Using partially synthetic data to replace suppression in the Business Dynamics Statistics: early results

Javier Miranda[1] and Lars Vilhuber[2]

[1] U.S. Bureau of the Census, Washington, DC, USA javier.miranda@census.gov
[2] Cornell University, Ithaca, NY, USA, lars.vilhuber@cornell.edu, Corresponding author

Abstract. The Business Dynamics Statistics is a product of the U.S. Census Bureau that provides measures of business openings and closings, and job creation and destruction, by a variety of cross-classifications (firm and establishment age and size, industrial sector, and geography). Sensitive data are currently protected through suppression. However, as additional tabulations are being developed, at ever more detailed geographic levels, the number of suppressions increases dramatically. This paper explores the option of providing public-use data that are analytically valid and without suppressions, by leveraging synthetic data to replace observations in sensitive cells.

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

1 Introduction

The Business Dynamics Statistics (BDS) were first released in 2008, providing novel statistics on business startups on a comprehensive basis for the U.S. economy [8]. They have been used in a number of recent publications, addressing questions of firm dynamics, who creates jobs, etc. [9].

The BDS are sourced from confidential microdata in the Longitudinal Business Database (LBD). It provides measures of business openings and closings, and job creation and destruction, by a variety of cross-classifications (firm and establishment age and size, industrial sector, and geography). Since the first release, additional cross-tabulations have been added each year: initially provided only based on firm charateristics, tabulations based on establishment characteristics were later added, as were additional geography cross-tabulations (Metropolitan Statistical Area, and Metro/Non-Metro). Sensitive data are currently protected through suppression. However, as additional tabulations are being developed, at

ever more detailed geographic levels, the number of suppressions increases dramatically.[3]

This paper explores the option of providing public-use data that are analytically valid and without suppressions, by leveraging synthetic data to replace observations in sensitive cells. The use of synthetic data in the provision of public-use tabulations has increased in the United States. [16] describe the use of synthetic data in the case of the OnTheMap data visualization, and the use of partially synthetic data in tabulations has been explored by others [1, 5, 18, 17, 2]. Few have attempted to provide synthetic data for business data - the cases we are aware of are the Synthetic LBD [15], on which we will rely heavily in this approach, and a synthetic version of the IAB Establishment Panel [3, 10, 4]. This is, to the best of our knowledge, the first attempt to integrate synthetic data into a public-use data tabulation for businesses.

We leverage the existence of a sophisticated partially synthetic data file the Synthetic LBD [19], henceforth SynLBD – in combination with the techniques first expressed in [7] and [6] to replace sensitive cells with tabulations based on synthetic data. We start by describing the extent of suppressions in the BDS, then lay out the algorithm to combine synthetic and confidential data for the purposes of tabulation. Preliminary results are discussed, and an outlook given on the next steps necessary to achieve a robust public-use tabulation.

2 Item suppression

BDS processing uses primary and secondary suppressions, derived from a P percent rule, as disclosure avoidance mechanism. All cells of a potential publication table are analyzed to make sure no identifying information about a particular business, household, or individual is released to the public. In the case of the BDS, cells where the top 2 firms account for more than P percent of the total value of the cell are flagged for suppression. The precise P value is not disclosed to minimize the possibility of reidentification by potential attackers. Secondary suppressions are identified so as to minimize the amount of information loss in a given table row or column. To this end, the search algorithm looks for candidate cells that contain the least amount of employment, and suppresses their content. Protecting these secondary cells might require a third round of supressions given the presence of column totals in the tables. Once the tables are analyzed and the necessary cells suppressed, each table row that contains

[3] The next set of expansions include plans to provide additional industry detail

a suppressions is flagged, and the modified table released to the public. Note that individual suppressed cells are not separately flagged, only the row that contains at least one suppressed cell. A necessary feature of this disclosure mechanism is that a large number of secondary suppressions are necessitated by the need to protect the cell that is the primary disclosing cell. The public-use data, of course, doesn't allow the identification of which suppressions are primary or secondary suppressions.

Table 1 describes the extent to which suppressions occur in the published establishment-level BDS, as available at `http://www.census.gov/ces/dataproducts/bds/data_estab.html` (Table 3 in the appendix also describes the similar pattern in firm-level statistics). The number of cells in each table is indicated, as are the percent of cells with suppression of some variable (`d_flag=1`), and the percent of cells where "Job Creation by Entrants" is suppressed. Other variables, also present on the establishment-level BDS, are never suppressed.

Table 1. Suppressions in establishment-level BDS

Type	Level	Number of cells	Suppressions (%)	
			Any	Job creation by entrants
Age	e	325	0.3	0.3
Age-Initial Size	e	2925	18.6	14.2
Age-Initial Size-SIC	e	25994	35.9	17.9
Age-SIC	e	2925	3	2.9
Age-State	e	18360	3.4	3.3
Age-Size	e	2925	26.8	16.2
All	e	35	0	0
Initial Size	e	315	0.3	0
Initial Size-SIC	e	2835	19.5	6.5
Initial Size-State	e	17847	26.8	11.2
SIC	e	315	0	0
State	e	1785	0	0
Size	e	315	0.3	0
Size-SIC	e	2834	28.1	11.3
Size-State	e	17848	31.9	14.6

Note: Cells are year x categories, where the number of categories varies by published table.

Clearly, while the usefulness of the data to users would seem to increase for more detailed cross-tabulations, that same detail, under current disclosure avoidance rules, leads to increased suppression, and thus less effective data utility. Suppression is worse for some variables than for oth-

ers. Establishment and firm counts are never supressed following County Business Patterns and Disclosure Review Board rules. By contrast employment, job creation and destruction are suppressed.

3 Synthetic Data as a Proposed Alternative to Item Suppression

The Synthetic LBD (SynLBD) is a synthetic dataset on establishments with proven analytic validity along several critical dimensions [15]. Additional improvements are currently being developed [13, 14]. A growing number of researchers have used the SynLBD, and their continued use contributes to the improvement of the SynLBD.

The use of the SynLBD for the purposes outlined in this paper is particularly appealing, because its analytic validity has been independently established, while maintaining a high level of data privacy. In fact, for many of the cross-tabulations identified in Table 1, no additional disclosure avoidance review would seem necessary. Only tabulations involving state and sub-state geography should require additional review since this variable was removed from the disclosure request that approved the release to the public of the SynLBD.[4]

The available SynLBD is released as a single implicate, and by design, may distort any single analysis by too large an amount. The use of additional implicates for the purposes of BDS table creation may be desirable and will be assessed in later work.

In this paper, we evaluate a simple algorithm to alleviate the problem of large numbers of suppression, while maintaining high, if not equivalent levels of disclosure protection. We then outline a second algorithm that improves on the first. An evaluation of the second algorithm is deferred to later work.

The first algorithm, which we will call the "drop-in algorithm", simply replaces a cell that has been suppressed with its synthetic-data equivalent, i.e., the equivalent table cell from a tabulation based on the SynLBD alone. The second algorithm, called "forward-longitudinal algorithm", is slightly more complicated. At any point in time t, if a (expanded) suppression algorithm identifies a cell that *would* be suppressed, all establishments that contribute to that cell in time period t are replaced by synthetic establishments that match on certain characteristics Z in periods $t-p$ through t, for t and the next n periods. Synthetic and observed

[4] The Census Disclosure Review Board has not pronounced itself on the disclosure avoidance methodology proposed here as of July 2014.

values are then tabulated to create the release statistics. If Z describes only the margin characteristics for the table in question (denoted by k below), and for $p = n = 0$, the algorithm reduces to the "drop-in" algorithm.

In this paper, we assess the time-consistency of the first algorithm for a single implicate. Assessing the impact of using multiple implicates is deferred to future work. Identifying acceptable values of Z, p, and n is deferred to a later version of this paper.

3.1 Definitions

The variable of interest is establishment employment e_{jt}, with establishments indexed by j and years indexed by t. All other variables (job creation and destruction from establishment entry, exit, expansion and contraction) are derived from that. For instance, an establishment is born at time t if employment is positive for the first time:

$$birth_{jt} = \begin{cases} 1 \text{ if } \ e_{jt} > 0 \ \text{ and } \ e_{jt-s} = 0 \ \ \forall s \geq 1 \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

We will denote aggregations using capital letters, so (national) employment is denoted as

$$E_{\cdot t} = \sum_{i=1}^{J} e_{it} \tag{2}$$

and (national) births are

$$Birth_{\cdot t} = \sum_{i=1}^{J} birth_{it}. \tag{3}$$

For any establishment j, the synthesized version of variable x_{jt} (from a single implicate) is denoted $\tilde{x}_j t$. Furthermore, an establishment j has certain time-varying characteristics $k_t(j)$, such as industry and geographic location, but also derived characteristics, such as establishment or firm age and size. In a slight abuse of notation, $j \in K'_t$ describes the set of firms at time t such that $k_t(j) = k'$. So generically,

$$X_{k't} = \sum_{j \in K'_t} x_{jt} \tag{4}$$

describes the different aggregations across establishments having characteristics k' at time t, for instance aggregations by establishment age

or metropolitan areas. Finally, suppression rules for (aggregate) variable X are captured by I_t^X, such that the releasable variable X^o under the current regime can be described by

$$X_{k't}^o = \begin{cases} X_{k't} \text{ if } I_{kt}^X = 1 \\ \text{missing otherwise} \end{cases} \tag{5}$$

For later reference, we denote the tabulations created as per (5) as $\mathbf{BDS}^o$.

3.2 Algorithm 1: Drop-in

We can now express the "drop-in" algorithm, leading to the released variable $X^{(i)}$, as:

---

if $I_t^X = 1$ **then**
  $X_{k't}^{(i)} = X_{k't}$
else
  $X_{k't}^{(i)} = \tilde{X}_{k't}$
end if

Thus, simply computing a "SynBDS", based on the SynLBD, in parallel to the computation of the BDS (based on the confidential LBD), and replacing suppressed cells with their fully synthetic counterparts, yields a dataset without missing observations. Variations can encompass using the average of multiple implicates as the replacement value. In general, increasing the number of implicates will improve the analytic validity, but reduce the protection provided by the synthesis process.

Because no time-consistency is imposed, this method can lead to seam biases or higher intertemporal variance. We will return to this issue in Section 4. For later reference, we denote the tabulations created by Algorithm 1 as $\mathbf{BDS}^{(i)}$.

3.3 Algorithm 2: Forward-longitudinal

In part to address the possible time-inconsistencies we propose an alternative algorithm. In order to minimize future seam issues, we remove establishments (or firms) that contribute to sensitive cells of tabulations with characteristics $k't$, for t and the next n periods. These establishments are replaced by synthetic establishments that match on characteristics $k't$, and we simply replace the observed values in the database x_{js} with the synthetic values $\tilde{x}_{js}$ (for all variables), for $s = t, \ldots, t+n$.[5] For conve-

[5] We thus re-use the index j for both observed and synthetic establishments.

nience, denote by $J^-_{k't}$ the set of establishments for which observed values x_{jt} do not contribute to any tabulations at time t. In its simplest form, the algorithm can be expressed as

---

Compute: $X_{k't} = \sum_{j \in K'_t} x_{jt}$
Compute: I^X_t
if $I^X_t = 0$ **then**
  Assign all $j \in K'_t$ to $J^-_{k't}$
  Assign all $j \in J^-_{k's}$ to $J^-_{k't}$ for $t > s > t - n$
end if
Compute:

$$X^{(ii)}_{k't} = \sum_{j \in \left\{K'_t \cap J^-_{k't}\right\}} \tilde{x}_{jt} + \sum_{j \in K'_t \wedge j \notin J^-_{k't}} x_{jt}$$

For $n = \infty$, J_t is an absorbing set, which seems undesirable. For $n = 1$, this reduces to Algorithm 1.[6] For reference, we denote the tabulations created by Algorithm 2 as $\mathbf{BDS}^{(ii)}$.

4 Analysis

We implemented Algorithm 1 for BDS tabulations by establishment age and size (`bds_e_agesz`). As noted in Table 1, about 26% of all cells have some suppression. For this version of the paper, we analyzed a single variable, "Job Creation by establishment births" (`job_creation_births`). (Additional analyses are pending release).

4.1 Extent of protection

Protection of the table relies in large part on the fact that the data replacing the suppressions is itself synthetic, and released (in the case of the examples in this paper) or (potentially) releasable (for tabulations with geography) to a broad audience [2]. No establishment's observed data is released in the SynLBD, and only the industry distribution of establishments is preserved exactly. However, in order to consider a broader notion of disclosure avoidance, we proceed as follows. In cell that would have been suppressed under the current regime $\mathbf{BDS}^0$, we compute the difference

[6] Alternatively to the combining rule described in Algorithm 2, we could also specify a per-establishment weight $w_{jt} \in [0, 1]$ that declines to 0 as s approaches $t-n$. w_{jt} is adjusted as a function of membership in $J^-_{k't}$, and we compute $X^{(ii)}_{k't} = \sum_j w_{jt}\tilde{x}_{jt} + (1 - w_{jt})x_{jt}$.

between the confidential values of the establishments contributing to this cell, and each of the values of the synthetic establishments contributing to the cell under $\mathbf{BDS}^{(i)}$, and assess the distribution of these differences.[7]

4.2 Analytical validity

In order to assess the analytical validity of each of the methods, we focus on simple time-series properties of the $X_{k't}$. In particular, we estimate a AR(2) process for each of $X_{k't}$, $X^s_{k't}$, and $X^{(i)}_{k't}$. We then assess the number of missing time-series estimates (repeated suppressions in $X^s_{k't}$ may lead to time-series that are too short), the number of significant coefficients for the first lag of the AR(2), estimated from both the confidential data (ρ_1) and the comparison data (ρ^s_1 and $\rho^{(i)}_1$), and finally two measures of utility: *coverage*, the percentage of regressions where the true ρ_1 lies within the confidence band around the coefficient estimated from the comparison ρ^s_1 and $\rho^{(i)}_1$, and the interval overlap measure J_k as suggested by [12]. Table 2 presents these results for `job_creation_births`.

Table 2. Analytic validity of published data

Variable	Number feasible	Missing		Percent significant			Coverage		Interval overlap	
	$X_{k't}$	$X^s_{k't}$	$X^{(i)}_{k't}$	ρ_1	ρ^s_1	$\rho^{(i)}_1$	ρ^s_1	$\rho^{(i)}_1$	J^s_1	$J^{(i)}_1$
job creation births	89	18	11.2	5.6	6.8	6.3	91.8	93.7	91.6	93.9

(Caveat: different definitions of "job creation births" in the BDS processing and our post-processing lead to incomplete filling in of missing cells. This will be fixed in later work.) For the one variable that has significant suppressions, the number of feasible regressions in the published data increases substantially (reduction in missing $X^{(i)}_{k't}$ relative to missing $X^s_{k't}$). The number of correctly estimated coefficients increases (in terms of assessing statistical significance of the coefficient), and utility increases, in terms of ρ_1 as well as J_1.

5 Concluding remarks

In this paper, we have described two alternate mechanisms to substitute for suppressions in small-cell tabulations of business microdata, with

[7] As of June 2014, this distribution had not been released.

the goal of improving analytic validity while maintaining a sufficiently high standard of disclosure limitation. Neither mechanism fundamentally changes the existing suppression methodology, rather, the mechanisms work to fill in the holes created by the suppression methodology.

Leveraging the availability of a high-quality synthetic datasets (the Synthetic LBD) with proven disclosure limitation efficiency and analytic validity [15], the first method is very simple, but may suffer from seam biases and time-inconsistency. The second method aims to improve on that by "blending in" synthetic establishments, which may slightly reduce analytic validity in time periods where the strict application of the suppression algorithms would no longer impose any constraints, but improving on the time-series properties of the released data.

Several limitations of the research presented here should be highlighted. The examples provided in this article rely on an earlier release of the Synthetic LBD [15]. Recent developments to improve the micro-level analytic validity of the SynLBD [14] should improve the analytic validity of the mechanisms proposed here as well. We also compare our proposed mechanisms to the actual published, but otherwise unmodified BDS. Comparing to post-publication improvements to a table with suppressions [11] will inevitably lead to an apparent reduction in the utility of this particular approach. Finally, the approach relies on continuous availability of synthetic microdata with analytical validity. Other approaches rely on fewer data points, and thus be favored due to lower implementation costs.

Future work for this paper involves assessing the procedure on a wider variety of variables, better synchronisation of the computational algorithms underlying the BDS and the SynBDS, and improved assessment at the microdata level of the protection afforded by Algorithm 1.

Acknowledgments. All authors were affiliated with the U.S. Census Bureau, Center for Economic Studies, when originally contributing to the contents of this paper. 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. Vilhuber acknowledges support through NSF Grant SES-1042181. This

project would not have been feasible without the valuable input from Saki Kinney and Jerry Reiter, and their valuable work on the Synthetic LBD.

References

1. Abowd, J.M., Gittings, K., McKinney, K.L., Stephens, B.E., Vilhuber, L., Woodcock, S.: Dynamically consistent noise infusion and partially synthetic data as confidentiality protection measures for related time-series. Tech. rep., Federal Committee on Statistical Methodology (January 2012), http://www.fcsm.gov/events/papers2012.html
2. Abowd, J.M., Vilhuber, L.: Synthetic data server (2010), http://www.vrdc.cornell.edu/sds/
3. Drechsler, J.: Synthetische Scientific-use-files der Welle 2007 des IAB-Betriebspanels. FDZ Methodenreport 201101_de, Institute for Employment Research, Nuremberg, Germany (Jan 2011), http://ideas.repec.org/p/iab/iabfme/201101_de.html
4. Drechsler, J.: New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey. Journal of Applied Statistics 39(2), 243–265 (April 2012), http://ideas.repec.org/a/taf/japsta/v39y2012i2p243-265.html
5. Drechsler, J., Reiter, J.P.: Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. , vol. 25, 589-603. Journal of Official Statistics 25(12), 589–603 (December 2009), http://ideas.repec.org/a/eee/csdana/v55y2011i12p3232-3243.html
6. Drechsler, J., Reiter, J.P.: Sampling with synthesis: A new approach for releasing public use census microdata. Journal of the American Statistical Association 105(492), 1347–1357 (2010), http://ideas.repec.org/a/bes/jnlasa/v105i492y2010p1347-1357.html
7. Gittings, R.K.: Essays in labor economics and synthetic data methods. Ph.d., Cornell University (2009)
8. Haltiwanger, J., Jarmin, R., Miranda, J.: Jobs created from business startups in the united states (2008), https://www.census.gov/ces/pdf/BDS_StatBrief1_Jobs_Created.pdf
9. Haltiwanger, J.C., Jarmin, R.S., Miranda, J.: Who creates jobs? small vs. large vs. young. Working Paper 16300, National Bureau of Economic Research (August 2010), http://www.nber.org/papers/w16300
10. Hethey, T., Schmieder, J.F.: Using worker flows in the analysis of establishment turnover : Evidence from German administrative data. FDZ Methodenreport 201006_en, Institute for Employment Research, Nuremberg, Germany (Aug 2010), http://ideas.repec.org/p/iab/iabfme/201006_en.html
11. Holan, S.H., Toth, D., Ferreira, M.A.R., Karr, A.F.: Bayesian multiscale multiple imputation with implications for data confidentiality. Journal of the American Statistical Association 105(490), 564–577 (2010), http://dx.doi.org/10.1198/jasa.2009.ap08629
12. KARR, A.F., KOHNEN, C.N., OGANIAN, A., REITER, J.P., SANIL, A.P.: A framework for evaluating the utility of data altered to protect confidentiality 60(3), 1–9 (2006)
13. Kinney, S.K., Reiter, J.: SynLBD: providing firm characteristics on synthetic establishment data. Presentation, World Statistics Conference (2013)
14. Kinney, S.K., Reiter, J., Miranda, J.: Improving the Synthetic Longitudinal Business Database. Working Paper 14-12, U.S. Census Bureau, Center for Economic Studies (2014)

15. Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., Abowd, J.M.: Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database. International Statistical Review 79(3), 362–384 (December 2011), http://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html
16. Machanavajjhala, A., Kifer, D., Abowd, J.M., Gehrke, J., Vilhuber, L.: Privacy: Theory meets practice on the map. International Conference on Data Engineering (ICDE) (2008)
17. Rodríguez, R.: Synthetic data disclosure control for american community survey group quarters (2007)
18. Sakshaug, J.W., Raghunathan, T.E.: Synthetic Data For Small Area Estimation In The American Community Survey. Working Papers 13-19, Center for Economic Studies, U.S. Census Bureau (Apr 2013), http://ideas.repec.org/p/cen/wpaper/13-19.html
19. U.S. Census Bureau: Synthetic LBD Beta version 2.0. [computer file], U.S. Census Bureau and Cornell University, Synthetic Data Server [distributor], Washington,DC and Ithaca, NY, USA (2011), http://www2.vrdc.cornell.edu/news/data/lbd-synthetic-data/

Appendix

Acronyms

BDS Business Dynamics Statistics
LBD Longitudinal Business Database
SynLBD Synthetic LBD, a synthetic microdata file at the establishment level

Additional tables

Table 3. Suppressions in firm-level BDS

Type	Level	No. of cells	Percent suppressed
all	f	35	0
metrononmetro	f	70	0
sic	f	315	0
age	f	325	0
agemetrononmetro	f	650	0
st	f	1785	0
agemsa	f	118950	0.3
szmsa	f	153688	1.4
agest	f	18360	1.8
agesic	f	2925	2.8
isz	f	420	9
iszmetrononmetro	f	840	9.8
sz	f	420	10.2
szmetrononmetro	f	840	11.1
iszst	f	23205	16.1
szst	f	23205	16.2
iszsic	f	3780	18.7
szsic	f	3780	19.9
ageisz	f	3874	24.2
agesz	f	3843	26.6
ageiszmetro	f	7647	29.1
ageszmetrononmetro	f	7575	30.8
ageiszsic	f	31500	41.3

Note: Cells are year x categories, where the number of categories varies by published table.

This version: $Id: synbds-noise-synthetic.tex 1541 2014-05-18 22:27:55Z vilhu001 $