Synthesizing Truncated Count Data For Confidentiality

Sam Hawala[1], Jerry Reiter[2] , Quanli Wang[2]
[1]U.S. Census Bureau, e-mail: sam.hawala@census.gov
[2]Duke University, e-mail: jerry@stat.duke.edu

Abstract

To maintain confidentiality national statistical agencies traditionally do not include small counts in publicly released tabular data products. They typically delete these small counts, or combine them with counts in adjacent table cells to preserve the totals at higher levels of aggregation. In some cases these suppression procedures result in too much loss of information. To increase data utility and make more data publicly available, we propose to generate synthetic values for the small counts from a Bayesian hierarchical model. We do not disturb the counts in the data tables that were considered safe. The same model allows for computation of several disclosure risk measures.

Keywords: Disclosure Limitation; Synthetic Data; Truncated Poisson.

Acknowledgements: We acknowledge Laura Zayatz and Marie Pees for bringing this problem to our attention. We also worked closely with Rodger Johnson, Roberto Marrero Cases, Charles Coleman, and Stephen Smith, all from the Census Bureau. We thank them for their cooperation.

1. Introduction

To maintain confidentiality national statistical agencies traditionally suppress information on small counts to publicly release frequency tallies (i.e. cross tabulations or contingency tables). For example, when releasing information on U.S. domestic annual county-to-county migration flows, the Census Bureau rolls up counts less than ten to release nationwide county-level migration tables. If, say only 5 people moved out of source county A_1, then these movers are added to the movers out of a neighboring source county A_2 . The agency releases the count from A_2 as long as it is higher than 10. This amounts to suppressing the information on A_1. Out of all the source-destination pairs of counties, more than 80% annually have less than 10 movers. In this and other examples, the suppressed information can be a considerable part of what is known to the agency which cannot publicly release it because of confidentiality concerns.

The problem of providing information about small counts in contingency tables, which, if published as actual counts, may create disclosure risk, has been the focus of research activity since the beginning of the previous decade. We mention in particular, controlled tabular adjustment, by Dandekar and Cox (2002), and synthetic data generation by Dandekar (2001), and Machanavajjhala *et al.,* (2008). In our paper we discuss a different

synthetic data approach in the context of county migration data. The method is adaptable to other frequency data products involving small counts.

Annually the Census Bureau derives area-to-area migration data in the United States based on data from the Statistics of Income Division (SOI) of the Internal Revenue Service (IRS). The bases for the data are year-to-year address changes reported on individual income tax returns filed with the IRS. The data products provide information on the mobility of Americans, where the people move to, and where they originate from. After compiling the data, the Census Bureau delivers the data to the IRS and to the Federal State Cooperative for Population Estimates (FSCPE). The new confidentiality protection procedure will benefit these and other data users interested in subnational population estimates. More state and local governments will see the approximate number of people moving in or out of their jurisdictions. They could better plan, allocate funds, and support research on the delivery of services. Businesses, economists, political scientists, and other researchers could use the additional information to analyze social and economic trends.

We organize the paper as follows: in section 2 we explain how we synthesize the data. In section 3 we develop disclosure risk measures.

2. Data Synthesis Methodology

Notation:

Let *k* represent a particular year. For the moment, we will omit the index *k* and assume that *k*=2010 only. Let *i* represent a source county and *j* be a destination county, where $i = 1, \cdots, 3142$ and $j = 1, \cdots, 3142$.

- n_{ij} represent the number of migrations from county *i* to county *j* in year *k*.
- N_k is the set of *ij* such that $1 \le n_{ij} \le 9$.
- *n* is the cardinality of N_k.
- $y_{ij} = n_{ij} - 1$, for all $ij \in N_k$.
- *Y* represents all y_{ij} in N_k.
- *B* is the number of distinct values of β_i,
- *G* is the number of distinct values of γ_j.

We model y_{ij} only for those cases with $ij \in N_k$. We use a truncated Poisson distribution. Let $F_8 = \Pr(0 \le y_{ij} \le 8)$. We have

$$\Pr(y_{ij} = y \mid 0 \le y_{ij} \le 8) = \frac{\exp(-\lambda_{ij})\lambda_{ij}^{y}}{y!F_8^{-1}} = \frac{\lambda_{ij}^{y}}{y!\sum_{r=0}^{8}\frac{\lambda_{ij}^{r}}{r!}} \quad (1)$$

2.1 Hierarchical Prior

We smooth the λ_{ij} by assuming that it is the sum of two main effects, β_i for source county i and γ_j for destination county j. We then set hierarchical priors on the parameters and hyper-parameters:

$$\begin{aligned} \log(\lambda_{ij}) &= \beta_i + \gamma_j \\ \beta_i &\sim N(\mu_\beta, \phi_\beta) \\ \gamma_j &\sim N(\mu_\gamma, \phi_\gamma) \\ \mu_\beta &\sim N(0, 1/25) \\ \mu_\gamma &\sim N(0, 1/25) \\ \phi_\beta &\sim Gamma(.01, .01) \\ \phi_\gamma &\sim Gamma(.01, .01) \end{aligned} \quad (2)$$

Here, there are as many β_i as there are distinct i values in N_k, and there are as many γ_j as there are distinct j values in N_k. Also, interpret ϕ_β, ϕ_γ, and 1/25 as a precision, not variance. We run this model within combinations of source-destination states (51^2 = 2601 models).

The conditional distributions for a Metropolis within Gibbs sampler are as follows:

1. For each $i' = 1, \cdots, B$, sample $\beta_{i'}$ from $f(\beta_{i'} | Y, \gamma_1, \cdots, \gamma_G)$ using a Metropolis algorithm.

   $$f(\beta_{i'} | Y, \gamma_1, \cdots, \gamma_J) \propto \left(\prod_{i,j:i=i'} \frac{(\exp(\beta_i + \gamma_j))^{y_{ij}}}{\sum_{r=0}^{8} (\exp(\beta_i + \gamma_j))^r / r!} \right) \exp(-\phi_\beta (\beta_i - \mu_\beta)^2 / 2)$$

   Therefore, at iteration s we sample a value of β_i from some proposal distribution, for example a $N(\beta_{i'}^{(s-1)}, c^2)$ with c^2 some tuning constant variance. Then we evaluate the acceptance ratio and do the M-H accept/reject step.
2. For each $j' = 1, \cdots, G$, sample $\gamma_{j'}$ from $f(\gamma_{j'} | Y, \beta_1, \cdots, \beta_B)$ using a Metropolis algorithm.

$$f(\gamma_{j'}|Y,\beta_1,\cdots,\beta_B) \propto \left(\prod_{i,j:j=j'} \frac{(\exp(\beta_i+\gamma_j))^{y_{ij}}}{\sum_{r=0}^{8}(\exp(\beta_i+\gamma_j))^r / r!} \right) \exp(-\phi_\gamma(\gamma_j-\mu_\gamma)^2/2)$$

Therefore, at iteration *s* we sample a value of $\gamma_{j'}$ from some proposal distribution, for example a $N(\gamma_{j'}^{(s-1)}, d^2)$ with d^2 some tuning constant variance. Then we evaluate the acceptance ratio and do the M-H accept/reject step.

3. Sample $\mu_\beta^{(s)}$ from $N(\mu_\beta^*, \phi_\beta^*)$ where

   $$\mu_\beta^* = \frac{B\phi_\beta \bar{\beta}^{(s)}}{B\phi_\beta + 1/25}$$

   $$\phi_\beta^* = B\phi_\beta + 1/25$$

4. Sample $\mu_\beta^{(s)}$ from $N(\mu_\beta^*, \phi_\beta^*)$ where

   $$\mu_\gamma^* = \frac{G\phi_\gamma \bar{\gamma}^{(s)}}{G\phi_\gamma + 1/25}$$

   $$\phi_\gamma^* = G\phi_\gamma + 1/25$$

5. Sample $\phi_\beta^{(s)}$ from $Gamma(a,b)$ with

   $$a = .01 + B/2$$

   $$b = .01 + \sum(\beta_i - \bar{\beta}^{(s)})^2/2$$

6. Sample $\phi_\gamma^{(s)}$ from $Gamma(a,b)$ with

   $$a = .01 + G/2$$

   $$b = .01 + \sum(\gamma_i - \bar{\gamma}^{(s)})^2/2$$

For maximum protection, we generate each synthetic dataset from a random draw of λ_{ij}, not the posterior mean of λ_{ij}. At this time the Census Bureau plans to release only one synthetic dataset. So, only the one dataset is used for disclosure risk computations, which we repeat multiple times to get a sense of the variability in the risk measures.

3. Measuring Disclosure Risk

In this section, we describe an approach to assessing disclosure risks of the partially synthetic data. We focus on computing the probabilities that the true source-destination combination cell counts (after subtracting one) can be learned from the released data, building on ideas developed by Duncan and Lambert (1989) and applied subsequently by several authors Feinberg et al. (1997), Reiter (2005), Drechsler and Reiter (2008), and Reiter and Mitra (2009). We use functions of these probabilities to create risk metrics for attribute disclosure (an intruder learns the value of true cell count).

The Bayesian approach allows to formally combine any information an intruder may have, prior to data release, with the information contained in the data release to derive some useful disclosure risk measures. These measures attempt to capture the increased risk of disclosure incurred by the data release, in some of the worst-case scenarios of an intruder possessing a large amount of information, i.e. everything but the count y_{ij} in a particular table cell.

To learn any y_{ij}, we assume that the intruder utilizes all information at her or his disposal. This includes information released about the synthetic data model, which we denote with *M*. For example, *M* could include mathematical descriptions corresponding to the model in Section 2.1, including the names of the counties. Alternatively, *M* could include the code used to fit the model without parameter estimates (including parameter estimates could leak too much information about true cell counts). The intruder also may possess auxiliary information about the cell counts on the file, which we denote with *A*. For example, *A* could include the counts of some subset of cells in the implied contingency table, or it could be empty.

Using this information, the intruder seeks to determine the probable values of y_t for one or more cells $t = (i, j)$ in the data. Mathematically, this can be characterized as follows. Restricting to the state including cell *t*, let D represent the observed counts and $\tilde{D}$ represent the *m* synthetic datasets. For any particular *t* and potential count *y*, the intruder seeks to estimate

$$\begin{aligned} \rho_t^y &= P(y_t = y|\tilde{D}, A, M) = cP(\tilde{D}|y_t = y, A, M)P(y_t = y|A, M) \\ &= c\left(\int P(\tilde{D}|y_t = y, A, M, \boldsymbol{\lambda})P(\boldsymbol{\lambda}|y_t = y, A, M)d\boldsymbol{\lambda}\right)P(y_t = y|A, M) \end{aligned} \quad (3)$$

over all feasible *y*, where *c* is a normalizing constant. We assume that the intruder selects the *y* yielding the maximum ρ_t^y as a best guess for y_t.

Conceptually, $P(y_t = y|A, M)$ represents the intruder's prior beliefs about the count in cell *t*. It is impossible for agencies to know any particular intruder's prior beliefs. Instead, agencies can adopt the recommendation of Skinner and Shlomo (2008) and evaluate risks under reasonable prior distributions. For example, the agency can use a uniform

distribution over all possible counts between zero and eight. This reflects vague prior knowledge about y_t.

Similarly, it is impossible for the agency to know the auxiliary information possessed by intruders. One approach, which we adopt here, is to evaluate risks under a ``worst case'' scenario by assuming that the intruder knows the counts of all cells except *t*, i.e., the intruder knows $y_{t'}$ for all $t' \neq t$. Call this set Y_{-t}. In addition to offering risk estimates for intruders with very strong prior knowledge, setting $A = Y_{-t}$ greatly facilitates computation, as we describe in the next section.

3.1 Computational Methods

The form of (3) when $A = Y_{-t}$ suggests a Monte Carlo approach to estimation of ρ_t^y. First, for any proposed value *y*, the agency replaces y_t with *y* to form a new set of counts, $D_t^y = (y_t = y, Y_{-t})$, with the same source-destination pairs. Second, treating D_t^y as if it were the collected data, the agency samples many values of $\boldsymbol{\lambda}$. Third, for each sampled $\boldsymbol{\lambda}$, the agency computes the probability of generating the released $\tilde{D}$, and averages these probabilities. The agency repeats these three steps for all values of *y*, which allows computation of the normalizing constant in (3) and hence ρ_t^y for all *y*.

To draw new $\boldsymbol{\lambda}$ for each D_t^y, one approach is to re-estimate the model in (2) in Section 2.1. This could be computationally intensive, however, as the agency needs to estimate 9 models per *t*. Another option is to use the sampled values of $\boldsymbol{\lambda}$ from $p(\boldsymbol{\lambda}|D)$ as proposals for an importance sampling algorithm. As a brief review of importance sampling, suppose we seek to estimate the expectation of some function $g(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} \sim f(\boldsymbol{\lambda})$. Further suppose that we have available a sample $(\boldsymbol{\lambda}^{(1)}, \cdots, \boldsymbol{\lambda}^{(L)})$ from a convenient distribution $f^*(\boldsymbol{\lambda})$ that differs from $f(\boldsymbol{\lambda})$. We then can estimate $E_f(g(\boldsymbol{\lambda}))$ using

$$E_f(f(\boldsymbol{\lambda})) \approx (1/L)\sum_{j=1}^{L} g(\boldsymbol{\lambda}^{(j)}) \frac{f(\boldsymbol{\lambda}^{(j)})/f^*(\boldsymbol{\lambda}^{(j)})}{\sum_{l=1}^{L} f(\boldsymbol{\lambda}^{(l)})/f^*(\boldsymbol{\lambda}^{(l)})} \qquad (4)$$

We note that (4) only requires that $f^*(\boldsymbol{\lambda})$ and $f(\boldsymbol{\lambda})$ be known up to a normalizing constants.

We can implement importance sampling algorithms to approximate the integral in (3). For any proposed $y_t = y$, we set $g(\boldsymbol{\lambda}) = cP(\tilde{D}|D_t^y, M)$ and seek to approximate its expectation with respect to $f(\boldsymbol{\lambda}) = P(\boldsymbol{\lambda}|D_t^y, M)$. To facilitate computation, we work with each synthetic dataset $\tilde{D}^{(l)}$ (assuming $m>1$ datasets) separately, since

$$P(\tilde{D}|D_t^y, M) = \prod_{l=1}^{m} P(\tilde{D}^{(l)}|D_t^y, M) \qquad (5)$$

Let $\tilde{y}_k^{(l)}$ be the synthetic value for y_k in $D^{(l)}$, where $k = 1, \cdots, 9$. Given a sampled value of $\boldsymbol{\lambda}$, we have

$$P(\tilde{D}^{(l)}|D_t^y, M, \boldsymbol{\lambda}) = \prod_{k=1}^{n} \frac{\lambda_k^{\tilde{y}_k^{(l)}}}{\tilde{y}_k^{(l)}! \sum_{r=0}^{8} \frac{\lambda_k^r}{r!}} \qquad (6)$$

We next set $(\boldsymbol{\lambda}^{(1)}, \cdots, \boldsymbol{\lambda}^{(L)})$ equal to L draws of $\boldsymbol{\lambda}$ already available from the estimated posterior distribution based on D; hence, we set $f^*(\boldsymbol{\lambda}) = f(\boldsymbol{\lambda}|D, M)$. Following the models in Section 2.1, the only differences in the kernels of $f^*(\boldsymbol{\lambda})$ and $f(\boldsymbol{\lambda})$ include

i. the components of the likelihood associated with the counts y and y_t for cell t and
ii. the normalizing constants for each density.

Hence, after computing the normalized ratio in (4), we are left with the expression,

$$P(\tilde{D}^{(l)}|D_t^y, M) = (1/L) \sum_{s=1}^{L} \left(\prod_{k=1}^{n} \frac{(\lambda_k^{(s)})^{\tilde{y}_k^{(l)}}}{\tilde{y}_k^{(l)}! \sum_{r=0}^{8} \frac{(\lambda_k^{(s)})^r}{r!}} \right) \left(\frac{(\lambda_t^{(s)})^{y-y_t}}{\sum_{h=1}^{L} (\lambda_h^{(s)})^{y-y_t}} \right) \qquad (7)$$

We repeat this computation for $l = 1, \cdots, m$ times, plugging the m results into (5).

Finally, to approximate ρ_t^y, we compute (5) for each y and multiply each resulting value by $P(y_y = y \mid \mathrm{Y}_{-t}, M)$, and we normalize the collection of n results (hence, computation of c is never required). As a note on computation, the terms in (6) other than those for t cancel when normalizing, so that one can replace the expression in (6) with

$$\frac{(\lambda_t^{(s)})^{\tilde{y}_t^{(l)}}}{\tilde{y}_t^{(l)}! \sum_{r=0}^{8} \frac{(\lambda_t^{(s)})^r}{r!}} \qquad (8)$$

Setting $A = Y_{-t}$ simplifies computation immensely, in that we have to impute new values only for y_t when computing ρ_t^y. In contrast, to compute ρ_t^y when $A \neq Y_{-t}$, the intruder needs to impute possible values for all unknown counts. This introduces a potentially large number of computations. One case of particular interest is when A is empty, representing no intruder knowledge. To avoid imputing all of Y, one rough approximation is to use each of the m sets of $\tilde{D}_{-t}^{(l)}$ as a draw of Y_{-t}, and average the m resulting values of $\rho_t^{y(l)}$.

3.2 Summary Measures

After obtaining the posterior probabilities, agencies need to summarize these probabilities to evaluate individual and file level disclosure risks. We now present four such risk measures. Each is based on the assumption that the intruder uses the count *y* with maximum ρ_t^y as the best guess for y_t.

The first measure assesses the risks that intruders learn true counts given the synthetic data; hence, it is an attribute disclosure risk measure. For all cells $t = 1, \cdots, n$ in the file, let $r_t = 1$ if the maximum posterior probability for cell *t* happens to be on the true y_t (with no ties), and let $r_t = 0$ otherwise. That is, for all *t*, let

$$r_t = \mathbf{1}_{\arg\max_s\left(\rho_t^y\right)} = y_t$$

A file level risk measure is the percentage of records with $r_t = 1$, i.e.,

$$R_{all} = \sum_{t=1}^{n} \frac{r_t}{n} \qquad (9)$$

Intuitively, smaller values of R_{all} are preferable to larger values for confidentiality protection.

As noted by many experts in disclosure estimation, Shlomo and Skinner (2010), agencies pay special attention to risks for records with unique combinations of variables in the sample (although arguably uniqueness in the population is more relevant). Singletons are more likely to be identified, since matches to external data are guaranteed to be correct (assuming no errors in matching). With this issue in mind, we introduce a measure that focuses on counts with unique source-destination combinations. Formally, for all $t = 1, \cdots, n$, let $a_t = 1$ if source county *i* and destination county *j* that comprise *t* are represented only once in N_k, and let $a_t = 0$ if $a_t > 1$. The second risk measure is the percentage of counts belonging to such singletons that the intruder correctly estimates,

$$R_{unq} = \frac{\sum_{t=1}^{n} a_t r_t}{\sum_{t=1}^{n} a_t} \qquad (10)$$

Intuitively, R_{unq} will be near one if there are a lot of similar points within the same grid cell, the intruder cannot guess which one of them is the one that he is trying to identify. It can be seen as an adjusted version of r_t. If a point gets correctly identified and it is alone ($a_t = 1$), it contributes more to the average risk than if it is close to other similar points ($a_t < 1$). By using this measure, we are reducing the effect of the grid size. If we have few grid cells, it is less likely to have alone individuals than in a thin grid. That means that

when *m* decreases, we expect that the frequency of $a_t = 1$ will also decrease. Thus, the average $\bar{a}$ gives a better idea of the average risk.

Both R_{all} and R_{unq} do not distinguish between intruders whose best guess is close (but not equal) to the actual count and whose best guess is far from the actual grid cell. To distinguish these, we present a third risk measure based on distances. For each *t*, let d_t be the distance between y_t and the count with the maximum probability, so that

$$d_t = \left\| y_t - \arg\max_y \left(\rho_t^y \right) \right\| \qquad (11)$$

The agency can assess the distributions of d_t over all *t* to determine if, for example, errors tend not to be concentrated at small values. We compute d_t as the Euclidean distance between y_t and the count with maximum probability. Note that the maximum d_t could equal $8^2 = 64$, so all differences need to be interpreted relative to that.

While R_{all} , R_{unq}, and the distribution of d_t summarize risks that intruders learn true counts, they are not readily interpretable as measures of identification disclosure risk. One could repeat this exercise for individual cases. In particular, in some grid cells many records have the same attribute pattern *b*, so that intruders cannot distinguish between them. For an extreme example, consider using only one grid cell for the entire area. Here, $r_t = 1$ for all *t*, since *a priori* everyone is guaranteed to be in the cell. Thus, $R_{all} = R_{unq} = 1$ and all d_t are equal. However, since coordinates are sampled randomly within the single cell, releasing $\tilde{S}$ introduces zero risks that individual records will be identified (assuming the intruder already knows the study area).

References

Dandekar R. A. (2001). Synthetic Tabular Data: A Better Alternative To Complementary Data Suppression. CENEX-SDC Project International Conference, PSD2006, Rome, Proceedings ISBN: 84-690-2100-1.

Dandekar R. A. and Cox L. H. (2002). Synthetic Tabular Data: An Alternative to Complementary Cell Suppression. Manuscript, Energy Information Administration, U. S. Department of Energy.

Drechsler, J. and Reiter, J. P. (2008 pp. 227--238). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In Privacy in Statistical Databases (LNCS 5262), (ed). J. Domingo-Ferrer, and Y. Saygin, New York: Springer-Verlag.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. Journal of Business and Economic Statistics, 7, 207--217.

Fienberg, S. E. and Makov, U. E. and Sanil, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. Journal of Official Statistics, 13, 75—89.

Machanavajjhala, A., Kifer, D., Abowd, M. J., Gehrke, J. & Vilhuber, L. (2008). Privacy: Theory meets Practice on the Map. *IEEE 24th International Conference on Data Engineering*, **2008**, 277-286

Reiter, J. P. (2005). Estimating identification risks in microdata. Journal of the American Statistical Association, 100, 1103—1113.

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. Journal of Privacy and Confidentiality, 1, 99—110.

Shlomo, N. & Skinner, C. J. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. Annals of Applied Statistics, 4, 1291-1310

Skinner, C. J. and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. Journal of the American Statistical Association, 103, 989—1001.