USC Dornsife
Center for Economic and Social Research

USC Schaeffer
Leonard D. Schaeffer Center for Health Policy & Economics

LASI Synthetic Data Report

Joshua Snoke, Erik Meijer, Drystan Phillips, Jenny Wilkens, Jinkook Lee

Paper No: 2023-009

CESR-SCHAEFFER WORKING PAPER SERIES

cesr.usc.edu **healthpolicy.usc.edu**

Synthesizing Surveys with Multiple Units of Observation: An Application to the Longitudinal Aging Study in India

Joshua Snoke [1], Erik Meijer[2], Drystan Phillips[2], Jenny Wilkens[2], and Jinkook Lee[2]

[1] *RAND Corporation, jsnoke@rand.org*

[2] *University of Southern Califonia, erik.meijer@usc.edu, drystanp@usc.edu, jwilkens@usc.edu, jinkookl@usc.edu*

Keywords— synthetic data, survey data, multiple units of observation, disclosure risk, aging studies

Abstract: We present an application of novel methodology to create a publicly-releasable synthetic version of the Longitudinal Aging Study in India (LASI). The LASI, a health and retirement survey, is used for research and educational purposes, but it cannot be shared due to privacy considerations. We present new methods to synthesize the survey, maintaining its three levels of observation, both continuous and categorical data, and complex survey weights. We show that the synthetic data maintains the distributional patterns of the confidential data and largely mitigates identification and attribute disclosure risk. We also present a novel method for controlling the risk and utility tradeoff of the synthetic data, motivated by the specific disclosure risks presented by the LASI collection methods. We introduce a method to down-weight records that have a high likelihood of being uniquely identifiable in the population due to the released demographic information and the sampling rates. We show this approach reduces both identification and attribute risk for records while preserving better utility over another common approach of coarsening records. Our methods and evaluations provide a foundation for creating a synthetic version of the LASI which can be shared publicly and increase research access.

Statement of Significance: We present novel methodology for synthesizing survey data that are collected with multiple units of observation. Prior synthetic data methods were insufficient due to restrictions on the units of observations or the types of data. Our approach can utilize any common method of sequential synthesis, but we simultaneously synthesize individuals, couples, and households by first restructuring the data and synthesizing structural variables. Our methods can be easily implemented using standard public software packages, so others can easily apply our methods to their

own survey data. Additionally, we provide new means of mitigating disclosure risk that are specific to the types of risk inherent in synthetic data, which is an area that currently has little research.

1 Introduction

Researchers and data maintainers who collect survey data increasingly seek to share the data for purposes of education, training, reproducibility, and enabling secondary analyses. Prior to sharing survey data, data maintainers commonly apply statistical disclosure control (SDC) methods to confidential data (Hundepool et al., 2012). Approaches include data reduction methods such as suppressing certain variables, records, or cells (for tabular data) or coarsening variables by collapsing categories. Other methods perturb individual records to protect privacy, such as data swapping which randomly switches values between records while preserving the marginal distributions, bottom- or top-coding that replaces outliers with selected minimum or maximum values, and random noise addition that adds noise to variables from known distributions. These methods attempt to reduce the risk to breaches of confidentiality in order to share microdata, tables, or summary statistics derived from surveys that otherwise could only be accessed through restricted means.

Synthetic data was proposed as an alternative method to traditional SDC approaches, with formative work by Rubin (1993), Little (1993), and Raghunathan et al. (2003). It differs from other SDC approaches in its conceptual basis. Rather than starting from the entire confidential sample and attempting to maintain as much of the original data values as possible, synthetic data starts with an assumed data generating process, estimates sample parameters based on the confidential data, and draws fully new records based on a model using these sample parameters. In this way, a model and a data generating process are essential to synthetic data in a way that they are not for other statistical disclosure control methods.

Synthetic data have been commonly applied to survey data in a variety of settings. For example, Benedetto et al. (2018) created a synthetic version of the Survey of Income and Program Participation for the U.S. Census, Hu et al. (2018) applied synthesis methods on the American Community Survey, Drechsler and Reiter (2009) and Kim et al. (2021) synthesized establishment surveys, Hu et al. (2021) evaluated synthesis methods using the Consumer Expenditure Survey from the Bureau of Labor Statistics, and Drechsler and Reiter (2012) offered general methods for

synthesizing large surveys. Less work has considered the problem of synthesizing structured surveys, for example those that contain individuals within couples or households. Prior work such as Manrique-Vallier and Hu (2018) and Hu et al. (2018) proposed latent class models to handle hierarchical data or structural zeros, but these methods have only been applied to categorical data. Alternatively, Benedetto and Totty (2023) proposed to create structure in synthetic data sets by linking records, e.g., couples, after synthesizing individual characteristics.

In this work, we propose a novel method for synthesizing multiple levels of structure, and we do so while using common synthesis models that are computationally efficient. We show that high utility and low disclosure risk can be maintained simply by re-imagining the joint distribution of interest.

We organize the remainder of this paper as follows: section 2 describes the Longitudinal Aging Study in India, section 3 provides an overview of synthesis models for data with household or similar structures, section 4 discusses downweighting records for additional disclosure control, section 5 details our synthesis model for the LASI data set, section 6 gives the utility and risk measures that we use to evaluate our synthesis, section 7 provides empirical results, and section 8 concludes with a discussion.

2 Gateway to Global Aging Data and the Longitudinal Aging Study in India

The Longitudinal Aging Study in India (LASI) is a multidisciplinary survey of individuals age 45 and over and their spouses of any age in India (Bloom et al., 2021; Perianayagam et al., 2022). Wave 1 was mostly conducted between 2017 and 2019, and Wave 2 is in preparation as of February 2024. LASI was designed to be nationally representative, as well as representative of each of the 36 states and union territories in India. Therefore, the sampling was stratified by state and urbanicity (i.e., urban or rural) within state, with a three-stage clustered sampling design in rural areas and a four-stage clustered sampling design in urban areas. Sampling weights are provided that reflect the sampling design as well as differential response rates (although the latter are high compared to most other surveys, varying between 77% and 96% across states and territories). Once a household had been included into the sample, all household members age 45 and over and their spouses were asked to participate. Thus, the number of respondents per household varies across households, and is often higher than in comparable studies in the U.S. and other high-income countries.

As part of collecting the LASI, respondents are ensured that their privacy and confidentiality will be

maintained throughout any downstream publications or data sharing. This is vital both to preserve the ethical right to privacy and also to ensure the viability of future data collections. Respondents might be concerned about re-identification, such that someone could ascertain that they participated in the survey. Along with this, they may be concerned about disclosure of certain sensitive attributes which may result from re-identification. In other cases, attributes may be learned by inference, such as the income of the wealthiest person in a small town.

With a very large underlying population, such as India's, it may be tempting to think that the risk of such disclosures is minimal. But given enough information, particularly geographically identifying information in low population regions, studies have shown that disclosure is quite possible (e.g., Rocher et al. (2019)). Additionally, the LASI relied on substantial oversampling of places with small populations in order to allow geographically representative analyses. While in some large states, the sampling fraction is very small, in smaller states it is larger. Additionally, even with small sampling fractions very unusual records can lead to disclosures.

2.1 Data Access Through Synthetic Data

Using synthetic data opens the door to a few potential means for increasing data access. While there are many potential uses for synthetic data, we have envisioned three major avenues for the use of synthetic Harmonized LASI data at this time.

1. User trainings: The Gateway conducts quarterly user trainings to teach students or researchers how to utilize harmonized data for their analyses. Rather than requiring participants to pre-register for a study, which is usually met with limited success, the utilization of synthetic data would limit the need for pre-registration.
2. Exploratory research: Researchers could conduct exploratory research on synthetic data made easily available on the Gateway website and determine whether the data would suit their research plan. If they find it suitable, they could then apply for the Harmonized LASI and use it in place of the synthetic data for their analyses.
3. Restricted data research: We could use synthetic data to protect respondent anonymity when using restricted data. For example the LASI may be linked to data on pollution or severe weather, which are very closely tied to specific geographic locations, potentially making

respondents identifiable, especially when historic records and residential history are available. The use of synthetic data in these cases would allow researchers to conduct their analyses in full detail, while protecting respondent anonymity.

We synthesize the variables listed in Table 1. The file contains both household (HH) and individual level characteristics, as well as couple identifiers indicating spousal respondents. There is one auxiliary variable, Head of HH gender, which was not part of the original Harmonized LASI file, but was derived to be used in the synthesis process to re-weight the data after synthesis to match the original weight totals. It will not be released in the synthetic data.

Measurement Unit	**Characteristic**	**Description**
Household	HH ID	42,311 households
Household	State	35 states/territories
Household	Rural Indicator	Binary
Household	HH # Residents	Continuous
Household	HH # Respondents	Continuous
Household	HH survey weight	Continuous
Household	Head of HH Gender (not released)	Binary
Couple	Couple ID	1-4 couples per HH
Individual	Individual ID	1-8 individuals per HH
Individual	Gender	Binary
Individual	Age	Binned into 9 categories
Individual	Education years	Continuous
Individual	Education category	10 categories
Individual	Self-reported health	Ordered 1-5
Individual	Activities of daily living	Order 0-5
Individual	Working status	Binary
Individual	Individual earnings	Continuous
Individual	Individual survey weight	Continuous

Table 1: Variables Included in the Synthesis Process and Evaluations. ID values are arbitrarily labeled in the synthetic data, not based on confidential IDs.

3 Synthesizing Multi-level Data Structures with Sequential Synthesis

The LASI survey contains three levels of observation: household, couples within households, and individuals within couples. For the synthetic data, we want to maintain both the household and couple information in the synthetic data and ensure the relationships between individuals in the same households or couples are maintained. To do this we propose a novel approach which reshapes the data in particular ways and adds some structural variables to the synthesis process, so we can use

common conditional synthesis approaches.

Prior work on synthesizing household structure is limited. Hu et al. (2018) proposed using hierarchical Bayesian models to draw individual and household characteristics from a Dirichlet process mixture of products of multinomials. The model assumes individuals and households are each members of nested latent classes, such that the relationships between the two can be modeled. Their model does not accommodate continuous variables or sampling weights, which the LASI includes. Additionally, these models are computationally intensive and require defining complex joint distributions for the variables in the data. Benedetto and Totty (2023) presented a method where couples are matched after synthesizing individual characteristics. For each synthesized individual who should be coupled, they greedily draw candidate matches from the pool of possible matches using distance measures, until all individuals are matched. The structure in our data set, which contains both households and couples, is more complex than in either of those prior works, and we do not see a straightforward means of applying those methods to our data.

Given the limitations of these prior methods for our application, we design a simpler but intuitive approach to maintaining the structure in the data in our synthesis model. Rather than considering multiple observation levels, we flatten the data to a single level, the household level, and we explicitly incorporate couple and individual level information. We want to synthesize the data using a sequence of conditional models, as is common in the field, e.g., using a sequence of regressions (Nowok et al., 2016; Raghunathan et al., 2001) or CART models (Drechsler and Reiter, 2011; Reiter, 2005).

To understand the new data structure, let N be the number of households in our data, and let there be M measured household characteristics, Q measured individual characteristics, and P possible couples[1] in each household. We define:

1. $Y_{ik} : i = 1, ..., N; k = 1, ..., M$ as the household characteristics.
2. $X_{ijlr} : i = 1, ..., N; j = 1, ..., Q; l = 1, ..., P; r = 1, 2$ as the individual characteristics.
3. $Z_{ilr} : i = 1, ..., N; j = 1, ..., Q; r = 1, 2$ as binary flags indicating the existence of person r in couple l in household i.

[1]Note that many "couples" will only have one individual who is without a partner.

The last set of variables define the structure of the households. First, the number of individual respondents that exist in a household is restricted by the total respondents, which is a household feature. Second, couples must be filled in order, such that couple 3 can only exist if couple 2 exists, and so on. Third, the second person in a couple can only exist if the first person in the couple exists.

In other words, if there are h_i total respondents then by definition $Z_{ilr} = 0$ for all $l > h_i$ or $(r * l)/2 >= h_i$. By definition if $Z_{ilr} = 0$, then X_{ijlr} is set to *missing* for all j. While N, M, and Q are given based on confidential data, the selection of P is less straightforward. We discuss the selection of P as a synthesis tuning parameter in Section 7.1.

Because certain numbers of total household respondents allow for different arrangements of these individuals into couples, we provide an example illustration of the possible person combinations that exist with a maximum of three respondents in a household in Table 2.

	Age				
	Couple 1		**Couple 2**		**Couple 3**
# HH Respondents	**Person 1**	**Person 2**	**Person 1**	**Person 2**	**Person 1**
1	50	-	-	-	–
2	45	47	-	-	–
2	46	-	65	-	–
3	47	45	66	-	–
3	70	-	48	46	–
3	70	-	65	-	47

Table 2: Notional Table of All Possible HH Person Patterns with Age as an Example Variable

In order to correctly synthesize only the individual characteristics for individuals who exist in the households, we must carefully construct the synthesis order, i.e., the order of conditional models for synthesizing each variable. We arrange the conditional models as follows[2]:

1. *Household characteristics:* for every k, fit $f(Y_{ik}|Y_{i1}, Y_{i2}, ..., Y_{i(k-1)})$
2. *First individual in the household:* for every j and $l; r = 1$, fit $f(X_{ij11}|X_{i111}, X_{i211}, ..., X_{i(j-1)11}, Y)$
3. *Remaining individuals:* starting with $l = 1; r = 2$, for every l and $r \in \{1, 2\}$:
   (a) Fit $f(Z_{ilr}|Z_{i11}, Z_{i12}, Z_{i21}, ..., Z_{i(lr-1)}, Y)$[3]

[2]Model parameters not included to simplify notation.
[3]We abuse the notation here slightly, since the last prior person could be either $l - 1$ or $r - 1$.

(b) For every j, fit

$$f(X_{ijlr}|X_{i1lr}, X_{i2lr}, ..., X_{i(j-1)lr}, X_{ij11}, X_{ij12}, X_{ij21}, ..., X_{ij(lr-1)}, Y, Z_{ilr} = 1)$$

Once we fit the models, we generate synthetic household records, and only draw synthetic X_{ijlr} values if $Z_{ilr} = 1$. Because we condition the Z models on the number of respondents as a household characteristic and the prior Z variables, we ensure that each household has the correct number of individuals with synthesized characteristics.

Given this approach, we can apply standard synthesis models and software (Nowok et al., 2016) to perform conditional sequential synthesis. After completing the synthesis we convert the structure back to the structure of the confidential individual level file to produce information in a familiar format for users. Code to replicate our synthesis approach is available at `github.com/XXX`.

4 Down-weighting Records Using Survey Weights

A few states and territories in the data have very small populations, but they were significantly over-sampled in order to create a sample of every Indian state and territory with a minimum number of records to perform state and territory specific analyses. For these locations, the disclosure risk from identifying synthetic data records similar to the confidential records is higher because there are fewer other potential candidate households from these states. We consider two approaches to mitigate this risk.

First, we consider grouping small states with larger neighboring states, i.e., coarsening. This straightforward approach removes any unique relationships between characteristics in the data that exist within the smaller states and instead draws values based on the average relationship across the combined smaller and larger states. This decreases the expected utility of both the smaller and larger states while decreasing the expected identifiability of records in the over-sampled states.

The second, more novel, approach uses an estimate of identification risk to down-weight records in the synthesis process for households that are at higher risk. Hu et al. (2021) proposed a similar approach that down-weighted risky records based on their contribution to the likelihood in a Bayesian model. To adapt this concept to our models, we use the estimated risk of being unique in the population (detailed in Section 6.2) as case-weights in the CART models implemented in the

rpart package (Breiman et al., 1984; Therneau et al., 2015) in **R**. We describe the specific way we define weights based on the risk values in Section 7.4 when we show evaluations of the method.

The models are estimated over groups of states, so down-weighting records will have the effect of capturing relationships in small states that are similar to those in larger states but reducing the ability of the model to learn unique relationships in smaller states and reflect these in the synthetic data. It should not have a substantial impact on larger states. Our empirical results given in Section 7 show utility and risk results for both an unweighted and weighted synthesis, as well as synthesis using coarsened states and territories.

Down-weighting records also makes sense for synthetic data, particularly since the membership inference attack (Stadler et al., 2022) has become a more common measure of disclosure risk for synthetic data. We choose not to evaluate this metric for our paper because we believe it is more likely for an attacker in our context to use a linkage attack, but we note that by definition down-weighting records directly reduces the type of risk model by the membership inference attack.

5 Synthesizing the LASI

In the prior sections, we described the general format for synthesizing structured survey data. Here we detail some of the specific decisions we made regarding the functional forms for the synthesis models and the tuning of those models.

5.1 Synthesis Model and Predictor Selection

We use flexible non-parametric synthesis models using the CART algorithm first proposed for synthesis by Reiter (2005) and implemented in the synthpop package in **R** (Nowok et al. (2016),Nowok et al. (2016)). All selected variables are synthesized except for State and Rural indicator. By not synthesizing state and rural/urban indicators, we maintain the same number of households by state and rural/urban areas in the synthetic data as are in the confidential data. These two geographic variables formed the core of the sampling frame of the confidential survey data, so we keep the size of the sample and strata fixed.

We exclude household survey weight as a predictor of the individual variables because we found that resulted in overfitting the relationship between the household weight and individual characteristics.

We also only include individual characteristics of previous individuals as predictors if it is the same characteristic. For example, when synthesizing the third respondent in the household's age, we include the first and second respondents' ages as predictors, but we do not include any other characteristics of the first and second people. This simplifies the number of features in each model, which we found led to overfitting. Finally, individual survey weights are synthesized conditional on household survey weights, previous persons' individual survey weights, and the corresponding individual's other characteristics only.

5.2 Additional Tuning of the Synthesis Model

We made a few practical decisions when it came to tuning the CART models. Some of these were made for computational purposes while others are possible parameters to control the trade-off of utility and risk. First, due to the large number of unique states in the data, we stratified our synthesis into four subgroups of states based on regions (North, Central, East, South). This significantly increased computational speeds, and due to the inclusion of state indicators in our predictors it should not have adverse effects on the accuracy of the models. In other words, we still model state-level variation within each region. The potential drawback is that by fitting smaller models, we lose some potential privacy protection from smoothing over a large set of households across a larger geographic area. We may also lose statistical precision for modeling the population when using subsets, but each region still represents a large, diverse sample of states and households.

Second, we binned age into 5-year bins (with open-ended bins below 45 and above 80) in order to provide additional protection. We chose this rather than smoothing approaches for continuous values drawn from CART models, such as those implemented by Bowen et al. (2022). Lastly, we tested a few different values for the CART model minimum bucket size parameter. This controls how many observations each terminal node must have in order to make an additional split. We use a minimum bucket size of 15. All other tuning parameters were left at their defaults in the *synthpop* package.

5.3 Re-calibrating Synthetic Survey Weights

After generating the synthetic data, we perform a final re-calibration of the individual and household survey weights. We do this so that the synthetic data have the same representativeness as the confidential data. Detailed information on the generation of LASI individual and household weights

can be found in Chien et al. (2023). We use the same raking and trimming approach, and we use the household and individual weights generated from the synthesis process as our starting weights in order to capture the original variation in the confidential starting base weights that is modeled by the synthesis process.

The household weights are re-calibrated to the household population for one factor within each state:

1. Head of household gender × Rural indicator

The individual weights are re-calibrated to the individual population for three factors:

1. Gender × Age
2. Rural indicator
3. Gender × Education

The resulting weights provide representativeness on the weighted characteristics both within states and nationally across all states. We use the weight distributions in the confidential data to estimate the population distributions across the raking factors.

6 Measures of Statistical Utility and Disclosure Risk

6.1 Utility Evaluation Metrics

The earliest papers proposed releasing synthetic data in place of the confidential data (Rubin, 1993), and they provided rules for valid inference such that analysts could fit valid models without ever accessing the confidential data (Raghunathan et al., 2003; Reiter, 2003; Reiter and Raghunathan, 2007). But inferences from synthetic data are only valid if the models that are used to synthesize the data correspond to the true underlying data generating process (Raab et al., 2016). We do not cover valid inference methods in this report, since that is not the goal for the synthetic data we create. But Drechsler (2011) and Raab et al. (2016) provide extended details on this topic.

Synthetic data are commonly used in two other ways, such as are described in Snoke et al. (2018) and Arnold and Neunhoeffer (2020). First, synthetic data may be used to explore specific models without drawing inference directly from the synthetic data. After testing models using the synthetic data, researchers will confirm results using restricted access data or verification servers (Reiter et al., 2009).

In this case, the goal of those using the synthetic data is not to make inference with respect to the population target, Q, but rather to make inference with respect to the confidential sample value, $\hat{Q}$. This is commonly known as *specific utility*, and is measured through metrics such as the confidence interval overlap (Karr et al., 2006; Snoke et al., 2018).

Second, synthetic data may be used as a general purpose tool to understand the univariate and multivariate distributions in the data. Students, developers, and methodologists may particularly benefit from this kind of synthetic data, since realistic test data sets are often hard to find, particularly in the social sciences. We measure this kind of value of the synthetic data using distributional distance measures, and it is commonly referred to as *general utility*.

Our utility assessment focuses on general utility, and we evaluate the utility of the synthetic data we generate using summary statistics, visual depictions of the distributions, and a distributional distance measure, the *pMSE ratio* (Snoke et al., 2018). The *pMSE ratio* is a metric developed by Snoke et al. (2018) that extended the *pMSE*, originally proposed (without a name) by Woo et al. (2009). The *pMSE* is a distributional distance measure that computes the mean-squared error from predicted probabilities of records belonging to the synthetic data versus the confidential data. Since the proposal of the *pMSE*, a broad class of general utility distance measures has been developed based on discriminate models between the confidential and synthetic data, such as those described in Bowen et al. (2021) and Raab et al. (2021). Raab et al. (2021) show that the *pMSE ratio* is both among the most powerful and the most versatile statistics in this class of metrics.

We obtain the *pMSE ratio* by dividing the observed pMSE by its expectation under the null that the synthetic data are drawn from the same true data generating process for the confidential sample. A ratio of 1 implies the synthetic data are distributionally as close as a new sample from the same underlying population, and larger values imply worse synthesis. Raab et al. (2021) suggest anything below 10 represents acceptable synthesis as a rule of thumb.

We compute distributional comparisons both for the entire data set and for all 1-, 2-, and 3-way combinations of characteristics. When computing the *pMSE ratio* for 1-, 2-, or 3- dimensional variable comparisons, we use fully saturated logistic models, which is recommended for low-dimensional distributional comparisons (Raab et al., 2021). When we compare the distribution of

the entire data set, we use CART models to discriminate between synthetic and confidential data. We utilize two adaptations to compute the *pMSE ratio* recommended by Bowen and Snoke (2021): first, we use a training-test split to choose the optimal CART model with the highest AUC and second we estimate the null by resampling the confidential data only rather than the confidential and synthetic data combined.

6.2 Disclosure Risk Metrics

We evaluate two types of disclosure risk: probability of identification and attribute disclosure. To measure identification disclosure risk we utilize matching on quasi-identifiers (QIs), such as is proposed in Reiter and Mitra (2009). Because we are synthesizing survey data, we do not assume the attacker already knows whether households participated, so we also consider the likelihood that a match is unique in the population as part of the risk estimate.

To estimate population uniqueness, we follow prior work using an adapted form of the log-linear risk estimation from Skinner and Shlomo (2008). We extend their work by using a penalized lasso regression, and we select the optimal penalization parameter based on the model selection criteria laid out in Skinner and Shlomo (2008). This approach allows us to estimate the expected number of households with the same QIs in the population based on the distribution of households in the survey and the sampling rate. Specifically, we estimate the risk value:

$$r_i = E[1/F_k | f_k = 1, QI_i], \forall \;\; k : f_k = 1$$

where F_k and f_k are the number of households in the population and sample respectively sharing the same characteristics.

For both the linkage attack and the log-linear risk estimates, we assume that the attacker seeking to identify records knows only certain QIs from the data. Specifically, we assume they know the state, rural vs. urban designation, number of household residents, the ages, gender, educational attainment, and employment status of the first two respondents in the household[4]. After computing r estimates

[4]The log-linear models were too sparse when we included information on more respondents, which leads to inaccurate results. Also, the likelihood of a match in the synthetic data decreases with additional QIs, so we assume the first two respondents is the optimal amount of information that the attacker may choose to utilize.

for each sample unique record, we link records with the synthetic data and estimate how many records have unique synthetic data matches, commonly referred to as unique-uniques. Finally, we estimate the number of unique-uniques by their r value, since records with low r values are unlikely to be identified in the population even if they are uniquely matched in the sample.

Given the ability to identify a household, an attacker may wish to learn about particular sensitive attributes. We estimate the risk of learning information about the total household income among households with high identification risk. We compute two measures of attribute risk. First, we compute the number of records where the observed synthetic value (e.g., income) is within a certain percentage of the confidential value:

Definition 6.1. Define the a_1 attribute risk metric as:

$$a_1(b, R) = \Sigma_i \ \ I(r_i > R) I(x_i^C b > x_i^S > \frac{x_i^C}{b})$$

where b is the percentage range around the confidential value, x_i^C are the confidential values, x_i^S are the synthetic values, and R is the threshold value that determines the set of risky records for which we compute this metric.

Second, we utilize similar methods to those recommended by Reiter, Wang, and Zhang (Reiter et al.) to estimate a distribution over the synthetic values rather than simply taking the value in the data. In this case, an attacker models the relationship between income, for example, and household characteristics to estimate a posterior distribution of income values for a record. We then compute the risk metric for the mean probability that the confidential income values are within a certain range given this distribution. That is,

Definition 6.2. Define the a_2 attribute risk metric as:

$$a_2(b, R) = \frac{\Sigma_i \ \ I(r_i > R) p(x_i^C b > \hat{x}_i^S > \frac{x_i^C}{b} | S)}{\Sigma_i \ \ I(r_i > R)}$$

where S is the synthetic data and $\hat{x}_i^S$ is the estimated value for household i.

7 Empirical Results

We provide detailed utility evaluations of synthetic data generated using the methods described in the previous sections. First, we perform a general utility evaluation to select the maximum number of household respondents allowed in a synthetic record, as discussed in Section 3. After this we provide an extensive utility analysis on a number of different distributional measures. Finally, we evaluate the risk-utility tradeoff for using downweight or grouping states to reduce risk as discussed in Section 4.

7.1 Selecting the Maximum Number of Respondents in a Household

A potential drawback of our proposed synthesis model is that we must choose a priori a fixed maximum number of respondents per household. The synthesis models for households with many respondents will only be based on a very small number of households, so we expect to see a bias-variance tradeoff between setting the maximum number of respondents too low (i.e., truncating the data) versus too high (i.e., sampling from noisy models).

In order to set the maximum number for our remaining evaluations, we test the general utility for syntheses with varying numbers of allowed household respondents, varying from two to 12, which is the observed maximum in the confidential data. We generate 30 replicates of synthetic data sets using each maximum allowed number, and we measure the general distributional similarity to the confidential data. We use the general *pMSE ratio* statistic described in Section 6, since we are interested in the overall distributional similarity of the household-level synthetic data using different maximum numbers of respondents. Figure 1 shows the results.

We see that as expected, the utility increases (smaller values) as the number of respondents per household increases, but that the utility plateaus around 7 or 8 respondents. We do not see any gains after that, so it makes sense to set the maximum lower than the observed maximum number of household respondents in the data. This is likely because there are very few households with more than 7 or 8 respondents. For the remaining evaluations, we set the maximum number of respondents at 8 for all syntheses.

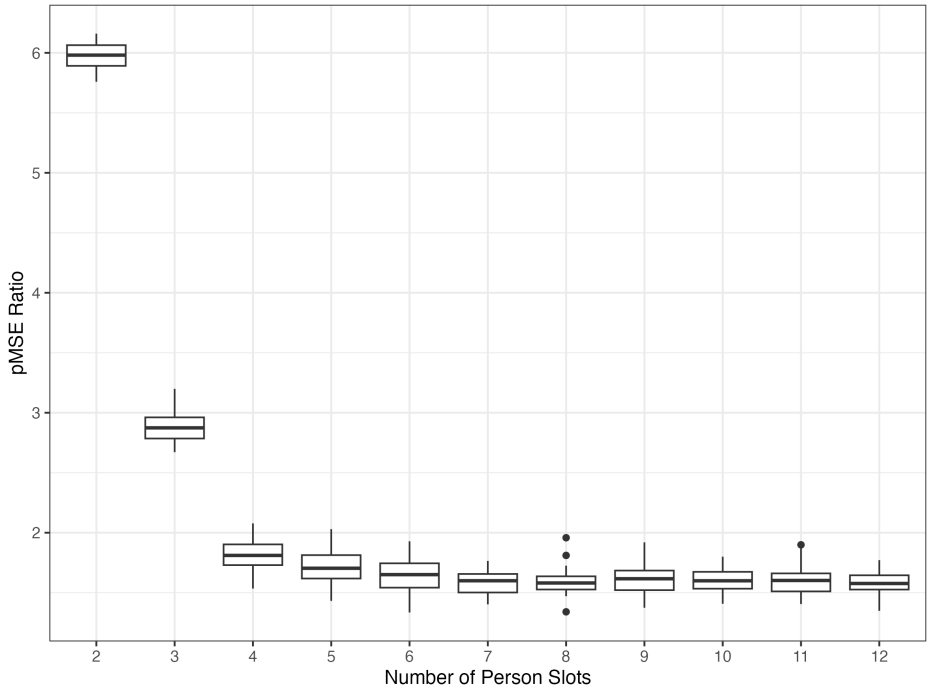

Figure 1: Boxplots summaries of the pMSE ratio values for 30 synthetic replicates at each maximum allowed number of household respondents. Lower values indicate closer distributional similarity with the confidential data.

7.2 Utility of the Synthetic LASI

We measure the utility of the synthetic data using distributional comparisons of individual, spousal, and household characteristics. We primarily focus on univariate and bivariate comparisons of the variables in the data, but we also compare the results of specific analyses one might run using the synthetic data. We compare both weighted and unweighted results.

7.2.1 Univariate Distributional Comparisons

We start by comparing all univariate distributions for each of the variables in the data. Table 2 provides some selected summary statistics. We do not see any significant differences in the distributions. It is common for synthetic data to match the marginal distributions, but recall that we synthesize households and most of the characteristics shown in the table are measured for individuals. This means our synthesis model is able to capture the individual distributions within households while synthesizing at the household level.

Variable	Confidential Data	Synthetic Data
Gender		
% Female	57.8 *[57.4, 58.1]*	57.7 *[57.3, 58.1]*
Education Category		
% Less than lower secondary	70.4 *[70.1, 70.8]*	70.6 *[70.3, 71.0]*
% Upper secondary/vocational	24.4 *[24.0, 24.7]*	24.3 *[24.0, 24.6]*
% Tertiary	5.22 *[5.06, 5.39]*	5.08 *[4.92, 5.25]*
Working status		
% Working	43.4 *[43.0, 43.8]*	43.3 *[42.9, 43.6]*
Age Group		
Under 45	9.33 *[9.12, 9.55]*	9.19 *[8.98, 9.41]*
45-49	18.3 *[18.0, 18.6]*	18.2 *[17.9, 18.5]*
50-54	15.2 *[14.9, 15.5]*	15.1 *[14.9, 15.4]*
55-59	13.9 *[13.7, 14.2]*	13.9 *[13.6, 14.1]*
60-64	14.1 *[13.8, 14.3]*	14.0 *[13.8, 14.3]*
65-69	12.2 *[12.0, 12.5]*	12.5 *[12.3, 12.7]*
70-74	7.90 *[7.70, 8.10]*	7.88 *[7.69, 8.08]*
75-79	4.58 *[4.43, 4.73]*	4.70 *[4.55, 4.86]*
Over 79	4.41 *[4.26, 4.57]*	4.49 *[4.34, 4.65]*
Mean Education years	4.32 *[4.29, 4.36]*	4.30 *[4.27, 4.34]*
Mean Self-reported health (out of 5)	3.19 *[3.18, 3.20]*	3.19 *[3.18, 3.20]*
Mean Activities of daily living (out of 5)	1.26 *[1.26, 1.27]*	1.26 *[1.25, 1.27]*
Mean Individual earnings	27,800 *[26,600, 29,000]*	27,800 *[26,600, 29,000]*
Mean # HH Residents	4.80 *[4.77, 4.82]*	4.77 *[4.75, 4.80]*
Mean # HH Respondents	1.68 *[1.67, 1.69]*	1.68 *[1.67, 1.68]*

Table 3: Variables Included in the Synthesis

7.2.2 Two-Way Distributional Comparisons

Next, we consider the similarity of bivariate distributions in the synthetic and confidential data to assess how well the synthetic data model captures correlations between pairs of variables. We summarize results using the same pMSE ratio distance metric, where 1 is the target value. This metric can be used across both continuous and categorical or binary variables, which makes it more applicable than directly comparing correlations.

Figure 2 shows the results for all two-way interactions. We see that the average pMSE values for bivariate relationships are below 10 (the rule of them from Raab et al. (2021)) for all but one variable, number of difficulties with activities of daily living (ADL), a commonly used measure of health

decline at older ages. While the marginal distribution of the number of ADLs was captured well, the bivariate relationships are somewhat less accurate for this variable than others in the data. Analyses using ADL are likely to be less similar to the confidential data than analyses using other variables in the data. These results suggest that the synthetic data models could be improved when it comes to capturing the distribution of ADL conditional on other characteristics. Working status and income also have higher discrepancy between the synthetic and confidential data. All individual two-way values are provided in Figure 7 in Appendix A.

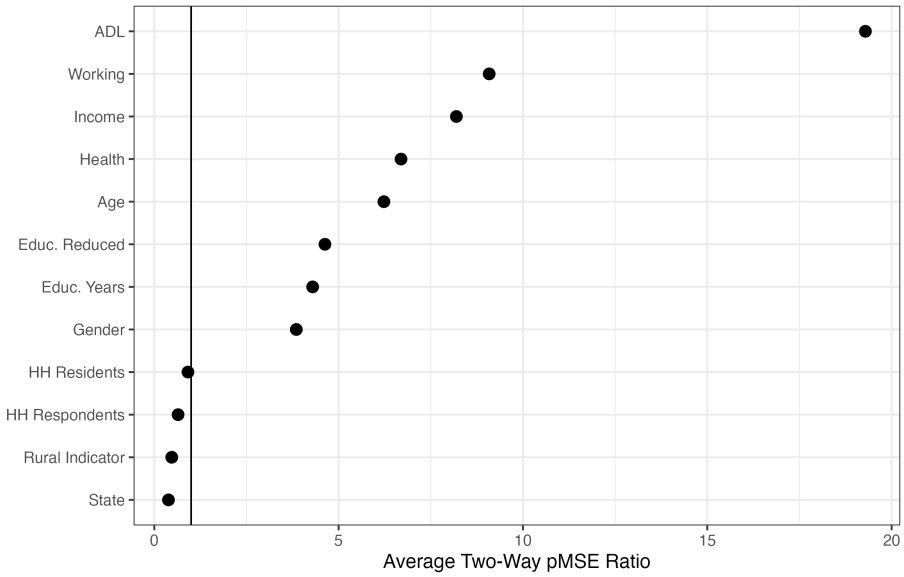

Figure 2: Average values across all two-way pMSE ratio values for each variable in the synthetic data. Lower values indicate closer distributional similarity with the confidential data. The vertical line is placed at 1, which implies ideal synthesis. Values can fall below 1 due to random variation or because they include non-synthesized variables, such as State and Rural Indicator.

7.2.3 Within Couple and Household Distributional Comparisons

Lastly, we evaluate how well the synthetic data capture the within couple and within household distributions in the confidential data by looking at the relationships between the male and female partner in the couple[5] and different persons in the household. First, Figure 3 shows the distribution of ages of the female and male partners in the couple in the left panel and the distribution of the oldest and youngest respondents in the household in the right panel, among households with more

[5]There are a small number of same-sex couples in the data, but we do not report results on these individuals due to confidentiality concerns. The synthetic data captures approximately the same rate of same-sex couples observed in the confidential data.

than one respondent[6]. Apart from a few couples where the female partner is significantly older than the male, the synthetic data accurately captures distribution in the confidential data. The synthetic data does an even better job, relative to the couples, of preserving the household age relationships.

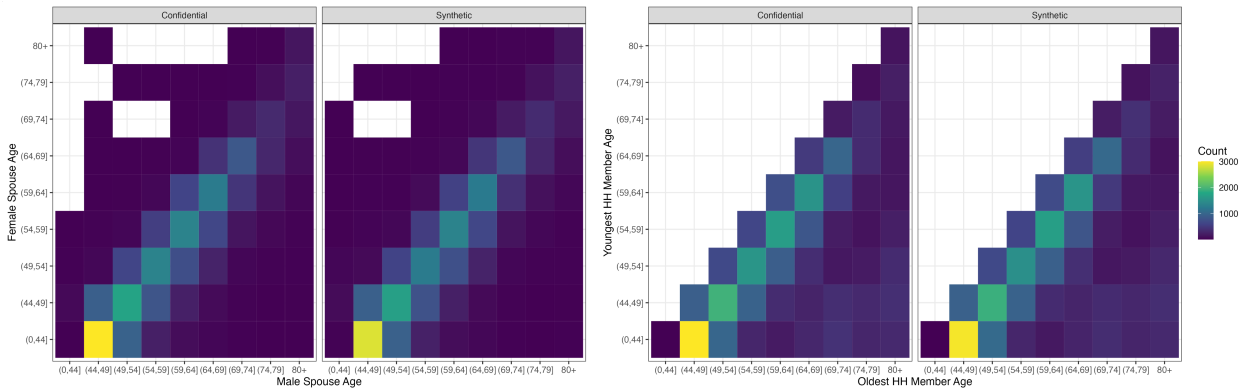

Figure 3: Left panel: distribution of the ages of the female and male partners in the couple in the confidential and synthetic data. Right panel: distribution of the ages of the oldest and youngest respondents in the household in the confidential and synthetic data. Both are only among households with more than one respondent.

Next, Figure 4 shows the difference between the educational attainment for the male and female partners in each couple in the left panel, and the difference between the most and least years of education among respondents in the household in the right panel, among households with more than one respondent. Again, the synthetic data captures the distribution accurately. It slightly under-predicts the number of couples with the same number of years of education, and the errors appear more likely to predict women having more educational years. We see that similarly to the couples, the synthetic data slightly under-predicts the number of households where the difference is zero, but overall the synthetic data does a very accurate job of replicating the distribution of educational attainment for individuals within the same household.

7.3 Disclosure Risks of the Synthetic LASI

We evaluate disclosure risks on the same synthetic data set for which we presented utility evaluations. We utilize the measures of identification and attribute disclosure risk described in Section 6.2. Recall that we assume an attacker has access to information about the state, rural vs. urban designation, number of household respondents, the ages, gender, educational attainment, and employment status

[6]Note that not all household members are respondents: only individuals age 45+ and their spouses were eligible.

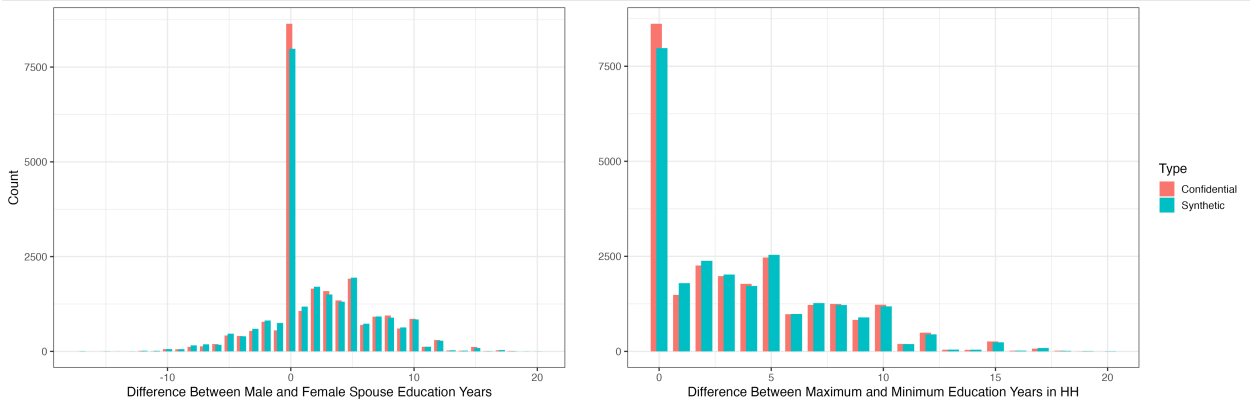

Figure 4: Left panel: distribution of the educational attainment (in years) of the female and male partners in the couple in the confidential and synthetic data. Right panel: distribution of the educational attainment (in years) of the respondents in the household with the most and least educational attainment in the confidential and synthetic data. Both are only among households with more than one respondent.

of the first two respondents in households they wish to identify. They link information of the households to the synthetic data and look for unique matches in the synthetic data.

For households that are unique in the confidential data, we compute the estimated risk of being unique in the population, defined in Section 6.2. Identification risk from this type of matching attack is higher for records that are more likely to be unique in the population, since the attacker is more likely to be identifying the correct household.

Households	N	Percent
Total	42311	100
Synthetic Unique-Confidential Unique	2505	5.92
SU-CU: $r > 0.1$	686	1.62
SU-CU: $r > 0.5$	215	0.508
SU-CU: $r > 0.95$	86	0.203

Table 4: Number of households with different thresholds of identification risk. Risk defined as the likelihood of being unique in the population given a set of QIs. SU-CU defined as records that share a unique set of QIs in both confidential and synthetic data.

Table 4 gives the results for the synthetic data. We see that while almost 6% of records are synthetic unique-confidential uniques (SU-CUs), the number of SU-CU records with high risk values declines. At the highest risk level, there are 86 household records in the synthetic data which are unique and can be matched to confidential records. As we will show in Section 7.4, we can expect a certain amount of matching simply by random chance, so we should not expect to see zero households in the

highest risk categories unless we manually edit the records. Depending on the context, the data maintainer may decide whether or not further protections are merited.

To help determine the potential harm from identifying a household, we estimate the risk of learning information about the total household income among households with high identification risk. We compute two measures as described in Section 6.2. First, we compute $a_1(b, R)$, the number of records where the observed synthetic income falls within a certain percentage difference, b, of the confidential income. Second, we compute $a_2(b, R)$, the probability of a percent range, b, of the confidential income given the estimated distribution of synthetic values.

	SU-CU Households		
	$R = 0.1$	$R = 0.5$	$R = 0.95$
Total Records	294	91	34
$a_1(50, R)$	67	23	9
$a_1(20, R)$	34	12	4
$a_1(10, R)$	20	6	1
$a_2(50, R)$	0.30	0.22	0.22
$a_2(20, R)$	0.11	0.073	0.075
$a_2(10, R)$	0.06	0.034	0.038

Table 5: Summary of total household income attribute risk values for households with high identification risk. Only households with nonzero confidential income are evaluated.

Table 5 provides the results. We see that the number of households with high attribute risk decreases as identification risk increases. We find only one household with the highest level of risk where the synthetic income is within 10% of the confidential income. For households with lower identification risk, we still find only 20 households where the synthetic income is close to the confidential value. We also see that the mean probabilities that the confidential value falls within the distribution estimated from the synthetic data are also low. For the highest identification risk group, we find a mean probability of 0.22 for a $\pm 50\%$ range around the confidential value, suggesting the attackers would likely learn little about even the rough magnitude of household income for particular records.

Having evaluated the utility and risk of one synthetic data set, we now turn to evaluate other synthesis options that may provide more or less utility and risk, depending on the goals of the data maintainer.

7.4 Risk-Utility Tradeoff for Additional Protections of States with Small Populations

The previous evaluations considered only one synthetic data set, which utilized our novel method for synthesizing household records but did not utilize the additional protections discussed in Section 4. The risk results in the previous section showed low risk, but these risks are not evenly distributed across geographic locations due to the survey oversampling. To better protect individuals in small locations, we evaluate two possible methods. First, one could simply coarsen the geographic information by grouping oversampled (small) states with larger neighbors. Alternatively, we propose to down-weight records based on the r measure of identification risk.

We evaluate three levels of down-weighting corresponding to different levels of protection. We give zero weight to any records with r values above one of the cutoffs $\{0, 0.5, 0.9\}$, effectively suppressing them from the synthesis. This removes any contribution of records sampled with certainty from the synthesis model. The lower the cutoff, the more suppression from the synthesis. Recall that r estimates the likelihood that a household record is unique in the population, so replicating records in the synthetic data that have higher values of r is more risky. For records with risk values below the cutoff we set the weight to $1 - r^2$. For records that have no expected value of being unique in the population, i.e., they are not unique in the confidential data, we set the weight to 1.

We evaluate the risk and utility trade-off for the different approaches by estimating (1) the overall $pMSE$ ratio for the distributional similarity of the entire synthetic data set to the confidential data set and (2) the identification risk based on the percentage of SU-CU records among those with different values of r. For each synthesis method, we replicate the process 30 times to avoid differences due to random chance and report the mean results.

Figure 5 plots the risk and utility values against each other. First note that the identification risk in the y-axis, as measured by the percentage of unique confidential records replicated uniquely in the synthetic data, drops for records with a higher likelihood of being unique in the population simply due to there being fewer records with high r values. We do not compare across panels, but we see a similar pattern when comparing different synthesis approaches within panels.

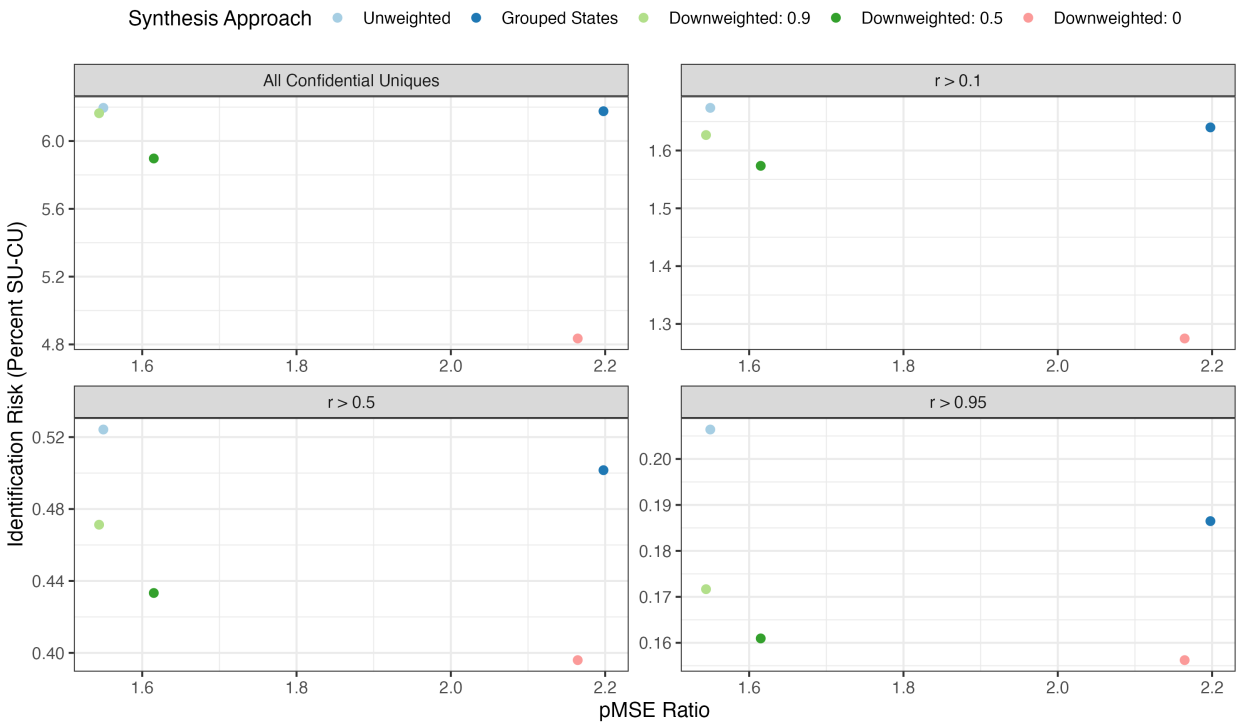

Figure 5: Risk and utility for different synthesis methods. Risk, on the y-axis (higher is more risk), is measured by the percentage of SU-CU records with certain risk thresholds. Utility, on the x-axis (lower is more utility), is measured by the pMSE ratio for the distributional closeness of the entire synthetic data set to the confidential data.

We see that the unweighted approach is in the upper left corner, signifying both higher utility and higher risk. The approach of grouping states does little to mitigate the risk, but it decreases the utility. The down-weighting approaches give progressively more protection in exchange for less utility, as we would expect. When only down-weighting the riskiest records, there is no drop in utility in exchange for increased protections, particularly among the riskiest records. As we increase the down-weighting threshold, we still maintain some utility while decreasing the risk even further. Results showing utility versus the attribute risk measures are given in Appendix B.

Figure 6 shows more granular utility results for two states, one smaller and one larger, that were grouped together in the coarsening approach. The small state has a high percentage of records with high risk values. In each figure the left panel shows the smaller state utility and the right panel shows the larger state utility. Each show both the results for (1) the full state synthesis (same as shown above), (2) the grouped synthesis, and (3) the down-weighted synthesis with the cutoff of $r > 0.9$.

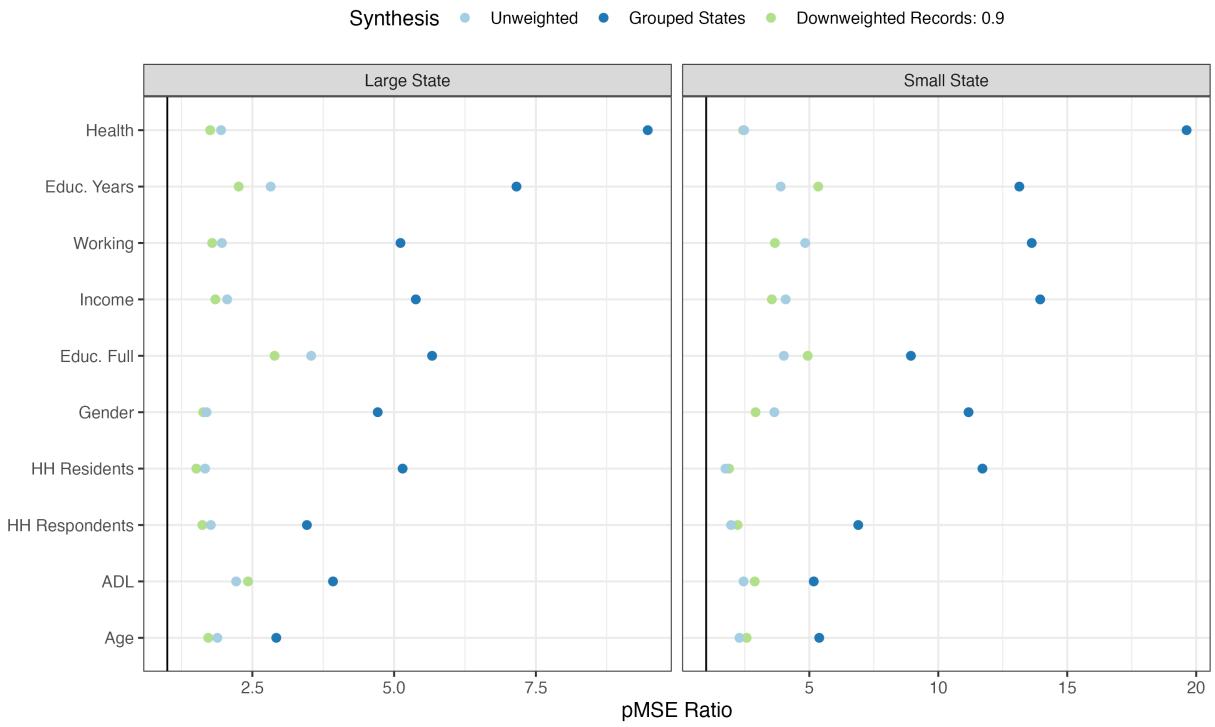

Figure 6: Average values across all three-way pMSE ratio values for individual or household variables in the synthetic data. Lower values indicate closer distributional similarity with the confidential data. Showing synthesis models using all states, grouped states, and down-weighted records. Left panel showing a small state and the right panel showing a large state, which were grouped together in the grouped synthesis

We see that the down-weighting preserves the accuracy of the two-way relationships for every variable in both the small and large states, while grouping the states distorts the distributions in each of the individual states. The fractions of SU-CU records with different risk levels in the small state (not shown) display a similar pattern to the overall results shown in Figure 5. In grouped states, the down-weighting approach is particularly better at preserving utility while reducing disclosure risk over a coarsening approach.

8 Discussion

We presented a novel approach for synthesizing individuals, couples, and households using simple sequential synthesis methods. This method provides high utility, capturing overall distributional similarity and maintaining the relationships between records in the data. This approach enables synthesis of this type of multi-structured survey that could not be accomplished using prior methods.

We also provide a novel method for managing the risk-utility tradeoff by down-weighting records using the record weight feature of CART models. We down-weight records with high probabilities of population uniqueness based on identifying features and sampling rates. For surveys that utilize significant oversampling, some records will be much more identifiable than others, so we need to mitigate their risk. We find that the down-weighting approach provides both better utility and less risk than the alternative of coarsening the geographic information.

We apply our methods to the first wave of the LASI, and we show that we can produce a synthetic version of this survey that maintains high levels of univariate, bivariate, and higher order distributional similarity to the confidential data. We also show that the identification and attribute risk of synthetic data released from this model are low.

Future work on this data set could expand the number of variables, either by including more questions from the LASI or by linking the variables here to additional data to synthesize jointly. As we do this, we will need to monitor different parts of the joint distribution of variables to ensure that it continues to capture the distribution observed in the confidential data. We may also target improvements on variables in the current file which have worse utility, such as ADL or working status.

Moving towards releasing synthetic data might require more considerations of the precise use-cases of the synthetic data, which will inform what variables are included in the data, the target accuracy requirements, and the level of risk mitigation required prior to releasing the synthetic data. The desired uses will also determine whether the data are released publicly or with additional access restrictions. Broadly speaking these decisions require policy decisions. The methods presented in this paper provide a foundation from which synthetic data versions of the LASI can be provided and evaluated.

Acknowledgements

The research in this study has been supported by a grant from the National Institute on Aging (R01 AG030153). We thank participants of the 2022 Virtual Gateway Advisory Meeting and the 2023 UCLA Synthetic Data Workshop for helpful discussions.

This analysis uses data or information from the Harmonized LASI dataset and Codebook, Version

A.3 as of April 2023, developed by the Gateway to Global Aging Data (https://doi.org/10.25549/h-lasi). The development of the Harmonized LASI was funded by the National Institute on Aging (R01 AG042778, 2R01 AG030153, 2R01 AG051125). For more information about the Harmonization project, please refer to https://g2aging.org/.

This document uses data from the 2017 – 2019 Wave 1 of LASI, Version B. LASI is a joint project of three partnering institutions: International Institute for Population Sciences (IIPS), Harvard T.H. Chan School of Public Health (HSPH), and University of Southern California (USC). LASI Wave 1 was funded by the Ministry of Health and Family Welfare, Government of India, the National Institute on Aging (R01 AG042778), and United Nations Population Fund, India.

References

Arnold, C. and M. Neunhoeffer (2020). Really useful synthetic data – a framework to evaluate the quality of differentially private synthetic data. *arXiv preprint arXiv:2004.07740*.

Benedetto, G., J. C. Stanley, E. Totty, et al. (2018). The creation and use of the SIPP synthetic beta v7.0. *US Census Bureau*.

Benedetto, G. and E. Totty (2023). Synthesizing familial linkages for privacy in microdata. *Journal of Privacy and Confidentiality 13*(1).

Bloom, D. E., T. Sekher, and J. Lee (2021). Longitudinal Aging Study in India (LASI): new data resources for addressing aging in india. *Nature Aging 1*(12), 1070–1072.

Bowen, C. M., V. Bryant, L. Burman, J. Czajka, S. Khitatrakun, G. MacDonald, R. McClelland, L. Mucciolo, M. Pickens, K. Ueyama, et al. (2022). Synthetic individual income tax data: methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases*, pp. 191–204. Springer.

Bowen, C. M., F. Liu, and B. Su (2021). Differentially private data release via statistical election to partition sequentially: statistical election to partition sequentially. *Metron 79*(1), 1–31.

Bowen, C. M. and J. Snoke (2021). Comparative study of differentially private synthetic data

algorithms from the NIST PSCR differential privacy synthetic data challenge. *Journal of Privacy and Confidentiality 11*(1).

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and regression trees*. Chapman and Hall/CRC.

Chien, S., C. Young, D. Phillips, J. Wilkens, Y. Wang, A. Gross, E. Meijer, M. Angrisani, and J. Lee (2023). Harmonized LASI documentation, version A.3 (2017–2021).

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: theory and implementation*. Springer.

Drechsler, J. and J. P. Reiter (2009). Disclosure risk and data utility for partially synthetic data: an empirical study using the German IAB Establishment Survey. *Journal of Official Statistics 25*(4), 589–603.

Drechsler, J. and J. P. Reiter (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis 55*(12), 3232–3243.

Drechsler, J. and J. P. Reiter (2012). Combining synthetic data with subsampling to create public use microdata files for large scale surveys. *Survey Methodology 38*, 73–79.

Hu, J., J. P. Reiter, and Q. Wang (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis 13*(1), 183–200.

Hu, J., T. D. Savitsky, and M. R. Williams (2021). Risk-efficient Bayesian data synthesis for privacy protection. *Journal of Survey Statistics and Methodology 10*(5), 1370–1399.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf (2012). *Statistical disclosure control*. Wiley.

Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician 60*(3), 224–232.

Kim, H. J., J. Drechsler, K. J. Thompson, et al. (2021). Synthetic microdata for establishment surveys under informative sampling. *Journal of the Royal Statistical Society Series A: Statistics in Society 184*(1), 255–281.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics 9*(2), 407–426.

Manrique-Vallier, D. and J. Hu (2018). Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *Journal of the Royal Statistical Society Series A: Statistics in Society 181*(3), 635–647.

Nowok, B., G. Raab, J. Snoke, and C. Dibben (2016). *synthpop: generating synthetic versions of sensitive microdata for statistical disclosure control.* R package version 1.8-0.

Nowok, B., G. M. Raab, and C. Dibben (2016). synthpop: bespoke creation of synthetic data in R. *Journal of Statistical Software 74*, 1–26.

Perianayagam, A., D. Bloom, J. Lee, S. Parasuraman, T. Sekher, S. K. Mohanty, A. Chattopadhyay, D. Govil, S. Pedgaonkar, S. Gupta, et al. (2022). Cohort profile: the Longitudinal Ageing Study in India (LASI). *International Journal of Epidemiology 51*(4), e167–e176.

Raab, G. M., B. Nowok, and C. Dibben (2016). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality 7*(3), 67–97.

Raab, G. M., B. Nowok, and C. Dibben (2021). Assessing, visualizing and improving the utility of synthetic data. *arXiv preprint arXiv:2109.12717*.

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, P. Solenberger, et al. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology 27*(1), 85–96.

Raghunathan, T. E., J. P. Reiter, and D. B. Rubin (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics 19*(1), 1–16.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology 29*(2), 181–188.

Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics 21*(3), 441–462.

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

Reiter, J. P., A. Oganian, and A. F. Karr (2009). Verification servers: enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis 53*(4), 1475–1482.

Reiter, J. P. and T. E. Raghunathan (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association 102*(480), 1462–1471.

Reiter, J. P., Q. Wang, and B. E. Zhang. Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *Journal of Privacy and Confidentiality 6*(1), 17–33.

Rocher, L., J. M. Hendrickx, and Y.-A. De Montjoye (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications 10*(1), 1–9.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics 9*(2), 461–468.

Skinner, C. and N. Shlomo (2008). Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association 103*(483), 989–1001.

Snoke, J., G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society 181*(3), 663–688.

Stadler, T., B. Oprisanu, and C. Troncoso (2022). Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1451–1468.

Therneau, T., B. Atkinson, B. Ripley, and M. B. Ripley (2015). *rpart: Recursive Partitioning and Regression Trees.* R package version 4.1.16.

Woo, M.-J., J. P. Reiter, A. Oganian, and A. F. Karr (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality 1*(1), 111–124.

A Appendix A: Additional Synthesis Evaluations

Evaluations comparing distributional similarity of all two-way interactions are shown in Figure 7. They show a similar pattern as the two-way results for the average value across each variable, but they highlight the bivariate relationships that are captured less well in the synthesis. We see that in particular, the interaction of ADL with working status, income, age, health, and education are poor relative to the other interactions.

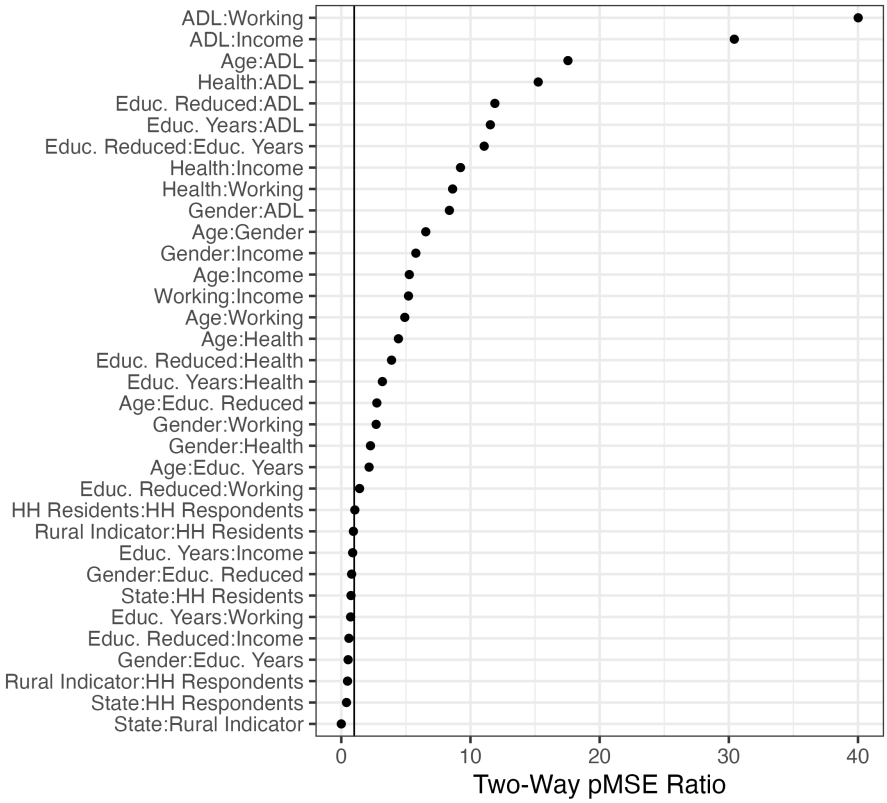

Figure 7: Two-way pMSE ratio values for all combinations of variables in the synthetic data. Lower values indicate closer distributional similarity with the confidential data. The vertical line is placed at 1, which implies ideal synthesis. Values can fall below 1 due to random variation or because they include non-synthesized variables, such as State and Rural Indicator.

Evaluations comparing distributional similarity of the average of all three-way interactions are shown in Figure 8. They show a similar pattern as the two-way results for the average value across each variable. We see that in particular, the distributions of ADL, working status, and income are poor

relative to the other variables on average in three-way distributions. The average pMSE ratio values are somewhat moderated compared to the two-way values, since each measure includes a third variable that has better utility.

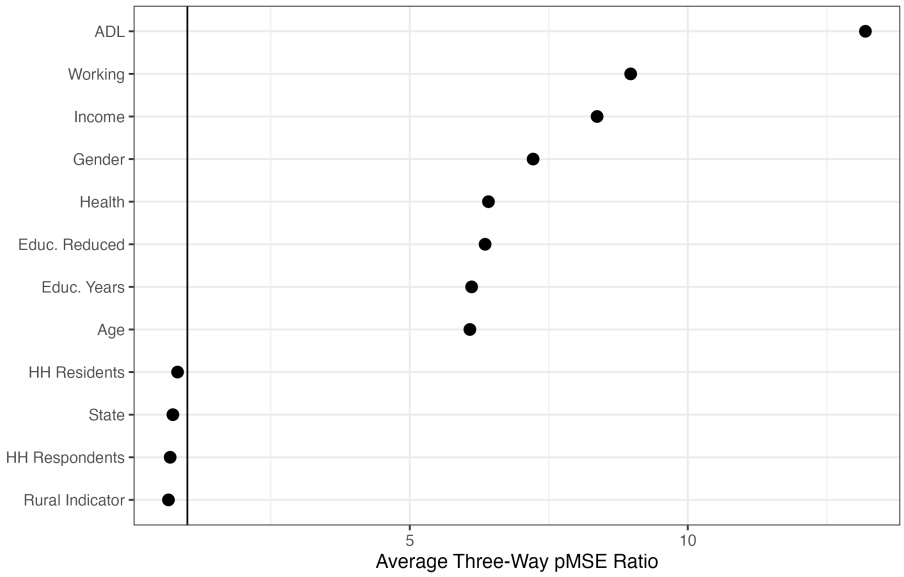

Figure 8: Average values across all three-way pMSE ratio values for each variable in the synthetic data. Lower values indicate closer distributional similarity with the confidential data. The vertical line is placed at 1, which implies ideal synthesis. Values can fall below 1 due to random variation or because they include non-synthesized variables, such as State and Rural Indicator.

B Appendix B: Additional Risk-Utility Tradeoff Results

Figures 9 and 10 plot the attribute risk and utility values against each other. We see that the unweighted approach is in the upper left corner, signifying both high utility and high risk. The percentage of records with synthetic incomes within the range of the confidential incomes shrink as we narrow the percent range. The approach of grouping states does not decrease the risk, but it decreases the utility. The down-weighting approaches provides more protection in exchange for less utility.

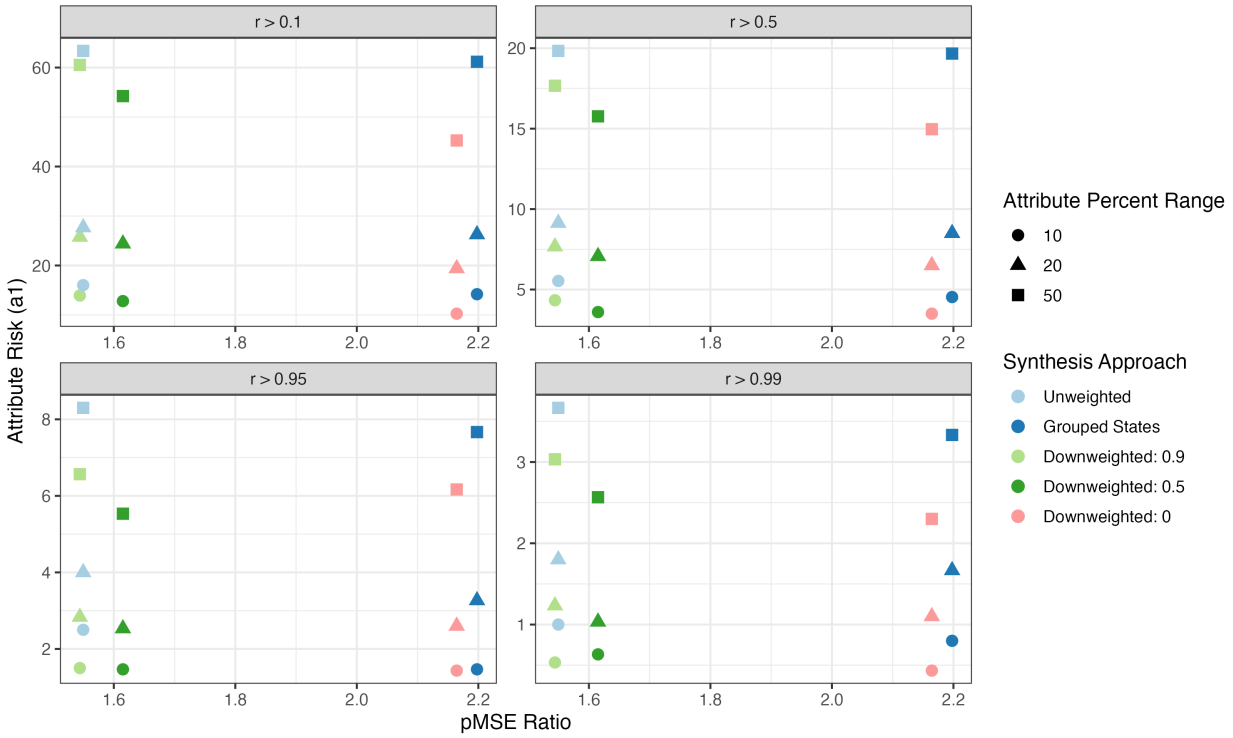

Figure 9: Risk and utility for different synthesis methods. Risk, on the y-axis (higher is more risk), is measured by the a_1 attribute risk measure for three levels of closeness on Income. Utility, on the x-axis (lower is more utility), is measured by the $pMSE$ ratio for the distributional closeness of the entire synthetic data set to the confidential data.

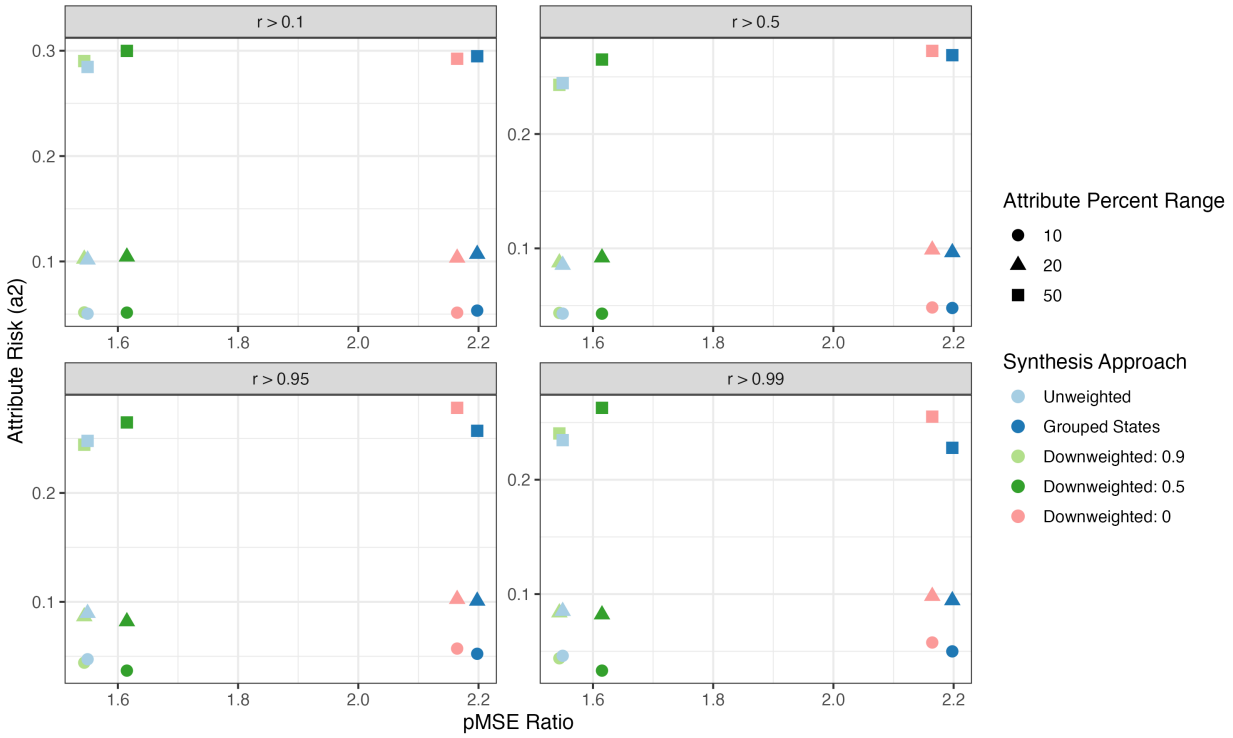

Figure 10: Risk and utility for different synthesis methods. Risk, on the y-axis (higher is more risk), is measured by the a_2 attribute risk measure for three levels of closeness on Income. Utility, on the x-axis (lower is more utility), is measured by the $pMSE$ ratio for the distributional closeness of the entire synthetic data set to the confidential data.