ADEP WORKING PAPER SERIES

Final Report: Economic Census Synthetic Data Project Research Team

Katherine Jenny Thompson
Economic Statistical Methods Division
Hang Kim
University of Cincinnati
Noah Bassel
Economic Statistical Methods Division
Kevin Bembridge
Economic Statistical Methods Division
Charles Coleman
Economic Statistical Methods Division
Michael Freiman
Center for Enterprise Dissemination
Maria Garcia
Center for Survey Methods
Stephen Kaputa
Economic Statistical Methods Division
Steven Riesz
Economic Statistical Methods Division
Phyllis Singer
Center for Enterprise Dissemination
Eric Valentine
Economic Statistical Methods Division
Thomas Kirk White
Center for Economic Studies
Daniel Whitehead
Economic Statistical Methods Division

Working Paper ADEP-WP-2020-05

October 2020

Associate Directorate for Economic Programs
U.S. Census Bureau
Washington DC 20233

Disclaimer: Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau. The Census Bureau has reviewed this paper for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied. (Approval ID: CBDRB-FY20-ESMD002-039)

Final Report: Economic Census Synthetic Data Project Research Team

Katherine J. Thompson, ESMD, Katherine.J.Thompson@census.gov
Hang Joon Kim, University of Cincinnati, kim3h4@ucmail.uc.edu
Noah Bassel, ESMD, Noah.Bassel@census.gov
Kevin Bembridge, ESMD, Kevin.Bembridge@census.gov
Charles Coleman, ESMD, Charles.D.Coleman@census.gov
Michael Freiman, CED, Michael.Freiman@census.gov
Maria Garcia (retired), CSM
Stephen Kaputa, ESMD, Stephen.Kaputa@census.gov
Steven Riesz, ESMD, Steven.Riesz@census.gov
Phyllis Singer, CED, Phyllis.Singer@census.gov
Eric Valentine, ESMD, Eric.Valentine@census.gov
Thomas Kirk White, CES, Thomas.Kirk.White@census.gov
Daniel Whitehead, ESMD, Daniel.Whitehead@census.gov

Working Paper ADEP-WP-2020-05
October 2020

Abstract

In May 2017, the Associate Director of Economic Programs (ADEP) and the Associate Director of Research and Methodology (ADRM) established a cross-directorate team to investigate the feasibility of developing synthetic establishment-level micro-data with sufficiently high utility and privacy protection features for public dissemination from a subset of Economic Census industries defined by six-digit 2012 North American Industry Classification System (NAICS) codes. The investigation presented in this report is more comprehensive, covering 42 industries in eighteen economic sectors covered by the Economic Census. These industries are not a random sample. This research project was designed as a "proof of concept," with understanding from upper management that post-research activities such as implementation of the recommended procedures in a production setting and development of a validation server were out of scope. This report presents the results of this research.

Keywords: Synthetic data, Economic Census, DP normal mixture models
JEL Classification Codes: Y9

The authors of this report thank the following people for their contribution to this research project: Aref Dajani (CED), Nelson Chung (CED), Yarissa Gonzalez (DSMD), Ryan Cumings (CED), and Stephen Clark (CED). The team is grateful for to Carol Caldwell (ESMD), John Abowd (ADRM), and Robert Sienkiewicz (ADRM) for supporting of research. Finally, the team thanks Carol Caldwell, William Davie, Jr., D. Mark Bauder, and Gary Benedetto for their careful review of earlier versions of this manuscript.

1. Introduction

In May 2017, the Associate Director of Economic Programs (ADEP) and the Associate Director of Research and Methodology (ADRM) established a cross-directorate team to investigate the feasibility of developing synthetic establishment-level micro-data with sufficiently high utility and privacy protection features for public dissemination from a subset of Economic Census industries defined by six-digit 2012 North American Industry Classification System (NAICS) codes. Specifically, the team was charged with investigating the synthetic data generator introduced in Kim, Reiter, and Karr (2016) and Kim, Karr, and Reiter (2015). This synthetic data generator fits Dirichlet process (DP) Gaussian mixture models to the irregular and skewed distributions of the provided business data, then selects repeated draws of synthetic data that satisfy the complete set of provided edits. The proposed generator builds on the editing and imputation methodology for economic microdata presented in Kim et al. (2015), which was tested on empirical data from two industries in one economic sector. The investigation presented in this report is more comprehensive, covering 42 industries in eighteen economic sectors covered by the Economic Census. These industries are not a random sample, as further discussed in Section 4.4.

This research project was designed as a "proof of concept," with understanding from upper management that post-research activities such as implementation of the recommended procedures in a production setting and development of a validation server were out of scope. In scope activities for the team included:

- Determination of study industries and associated variables, in collaboration with program managers, classification experts, and economists.
- Obtaining all necessary edit parameters for the applicable data items (primarily ratio edits and balance edits), as well as development of (research and validation) micro-data from the 2012 Economic Census in the selected industries.
- Review and modification of proposed synthetic data generators
- Development, testing, and implementation of software in a research environment
- Evaluation activities (utility and privacy protection)
- Issuance of research report

When established, the team consisted of ten ADEP representatives, all from the Economic Statistical Methods Division (ESMD) and six ADRM representatives (four from the Center for Enterprise Dissemination (CED), one from the Center for Economic Studies (CES), and one from the Center for Statistical Research Methods (CSRM)). Over time, ESMD participation dropped down to eight members with one new member rotating in, and CED participation ultimately dropping to two members. Throughout the project, Dr. Hang Kim of the University of Cincinnati consulted and collaborated with the team, first as an unpaid consultant (SUMMER AT CENSUS in June 2017, personal visit in November 2017), then as a part-time Census employee after being awarded an ASA/NSF/Census Fellowship from March 2018 through January 2020. As a Census employee, Dr. Kim had limited access offsite to Census hardware and data, but conducted six long-term onsite visits. During that time, he conducted extensive training and participated in several directed research projects.

This report describes the scope of the project along with the considered methodology. Section 2 gives a brief overview of the Economic Census. Section 3 discusses the considered synthetic data generators, focusing on the lessons learned in the applications in terms of input data restrictions and edit rules.

Section 4 describes the utility metrics and provides results. Section 5 discusses the privacy protection metrics and provides results. Section 6 concludes the report with recommendations.

2. Background on the 2012 Economic Census

2.1. Design

Every five years, the Census Bureau conducts an Economic Census, providing official benchmark measures of American business and the economy. Statistics from the Economic Census are used by policymakers and trade and business associations, as well as individual business owners. The totals are inputs to key measures of the U.S economy such as the Gross Domestic Product (GDP), the National Income and Product Accounts (NIPAs), and the Producer Price Index (PPI). As with the population census, ongoing sample surveys use the Economic Census data in their sampling frames. The microdata are used extensively by researchers at the Census Bureau and in the Federal Statistical Research Data Centers.

The term "Economic Census" is a bit of a misnomer. The Economic Census is both a census and a probability sample. The majority of sectors select a probability sample of the smallest single-unit establishments; larger single-unit establishments and all of the multi-unit establishments[1] are sampled with probability one i.e. included with certainty. The key statistics produced for the Economic Census include Total Number of Establishments; Primary Business Activity; Total Number of Employees in the 1st Quarter (Emp1Q) ; Value of Sales, Shipments, Receipts, Revenue (Sales); Total Annual Payroll (AnnPay); Total First Quarter Payroll (Pay1Q); and sector or industry-specific data items. Primary business activity is verified during the collection process. The industry-specific data items are collected from sampled units located in the appropriate industries, hence the probability sample. However, with the exception of the construction sector, all sectors construct a *complete universe* of the other four statistics values by using administrative data in place of respondent data for unsampled single-unit establishments, literally creating a census of establishments. We use the term "general statistics" to denote items produced within a sector for using data from all eligible establishments, whether directly collected, obtained (by design) from administrative data, or imputed using an industry model.

To be eligible for the Economic Census, a business must be in operation during the reference year. In most industries, the majority of businesses are ongoing institutions that are in operation for the entire year (i.e. are full year reporters). However, data are also collected from "part-year reporters," specifically births (businesses established in the reference year) and the deaths (businesses that ended in the reference year), whose contribution to industry totals can be substantive for selected variables. Table 1 distinguishes among the four types of units:

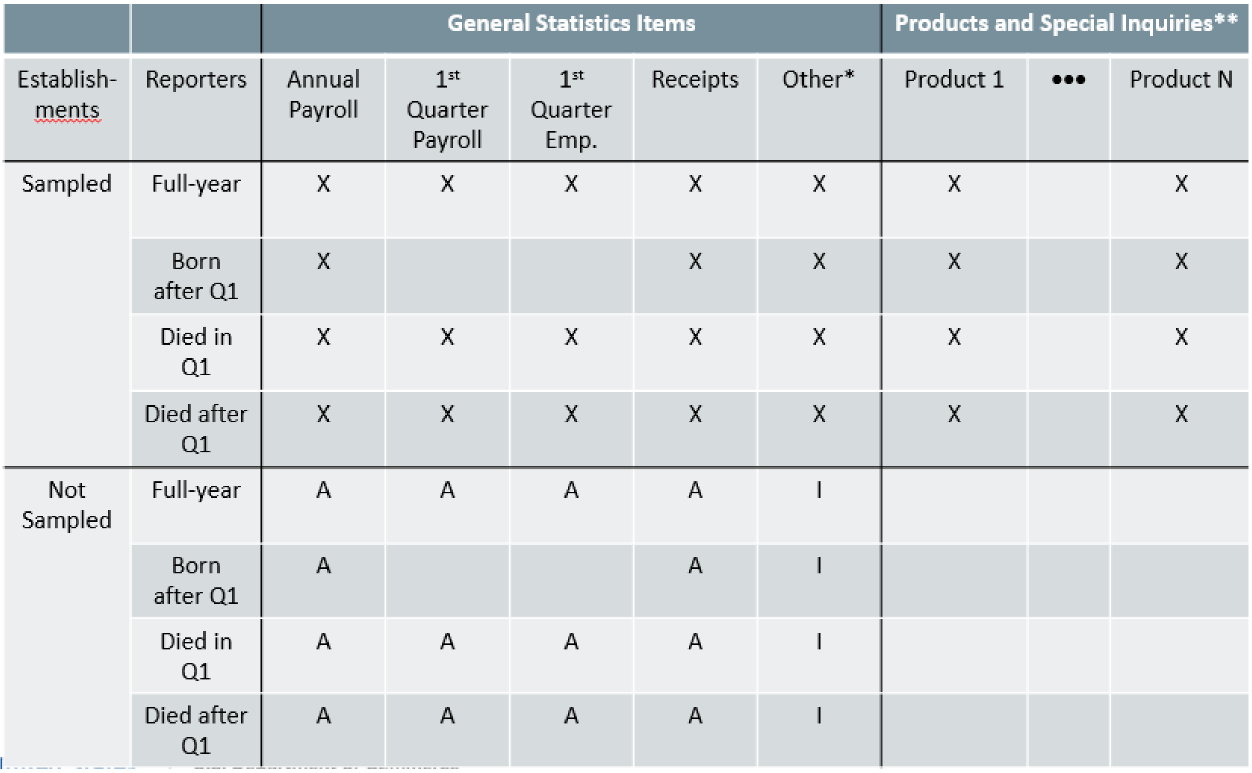
Table 1: Full and Part-Year Reporter Definitions. MIB indicates the number of active Months-in-Business.

Category	Description	Minimum MIB
Full Year Reporter	Active business at the end of the reference year	10+
Birth After 1st Quarter	Active business at the end of the reference year, start-up date after March 12	1
Death in 1st Quarter	Closed business in first quarter of the reference year	3
Death After 1st Quarter	Closed business at the end of the reference year	4

[1] A single-unit (SU) establishment owns or operates a business at a single location, whereas multi-unit (MU) establishments comprise two or more establishments that are owned or operated by the same company.

In many industries, there are seasonal businesses that do not operate for the full calendar year. Even within the same industry, the multivariate data patterns for seasonal businesses are difficult to generalize. We ignore seasonal businesses in this paper, assuming that seasonality is often inherent in the industry classification and therefore, the processing procedures (and associated parameters) implicitly account for seasonal businesses.

Figure 1 depicts the Economic Census data collection for sampled and unsampled units by type of reporting unit. Attempted collection (reported data) is depicted by an 'X'; automatic administrative data substitution is depicted by an 'A'; and automatic imputation is depicted by an 'I.' Notice that neither 1st Quarter Payroll nor Total Number of Employees in the 1st Quarter are collected or imputed for births. Businesses that died in the 1st Quarter have the same values of 1st Quarter Payroll and Annual Payroll.



		General Statistics Items					Products and Special Inquiries**		
Establish-ments	Reporters	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Receipts	Other*	Product 1	•••	Product N
Sampled	Full-year	X	X	X	X	X	X		X
	Born after Q1	X			X	X	X		X
	Died in Q1	X	X	X	X	X	X		X
	Died after Q1	X	X	X	X	X	X		X
Not Sampled	Full-year	A	A	A	A	I			
	Born after Q1	A			A	I			
	Died in Q1	A	A	A	A	I			
	Died after Q1	A	A	A	A	I			

Figure 1: Economic Census Data Collection

* Other items include – but are not exclusively limited to – beginning and ending inventories (wholesale trade, manufacturing, and mining sectors), operating expenses (tax-exempt services industries), purchase costs (wholesale trade and manufacturing sectors), plant hours worked by production workers (manufacturing). See Appendix 1.

** Products and industry-specific data items are collected only from sampled units.

After extensive discussion, the research team decided to limit the scope of research to general statistics items for full year reporter establishments in 54 industries recommended by our subject matter experts. Kim, Dreschler, and Thompson (forthcoming in 2020) presents a separate research report that investigates and proposes a synthetic data generator for the items collected only from *sampled* units (products and special inquiries). Ultimately, the team dropped twelve industries from the research for the following reasons:

- Auxiliary reporting units for total sales. In these cases, the parent company reports total sales for the company, and the edited and imputed company by industry values are allocated to the company's establishments. At the establishment level, all reported values for sales are missing (8 industries)
- Missing classification variable to distinguish warehouses from storage facilities (2 industries)
- No records whose reported data satisfy all of the production edits (1 industry)
- No records provided to 2012 Economic Census, perhaps because an incorrect 2012 NAICS code was given to the research team (1 industry)

2.2. Editing and Imputation

The Economic Census programs have procedures for detecting errors and inconsistencies in the reported data and various methods for replacing erroneous values with plausible data values. The microdata are subjected to range, ratio, and balance edits as part of the overall data review process. Figure 2 presents a greatly simplified depiction of the Economic Census microdata processing flow. Only Plain Vanilla (PV) data preparation activities are presented; there are over 39 separate post-PV edit and imputation procedures.

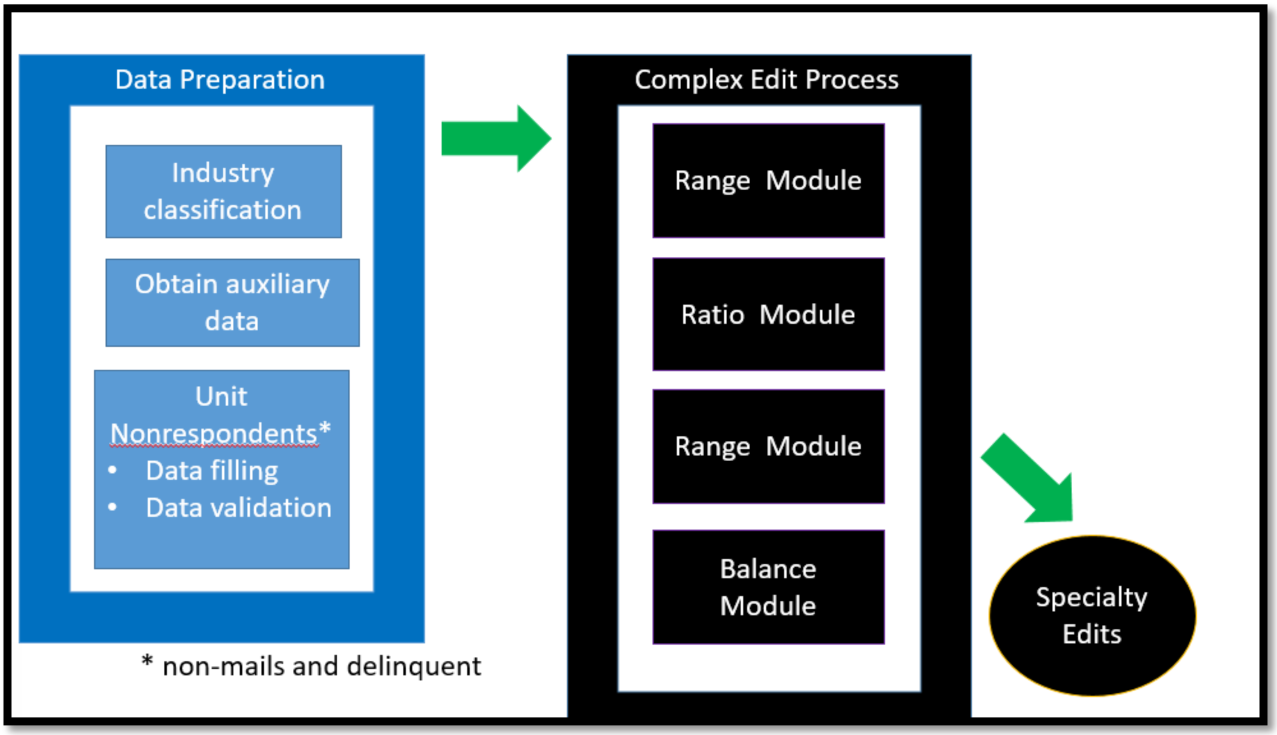


Figure 2: Simplified Depiction of Economic Census Microdata Edit and Imputation Processing Flow

The Census Bureau's Plain Vanilla (PV) suite of generalized edit and imputation modules is the primary vehicle for ensuring consistent microdata among general statistics items and selected miscellaneous industry-specific items. These modules are employed after an extensive set of data preparation activities that include industry classification of each establishment; assembling establishment-specific auxiliary microdata, primarily administrative tax data and historic economic census data plus prior year data for manufacturing establishments in the Annual Survey of Manufactures; and data filling for unsampled small businesses with planned administrative data imputation and unit nonrespondents. Many sectors perform extensive pre-PV validation and imputation activities on Annual Payroll, which tends to have consistent values reported to both the economic census and administrative data sources (e.g. the Internal Revenue Service). In the manufacturing and mining sectors, the Annual Payroll item is generally fixed (“goldplated”) in subsequent editing/imputation procedures after the pre-edit. The practice of goldplating a single extensively-validated item prior to subsequent editing and imputation procedures is referred to in-house as “anchoring.”

Figure 2 includes three of the four separate edit and imputation programs in the PV suite: a ratio edit module, a balance edit module, a verification module, and a range edit module. The ratio edit module is the "core" component of PV, used to validate most sectors’ basic data items. A ratio edit compares the ratio of two highly correlated items to previously established upper and lower bounds (tolerances). Ratios that fall outside of bounds are edit failures, and one or both items in a failing ratio are either imputed or flagged for an analyst's review. The ratio module implements the Fellegi-Holt model of editing in which the complete set of edits is considered simultaneously to determine a minimum number of fields (items) to change so the imputed record satisfies the edits. Sigman (1997) describes the development of this system, and Wagner (2000) provides a general overview of the system as originally implemented for the 1997 Economic Census.

Range edits compare an item to lower and upper bounds (tolerances). The PV range edit module includes single item tests and ratio tests that designate one data item as fixed (goldplated). In the latter case, the module will impute only the non-fixed edit-failing item. The tolerances used by the PV range edit module resemble the PV ratio edit tolerances; they are provided as ratio edits and are determined at the imputation cell level. However, each establishment has customized tolerances obtained by multiplying the imputation cell limits by the fixed data item’s value. The PV range edit module is utilized before and after the PV ratio edit module. PV range editing is used *before* ratio editing to correct rounding errors (values reported in $1 instead of $1000) in general statistics items such as Annual Payroll, 1st Quarter Payroll, Sales/Receipts, or Operating Expenses. Businesses tend to consistently (mis)report all dollar values in the same units. Consequently, the ratio tests for the dollar values do not detect any rounding errors (the factor of $1000 cancels in the ratio). However, ratio tests of dollar values to 1st quarter employment do. Consequently, the PV ratio module will inflate the 1st quarter employment value by 1000 to create a consistent (non-edit-failing record). The PV range edit module is used *after* the ratio edit module to validate industry-specific specialty items that tend to be poorly or unreliably reported but have a weak positive association with a well-reported general statistics item (e.g. number of hotel rooms (poorly reported) and receipts/sales(frequently reported)). The PV range edit module is also used to edit part-year reporters (births and deaths), which are not subjected to simultaneous ratio edits. The ratio edit module subjects each record to all edits and uses the same edit tolerances for all records, unless an establishment-level multiplier is provided to widen the tolerances in advance. However, many ratio edits are simply not applicable to births after the first quarter or deaths in the first quarter. In addition, the tolerances for part-year reporters are generally modified to reflect months-in-business, as are the imputation parameters.

Notice that the PV balance edit module is applied as the final module in the sequence presented in Figure 2. A balance edit enforces an additivity constraint on two or more items to a (reported) item total. One or more items in a failing balance edit must be adjusted. In most Economic Census applications, the data item that contains the balance edit’s total is goldplated after validation in the Ratio or Range module, and only the detail items are modified.

With the ratio and range modules, imputation is performed using a variety of methods and models. Items are imputed in a pre-specified order, with the order determined by (1) the historical reporting reliability of the data item[2] and (2) the availability of reliable auxiliary data for imputation. Subject-matter experts provide a list of imputation methods for each item. Logical, auxiliary data, and establishment-level historical imputation models always precede industry average and midpoint imputation. As a rule, "imputed" values that use data provided by the same business on the same census form are preferable, and auxiliary data imputation (direct substitution) of auxiliary data on the same establishment is preferred over model imputation. Before attempting any of the pre-listed imputation models, the PV ratio module always attempts to "impute" the originally provided value, then the originally provided value/1000.

The frequency of use of different imputation methods varies across industry sectors and items. For example, in the manufacturing sector, the most frequently used imputation method for total value of shipments and total cost of materials is a univariate regression model, while the most common method for replacing dubious survey data for annual payroll is to use administrative records data (White, Reiter, and Petrin 2018). In part, these differences are driven by the availability of alternative data sources. The Census Bureau's Business Register includes administrative record data for annual payroll, quarterly payroll, 1st quarter employment, and EIN-level revenue. Although some imputation methods (such as using administrative records data for annual payroll) may preserve the dispersion and correlations that we see in the edit-passing reported data, other methods, such as univariate regression, produce significantly less dispersion in the imputed data than in the edit-passing reported data (White, Reiter, and Petrin 2018) and do not preserve the correlations that we see in the reported data (Kim et al. 2015). Appendix 1 displays the studied industries and data items, along with the edits by processing trade area. The first column provides the trade areas. The second column lists the studied industries for each trade area. Column 3 displays the items in PV ratio edits for each industry; each of these items appears in at least one explicit ratio edit supplied by a subject matter expert. For example, annual payroll (AnnPay) and employment in the first quarter (Emp1Q) are ratio edited by $L \leq AnnPay/Emp1Q \leq U$, for values of L and U specific to the trade area and industry (or imputation cell within industry). Column 4 indicates the items that are subjected to single item range edits developed by the synthetic data research team. The range edits are not included in the production system but are necessary to prevent generating overly large or negative synthetic data items. For example, the value of the annual payroll must lie within industry-specific lower and upper bounds, $L \leq AnnPay \leq U$. [Note: If there are no imputation cell-specified bounds we set the lower bound to 0 and the upper bound to infinity (represented by a very large number.)] The last column (Column 5) displays the items that must satisfy an additivity constraint and the corresponding balance edits. There are two balance edits for the general statistics data items in the manufacturing (MAN) and mining (MIN) trade areas, and no balance requirements for general statistics items in other trade areas. For example, in the MIN trade area, the reported annual payrolls for production workers (AnnPayPW) and other workers (AnnPayOM) must add to the total reported annual payroll (AnnPay) for each record, i.e. $AnnPay = AnnPayPW + AnnPayOM$.

3. Synthetic Data Generators

We considered two different synthetic data generators:

- The two-step data generator that produces fully synthetic data discussed in Section 3.1.
- A one-step data generator that produces partially synthetic data discussed in Section 3.2.

[2] Determined analytically and via onsite visits and interviews with businesses.

In this context, fully synthetic data have entirely modeled synthetic values and there is no concordance between the establishments in the original (input) data and the synthetic datasets. With the partially synthetic data, there is a one-to-one correspondence between the original data establishments and the synthetic data establishments. As with the fully synthetic data, all of the sensitive items are replaced with modeled synthetic values.

3.1. Fully synthetic data

3.1.1. Model

We used a two-step process to draw synthetic microdata $\boldsymbol{Y}_{Syn}^{*} = \{\boldsymbol{y}_1^{*}, \cdots, \boldsymbol{y}_n^{*}\}$ from its posterior predictive distribution:

$$f\left(\boldsymbol{Y}_{Syn}^{*} \middle| \widetilde{\boldsymbol{Y}}_{O}, \boldsymbol{E}\right) = \int f\left(\boldsymbol{Y}_{EI} \middle| \widetilde{\boldsymbol{Y}}_{O}, \boldsymbol{E}\right) f\left(\boldsymbol{Y}_{Syn}^{*} \middle| \boldsymbol{Y}_{EI}, \boldsymbol{E}\right) d\boldsymbol{Y}_{EI}$$

where

$\widetilde{\boldsymbol{Y}}_{O} = \{\widetilde{\boldsymbol{y}}_1, \cdots, \widetilde{\boldsymbol{y}}_n\}$ denotes the original confidential microdata

$\boldsymbol{Y}_{EI} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$ denotes the post edit-imputation microdata

$\boldsymbol{E}$ = the edit rules for the imputation cell

We generate *n x m* sets of synthetic microdata $\boldsymbol{Y}_{Syn}^{*}$ y:

1. Drawing $\boldsymbol{Y}_{EI}$ from $f\left(\boldsymbol{Y}_{EI} \middle| \widetilde{\boldsymbol{Y}}_{O}, \boldsymbol{E}\right)$ given the confidential and microdata $\widetilde{\boldsymbol{Y}}_{O}$, creating *n* completed census datasets
2. Draw *m* sets of $\boldsymbol{Y}_{Syn}^{*}$ from $f\left(\boldsymbol{Y}_{Syn}^{*} \middle| \boldsymbol{Y}_{EI}, \boldsymbol{E}\right)$ given each of the *n* sets of post edit-imputation microdata $\boldsymbol{Y}_{EI}$

Following Kim, Reiter, and Karr (2018), we used the Dirichlet process (DP) Gaussian mixture model as a data synthesis engine, with the posterior predictive distribution specified as

$$f(\tilde{y}_i | \{\Theta_k\}, Y_n) = c_1(\Upsilon, \{\Theta_k\}) \sum_{k=1}^{K} \eta_k^{*} \, N(\log y_i^{*}; \mu_k^{*}, \Sigma_k^{*}) I(y_i^{*} \epsilon \, \Upsilon) \qquad (3.1)$$

where $\Theta_k = \{\eta_k, \mu_k, \Sigma_k\}$ is the k^th mixture component, Υ represents the feasible region defined by the set of edits applied in the imputation cell, y_i^{*} are the data items to be modeled, and $\eta_k^{*}$is the mixture-component weight. The number of mixture components is set as an upper bound in each individual application; not all mixture components will be occupied.

We used an MCMC process to draw values of model parameters and then drew edited/imputed and synthetic datasets conditional on those values of the model parameters with a *minimum* burn-in of 5,000 and drawing edited/imputed or synthetic datasets at each 200th iteration. This approach preserves multivariate relationships between items but does not preserve the marginal distributions, leading in turn to variable totals. Noninformative priors are used for each item. In future research, this approach could be modified to take the relative reliability of the data-items into account). The number of occupied mixture components varied greatly by industry/imputation cell.

Estimates and variance estimates for any estimator $\hat{\theta}$ constructed from the multiple sets of synthetic data are given by

Estimator $\hat{\theta} = \frac{1}{n}\sum_{l=1}^{m}\hat{\theta}^{(l)} = \frac{1}{mn}\sum_{l=1}^{n}\sum_{s=1}^{m}\hat{\theta}^{(l,s)}$ where $\hat{\theta}^{(l,s)}$ is computed within each synthetic dataset and $\hat{\theta}^{(l)}$ is the synthetic dataset average estimate drawn from the *l*th edited/imputed dataset

Variance $\hat{V} = \left(1 + \frac{1}{n}\right)B_n - \frac{b_n}{m}$

$$B_n = \frac{1}{n-1}\sum_{l=1}^{n}(\hat{\theta}^{(l)} - \hat{\theta})^2$$

$$b_n = \frac{1}{n(m-1)}\sum_{l=1}^{n}\sum_{s=1}^{m}(\hat{\theta}^{(l,s)} - \hat{\theta}^{(l)})^2$$

Our applications omit the third variance component (within-synthetic-dataset variance estimate as given by $\bar{u}_n = \frac{1}{nm}\sum_{l=1}^{n}\sum_{s=1}^{m}u^{(l,r)}$) because we are creating synthetic finite populations instead of synthetic samples (Vink and Van Buuren 2014 and Kim et al 2018).

3.1.2. Input Data and Parameters

As mentioned in Section 2, this research is limited to general statistics items (items available for all establishments in a sector) that are edited with the Plain Vanilla modules. Input data are restricted to full year reporter units[3].

The edit/imputation programs have two input data requirements:

- Input data must contain at least one record whose data satisfy the entire set of edits
- Input records must include at least one non-missing item. If a record consists of **one** non-missing item, then the programs will impute a complete record that satisfies the entire set of edits. In this case, the imputed record retains the original item value.

Initially, we planned to edit/impute/synthesize all of the variables that are edited and imputed using Plain Vanilla. Ultimately, we ended up dropping selected variables. For example, in the manufacturing and mining sectors, many total items are collected along with detailed breakdowns (e.g. total employees = production workers + other employees). Because of low response and high imputation rates, we ended up dropping most or all of the detail items from this process. We also found some collection inconsistencies with selected "specialty" items. For example, there were several establishments in tax-exempt services industries that did not report operating expenses and whose imputed operating expenses were $0. In these cases, we learned that an additional classification variable value was used in the production edit system. We did not have this variable and were not provided with the conditions governing operating expenses imputation, so we omitted the edit constraints for operating expenses and did not synthesize the data item.

Of course, many eligible (sampled) establishments did not respond to the economic census. If administrative payroll data was available for these delinquent units, we imputed it in the input data, as long as the value was greater than $12,000 (the assumption was that a full-year reporting establishment with one employee would pay at least $1000 per month.) We enforced the same restriction on establishments that only reported annual payroll, substituting administrative data if the reported value

[3] Any establishment that is in business for 10 or months during the reference year and is not explicitly flagged as a death is included (i.e. first quarter births are included, as are some seasonal operators).

was less than \$12,000, and dropping all single-item records with (reported or administrative-data imputed) values less than this lower bound.

Lastly, we corrected the rounding errors (values reported in \$ instead of \$1000) in the input data for annual payroll, 1st quarter payroll, and sales. This prevented the imputation and subsequent synthesis of overly large values of 1st quarter employment.

Subject matter experts provided the ratio tests and tolerances from the 2012 Economic Census Plain Vanilla parameter files (Ratio and Range edits, as applicable), along with detailed instructions for defining the imputation cells within industry. When possible, we used the balance edits provided in the manufacturing and mining production edit scripts, although many of the detail items were dropped in our process due to input data restrictions.

We created three additional range edits per imputation cell for annual payroll, 1st quarter employment, and sales: (0, maximum value of item *x* 1.25). This reduced the synthesis of unreasonably large unit-level values in the edit/imputed and synthetic datasets, although it did not always prevent it.

It took a while for the team to develop comprehensive criteria for the input data. The editing and imputation process can be very slow when the models contain more than four or five variables, and it was at times very difficult to detect whether this was due to the number of variables in the datasets, the number of observations in the imputation cell or industry, or nuances in the data. Because the modeling process does not take relative reporting reliability of different items into account, we had to make the decisions at the front end on whether to include or drop variables. Including balance edits in the edit set often proved too restrictive, which led to dropping the balance edits and associated items in many industries. In fact, we did not including any balance edits or associated detail items in the partial data synthesis procedures described in the next section. Finally, we had to make difficult decisions on when to replace reported data values with analyst corrections, as the PV analyst correction flag does not distinguish between an analyst "correction" and an analyst "imputation." With the former, the correction was almost always some form of rounding (divide by 1000, 100, or 10). With the latter, the differences between the reported and corrected values tended to be negligible, and we retained the reported values.

In a production setting, including subject matter experts in the decision making process for input data requirements and edit inclusion would be required. Our decisions relied on our own data analysis and could violate (unknown) production requirements.

3.2. Partially Synthetic Data

3.2.1. Model

We used a one-step process to draw *m* sets of partially synthetic microdata $\boldsymbol{Y}_{Syn}^{*} = \{\tilde{\boldsymbol{y}}_1^{*},\ \cdots, \tilde{\boldsymbol{y}}_n^{*}\}$ from the posterior predictive distribution:

$$f(\tilde{y}_i \mid \{\Theta_k\}, Y_n) = c_1(\Upsilon, \{\Theta_k\}) \textstyle\sum_{k=1}^{K} \eta_k^{*}\, N(\log y_i^{*}; \mu_k^{*}, \Sigma_k^{*}) I(y_i^{*} \epsilon\, \Upsilon) \qquad (3.2)$$

where $\Theta_k = \{\eta_k, \mu_k, \Sigma_k\}$ for the kth mixture component, and $Y_n = \{y_{surv,i},\ MOS_i: i = 1, \ldots, n\}$, the survey data (to be synthesized) and a Measure of Size (MOS) variable (to aid in the synthesis). This unit-level MOS variable is *not* simulated and is included on the output synthetic datasets. These are partially synthetic data, as the synthetic data are generated from the conditional predictive distribution given the observed values of MOS. Otherwise, the modeling procedures are identical to those described in Section 3.1.1. for fully synthetic data. In general, fewer mixture components were needed than for the fully synthetic counterparts.

Synthetic data estimates for any estimator $\hat{\theta}$ and variance estimates are given by $\hat{\theta} = \frac{1}{m}\sum_{l=1}^{m} \hat{\theta}^{l}$ (estimate) and $\hat{V} = \frac{b_m}{m}$, where *m* is the number of synthetic datasets. Again, the within-synthetic dataset variance is dropped from the computations, as the finite population is synthesized.

3.2.2. Input Data and Parameters

For comparability with the fully synthetic data products, the team used the *set* of full-year reporter establishments and general statistics items as input data. However, the input data consisted of the final (edited/imputed, tabulated) microdata associated with each matched unit ID. Because the final data often contain values of items that fail at least one edit (goldplated items), we removed the requirement input data that the dataset must contain at least one record whose data satisfies the entire set of provided edits; we still required that the synthetic data satisfy the entire set of edits.

We used the same sets of ratio and range edits for partially synthetic data generation as with the fully synthetic data: (1) ratio tests and tolerances used in the 2012 Economic Census Plain Vanilla parameter files (Ratio and Range edits, as applicable) and (2) the imputation cell level range edits for annual payroll, 1st quarter employment, and sales. In the process of developing fully synthetic data, we observed some detrimental effects on the synthesized totals data items (annual payroll and 1st quarter employment) when including the associated detail items and balance edits in the synthetic data generation process. The generators tended to model the detail data items separately and derive the totals data items by additions, leading to “noisy” totals estimates. This disadvantage – coupled with greatly simplified programming requirements – led us to drop all balance edits from the partially synthetic data generator.

4. Utility Metric Analysis

4.1. Definition

The team developed a single data utility score applied to each industry based on three different statistics:

(i) the absolute value of differences between correlations of key (log-transformed) variables in the synthetic data vs. the edit-passing reported data;

(ii) the relative bias of industry totals from the synthetic data taking the totals from Census-edited/imputed data as the true values; and

(iii) the relative bias of average ratios of key items, calculated from the synthetic data, taking the average ratios from the Census edited/imputed data as the true values.

A separate utility metric was obtained for each imputation cell within an industry, with the imputation cell’s contribution to the industry total given by (aggregate sales for imputation cell)/(aggregate sales for industry) computed from final tabulated (publication data). Table 2 outlines the components for the utility scores. Note that the point scoring system used subjectively determined critical values, chosen *before* the evaluations.

Within an imputation cell, the final score is the sum of these three components, divided by the maximum number of points (7 for manufacturing and mining; 10 for all other sectors). Weighted scores for each imputation cell were aggregated to obtain an industry total score, valued between (0,100), with a higher score indicating higher utility. Imputation cells whose data could not be synthesized were assigned total scores of zero.

Table 2: Components of Utility Scores

Statistic	Definition	Comparisons
Correlation	DiffCorr = (Synthetic Data Correlation – True Correlation) Points: 1 if $\|\text{Diff}_{\text{Cor}}\|< 0.10$; 0 otherwise • With correlation is computed from (ln(X),ln(Y)) • "True Correlation" computed from reported data that satisfied all ratio edits • Drop records containing zeros or missing values	(AnnPay,Emp1Q) (Sales,AnnPay) (AnnPay,Pay1Q)*
Total	Relative $\text{Bias}_{\text{Item}}$ = (Synthetic Item Value – True Item Value)/(True Item Value) Points: 1 if $\|\text{Relative Bias}_{\text{Item}}\|\leq 0.05$; 0 otherwise • "True" Item Value computed from final edited/imputed data for input records (not published totals)	AnnPay Pay1Q* Emp1Q Sales
Ratio	Relative $\text{Bias}_{\text{Ratio}}$ = (Synthetic Ratio Value – True Ratio Value)/(True Ratio Value) Points: 1 if $\|\text{Relative Bias}_{\text{Ratio}}\|\leq 0.05$; 0 otherwise • "True" Value computed from final edited/imputed data for input records (not published totals) – same Totals used for Ratio and Totals components	AnnPay/Emp1Q Sales/AnnPay AnnPay/Pay1Q*

*Pay1Q is not collected in the manufacturing or mining sectors.

We computed utility scores from the fully synthetic data with five edited/imputed datasets (*n* = 5) and 5 or 10 synthetic datasets generated from each edited/imputed dataset (*m* = 5 and *m = 10*) and from the partially synthetic data with *m* = 5 and *m = 10.* We examined these utility scores to answer two different questions:

- For a given synthetic data generator, how many synthetic datasets should we create to "maximize" utility over the studied industries?
- Which synthetic data generator has better performance in terms of this utility metric over the studied industries?

One caveat: we developed this utility score with a general-purpose framework, with an emphasis on data characteristics known to be important to a wide-variety of data users. We suggest considering additional utility metrics for the synthetic data once we have identified a single generator.

4.2. Results

The following section summarizes utility metric scores within method for all 42 studied industries. Utility scores are rounded to zero decimal places for all comparisons.

4.2.1. Fully Synthetic Data

For this two-stage development procedure, we used five multiply-imputed datasets for the 1st (edit/impute) stage as recommended in Kim et al. (2015). Figure 3 compares utility scores obtained creating five synthetic datasets per edited/imputed dataset (FSYN5) to ten synthetic datasets per edited/imputed dataset (FSYN10) i.e. 25 synthetic datasets versus 50 synthetic datasets.

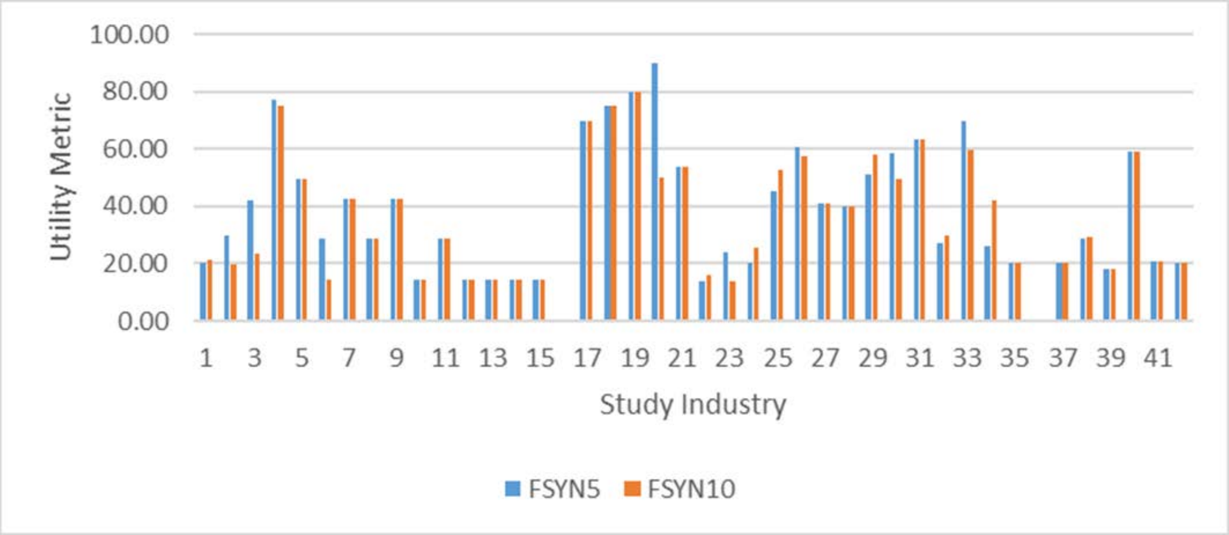


Figure 3: Comparison of Utility Scores for Fully Synthetic Data

Figure 3 depicts 26 ties between FSYN5 and FSYN10, 9 industries where the FSYN5 has higher utility than the FSYN10 counterpoint, and 7 industries where the reverse is true. In the majority of industries, the correlation component of the utility was generally good (by design), but the industry totals were far from the true totals, and consequently the ratios were equally inadequate. Adding synthetic datasets did not improve these latter properties.

4.2.2. Partially Synthetic Data

For this one-stage development procedure, we considered generating five and ten synthetic datasets per industry. Figure 4 compares utility scores obtained creating five synthetic datasets (PSYN5) to ten synthetic datasets (PSYN10).

Figure 4 depicts 23 ties between PSYN5 and PSYN10, 12 industries where the PSYN5 has higher utility than the PSYN10 counterpoint, and 7 industries where the reverse is true. In general, the totals tended to be closer to the true totals than with the fully synthetic data, and there was little apparent degradation in the correlation results.

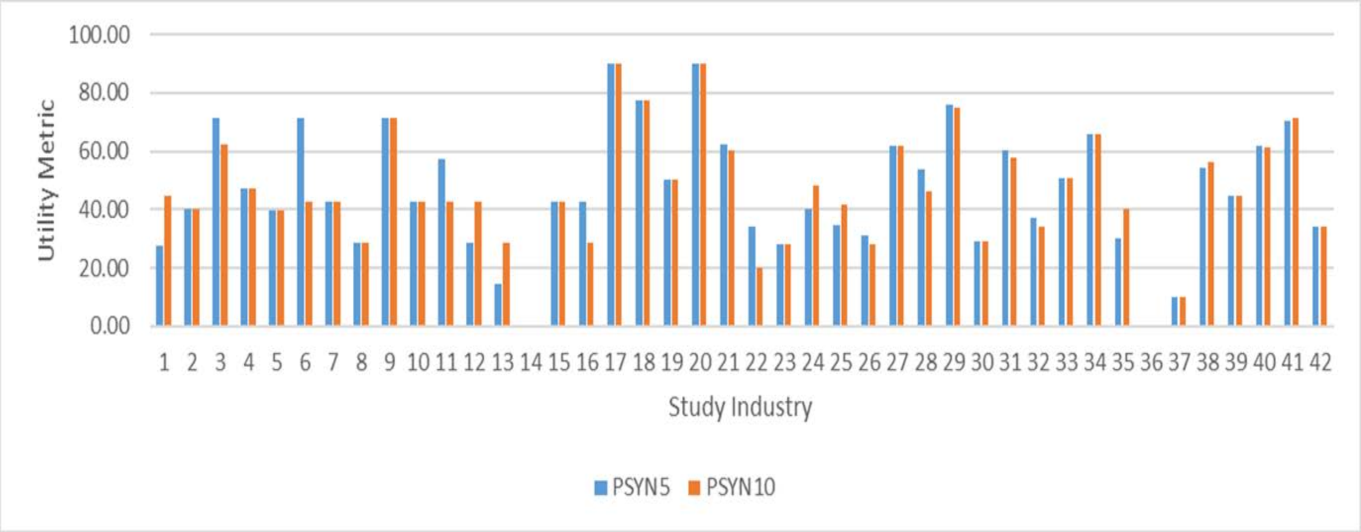


Figure 4: Comparison of Utility Scores for Partially Synthetic Data

4.2.3. Fully Synthetic Data (25 Datasets) Versus Partially Synthetic Data (10 Datasets)

Figure 5 compares utility scores within industry for 5 fully synthetic datasets (FSYN5) to scores for 10 partially synthetic dataset (PSYN10). There are four industries where the two methods tie. In 28 industries, the partially synthetic data have higher utility, with the minimum difference between corresponding datasets of two, a maximum difference of 50, and a median difference of 20. In 10 industries, the fully synthetic data has higher utility, with a minimum difference between corresponding datasets of four, a maximum difference of 33, and a median difference of 16.5.

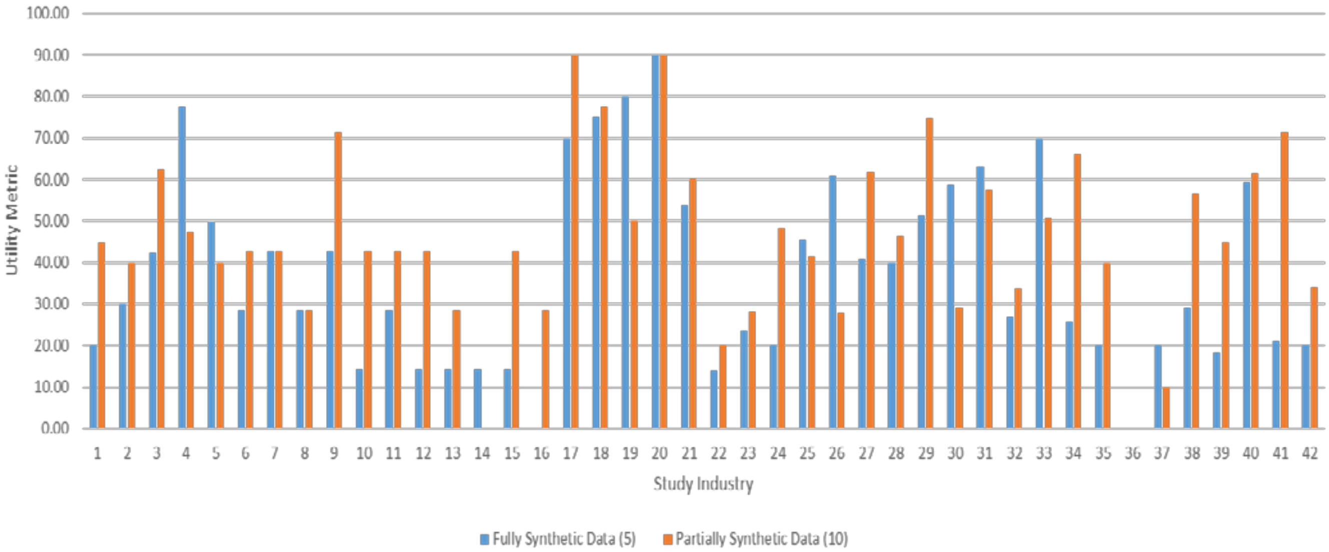


Figure 5: Comparison of Utility Scores for Fully Synthetic Data (25 datasets) and Partially Synthetic Data (10 datasets)

In the majority of industries, the partially synthetic data has improved utility over the fully synthetic data. Furthermore, the differences in utility are less pronounced, with the fully synthetic data having higher utility. Recall that the fully synthetic data generator is designed to preserve multivariate data relationships and does not preserve marginal moments. Because it does preserve the totals, the generator does not capture the linear relationship used for ratio editing/imputation in many cases. In contrast, the partially synthetic data generator preserves multivariate relationships and marginal moments, as long as there is a strong association between the measure of size and set of study variables.

There are other production considerations. The fully synthetic data generator places restrictions on input data. Ultimately, the team had to exclude a variety of units because of these restrictions; in some cases, entire imputation cells were dropped. These restrictions on the input data are not applicable to the partially synthetic data generator. More bookkeeping is required for variance estimation with the fully synthetic data than for the corresponding partially synthetic data. Lastly, there are straightforward modifications to the partially synthetic data generator to include part-year reporter establishments (births and deaths); the modifications are more challenging with the fully synthetic data generator. Consequently, the team decided to restrict further analyses to partially synthetic data.

4.3. Relationship Between Characteristics Of Input Data And Partially Synthetic Data Model "Success"

This section discusses exploratory analyses designed to identify factors/characteristics in the input data that were related to the level of the utility metric; see Table 3 for the complete set of independent analyses methods. For each analysis, the utility score (p_score10) is the dependent variable. Table 4 lists the "imputation cell" level variables provided and summarizes the variables included in each independent evaluation.

Table 3: Methods Used to Identify Industry Characteristics to Predict Utility

	Method	Selection Procedure	Selection Criteria
1	OLS Regression (no intercept)	Stepwise	SS2 SSE AIC
2	OLS Regression (no intercept)	Forward Selection	SS2 SSE AIC ADJRSQ
3	OLS Regression (no intercept)	Backwards Elimination	SS2 SSE AIC
4	Adaptive LASSO	Prescreen with random forest	AIC
5	Regression Tree	Iterative greedy algorithm	SSE
6	Regression Tree	Iterative greedy algorithm	Mean squared error
7	Generalized Linear Models (no intercept)	Stepwise	AIC
8	OLS Regression	Stepwise	SS2 SSE AIC ADJRSQ
9	Generalized Linear Models (no intercept)	LASSO	AIC
10	OLS Regression	Backwards Selection	AIC
11	Generalized Linear Models	Stepwise (with categorical as class variables)	Not Provided
12	Neural Networks	3 variable samples	Lowest Average SSD
13	Neural Networks	4 variable samples	Lowest Average SSD
14	Neural Networks	5 variable samples	Lowest Average SSD

Taken collectively, three characteristics substantively contribute to the utility of the partially synthetic data: number of establishments, proportion of multi-unit establishments in the imputation cell, and item response rate for sales or annual payroll. The consistently positive association between number of establishments and increased utility is intuitive. The consistently positive association between proportion of multi-unit establishments and utility is not. In the studied industries, companies are requested to complete a census form for each of their establishments; we excluded industries that report consolidated sales for all establishments. As the company size increases, it is more likely that the census forms are completed in a single accounting office, in turn potentially leading to a smoothed multivariate distribution. Equally likely, the industry average ratios that are used to determine the ratio edit limits will increasingly reflect the larger company distributions as the proportion of multi-units increase. Lastly, in many sectors, the Census Bureau performs an extensive "pre-edit" on annual payroll. If the original reported value is modified, the remaining items may be modified in subsequent edits. Since the Economic Census imputation procedures are univariate (one-variable-at-a time), the multivariate relationships in the final edited and imputed data can be extremely irregular. Consequently, it makes sense that a high response rate for annual payroll would lead to improved modeling utility. Along the same lines, sales are usually lower in the imputation order than the payroll or employment items. Thus, a multivariate observation for an observation with an edited/imputed value of sales could be very isolated from the bulk of the multivariate distribution, depending on the imputation model used and the reported data status of the other items.

Table 4: Independent variables available for independent analysis. “Included” counts are the number of independent analyses in which the variable was considered; “significant” counts are the proper subset of items that appeared in the final analysis model.

Item	**Description**	**Analyses**	
		Included (Count)	**Significant (Count)**
imputation_cell	2012 EC imputation cell (9 characters)	1	0
Trade	MIN, MAN, RET, SER, WHO, FIR, UTL	1	0
max_num_var	Maximum number of variables in synthetic dataset	10	0
est_count	Number of establishments	11	9
est_count_2500	Number of establishments top-coded at 2500	7	1
Size	Size category based on total establishments	11	3
prop_mu	Proportion of multi-unit establishments	14	13
unit_resp_rate	Proxy for unit response rate	13	0
AnnPay_resp_rate	Proxy for item response rate for AnnPay	13	8
Sales_resp_rate	Proxy for item response rate for Sales	14	6
birth_rate	Proportion of births	10	0
death_rate	Proportion of deaths	11	4
full_year_rate	Proportion of full reporters and births in quarter 1	14	4
AnnPay_spread	Standardized measure of spread for AnnPay	13	0
Emp1Q_spread	Standardized measure of spread for Emp1Q	14	1
Sales_spread	Standardized measure of spread for Sales	13	4
Corr_ANNPAY_Sale	Correlation between logged ANNPAY and logged Sales	13	0
Corr_ANNPAY_Em	Correlation between logged ANNPAY and logged	13	3
Corr_Sales_Emp1Q	Correlation between logged Emp1Q and logged Sales	10	0
Weight	Weight used to compute utility score	10	0

4.4. Limitations of the Research

One purpose of this project was to investigate the feasibility and usefulness of synthetic Economic Census microdata. Here we describe some limitations of the research. Management and future researchers may want to take these limitations into account when evaluating the findings and/or planning subsequent synthetic data projects.

The 2012 Economic Census included data for over 950 6-digit NAICS industry classifications. Ratio edit rules in the Economic Census are specific to an imputation cell, and in many industries there are multiple imputation cells per industry. Developing synthetic data for every industry was beyond the scope of the project. The team initially chose 54 industries (across 18 sectors) recommended by Census Bureau industry classification experts, and settled on 42 industries with adequate data. These industries had consistent NAICS definitions for the 2012 and 2017 Economic Censuses, varied in size and scope, and reported varying numbers of detailed products (note: products not considered in this research). The point of selecting certain industries based on their characteristics was to be able to assess the usefulness of the new methodology for generating synthetic data for these studied industries using 2017 Economic Census data. However, this set of industries *not* representative of the Economic Census as a whole.

The Economic Census collects data on hundreds of variables, many of which are relevant for only a few industries. Given the exploratory nature of the project, the project focused on a limited set of variables that are common across industries within each sector. The team also limited the research to items that are edited for the Economic Census within the Plain Vanilla editing and imputation system. This latter restriction excludes some items that are common across all industries in certain sectors, such as the value of assets and capital expenditures in mining and manufacturing. To the extent that these variables have different properties than variables we selected (e.g., if they have high rates of missingness and are not highly correlated with the other variables), then our research conclusions may not extend to those variables.

Another potential limitation of the current research is that it is not entirely clear who would use the synthetic data created using the methods described in this report. Although there is precedent for the Census Bureau releasing synthetic demographic microdata, such as the Synthetic Survey of Income and Program Participation (SIPP), synthetic business microdata released by the Census Bureau has been limited to the synthetic Longitudinal Business Database (LBD), which was first released in 2007. A relatively small number of researchers have used the synthetic LBD (Miranda and Vilhuber 2014), and the Census Bureau currently has no plans to produce a more up-to-date version. This is relevant for the synthetic Economic Census data, because the utility of any dataset depends to some extent on what questions researchers want to use it to answer. Although external researchers can currently access the confidential microdata for the Economic Censuses from 1977 to 2012, the process for gaining approved access to the data via the Federal Statistical Research Data Centers (FSRDC) can take several months to a year, and not all research institutions have access to an FSRDC.[4] One possibility is that researchers or other data users may want to use the synthetic data for preliminary analyses prior to submitting a proposal for access to the confidential data via the FSRDCs. Obviously, the analyses that researchers could do with the synthetic data would be much more limited than what they might do with the confidential microdata, because of the limited number of variables and industries in the synthetic data. On the other hand, the time between coming up with a research idea and being able to publish results using synthetic data could be a fraction of the time that it takes to do the same thing with confidential Census microdata. The synthetic data could also be of use to a different academic audience. Statisticians are very interested in realistic datasets for developing estimators and validating imputation models (e.g. hot deck) and many data users have expressed interest in using microdata for exploratory analysis without wanting to write a specific justification for an FSRDC. Furthermore, many researchers do not have easy access to an FSRDC.

Interpretation of the utility metric discussed above in Section 4 of this report is related to the issue of who will use the data. Our team's subjective definition of utility attempted to encompass a wide-range of desirable data characteristics. Of course, whether or not these are the "right" utility metrics depends on what question the researcher wants to use the data to answer. For example, even if an industry has a high score for this utility metric, if a researcher is primarily interested in a relationship between two variables that are not accounted for in the utility metric (e.g., the difference between the total value of shipments and the total cost of materials in manufacturing), then the data may not have high utility for that researcher. Furthermore, the studied associations are only valid for linear models, either on the original scale or on log-transformed values.

On the other hand, if the researcher does not care if the synthetic data matches the industry totals — for example, if the interest is restricted to a subset of the establishments within the industry as in

[4] However, the number of institutions with access to an FSRDC has increased greatly since the synthetic LBD was first released in 2007. In 2007, there were eight FSRDCs. At the time of writing, there are 29 FSRDCs in 18 states.

Foster, Haltiwanger, and Syverson (2008)— then the synthetic data may have high utility for that researcher even though it scores low on the single metric, as long as associations between variables are preserved.

5. Privacy Protection Assessments

The team considered five univariate privacy metrics and one multivariate metric, again focusing on the general statistics items reported in all sectors. The table below briefly summarizes key properties of the generated data.

Categorical variables	Included in Synthetic Data	6-digit NAICS Type of Operation (Wholesale) Tax Exempt Status (Services)
	NOT Included or Modeled in Synthetic Data	Company identifier Geographic identifiers Establishment type (Full year, birth, death)
Continuous variables	Modeled	General statistics items (3 for all sectors, 4 for all but mining and manufacturing) • Sampled units – directly collected • Unsampled units – administrative data substitution when available Other variables collected from all eligible sampled units by sector • Some variables dropped for synthesis due to sample size constraints (original data)
	Not modeled	Products and special inquiries (see Kim, Dreschler, Thompson 2020)

It should be noted that the lack of geographic and company identifiers in the synthetic generation process greatly reduces the synthetic data tabulations' sensitivity (and utility) over comparable tabulations published by the economic programs. The metrics described below assess the disclosure risk of the *establishments*, not their parent companies (firms). The privacy protection metrics for the economic tabular data evaluate disclosure risk with respect to the latter.

We make the following assumptions in measuring the disclosure risk in the partially synthetic datasets:

- The intruder assumes that the population coverage is complete, i.e., the synthetic datasets contain one record for each ***full-year reporter*** establishment in the industry. Furthermore, industry classification, merchant wholesaler type of operation (warehouse, agent, broker) and tax-exempt status are publicly available and are not required to be protected.
- The intruder is interested in gaining information about data item *attributes* (values).
- The intruder has previous knowledge about the exact value of one or more data items in the original data for one or more identified establishments.

All of the privacy protection metrics discussed below measure disclosure sensitivity with respect to the attribute on the ***original scale***. This poses challenges, as three of our variables are collected in thousands of dollars, whereas 1st quarter employment is a count variable. A variety of robust and resistant multivariate statistical outlier detection methods effectively identifies unique and isolated units. However, these methods require standardization. An attribute identification based on a standardized variable seems unrealistic, as the intruder would need access to the *complete* original data. Furthermore, an attribute identification based on power-transformed data would be equally questionable; for example, the log transformation compresses the largest values and leaves the smallest values essentially unchanged; an "isolated" multivariate observation on the original scale might be quite consistent with the rest of the distribution on the transformed scale. Lastly, there is a "smell test," in the sense that it seems highly unlikely that an intruder would first apply a power transformation to their known information. If s/he did, it seems unlikely that claim of a disclosure based on a power-transformed value would stand up in a court of law, given the caveats above.

5.1 *Metrics*

Univariate Metrics
Metric 1 compares the estimated largest value of each studied data item from the synthetic data (averaged across all replicates) in a given industry to the largest value in the original "true" data, defined for industry *i* as

$$RB_{i_{(n)}} = (\hat{Y}_{i_{(n)}} - Y_{i_{(n)}})/Y_{i_{(n)}}$$

where $Y_{i_{(n)}}$ is the largest value of data item *Y* in the original ("true") data and $\hat{Y}_{i_{(n)}} = \frac{1}{S}\sum_s \hat{Y}^s_{i_{(n)}}$ is the average value of the largest order statistic for data item *Y* in each of the *S* synthetic datasets. We consider this value to be a high level of disclosure risk when$\left|RB_{(n)}\right| \leq 0.05$.

Metrics 2-4 are patterned after the p-percent rule used widely in the Census Bureau's Economic Directorate, to determine the sensitivity of a tabulation cell to target the cell for primary suppression. The p-percent rule assumes that the intruder (plus up to *c* = 2 collaborators) uses industry tabulations obtained from the synthetic data to bound the value for the largest establishment. To obtain a lower bound on the precision of the intruder's (s') attack, let $Y_{ijk_{(n-1+c)}}$ be the true value of the second largest establishment[5] in the original input data, $\hat{T}_{i(y)} = \frac{1}{S}\sum_s \sum_{k \in i} \sum_{j \in k} \tilde{y}_{ijks}$ be the estimated population total for item *j* obtained by averaging the totals from the *S* synthetic datasets, and$\tilde{Y}^c_{ijk_{(n)}} = \hat{T}_{i(y)} - \sum_c Y_{ijk_{(n-1+c)}}$ be the attacker's estimate of the largest establishment's value for item *i*. Metrics 2 through 4 are defined as

$$P_{(n)} = (\tilde{Y}^c_{ijk_{(n)}} - Y_{ijk_{(n)}})/Y_{ijk_{(n)}}$$

for *c = 0 (Metric 2), 1 (Metric 3), or 2 (Metric 4).* We consider this value to be a high disclosure risk when$\left|P_{(n)}\right| \leq 0.05$.

Metric 5 is designed to compare the utility in terms of relative L1 error of the partially synthetic industry-level item totals of annual payroll, 1st quarter employment, and receipts to correspondingly ε-differentially private totals obtained from the input microdata. This metric assesses whether the estimated *synthetic totals* are comparably noisy to totals obtained using formal privacy protection methods such as differential privacy, which provide provable privacy guarantees limiting inferential disclosures.

With skewed economic data, straightforward implementation of differential privacy methods is often not feasible due to the amount of noise required to guarantee sufficient protection (Haney et al 2017). However, using the upper limits of the range edits used for the synthesis to obtain L_1 sensitivity, we can

[5] and 3rd or 4th, depending on the number of collaborators

obtain the ε-differentially private totals for each item for an industry, setting the overall privacy budget as ε, equally divided between the *l* imputation cells within the industry. Noisy totals are obtained by adding draws from the Laplace distribution proportional to the imputation cell's sensitivity and privacy budget, i.e $\ddot{T}_{ij} = \left(T_{ij}\right) + \text{Lap}(b)$ where $b = \Delta_q/(\varepsilon\,/l)$ and *q* represents the (queried) total. ε-differentially private totals are computed as $\ddot{T}_{iJ} = \sum_{j\in J}\ddot{T}_{ij}$. Relative L1 error is given as $RL1_{iJ} = \frac{|\ddot{T}_{iJ} - T_{ij}|}{T_{ij}}$.

This metric allows comparison of the aggregate noise needed to obtain differentially private totals for ε = 0.5, 1, 1.5, and 2 to that created in our synthesis process. Our choices of privacy budget are arbitrary, with ε=0.5 representing an unrealistically high level of protection, and ε=2 being the lower limit of what might provide acceptable utility with highly skewed economic data. For example, the LEHD "On the Map" program has used ε=4.6; see Section VI. Experiments, https://lehd.ces.census.gov/doc/help/ICDE08_conference_0768.pdf. Having comparable relative L1 error loss for the differentially private and synthetic data totals provides evidence of strong privacy protection from the partially synthetic data generation process. We also use this method to assess the reduced protection in using ten synthetic datasets instead of five synthetic datasets.

Each differentially private total is assessed after 200 random draws of Laplace noise (i.e., by averaging 200 ε-differentially private totals). Synthetic data totals are derived from (1) all 10 synthetic datasets and (2) from all possible combinations of five synthetic datasets (254 total). Note that industries 312111 and 336612 were excluded from this analysis because no range edit files were provided.

Multivariate Metric

The multivariate metric employs a holistic view of the synthetic data, assessing the proximity of original data multivariate observations to their synthetic data counterparts. We define a disclosure to be when there is a strong probability of a "match" between a unique (isolated) observation in the original data set and an estimate of the corresponding observation from the synthetic data on the *original untransformed scale*. The primary assumption underlying this metric is that an intruder has exact information on the value of at least one variable for a given establishment that s/he will use to obtain attribute information (i.e. close approximations of value) about the other items.

The "match" depends on the attributes of the set of variables associated with the observation. We define a match as follows:

Let α = predetermined percentage of item value (tolerance around true value)
$\{X_D\}$ = set of variables used to define the pattern for *each establishment*
$\{X_{Di}\}$ = values of the variables for each establishment

For each establishment *i*, compute the disclosure pattern on the original data as $\text{DP}_i = \{X_{Di} \pm \alpha X_{Di}\}$. Figure 6 provides an illustration on a trivial example using bivariate data generated from a fictional multivariate lognormal distribution. The patterns are unequal as the length and width are defined by the individual establishment's data items values and the value of alpha (α). Consequently, Disclosure Pattern A is a square and contains two clustered observations. Disclosure Pattern B is a large rectangle, containing a single observation that is isolated from the bulk of the distribution and is presumably easier for a knowledgeable intruder to identify.

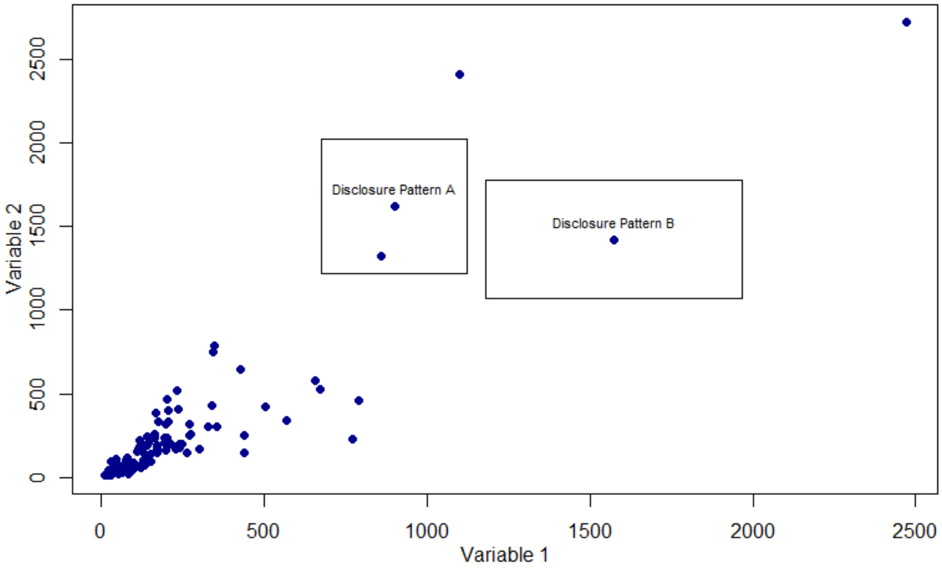


Figure 6: Fictional example illustrating disclosure patterns on bivariate data

In our evaluation, the pattern for each establishment uses annual payroll, employment, and sales/receipts. As the correlation between items decreases, it is more likely that there are isolated multivariate observations. Since 1st quarter payroll is nearly perfectly correlated with annual payroll for full-year reporters, we did not include it in our metric. A "match" is highly dependent on the tolerance set around the true observation, so the sensitivity of the metric should be assessed with a range of tolerances. We use α = 0.05, 0.10, 0.15, 0.20, 0.25. Note that the values of α that are smaller than 0.10 are inefficient for identifying observations with small values of 1st quarter employment. On the other hand, values of α that are larger than 0.15 are inefficient for identifying observations with large values of annual payroll and/or sales/receipts.

Definitions:

Uniqueness (U_i) measures the isolation of a multivariate observation in the original data from the bulk of the study distribution *on an ordinal scale*, given a predetermined α.

Value	**Definition**	**Isolation in Data**
1	1 observation in pattern associated with establishment *i* (*DP_i*)	Very isolated/unique
2	2 observations in pattern	↓
3	3 observations in pattern	
4	4 observations in pattern	
5	5 observations in pattern	
6	6+ observations in pattern	Very nonspecific

Proximity (PR_i) measures the proximity of the corresponding multivariate observation in the *synthetic data* from its original data counterpart *on an ordinal scale*, given a predetermined α and disclosure pattern (DP_i): $PR_i = \sum_{s=1}^{S} I_{is}$ where $I_{is} = 1 \leftrightarrow \{X_{Di,s}\} \in DP_i$, *s* =1, ..., *S* synthetic datasets.

Incidence (IN_i) is the percentage of synthetic observations *i* that fall within the corresponding disclosure pattern in the original data: $IN_i \approx \sum_{s=1}^{S} I_{is} / S$.

Proximity Value	**Definition**	**Incidence Value**	**Isolation in Data**
1	10/10 observations within *DP_i* bounds for establishment *i*	10/10	Perfectly approximated ↓ Poorly approximated
2	9/10 observations within bounds	9/10	
3	8/10 observations within bounds	8/10	
4	7/10 observations within bounds	7/10	
5	6/10 observations within bounds	6/10	
6	5/10 observations within bounds	5/10	
7	4/10 observations within bounds	4/10	
8	3/10 observations within bounds	3/10	
9	2/10 observations within bounds	2/10	
10	1/10 or 0/10 observations within bounds	1/10	

We consider an establishment as highly sensitive to disclosure risk when both its uniqueness and its incidence approach 1, i.e., it is very identifiable (isolated) in the original dataset and it is well approximated in the synthetic data for a given α. We define Attribute Identification Risk (AIR) for a given value of α as $AIR_{i(\alpha)} = \left(\frac{1}{U_{i(\alpha)}}\right) IN_{i(\alpha)}$.

5.2. Results (Partially Synthetic Data)

5.2.1 Univariate Metric Results

Appendix 2 presents the values for Metrics 1 through 4 for industry-variable pairs, by trade area. The red highlights indicate which industry-variable pairs exhibit high disclosure risk with Metric 1. The yellow highlights indicate which industry-variable pairs exhibit high disclosure risk for Metrics 2 through 4.

In general, the disclosure risks for Metric 1 are at acceptable levels, with the following exceptions:

- First quarter employment in two manufacturing industries (312120, 327320), one retail trade industry (45291020), and one service industry (712110).
- Annual payroll in one service industry (621111).
- First quarter payroll in one retail trade industry (44821050) and one service industry (712110).
- Sales in two manufacturing industries (312111, 334513), one finance-insurance-real estate industry (524113), two mining industries (211111, 211112), one retail trade industry (447110), four service industries (561720, 621111, 712110, 812331), and one utilities-transportation-warehousing industry (48423020).

It is interesting that:

- Both annual payroll and sales demonstrated high disclosure risk with Metric 1 for one service industry (621111).
- First quarter employment, first quarter payroll, and sales demonstrated high disclosure risk with Metric 1 for one service industry (712119).

Excluding industry-variable pairs that are not applicable, about 99.4 percent of industry-variable pairs did not exhibit high disclosure risk using Metrics 2-4 as criterion; about 88.5 percent of industry-variable pairs did not did not exhibit high disclosure risk using Metric 1 as criteria.

One wholesale trade industry (42393012) exhibited high disclosure risk for sales with Metrics 2, 3, and 4. Otherwise, there were no industry-variable pairs whose Metrics 2-4 values were indicative of high disclosure risk. About 99.4 percent of industry-variable pairs did not exhibit high disclosure risk using Metrics 2-4. This is consistent with results presented in Kim, Drechsler, and Thompson (2020, forthcoming).

Figure 7 presents the median relative L1 error loss by item and industry of the 200 differentially private totals (ε = 0.5, 1, 1.5, and 2), the 254 different combinations of synthetic datasets, and the single combination of 10 synthetic datasets. For presentation, values of relative L1 error loss greater than 3 are excluded.[6] These large values are obtained from differentially private totals; in these cases, the synthetic data are not comparable in terms of privacy protection to the differentially private data. The left panel of Figure 7 presents the results for industries with one imputation cell (industry = imputation cell); the right panel presents the results for industries that contain more than one imputation cell.

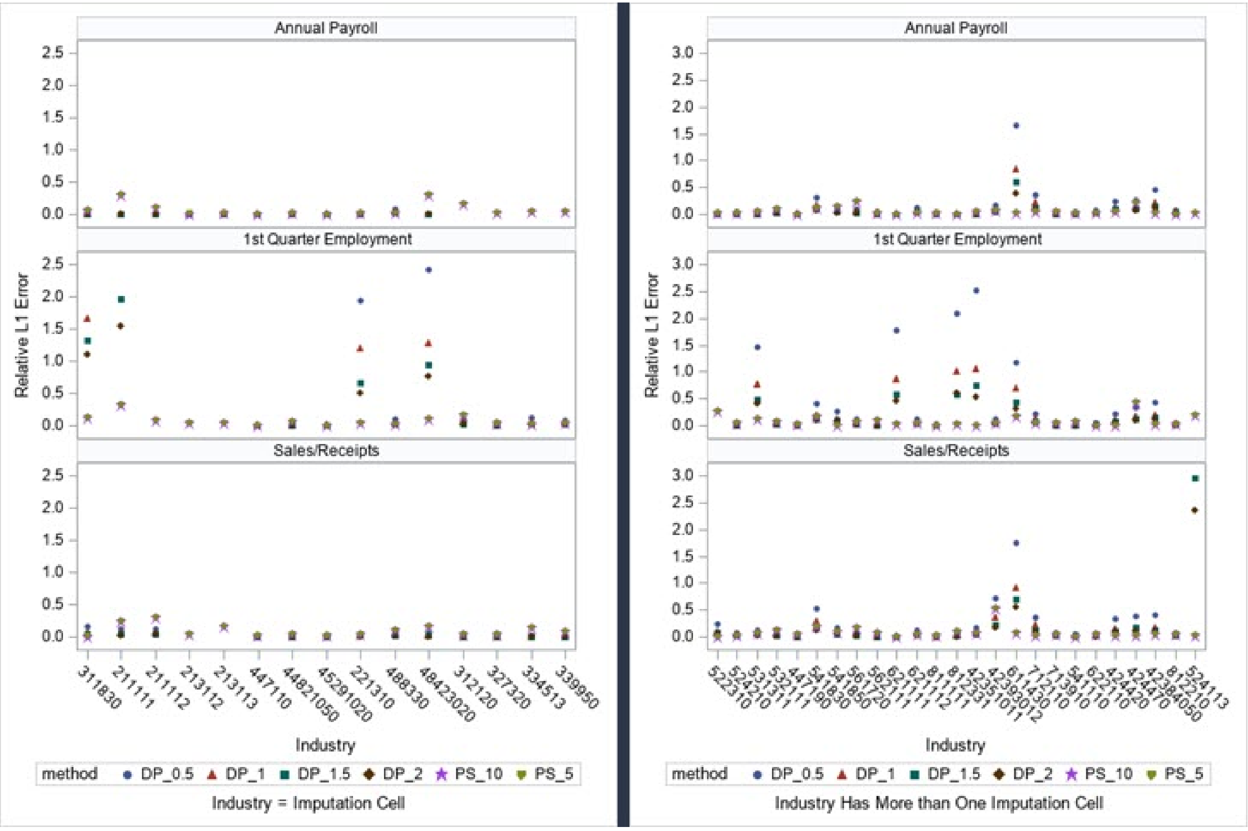


Figure 7: Comparison of median relative L1 error loss for differentially private totals with ε = 0.5 (DP_0.5), ε =1 (DP_1), ε = 1.5 (DP_1.5), ε = 2 (DP_2), and for synthetic totals with 10 and 5 datasets respectively (PS_10, PS_5)

[6] The following industries are excluded from Figure 7:
Annual Payroll: 311830 (DP_0.5), 211111 (DP_0.5, DP_1.0), 211112 (DP_05, DP_1.0, DP_1.5, DP_2.0), 213112 (DP_05, DP_1.0, DP_1.5, DP_2.0), 213113(DP_05, DP_1.0, DP_1.5, DP_2.0), 522310 (DP_05, DP_1.0, DP_1.5, DP_2.0), and 524113 (DP_05, DP_1.0, DP_1.5, DP_2.0)
1st Quarter Employment: 312120 (DP_05, DP_1.0, DP_1.5, DP_2.0), 327320 (DP_05, DP_1.0, DP_1.5, DP_2.0), 334513 (DP_05, DP_1.0, DP_1.5, DP_2.0), 339950 (DP_05, DP_1.0, DP_1.5, DP_2.0), 524113 (DP_05, DP_1.0, DP_1.5, DP_2.0)
Sales/Receipts: 213112 (DP_05, DP_1.0, DP_1.5, DP_2.0), 213113 (DP_05, DP_1.0, DP_1.5, DP_2.0), 524113 (DP_05, DP_1.0)

Table 5 summarizes the comparisons by item across industry. The final row presents counts of industries where the synthetic data Relative L1 error loss is greater than the corresponding differentially private value for all three data items.

Table 5: Number of Industries with Relative L1 Error Loss Greater than Differentially Private Error Loss for Specified Value of ε [Note: Median Relative L1 Error Loss Presented with Five Synthetic Datasets]

Data Item	**Number of Partially Synthetic Datasets**	**Privacy Loss Budget (ε)**			
		0.5	1.0	1.5	2.0
Annual Payroll	5	20	25	26	27
	10	20	25	26	27
1st Quarter Employment	5	12	16	19	22
	10	12	16	19	22
Sales/Receipts	5	19	26	29	30
	10	19	26	29	30
All Variables Combined	5	4	8	11	13
	10	4	8	11	13

For annual payroll and for sales/receipts, the synthetic data has consistently larger relative L1 error than the differentially private counterparts in 25 and 26 industries of the total studied 40 industries, respectively, for all values of $\varepsilon \geq 0.10$. In several other industries, the relative L1 error loss for the ε-1.5 and ε-2 differentially private annual payroll and sales/receipts totals are very comparable to the synthetic data counterparts. However, even with a large privacy budget of $\varepsilon = 2$, there are 13 and 10 industries whose synthetic estimates of total Annual Payroll and total Sales/Receipts would not be adequately protected under this paradigm.

With 1st quarter employment, less than half of the synthetic data totals have larger relative L1 error loss than the differentially private totals for $\varepsilon \leq 1.5$; slightly more than half (22 industries of 40) of the synthetic data totals have larger relative L1 error loss than the differentially private totals for $\varepsilon = 2$. Marginal employment distributions are highly skewed, and larger values in the right hand tail tend to be isolated and unique. Consequently, it is not unreasonable to require more noise to protect the 1st quarter employment totals.

The additive noise for the differentially private totals is implemented independently for each data item within industry/imputation cell. Taken collectively, however, there are only 13 of the 40 studied industries that yield synthetic data totals with similar levels of noise as their ε=2 differentially private counterparts for all studied items. This highlights a deficiency in applying *univariate* privacy protection metrics to a multivariate data set.

Finally, there is very little difference in relative L1 error loss between the PS_5 and PS_10 relative L1 error loss for all variables. Thus, the loss in utility due to using a smaller number of synthetic datasets is minimal. This is consistent with the results presented in Section 4.

5.2.2. Multivariate Metric Results

Within each industry and for each value of alpha (α), we categorized the AIR values as follows:

Risk Category	Definition	Contains
1 (High)	$AIR_i \geq 0.6$	$U_i = 1$ and synthetic data values (all items) within α pattern in at least 6 of 10 synthetic datasets
2 (Medium)	$0.4 \leq AIR_i < 0.6$	$\{U_i = 2$ and $IN_i \geq 8/10\}$ or $\{U_i = 1$ and $4/10 \leq IN_i \leq 5/10\}$
3 (Low)	$AIR_i < 0.4$	All other conditions

Table 6 provides the results for industries that had one or more observations that fell in risk category 1 for a given alpha. Appendix 3 contains the complete table.

Table 6: Industries in Risk Category 1 for a $0.05 \leq \alpha \leq 0.25$

Risk Category	Trade Area	Industry	Alpha (α)				
			0.05	0.10	0.15	0.20	0.25
1	FIR	532111	0	0	0	0	1
1	RET	447110	0	0	0	0	1
1	RET	45291020	0	0	1	1	0
1	SER	541110	0	0	0	0	1
1	SER	712110	0	0	0	0	1

Regardless of industry or level of alpha, the highest proportion of establishments fall into risk category 3. Five industries had observations that fell in the highest risk category, all with values of alpha greater than 0.10; only one of the four industries flagged values with alpha less than 0.25. To understand the extent of the disclosure risk, we inspected the flagged values.

The annual payroll and sales values of the flagged unit in industry 532111 are the industry maxima, and the 1st quarter employment value is greater than the 99th percentile in the industry but not the industry maximum. Furthermore, the wage-per-employee and sales-to-payroll ratios for this unit are very close to the corresponding industry averages. This is a large business whose multivariate distribution is very consistent with the industry distribution.

None of the data item values for the flagged unit in industry 447110 is the industry maximum value. However, all three values are greater than the 99th percentile value in the industry. Consequently, this unit is unique and isolated. With α=25%, the disclosure patterns for Annual Payroll and Receipts are very wide. However, the tolerance range for employment is quite narrow as the industry values of employment are small (95% of the establishments have less than 20 employees).

The 1st quarter employment and sales values of the flagged unit in industry 45291020 are the industry maximum values. However, the annual payroll value is less than the industry median. This is a business with a low (but acceptable) wage-per-employee ratio and a high sales-per-payroll dollar ratio.

The annual payroll value of the flagged unit in industry 541110 is the industry maximum value. The values of 1st quarter employment and sales are above their respective 99th percentiles, but are both well below the industry maxima. This unit has an atypically large wage-per-employee ratios and an atypically low sales-to-payroll ratio; the wage-per-employee ratio for this unit is almost three times as large as the corresponding industry average ratio and the sales-to-payroll ratio for this unit is almost half as large as the corresponding industry average ratio. This business pays its employees extremely well, but is not as profitable as other businesses in the industry.

All three of the data item values of the flagged unit in industry 712110 are their industry maximum value.

In these cases, the choice of a single α to identify sensitive values is difficult, as the items are collected in different units and are on different scales. We attempted to work around this deficiency in the metric by examining a range of α, with the lower values designed to identify either very isolated units or isolated units near the origin. All of the units are definitely isolated in the original data; their uniqueness measure is 1, regardless of alpha. However, the accuracy of the approximation of the synthetic data is questionable, as these values are only identified as potential disclosure risks with large values of alpha.

6. Conclusion/Recommendations/Next Steps

The synthetic data research team was originally convened with the charge of assessing the performance of the fully synthetic data generator proposed in Kim, Reiter, and Karr (2016) and Kim, Karr, and Reiter (2015) on economic census data from a variety of industries. Previously, these methods were vetted on a limited number of industries in a single sector, with no input from economic census data users. There was no discussion of utility requirements for industry totals in the earlier research, and the synthetic data generator did not have this capability.

The team's assessments of the fully synthetic data revealed several deficiencies with the original generator and software. Some of these deficiencies were entirely due to the software. Dr. Kim fixed the software deficiencies during his onsite visits. The team uncovered other important issues in implementation. First, the generator requires at least one input record whose data satisfies all of the provided edits. This is not necessarily reality when using originally reported data, as some items may be "goldplated." Second, the generator does not account for unit nonresponse. Kim, Dreschler, and Thompson (2020, forthcoming) address this by developing a synthetic data generator that produces "populations" from samples; in our setting, nonresponse adjustment weights could serve in place of sampling weights. Finally, consideration must be given to the input data and the input edits:

- Rounding errors should be corrected in dollar values (reported data in \$1 instead of \$1000) to prevent generating unusually large individual records;
- Negative values must be deleted;
- Range edits on individual items should be included to prevent generating unusually small or large individual records. Ratio edits ensure consistency between items in the multivariate synthetic data.

Even with these modifications, the industry totals produced from the fully synthetic data tended to be very different from the corresponding Economic Census tabulations, especially when the input data contained several edit-failing records. The two-step process replaced edit-failing items with entirely different imputation models from the production system used by the Economic Census, often yielding different totals in the input data. These differences were magnified by the synthetic data generation process.

Of course, the scope of the original project did not include a comparison of alternative imputation methods with the production methods. Instead, the synthetic data could be generated from the final edited/imputed data (used for tabulations), without imposing any edit restrictions on the input data (these restrictions are retained on the output synthetic data). Kim, Dreschler, and Thompson (2020, forthcoming) presented promising results on three study industries, producing partially synthetic data by including a measure of size variable (not synthesized) in the original generator and using the final edited/imputed data as input. Overall, the team had similar success with this approach, using a (derived) composite measure of establishment size from the real data. Several team members were concerned that generating synthetic data from the final edited/imputed data might compromise the multivariate associations in the synthetic data. This did not appear to be the case, although the association measure (correlation) was computed from log-transformed data, in part to retain consistency with earlier studies. For most industries, there was little improvement in utility with 10 synthetic datasets over five. However, there may be some privacy advantage in releasing the smaller number of partially synthetic datasets (5).

Determining viable privacy protection metrics proved problematic. In part, this was due to the team's composition; very few members had worked on privacy protection. The Economic Directorate does not have much experience producing synthetic datasets, and the existing synthetic datasets are primarily used for testing software programs. Literature searches and internal discussions suggested univariate metrics – such as the p-percent rule, Metrics 1-4 in Section 5, or relative L1 error. With the exception of Metric 1, these metrics measure the disclosure avoidance risk in tabular data, not micro-data. Furthermore, the synthetic data are multivariate, and the privacy protection metric should evaluate the set of information released with each observation. Consequently, the research team developed its own multivariate assessment metric, which may not be acceptable to the larger data science community.

Data privacy is a very active area of research in data science, and this research frontier is constantly changing. As a result, the rules and standards for disclosure avoidance are in flux. Differential privacy (Dwork et al. 2014) is currently the gold standard for privacy. With highly skewed data, Haney et al (2017) states that "state-of-the-art differentially private algorithms add too much noise for the output to be useful." Moreover, although there has been some very recent research on creating differentially private synthetic data (Snoke and Slavkovic 2018), to date it has not been used to create the highly skewed distributions that we see in the Economic Census data.

This report provides promising results, demonstrating that the recommended synthetic data generator yields data with high utility under the proscribed conditions. Depending on the assumed privacy protection metric, these synthetic data are generally sufficiently noisy to avoid the risk of disclosure. Given these results, logical next steps are to modify the synthetic data generator to account for part-year reporter establishments (births and deaths), to combine the partially synthetic population data of general statistics items with fully synthetic population data of industry products, and to test the synthetic data generator on other industries. At the time of writing this report, there is an ongoing discussion on the privacy protection requirements of establishment counts. If it is determined that establishment counts are protected under Title 26, then subsampling and weighting procedures will need to be implemented as well, before the synthetic data can be released.

References

Dwork, C, and A. Roth (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* **9** (3-4), 211-407.

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, firm turnover, and efficiency: selection on productivity or profitability? *American Economic Review* **98**(1): 394-425.

Haney, S., Machanavajjhala, A., Kutzbach, M., Graham, M., Abowd, J. and Vilhuber, L. (2017). Utility cost of formal privacy for releasing national employer-employee statistics. In ACM SIGMOD.

Kim, H.J., Cox, L.F., Karr, A.F., Reiter, J.P., and Wang, Q. (2015). Simultaneous edit-imputation for continuous microdata. *Journal of the American Statistical Association* **110**, 987–999.

Kim, H.J., Drechsler, J., and Thompson, K.J. (forthcoming in 2020). Synthetic microdata for establishment surveys under informative sampling. *Journal of the Royal Statistical Society Series A: Statistics in Society.*

Kim, H. J., Reiter, J. P., and Karr, A. F. (2018). Simultaneous edit-imputation and disclosure limitation for business establishment data*. Journal of Applied Statistics*.

Kim, H. J., Karr, A. F., and Reiter, J. P. (2015). Statistical disclosure limitation in the presence of edit rules. *Journal of Official Statistics* **31**, 121-138.

Miranda, J. and L. Vilhuber (2014). Looking back on three years of using the synthetic LBD beta. Census Bureau Center for Economic Studies Discussion Paper CES-14-11.

Sigman, R. (1997). Development of a plain vanilla system for editing economic data. UNECE Work Session on Statistical Data Editing. WP24.

Snoke, J. and Slavkovic, A. (2018). pMSE mechanism: Differentially private synthetic data with maximal distributional similarity. In Privacy in Statistical Databases (J. Domingo-Ferrer and F. Montes, eds), pp. 138-159: Springer, Cham.

Vink, G., & van Buuren, S. (2014). Pooling Multiple Imputations when the Sample Happens to be the Population*. arXiv preprint arXiv:1409.8542*.

Wang, Q., Kim. H. J., Reiter, J. P., Cox, L. H., and Karr, A. F. (2016). EditImputeCont: Simultaneous edit-Imputation for continuous microdata. R package version 1.0.1.

Wagner, Dennis L. (2000). Economic Census general editing – Plain Vanilla," *Proceedings of the Second International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, pp. 561-570.

White, T. Kirk, Reiter, J., and Petrin, A. (2018). Imputation in U.S. manufacturing data and implications for within-industry productivity dispersion. *Review of Economics and Statistics 100(3):* 502-509.

Appendix One
Study Industries (6-digit NAICS) and Ratio, Range, and Balance Edits

Trade area	Industries	Items in Ratio Edits	Items in Range Edits	Items in Balance Edits
FIR	522310 524210 531311 532311	Emp1Q AnnPay Pay1Q Sales	Emp1Q AnnPay Sales Pay1Q in [0, ∞)	No balance edits

MAN	311830
	312111
	312120
	327320
	334513
	336612
	339950
		EmpBens
		TotCstM
		Emp1QAVPW
		Emp1QOM
		Emp1Q
		TotHrsM
		TotInvB
		TotInvE
		AnnPay
		AnnPayOM
		AnnPayPW
		Sales
			Emp1Q
			AnnPay
			Sales
			
			All other items in [0, ∞)
				Emp1Q = Emp1QAVPW + Emp1QOM
				AnnPay = AnnPayOM + AnnPayPW

MIN	211111
	211112
	213112
	213113
		EmpBens
		TotCstM
		Emp1QOW
		Emp1QPW
		Emp1Q
		TotHrsM
		AnnPay
		AnnPayOM
		AnnPayPW
		Sales
			Emp1Q
			AnnPay
			Sales
			
			All other items in [0, ∞)
				Emp1Q = Emp1QPW + Emp1QOW
				AnnPay = AnnPayOM + AnnPayPW

RET	447110 447190 44821050 45291020	Emp1Q AnnPay Pay1Q Sales	Emp1Q AnnPay Sales Pay1Q in [0, ∞)	No balance edits
SER	541110 541830 541850 561720 562111 611430 621111 621112 622110 712110 713110 811111 812210 812331	Emp1Q AnnPay Pay1Q Sales Note: Some imputation cells have OpExp	Emp1Q AnnPay Sales Pay1Q and OpExp in [0, ∞)	No balance edits

Appendix One
Study Industries (6-digit NAICS) and Ratio, Range, and Balance Edits

Trade area	Industries	Items in Ratio Edits	Items in Range Edits	Items in Balance Edits
UTL	221310 488330 48423020	Emp1Q AnnPay Pay1Q Sales	Emp1Q AnnPay Sales Pay1Q in [0, ∞)	No balance edits
WHO	424420 424470 42351011 42384050 42393012	Emp1Q AnnPay Pay1Q Sales Note: some imputation cells have OpExp, CstMerch, TotInvE, TotInvB	Emp1Q AnnPay Sales All other items in [0, ∞)	No balance edits

Appendix Two
Privacy Protection Assessment Results: Univariate Metrics 1 through 4 (Complete Results)

		Metric 1: Maximum Value				Metric 2: p-percent (0 collab)				Metric 3: p-percent (1 collab)				Metric 4: p-percent (2 collab)			
		EMP1Q	ANNPAY	PAY1Q	SALES	EMP1Q	ANNPAY	PAY1Q	SALES	EMP1Q	ANNPAY	PAY1Q	SALES	EMP1Q	ANNPAY	PAY1Q	SALES
MAN	311830	-0.18	-0.18	NA	-0.33	12.83	14.58	NA	10.06	12.13	13.88	NA	9.36	11.45	13.27	NA	8.70
	312111	-0.36	-0.53	NA	-0.04	46.47	45.41	NA	46.57	45.92	44.86	NA	45.68	45.39	44.31	NA	44.89
	312120	0.00	-0.07	NA	-0.13	14.14	17.78	NA	7.00	13.39	17.11	NA	6.23	12.75	16.45	NA	5.54
	327320	0.01	-0.23	NA	-0.10	187.99	182.62	NA	182.26	187.10	181.83	NA	181.45	186.24	181.24	NA	180.69
	334513	-0.19	-0.43	NA	-0.02	14.91	13.59	NA	11.04	14.06	13.05	NA	10.48	13.45	12.62	NA	9.98
	336612	-0.32	-0.27	NA	-0.11	28.46	22.68	NA	21.06	27.63	22.03	NA	20.21	26.86	21.46	NA	19.49
	339950	-0.09	-0.48	NA	0.05	22.89	18.97	NA	27.18	22.63	18.73	NA	26.75	22.40	18.51	NA	26.40
FIR	522310	-0.87	0.36	0.08	-0.56	115.32	49.37	39.86	23.31	114.38	48.78	39.39	22.63	113.68	48.26	39.14	21.97
	524113	-0.51	0.08	0.30	0.04	15.36	19.32	23.94	17.00	15.11	18.95	23.28	16.45	14.87	18.65	22.70	16.10
	524210	-0.80	-0.15	-0.33	0.29	235.87	98.28	50.78	133.05	235.12	97.94	50.47	132.13	234.48	97.63	50.25	131.24
	531311	-0.76	-0.34	-0.38	-0.14	75.45	117.49	118.08	30.55	74.82	116.77	117.43	30.28	74.19	116.11	116.78	30.03
	532111	-0.42	-0.14	-0.11	0.08	43.31	36.20	34.71	43.30	42.41	35.35	33.87	42.71	41.62	34.56	33.16	42.16
MIN	211111	-0.93	-0.66	NA	-0.03	12.75	14.68	NA	17.51	12.18	14.03	NA	16.96	11.73	13.43	NA	16.53
	211112	-0.68	-0.66	NA	-0.05	9.95	11.17	NA	15.78	9.27	10.57	NA	15.16	8.65	10.01	NA	14.65
	213112	-0.41	-0.27	NA	0.14	15.68	10.14	NA	8.73	15.43	9.91	NA	8.46	15.20	9.70	NA	8.31
	213113	-0.11	-0.37	NA	-0.10	16.26	9.19	NA	13.99	15.48	8.58	NA	13.45	14.80	8.09	NA	12.91
RET	447110	-0.35	-0.31	-0.12	-0.03	2355.42	1568.22	1893.53	783.68	2354.58	1567.37	1892.64	783.50	2353.77	1566.72	1891.77	783.33
	447190	-0.59	-0.43	-0.50	0.16	392.54	275.84	259.77	381.26	391.73	275.20	259.27	380.59	390.94	274.63	258.78	379.95
	44821050	-0.25	-0.25	0.02	-0.14	154.69	88.56	101.34	129.01	154.15	88.04	100.90	128.12	153.65	87.67	100.48	128.12
	45291020	-0.04	0.08	-0.11	0.12	1546.73	1422.15	1371.37	1377.01	1545.81	1421.27	1370.50	1376.22	1544.93	1420.47	1369.71	1375.45
SER	541110	-0.10	-0.33	-0.61	-0.17	413.93	267.29	242.76	173.02	413.07	266.39	241.87	172.17	412.27	265.53	241.02	171.35
	541830	-0.59	-0.49	-0.44	-0.44	7.35	8.84	8.55	6.67	6.68	7.98	7.66	6.38	6.22	7.54	7.23	6.12
	541850	-1.00	-0.99	-0.99	-0.98	25.17	77.32	100.21	60.13	24.64	76.58	99.30	59.29	24.16	75.93	98.42	58.49
	561720	-0.58	-0.30	-0.20	0.03	37.65	39.07	48.21	63.58	36.84	38.29	47.43	62.96	36.47	37.61	46.79	62.35
	562111	-0.33	-0.34	-0.34	-0.13	164.19	113.88	120.78	169.15	163.47	113.17	119.95	168.43	162.78	112.54	119.30	167.71
	611430	-0.97	-0.97	-0.99	-1.00	190.81	135.02	106.06	116.27	189.98	134.21	105.41	115.58	189.18	133.45	104.86	114.96
	621111	-0.47	0.01	-0.26	-0.04	415.76	297.11	290.19	377.60	414.90	296.18	289.24	376.64	414.08	295.26	288.35	375.71
	621112	-0.30	0.13	0.26	-0.07	59.30	51.51	56.95	52.53	58.72	50.88	56.35	51.99	58.18	50.32	55.77	51.55
	622110	-0.12	-0.14	-0.06	0.07	161.78	126.38	130.35	185.05	161.15	125.66	129.64	184.32	160.59	125.06	129.00	183.67
	712110	0.04	-0.12	-0.02	-0.03	30.27	14.33	17.18	13.19	29.60	13.85	16.67	12.72	28.98	13.48	16.24	12.31
	713910	-0.40	0.05	-0.58	-0.39	271.28	286.37	193.59	93.23	270.53	285.48	192.63	92.90	269.89	284.61	191.71	92.62
	811111	-0.99	-0.99	-0.99	-0.99	425.57	401.54	363.79	615.06	425.10	400.92	363.23	614.29	424.65	400.47	362.80	613.71
	812210	-0.46	-0.09	-0.20	0.26	452.68	296.95	260.04	535.97	451.82	296.33	259.49	535.15	451.10	295.75	258.95	534.38
	812331	-0.29	-0.23	-0.11	-0.03	71.62	69.78	72.93	107.63	70.92	69.16	72.31	106.92	70.25	68.62	71.73	106.26
UTL	221310	-0.45	-0.52	-0.56	-0.42	73.12	42.83	33.09	39.04	72.21	41.90	32.31	38.24	71.31	41.14	31.69	37.45
	488330	-0.39	-0.47	-0.56	-0.32	21.97	21.11	18.95	24.55	21.35	20.38	18.28	23.91	20.78	19.67	17.63	23.27
	48423020	-0.58	-0.56	-0.51	-0.04	19.45	17.02	20.17	30.67	18.83	16.14	19.42	30.14	18.32	15.69	18.96	29.72
WHO	424420	-0.56	-0.21	-0.25	-0.16	85.94	72.30	66.52	48.89	85.12	71.47	65.68	48.21	84.36	70.65	64.94	47.55
	424470	-0.79	-0.55	-0.48	-0.48	95.87	138.35	136.52	87.90	95.39	137.51	134.95	87.08	94.99	136.82	134.95	86.40
	42351011	-0.62	-0.21	-0.24	0.13	98.92	113.44	112.10	46.26	98.43	113.83	111.35	45.61	98.05	113.27	110.67	45.08
	42384050	-0.12	-0.11	-0.97	0.98	45.85	33.16	32.11	15.48	45.11	32.37	31.43	15.13	44.40	31.78	30.80	14.81
	42393012	-0.34	-0.28	-0.35	-0.87	37.06	30.95	28.36	1.71	36.63	30.31	27.72	0.86	36.23	29.81	27.10	0.11

Appendix Three
Privacy Protection Assessment: Multivariate Metric (Complete Results)

Risk	Trade	Industry	Alpha				
Category	Area		5	10	15	20	25
1	FIR	522310	0	0	0	0	0
1	FIR	524113	0	0	0	0	0
1	FIR	524210	0	0	0	0	0
1	FIR	531311	0	0	0	0	0
1	FIR	532111	0	0	0	0	1
1	MAN	311830	0	0	0	0	0
1	MAN	312111	0	0	0	0	0
1	MAN	312120	0	0	0	0	0
1	MAN	327320	0	0	0	0	0
1	MAN	334513	0	0	0	0	0
1	MAN	336612	0	0	0	0	0
1	MAN	339950	0	0	0	0	0
1	MIN	211111	0	0	0	0	0
1	MIN	211112	0	0	0	0	0
1	MIN	213112	0	0	0	0	0
1	MIN	213113	0	0	0	0	0
1	RET	447110	0	0	0	0	1
1	RET	447190	0	0	0	0	0
1	RET	44821050	0	0	0	0	0
1	RET	45291020	0	0	1	1	0
1	SER	541110	0	0	0	0	1
1	SER	541830	0	0	0	0	0
1	SER	541850	0	0	0	0	0
1	SER	561720	0	0	0	0	0
1	SER	562111	0	0	0	0	0
1	SER	611430	0	0	0	0	0
1	SER	621111	0	0	0	0	0
1	SER	621112	0	0	0	0	0
1	SER	622110	0	0	0	0	0
1	SER	712110	0	0	0	0	1
1	SER	713910	0	0	0	0	0
1	SER	811111	0	0	0	0	0
1	SER	812210	0	0	0	0	0
1	SER	812331	0	0	0	0	0
1	UTL	221310	0	0	0	0	0
1	UTL	488330	0	0	0	0	0
1	UTL	48423020	0	0	0	0	0
1	WHO	424420	0	0	0	0	0
1	WHO	424470	0	0	0	0	0
1	WHO	42351011	0	0	0	0	0
1	WHO	42384050	0	0	0	0	0
1	WHO	42393012	0	0	0	0	0

Appendix Three

Privacy Protection Assessment: Multivariate Metric (Complete Results)

Risk	Trade	Industry	Alpha				
Category	Area		5	10	15	20	25
2	FIR	522310	0	0	0	0	0
2	FIR	524113	0	0	0	0	0
2	FIR	524210	0	0	0	0	2
2	FIR	531311	0	0	0	1	1
2	FIR	532111	0	0	0	1	0
2	MAN	311830	0	0	0	0	0
2	MAN	312111	0	0	0	0	0
2	MAN	312120	0	0	0	0	1
2	MAN	327320	0	0	0	0	0
2	MAN	334513	0	0	0	0	0
2	MAN	336612	0	0	0	0	0
2	MAN	339950	0	0	0	0	1
2	MIN	211111	0	0	0	0	0
2	MIN	211112	0	0	0	0	0
2	MIN	213112	0	0	0	0	0
2	MIN	213113	0	0	0	0	0
2	RET	447110	0	0	0	1	0
2	RET	447190	0	0	0	0	0
2	RET	44821050	0	0	0	0	1
2	RET	45291020	0	0	0	1	1
2	SER	541110	0	0	0	0	0
2	SER	541830	0	0	0	0	0
2	SER	541850	0	0	0	0	0
2	SER	561720	0	0	0	0	0
2	SER	562111	0	0	0	0	0
2	SER	611430	0	0	0	0	0
2	SER	621111	0	0	0	0	1
2	SER	621112	0	0	0	0	1
2	SER	622110	0	0	0	1	1
2	SER	712110	0	0	0	0	1
2	SER	713910	0	0	0	0	0
2	SER	811111	0	0	0	0	0
2	SER	812210	0	0	0	0	1
2	SER	812331	0	0	0	0	0
2	UTL	221310	0	0	2	0	0
2	UTL	488330	0	0	0	0	0
2	UTL	48423020	0	0	0	0	0
2	WHO	424420	0	0	0	0	0
2	WHO	424470	0	0	0	1	1
2	WHO	42351011	0	0	0	0	0
2	WHO	42384050	0	0	0	0	0
2	WHO	42393012	0	0	0	0	0

Appendix Three
Privacy Protection Assessment: Multivariate Metric (Complete Results)

Risk	Trade	Industry	Alpha				
Category	Area		5	10	15	20	25
3	FIR	522310	4211	4211	4211	4211	4211
3	FIR	524113	6586	6586	6586	6586	6586
3	FIR	524210	25796	25796	25796	25796	25794
3	FIR	531311	10682	10682	10682	10681	10681
3	FIR	532111	6448	6448	6448	6447	6447
3	MAN	311830	125	125	125	125	125
3	MAN	312111	275	275	275	275	275
3	MAN	312120	173	173	173	173	172
3	MAN	327320	3260	3260	3260	3260	3260
3	MAN	334513	331	331	331	331	331
3	MAN	336612	334	334	334	334	334
3	MAN	339950	1809	1809	1809	1809	1808
3	MIN	211111	3304	3304	3304	3304	3304
3	MIN	211112	178	178	178	178	178
3	MIN	213112	4237	4237	4237	4237	4237
3	MIN	213113	140	140	140	140	140
3	RET	447110	48894	48894	48894	48893	48893
3	RET	447190	6923	6923	6923	6923	6923
3	RET	44821050	4884	4884	4884	4884	4883
3	RET	45291020	3732	3732	3731	3730	3731
3	SER	541110	25002	25002	25002	25002	25001
3	SER	541830	343	343	343	343	343
3	SER	541850	1676	1676	1676	1676	1676
3	SER	561720	6871	6871	6871	6871	6871
3	SER	562111	3856	3856	3856	3856	3856
3	SER	611430	5274	5274	5274	5274	5274
3	SER	621111	49570	49570	49570	49570	49569
3	SER	621112	2372	2372	2372	2372	2371
3	SER	622110	3973	3973	3973	3972	3972
3	SER	712110	1299	1299	1299	1299	1297
3	SER	713910	3174	3174	3174	3174	3174
3	SER	811111	43670	43670	43670	43670	43670
3	SER	812210	6344	6344	6344	6344	6343
3	SER	812331	788	788	788	788	788
3	UTL	221310	2803	2803	2801	2803	2803
3	UTL	488330	515	515	515	515	515
3	UTL	48423020	589	589	589	589	589
3	WHO	424420	2345	2345	2345	2345	2345
3	WHO	424470	4320	4320	4320	4319	4319
3	WHO	42351011	4825	4825	4825	4825	4825
3	WHO	42384050	841	841	841	841	841
3	WHO	42393012	853	853	853	853	853