WP. 49
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (ix): Statistical disclosure limitation for table and analysis servers: how to make outputs of modern data access infrastructures safe

Adaptation of EZS Disclosure Method to the Quarterly Census of Employment and Wages program (QCEW)

Prepared by Michael Buso, Shail Butani, David Hiles, Spencer Jobe, Ali Mushtaq, Santanu Pramanik and Michael Yang, U.S. Bureau of Labor Statistics and Fritz Scheuren, NORC at the University of Chicago, U.S.A.

Adaptation of EZS Disclosure Method to QCEW

Michael Buso*, Shail Butani, David Hiles, Spencer Jobe [1]
Ali Mushtaq, Santanu Pramanik, Fritz Scheuren** and Michael Yang [1]

* U.S. Bureau of Labor Statistics, 2 Massachusetts. Ave., NE # 4985, Washington, D.C. 20212, USA, butani.shail@bls.gov

** NORC at the University of Chicago, 55 East Monroe # 2000, Chicago, IL 60603-5713, USA, scheuren-fritz@norc.org

1 Introduction

The Bureau of Labor Statistics (BLS) has contracted with NORC at the University of Chicago (NORC) to develop alternative disclosure limitation methodologies for the Quarterly Census of Employment and Wages program (QCEW). The objectives of the NORC-BLS joint effort are: (1) Increase the amount of QCEW data that is disclosed (reduce suppressions); (2) Reduce the time spent in performing disclosure processing; (3) Develop a disclosure limitation approach that would reflect best practices of the statistical community; and 4) Improve the ability to safely use QCEW data for alternate aggregations.

The Quarterly Census of Employment and Wages (QCEW) program is one of the Bureau of Labor Statistic's Federal-State cooperative programs. The program receives establishment level reports of employment and wages for employers covered by Unemployment Insurance (UI) programs in the 50 States, the District of Columbia, Puerto Rico, and the U.S. Virgin Islands. These reports are received quarterly for each of 9 million covered establishments that employ nearly 130 million workers. Each establishment report is coded by location, and by detailed industry of activity. The employment and wage data are tabulated at a variety of aggregation levels for publication and analysis. The individual employer reports are considered sensitive, and as a result, the publication process poses a significant tabular disclosure challenge. The earliest publications of the 70 year old program show that primary cell suppression and complementary cell suppression have been used to protect employers since those early years.

The detail that is available in the QCEW program is in great demand, not just for current data products, but for additional or alternative levels. For example, the finest geographic detail that is currently published is county level, but it is desired below that level for a variety of areas such as legislative districts, cities, and central business districts. Many statistical programs face a trade off between disclosure risk and data utility, and this is very much the case for the QCEW. Even given the detail

[1] The views expressed are those of the authors and do not represent official positions of BLS and NORC .

available from the current products, the Bureau of Labor Statistics has been concerned for some time about the disclosure limitation methodology used with the QCEW program. The increased computational capacities and the improved analysis methods available to data crackers have led to vulnerabilities that were not considered significant even a decade ago.

2 Current status

The QCEW program prepares tabulations of monthly employment levels, total wages, and the number of establishments by industry and area. Although sometimes referred to as estimates, QCEW data are better called total counts. They are not estimated totals based on a sample, but are instead exact totals of mandatory data reported from a quarterly universe count. This exactness is valuable, but it comes at a cost. Approximately 60 percent of QCEW totals are "withheld" and cannot be released under the current disclosure limitation process.

Industry detail is at the 6-digit North American Industry Classification System (NAICS) level, meaning over 1,200 detailed industries. Higher levels of industry aggregation are prepared as well, for a total of nearly 2,400 industries at various levels. In addition to industry codes, all establishments are assigned an ownership code, depending on whether they are a private sector establishment or a Federal, State, or local government establishment. A relatively small number of tabulations are stratified by establishment size, the strata set by employment levels. QCEW data are prepared and released on a quarterly basis through the BLS web site, and are also available from many state agency web portals and publications.

Geographic detail is at the county, Metropolitan Statistical Area (MSA), state, and national levels, for a total of nearly 4,000 areas. In addition, area totals without industry detail are produced for Micropolitan Statistical Areas (MicroSAs) and Combined Statistical Areas (CSAs). The vast majority of the underlying QCEW establishment reports are assigned latitude and longitude coordinates based on reported addresses. This makes them a possible base for rich sub-county economic data reports and analysis.

Out of about 3.6 million QCEW cells, approximately one-half are suppressed in the primary disclosure processing phase. The secondary/complementary phase suppresses about another 10 percent of the cells, leaving a total of about 40 percent published. The percent disclosed varies by industry and area detail, with the more detailed levels having higher suppression rates.

3 The EZS Noise Method

Research at the Census Bureau in the 1990s led to the 1998 Journal of Official Statistics article by Timothy Evans, Laura Zayatz, and John Slanta, *Using Noise for Disclosure Limitation of Establishment Tabular Data.* This work set a pathway for alternatives to cell suppression that has been pursued by a number of programs both in the U.S. and beyond. That seminal work proposed a class of methods, techniques, and principles that have been expanded on by other studies, and implemented in a number of programs. These methods and principles have come to be known as the EZS noise method.

The QCEW program has been researching noise method techniques to both increase the published detail and to respond to the disclosure vulnerabilities of cell suppression methods. Although, to the extent that tabulations might depart from the exactness that characterized data from the incumbent system, it was understood that the method would need to still result in a reasonable, perhaps measurably accurate picture of local and national economies, both in levels and trends. That is, products of the new method should be reasonably comparable to the products of the past, and support reasonably detailed comparisons over time involving both the past and future data.

4 Basic Elements

It isn't appropriate here to review all aspects of the EZS noise method or model, but a few key features of that model will be enumerated as they are especially relevant to the adaptability of the method to QCEW.

1) The model is oriented to data for tables of magnitude variables derived from establishment programs. Frequency data are not discussed.
2) The model perturbs all responses. Responses that contribute to both sensitive and non-sensitive cells are perturbed.
3) The perturbation is by multiplication by randomly selected factors.
4) The perturbation factors are selected from a symmetrical distribution. Changes can occur in either a positive or negative direction.
5) The model allows for a minimal perturbation threshold to be set. If used, these thresholds ensure that all responses are perturbed a minimal percentage. Enforcing a minimal perturbation seems to be a common aspect of most implementations of EZS.
6) Consistency can be provided in the application of EZS for a given set of company totals and its establishment responses by drawing all perturbations for a company from the same side of the symmetric distribution.
7) The perturbed establishment reports are aggregated to publication cells. If the program is a sample survey, the perturbation factor is diminished in

proportion to the sample weight. A feature that does not applicable to QCEW, as all of the data is self-representing, that is, has a weight of one.

8) The unperturbed data are also examined to identify sensitive cells, that is, those cells that fail dominance tests or other criteria for sensitivity. The results of those tests are transferred as flags to the perturbed aggregates which are subsequently suppressed for publication.
9) Unperturbed cell aggregates are also compared to the perturbed cell aggregates. For those cells whose perturbed and unperturbed aggregates differ by more than a threshold percentage the perturbed aggregates are also flagged and suppressed. The flags for sensitive cells and high perturbation cells are not distinguished in publications.
10) The EZS model, unlike current practice of QCEW and many other programs, calls for no complementary suppressions to be performed.
11) The model's suppression of data for sensitive cells provides an appearance of protection beyond that provided by the random factors. However, without complementary suppressions, some perturbed values for sensitive cells might be more or less readily derived arithmetically.
12) The model does not ratio adjust or otherwise force perturbed aggregates to match or closely approximate corresponding unperturbed aggregates. Several extensions to the core EZS method have addressed this issue, and it remains an active research area. BLS expects the QCEW noise research to result in some mechanism to address this.

5 QCEW Characteristics that Challenge the Basic EZS Model

5.1 Employment as Both a Frequency Variable and a Magnitude Variable

Employment has characteristics of both a frequency variable and a magnitude variable. This is often true of many variables that are treated exclusively as magnitude variables, but, in the case of QCEW employment values, the dualistic nature is much more relevant. For cell totals, especially for cells with numerous contributors, employment is well treated as a magnitude variable. It has discrete values, but, so do the rounded values of a number of continuous magnitude variables. On the other hand, employment reported by small establishments must often be considered a frequency variable. The establishment either had an employee, or it did not; had two, or did not, had two and added one; etc. To the individual employee added (hired) or lost (dismissed), the difference is significant and seen as an individual attribute. To the extent that employment can be treated as a magnitude variable, the EZS model may be more relevant. To the extent that it is a frequency variable, the situation is less clear. In our review of the EZS literature we found little discussion of the relevance of the method to frequency variables.

5.2 Small Establishment Issues

There are three issues related to small establishments that are particularly challenging: (1) Employment levels do not perturb well multiplicatively, since they have to be rounded to an integer value. (2) There are a large number of zeroes in the population and (3) the distribution is grossly skewed. See Figure 1, for a depiction of the distribution of QCEW establishments by employment size.

Over sixty percent of the nearly 9 million establishment reports in the QCEW program are of employment of less than five. These values do not perturb well with EZS type multiplicative factors. If factors were used which were large enough to displace the values of these small establishments from their reported values, they would have an enormous impact on larger values.

It is typical for about fifteen (15) percent of establishments in any given period to report employment of zero. These are a mix of establishments. Some are reporting zero for the reference period – the pay period that includes the twelfth of the month – while they have positive employment in other periods. Other zero employment establishments are start-ups that as yet have no employees. Another type are seasonal establishments or those subject to business cycle shutdowns. All of these are valid employment reports by one measure or another. Zero is clearly a number that does not displace well by multiplicative factors. Overall, then, the left tail of the highly skewed QCEW establishment employment reports poses a challenge to the EZS and we sought another approach.

In summary, the EZS noise method provides for protecting the data of medium and large establishments, yet falls short in the protection and usefulness of data from small establishments.

	Establishment Size	
	Small	**Medium and Large**
Performance	Research	"Solved"
Protection	Research	"Solved"

Seam

6 NORC's Mixed Approach: Deterministic mixture and synthetic and noise-treated data

The BLS contracted with NORC at the University of Chicago, to help address this problem NORC suggested an approach that combined the use of multiplicative noise to protect medium and larger establishments values with the use of synthetic data to protect the smaller establishments. NORC has developed and studied two prototypes that take significantly different variations on that approach. Details on the methodologies and results are in a forthcoming paper Evaluation of Four Disclosure Limitation Models for the QCEW Program (Yang, M., et. al.) planned for International Conference on Establishment Surveys, June 2012. That paper compares their two approaches with two earlier prototypes based more closely on the EZS approach.

One critical goal for the project overall, was to have employment and wage distribution of the fuzzy data very closely approximate the reported distribution at the establishment level, and the prototypes have been able to achieve that. See figures 2 through 5 for a depiction of the perturbed employment levels that result from the recent models.

The NORC prototypes have also implemented a raking procedure to ensure that many of the high level aggregates generated from the perturbed micro data more closely approximate the results of aggregating the unperturbed data than would result from the random noise method alone.

Overall, the results of the NORC research, and the performance of the prototypes they have developed, while preliminary, have been very promising.

7 Tables, figures and references

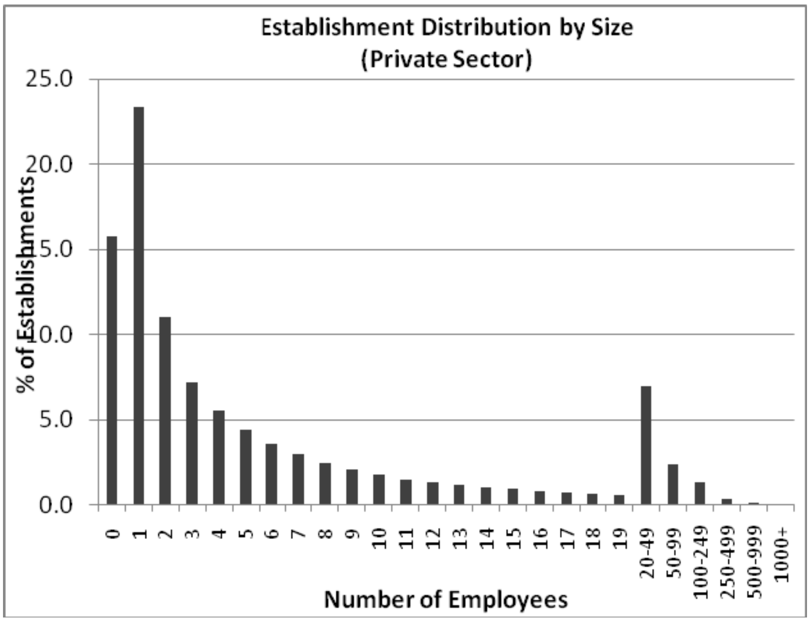

Fig 1 Establishment distribution by employment size

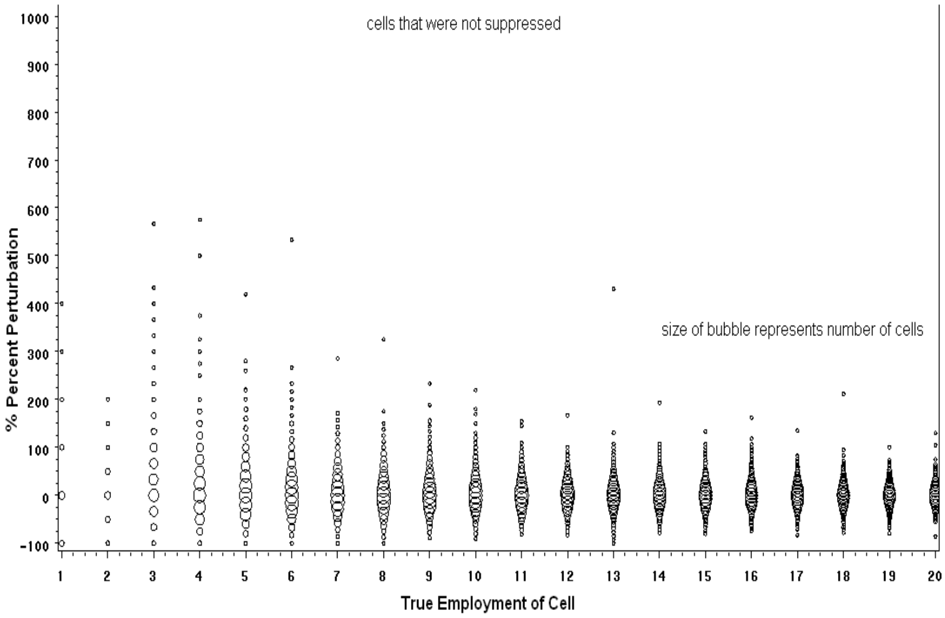

Fig 2 Perturbation of published cells with 1-20 employees

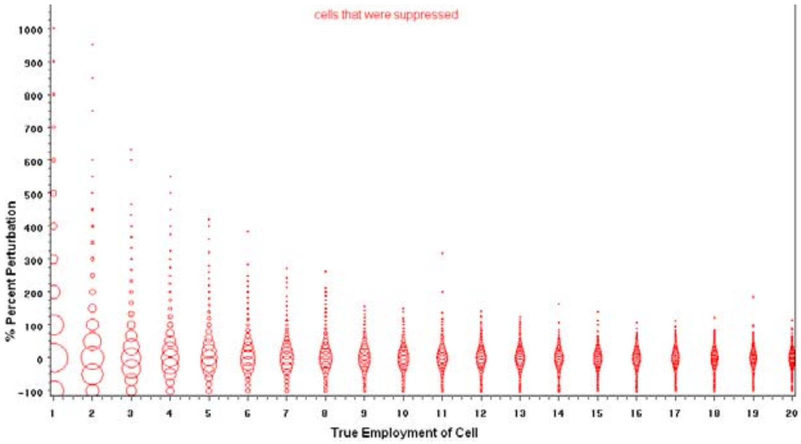

Fig 3 Perturbation of suppressed cells with 1-20 employees

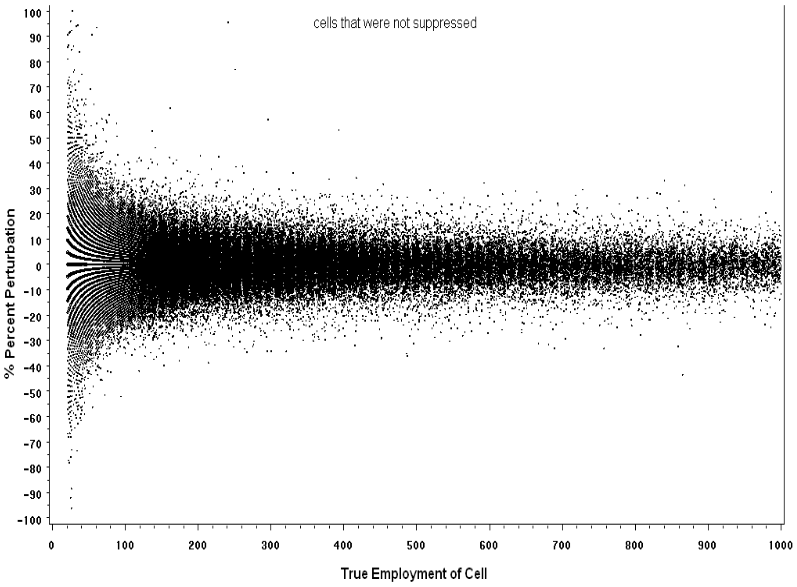

Fig 4 Perturbation of published cells with 21-1000 employees

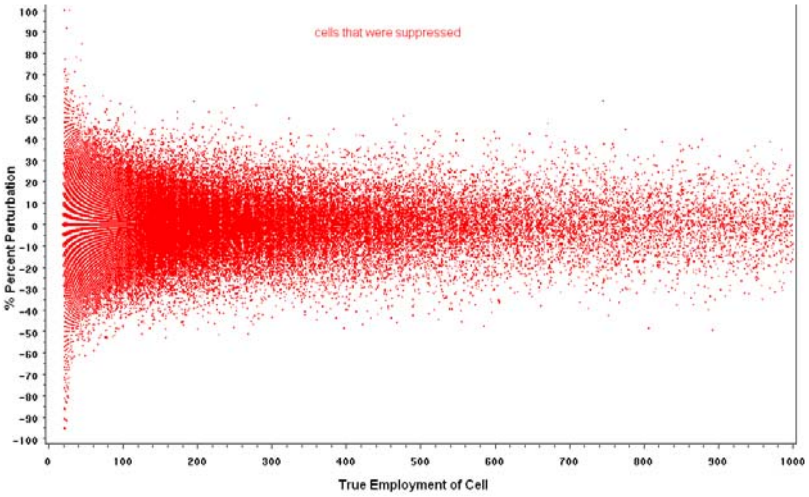

Fig 5 Perturbation of suppressed cells with 21-1000 employees

References

Business Employment and Dynamics (BED) Bulletin. Bureau of Labor Statistics, http://www.bls.gov/bdm/ .

Evans, B. Timothy, (1997). *Effects on Trend Statistics of the use of Multiplicative Noise for Disclosure Limitation*, Proceedings of the Section on Government Statistics, American Statistical Association, pp. 303 - 307 www.amstat.org/meetings/ices/2000/proceedings/S49.pdf

Evans, B.T., Zayatz, L., and Slanta, J. (1998). *Using Noise for Disclosure Limitation of Establishment Tabular Data*, Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 537—551.

Federal Committee on Statistical Methodology (1994), Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology, Washington, DC: US Office of Management and Budget.

Federal Committee on Statistical Methodology (2005), Statistical Policy Working Paper 22 (Second version, 2005): Report on Statistical Disclosure Limitation Methodology, Washington, DC: US Office of Management and Budget.

Jiang, J. & Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15, pp. 1-96.

Pursey, S., Labelle-Blanchet, S. (2011). *Disclosure Control: the Application of Random Noise to the Microdata in the Release of Aggregate Estimates"*, Working Paper, Business Survey Methods Division, Statistics Canada.

Quarterly Census of Employment and Wages (QCEW) Bulletin. Bureau of Labor Statistics, http://www.bls.gov/cew/cewbultn09.htm

Rao, J. N. K. (2003). *Small area estimation*. Wiley Series in Survey Methodology. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons].

Schafer, J. (2001). *A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data*, Journal of the American Statistical Association, Vol. 96, No. 454, pp. 730—745.

Yang, M., et al forthcoming (2012). *Evaluation of Four Disclosure Limitations Models for the QCEW Program*, Proceedings of the International Conference on Establishment Surveys.