4-24-2012

Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series

John Abowd
Cornell University, John.Abowd@cornell.edu

Kaj Gittings
Louisiana State University

Kevin L. McKinney
U.S. Census Bureau

Bryce E. Stevens
U.S. Consumer Finance Protection Bureau

Lars Vilhuber
Cornell University, lv39@cornell.edu

See next page for additional authors

Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series

Abstract
The Census Bureau's Quarterly Workforce Indicators (QWI) provide detailed quarterly statistics on employment measures such as worker and job ows, tabulated by detailed worker characteristics in various combinations. The data are released for detailed NAICS industries and for several levels of geography, the lowest aggregation of which are counties. OnTheMap, another Census Bureau product, provides a subset of these tabulations at the tract level. Disclosure avoidance methods are required to protect the information about individuals and businesses that contribute to the underlying data. The QWI disclosure avoidance mechanism we describe here relies heavily on the use of noise infusion through a permanent multiplicative noise distortion factor, used for magnitudes, counts, differences and ratios. There is minimal suppression and no complementary suppressions. To our knowledge, the release in 2003 of the QWI was the first large-scale use of noise infusion in any official statistical product. We show that the released statistics are analytically valid along several critical dimensions -- measures are unbiased and time series properties are preserved. We provide an analysis of the degree to which con dentiality is protected. Furthermore, we show how the judicious use of synthetic data, injected into the tabulation process, can completely eliminate suppressions, maintain analytical validity, and increase the protection of the underlying con dential data.

Keywords
noise infusion, synthetic data, statistical disclosure limitation, time-series, local labor markets, gross job

Comments
Suggested Citation

Abowd, J.M., Gittings, K., McKinney, K.L., Stephens, B.E., Vilhuber, L. & Woodcock, S. (2012, April). *Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series.* Presented at FCSM.

Authors
John Abowd, Kaj Gittings, Kevin L. McKinney, Bryce E. Stevens, Lars Vilhuber, and Simon Woodcock

Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series[☆]

John M. Abowd[a,b,c,g,h], Kaj Gittings[d], Kevin L. McKinney[c], Bryce E. Stephens[f], Lars Vilhuber[a,b,c], Simon Woodcock[e,g]

[a] *Cornell University, Economics Department*
[b] *Labor Dynamics Institute, ILR School, Cornell University*
[c] *U.S. Census Bureau, Center for Economic Studies*
[d] *Louisiana State University*
[e] *Simon Fraser University*
[f] *US Consumer Finance Protection Bureau*
[g] *Institute for the Study of Labor (IZA)*
[h] *National Bureau of Economic Research (NBER)*

Abstract

The Census Bureau's Quarterly Workforce Indicators (QWI) provide detailed quarterly statistics on employment measures such as worker and job flows, tabulated by detailed worker characteristics in various combinations. The data are released for detailed NAICS industries and for several levels of geography, the lowest aggregation of which are counties. OnTheMap, another Census Bureau product, provides a subset of these tabulations at the tract level. Disclosure avoidance methods are required to protect the information about individuals and businesses that contribute to the underlying data. The QWI disclosure avoidance mechanism we describe here relies heavily on the use of noise infusion through a permanent multiplicative noise distortion factor, used for magnitudes, counts, differences and ratios. There is minimal suppression and no complementary suppressions. To our knowledge, the release in 2003 of the QWI was the first large-scale use of noise infusion in any official statistical product. We show that the released statistics are analytically valid along several critical dimensions – measures are unbiased and time series properties are preserved. We provide an analysis of the degree to which confidentiality is protected. Furthermore, we show how the judicious use of synthetic data, injected into the tabulation process, can completely eliminate suppressions, maintain analytical validity, and increase the protection of the underlying confidential data.

JEL categories: C82, J21, J23, J40

Keywords: noise infusion, synthetic data, statistical disclosure limitation, time-series, local labor markets, gross job flows, gross worker flows, confidentiality protection

[☆]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. All results have been reviewed to ensure that no confidential information is disclosed. The views expressed herein are attributable only to the authors and do not represent the views of the U.S. Census Bureau, its program sponsors, Cornell University, the Director of the Consumer Financial Protection Bureau nor those of its staff, nor any of the data providers. This document draws on Gittings (2009) and Stephens (2007).

1. Introduction

Statistical disclosure limitation is the set of methods used to protect the confidentiality of the identity and attributes of the individuals, businesses or other entities that supplied the micro-data used to create public-use data products.[1] Disclosure avoidance protocols have come under intense scrutiny as improvements in information technologies have permitted increasingly sophisticated threats to the integrity of extant protection systems. At the same time, statistical agencies have been asked to ensure that their public-use data provide great levels of detail and meet analytical validity standards, creating a constant tension with the disclosure limitation procedures. The widely-cited, elegant, and still very relevant National Research Council study *Private Lives and Public Policies* (Duncan et al., 1993) noted that "[i]n choosing among different disclosure limitation techniques, agencies should take account of the level of protection provided and the effects on the ability of users to draw valid inferences." Introducing the second special issue of the *Journal of Official Statistics* on disclosure limitation (Fienberg and Willenborg, 1998, pg. 338) note that "[o]n the one hand, there is the agencies' public obligation to provide maximum information to society, while on the other hand, the agencies must ensure that the privacy of individual entities represented in the data is sufficiently protected." Finally, the Federal Committee on Statistical Methodology in its 2005 compendium of methods for disclosure limitation notes "agencies should consult data users on issues relating to: balancing the risk of disclosure against the loss in data utility," (Federal Committee on Statistical Methodology, 2005, page 99) although there are no proposed standards for "data utility."

Evidently, a critical component of any agency system using confidential micro-data for statistical purposes is to produce detailed, valid products without compromising the confidentiality of the original data. Duncan and Lambert (1986), building on the pioneering methodology of Dalenius (1977), proposed the formalization of this objective by showing that common procedures are all special cases of a disclosure limitation protocol that bounds the posterior predictive distribution away from zero or one and that these bounds also measure the information loss from the procedure. Hence, their analysis clearly focused attention on the formal trade-off. Statisticians now refer to this as the risk-utility trade-off as formalized by Duncan et al. (2001b). Economists (see Abowd and Lane, 2004) call this a production possibility frontier between confidentiality protection and information release. In their monograph, Willenborg and de Waal (2001) devote a chapter to assessing the consequences of micro-data disclosure control on the analytical validity of standard inferential statistical procedures using the Kullback-Leibler (1951) relative entropy measure. The Doyle et al. (2001) collection contains papers by Abowd and Woodcock (2001), who assess analytical validity of partially synthetic data using univariate, bivariate, and selected multivariate statistical models, Domingo-Ferrer and

[1] This is also sometimes called "disclosure proofing" or "disclosure avoidance" at statistical agencies.

Torra (2001), who assess information loss for micro-data disclosure avoidance protocols using a host of validity measures, and Duncan et al. (2001a), who assess the analytical validity of tabular suppression methods using mean squared precision.

This paper presents a protection technology that relies primarily on a single infusion of noise into the entire longitudinal history of an entity in the confidential underlying micro-data. The procedure is an extension of the single-period method proposed by Evans et al. (1998). The procedure described here extends the single noise infusion model so that it is dynamically consistent, *i.e.*, preserves time-series as well as cross-sectional analytical validity. As a consequence, the public use data preserves stock-flow relationships in the underlying micro-data. In addition, our dynamically-consistent method extends the cross-sectional methods by showing how they can be used for magnitudes, differences in magnitudes, and ratios.

The procedure we developed was subsequently implemented in a novel statistical product. Since 2003, the Census Bureau has published a collection of statistical series called the Quarterly Workforce Indicators (QWI).[2] The underlying micro-data infrastructure was designed by the Longitudinal Employer-Household Dynamics (LEHD) Program at the Census Bureau (Abowd et al., 2004) and is described in detail elsewhere (Abowd et al., 2006, 2009). At its core, the QWI system uses administrative records data on jobs (employer-employee pairs) and establishments (work locations) collected from 49 states.[3] The administrative records data are enhanced with information from other micro-data at the Census Bureau. Consequently, the public-use QWI offer unprecedented demographic and economic detail on the local dynamics of labor markets. As of 2011, data are released for three types of cross-tabulations: eight age groups by sex, six race categories by hispanic origin, and four education categories by sex, in tabulations that further control for ownership category, detailed sub-state geography, and NAICS industry group. The released data can be aggregated; however, published aggregates, prepared by the agency, are less distorted than customized post-release aggregates, prepared by a user.

Because of the fine detail offered by the published statistics and the confidential nature of the micro-data used to compile the indicators, confidentiality protection is a critical and integral part of the design of the QWI system. We quantify the protection provided by the system and show that the analytical validity of the data remains high in comparison with the indicators prepared directly from the confidential micro-data without noise infusion. In particular, we provide evidence that the time-series properties of the disclosure-protected data remain intact, and that the disclosure-protected data are not biased. Of course, the noise

[2] A variant of the same noise-infusion mechanism has been used since 2007 to protect the confidentiality of data underlying the Census Bureau's County Business Patterns (Massell and Funk, 2007) and was tested for application to the Commodity Flow Survey (Massell et al., 2006).

[3] Data are available for participating states which have joined the Federal/State Local Employment Dynamics Cooperative Program and have regularly delivered data to the Census Bureau. Current information can be found at http://lehd.did.census.gov/led.

infusion system makes the public use data have greater variance than the same indicators prepared directly from the unprotected micro-data. We quantify this increase in relative terms.

The confidentiality protection system described in this article results in the release of some public-use items that are flagged as "significantly distorted to preserve confidentiality" because they differ from the undistorted item by a significant proportion. Even for the significantly distorted items, the data remain analytically valid for time series properties.

Magnitude data based on a few entities 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 disclosure avoidance system, 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 magnitudes like employment counts that can take values too small to be protected in the first sense. Item suppression is still used when the employment count is too small to be afforded either type (1) or type (2) protection.

These two conditions are met by the multiple layers of confidentiality protection in the QWI 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 (establishment) measures. The noise is designed to have three very important properties. First, every data item is distorted by some minimum amount. 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 over establishments, the effects of the distortion cancel out for the vast majority of the estimates, preserving both cross-sectional and time-series analytical validity. After this noise infusion, the distorted data item is used in all the publication QWIs.

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-wide beginning-of-quarter employment for all private employers to match the state-wide 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. Employment count data are subject to suppression because they can take small integer values that are not adequately protected using the criteria above. Continuous dollar measures like payroll are not suppressed because they have all the features of the magnitude data originally modeled by Evans et al. (1998). Some 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.

In addition to the analysis of the noise infusion and cell suppresssion system as it is implemented in the QWI released as of 2011, we also describe an experimental mechanism that addresses the suppression and distortion. Synthetic values are generated by sampling from the posterior predictive distribution of the underlying confidential data, not the released data, given its history and the rules that cause the suppression. The synthetic values are then combined with the tabulation from the noise-infused data to create the publication tabulations. The use of synthetic data in this application improves the analytical validity of the QWI because the user no longer has to model the suppressions separately. We demonstrate that the combination of the two protection systems meets the standard of providing a minimum probability that a particular count is not the true count.

The remainder of this article is structured as follows. Section 2 describes the dynamically-consistent multiplicative noise infusion model. Section 3 details its integration into the computation of the QWI. Section 4 provides an overview of the imputation procedures and Section 5 describes the weighting procedures used in the development of the QWI employment and earnings measures. The algorithm underlying the item suppression is outlined in Section 6. Sections 7 and 8 provide evidence on the extent of the protection and the analytical validity, respectively. Section 9 describes the synthetizer used for the combined system, the algorithm used to combine the protected and the synthetic data, and results from a comparison of the combined system to the protection-only system. Section 10 concludes.

2. The dynamically-consistent multiplicative noise infusion model

The idea underlying noise infusion is to permanently perturb all the inputs to the statistical query on the underlying confidential data. Then, each entity is afforded protection to the extent that the input perturbation distorts the entity's data by a minimum amount (see Federal Committee on Statistical Methodology, 2005, page 72). Multiplicative noise infusion distorts the input data by multiplying each magnitude item

for a given entity by a random distortion factor, called a fuzz factor, that is centered around unity. The multiplicative distortion guarantees minimal and maximal input distortion by sampling the fuzz factor from a specially constructed distribution that has no support surrounding unity. Dynamically consistent noise infusion uses the same fuzz factor for each period that the entity contributes micro-data to the analysis.

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

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

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

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.[4] 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.

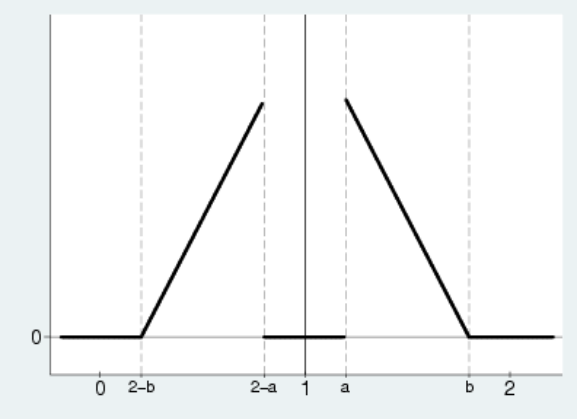
A fuzz factor is drawn once for each employer, and once for each of the establishments associated with that employer. Fuzz factors are permanently attached to each employer and establishment. They are retained for all time periods and for all revisions of the QWI public use data. Although fuzz factors 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.

3. Applying the fuzz factors to different types of indicators

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. (2009), and a list of definitions for the statistics mentioned in this section, and the formulae for their computation is provided in the appendix of that publication and in Abowd et al.

[4] The exact numbers are confidential.

Figure 1: Distribution of Fuzz Factors



(2006). 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 magnitudes (counts or dollar values) by scaling of the true establishment level statistic

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

where X_{djt} is an establishment-level statistic for a given demographic group d (e.g., age group a by sex s, or some other valid combination) 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 (H_3), total payroll (W_1), payroll associated with E (W_2), with F (W_3), with new full-quarter accessions (WFA), with new full-quarter accesions who were also new hires (WH_3), with new full-quarter separations (WFS), 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

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

where ZY_{djt} is a statistic related to a total Y_{djt}, 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, $ZWFA$, ZWH_3, $ZWFS$), 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 by industry by age by 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 and separations). To protect average wage changes, the undistorted dollar wage changes were divided by the undistorted base, then multiplied by the ratio of the distorted denominator to the undistorted 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 ($Z\Delta WA$ and $Z\Delta WS$).

4. Multiple imputation of missing establishment characteristics

Because the employer identifier on the unemployment wage records refers to a UI account, they do not contain information on the establishment's economic activity (industry code) nor its geographic location (address), except for data provided by Minnesota. For single-unit employers–those with a single establishment–this information can be derived from the employer-level information on the QCEW records. However, approximately 30 to 40 percent of state-level private employment is at establishments that are part of a multi-unit employer. While the QCEW has information on all establishments, derived from auxiliary reports called the Multiple Worksite Report, it does not have information on the establishment at which a particular employee reports to work.

In order to impute establishment-level characteristics to job histories of multi-unit employers, a non-ignorable missing data model with multiple imputation was developed. The model is described in detail in Abowd et al. (2009) and Stephens (2007). The model multiply imputes establishment-of-employment

based on (1) the distribution of employment across establishments of multi-unit employers and (2) distance between place-of-work and place-of-residence. The distance to work model is estimated using data from Minnesota, where both the SEIN and SEINUNIT identifiers appear on a UI wage record. For other states, the posterior distribution of the parameters from this estimation, combined with SEIN and SEINUNIT employment histories from the QCEW data, are used to multiply impute SEINUNIT within SEIN, and thus its associated characteristics, to a particular job (worker-employer combination). The implicates are then used in the downstream processing of the QWI. The basic proportions in this imputation are the proportions of employment in each establishment whose entity demography is consistent with the complete wage record history of the individual-employer being imputed. The distance-to-work model adjusts these proportions.

Thus, for all states but Minnesota, the imputation of establishment-level characteristics to jobs is based on a purely statistical missing data model, rather than the actual, unobserved value. The imputed data thus provide an additional, indirect, level of protection similar to synthetic micro-data. However, the influence of the missing data imputation on the statistical disclosure limitation methods used in the QWI will not be analyzed in this paper, which focusses solely on the contribution of the noise-infusion and cell suppression method that are directly applied at the tabulation stage.

5. Generation of QWI weights

The economic concepts underlying the Quarterly Census of Employment and Wages (QCEW) statistics, published by Bureau of Labor Statistics (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 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.

Weighting is not used to control sub-state geography and industry for two reasons. First, the LEHD establishment-level edits to the QCEW data differ from the BLS edits, which implies that published BLS totals are not the appropriate controls. Although the confidentiality-protected LEHD-produced controls could be used, this would not address the problem of two different control totals–LEHD and BLS versions. Second, the multiple imputation procedure for the missing workplace characteristics of wage records associated with multi-unit employers is not easily adapted to such controls because of the way it handles failures–specifically, the use of modal attributes for the employer when the wage record fails the multiple imputation. Early versions of the QWI attempted sub-state geography controls.

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. Item suppression

Despite the noise infusion and quarterly reweighting described in the previous sections, some disclosure risk remains for employment counts based on very few entities in a cell because their values must be whole numbers. The current noise infusion system cannot adequately protect these small whole numbers, as we demonstrate in Section 7. Hence, certain employment counts, and related flows, are suppressed for this reason. In addition, although the Evans et al. (1998) procedure does provide adequate protection for employment counts and flows based on one or two employing establishments, the current public-use file also suppresses those items. In Section 9, we propose a method to avoid these suppressions, but data published as of 2011 still have item suppression. Item suppression affects B, E, M, F, A, S, H, R, FA, FH, FS, JC, JD, JF, FJC, FJD, FJF, and associated average earnings or average change in earnings variables.

Consider cell k in time period t, where a cell k represents a particular combination of geography $\times$ industry $\times$ age $\times$ sex.

- Check the conditions leading to a disclosure status 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 Bureau publication 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.
  - For average earnings variables (ZW_2, ZW_3, ZWH_3, $ZWFA$, $ZWFS$) and change in average earnings variables ($Z\Delta WA$ and $Z\Delta WS$), determine whether the (rounded) denominator is zero.
  - For all variables, check whether the data used to construct the cell kt value were based on 1 or 2 individuals.
  - For all variables, check whether the data used to construct the cell kt value were based on 1 or 2 employers.

  If any of these conditions is met, set the disclosure status flag to 5 and set the item to missing in the public-use file.
- Check whether the item rounds to zero. If so, set the disclosure status flag to 0.
- Check whether the distortion of cell kt value exceeds the limit set by the Census Bureau Disclosure Review Board[5]. If so, set the disclosure status flag to 9 and copy the distorted value to the release file.
- Otherwise, set the disclosure status flag to 1 and copy the distorted value to the release file.

Because the noise infusion and weighting previously discussed protect all of the non-suppressed items, no complementary suppressions are needed. All of values based on three or more individuals or employers are adequately protected. Any estimate of the suppressed item computed by subtraction is also protected.

Table 1 lists all possible flag values.

7. 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 comparisons were computed using both custom internal tabulations as well as published numbers, for two states (Illinois and Maryland). Each cell underlying the tabulation is for a statistic X_{kt} for k defined by a combination of age, gender, industry (SIC3), and geography (county), and for all released

[5]The precise value is confidential.

Table 1: Disclosure avoidance status 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

time periods for the states at the time of these experiments.[6] For any given state, the number of cells will differ by the number of geographical areas within the state, and the number of quarters of available data. However, experiments showed that Illinois (a medium-to-large state) and Maryland (a small state with the longest time series) have typical results.

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, 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) illustrates 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 a 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 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

[6] These experiments were run in 2003.

Table 2: Small Cells: B, Undistorted and Unweighted (Raw) vs. Undistorted and Weighted, SIC3

(a) Illinois

	Undistorted and Weighted Count					
Undistorted and 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

	Undistorted and Weighted Count					
Undistorted and 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 and Weighted vs. Distorted and Weighted, SIC3

(a) Illinois

	Weighted and Distorted Count					
Weighted and 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

	Weighted and Distorted Count					
Weighted and Undistorted 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, Undistorted and Unweighted (Raw) vs. Published, SIC3

(a) Illinois

	Published Count						
Undistorted and Unweighted Count	Suppressed	0	1	2	3	4	5 or more
0	0.66	99.21	0.13	0.00	0.00	0.00	0.00
1	99.89	0.08	0.02	0.00	0.00	0.00	0.00
2	91.51	0.01	0.00	2.51	5.87	0.09	0.01
3	32.13	0.01	0.00	2.19	47.75	16.98	0.94
4	25.83	0.00	0.00	0.04	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						
Undistorted and Unweighted Count	Suppressed	0	1	2	3	4	5 or more
0	0.90	98.94	0.16	0.00	0.00	0.00	0.00
1	99.88	0.09	0.02	0.00	0.00	0.00	0.00
2	80.81	0.04	0.00	4.90	13.90	0.32	0.02
3	22.61	0.03	0.00	0.93	40.18	33.60	2.65
4	18.05	0.02	0.00	0.01	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.

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 × county × time × sex × age), around 25% of the beginning of period employment cells are suppressed in both the states analyzed here. We will return to this high level of suppressions in Section 9. 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 dollar magnitude, not an employment 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 by 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.

8. Analytical validity

The noise infusion described in Section 2 is designed to preserve the analytical 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 × industry × age × sex cell kt. Sub-state geography in all cases is a county, whereas the industry classification is SIC. For our purposes, analytical validity is obtained when the data display no bias and the additional dispersion due to the confidentiality protection system can be quantified so that statistical inferences can be adjusted to accommodate it.

8.1. Time-series properties of distorted data

To analyze the impact on the time series properties of the weighted, 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 next 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 19 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 central tendency of the bias (as measured by the median of the Δr distribution) is never greater than 0.001 (raw versus distorted or raw versus 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[7]. The distribution of errors is similar when considering raw versus published data (Table 6 on page 19). Tables 7–10 repeat the analysis of bias for more detailed SIC 2 x county and SIC 3 x county cells. The tables show that 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.

[7]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	B Beginning of Quarter Employment	A Accessions	S Separations	F Full Quarter Employment	JF 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	B Beginning of Quarter Employment	A Accessions	S Separations	F Full Quarter Employment	JF 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 × County, Raw vs. Distorted Data

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

Percentile	B Beginning of Quarter Employment	A Accessions	S Separations	F Full Quarter Employment	JF 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 × County, Raw vs. Published Data

$\Delta r = r - r^*$

Percentile	B Beginning of Quarter Employment	A Accessions	S Separations	F Full Quarter Employment	JF 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.

Table 9: Distribution of the Error in the First-order Serial Correlation: Three-digit SIC × County, Raw vs. Distorted Data

	$\Delta r = r - r^*$				
	B	A	S	F	JF
	Beginning of Quarter			Full Quarter	Net Job
Percentile	Employment	Accessions	Separations	Employment	Flows
			IL SIC3		
01	-0.069422	-0.059554	-0.061773	-0.066439	-0.021332
05	-0.036533	-0.035072	-0.037855	-0.034509	-0.008231
10	-0.023716	-0.025104	-0.026586	-0.022499	-0.005039
25	-0.008352	-0.010908	-0.010209	-0.008086	-0.001631
50	0.000000	0.000001	0.000533	0.000000	-0.000051
75	0.009779	0.010971	0.012838	0.008914	0.001456
90	0.025995	0.025771	0.027628	0.024343	0.004120
95	0.039350	0.035535	0.038975	0.037117	0.007231
99	0.078006	0.057571	0.062574	0.072630	0.015415
			MD SIC3		
01	-0.056972	-0.055866	-0.060231	-0.052390	-0.230760
05	-0.033133	-0.035893	-0.038862	-0.029267	-0.013809
10	-0.022887	-0.026384	-0.027551	-0.020339	-0.007502
25	-0.009078	-0.011608	-0.012282	-0.008090	-0.003020
50	0.000000	0.000058	-0.000000	-0.000000	-0.000416
75	0.008957	0.012534	0.012177	0.007787	0.001211
90	0.022707	0.028441	0.026675	0.020442	0.004219
95	0.033875	0.039179	0.037298	0.030153	0.007328
99	0.060929	0.060199	0.062728	0.055435	0.013156

Unit of observation is a cell. Industry aggregation is SIC3, geography aggregated to county level. All tabulations are weighted using the QWI weights. For more details, see text.

Table 10: Distribution of the Error in the First-order Serial Correlation: Three-digit SIC × County, Raw vs. Published Data

$\Delta r = r - r^*$

Percentile	B Beginning of Quarter Employment	A Accessions	S Separations	F Full Quarter Employment	JF Net Job Flows
			IL SIC3		
01	-0.169394	-0.120104	-0.114040	-0.154348	-0.030423
05	-0.063777	-0.062527	-0.059499	-0.061869	-0.008936
10	-0.041526	-0.042022	-0.040601	-0.040379	-0.005339
25	-0.017723	-0.019520	-0.018339	-0.017450	-0.001938
50	-0.002337	-0.002810	-0.001131	-0.002643	-0.000047
75	0.009844	0.011742	0.013859	0.008851	0.001542
90	0.026970	0.029207	0.032437	0.025157	0.004658
95	0.041346	0.042109	0.047260	0.038599	0.007741
99	0.083776	0.090491	0.103535	0.077563	0.015973
			MD SIC3		
01	-0.134109	-0.107447	-0.111710	-0.127754	-0.232744
05	-0.058736	-0.059315	-0.062445	-0.055274	-0.015091
10	-0.038857	-0.042220	-0.044345	-0.035878	-0.007920
25	-0.017310	-0.020629	-0.021377	-0.016571	-0.003322
50	-0.002988	-0.002758	-0.003140	-0.003188	-0.000502
75	0.008371	0.013841	0.013680	0.007050	0.001371
90	0.023079	0.032981	0.033616	0.020351	0.004212
95	0.035196	0.048446	0.052044	0.030612	0.007344
99	0.070017	0.119878	0.156250	0.059618	0.013653

Unit of observation is a cell. Industry aggregation is SIC3, geography aggregated to county level. For more details, see text.

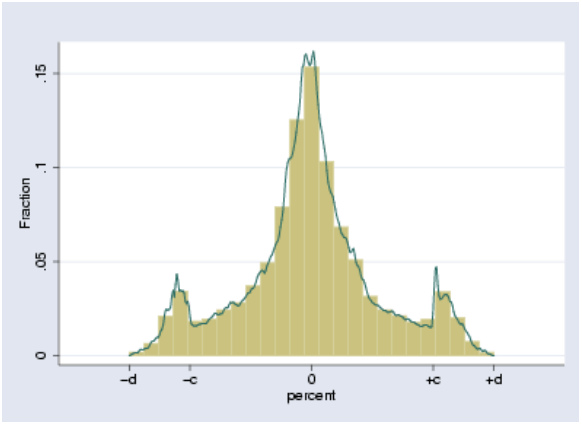
8.2. Cross-sectional unbiasedness of the distorted data

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).

(a) Illinois



(b) Maryland

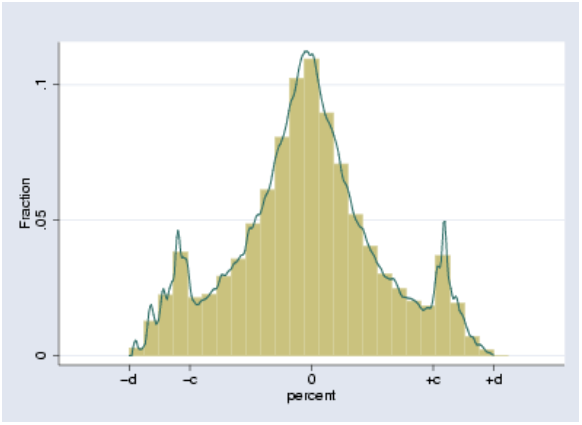
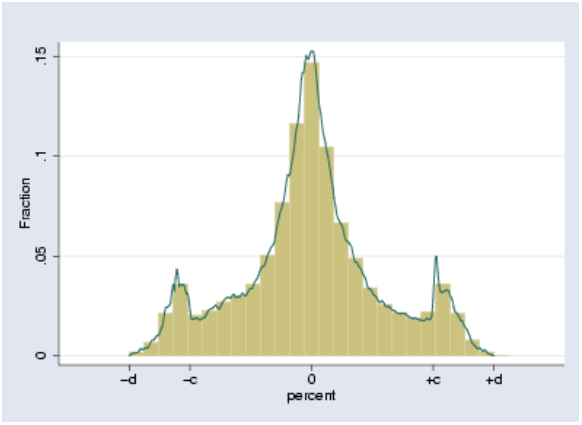


Figure 2: Distribution of Noise, B, SIC3

(a) Illinois



(b) Maryland

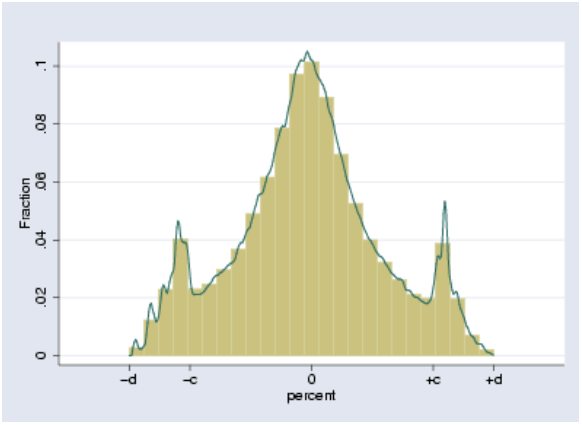


Figure 3: Distribution of Noise: W_1, SIC3

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.

9. Synthetic Data as a Proposed Alternative to Item Suppression

As we have described, the QWI statistics released to the public as of 2011 incorporate item suppressions as part of the confidentiality protection measures. Suppressed items occur in cells with particular characteristics, but not all items in a cell are suppressed. At the most detailed levels of industry and geography, a significant number of suppressions and distorted estimates are released in the current QWI. For example, in the most recent release of the QWI (R2011Q4), 16% of cells have suppressed B values are in Maryland, and 24% in Illinois. No cells have suppressed W_1 values.[8]

As an alternative to item suppression, we developed a synthetic data model that replaces suppressed values with draws from an appropriate posterior predictive distribution.[9] The system incorporating both noise-infused and synthetic data will be referred to as a "hybrid" system, leading to released data without suppressions. We show that the confidentiality protection provided by the hybrid system without suppressions is comparable to the protection afforded by the system using the noise infusion system with suppressions, but the analytical validity of the experimental system is improved because the synthetic data are better than the best inference an external user can make regarding the suppressions. This experimental system has not been implemented by the Census Bureau.

9.1. Synthetic Data Model

To synthesize the core set of variables and maintain dynamic consistency we need a model that satisfies all of the definitions and identities (Abowd et al., 2009). However, the process is complicated by the desire to use the synthetic data only when the noise infusion does not adequately protect an estimated employment count. Since all of the released data will be based on either noise infusion or synthesis, preserving dynamic consistency requires that the noise-infused values and the synthetic values be consistent. This is done by choosing an appropriate set of conditioning variables and sampling from the correct joint posterior distribution to ensure that the identities hold.

The synthetic data model is based on a multinomial likelihood with Dirichlet priors. Specifically, denote Y_{djt} as the set of QWI variables to be jointly synthesized (e.g. $Y_{djt} = (B_{djt}, H_{djt}, R_{djt})$) and Y^r_{djt} as the resulting set of synthetic values for a given demographic group d (e.g., age group a by sex s, or some other valid combination) and firm j. Each element of Y_{djt} takes on the values of $0, 1, 2, 3$ or 4^+. We denote the conditioning set as Ω_{djt} that contains Y_{djt-1}, Y_{djt+1} as well as job flows, JF^*_{djt}, which has a feasible range of -4 to 4. Letting θ be the vector of multinomial probabilities and $\alpha_{1|djt}, ..., \alpha_{L|djt}$ be the shape parameters

[8] In contrast to the data reported in Tables 2 through 4, results in this section are reported for NAICS-based tabulations. For NAICS sub-sectors (NAICS3) by county by age by sex cells, 16% of cells are suppressed for B in Maryland, and 24% in Illinois.

[9] This section draws on Gittings (2009).

of the Dirichlet for the L possible outcomes, the likelihood and Dirichlet prior can be summarized by the following two equations:

$$p(n_{djt}|\Omega_{djt}, \theta_{djt}) \propto \prod_{l=1}^{L} \theta_{l|djt}^{n_{l|djt}} \tag{3}$$

$$\theta_{djt} \tilde{} \; Dirichlet(\alpha_{1|djt}, ..., \alpha_{L|djt}) \tag{4}$$

where n_{djt} is the vector of counts of Y_{djt} for characteristics d in establishment j in quarter t. The prior shape is given by the vector $\left(\alpha_{1|djt}, ..., \alpha_{L|djt}\right)$. The resulting posterior can then be written as:

$$\theta_{djt}^{pos} \tilde{} \; Dirichlet(\alpha_{1|djt} + n_{1|djt}, ..., \alpha_{L|djt} + n_{L|djt}) \tag{5}$$

$$p(n_{djt}^{pos}|\Omega_{djt}, \theta^{pos}) \propto \prod_{l=1}^{L} \left(\theta_{l|djt}^{pos}\right)^{n_{l|djt}^{pos}} \tag{6}$$

Replacing the suppressions with synthetic data, then, requires first sampling from the posterior distribution of the probabilities, then using that draw to sample an outcome Y_{djt}^{r} for establishment j.[10] The resulting outcome Y_{djt}^{r} is used to compute the remaining QWI variables via the definitions and identities.

9.1.1. Estimation of the likelihood component

To illustrate the construction of the likelihood, consider the synthesizer that draws Y_{djt}^{r} conditional on Ω_{djt}. For each age group and sex $d = (a, s)$ with data configuration Ω_{djt} we estimate the likelihood contribution separately for each quarter t as follows. In quarter t, select only those establishments for which the values of Y_{djt} lie in the allowable outcome space (all non-negative). Stratify these establishments according to the observed combinations of Y_{djt-1}, Y_{djt+1} and JF_{djt}^{*}. Let $\eta_{l|dmt}$ be the establishment count for each possible combination l in the feasible outcome space of Y_{djt}, where m designates each unique combination of Y_{djt-1}, Y_{djt+1} and JF_{djt}^{*}. Then, $\eta_{\bullet|dmt}$ is the total number of establishments with configuration m, and the subscript $\bullet$ denotes the operation of summing. The estimator $\widehat{\theta}_{l|dmt} = \frac{\eta_{l|dmt}}{\eta_{\bullet|dmt}}$ denotes the maximum likelihood estimator of the non-zero outcome probabilities.

9.1.2. Specification of the Dirichlet Prior

We want to use a data-based informative prior for the probabilities. To do this, we aggregate over the conditioning variables Y_{djt-1} and Y_{djt+1} so that the conditioning set consists of groups and job flow counts, JF_{djt}^{*}. The aggregated data are then pooled across the quarter being synthesized and three additional seasonally-consistent quarters–historical if available, future otherwise–designated by the set of quarters Q_t.

[10] For algorithms to empirically compute the Dirichlet posterior, see Gelman et al. (2003).

To ensure that the posterior receives positive weight on all feasible outcomes, we blend the data-based prior with a uniform prior denoted as u. The Dirichlet prior shape parameters are estimated by

$$\widehat{\rho}_{l|dmt} = 0.99\left(\frac{\sum_{t\in Q_t}\eta_{l|dmt}}{\sum_{t\in Q_t}\eta_{\bullet|dmt}}\right) + 0.01u \tag{7}$$

The specification of the Dirichlet is completed by assigning a prior sample size. The results shown below are for a prior sample size of 1. The exact prior sample size used in any implementation may be confidential. Denoting the prior sample size by α_0, the Dirichlet prior can be completely specified by $\alpha_{l|dmt} = \alpha_0\widehat{\rho}_{l|dmt}$, where l ranges over all values in the feasible set that have positive prior probability.

9.1.3. Sampling from the Dirichlet Posterior

For each observed value of (d, m, t), the probabilities $\theta^{pos}_{l|dmt}$ have a posterior Dirichlet distribution with parameters $\alpha_{1|dmt} + n_{1|dmt}, ..., \alpha_{L|dmt} + n_{L|dmt}$, where events with zero posterior probability have been removed from the feasible outcome space. For each (d, m, t), sample $\theta_{l|dmt}$ from the posterior Dirichlet. Then, for each establishment j, sample Y^r_{djt} from these probabilities. Compute the remaining QWI variables from the identities and definitions above. If the computed values from the identities are infeasible (*i.e.*, are negative), reject and draw again. The synthetic data sampling can be performed multiple times; however the results reported below are for single synthetic data samples.

9.2. Combining Synthetic Data with Protected Data

Since the QWI are linked through a series of identities and inequality constraints, the quantities must be partitioned into a subset that is synthesized and a subset that is evaluated using the identities. Furthermore, this allows for the creation of multiple sets of synthetic data depending on which subsets are synthesized and calculated. A decision rule on which set to use minimizes the amount of synthesis necessary for protection.

The synthesis is conducted in two stages with the full-quarter variables being synthesized last. Let $(B_{djt}, H_{djt}, R_{djt})$ be *group 1* variables to synthesize and $(E_{djt}, A_{djt}, S_{djt})$ be *group 2* variables to synthesize. When *group 1* is synthesized, $E_{djt}, A_{djt}, S_{djt}, M_{djt}$ are evaluated using the identities and likewise, and when *group 2* is synthesized, $B_{djt}, H_{djt}, R_{djt}, M_{djt}$ are evaluated using identities. The second stage consists of synthesizing the values $(F_{djt}, FA_{djt}, H3_{djt})$ and evaluating FS_{djt}. Note that the second stage draws the full-quarter synthetic values conditional on the synthetic set obtained from the first stage.

Let Y^{1r}_{djt} denote the synthetic value of variables $B_{djt}, E_{djt}, H_{djt}, R_{djt}, JF_{djt}, JC_{djt}, JD_{djt}, A_{djt}, S_{djt}, M_{djt}$ resulting from *group 1* synthesis, and let Y^{2r}_{djt} denote the values from *group 2* synthesis. Let Z^{1r}_{djt} denote the synthetic values of full-quarter variables $F_{djt}, FA_{djt}, FS_{djt}, H3_{djt}, FJF_{djt}, FJC_{djt}, FJD_{djt}$ that result from conditioning on Y^{1r}_{djt} in the second stage, and Z^{2r}_{djt} the synthetic values that result from conditioning on Y^{2r}_{djt} in the second stage. This results in two full sets of synthetic data we label $X^{1r}_{djt} = [Y^{1r}_{djt}\ Z^{1r}_{djt}]$ and $X^{2r}_{djt} = [Y^{2r}_{djt}\ Z^{2r}_{djt}]$.

The overall algorithm creates a blend of noise-infused and partially synthetic data (Raghunathan et al., 2003; Reiter, 2004). It can be summarized as follows:

Stage 1

- Draw $(B^r_{djt}, H^r_{djt}, R^r_{djt})$ given $(B_{djt-1}, H_{djt-1}, R_{djt-1})$, $(B_{djt+1}, H_{djt+1}, R_{djt+1})$, JF^*_{djt}
- Draw $(E^r_{djt}, A^r_{djt}, S^r_{djt})$ given $(E_{djt-1}, A_{djt-1}, S_{djt-1})$, $(E_{djt+1}, A_{djt+1}, S_{djt+1})$, JF^*_{djt}
- For each synthesized group, evaluate the remaining variables using the identities (note that use of the definitions insures that $JC^r_{djt} = JC^*_{djt}$and $JD^r_{djt} = JD^*_{djt}$).
- Check that $Y^{1r}_{djt}, Y^{2r}_{djt} \geq 0$ for all variables. If not, redraw the appropriate synthesis group and reevaluate the identities.
- For $Y^r_{djt} = JF_{djt}, JC_{djt}, JD_{djt}$ calculate $Y^r_{\bullet sjt} = Y^*_{\bullet sjt}, Y^r_{a\bullet jt} = Y^*_{a\bullet jt}$, and $Y^r_{\bullet\bullet jt} = Y^*_{\bullet\bullet jt}$.[11]
- For the remaining variables in Y^r_{djt}, calculate $Y^r_{\bullet sjt}, Y^r_{a\bullet jt}$, and $Y^r_{\bullet\bullet jt}$.

Stage 2

- Draw $(F^r_{djt}, FA^r_{djt}, H3^r_{djt})$ given $(F_{djt-1}, FA_{djt-1}, H3_{djt-1})$, $(F_{djt+1}, FA_{djt+1}, H3_{djt+1})$ (note that conditioning on FJF^*_{djt} here would fully constrain F^r_{djt}).
- Evaluate $FS^r_{djt}, FJC^r_{djt}, FJD^r_{djt}$ ($FS^r_{djt} = FA^r_{djt} - FJF^*_{djt}$ and $FJF^r_{djt} = FJF^*_{djt}, FJC^r_{djt} = FJC^*_{djt}$, $FJD^r_{djt} = FJD^*_{djt}$).
- Check that $Z^{1r}_{djt}, Z^{2r}_{djt} \geq 0$ for all variables. If not, redraw the appropriate synthesis group and reevaluate the identities.
- For $Z^r_{djt} = FJF_{djt}, FJC_{djt}, FJD_{djt}$ calculate $Z^r_{\bullet sjt} = Z^*_{\bullet sjt}, Z^r_{a\bullet jt} = Z^*_{a\bullet jt}$, and $Z^r_{\bullet\bullet jt} = Z^*_{\bullet\bullet jt}$.
- For the remaining variables in Z^r_{djt}, calculate $Z^r_{\bullet sjt}, Z^r_{a\bullet jt}$, and $Z^r_{\bullet\bullet jt}$.

At the end of the synthesizing algorithm, each establishment has a complete set of $X_{djt}, X^*_{djt}, X^{1r}_{djt}$, and X^{2r}_{djt} and all dynamic identities hold. Furthermore, the dynamic identities and the intra-establishment marginal employment counts are all consistent between the noise-infused and synthetic data.

[11] In the implementation of this experimental hybrid system, we used $d = (a, s)$. Marginal values of the $a \times s$ cells were calculated. This is the reason the subscripts a and s occur in this section.

9.3. Forming Hybrid Quarterly Workforce Indicators

The QWI are created by aggregating the micro-data for establishments j into ownership $\times$ geography $\times$ industry categories using the establishment weight w_{jt}. Call an item of a particular aggregation k. Thus, the subscript k replacing the establishment subscript indicates that all the establishments meeting a particular set of ownership, geography, and industry criteria have been summed. For each k, consider the QWI X_{askt}, X^*_{askt}, and X^r_{askt}. A hybrid QWI algorithm consists of specifying which synthetic value X^r_{askt} to use, and when to replace X^*_{askt} with X^r_{askt} in lieu of suppressing certain items in X^*_{askt}.

At each level of aggregation, the QWI are calculated from X_{askt} and X^*_{askt}, respectively. Then the values computed from X^*_{askt} are evaluated, item by item, according to the item suppression rules described in section 6. The rules are searched in the order of the variables in group 1, then group 2. If any of the items in variable group 1 would be suppressed using the values calculated from X^*_{askt}, then the synthetic data X^1_{askt} are used below. Else if any of the items in variable group 2 would be suppressed using the values calculated from X^*_{askt}, then the synthetic data X^2_{askt} are used below. Otherwise, the QWI are computed as usual from X^*_{askt}.

Given the experiments with synthetic data described in this paper, the experimental rules are:

- if any $X_{askt} \in \{1, 2\}$ then release the QWI computed from X^r_{askt} for the appropriate r and set the item disclosure status flag to 9;
- else if $abs\left(\frac{X^*_{askt} - X_{askt}}{X_{askt}}\right) \geq \beta$ then release the QWI computed from X^*_{askt} and set the item disclosure status flag to 9;[12]
- else release QWI computed from X^*_{askt} and set the disclosure status flag to 1.

Under these experimental rules with a synthetic data protection component, there are item suppressions (items with disclosure status flag 5); however, significant distortion can arise from either noise infusion or synthesis. Hence, the revised release rules meet either the magnitude inference distortion or the probability inference distortion conditions set forth in the introduction.

9.4. Results for Experimental Protection Rules

The results summarize the effects of the various layers of the protection system with the incorporated synthetic data. The data item to be protected is the value in the unweighted, undistorted micro-data, which corresponds to a particular variable in X_{askt} for aggregations k and individuals in demographic category (a, s). The aggregations presented here are for county-level geography and NAICS industry group (4-digit),

[12] β is the confidential noise limit in the QWI disclosure avoidance system.

although the tables have also been computed for the other industry classifications.[13] The NAICS industry group classification was chosen at the county geography because this stratification has the largest number of small items among the QWI publication tables and therefore the most item suppressions in the released data. The first set of results are cross-tabulations that show how the values of the unprotected microdata are perturbed by each of the protection layers, as we illustrated in section 7 for the currently published QWI. The second set of results illustrate the how the time series properties the data are affected, as we illustrated in section 8 for the currently published QWI.

9.4.1. Impact on the Extent of Protection

The three panels in Table 11 show how the unweighted, undistorted data are affected by distortion, distortion plus weighting and then synthesis, respectively, for the variable B_{djt}.[14] The data are rounded to the nearest unit in the column, where 5^+ contains data values of 5 or more. The main elements of interest in the tables are the percentages along the diagonal, which show how often the value of the confidential micro-data (the rows of the table) is unchanged by the particular protection method (the columns of the table). If the value on the diagonal is too high, the data are insufficiently protected.

By reviewing the rows of Table 11(a), it is clear that the noise infusion does not adequately protect single individuals in an age/sex category and, by extension, beginning-of-period employment in cells of size 1. It is also clear that values of 2 are adequately protected if the required inference error rate is set at 10% or more. Of course, one cannot suppress only values of 1 which is why more than one item must be suppressed in the current protocols. Table 11(b) shows the effects of combining noise infusion with weighting. Again, the percentage on the diagonal for values of 1 is still high. Finally, Table 11(c) shows that introducing our synthetic data methods adequately protects the small values and suppression is no longer needed. The results for the variables not presented here are similar.

9.4.2. Analytical Validity of Time-series Properties

The conclusion that the synthesizer sufficiently protects the QWI micro data is positive, but it is also of interest how the statistical properties of the data hold up. The current use of suppression is problematic for users of the data because they are forced to model the missing data themselves based on the released data, or ignore it. Here we show that replacing the suppressions with synthetic data not only retains the

[13]The results presented in Sections 7 and 8 are based on an earlier vintage of the QWI measures that were, at that time, not reported by NAICS. The experimental results presented in this section are based on a later vintage for which NAICS codes were available. The core data processing procedures between these two vintages are roughly identical.

[14]In addition to B, the same analysis has been computed for each of the remaining ten synthesized variables E_{djt}, H_{djt}, R_{djt}, A_{djt}, S_{djt}, M_{djt}, F_{djt}, FA_{djt}, FS_{djt}, and $H3_{djt}$. Tables for B_{djt}, H_{djt}, S_{djt}, F_{djt}, FS_{djt}, and $H3_{djt}$ for Maryland are provided by Gittings (2009). The conclusions are qualitatively the same for all variables.

Table 11: Extent of protection in hybrid system: B

(a)

Unweighted/Undistorted vs. Unweighted/Distorted

	0	1	2	3	4	5
0	99.61	0.39	0	0	0	0
1	0	98.57	1.43	0	0	0
2	0	1.04	96.1	2.85	0	0
3	0	0	2.19	93.21	4.6	0
4	0	0	0	7.3	82.52	10.18
5	0	0	0	0	1.6	98.4

(b)

Unweighted/Undistorted vs. Weighted/Distorted

	0	1	2	3	4	5
0	99.19	0.81	0	0	0	0
1	0.14	89.29	10.56	0.01	0	0
2	0.04	1.39	67.45	30.7	0.42	0
3	0.03	0.04	2.19	50.99	42.76	3.99
4	0.03	0.02	0.03	3.07	41.04	55.81
5	0.01	0	0	0.02	0.33	99.64

(c)

Unweighted/Undistorted vs. Synthesized

	0	1	2	3	4	5
0	99.17	0.82	0.01	0	0	0
1	7.85	84.74	6.62	0.78	0.01	0
2	0.51	11.93	61.06	24.14	2.24	0.12
3	0.06	0.76	7.53	47.5	39.13	5.02
4	0.03	0.11	0.93	7.4	38.84	52.69
5	0.01	0.01	0.01	0.11	0.71	99.16

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent row percentages and sum to 100.

statistical properties of the underlying data but also yields an improvement over modeling the missing data externally.

Table 12: Distribution of the Error in the First Order-serial Correlation: Unweighted/Undistorted vs. Unweighted/Distorted

$\Delta r = r - r^*$

Percentile	B	H	R	E	A	S
99	0.067	0.058	0.042	0.066	0.061	0.061
95	0.036	0.035	0.025	0.036	0.036	0.036
90	0.025	0.025	0.015	0.024	0.025	0.025
75	0.009	0.009	0.005	0.009	0.01	0.01
50	0	0	0	0	0	0
25	-0.008	-0.011	-0.005	-0.009	-0.01	-0.01
10	-0.023	-0.028	-0.016	-0.023	-0.026	-0.026
5	-0.035	-0.038	-0.025	-0.037	-0.036	-0.038
1	-0.061	-0.063	-0.045	-0.065	-0.059	-0.065

	F	FA	FS	H3
99	0.059	0.056	0.054	0.056
95	0.032	0.034	0.033	0.035
90	0.021	0.024	0.023	0.024
75	0.008	0.009	0.008	0.008
50	0	0	0	0
25	-0.008	-0.009	-0.008	-0.009
10	-0.022	-0.026	-0.024	-0.026
5	-0.033	-0.037	-0.035	-0.038
1	-0.061	-0.061	-0.059	-0.062

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent the difference between the autocorrelation coefficients the percentile designated by the rows.

Our analysis of the validity of the time-series inferences is based on the first-order serial correlation coefficient eistimated using maximum likelihood for each variable for age/sex groups at the county × NAICS Industry Group level of aggregation. To judge the analytical validity we consider the distribution of the difference between the serial correlation coefficient using the unweighted/undistorted data and the estimate produced when using one of the layers of protection. A difference of zero between these two estimates would indicate no bias and preservation of the time-series properties. Tables 12 to 14 show the distribution of

Table 13: Distribution of the Error in the First-order Serial Correlation: Raw vs. Published Data

$\Delta r = r - r^*$

Percentile	B	H	R	E	A	S
99	0.318	0.592	0.538	0.325	0.602	0.652
95	0.153	0.326	0.221	0.158	0.294	0.304
90	0.086	0.196	0.136	0.084	0.173	0.181
75	0.021	0.06	0.049	0.019	0.053	0.053
50	-0.002	-0.006	0	-0.003	-0.006	-0.008
25	-0.029	-0.075	-0.063	-0.031	-0.071	-0.078
10	-0.085	-0.188	-0.162	-0.091	-0.17	-0.19
5	-0.15	-0.289	-0.242	-0.154	-0.258	-0.288
1	-0.393	-0.551	-0.474	-0.426	-0.505	-0.531

Percentile	F	FA	FS	H3
99	0.366	0.606	0.656	0.648
95	0.181	0.317	0.352	0.359
90	0.108	0.196	0.218	0.229
75	0.025	0.068	0.07	0.087
50	-0.001	-0.004	-0.004	-0.001
25	-0.026	-0.07	-0.084	-0.082
10	-0.077	-0.169	-0.2	-0.19
5	-0.141	-0.262	-0.3	-0.286
1	-0.419	-0.499	-0.538	-0.527

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent the difference between the autocorrelation coefficients the percentile designated by the rows.

Table 14: Distribution of the Error in the First-order Serial Correlation: Raw vs. Hybrid Data

	$\Delta r = r - r^*$					
Percentile	B	H	R	E	A	S
99	0.107	0.246	0.223	0.096	0.238	0.263
95	0.056	0.144	0.126	0.051	0.134	0.166
90	0.036	0.104	0.092	0.033	0.099	0.12
75	0.011	0.05	0.043	0.01	0.047	0.058
50	-0.004	0.003	0.007	-0.005	0.005	0.007
25	-0.03	-0.043	-0.013	-0.03	-0.032	-0.032
10	-0.065	-0.109	-0.067	-0.064	-0.089	-0.095
5	-0.092	-0.167	-0.12	-0.093	-0.142	-0.145
1	-0.166	-0.295	-0.257	-0.176	-0.257	-0.268

	F	FA	FS	H3
99	0.227	0.26	0.157	0.294
95	0.112	0.145	0.076	0.185
90	0.075	0.108	0.049	0.133
75	0.031	0.051	0.018	0.066
50	0.002	0.007	-0.003	0.009
25	-0.018	-0.03	-0.034	-0.037
10	-0.056	-0.095	-0.082	-0.128
5	-0.088	-0.144	-0.116	-0.194
1	-0.187	-0.277	-0.233	-0.343

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent the difference between the autocorrelation coefficients the percentile designated by the rows.

this difference for each variable under each protection scheme. Table 12 shows the difference between the autocorrelation coefficient estimated from the unweighted/undistorted data versus the unweighted/distorted data. Table 13 compares the underlying micro data with the weighted/distorted data that suppresses the appropriate small values, and Table 14 displays the results when the suppressions are replaced with synthetic data.

It is no surprise that there is almost no bias when only distortion is used since it was designed to preserve these time series properties. However, the results comparing the synthetic data (Table 14) are clearly superior to those with suppressions (Table 13). Use of the synthetic data introduces very little bias compared to Table 12. with almost all of the bias being in the tails, whereas the difference is largely zero for much of the distribution. Across the board, this bias is less than that shown in Table 13 and the results clearly demonstrate that the time-series properties of the data are preserved when the synthesizer component is added to the dynamically consistent noise infusion.

10. Concluding remarks

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

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 induced on average by the confidentiality protection mechanism, and the distributions of bias are tightly centered around the modal/median bias of zero. Similar results were found by Cohen and Li (2006), who studied the Evans, Zayatz and Slanta procedure and found that it had statistical disclosure limitation properties that were never worse than primary/complementary cell suppression, and analytical validity that was always better.

A data synthesizer was proposed and experimentally implemented. This synthesizer replaces sensitive item values that are suppressed in the production Quarterly Workforce Indicators. We show that the hybrid system using synthetic data to replace the suppressed items and items that are linked by identity to those suppressed items offers sufficient protection and also preserves the time-series properties of the underlying confidential data. In particular, noise infusion and weighting alone often do not provide adequate protection for small data items, but the hybrid synthesizer clearly protects those values sufficiently. Furthermore, not only are the data protected but the time-series properties are retained in the hybrid synthetic data and yield

a substantial improvement over the current published data with suppressions.

Acknowledgements

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, Erika McEntarfer, Stephen Tibbets for helpful comments on many earlier versions of this paper. Parts of this paper were previously presented under the title "Confidentiality Protection in the Census Bureau's Quarterly Workforce Indicators" at the Joint Statistical Meetings, 2005. Parts of this paper were available as LEHD TP 2006-02, and as parts of the theses of Stephens (2007) and Gittings (2009). This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD) at the Center for Economic Studies (CES), which has been 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 AG018854-02, and the Alfred P. Sloan Foundation. Abowd and Vilhuber acknowledge additional funding through NSF Grants SES-0922005, SES-1042181, SES-0922494, and TC-1012593. Some or all of the data used in this paper are confidential data from the LEHD Program. All results have been reviewed to ensure that no confidential information is disclosed. The U.S. Census Bureau supports external researchers' use of these data through the Research Data Centers (see www.census.gov/ces).

References

Abowd, J. M., Haltiwanger, J. C., Lane, J. I., May 2004. Integrated Longitudinal Employee-Employer Data for the United States. American Economic Review 94 (2).

Abowd, J. M., Lane, J., 2004. New approaches to confidentiality protection: Synthetic data, remote access and Research Data Centers. In: Domingo-Ferrer, J., Torra, V. (Eds.), Privacy in Statistical Databases. Springer-Verlag, Berlin, pp. 282–289.

Abowd, J. M., Stephens, B. E., Vilhuber, L., Andersson, F., McKinney, K. L., Roemer, M., Woodcock, S. D., 2006. The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. Technical paper TP-2006-01, U.S. Census Bureau, LEHD and Cornell University.

Abowd, J. M., Stephens, B. E., Vilhuber, L., Andersson, F., McKinney, K. L., Roemer, M., Woodcock, S. D., 2009. The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. In: Dunne, T., Jensen, J. B., Roberts, M. J. (Eds.), Producer Dynamics: New Evidence from Micro Data. University of Chicago Press.

Abowd, J. M., Woodcock, S. D., 2001. Disclosure limitation in longitudinal linked data. In: Doyle et al. (2001), Ch. 10, pp. 215–278.

Cohen, S. H., Li, B. T., December 2006. A comparison of data utility between publishing cell estimates as fixed intervals or estimates based upon a noise model versus traditional cell suppression on tabular employment data. Tech. rep., Bureau of Labor Statistics, Washington, D.C.
URL http://www.bls.gov/ore/pdf/st060100.pdf(citedonMarch20,2007)

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

Dalenius, T., 1977. Towards a methodology for statistical disclosure control. Statistisk tidskrift Statistical review 15, 429–444.

Domingo-Ferrer, J., Torra, V., 2001. Disclosure control methods and information loss for microdata. In: Doyle et al. (2001), Ch. 5.

Doyle, P., Lane, J. I., Theeuwes, J. J., Zayatz, L. V. (Eds.), 2001. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., Roehrig, S. F., 2001a. Disclosure limitation methods and information loss for tabular data. In: Doyle et al. (2001), Ch. 7, pp. 135–166.

Duncan, G. T., Jabine, T. B., de Wolf, V. A. (Eds.), 1993. Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. National Academy Press, Washington, D.C.

Duncan, G. T., Keller-McNulty, S. A., Stokes, S. L., December 2001b. Disclosure risk vs. data utility: The R-U confidentiality map. Technical Report 121, National Institute of Statistical Sciences, Research Triangle Park, NC, http://www.niss.org/technicalreports/tr121.pdf (cited on March 20, 2007).
URL http://www.niss.org/technicalreports/tr121.pdf

Duncan, G. T., Lambert, D., 1986. Disclosure-limited data dissemination. Journal of the American Statistical Association 81, 10–28.

Evans, T., Zayatz, L., Slanta, J., december 1998. Using noise for disclosure limitation of establishment tabular data. Journal of Official Statistics 14, 537–551.

Federal Committee on Statistical Methodology, December 2005. Statistical policy working paper 22 (second version, 2005): Report on statistical disclosure limitation methodology. Tech. rep., Office of Management and Budget, Washington, D.C.
URL http://www.fcsm.gov/working-papers/SPWP22_rev.pdf(citedonMarch20,2007)

Fienberg, S. E., Willenborg, L. C. R. J., December 1998. Introduction to the special issue: Disclosure limitation methods for protecting the confidentiality of statistical data. Journal of Official Statistics 14 (4), 337–345.

Gelman, A. B., Carlin, J. S., Stern, H. S., Rubin, D. B., 2003. Bayesian Data Analysis, 2nd Edition. Chapman and Hall.

Gittings, R. K., 2009. Essays in labor economics and synthetic data methods. Ph.d., Cornell University.

Kullback, S., Leibler, R., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–86.

Massell, P., Zayatz, L., Funk, J., 2006. Protecting the confidentiality of survey tabular data by adding noise to the underlying microdata: Application to the commodity flow survey. Springer, New York.

Massell, P. B., Funk, J. M., 2007. Recent developments in the use of noise for protecting magnitude data tables: Balancing to improve data quality and rounding that preserves protection. In: Federal Committee on Statistical Methodology Conference.
URL http://www.fcsm.gov/07papers/Massell.IX-B.pdf

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

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

Stephens, B., 2007. Essays on firm compensation policy and confidentiality protection and imputation in the Quarterly Workforce Indicators. Ph.d., University of Maryland, College Park.

Willenborg, L., de Waal, T., 2001. Elements of Statistical Disclosure Control. New York: Springer-Verlag.

This version: $Id: fcsm2012-noise-synthetic.tex 2843 2012-04-17 03:54:06Z vilhu001 $