IMPROVING THE SYNTHETIC LONGITUDINAL BUSINESS DATABASE

by

Satkartar K. Kinney
National Institute of Statistical Sciences

Jerome P. Reiter
Duke University

Javier Miranda
U.S. Census Bureau

CES 14-12 February, 2014

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.census.gov/ces or contact Fariha Kamal, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 2K132B, 4600 Silver Hill Road, Washington, DC 20233, CES.Papers.List@census.gov.

Abstract

In most countries, national statistical agencies do not release establishment-level business microdata, because doing so represents too large a risk to establishments' confidentiality. Agencies potentially can manage these risks by releasing synthetic microdata, i.e., individual establishment records simulated from statistical models de- signed to mimic the joint distribution of the underlying observed data. Previously, we used this approach to generate a public-use version—now available for public use—of the U. S. Census Bureau's Longitudinal Business Database (LBD), a longitudinal cen- sus of establishments dating back to 1976. While the synthetic LBD has proven to be a useful product, we now seek to improve and expand it by using new synthesis models and adding features. This article describes our efforts to create the second generation of the SynLBD, including synthesis procedures that we believe could be replicated in other contexts.

The set of papers on the Synthetic LBD (CES-WP-14-10, CES-WP-14-11, CES-WP-14-12, and CES-WP-14-13) were first presented in a session at the the World Statistical Congress 2013 in Hong Kong (Session STS062) on August 28, 2013 (for more information, see http://www2.vrdc.cornell.edu/news/?p=2150). Lars Vilhuber, Jörg Drechsler, John M. Abowd (for Saki Kinney), and Thomas A. Louis presented the papers. The organization of this session was partially funded through NSF grant SES-1042181. The creation of version 2 of the Synthetic LBD itself was funded by NSF Grant SES-0427889. The Synthetic Data Server (http://www.vrdc.cornell.edu/sds/)is funded through NSF grant SES-1042181 with support by the U.S. Census Bureau.

1 Introduction

The Longitudinal Business Database (LBD) is a census of business establishments in the United States with paid employees. Updated on an annual basis by the U. S. Census Bureau, the LBD combines administrative data from business income and payroll Internal Revenue Service filings with Census collections including the Economic Census, the Company Organization Survey, and the Annual Survey of Manufactures. As of this writing, it includes over 25 million establishments with employees that were active between 1976 and 2011. The LBD supports research on business entry and exit, gross employment flows, employment volatility, industrial organization and other topics that cannot be adequately addressed without establishment-level data.

The data in the LBD is protected by federal laws (Title 13 and Title 26) that prohibit the release of confidential microdata. Thus, researchers seeking to use the LBD must apply for access to one of the Census Bureau's research data centers (RDCs) and pass background checks. This can be costly and time-consuming for researchers, particularly if there is not an RDC in their area.

To allow researchers to access data more efficiently, the Census Bureau and Internal Revenue Service decided to create a synthetic, public use version of the LBD, known as the SynLBD (Kinney et al., 2011). The basic procedure for generating the SynLBD was to fit models for the sensitive information in the confidential data, simulate replacement values from these models, and release the simulated data for public use. This can protect confidentiality, since identification of establishments' sensitive data is difficult when the released data are not actual, collected values. Furthermore, with appropriate data generation methods, the approach enables data users to make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Of

course, the validity of inferences on the released data depend critically on the accuracy of the synthesis models. For additional discussion of the SynLBD, see Miranda and Vilhuber (2014). For information on accessing an initial version of the SynLBD, see `http://www.census.gov/ces/dataproducts/synlbd/index.html`. For further discussion of synthetic data approaches in general, see Reiter (2003, 2004), Reiter and Raghunathan (2007), Dreschler et al. (2008), and Abowd and Vilhuber (2008), Abowd et al. (2009). For a discussion of the expanding role of synthetic data at the U.S. Census Bureau, see Jarmin et al. (2014).

The SynLBD was the first set of business microdata publicly released in the United States by a federal agency. While a useful product, the Census Bureau now seeks to improve (e.g., preserve more inferences) and extend (e.g., include additional features) the SynLBD in a second version. Key Phase 2 improvements include (i) extending the series from 2001 to 2009, (ii) switching from SIC to NAICS coding, (iii) more detailed longitudinal history of multi-unit status, (iv) improving longitudinal distributions of employee size and payroll, and (v) adding firm structure to the data.

In this article, we describe the development of the SynLBD, focusing primarily on descriptions of the new models and checks of analytical validity developed in Phase 2. This work is still in progress and updated microdata has not yet been submitted for disclosure approval. The remainder of the article is organized as follows. Section 2 describes the LBD in more detail. Section 3 describes the synthesis procedures. Section 4 discusses analytical validity of the SynLBD and Section 5 describes the disclosure protection properties of the synthetic LBD. Section 6 concludes with a discussion of project status and plans.

2 The Longitudinal Business Database and SynLBD

The LBD trades depth of information typical of surveys in exchange for complete industry and geographic coverage. Data elements on each establishment include employment, payroll, industry of activity, location, as well as firm and establishment identifiers. Unique features set it apart from similar databases including: 1) the length of the time series –currently covering multiple business cycles; 2) the ability to track establishments and firms over time; 3) its wide coverage including practically all the private non-farm activity in the U.S. and 4) the ability to link to hundreds of other databases from the U.S. Census Bureau including economic census and surveys, as well as external databases such as the COMPUSTAT, VentureExpert, or patent data from the USPTO to name a few.

The Census Bureau defines an establishment as the physical location where business is conducted, and the firm as the legal unit with operational ownership and control. In the case of the simplest firms, single-unit firms, the establishment and the firm are one and the same. Other firms have many establishments under common ownership and control, and can span many different locations and industries. Establishment IDs function as unit identifiers and thus are present in the SynLBD, though replaced with psuedo-identifiers for confidentiality purposes. Firm IDs were not synthesized in the SynLBD; including them is a major thrust of Phase 2. A key point here is that each establishment ID is designed to be a longitudinally consistent identifier that remains unchanged for the duration of the life of an establishment, regardless of any changes in ownership and firm reorganizations. The same is not true of firm IDs, since they respond to the prevalent legal structure associated with a particular group of establishments.

The longitudinal nature of the LBD enables research into firm formation and growth, entrepreneurship, the nature of innovation, company organization and business cycle dynamics

amongst other research that is tied to the behavior of firms as decision making units.This is a feature of the data that we are particularly interested in replicating.

3 Synthesis Methodology

The SynLBD is generated using the variables described in Table 1. Since the LBD is a universe file, there are no sampling weights. We note that several data cleaning steps were applied prior to the modeling and synthesis.

In Table 1, variables denoted with y_i are synthesized, i.e., their values are replaced with imputed values, and variables denoted with x_i are not synthesized. Synthetic data are generated separately for each industry group by drawing from successive conditional distributions to obtain the joint posterior predictive distribution of all y_i. In particular, we draw first year from $f(y_1|x_1, x_2)$, last year from $f(y_2|y_1, x_1, x_2)$, etc. The synthesis process is done separately for each group using models fit to the data for that group.

We describe each of these conditional models in the sections below.

3.1 Categorical variable synthesis

The strategy for synthesizing the variables y_1 and y_2, the first year and last year each establishment has nonzero payroll, is the same as the one used in Phase 1. Within each county and NAICS group, we generate possible values of Firstyear using a multinomial distribution with an informative Dirichlet prior distribution. An informative prior distribution is used for the purposes of disclosure protection. We generate Lastyear using the Dirichlet-multinomial approach with a flat prior, conditional on synthetic values of Firstyear but not on county. See Kinney et al. (2011) for details.

Table 1: Synthetic LBD Variable Descriptions. Here, we use superscript t to indicate that the variable changes annually.

Name		Description	Action	SynLBD
ID		Unique Random Number for Establishment	created	yes
County	x_1	Geographic Location	not released	no
SIC	x_2	Industry Code	unmodified	3-digits
NAICS	x_2	Industry Code	unmodified	no
Firstyear	y_1	First Year Establishment is Observed	synthesized	yes
Lastyear	y_2	Last Year Establishment is Observed	synthesized	yes
Year		Year dating of annual variables	created	yes
Multiunit	$y_3^{(t)}$	Multiunit Status (annual)	synthesized	categorical
Employment	$y_4^{(t)}$	Number Employees on March 12 (annual)	synthesized	yes
Payroll	$y_5^{(t)}$	Total Payroll (annual)	synthesized	yes
Firm ID	$y_6^{(t)}$	Unique firm identifier (annual)	synthesized	no

Multiunit status is a longitudinal binary variable indicating whether an establishment is part of a multiunit firm in a given year. In Phase 2, we generate the longitudinal history by first generating a binary indicator for the multiunit status of each establishment in its first year, and then for each year we generate Bernoulli indicators for whether or not each active establishment's multiunit status changes. We put these together to recreate longitudinal binary indicators of multiunit status. The synthesis method for all of the binary indicators was a reduction of the Dirichlet-multinomial method to a Beta-Bernoulli model with a flat prior.

This represents a significant departure from the synthesis of multi-unit status in Phase 1. There, multiunit status was converted from a longitudinal binary variable into a five-category nominal variable summarizing the lifetime multiunit status of establishments.

3.2 Employment and Payroll

Employment and Payroll are generated year by year. To generate the first year employment for all establishments within a NAICS group entering in year t, we adapt the CART synthesizer of Reiter (2005). Specifically, we fit a regression tree for employment conditional on the observed $(y_1, y_2, y_3^{(t)})$ using only data for establishment births in year t. We run each record's synthetic values of $(y_1, y_2, y_3^{(t)})$ down the tree to find the predicted leaf. We sample randomly from a smoothed estimate of the density of the values in the leaf, adding positive probability of picking a value that exceeds the maximum value in the leaf. Similarly, for payroll in year t, we estimate a regression tree of payroll on the observed $(y_1, y_2, y_3^{(t)}, y_t^{(t)})$ using only establishment births at time t. We run each record's synthetic values of $(y_1, y_2, y_3^{(t)}, y_4^{(t)})$ down the tree, and sample using a smoothed density estimate.

For establishments that were active at time $t-1$ and time t (so called "continuers"), we synthesize how much their employment and payroll change over the year. To do so, we use CART synthesizers based on regression trees for predicting annual changes (sequentially for employment and payroll). We draw values of changes from the models, and let the continuers' synthetic $y_4^{(t)} = y_4^{(t-1)} +$ sampled change. Here, we propose not to do any smoothing, so that we sample actual changes existing in the data. We use a similar process to get the continuers' synthetic $y_5^{(t)}$.

This is a substantial departure from the synthesis strategy in SynLBD. Previously, we generated employment levels, not changes, using normal linear regression models with kernel density estimator transformations applied to the outcomes (Abowd and Woodcock, 2004). This included automated model selection procedures and *ad hoc* rules for segmenting records. The use of CART models eliminates the need for these variable selection and subsetting procedures, making the code more robust.

One aspect of the SynLBD Employment and Payroll synthesis procedure that remains unchanged is the synthesis of zeroes for first-year employment. As Employment is defined as the number of employees of March 12 in any given year, a large fraction of establishments in their first year have zero employment, simply due to commencing business after March 12. We impute zero or non-zero first year employment with a logistic regression. Establishments in their second year with zero employees in the first year are treated as if they were in their first year. Thus, second-year employment does not depend on first-year employment when first-year employment is zero.

3.3 Firm links

Synthesis of firm characteristics is complex because we seek to reproduce the distribution of firm characteristics by aggregating establishments' synthetic values. Additionally, establishments can join, change, or leave firms over time; firms can enter and exit over time; and, firms interact with other firms, e.g., by merging, acquiring, divesting, etc.

Our proposed strategy, still under refinement, uses multivariate regression trees and propensity score matching to synthesize Firm IDs. For any synthetic establishment that was not part of a firm at time $t-1$ and is so at time t (based on the synthesized multi-unit status), we synthesize the ID of the firm it joins. For establishments in firms at time $t-1$ and time t, we synthesize whether or not the establishment changes firms. If it does change firms, we synthesize its new Firm ID; otherwise we keep the Firm ID.

We propose a novel procedure to synthesize Firm IDs. We first form multivariate firm-level characeristics (e.g., number of establishments, total employment, firm age) based on the actual firms. We then predict these firm level characteristics from establishment-level characteristics using a multivariate regression tree. We run each record's synthetic values of

$(y_1, y_2, y_4^{(t)}, y_5^{(t)})$ down the tree, coming up with a set of potential predicted firm characteristics for each establishment. We then match these predicted firm characteristics to actual firms, and assign the synthetic Firm ID to be that of the matched real firm. We match without replacement to replicate marginal distributions of the Firm IDs somewhat closely; otherwise we would tend to under-represent the number of firms with small numbers of establishments. At this point, the values of synthetic Firm ID are actual Firm IDs, assigned to different units. Once the full synthesis is complete, the synthetic Firm IDs will be replaced with pseudo-identifiers, so that no actual Firm IDs are contained in the synthetic data. For purposes of logical consistency as well as disclosure protection, we delete the synthetic firm characteristics from synthetic data file. These characteristics can be recomputed using the synthetic Firm ID and synthetic establishment characteristics.

4 Analytical properties of the SynLBD

In this section we explore the analytical validity of a sample subgroup of the Phase 2 SynLBD for reproducing key analyses obtained from the same sample in the LBD. For this portion of the analysis we restrict our attention to a single NAICS industry. This evaluation is limited by necessity at this point—for example, we do not present results involving firms due to insufficient time for review of disclosure risks— and focuses on key discrepancies identified in Phase 1 (see Kinney et al., 2011). Generally, it is expected that high-level analyses involving a large number of establishments will be well preserved, whereas analyses involving small groups or high-dimensional inferences will require access to the confidential data.

4.1 Establishments Characteristics

The sample file generally provides inferences on aggregate means and correlations that are similar to what would be obtained from the LBD. For example, Figure 1 shows gross employment levels for each year are very close to those in the LBD. The average discreprancy over the 1977-2009 period is approximately 5% with an opening gap starting in the late 1990s. Other marginal and conditional distributions that are modeled (as described in Section 3), such as first year and last year, are well preserved, as are the distributions of numbers of establishments per year and their lifetime (defined as Lastyear - Firstyear).

4.2 Lifetime: Firstyear and Lastyear

We examine the validity of the synthesized lifetime of establishments by matching establishment entry and exit rates in the SynLBD against observed ones in the LBD for every year between 1977 and 2009. We weight these statistics by the size of the establishment to obtain job creation and destruction rates from entry and exit. We find job creation and destruction rates from establishment entry and exit match up well and follow the same trends and inflection points. Figures 2 and 3 show average job creation/destruction respectively from establishment entry/exit in the SynLBD during this period is 47.5% and 48.7% respectively. Those numbers constrast with 44.7% and 42.5% in the LBD.

4.3 Dynamics of job flows

One of the most important applications of the LBD is to generate statistics that describe the amount of job creation and destruction taking place in the economy. Job creation and destruction statistics are particularly good candidates for our testing purposes because 1)

they make intensive use of the data by requiring computation of flow measures for each and every establishment in the file, and 2) they require reproduction of cross-sectional as well as longitudinal features of the data.

Job creation and destruction are defined as in Davis et al. (1996):

$$JC_t = \sum_e \left(\frac{Z_e^{(t)}}{Z^{(t)}} \right) \left| \max \left\{ 0, g_e^{(t)} \right\} \right| = \sum_e \left| \max \left\{ 0, y_{4e}^{(t)} - y_{4e}^{(t-1)} \right\} \right| / Z^{(t)}$$

$$JD_t = \sum_e \left(\frac{Z_e(t)}{Z^{(t)}} \right) \left| \min \left\{ 0, g_e^{(t)} \right\} \right| = \sum_e \left| \min \left\{ 0, y_{4e}^{(t)} - y_{4e}^{(t-1)} \right\} \right| / Z_t$$

$$NET_t = JC_t - JD_t$$

where JC_t is the sum of all employment gains from expanding establishments from year $t-1$ to year t including establishment startups, $Z_e(t) = 0.5 * (y_{4e}^{(t)} + y_{4e}^{(t-1)})$ is a measure of size of employer e, y_{4e} denotes the number of employees of employer e and $g_e^{(t)} = \left(y_{4e}^{(t)} - y_{4e}^{(t-1)} \right) / Z_e^{(t)}$, is the growth rate from $t-1$ to t of employer e. Job creation is expressed as a rate by dividing through by total employment defined as the average of the total jobs in years $t-1$ and t, $Z^{(t)} = \sum Z_e^{(t)}$. Similarly, JD_t is the sum of all employment losses in year t including the sum of employment over (a) all establishments that are last observed in year $t-1$ and (b) employment losses for establishments that contracted between year $t-1$ and year t. Job destruction is expressed as a rate by dividing by total employment defined as the average of the total jobs in years $t-1$ and t. Net job creation is the job creation rate minus the job destruction rate.

These statistics were particularly difficult to reproduce, due to complexities in modeling continuing businesses that are highly dynamic (experiencing changes in employment on an

annual bases) simultaneously with those that do not experience substantial employment changes from one year to the next. The statistics in SynLBD were consistently 10 percentage points higher relative to those in the LBD in Kinney et al. (2011). In Phase 2 we have altered the synthesis method, as described in Section 3, to better preserve job creation and destruction in the SynLBD.

Figure 4 shows the job creation rates from the SynLBD and compares them against the job creation rates from the LBD. Figure 4 shows that the SynLBD reproduces the year-to-year movements in the job creation rate rather well. The job creation rate from the SynLBD stands at an average of 21.9% over the 1977 to 2009 period. By contrast job creation in the LBD stands at 22.1%. Patterns are similar for the job destruction margin as illustrated in Figure 5. These represent substantial improvement over the results in Kinney et al. (2011).

5 Disclosure limiting properties of the SynLBD

For the Phase 2 SynLBD, we expect to perform similar evaluations of the disclosure risks as in Phase 1, which was approved for release by the Census Disclosure Review Board and Internal Revenue Service. To summarize, for Firstyear and Lastyear, Kinney et al. (2011) showed that establishments' synthetic first and last years typically did not match the observed values of the corresponding units in the LBD, and that typically there was ample probability of generating a wide range of synthetic values. Given that the same synthesis procedure is used in Phase 2 for these variables, we anticipate similar findings. For Employment and Payroll, Kinney et al. (2011) found that Pearson correlations between the observed and synthetic values within each year and industry were almost uniformly small, indicating a large amount of unit-level variability between the the observed and synthetic data. We

expect similar results in Phase 2. Kinney et al. (2011) also found that outlying values, which are considered to be at high risk of disclosure, were generally far from their corresponding true values with no systematic way of knowing which outliers were close to their observed values. Here we may find additional risk, since we appear to do a better job of preserving the marginal distribution of annual employment/payroll than in Phase 1. Once the full LBD is synthesized, we will investigate this issue carefully.

6 Concluding Remarks

The SynLBD has already shown itself to be a valuable research tool. However, as with all recurring data products, we believe it can be upgraded to be more accurate and include more features. By the time this article goes to print, we expect to have scaled up the synthesis to the whole LBD, including firm characteristics. Of course, any proposed release must satisfy the disclosure review boards of both the Census Bureau and the Internal Revenue Service.

As is the nature of a longitudinal database, the LBD continues to grow. Ultimately the Census Bureau would like to have a mechanism by which the SynLBD can be updated regularly as the LBD is updated. The nature of the synthesis, at least in its current form, is that adding years of data will require repeating the entire synthesis process since the Firstyear and Lastyear variables are synthesized. Otherwise, all new synthetic establishments would enter in their actual year, which would pose unacceptable disclosure risks. We do not see much alternative to re-synthesizing the LBD; however, there may be unanticipated disclosure risks of repeated synthesis. Characterizing such risks represents an interesting and important area of research.

Acknowledgments

The research in this paper was conducted by U.S. Census Bureau employees and Special Sworn Status researchers at the Triangle Census Research Data Center. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Phase 1 was supported by NSF grant ITR-0427889. We are greateful to the National Science Foundation for their financial support. Phase 2 is sponsored by the U. S. Census Bureau and uses methodology developed as part of the NSF NCRN award to Duke University and NISS (SES 11-31897). We also wish to thank John Abowd, Ron Jarmin, Arnold Reznek, John Haltiwanger, Nick Greenia, Karen Masken, Kevin McKinney, Lars Vilhuber, and Laura Zayatz.

References

Abowd, J. and Vilhuber, L. (2008), "How Protective are Synthetic Data?" in *Privacy in Statistical Databases 2008*, eds. Domingo-Ferrar, J. and Saygin, Y., Springer, no. 5262 in Lecture Notes in Computer Science (LNCS), pp. 239–246.

Abowd, J. M., Gehrke, J., and Vilhuber, L. (2009), "Parameter Exploration for Synthetic Data with Privacy Guarantees for *OnTheMap*," in *Joint UNECE/Eurostat work session on statistical data confidentiality.*

Abowd, J. M. and Woodcock, S. D. (2004), "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data," in *Privacy in Statistical Databases*, eds. Domingo-Ferrer, J. and Torra, V., New York: Springer-Verlag, pp. 290–297.

Davis, S., Haltiwanger, J., and Schuh, S. (1996), *Job Creation and Destruction*, MIT Press.

Dreschler, J., Bender, S., and Rassler, S. (2008), "Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB establishment panel," *Transactions on Data Privacy*, 1, 105–130.

Jarmin, R., Louis, T. A., and Miranda, J. (2014), "Expanding the Role of Synthetic Data at the U.S. Census Bureau," Working Paper 14-10, U.S. Census Bureau, Center for Economic Studies.

Kinney, S., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011), "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database," *International Statistical Review*, 79, 362–384.

Miranda, J. and Vilhuber, L. (2014), "Looking back on three years of using the Synthetic LBD," Working Paper 14-11, U.S. Census Bureau, Center for Economic Studies.

Reiter, J. P. (2003), "Inference for Partially Synthetic, Public Use Microdata Sets," *Survey Methodology*, 29, 181–189.

— (2004), "New approaches to data dissemintation: A glimpse into the future (?)," *Chance*, 17, 12–16.

— (2005), "Using CART to Generate Partially Synthetic, Public Use Microdata," *Journal of Official Statistics*, 21, 441–462.

Reiter, J. P. and Raghunathan, T. E. (2007), "The multiple adaptations of multiple imputation," *Journal of the American Statistical Association*, 102, 1462–1471.

Figures

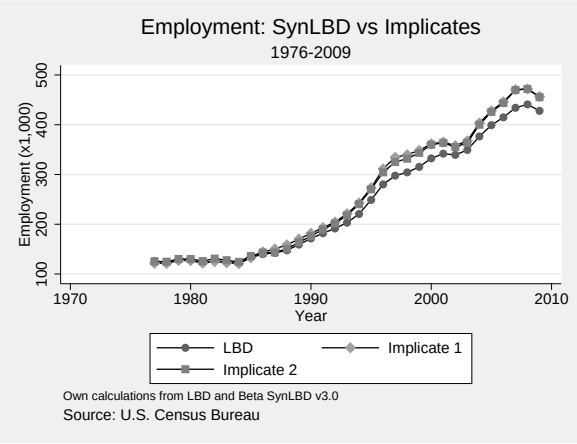

Figure 1: Gross Employment Level by Year

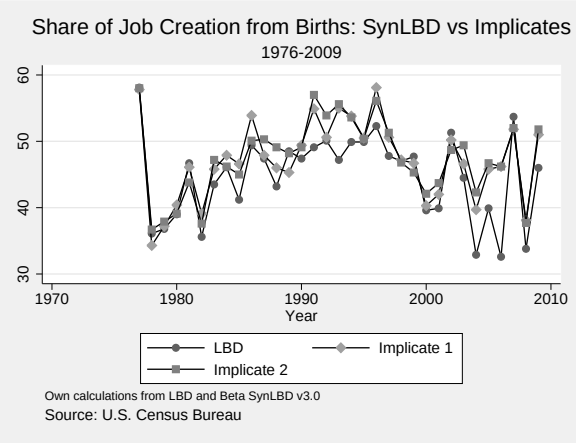

Figure 2: Job Creation Rate from Births

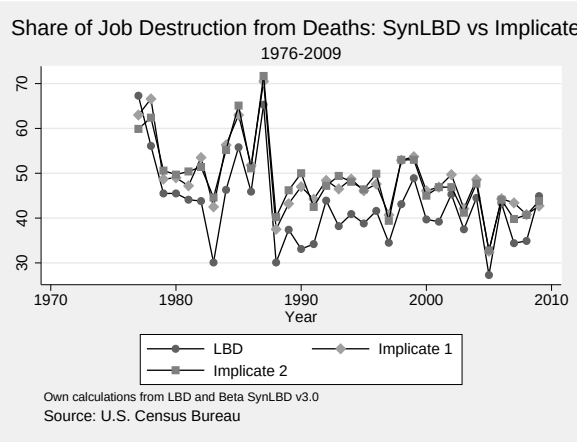

Figure 3: Job Destruction from Deaths

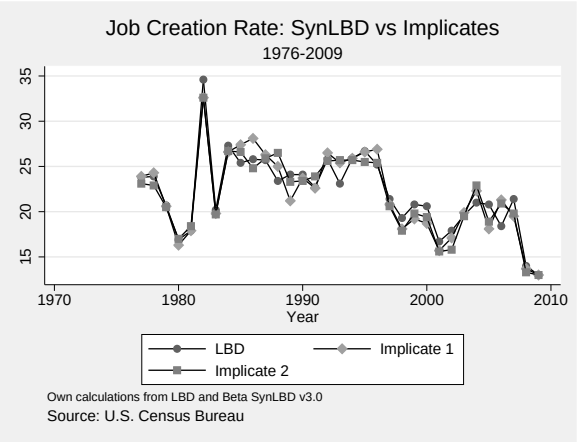

Figure 4: Job Creation Rate

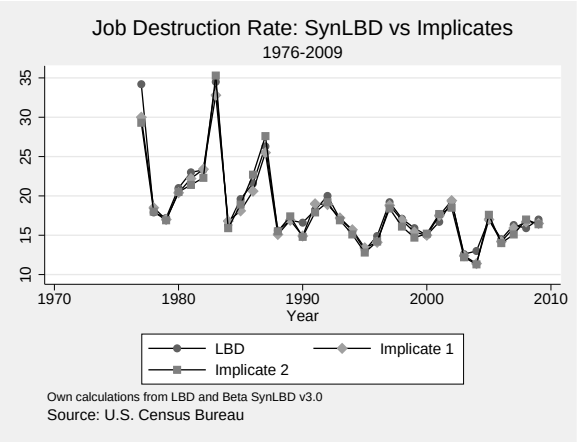

Figure 5: Job Destruction Rate