Modeling Aggregates for Reliability and Confidentiality of Output with Application to the QCEW program of BLS

Avi Singh, Santanu Pramanik, Michael Yang, Fritz Scheuren[1]

[1]NORC at the University of Chicago, 55 E. Monroe Street, 30th Floor, Chicago, IL 60603

Abstract

The QCEW program of BLS publishes tabulations of employment and wages by industry and geography. The BLS has been concerned for some time about the current cell suppression method for disclosure limitation because it results in substantial data suppression that compromises the quality and utility of the QCEW data. To address such concerns, BLS has been conducting research on the application of the random noise method (input treatment) to QCEW as an alternative to cell suppression. The goal is to release significantly more data and to respond to new disclosure vulnerabilities. In this paper, we explore another alternative based on the application of small area modeling techniques to disclosure limitation for the QCEW by modeling aggregates. In this new application of small area modeling we exploit the built-in perturbation of direct estimates (here, true totals) by the synthetic component. The current research is in progress and requires further empirical validation.

Key Words: disclosure limitation, cell suppression, random noise method, small area estimation, output treatment

1. Introduction

The Quarterly Census of Employment and Wages (QCEW) program of the Bureau of Labor Statistics (BLS) collects 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 (QCEW bulletin). The employment and wage data are tabulated at a variety of aggregation levels, defined by industry and geography, for publication and analysis. In accordance with BLS policy, data provided to the Bureau in confidence are used only for specified statistical purposes. In particular, the published tabular data requires protection. At the same time it is important to preserve the analytical utility of the data as QCEW program reflects the economic picture of the nation at a very detailed level.

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. Geographic detail is at the county, Metropolitan Statistical Area (MSA), state, and national levels, for a total of nearly 4,000 areas. QCEW data are prepared and released on a quarterly basis at the aggregate (cell) level where cell is defined by industry, geography and ownership. Currently, to protect the tabular data from possible disclosure vulnerabilities, BLS uses cell suppression method.

The BLS has been concerned for some time about the current cell suppression method used with the QCEW program. The most obvious disadvantage of the cell suppression method is that it has resulted in a large amount of data being suppressed, compromising the quality and utility of the QCEW data products. In particular, this method suppresses much information that is not at risk for disclosure. Any cell that is used as a complementary suppression represents data that could have been published if there were other ways of protecting the sensitive cells. Overall, approximately 60% of 3.6 million QCEW cells are suppressed under the current disclosure protection process.

QCEW data are in great demand, not only for the current data products, but also in greater detail. For example, tabulations for sub-county areas will be very useful for policy studies involving legislative districts, cities, central business districts, and so on. However, data publication for more detailed geographic areas will be subjected to even higher suppression rates under the current cell suppression method. To address such concerns, BLS has been conducting research on the random noise method toward extending that model for application to QCEW. The goal is to release significantly more data and to respond to new disclosure vulnerabilities.

In recent years, the random noise method has been gaining wider use in statistical agencies to protect respondent data from unintended disclosure. The original random noise method was developed in the late 1990s by Tim Evans, Laura Zayatz, and John Slanta (Evans et al. 1998). The so-called EZS noise method takes a micro approach to disclosure limitation: a multiplier, or noise factor, is applied at the unit level rather than at the cell level and falls under the category of *input treatment* for disclosure limitation. Under this method, a noise factor is applied to each unit prior to any tabulation, which guarantees that different tabulations, from the lowest to the highest level, are consistent. Under a BLS contract, NORC at the University of Chicago (NORC) has 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. Details on the methodologies and results are in a forthcoming paper (Yang et al. 2012), planned for International Conference on Establishment Surveys, June 2012. NORC's approach can also be viewed as an input treatment for disclosure limitation.

It is well-known that the perturbed totals, obtained after applying random noise method, are unbiased estimates of true totals, under suitably chosen parameters of the noise distribution (Evans et al. 1998, Yang et al. 2010). However, in practice, the random noise method can lead to considerable bias for some cells as this method does not provide any direct control at the cell level. While perturbing, this method does not take into account the variability of the study attributes (employment, wages etc.) at the cell level, size of the cell (number of establishments in the cell). In this paper, we propose an alternative based on small area modeling techniques for *output treatment* of disclosure. Output refers to domains (cells) of interest defined by cross-classifying geography, industry and ownership, such as, county by 6-digit NAICS code (NAICS6) at the private sector, county by 2-digit NAICS code (NAICS2), state by NAICS6 etc.

2. MARC Method: an Application of Small Area Estimation Technique

Modeling Aggregates for Reliability and Confidentiality (MARC) method can be viewed as a new application of small area estimation (SAE, Rao 2003) in the context of disclosure control. Small area models are known to produce precise estimates of small domains, compared to the direct survey estimates, by combining survey data with administrative records. The gain in precision is substantial for domains having small sample size. Small area technique has the ability to borrow strength from similar areas to compensate for small sample sizes in some domains. It basically produces an estimator which is a weighted combination of direct survey estimate and synthetic estimate, weights being proportional to the relative precision of the direct and synthetic estimate. In other words, for a large domain, it gives more weight to the direct estimate and less to the synthetic; for small domains it's the opposite. For example, under a typical area level small area model, the indirect estimate of true small area means θ_i is given by $\widehat{\theta_i} = (1 - B_i)y_i + B_i x_i'\hat{\beta}$, where $B_i = V_i/(\sigma_v^2 + V_i)$, is the shrinkage factor that shrinks the direct survey estimate y_i (having sampling variability V_i) to the regression synthetic estimate $x_i'\hat{\beta}$ with σ_v^2 being the model variance. We have exploited this built-in perturbation of direct estimates (here, true totals) by the synthetic component in the context of the QCEW program. Small area estimates (SAEs) have dual property of improving reliability as well as confidentiality. This method of output treatment has the potential to be more precise than the random noise method of input treatment. Here we have direct control on the magnitude of perturbation at the cell level.

3. Building Block Idea for MARC Method

The application of small area shrinkage technique is not straightforward in the context of QCEW program since we are interested in publishing estimates at various levels of aggregation. The question boils down to at which level the model should be defined? This is a common problem with small area estimation. Often the choice of the level of aggregation in small area modeling is not governed by adequacy of modeling assumptions but by user needs which varies from user to user. This may have a serious impact on validity of the underlying exchangeability assumption of area-specific random effects in small area models (Singh and Yuan 2010). For example, in the context of the Small Area Income and Poverty Estimates (SAIPE) program of the U.S. Census Bureau, to produce estimates of number of poor school age (5-17) children at the school district, county, and state level, 3 different models are assumed. The model is of the form:

$$y_i = x_i'\beta + v_i + e_i, i = 1 \ldots, m \text{ (number of domains)}$$

In the above model, x_i' is a vector of covariates, β is a vector of unknown regression coefficients. Area specific effects v_i's are random and it is assumed that $v_i \sim \text{iid } N(0, \sigma_v^2)$, σ_v^2 is an unknown variance component. It is also assumed that $e_i \sim \text{ind } N(0, V_i)$, where the sampling variances, V_i's are assumed to be known. Though, in practice, they are estimated by some suitable method. The exchangeability assumption of random effects at all three different level of geography (in three separate models) is generally applied across small areas. But this assumption may be incorrect and, without alternatives to this assumption, will remain throughout the inference (Malec and Muller 2008).

Ideally, one single model should drive the estimation procedure and it should be defined at the lowest level where the covariates are available. This model is termed as building-block (B-level) model (Singh and Yuan 2010). Having such a model at a very low level comes close to unit-level modeling and helps to justify, at least heuristically, the exchangeability of area-specific random effects because of similarity of building-blocks in terms of population counts. Also because of the high predictive power of the covariates at the building-block level, exchangeability assumption is more likely to hold. Model parameters are estimated at this level or at the group level (G-level) after grouping of B-level areas to avoid the problem of zero sample size or very unstable direct estimates.

4. Implementation of MARC Method in the Context of QCEW Program

For the QCEW program, two main variables of interest are employment and wages. QCEW program collects information at the establishment level and hence we have some covariates available specifically at the establishment level. To use that information effectively, we define the model at the establishment level, which essentially constitute the building blocks. Two separate univariate models are considered for employment and wages, as opposed to considering a joint model (see the discussion later).

4.1 Modeling Employment

We have explored several alternative models for employment. In this paper, we describe one of them. To this end, we divide the whole country into number of strata. Stratum is so chosen that the total employment at the stratum level is not subject to disclosure risk. We have considered state as our stratum. Our model is motivated from a multinomial-loglinear model within a stratum. For each stratum, in level 1, we model the distribution of the total number of employees over building blocks (establishments) as multinomial under a superpopulation model for the quarterly census data. In level 2, we define an establishment level log-linear model for the true proportion of employees across stratum as a function of establishment level and higher aggregate level covariates. A building-block level random effect is added to capture any residual building-block specific effect as in SAE. The advantage of considering multinomial model in level 1, as opposed to Poisson model, is that it can capture the dependencies among establishments belonging to a stratum (Lang 1996). Based on the above description, we define the model for employment as follows.

Let's assume that we know the total number of employees (N) in a particular stratum and this information is not subject to disclosure risk. Let y_b be the (true) number of employees in the b-th establishment, $b = 1, \dots, B$. The proposed model is given by:

$$y_b = x_b'\beta + \eta_b + e_b,$$

where $Var(e_b) = N\pi_b(1-\pi_b)$, $Cov(e_b, e_{b'}) = -N\pi_b\pi_{b'}; b \neq b'$, $\sum_{b=1}^{B}\pi_b = 1$, $\sum_{b=1}^{B} y_b = N$. In the above variance-covariance (V-C) structure of the random error term, the quantity π_b is the true proportion of employees in the bth establishment. The establishment specific random effect $\eta_b \sim N(0, \sigma_\eta^2)$, where σ_η^2 is an unknown variance component. The above model includes the following covariates (x_b'): broader level of industry code (2-digit NAICS, having 25 categories), size class (having 9 categories, based on the establishment-level number of employees), ownership (3 categories, state

govt., local govt., private sector- note that Federal govt. records are not subject to disclosure risk), and MSA-status (whether the establishment belongs to a MSA or not) with β being the unknown regression coefficient.

As mentioned earlier, in a typical small area model, the V-C structure is estimated based on the available information, but assumed to be known throughout the estimation procedure (Fay and Herriot 1979, Rao 2003). To assume the V-C structure to be known, we need to replace π_b by the observed proportion in the data. Once we know the V-C structure, the unknown parameters of the building-block model are β, η_b, and σ_η^2. Note that, at this B-level, we will not get a stable estimate of π_b as many of the y_b's are zero or very small. To obtain a reliable estimate of variance-covariance components, we propose to group the building blocks at a higher level.

4.2 Fitting the Model at the Group Level (G-level)

For model fitting, the building blocks (here establishments) can be suitably grouped so that observed proportions of employees in each group can be used to provide a stable estimate of the multinomial V-C matrix. We define our group based on the cross-classification of county and NAICS6. Here we need to aggregate the building-block model (Section 4.1) at the G-level, which means we sum over the establishments belonging to the same county and having the same 6-digit NAICS code. Even after this grouping, the multinomial V-C structure is retained (with π_b replaced by π_g, proportion at the group level) , which follows from the property of multinomial distribution. It is worth mentioning that, at the G-level model, the unknown parameters (β, η_b, and σ_η^2) of the B-level model remains the same. More importantly, the random effect term η_b is still defined at the building block level, where the exchangeability assumption of η_b is more likely to hold. All these parameters can be estimated using the G-level model.

Even at the G-level, there might be some groups for which total employment is zero or very small. Under such a scenario, to obtain reliable estimate of π_g, we need to collapse only those county X NAICS6 groups that do not have a minimum of 10 employees. This rule of thumb is probably reasonable as far as normal approximation to multinomial goes. We need an objective method for collapsing. For example, for fixed NAICS6 code, we will collapse by neighboring counties or for a fixed county, collapse by consecutive NAICS6 codes.

Based on 2006 quarter1 QCEW data, the total employment for the state Maryland is $N = 2{,}486{,}354$. According to our B-level model, N is distributed over 159,930 establishments. In order to get reliable estimates of proportions, we aggregate these establishments into 12,952 groups, defined by county X NAICS6. So, now N is multinomially distributed across 12,952 groups. Out of these 12,952 groups, 34% has less than 10 employees. We will collapse these 34% groups based on the objective criteria mentioned above. If our domain of interest (the aggregate level at which we want to publish the employment totals) is county X NAICS6 (which is , in fact, a crucial domain of interest for the QCEW program), then for the majority (remaining 66%) of the groups, perturbed employment total would be the Best Linear Unbiased Predictors (BLUPs), under the G-level model. BLUPs are essentially mixture of true employment totals and regression synthetic estimates. However, for the groups, for which collapsing was necessary, perturbed employment total would only be synthetic. This makes sense as the small domain sizes make these groups sensitive.

For higher level of domains, such as domains defined by county by NAICS2, state by NAICS2, we need to aggregate from the G-level model to the higher level; following the same method we adopted to derive the G-level model from B-level.

4.3 Benchmarking for Internal Consistency

For internal consistency, we need to apply hierarchical benchmarking to adjust the perturbed totals using a top-down approach. Using a ratio adjustment, we would like to ensure that the county level domain would add up to the true state total and so on.

4.4 Modeling Wages

Although employment and wages are highly correlated, we model them separately (instead of using a bivariate joint model). Conditional on employment, the wages are independent of employment is a reasonable assumption. In other words, after including employment as a covariate in the model for wages, we can capture most of the dependencies that exist between the two variables. This would keep things simpler. For modeling wages, we have considered linear mixed model (Rao 2003, Jiang and Lahiri 2006), after suitable transformation of the wage variable. For the wage model, we included employment, ownership, 2-digit NAICS code, and MSA-status as the fixed effects covariates. To obtain perturbed total wages at various aggregation levels, we follow the same procedure as we did for employment.

5. Conclusion

In this paper, we presented an outline of the output disclosure treatment as an alternative to input treatment for the QCEW program of BLS. We believe this method has the potential to produce more precise perturbed totals at the aggregate level than that of random noise method. Our output treatment can be viewed as a new application of small area estimation technique. We have exploited the inherent dual property of reliability and confidentiality of small area models. The proposed method is useful when not too many attributes require protection. Hence, makes it applicable to the QCEW program, where only attributes of interest are employment and wages. The key idea of building-block BLUPs was used to develop models for domains at various levels. These models can also be generalized to nonlinear models.

Acknowledgements

This work needs further research. We thankfully acknowledge the encouragement we received from BLS to continue working on this (in spite of not being part of the contract). In collaboration with BLS, we hope to pursue our research in future.

References

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.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74:269-277.

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

Lang, J.B. (1996). On the Comparison of Multinomial and Poisson Log-Linear Models, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1, pp. 253-266.

Malec, D. and Muller, P. (2008). A Bayesian semi-parametric model for small area estimation. Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, ed. by Clarke, B.S. and Ghosal, S. Institute of Mathematical Statistics.

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].

Singh, A.C., and Yuan, P. (2010). Building-Block BLUPs for Aggregate level small area estimation for survey data. *JSM Proceedings, Sec. Surv. Res. Meth.*

Yang, M., Mushtaq, A., Pramanik, S., Scheuren, F. (2010), Report on Existing BLS Work, *Client Report*.

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