LONGITUDINAL EMPLOYER - HOUSEHOLD DYNAMICS

TECHNICAL PAPER NO. TP-2006-02

Confidentiality Protection in the Census Bureaus Quarterly Workforce Indicators

Date : December 5, 2005
Prepared by : John M. Abowd, Bryce E. Stephens, Lars Vilhuber
Contact : U.S. Census Bureau, LEHD Program
FB 2138-3
4700 Silver Hill Rd.
Suitland, MD 20233 USA

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 ongoing research and to encourage discussion of work in progress. 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 Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant 5 R01 AG018854-02, and the Alfred P. Sloan Foundation. The views expressed herein are attributable only to the author(s) and do not represent the views 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. http://lehd.dsd.census.gov.

Confidentiality Protection in the Census Bureau's Quarterly Workforce Indicators

John M. Abowd, Bryce E. Stephens and Lars Vilhuber[1]

December 5, 2005

[1]Abowd: Cornell University, CISER, and U.S. Census Bureau; Stephens: University of Maryland College Park; Vilhuber: Cornell University, CISER. The authors acknowledge the substantial contributions of the staff and senior research fellows of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program. We thank Melissa Bjelland for helpful comments. 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 Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant R01 AG018854-02, and the Alfred P. Sloan Foundation. This document reports the results of research and analysis undertaken by U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This document is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed herein are attributable only to the author(s) and do not represent the views of the U.S. Census Bureau, its program sponsors, Cornell University, 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, Director, 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).

Abstract

The Quarterly Workforce Indicators are new estimates developed by the Census Bureau's Longitudinal Employer-Household Dynamics Program as a part of its Local Employment Dynamics partnership with 37 state Labor Market Information offices. These data provide detailed quarterly statistics on employment, accessions, layoffs, hires, separations, full-quarter employment (and related flows), job creations, job destructions, and earnings (for flow and stock categories of workers). The data are released for NAICS industries (and 4-digit SICs) at the county, workforce investment board, and metropolitan area levels of geography. The confidential microdata – unemployment insurance wage records, ES-202 establishment employment, and Title 13 demographic and economic information – are protected using a permanent multiplicative noise distortion factor. This factor distorts all input sums, counts, differences and ratios. The released statistics are analytically valid – measures are unbiased and time series properties are preserved. The confidentiality protection is manifested in the release of some statistics that are flagged as "significantly distorted to preserve confidentiality." These statistics differ from the undistorted statistics by a significant proportion. Even for the significantly distorted statistics, the data remain analytically valid for time series properties. The released data can be aggregated; however, published aggregates are less distorted than custom postrelease aggregates. In addition to the multiplicative noise distortion, confidentiality protection is provided by the estimation process for the QWIs, which multiply imputes all missing data (including missing establishment, given UI account, in the UI wage record data) and dynamically re-weights the establishment data to provide state-level comparability with the BLS's Quarterly Census of Employment and Wages.

Keywords: Confidentiality ; Noise addition ; Quarterly Workforce Indicators ; Analytic validity ; Multiple imputation

1 Introduction

Disclosure proofing is the set of methods used by statistical agencies to protect the confidentiality of the identity of and information about the individuals and businesses that form the underlying data in the system. It is a critical component of any statistical system which uses confidential data to produce detailed public statistics, without compromising the confidentiality of the original micro-data.

Since 2003, the U.S. Census Bureau has published a new and novel statistical series: the Quarterly Workforce Indicators (QWI). The underlying data infrastructure was designed by the Longitudinal Employer-Household Dynamics Program at the Census Bureau (Abowd et al.; 2004) and is described in detail elsewhere (Abowd et al.; 2005). At its core, the QWI system uses administrative records data collected by a large number of states for both jobs and firms. This administrative database is enhanced with information from other micro-data sets at the Census Bureau. The QWI statistics offer unprecedented demographic and economic detail on the local dynamics of labor markets. Because of the fine detail offered by the published statistics and the confidential nature of the micro-data used to compile the statistics, confidentiality protection is a critical and integral part of the design of the QWI system. Only the application of state-of-the-art protection methods allows the Census Bureau to publish these statistics. In this article, we describe the fundamental components of the confidentiality protection system as it is used in the generation of the QWI. We also show that significant protection is provided by the system, but that the analytic validity of the data remains high. In particular, we provide evidence that the time-series properties of the disclosure-proofed data remain intact, and that the disclosure-proofed data is not biased.

In the QWI system, disclosure proofing is required to protect the information about individuals and businesses that contribute to the unemployment insurance (UI) wage records, the Quarterly Census of Employment and Wages (QCEW, also known as ES-202) reports,[1] and the Census Bureau demographic data that have been integrated with these sources. The primary concern of the confidentiality protection mechanism is thus with small cells, *i.e.*, cells that reflect data on few individuals or few firms.

Methods for protecting such data have been discussed before (*e.g.*, Cox and Zayatz (1993)). In general, data are considered protected when "aggregate cell values do not closely approximate data for any one respondent in the cell" (Cox and Zayatz; 1993, pg. 5). In the QWI confidentiality protection scheme, confidential micro-data are considered protected by noise infusion if one of the following conditions holds: (1) any inference regarding the magnitude of a particular respondent's data must differ from the confidential quantity by at least $c\%$ even if that inference is made by a coalition of respondents with exact knowledge of their own answers, or (2) any inference regarding the magnitude of an item is incorrect with probability no less than $y\%$, where c and y are confidential but generally "large." Condition (1) covers protection of magnitudes like total payroll. Condition (2) covers protection of counts assuming item suppression or some additional protection, like synthetic data, when the count is too small.

These two conditions are met by the multiple layers of confidentiality protection in the QWI

[1] The Quarterly Census of Employment and Wages (QCEW) statistics are published by BLS in cooperation with state Labor Market Information offices.

system. The first layer occurs when job-level estimates are aggregated to the establishment level. A job-level measurement pertains to a given individual at a given workplace. As the job-level estimates are aggregated to the establishment level, the QWI system infuses specially constructed noise into the estimates of all of the workplace-level measures. This noise is designed to have three very important properties. First, every data item is distorted by some minimum amount. After this noise infusion, the distorted data item is used in all the publication QWIs. Second, for a given workplace, the data are always distorted in the same direction (increased or decreased) by the same percentage amount in every period. Third, the statistical properties of this distortion are such that when the estimates are aggregated, the effects of the distortion cancel out for the vast majority of the estimates, preserving both cross-sectional and time-series analytic validity. The use of multiplicative noise infusion, similar to what we develop here, as a cross-sectional confidentiality protection mechanism was first proposed by Evans et al. (1998).

A second layer of confidentiality protection occurs when the workplace-level measures are aggregated to higher levels, *e.g.,* sub-state geography and industry detail. The data from many individuals and establishment are combined into a (relatively) few estimates using a dynamic weight that controls the state-level beginning of quarter employment for all private employers to match the first month in quarter employment as tabulated from the Quarterly Census of Employment and Wages (QCEW). The establishment-level weight is used for every indicator in the QWIs. Hence, an additional difference between the confidential data item and the released data item arises from this weight The weighting procedure, combined with the noise infusion, move the published data away from the value contained in the underlying micro-data, and thus contribute to the protection of the confidentiality of the micro-data.

Third, some of the aggregate estimates turn out to be based on fewer than three persons or establishments. These estimates are suppressed and a flag set to indicate suppression. Suppression is only used when the combination of noise infusion and weighting may not distort the publication data with a high enough probability to meet the criteria layed out above. Count data such as employment are subject to suppression. Continuous dollar measures like payroll are not. All published estimates are still substantially influenced by the noise that was infused in the first layer of the protection system. These distorted estimates are published and flagged as substantially distorted. Each observation on any one of the published QWI tables thus has an associated flag that describes its disclosure status.

The remainder of this article is structured as follows. Section 2 describes the multiplicative noise model, and Section 3 details its integration into the computation of the QWI. The algorithm underlying the item suppression is outlined in Section 4, whereas the computation of weights is shown in Section 5. Sections 6 and 7 provide evidence on the extent of the protection and the analytic validity, respectively. Section 8 concludes.

2 Multiplicative noise model

To implement the multiplicative noise model, a random fuzz factor δ_j is drawn for each establishment j according to the following process:

$$p(\delta_j) = \begin{cases} (b-\delta)/(b-a)^2, & \delta \in [a, b] \\ (b+\delta-2)/(b-a)^2, & \delta \in [2-b, 2-a] \\ 0, & \text{otherwise} \end{cases}$$

$$F(\delta_j) = \begin{cases} 0, & \delta < 2-b \\ (\delta+b-2)^2 / \left[2(b-a)^2\right], & \delta \in [2-b, 2-a] \\ 0.5, & \delta \in (2-a, a) \\ 0.5 + \left[(b-a)^2 - (b-\delta)^2\right], & \delta \in [a, b] \\ 1, & \delta > b \end{cases}$$

where $a = 1 + c/100$ and $b = 1 + d/100$ are constants chosen such that the true value is distorted by a minimum of c percent and a maximum of d percent.[2] Note that $1 < a < b < 2$. This produces a random noise factor centered around 1 with distortion of at least c and at most d percent. The distribution of δ is plotted in Figure 1 on the following page.

A fuzz factor is drawn once for each employer, and for each of the establishments associated with that employer. Although fuzz factors vary across establishments, the fuzz factors attached to all establishments of the *same* employer are drawn from the same (upper or lower) tail of the fuzz factor distribution. Thus, if the fuzz factor associated with a particular employer is less than unity, then all of that employer's establishments will also have fuzz factors less than unity.

It is important to point out that fuzz factors are permanently attached to each employer and establishment and are retained for all time periods and for all revisions of QWI statistics.

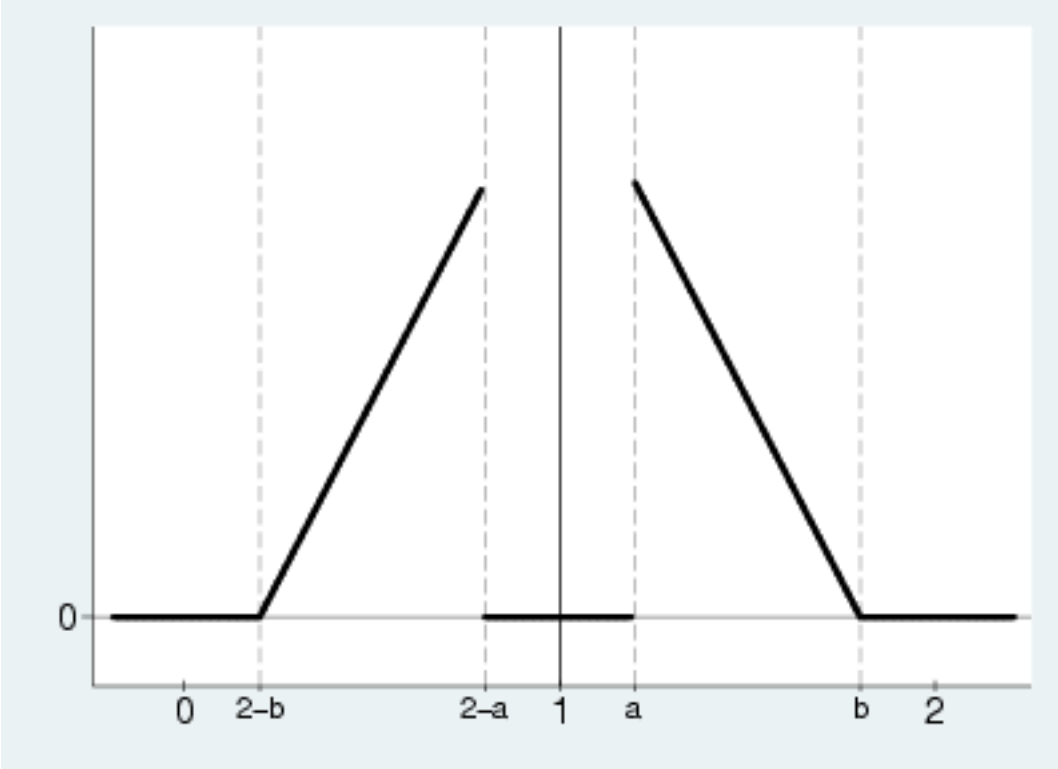
3 Applying the fuzz factors to estimates

Although all estimates are distorted based on the multiplicative noise model, the exact implementation depends on the type of estimate that is computed. A full discussion of how QWI estimates are computed can be found in Abowd et al. (2005), and a list of definitions for the statistics mentioned in this section, and the formulae for their computation is provided in Appendix A on page 23. In all cases, the noise infusion occurs at the level of an establishment estimate. By convention, distorted values are distinguished from their undistorted counterparts by an asterisk; *i.e.*, the true value of beginning-of-quarter employment is B and its distorted counterpart is B^*.

Distorting totals The fuzz factor δ_j is used to distort all establishment totals by scaling of the true establishment level statistic

[2] The exact numbers are confidential.

Figure 1: Distribution of Fuzz Factors



$$X^*_{jt} = \delta_j X_{jt},$$

where X_{jt} is an establishment level statistic among beginning-of-quarter (B), end-of-quarter (E) employment, flow employment (M), full-quarter employment (F), accessions (A), separations (S), new hires (H), recalls (R), flows into full-quarter status (FA), flows out of full-quarter status (FS), new hires into full-quarter status (FH), total payroll (W_1), payroll associated with E (W_2), with B (W_3), with new hires (WFH), periods of non-employment for accessions (NA), for new hires (NH), for recalls (NR), and for separations (NS).

Distorting averages of magnitude variables Averages are constructed from distorted numerators (totals) with undistorted denominators according to

$$ZY^*_{jt} = \frac{Y^*_{jt}}{B(Y)_{jt}} = \delta_j \frac{Y_{jt}}{B(Y)_{jt}},$$

where ZY_{jt} is a statistic related to a total Y_{jt}, and $B(Y)$ is the appropriate denominator for the calculation of the average. Statistics distorted by this method are average earnings for various groups ($ZW_2, ZW_3, ZWFH, ZWA, ZWS$), and average periods of non-employment for several groups (ZNA, ZNH, ZNR, and ZNS).

Distorting differences of counts and magnitudes Distorted net job flow (JF) is computed at the aggregate (k = geography, industry, or combination of the two for the appropriate age and sex categories) level as the product of the aggregated, undistorted rate of growth and the aggregated distorted employment:

$$JF^*_{kt} = G_{kt} \times \bar{E}^*_{kt} = JF_{kt} \times \frac{\bar{E}^*_{kt}}{\bar{E}_{kt}}.$$

This method of distorting net job flow will consistently estimate net job flow because it takes the product of two consistent estimators. The formulas for distorting gross job creation (JC) and job destruction (JD) are similar:

$$JC^*_{kt} = JCR_{kt} \times \bar{E}^*_{kt} = JC_{kt} \times \frac{\bar{E}^*_{kt}}{\bar{E}_{kt}}$$

and

$$JD^*_{kt} = JDR_{kt} \times \bar{E}^*_{kt} = JD_{kt} \times \frac{\bar{E}^*_{kt}}{\bar{E}_{kt}}.$$

where JCR_{kt} and JDR_{kt} are the aggregated growth rates for job creations and destructions, respectively. Exactly analogous expressions apply to full-quarter net job flows (FJF), full-quarter job creations (FJC) , and full-quarter job destructions (FJD) .

The same logic was used to distort wage changes for subgroups (accessions, separations, full-quarter accessions and separations). The undistorted total changes were divided by the undistorted

denominators then multiplied by the ratio of the distorted denominator to the undistorted denominator for the computation of average change in earnings. Averages are distorted by multiplying by the ratio of the distorted denominator to the true denominator. For example:

$$Z\Delta WY^*_{kt} = \frac{\Delta WY_{kt}}{Y_{kt}} \times \frac{Y^*_{kt}}{Y_{kt}}.$$

where, again, Y denotes a particular count, and $Z\Delta WY$ the average change in total earnings associated with that particular count.

4 Item suppression

Despite the noise infusion described in the previous sections, some disclosure risk remains for counts based on very few entities in a cell. For counts based on data from fewer than three individuals or employers, the fuzz factors may not provide sufficient protection. This condition applies to the variables B, E, M, F, A, S, H, R, FA, FH, FS, JC, JD, JF, FJC, FJD, FJF. The QWIs therefore also implement item suppression based on the number of either workers or the number of employers that contribute data for that item in a cell k in time period t, where a cell represents a particular combination of geography $\times$ industry $\times$ age $\times$ sex. Because of the noise infusion used previously, however, no complementary suppressions are needed since all of the values based on three or more individuals or employers are adequately protected. Any estimate of the suppressed item computed by subtraction is also protected.

The algorithm for item suppression for these variables is as follows:

- Check the conditions leading to a disclosure flag of -2 or -1 (data availability). If met, set the item to missing in the release file.
- Determine whether the value can be computed according to Census standards:
  - For the variables JC, JD, and JF, (respectively, $FJC, FJD,$ and FJF) check whether the denominator average employment ($\bar{E}_{kt}$; respectively, $\bar{F}_{kt}$) in the relevant cell kt rounds to zero.
  - Check whether the item in cell kt rounds to zero.
  - Check whether the data used to construct the cell kt value were based on 1 or 2 individuals.
  - Check whether the data used to construct the cell kt value were based on 1 or 2 employers.

  If any of these conditions are met, set the disclosure status to 5 and set the item to missing in the release file.

- Check whether the distortion of cell kt value exceeds the limit set by the Census Disclosure Review Board[3]. If so, set the disclosure status to 9 and copy the distorted value to the release file.
- Otherwise, set the disclosure status to 1 and copy the distorted value to the release file.

Table 1 lists all possible flag values.

Table 1: Disclosure flags in the QWI

Flag	Explanation
-2	no data available in this category for this quarter
-1	data not available to compute this estimate
0	no employment in this cell, or no positive denominator (OK to disclose a 0 for sum or count, missing for ratio)
1	OK, distorted value released
5	Value suppressed because it does not meet US Census Bureau publication standards.
9	data significantly distorted, distorted value released

5 Weighting the QWI

The economic concepts underlying the Quarterly Census of Employment and Wages (QCEW) statistics, published by BLS in cooperation with state Labor Market Information offices, and the QWI statistics, published by the U.S. Census Bureau, are similar, but not identical. While the QCEW reports employment on the 12th day of the month, for all months, as reported by employers for each establishment, the QWI has several measures of employment, all of which are derived from reports of quarterly employment and wages of individual workers at particular employers (state UI accounts). In particular, flow employment can be distinguished from point-in-time measures. Flow employment M_{jt} is defined as a simple count of employees who had positive, UI covered earnings and any time during quarter t at establishment j. Beginning of quarter employment B_{jt}, on the other hand, counts the number of employees present at establishment j in both quarter t and $t-1$, and by inference, on the 1st day of quarter t. By definition, flow employment will be higher than any point-in-time measure. The point-in-time measures in the QCEW and the QWI are comparable, and in particular, the QCEW report for employment on the 12th of the first month of a quarter (January, April, July, October) is comparable but not identical to the QWI measure of B.

These two measures are not identical because (a) they do not refer to exactly the same point in time, (b) the in-scope establishments differ slightly, and (c) they are computed from different

[3]The precise value is confidential.

universe data. The actual differences between these two measures are modeled and captured by the weighting scheme used in the QWI. To be precise, denote by $QCEW_{1,jt}$ the measured employment for the 12th of the first month on the QCEW report for establishment j in quarter t and let w_t denote the (state-specific) weight. Then the time-series of adjustment weights are defined by

$$w_t \sum_j b_{jt} = \sum_j QCEW_{1,jt} \tag{1}$$

for each time period t.

The weighting is not used to control to sub-state geography and industry because the characteristics of multi-unit employing establishments are multiply-imputed in the QWI data. Due to the way in which the UI wage records are collected at the state agencies, establishment identifiers are missing for multi-units. In the QCEW, sub-state geography and industry are coded directly at the establishment level.

The fact that workplace characteristics of geography and industry are multiply-imputed for multi-unit employers also has confidentiality protection implications. The establishment-level QWI micro-data for these entities were not provided by the responding firm (a UI account). Hence, there are no actual confidential micro-data measured at the establishment level. In effect, these establishments are protected by a form of synthetic data.

6 Extent of protection

The extent of the protection of the QWI micro-data can be measured by how many counts differ from their true values. The percentage deviation is a measure of the uncertainty about the true value that one can infer from the released value. The following tables show a series of comparisons designed to emphasize the contribution of each component of the QWI confidentiality protection mechanisms to the uncertainty about the true value. The contributions of weighting and noise-infusion can be separated by first comparing the undistorted, unweighted data with the undistorted, weighted data (Table 2), thus tabulating the number of cells that diverge from their true value solely due to weighting. The undistorted, weighted data are then compared to the distorted and weighted data (Table 3), highlighting the contribution of the noise infusion. Finally, a comparison of the undistorted, unweighted data with the published data (Table 4) brings out the combined contribution of weighting, noise infusion, and item suppression.

The tables display the row percentages and may be interpreted as the conditional probability of reporting the column entry given the row entry. A prominent feature of Tables 2 and 3 is the strong weight of the diagonal. The vast majority of cells is left unchanged by either noise infusion or weighting. Nevertheless, both weighting and noise infusion do affect significant number of cells. The changed cells in Table 2 are more likely to be found above the diagonal, demonstrating that the raw job-level wage records in the QWI system generally estimate lower beginning-of-quarter employment than month-one employment in the published establishment-record-based statistics in the QCEW. The changed cells in Table 3 are more symmetrically aligned around the diagonal, reflecting the symmetry of the noise distribution used to distort the data.

Table 2: Small Cells: B, Unweighted vs. Weighted

(a) Illinois

	Weighted count					
Unweighted count	0	1	2	3	4	5 or more
0	99.33	0.66	0.00	0.00	0.00	0.00
1	0.10	96.76	3.13	0.00	0.00	0.00
2	0.01	2.00	84.68	13.26	0.04	0.01
3	0.01	0.01	3.42	75.72	20.26	0.59
4	0.00	0.00	0.01	4.49	67.62	27.87
5 or more	0.00	0.00	0.00	0.01	0.59	99.39

Total number of cells: 14,229,968 . For details, see text.

(b) Maryland

	Weighted count					
Unweighted count	0	1	2	3	4	5 or more
0	99.10	0.90	0.00	0.00	0.00	0.00
1	0.11	94.36	5.52	0.00	0.00	0.00
2	0.04	0.53	73.83	25.45	0.13	0.02
3	0.03	0.03	1.42	55.47	41.79	1.25
4	0.02	0.02	0.04	1.85	41.39	56.69
5 or more	0.01	0.01	0.01	0.02	0.21	99.75

Total number of cells: 4,659,408 . For details, see text.

Table 3: Small Cells: B, Undistorted vs. Distorted

(a) Illinois

	Distorted count					
Undistorted count	0	1	2	3	4	5 or more
0	99.86	0.14	0.00	0.00	0.00	0.00
1	0.91	95.75	3.34	0.00	0.00	0.00
2	0.00	4.27	87.25	8.47	0.00	0.00
3	0.00	0.00	10.69	77.20	12.11	0.00
4	0.00	0.00	0.00	14.73	67.49	17.78
5 or more	0.00	0.00	0.00	0.00	1.93	98.07

Total number of cells: 14,229,968 . Both comparisons are for weighted data. For details, see text.

(b) Maryland

	Distorted count					
Weighted count	0	1	2	3	4	5 or more
0	99.83	0.17	0.00	0.00	0.00	0.00
1	0.73	92.35	6.91	0.00	0.00	0.00
2	0.00	5.07	80.45	14.48	0.00	0.00
3	0.00	0.00	12.51	71.21	16.27	0.00
4	0.00	0.00	0.00	17.62	65.74	16.63
5 or more	0.00	0.00	0.00	0.00	1.68	98.32

Total number of cells: 4,659,408 . For details, see text.

Table 4: Small Cells: B, Raw vs. Published

(a) Illinois

	Published count						
Unweighted count	Suppressed	0	1	2	3	4	5 or more
0	0.79	99.21	0.00	0.00	0.00	0.00	0.00
1	99.91	0.08	0.00	0.00	0.00	0.00	0.00
2	94.02	0.01	0.00	0.00	5.87	0.09	0.01
3	34.33	0.00	0.00	0.00	47.75	16.98	0.94
4	25.87	0.00	0.00	0.00	5.56	43.24	25.32
5 or more	15.20	0.00	0.00	0.00	0.03	0.82	83.95

Total number of cells: 14,229,968 . Raw is unweighted and undistorted. Published is after weighting, distorting, and suppression. For details, see text.

(b) Maryland

	Published count						
Unweighted count	Suppressed	0	1	2	3	4	5 or more
0	1.06	98.94	0.00	0.00	0.00	0.00	0.00
1	99.90	0.09	0.00	0.00	0.00	0.00	0.00
2	85.71	0.04	0.00	0.00	13.90	0.32	0.02
3	23.54	0.03	0.00	0.00	40.18	33.60	2.65
4	18.06	0.02	0.00	0.00	2.22	33.67	46.04
5 or more	8.44	0.01	0.00	0.00	0.02	0.26	91.26

Total number of cells: 4,659,408 . For details, see text.

Table 4 shows the amount of suppression after weighting and noise-infusion as it relates to the original raw value. Note that all single-individual cells have been suppressed. This is not true for two-person cells, some of which have a weighted value that lies above the suppression threshold causing the weighted distorted estimate to be released. The converse is true for cells with three individuals. Due to weighting, some of these cells have weighted, undistorted values that lie below the suppression threshold, and are consequently suppressed. While not explicitly detailed in these tables cells that contain count data based on fewer than three firms also generate suppressions, which are included in the suppression totals. Given the information in Tables 2 and 3, almost no cells with 4 or more individuals in the raw data have distorted and weighted data below 3 (a jump of two columns). Thus, for these cells, all suppressions are due to a small number of firms in a cell, or one of the other suppression conditions listed in Table 1. Overall, at the level of detail analyzed here (SIC3 $\times$ county $\times$ time $\times$ sex $\times$ age), around 25% of the beginning of period employment cells are suppressed in both the states analyzed here. For more aggregate tabulations, for instance at the SIC Division level, that percentage falls to between 5% and 10%.

Because total payroll, the other variable considered in detail in this paper, is a total (magnitude), not a count, it is never suppressed. The combination of weighting and distorting is sufficient to protect the confidentiality of this item without suppression because if the item is based on a single person or establishment, then the minimum distortion of the underlying micro-data applies. If the item is based on 2 employers or establishments then both micro-data items have been distorted at least the minimum percentage. Knowledge of one's own value does not help in inferring another's value because both data items were distorted in an unknown direction by an unknown minimum percentage. Even an accurate inference about one's own distortion factor supplies no information about the other parties distortion factor, thus protecting that item by at least the minimum distortion factor in each direction.

7 Analytic validity

The noise infusion described in Section 2 is designed to preserve the analytic validity of the data. In order to demonstrate how successfully this validity has been preserved, we provide in this section evidence on the time-series properties of the distorted data, as well as evidence on the cross-sectional unbiasedness of the published data. In each case, we used data from Illinois and Maryland. We concentrate on two estimates, beginning-of-quarter employment B, and total payroll W_1. The unit of analysis is an interior sub-state geography $\times$ industry $\times$ age $\times$ sex cell kt. Sub-state geography in all cases is a county, whereas the industry classification is SIC. For our purposes, analytic validity is obtained when the data display no bias and the additional dispersion due to the confidentiality protection system that can be quantified so that statistical inferences can be adjusted to accomodate it.

7.1 Time-series properties of distorted data

To analyze the impact on the time series properties of the distorted data, we estimated an AR(1) for the time series associated with each cell kt, using county-level data for all Illinois and Maryland

counties. Two AR(1) coefficients are estimated for each cell-time series. The first order serial correlation coefficient computed using undistorted data is denoted by r. The estimate computed using the distorted data is denoted by r^*. For each cell, the error $\Delta r = r - r^*$ is computed. Table 5 on the following page shows the distribution of the errors Δr across SIC-division $\times$ county cells, for B, A, S, F, and JF when comparing raw (confidential) data to distorted data, whereas Table 6 on page 15 compares the same variables between the raw and the published data, which excludes suppressed data items.

The tables show that the time series properties of all variables analyzed remain largely unaffected by the distortion. The maximum bias (as measured by the median of this distribution) is never greater than 0.001 (raw v. distorted or raw v. published). The error distribution is tight: the semi-interquartile range of the distortion for B in Maryland is 0.010, which is less than the precision with which estimated serial correlation coefficients are normally displayed. The maximum semi-interquartile range for any variable in any one of the two states is 0.012[4].

The distribution of errors is similar when considering raw versus published data (Table 6). Furthermore, although the overall spread of the distribution is slightly higher when considering two-digit SIC $\times$ county and three-digit SIC $\times$ county cells, which are sparser than the SIC-division $\times$ county cells, the general results hold there as well. We conclude that the time series properties of the QWI data are unbiased with very little additional noise, which is, in general, economically meaningless.

[4]The maximum semi-interquartile range for SIC2-based variables is 0.0241, and for SIC3-based variables, 0.0244.

Table 5: Distribution of the Error in the First Order Serial Correlation: SIC-division × County, Raw vs. Distorted Data

$$\Delta r = r - r^*$$

Percentile	Beginning of Quarter Employment	Accessions	Separations	Full Quarter Employment	Net Job Flows
			IL SIC Division		
01	-0.069373	-0.049274	-0.052155	-0.066461	-0.007969
05	-0.041585	-0.031460	-0.032934	-0.039787	-0.004651
10	-0.028849	-0.022166	-0.023733	-0.027926	-0.002785
25	-0.011920	-0.009996	-0.010161	-0.011913	-0.001003
50	0.000571	0.000384	0.000768	0.000306	-0.000044
75	0.013974	0.011806	0.012891	0.012632	0.000776
90	0.030948	0.025152	0.026290	0.028299	0.002263
95	0.044233	0.033871	0.037198	0.040565	0.004375
99	0.078519	0.054415	0.060327	0.074212	0.007845
			MD SIC Division		
01	-0.059390	-0.050060	-0.049160	-0.048983	-0.010339
05	-0.032436	-0.030694	-0.030720	-0.028823	-0.004482
10	-0.022176	-0.023042	-0.023525	-0.018979	-0.002589
25	-0.009125	-0.010831	-0.010199	-0.007936	-0.001161
50	0.000658	0.000726	0.001123	0.000788	-0.000073
75	0.011639	0.012500	0.012871	0.010200	0.001044
90	0.024883	0.024917	0.024511	0.022358	0.002256
95	0.035014	0.033517	0.033028	0.030864	0.003699
99	0.059709	0.049903	0.050689	0.047204	0.008619

Unit of observation is a cell. Industry aggregation is SIC Division, geography aggregated to county level. For more details, see text.

Table 6: Distribution of the Error in the First Order Serial Correlation: County x SIC-division × County, Raw vs. Published Data

$$\Delta r = r - r^*$$

Percentile	Beginning of Quarter Employment	Accessions	Separations	Full Quarter Employment	Net Job Flows
			IL County x SIC Division		
01	-0.085495	-0.092455	-0.098770	-0.079205	-0.008447
05	-0.047704	-0.046665	-0.045208	-0.046830	-0.004959
10	-0.034558	-0.031767	-0.032898	-0.033607	-0.003186
25	-0.015317	-0.014197	-0.015077	-0.015533	-0.001189
50	-0.000512	-0.000997	-0.000707	-0.001000	-0.000049
75	0.013438	0.011536	0.012457	0.011670	0.000861
90	0.030963	0.027037	0.028835	0.027970	0.002489
95	0.044796	0.037906	0.041862	0.040096	0.004801
99	0.080282	0.079122	0.083824	0.077419	0.007537
			MD County x SIC Division		
01	-0.065342	-0.072899	-0.072959	-0.058021	-0.009081
05	-0.035974	-0.036995	-0.040314	-0.030985	-0.004540
10	-0.024174	-0.027689	-0.028577	-0.021361	-0.002823
25	-0.010393	-0.013686	-0.012505	-0.009401	-0.001243
50	0.000230	-0.000542	0.000797	0.000279	-0.000025
75	0.011382	0.012628	0.013034	0.009429	0.001045
90	0.025160	0.026325	0.025272	0.022027	0.002799
95	0.035176	0.034114	0.034999	0.030152	0.004321
99	0.060042	0.056477	0.055043	0.049213	0.009208

Unit of observation is a cell. Industry aggregation is SIC Division, geography aggregated to county level. For more details, see text.

Table 7: Distribution of the Error in the First Order Serial Correlation: Two-digit SIC $\times$ County, Raw vs. Distorted Data

$$\Delta r = r - r^*$$

Percentile	Beginning of Quarter Employment	Accessions	Separations	Full Quarter Employment	Net Job Flows
			IL SIC2		
01	-0.070671	-0.052107	-0.057965	-0.068505	-0.017139
05	-0.039739	-0.033252	-0.035271	-0.036607	-0.006337
10	-0.026348	-0.023354	-0.024951	-0.024729	-0.003599
25	-0.009891	-0.010622	-0.010718	-0.009530	-0.001238
50	0.000333	-0.000023	0.000675	0.000212	0.000003
75	0.012089	0.010960	0.013107	0.011015	0.001185
90	0.029082	0.025055	0.028222	0.026441	0.003455
95	0.042054	0.034896	0.038768	0.039589	0.005497
99	0.077996	0.058780	0.065105	0.072694	0.011871
			MD SIC2		
01	-0.056975	-0.055872	-0.057173	-0.049496	-0.014149
05	-0.033605	-0.035727	-0.037286	-0.029605	-0.006805
10	-0.023911	-0.025826	-0.027422	-0.020951	-0.003828
25	-0.009977	-0.011753	-0.012791	-0.008451	-0.001427
50	0.000075	0.000332	-0.000282	0.000140	0.000082
75	0.010242	0.012439	0.011353	0.008987	0.001532
90	0.024432	0.026786	0.025800	0.021818	0.004062
95	0.035468	0.035693	0.035284	0.031619	0.006035
99	0.061907	0.055054	0.055839	0.054744	0.011731

Unit of observation is a cell. Industry aggregation is SIC2, geography aggregated to county level. For more details, see text.

Table 8: Distribution of the Error in the First Order Serial Correlation: Two-digit SIC $\times$ County, Raw vs. Published Data

$$\Delta r = r - r^*$$

Percentile	Beginning of Quarter Employment	Accessions	Separations	Full Quarter Employment	Net Job Flows
			IL SIC2		
01	-0.129094	-0.104500	-0.102003	-0.123819	-0.019439
05	-0.056734	-0.054465	-0.054423	-0.054914	-0.006630
10	-0.038474	-0.037901	-0.036443	-0.036726	-0.004058
25	-0.016431	-0.016847	-0.016628	-0.016082	-0.001277
50	-0.001610	-0.002131	-0.000789	-0.001742	0.000022
75	0.011486	0.011319	0.013833	0.010231	0.001235
90	0.029364	0.027751	0.031744	0.026192	0.003639
95	0.043912	0.039888	0.046670	0.040161	0.005915
99	0.082596	0.079321	0.098374	0.076498	0.014536
			MD SIC2		
01	-0.101585	-0.091941	-0.096422	-0.105893	-0.016338
05	-0.049849	-0.049707	-0.053894	-0.043979	-0.007201
10	-0.032742	-0.035509	-0.038168	-0.030164	-0.004159
25	-0.015218	-0.017011	-0.018759	-0.013736	-0.001780
50	-0.001978	-0.001817	-0.002780	-0.001532	0.000024
75	0.009548	0.013094	0.011995	0.008193	0.001590
90	0.024396	0.029727	0.028478	0.021555	0.004398
95	0.035172	0.041838	0.042422	0.032194	0.006325
99	0.065299	0.097201	0.105719	0.057076	0.012864

Unit of observation is a cell. Industry aggregation is SIC2, geography aggregated to county level. For more details, see text.

7.2 Cross-sectional unbiasedness of the distorted data

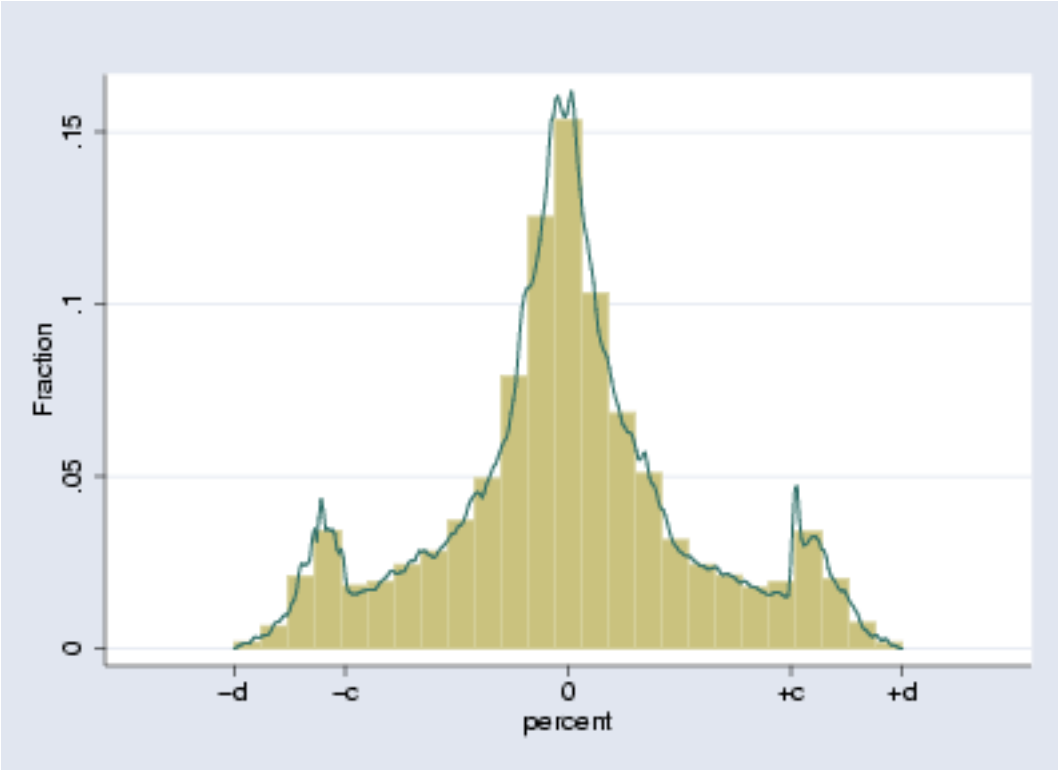
The distribution of the infused noise is symmetric, and allocation of the fuzz factors is random. The data distribution resulting from the noise infusion should thus be unbiased. Evidence of unbiasedness is provided by Figures 2 and 3. Each graph shows, for the states of Illinois (a) and Maryland (b) and a variable X, the distribution of the bias ΔX in each cell kt, expressed in percentage terms:

$$\Delta X_{kt} = \frac{X^*_{kt} - X_{kt}}{X_{kt}} \times 100 \tag{2}$$

where X is B or W_1. All histograms are weighted by B_{kt}. Industry classification is three-digit SIC (industry groups).

Both the distribution of ΔB and ΔW_1 have most mass around the mode at zero percent. Also, as is to be expected, both present secondary spikes around $\pm c$, the inner bound of the noise distribution.

(a) Illinois



(b) Maryland

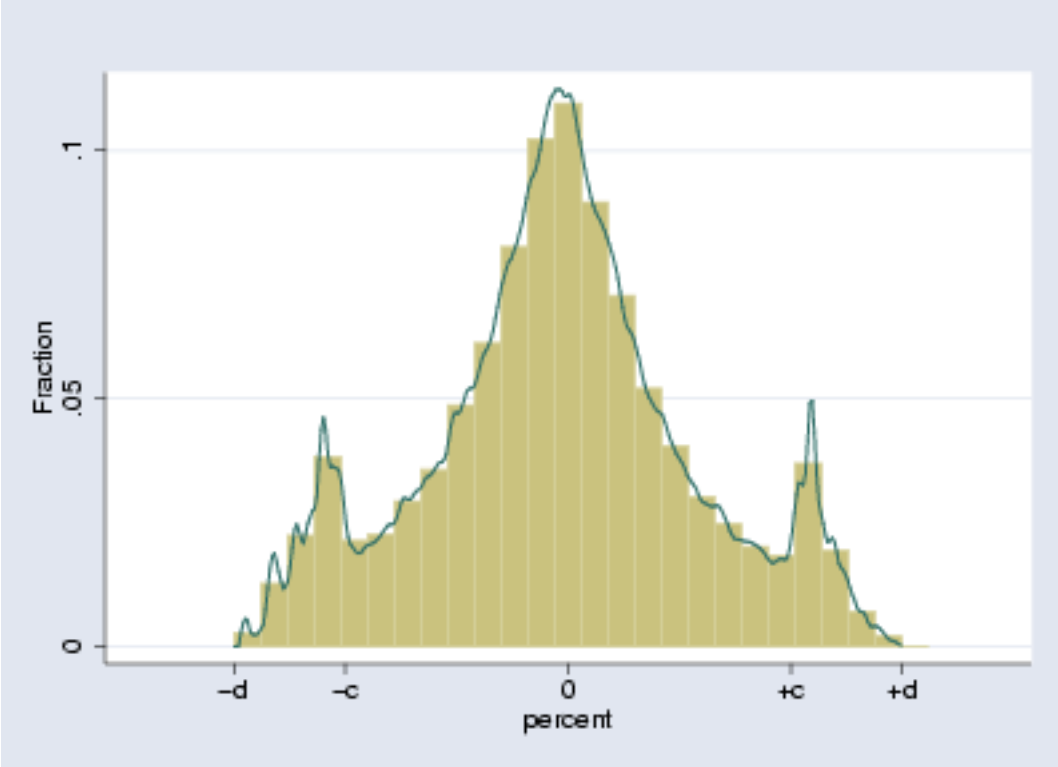
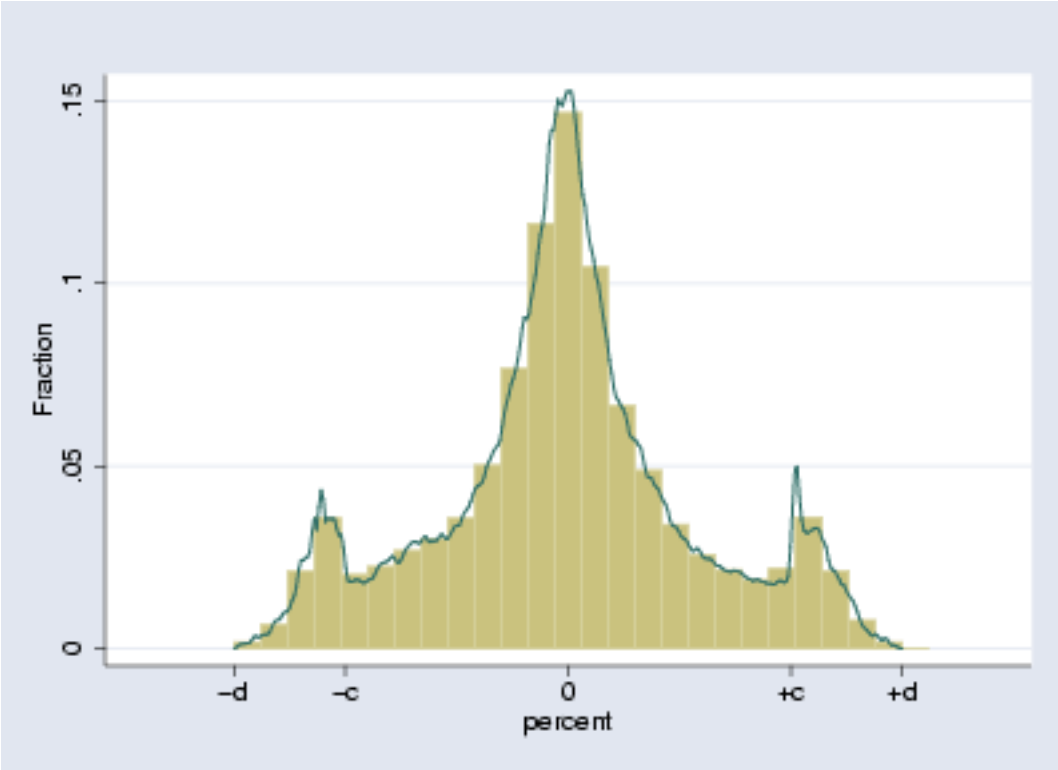


Figure 2: Distribution of Noise: B

(a) Illinois



(b) Maryland

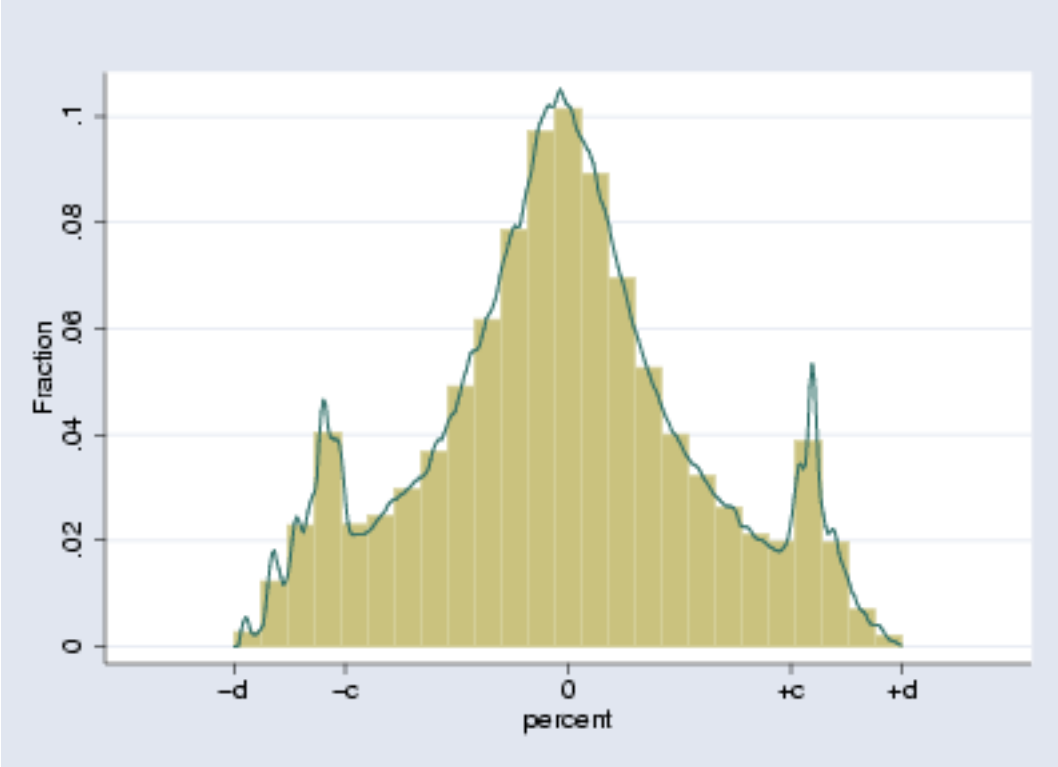


Figure 3: Distribution of Noise: W_1

8 Concluding remarks

In this paper, we provided a description of the confidentiality protection mechanism used in the generation of the Quarterly Workforce Indicators (QWIs). A notable feature of this disclosure proofing mechanism is the absence of table-level or complementary suppressions. Thus, although a significant number of count item values are indeed suppressed, the vast majority of counts are releasable data. All ratios and sums are released without suppression. To our knowledge, this is the first large-scale implementation of confidentiality protection by noise infusion.

Results from a comparison of the time-series characteristics of the undistorted and the distorted data shows remarkable consistency in the serial correlation coefficients between the two series at highly detailed levels. Furthermore, there is little or no bias in induced on average by the confidentiality protection mechanism, and the distributions of bias are tightly centered around the modal bias of zero.

Bibliography

Abowd, J. M., Andersson, F., McKinney, K. L., Roemer, M., Stephens, B. E., Woodcock, S. and Vilhuber, L. (2005). The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators, *mimeo*, U.S. Census Bureau, LEHD and Cornell University.

Abowd, J. M., Haltiwanger, J. C. and Lane, J. I. (2004). Integrated longitudinal employee-employer data for the United States, *American Economic Review* **94**(2).

Cox, L. H. and Zayatz, L. V. (1993). An agenda for research in statistical disclosure limitation, *Statistical Research Report Series LVZ93/01*, U.S. Census Bureau.

Evans, T., Zayatz, L. and Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data, *Journal of Official Statistics* .

A Definitions of fundamental LEHD concepts

We briefly explain some of the basic concepts underlying QWI processing, and indeed, much of the LEHD Infrastructure. A more exhaustive list of definitions can be found on the LEHD website at http://lehd.dsd.census.gov.

A.1 Fundamental Concepts

A.1.1 Dates

The QWI is a quarterly data system with calendar year timing. We use the notation YYYY:Q to refer to a year and quarter combination. For example, 1999:4 refers to the fourth quarter of 1999, which includes the months October, November, and December.

A.1.2 Employer

An employer in the QWI system consists of a single Unemployment Insurance (UI) account in a given state's UI wage reporting system. For statistical purposes the QWI system creates an employer identifier called an State Employer Identification Number (SEIN) from the UI-account number and information about the state (FIPS code). Thus, within the QWI system, the SEIN is a unique identifier within and across states but the entity to which it refers is a UI account.

A.1.3 Establishment

For a given employer in the QWI system, an SEIN, each physical location within the state is assigned a unit number, called the SEINUNIT. This SEINUNIT is based on the reporting unit in the ES-202 files supplied by the states. All QWI statistics are produced by aggregating statistics calculated at the establishment level. Single-unit SEINs are UI accounts associated with a single reporting unit in the state. Thus, single-unit SEINs have only one associated SEINUNIT in every quarter. Multi-unit SEINs have two or more SEINUNITS associated for some quarters. Since the UI wage records are not coded down to the SEINUNIT, SEINUNITs are multiply imputed as described in Abowd et al. (2005). A feature of this imputation system is that it does not permit SEINUNIT to SEINUNIT movements within the same SEIN. Thus, for multi-unit SEINs, the definitions below produce the same flow estimates at the SEIN level whether the definition is applied to the SEIN or the SEINUNIT.

A.1.4 Employee

Individual employees are identified by their Social Security Numbers (SSN) on the UI wage records that provide the input to the QWI. To protect the privacy of the SSN and the individual's name, a different branch of the Census Bureau removes the name and replaces the SSN with an internal Census identifier called a Protected Identity Key (PIK).

A.1.5 Job

The QWI system definition of a job is the association of an individual (PIK) with an establishment (SEINUNIT) in a given year and quarter. The QWI system stores the entire history of every job that an individual holds. Estimates are based on the definitions presented below, which formalize how the QWI system estimates the start of a job (accession), employment status (beginning- and end-of-quarter employment), continuous employment (full-quarter employment), the end of a job (separation), and average earnings for different groups.

A.1.6 Unemployment Insurance wage records (the QWI system universe)

The Quarterly Workforce Indicators are built upon concepts that begin with the report of an individual's UI-covered earnings by an employing entity (SEIN). An individual's UI wage record enters the QWI system if at least one employer reports earnings of at least one dollar for that individual (PIK) during the quarter. Thus, the job must produce at least one dollar of UI-covered earnings during a given quarter to count in the QWI system. The presence of this valid UI wage record in the QWI system triggers the beginning of calculations that estimate whether that individual was employed at the beginning of the quarter, at the end of the quarter, and continuously throughout the quarter. These designations are discussed below. Once these point-in-time employment measures have been estimated for the individual, further analysis of the individual's wage records results in estimates of full-quarter employment, accessions, separations (point-in-time and full-quarter), job creations and destructions, and a variety of full-quarter average earnings measures.

A.1.7 Employment at a point in time

Employment is estimated at two points in time during the quarter, corresponding to the first and last calendar days. An individual is defined as employed at the beginning of the quarter when that individual has valid UI wage records for the current quarter and the preceding quarter. Both records must apply to the same employer (SEIN). An individual is defined as employed at the end of the quarter when that individual has valid UI wage records for the current quarter and the subsequent quarter. Again, both records must show the same employer. The QWI system uses beginning and end of quarter employment as the basis for constructing worker and job flows. In addition, these measures are used to check the external consistency of the data, since a variety of employment estimates are available as point-in-time measures. Many federal statistics are based upon estimates of employment as of the 12th day of particular months. The Census Bureau uses March 12 as the reference date for employment measures contained in its Business Register and on the Economic Censuses and Surveys. The BLS Quarterly Census of Employment and Wages (QCEW)[5] series, which is based on the ES-202 data, use the 12th of each month as the reference date for employment. The QWI system cannot use exactly the same reference date as these other systems because UI wage reports do not specify additional detail regarding the timing of the wage payments. QWI research has shown that the point-in-time definitions used to estimate beginning and end of quarter employment track the QCEW month one employment estimates well at the

[5]The QCEW were formerly known as Covered Employment and Wages (CEW).

level of an employer (SEIN). For single-unit SEINs, there is no difference between an employer-based definition and an establishment-based definition of point-in-time employment. For multi-unit SEINs, the unit-to-worker imputation model assumes that unit-to-unit transitions within the same SEIN cannot occur. So, point in time employment defined at either the SEIN or SEINUNIT level produces the same result.

A.1.8 Employment for a full quarter

The concept of full quarter employment estimates individuals who are likely to have been continuously employed throughout the quarter at a given employer. An individual is defined as full-quarter-employed if that individual has valid UI-wage records in the current quarter, the preceding quarter, and the subsequent quarter at the same employer (SEIN). That is, in terms of the point-in-time definitions, if the individual is employed at the same employer at both the beginning and end of the quarter, then the individual is considered full-quarter employed in the QWI system.

Consider the following example. Suppose that an individual has valid UI wage records at employer *A* in 1999:2, 1999:3, and 1999:4. This individual does not have a valid UI wage record at employer *A* in 1999:1 or 2000:1. Then, according to the definitions above, the individual is employed at the end of 1999:2, the beginning and end of 1999:3, and the beginning of 1999:4 at employer *A*. The QWI system treats this individual as a full-quarter employee in 1999:3 but not in 1999:2 or 1999:4. Full-quarter status is not defined for either the first or last quarter of available data.

A.1.9 Point-in-time estimates of accession and separation

An accession occurs in the QWI system when it encounters the first valid UI wage record for a job (an individual (PIK)-employer (SEINUNIT) pair). Accessions are not defined for the first quarter of available data from a given state. The QWI definition of an accession can be interpreted as an estimate of the number of new employees added to the payroll of the establishment (SEINUNIT) during the quarter. The individuals who acceded to a particular employer were not employed by that employer during the previous quarter but received at least one dollar of UI-covered earnings during the quarter of accession.

A separation occurs in the current quarter of the QWI system when it encounters no valid UI wage record for an individual-employer pair in the subsequent quarter. This definition of separation can be interpreted as an estimate of the number of employees who left the employer during the current quarter. These individuals received UI-covered earnings during the current quarter but did not receive any UI-covered earnings in the next quarter from this employer. Separations are not defined for the last quarter of available data.

A.1.10 Individual concepts

The variable t refers to a sequential quarter.$qfirst$ refers to the first available sequential quarter of data for a state, and $qlast$ to the last available sequential quarter of data for a state. Unless otherwise specified a variable is defined for $qfirst \leq t \leq qlast$. Demographic characteristics are

defined as an $d \in D$, where D is defined as the union of all measured demographic dimensions. With a slight abuse of notation, we will treat d as designating a set of individuals that have a particular set of demographic characteristics and note membership in that set as $i \in d$. At present, D only has two dimensions, age and gender, but others are possible. Equivalently, establishment characteristics are defined as $k \in K$, and a firm with such characteristics is noted as $j \in k$. Again, K is currently two-dimensional, defined in terms of geography and industry (in this article, county and SIC-based, respectively).

Flow employment (m): for $qfirst \leq t \leq qlast$, individual i employed (matched to a job) at some time during period t at establishment j

$$m_{ijt} = \begin{cases} 1, \text{ if } i \text{ has positive earnings at establishment } j \text{ during quarter } t \\ 0, \text{ otherwise.} \end{cases} \tag{3}$$

Beginning-of-quarter employment (b): For $qfirst < t$, individual i employed at the end of $t-1$, beginning of t

$$b_{ijt} = \begin{cases} 1, \text{ if } m_{ijt-1} = m_{ijt} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{4}$$

Equivalently, *end-of-quarter employment* can be defined.

Accessions (a): For $qfirst < t$, individual i acceded to j during t

$$a_{ijt} = \begin{cases} 1, \text{ if } m_{ijt-1} = 0 \ \& \ m_{ijt} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{5}$$

A similar and symmetric definition can be made for separations s_{ijt}.

Full quarter employment (f): For $qfirst < t < qlast$, individual i was employed at j at the beginning and end of quarter t (full-quarter job)

$$f_{ijt} = \begin{cases} 1, \text{ if } m_{ijt-1} = 1 \ \& \ m_{ijt} = 1 \ \& \ m_{ijt+1} = 1 \\ 0, \text{ otherwise.} \end{cases} \tag{6}$$

Total earnings during the quarter (w_1): for $qfirst \leq t \leq qlast$, earnings of individual i at establishment j during period t

$$w_{1ijt} = \sum \text{all } UI \text{ covered earnings by } i \text{ at } j \text{ during } t \tag{7}$$

A.1.11 Establishment concepts

The above statistics, once defined for a single individual, are then aggregated to the establishment level across the demographic characteristics $d \in D$, then to higher levels across firm characteristics $k \in K$. Thus, we aggregate or compute a statistic x_{djt} for all individuals i with demographic characteristic d in period t at establishment j. For example, we can compute all accessions (a) for all "women aged 25-34" ($\tilde{d}$) for firm j as

$$a_{\tilde{d}jt} = \sum_{i \in \tilde{d}} a_{ijt} \tag{8}$$

In general,

$$x_{djt} = \sum_{i \in d} x_{ijt} \tag{9}$$

In this paper, b, a, s, and f generate establishment totals according to the formula above.

The key establishment statistic, however, is the average end-of-period employment growth rate g_{djt} for establishment j between periods $t-1$ and t. It is not computed directly from individual data, rather, it is computed using two other establishment-level statistics:

$$g_{djt} = \frac{jf_{djt}}{\bar{e}_{djt}} \tag{10}$$

where *net job flows* jf is the change in employment for establishment j during period t:

$$jf_{djt} = e_{djt} - b_{djt} \tag{11}$$

and $\bar{e}$ is average employment of establishment j between periods $t-1$ and t:

$$\bar{e}_{djt} = \frac{(b_{djt} + e_{djt})}{2} \tag{12}$$

and e and b are computed as in (9).

A.1.12 Aggregation

Finally, to arrive at an aggregated statistic for a particular group defined by $(d, k) \in D \otimes K$, for all variables including jf, the establishment-level statistics are summed over the relevant j during period t

$$X_{dkt} = \sum_{j \in k} x_{djt} \tag{13}$$

This version: $Id: confidentiality2005.tex 172 2005-12-05 16:55:01Z vilhuber $