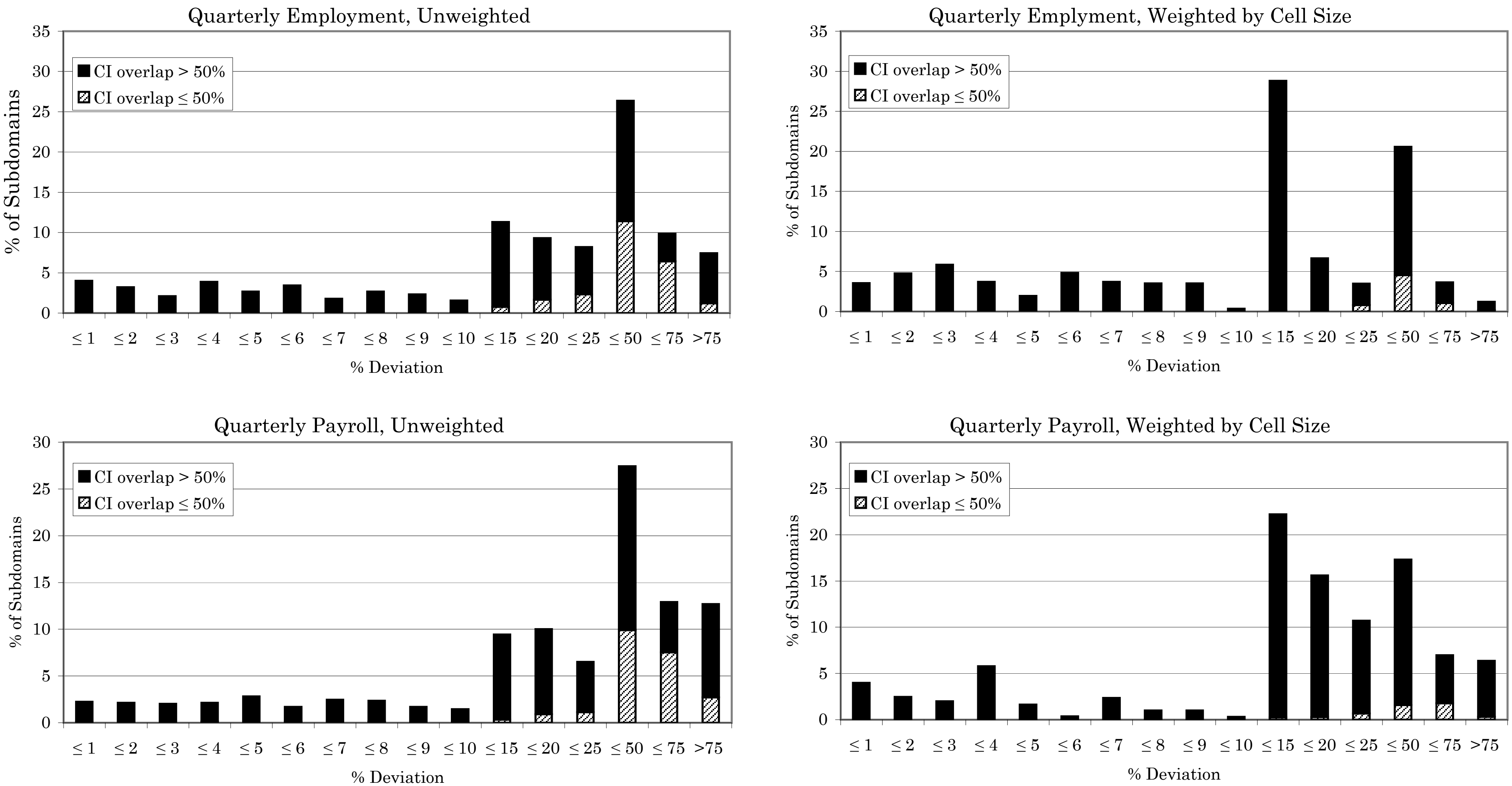


Figure 2: Percent Deviation of Synthetic Data Means from Completed Data Means on Subdomains, Job-Level Variables



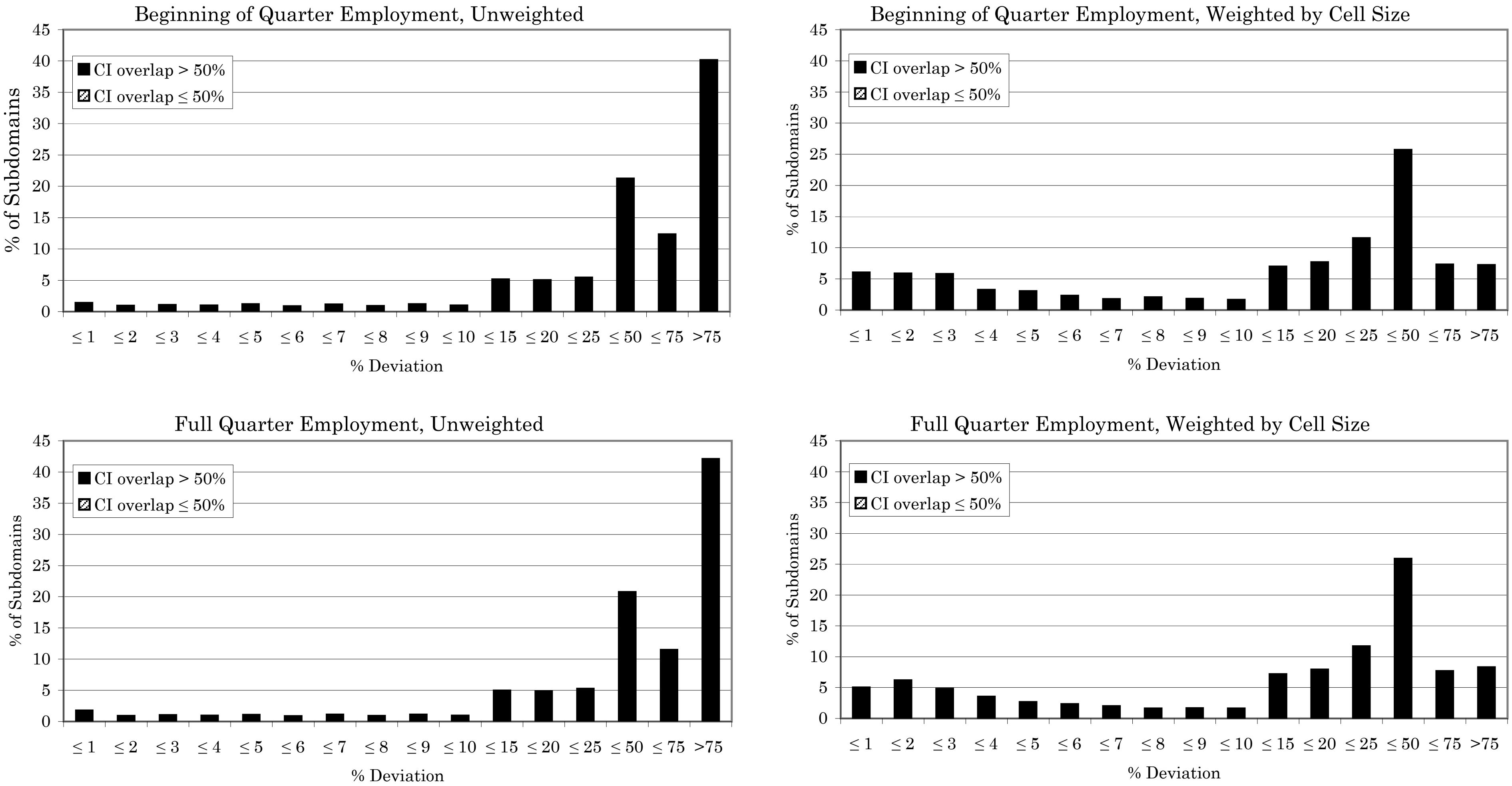
Note: Subdomains are: sex x race x age category; sex x employer county; sex x county of residence; sex x industry; age category; race; industry; sex; county of residence; and employer county.

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



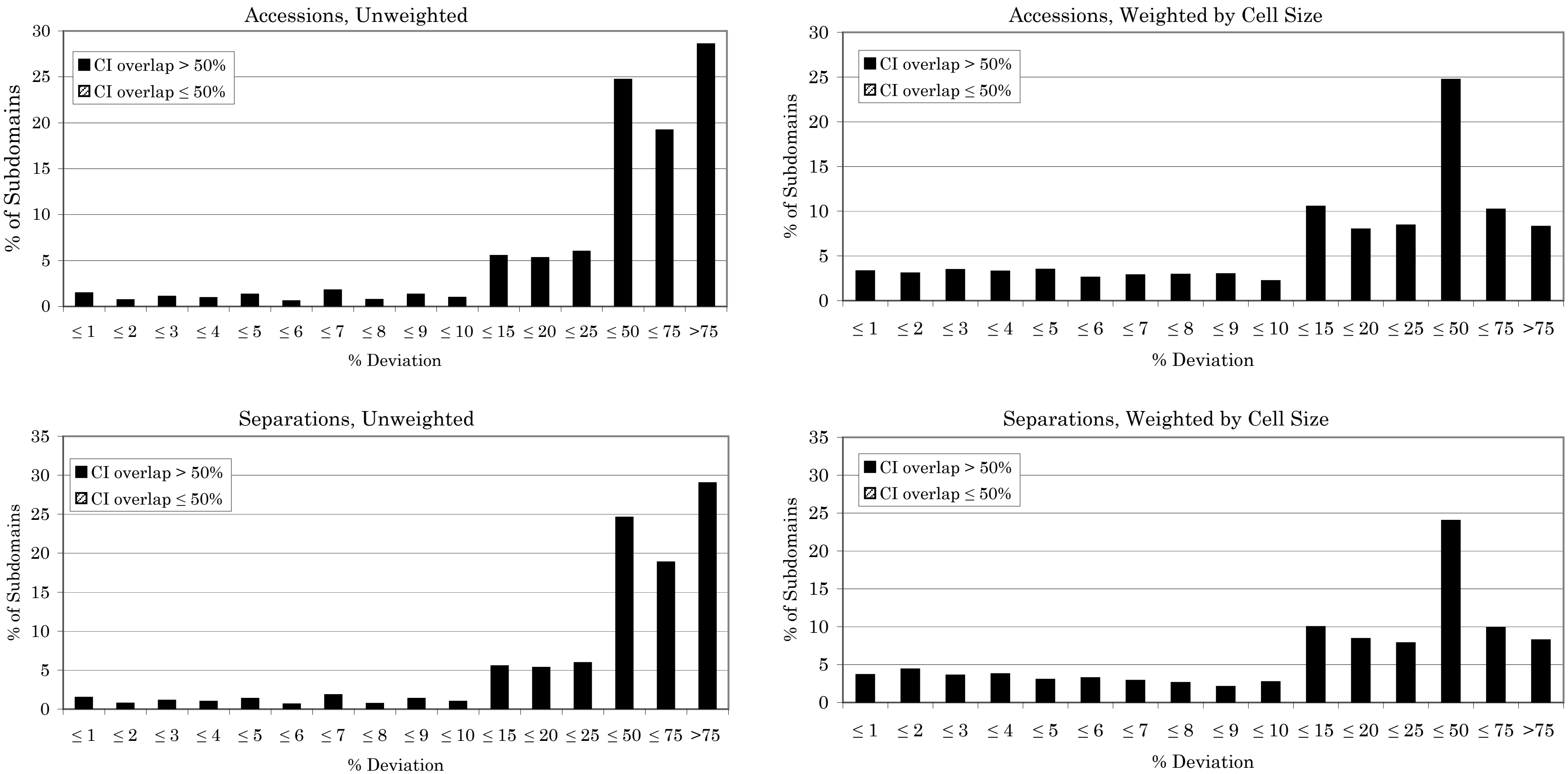
Note: Subdomains are: industry x county; industry; and county.

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



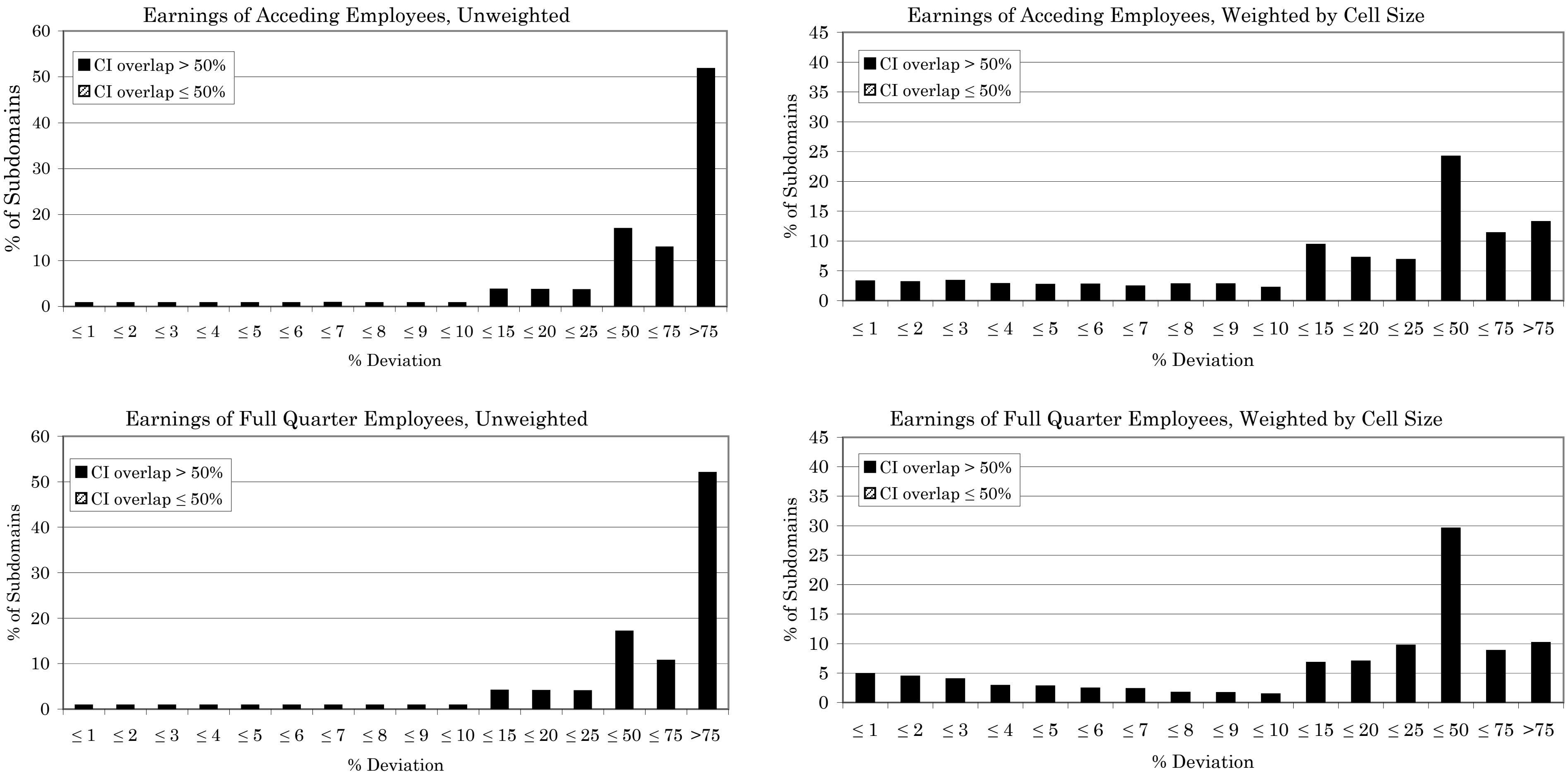
Note: Subdomains are: year x quarter x sex x age category x industry x employer county; year x quarter x employer county x county of residence; year x quarter x sex x race x age category; and year x quarter x industry x employer county.

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.

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

Attribute disclosure risk

We presume an intruder can link records across synthetic implicates.

In most applications, this would be conservative. Here, it is probably realistic.

Because we do not perturb the employment graph, some simple summaries of employment histories are replicated across partially-synthetic implicates

The number of distinct firms at which each individual was employed (R), coupled with the number of distinct employees (E) at each of those firms, 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, uniquely identifies about 80 percent of individuals.

Similar exercise will uniquely identify many firms.

Does this matter for risk of identity disclosure?

A measure of attribute disclosure risk

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: $\bar{y}_{k,i} = \sum_{m=1}^{M} \tilde{y}_{k,i}^{m}$.

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.

A second measure of attribute disclosure risk

Assume the intruder estimates $\bar{y}_{k,i}$ as before, 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.

Idea: predictions are more informative when the interval contains a small proportion of the empirical density (either the interval is narrow, or the prediction lies in a low-density region of the distribution).

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

Strategies to further reduce disclosure risk

Idea: reducing an intruder's ability to combine information across synthetic data implicates reduces risk (attribute and identity).

One possibility: release a sample of observations.

- Unique summaries of the employment graph in a sample do not guarantee uniqueness in the population, so intruder must assign probabilities that records with identical summaries correspond to the same unit.
- Most units will not appear in all samples, so an intruder has fewer implicates on which to base predictions about any unit's confidential values, and hence predictions are less precise.

Another possibility: slightly perturb the employment graph, e.g., multiply-impute the identity of a fraction of individuals' employers.

- Expect a fairly small number of such imputations will introduce enough between-implicate variability to render summaries of the employment graph non-unique.

Conclusion

Overall, our results thus far suggest data utility is quite good and attribute disclosure risk is quite low.

There remains much to do:

- Further assessment of data utility and attribute disclosure risk.
- Assess risk of identity disclosure (via re-identification).
- Results thus far indicate ways to improve the synthesis procedure (e.g., imputation model for earnings should include main effects for industry).
- Take steps to reduce the ability of an intruder to combine information across synthetic data implicates.