Two-Phase Income Synthesis Using IPUMS Data

May 2020

1 Abstract

The purpose of this study is to apply Bayesian methods to synthesize values of income, with respect to an individual's socioeconomic, ethnic, and health characteristics. This paper focuses on the utility and risk evaluation methods of a two-phase income synthesis approach applied to a publicly released dataset from 2018 annual National Health Interview survey, conducted by the United States Census. This two-phase approach utilized both logistic and linear regression models. Findings showed that the two-phase method reached a higher level of utility with a slightly higher level of risk compared to direct, single-phase income synthesis. We highlight several limitations and possible future research directions in order to assess the two-phase model, and its utility and risk evaluations.

2 Introduction

It is impossible to overstate the importance of data in today's world. Nearly every decision made by corporations and governments is based off conclusions drawn from data, in one way or another. However, many datasets structure information on an individual level (microdata) which leads to the possibility of identifying attributes of an individual or identifying the individual, known as disclosure risk [2]. Not only can this harm the individual but it may also lead to the disclosure of legally protected information, such as medical records. This is commonly avoided by using Bayesian methods to synthesize sensitive variables to reduce identification and attribute risks while maintaining the utility and relations between variables in the dataset.

The income variable is particularly sensitive due to its uniqueness and potential for outliers. With certain datasets, income will contain a significant number of zeros (individuals with no income) along with non-zeros (individuals with income), which typically results in a large spread between values. This can cause certain Bayesian synthesis models to lose effectiveness, thus yielding synthesized data with low utility. To prevent this, we attempt a two-phase income synthesis process.

For the first phase, we synthesize income categorically using logistic regression, where 0 indicates no income, and 1 indicates non-zero income. In the second phase, we synthesize all the non-zero income values using linear regression. Combining the synthetic 0s from the first phase and the synthetic income values from the second phase results in a completely synthesized income variable. This two-phase approach is implemented on data collected by the National Health Interview Survey of 2018 and assembled by the IPUMS Health Surveys[1], which collects information regarding health, healthcare access, and health behaviors of United States Citizens.

[1] https://nhis.ipums.org/nhis/

This paper is outlined as follows. In Section 3, we discuss the IPUMS dataset used to fit the two-phase income synthesis method, why the dataset was chosen, and the pre-processing of the dataset. In Section 4, we formally explain the two-phase income synthesis method. We present the results of the synthesis, including various utility and risk evaluation scores, in Section 5. Finally, we conclude with discussion of limitations and future direction in Section 6.

3 Background and Significance of the Research

Section 3.1 describes the IPUMS dataset, as well as the variables used in this study. Further, we outline the main ideas of partially synthetic data, followed by the data-cleaning process in Section 3.2.

3.1 The IPUMS Health Database

The IPUMS Health Surveys provide individual-level survey data for research purposes. The dataset includes extensive information on the demographic, socioeconomic, and health experiences of individuals living in the U.S. The data were self-reported by random participants representing the U.S. population. Although the IPUMS provides surveys to the public every year, we focus on the 2018 IPUMS Health Surveys for this study. 72,832 observations were collected from the United States Census. Specifically, the health and healthcare access information for this study was drawn from the National Health Interview Survey (NHIS) and the Medical Expenditure Panel Survey (MEPS).

The dataset is composed of a total of eight variables, including age, sex, race, education level, hours worked, health insurance coverage, hours of sleep, and frequency of worry. The dependent variable, income, is measured in binary and nominal terms. Specifically in the two-phase synthetic model, income of zeroes and non-zeroes are used in the logistic regression, and non-zero income are used in the linear regression. Age, hours worked, hours of sleep and income (linear regression) represents continuous variables, while sex, race, education level, health insurance coverage, frequency of worry, and income (logistic regression) represents categorical variables. The variables are all chosen based on self-intuition of most sensitive variables to an individual's income. The analyses for this research are restricted to samples that had all survey fields answered fully with no missing values to be a part of the sample size.

The income variable was chosen to be synthesized due to the high sensitivity and potential disclosure risk. When disclosing sensitive information such as income, there is high risk that an intruder will be able to derive the confidential information given their knowledge of other characteristics. Due to a wide range of possible values, income was deemed the most sensitive among the nine variables. Specifically, the relationship between income and hours worked, as well as income and education levels are the most important relationships to preserve. With the addition of six more health and socioeconomic variables, the identification disclosure risk rises. Thus, the income variable is partially synthesized due its sensitivity in order to protect the individual's privacy. In this case, partial synthesis refers to only replacing income values with simulated values, while the explanatory variables remain the same.

In order to protect the privacy of both individuals that receive and do not receive income, a two-phase synthesis measure is implemented. First, income is synthesized as binary values. This process is important to protect the privacy of unemployed individuals or students who receive zero income. Next, combining the synthetic 0s from the logistic regression with the non-zero income from the linear regression completes the full privacy protection method. Thus, the two-phase method can be applied to various datasets with a number of zero values that could potentially skew the distribution, in order to maintain high utility, and low disclosure risk.

Variable	Description	Type	Values	Synthesized
Income	Total earnings from previous calendar year	Categorical	0, 1	Yes
		Continuous	1 - 149,000	
Age	Age at time of survey	Continuous	18 - 85	No
Sex	Participant sex	Categorical	1 = Male 2 = Female	No
Race	Main racial background	Categorical	1 = White 2 = African American 3 = American Indian 4 = Asian 5 = Other races 6 = Two or more races	No
Education	Educational attainment	Categorical	1 = 4 years of high school or less 2 = 1 - 4 years of college 3 = 5+ years of college	No
Hours worked	Total hours worked last week or usually	Continuous	1 - 95+	No
Health insurance coverage	Health Insurance coverage status	Categorical	No, has coverage Yes, has no coverage	No
Hours of sleep	Usual hours of sleep per day	Continuous	0 - 24	No
Frequency of worry	How often feel worried, nervous, or anxious	Categorical	1 = Daily 2 = Weekly 3 = Monthly 4 = A few times a year 5 = Never	No

Table 1: Variables used. Data taken from the 2018 IPUMS public use dataset.

3.2 Data Preprocessing

The synthetic model was developed using the variables described in Table 1. The data cleaning process includes multiple steps. All missing or NA observations were removed. Next, NIU (Not In Universe) values, expressed as 0 and 00, were deleted from education, hours worked, health insurance coverage, hours of sleep, and frequency of worry. For education, hours worked, and hours of sleep, all rows that contained a variable value of 97 (Refused), 98 (Unknown- not ascertained), and 99 (Unknown - don't know) were removed. Similarly, this was done with health insurance coverage and frequency of worry for values of 7, 8, and 9, as well as for race but with values 970, 980, and 990. This reduced the dataset size from 72,832 observations to 14,287 observations. Because of computational limitations, we conducted our investigation using a random sample of 5000 entries. Finally, race and education were re-coded to the values described in Table 1.

The variable race was re-coded into 6 main racial backgrounds demonstrated in Table 1. The education variable was expressed by education attainment completed by grade, which was collapsed to 3 categories of 4 years of high school or less, 4 years of college, and 5+ years of college. Thus, each category is representative of a greater sample size.

4 Methods Used to Obtain and Analyze IPUMS Data

The paper's income synthesis method follows a two-phase approach. Section 4.1 explains generating the synthetic categorical income using a Bayesian logistic regression. The following step to generate the synthetic income requires a Bayesian simple linear regression, which is described in Section 4.2.

4.1 Generating the Synthetic Categorical Income

For the first step of the partial synthesis, income is synthesized categorically using a logistic regression. A dummy column is created for variable income with samples with non-zero income re-coded to 1, while samples with zero income remained as 0. Since income takes a binary out 0 or 1, the Bernoulli sampling model is used. A logistic regression is used in order to express the relationship between the binary dependent variable income and the other eight independent variables. Outcome variable $Y_i \in 0, 1$ represents a binomial random variable for i number of trials. $income_i$ is the success probability of event i taking $Y_i = 1$, or a non-zero income expressed as

$$Y_i \sim \text{Bernoulli}(income_i). \tag{1}$$

The logit of the Bernoulli probability is a linear combination of the predictors: age, sex, race, education, hours worked, health insurance coverage, hours of sleep and frequency of worry. In total, there are 21 parameters. Given the explanatory variables, a linear function of each variable in i number of trials for the logit of $income_i$ can be expressed as

$$\begin{aligned} \text{logit}(income_i) = \beta_0 + \beta_1 age_i + \beta_2 sex_{male_i} + \beta_3 sex_{female_i} + \beta_4 race_{w_i} + \beta_5 race_{b_i} + \\ \beta_6 race_{i_i} + \beta_7 race_{a_i} + \beta_8 race_{o_i} + \beta_9 educ_{1_i} + \beta_{10} educ_{2_i} + \beta_{11} educ_{3_i} + \beta_{12} hours_{wrk_i} + \\ \beta_{13} health_{cov_i} + \beta_{14} health_{nocov_i} + \beta_{15} hrsleep_i + \beta_{16} wor_{daily_i} + \beta_{17} wor_{weekly_i} + \\ \beta_{18} wor_{monthly_i} + \beta_{19} wor_{fewtimes_i} + \beta_{20} wor_{never_i}. \end{aligned} \tag{2}$$

We assume normal prior distributions for the regression parameters. With the Bayesian logistic regression, the Markov Chain Monte Carlo (MCMC) method is used to estimate the model through JAGS (Just Another Gibbs Sampler). There are relatively large auto-correlations displayed in the MCMC diagnostics. This could be due to the multi-parameter MCMC algorithms given the parameters used in the model, which are highly correlated. A smaller sample size may also contribute to the high auto-correlation. However, the trace plots indicate a significant amount of movement and very little stickiness. The posterior parameter draws are used to simulate synthetic data from the posterior predictive distribution. There are a total of $m = 20$ datasets generated. Table 2 demonstrates one of the 20 datasets generated. The synthesized income variable is renamed CatIncomeSyn (categorical income synthetic).

Table 2 displays the close resemblance between the distribution of synthetic values, CatIncomeSyn, and original values, CatIncome. The utility and risk evaluations are explained in the beginning of Section 5.

4.2 Generating the Synthetic Income

In the second phase of the synthesis model, a Bayesian simple linear regression is used to synthesize the non-zero income from the previous CatIncome. First, the 181 zero values from the original income, presented in Table 2 are removed. Then the non-zero income is logged. Let Y_i be the

	0	1	Total
CatIncome	181	4819	5000
CatIncomeSyn	168	4832	5000

Table 2: Original and Synthesized Income Comparison.

log(Income) and each X_i represent each variable for observation i. The Bayesian simple linear regression can be expressed as

$$Y_i \mid \mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma), \tag{3}$$

$$\begin{aligned}\mu_i = \beta_0' + \beta_1' age_i + \beta_2' sex_{male_i} + \beta_3' sex_{female_i} + \beta_4' race_{w_i} + \beta_5' race_{b_i} + \beta_6' race_{i_i} + \beta_7' race_{a_i} + \\ \beta_8' race_{o_i} + \beta_9' educ_{1_i} + \beta_{10}' educ_{2_i} + \beta_{11}' educ_{3_i} + \beta_{12}' hours_{wrk_i} + \beta_{13}' health_{cov_i} + \beta_{14}' health_{nocov_i} + \\ \beta_{15}' hrsleep_i + \beta_{16}' wor_{daily_i} + \beta_{17}' wor_{weekly_i} + \beta_{18}' wor_{monthly_i} + \beta_{19}' wor_{fewtimes_i} + \beta_{20}' wor_{never_i}.\end{aligned} \tag{4}$$

Due to limited prior information about each parameter, a weakly informative prior distribution is used. We must assume independence of the 22 parameters (21 β''s and 1 σ). The independence assumption can be noted as

$$\pi(\beta_0', \beta_1', \beta_2' ... \beta_{20}', \sigma) = \pi(\beta_0')\pi(\beta_1')\pi(\beta_2') \cdots \pi(\beta_{20}')\pi(\sigma). \tag{5}$$

Then, assign weakly informative priors for each parameter:

$$\beta_j' \sim \text{Normal}(\mu_j, s_j), \tag{6}$$

where $j = 0, 1, \cdots, 20$

$$1/\sigma^2 \sim \text{Gamma}(a, b), \tag{7}$$

where $\mu_0 = \mu_1 = \mu_2 = \cdots = \mu_{20} = 0$, $s_0 = s_1 = s_2 = \cdots = s_{20} = 100$, and $a = b = 1$.

JAGS is used in R to obtain 5000 posterior parameter draws. The MCMC diagnostics show low auto-correlation, which indicates the chain is mixing well. Further, the trace plots indicate a lot of movement. Given the posterior predictive distribution of the data values, we generate the synthetic values for income, called SynIncome (Synthetic Income). Given each explanatory variable and one set of posterior draws of the parameters $(\beta_0', \beta_1', \beta_2' ... \beta_{20}', \sigma)$ we could simulate synthetic values. There are a total of $m = 20$ synthetic datasets generated. Next, all non-zero income values in CatIncome, from the phase-one synthesis, are replaced with the SynIncome, synthesized income values. Thus, the synthesized zero and non-zero income values are merged together to generate full results in synthetic income.

5 Results of Analysis

The following two Sections, 5.1 and 5.2, outline the measures of utility and risk of the income synthesis processes, respectively. The utility of the synthetic income is measuring using both analysis-specific measures and global measures. The risk is analyzed in the context of identification disclosure. All measures are averages from 20 synthetic datasets.

Throughout the two sections, we compare the utility and risk measures of the two-phase income synthesis process alongside those of the single-phase income synthesis process. The latter was accomplished by synthesizing income directly (using the method described in Section 4.2), without

taking into account any zero values. All measures on the two-phase income synthesis are denoted with a "t" superscript, whereas all single-phase income synthesis measures are denoted with an "s" superscript.

From the results of the analysis, we conclude that the utility of the two-phase synthesized income is fairly high and the identification disclosure risk is relatively low. This is partly because the first phase of the two-phase synthesis demonstrates high utility, which makes it harder to discriminate between the original and synthetic data. In addition, there is a high identification disclosure risk evaluation in the first phase, which is mitigated in the second phase. Thus, the two-phase income synthesis method results in more data utility than single-phase income synthesis, although resulting in a slightly higher risk level.

5.1 Utility Measures

Sections 5.1.1, 5.1.2, and 5.1.3 describe the propensity score, cluster analysis measure, and empirical CDF measure, respectively, as well as list the scores calculated using these measures on our income synthesis dataset. All of these measures are global utility measures and were originally described in [3]. Next, in Section 5.1.4, we describe some analysis-specific measures, such as the mean and median, along with distribution comparisons using the interval-overlap measure.

5.1.1 Propensity Score

The first global utility measure calculated was the propensity score. Described in the context of data synthesis by Woo et al., the propensity score measure aims to quantify how distinguishable the masked (synthetized) data values are from the non-masked (original) data values [3]. This is achieved through merging both the original and synthetized datasets and adding a variable indicating membership of each unit to either dataset. Then, the probability of belonging to the synthesized dataset is computed for all values in the merged dataset. Finally, the distributions of synthetic-dataset membership are compared for both the original and synthesized values. Similar propensity score distributions indicate similar data distributions, and thus indicating high data utility [3]. This can be measured by percentile comparison, which is calculated as follows:

$$U_p = \frac{1}{N}\sum_{i=1}^{n}[p_i - c]^2, \tag{8}$$

where N is the number of units, p_i is the calculated propensity score for each unit, and c is the proportion of synthetic units in the merged dataset [3].

For both the two-phase and the single-phase income synthesis, $c = 1/2$ and $N = 5000$. The two-phase calculated propensity score for our merged dataset was as follows:

$$U_p^t = 2.41566e - 05, \tag{9}$$

which implies that $p_i \approx c$ across both the original and synthetic data, indicating high data utility. The single-phase propensity score, however, was:

$$U_p^s = 0.000567386, \tag{10}$$

which indicates a slightly lower utility when compared to the two-phase propensity score.

5.1.2 Cluster Analysis

Next, we calculated the cluster analysis measure [3]. The cluster analysis measure quantifies the distinguishability of the original and synthetic data from within clusters resulting from cluster analysis. The proportions of original and synthetic units are compared in each cluster, and then they are weighted and aggregated across all clusters. A higher score indicates larger differences between the original and synthetic data, implying a lower utility. In other words, a lower score indicates similar proportions between the original and synthetic data, implying a higher utility [3]. More formally, the cluster analysis measure can be stated as follows:

$$U_c = \frac{1}{G}\sum_{j=1}^{G}\left[\frac{n_j o}{n_j} - c\right]^2, \tag{11}$$

where G is the number of clusters, c is the proportion of synthetic data in the merged dataset, and $\frac{n_j o}{n_j}$ is the proportion of original units within each cluster j [3].

For our analysis, we set G, the number of clusters, equal to 50. The cluster analysis measure for two-phase income synthesis was calculated to be as follows:

$$U_c^t = 0, \tag{12}$$

which indicates high data utility. Similarly, the cluster analysis measure for single-phase income synthesis was:

$$U_c^s = 0, \tag{13}$$

again indicating high data utility.

5.1.3 Empirical CDF

Our final global utility measure calculated was the empirical CDF measure. Again described by Woo et al., the empirical CDF measure "assesses the differences between the empirical distribution functions obtained from the original and masked [synthetic] data." [3]. For original dataset X and synthetic dataset Y, let S_x and S_y be the respective empirical distributions. Additionally, let Z be the merged dataset. Thus, the following two values are calculated:

$$U_m = max_{1\leq i\leq N}|S_x(z_i) - S_y(z_i)|, \tag{14}$$

$$U_s = \frac{1}{N}\sum_{i=1}^{N}[S_x(z_i) - S_y(z_i)]^2, \tag{15}$$

where U_m is the maximum absolute difference and U_s is the averaged squared difference [3]. For both equations, lower values indicate higher utility.

Applying this measure to our original dataset and two-phase synthetic datasets resulted in the following values:

$$U_m^t = 0.10063 \tag{16}$$

$$U_s^t = 0.002671163, \tag{17}$$

and applying to our original dataset and single-phase synthetic datasets resulted in:

$$U_m^s = 0.24126 \tag{18}$$

$$U_s^s = 0.01930874. \tag{19}$$

Although U_m^t and U_s^t are relatively close to 0, their values indicate that our original and synthetic datasets have non-trivial differences between their distributions. Comparatively, U_m^s and U_s^s are significantly larger, indicating a decrease in utility.

5.1.4 Analysis-Specific Measures

For the 5000 original income values, the mean and median are as follows:

$$mean_{orig} = 50039.58, \tag{20}$$
$$median_{orig} = 40000, \tag{21}$$

and here are the mean and median for the averaged two-phase synthetic income values:

$$mean^{t}_{syn} = 50537.89, \tag{22}$$
$$median^{t}_{syn} = 33954.89. \tag{23}$$

The difference between $mean_{orig}$ and $mean_{syn}$ is small, indicating high utility. However, the difference between $median_{orig}$ and $median_{syn}$ is large, indicating that the distribution of income has changed during the synthesis process. This is evident in the histograms of the original and two-phase synthesized income values, depicted in Figure 1. The synthetic income histogram appears to be more skewed towards both 0 and $150,000$ when compared to the original income histogram. The skew seems to be heavier towards the lower values, which explains the decrease in the median.

This discrepancy is larger when considering the mean and median for the averaged single-phase synthetic income values:

$$mean^{s}_{syn} = 54297, \tag{24}$$
$$median^{s}_{syn} = 24570.07. \tag{25}$$

Figure 2 clearly shows that the original income and single-phase synthetic income are extremely different, and the distribution explains why the median ((25) above) is significantly lower than the original income median ((21) above).

Figure 1: Violin plot of original (0) and two-phase synthetic (1) income.

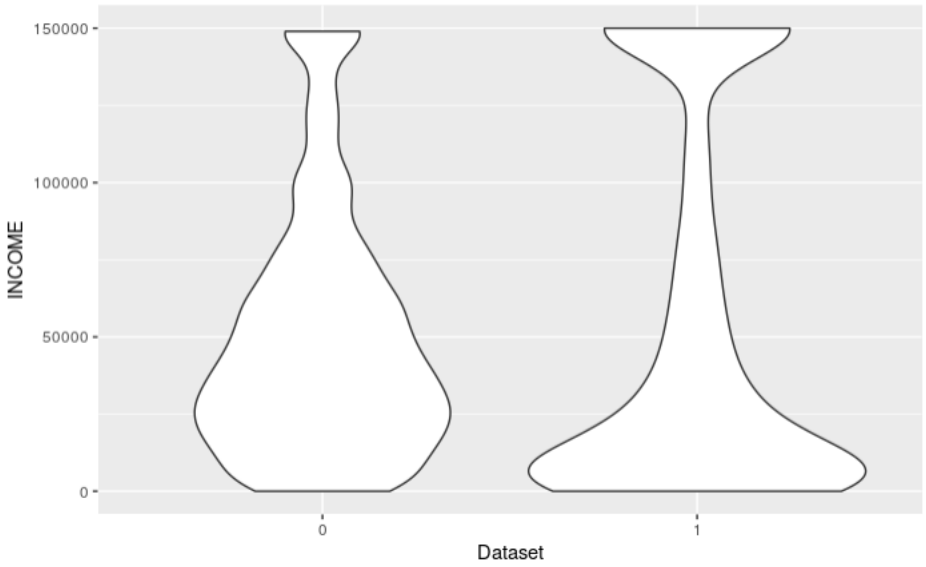

Figure 2: Violin plot of original (0) and single-phase synthetic (1) income.

In order to formally explore this disparity, we measured the variability between the 20 synthetic datasets for both phases, by calculating the following measures [1]:

$$\bar{q}_m = \sum_{i=1}^{m} \frac{q^{(i)}}{m}, \tag{26}$$

$$b_m = \sum_{i=1}^{m} \frac{(q^{(i)} - \bar{q}_m)^2}{m-1}, \tag{27}$$

$$v_m = \sum_{i=1}^{m} \frac{v^{(i)}}{m}, \tag{28}$$

where $q^{(i)}$ and $v^{(i)}$ are the point and variance estimates for the income in each synthetically generated dataset i. The above measures were used to construct an averaged 95% confidence interval across all synthetic datasets, which was then compared to the 95% confidence interval of the original dataset by measuring the overlap. For the two-phase synthetic datasets, the calculations resulted in the following:

$$\bar{q}_m^t = 50537.89 \tag{29}$$

$$interval_{95}^t = [49237.86, 51837.91]. \tag{30}$$

Additionally, measuring the single-phase synthetic datasets resulted in

$$\bar{q}_m^s = 54297 \tag{31}$$

$$interval_{95}^s = [52640.95, 55953.05]. \tag{32}$$

The original dataset had the mean listed in (20) above, and had the following 95% confidence interval:

$$interval_{95} = [48941.63, 51137.53]. \tag{33}$$

In order to compare the overlap between (30,32) and (33), we utilized the following interval overlap measure [1]:

$$I = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_s - L_s)}, \tag{34}$$

where $[L_o, U_o]$ is the original confidence interval, $[L_s, U_s]$ is the averaged synthetic confidence interval, $L_i = max(L_o, L_s)$, and $U_i = min(U_o, U_s)$. An interval overlap measure close to 1 indicates identical intervals and high utility, and a measure close to 0 indicates little overlap and low utility. Applying this measure to (30) and (33) gave us

$$I^t = 0.7978614, \tag{35}$$

which indicates relatively high utility, although it seems to reflect the disparities in median between the original and synthetic datasets mentioned above.

Applying the measure to (32) and (33) resulted in:

$$I^s = -0.5692832, \tag{36}$$

which is negative because the intervals do not actually overlap. This verifies that there was a significant distributional change between the original income and the single-phase synthetic income, thus indicating a decrease in utility.

5.2 Risk Evaluation

We evaluated the risk of our synthetic dataset by measuring identification disclosure through the expected match risk, the true match rate, and the false match rate [2]. To formally write the aforementioned measures, let c_i be the number of records with the highest match probability for target record i (records sharing same known variables). Let $T_i = 1$ be true if the true match is among c_i, otherwise $T_i = 0$. Additionally, let $K_i = 1$ if $c_i T_i = 1$ (if true match is unique), and $K_i = 0$ otherwise. Similarly, let $F_i = 1$ if $c_i(1 - T_i) = 1$ (if there exists unique match but it is not true match), and $F_i = 0$ otherwise. Finally, let N be the total number of records (5000 in our case) and let s be the number of uniquely-matched records. We assumed the known variables to be sex, race, and education level. Additionally, we considered a radius for acceptable synthetic income matches, namely $[0.1, 0.2, 0.5, 0.9]$ multiplied by the original income value.

Intuitively, the expected match risk quantifies the average likelihood of identifying a correct match for each record, which can be stated as follows:

$$E = \sum_{i=1}^{N} \frac{T_i}{C_i}. \tag{37}$$

Higher expected match risks indicate higher identification risk [2]. Next, we considered the true match rate, which indicates the percentage of true and unique matches. This can be quantified as

$$T = \sum_{i=1}^{N} \frac{K_i}{N}. \tag{38}$$

Higher true match rates indicate higher identification risk, and vice versa [2]. Note that the true match rate is bounded between 0 and 1. Finally, we considered the false match rate. The false match rate measures the percentage of unique matches that are false matches:

$$F = \sum_{i=1}^{N} \frac{F_i}{s}. \tag{39}$$

Unlike the first two measures, higher false match rates indicate lower identification risk [2]. Additionally, note that the false match rate is bounded between 0 and 1.

Radius	Measure	Single-Phase	Two-Phase
0.1	Expected Match Risk	0.264441	0.4580699
	True Match Rate	0	2e-05
	False Match Rate	1	0.9992752
0.2	Expected Match Risk	0.3296083	0.3916344
	True Match Rate	0	1e-05
	False Match Rate	1	0.9993243
0.5	Expected Match Risk	0.315921	0.3566615
	True Match Rate	0	0
	False Match Rate	1	1
0.9	Expected Match Risk	0.323256	0.3510375
	True Match Rate	0	0
	False Match Rate	1	1

Table 3: Risk measures for single-phase and two-phase income synthesis.

Table 3 lists the risk measures while varying the radius for both single-phase and two-phase income synthesis. For each radius value, there is a significant expected match risk, but the risk is slightly lower for the single-phase income values. Overall, the risk is slightly lower for single-phase income synthesis, as there is a lower correlation between the original and the synthesized income values.

6 Discussion

The two-phase approach is an innovative method for synthesizing income using a Bayesian logistic and linear regression. In our application, we found that a two-stage approach preserves the relationships of the variables, while maintaining low disclosure risks. Due to the wide availability of socioeconomic, demographic, and health characteristics, the risk for intruders to use the IPUMS database and derive confidential information is a concern for the agencies. With the objective of lowering disclosure risks, it is also important to maintain high utility so that key relationships are preserved and inferences can be made. Our results show that a two-stage approach yields higher utility and similar risk of the partially synthetic data compared to that of the single-phase synthesis.

In this paper, we highlight key steps to use a Bayesian logistic regression to synthesize zero and non-zero income. Then applying a Bayesian normal linear regression to the synthesized, non-zero income from the first step. Utility evaluations are estimated using the propensity score, cluster analysis, empirical CDF, and interval overlap measures. Then we present identification disclosure risk evaluations by calculating the expected match risk, true match rate, and false match rate. This study held strong implications for successfully synthesizing income, compared to just a one-phase approach.

The proposed synthesizer focuses on eight explanatory variables that encompass demographic, socioeconomic, and health characteristics. We chose the variables based on our own intuition due to possible correlations and sensitivity to the variable income. However, while significant, there may be additional variables that can be implemented to improve the utility evaluation and lower the risk measures. For example, variables related to medical care access, health behaviors, occupation, and family interrelationships can provide a more accurate model. Thus, an important future work direction is developing measures to assess what variables hold the most sensitive relationships with income. It is important to create a full model with as many significant variables, without resulting

in autocorrelation.

Another limitation is that we assumed random missing for missing values and removed those observations. However, there might be situations where missing values carry information about the observations themselves. The removal of certain samples reduces the overall sample size to 14,287. Due to the still relatively large sample size, the code could not run completely for the utility and risk evaluations. Thus, we were not able to use the full sample, which may cause the results to be less accurate. Thus, further exploration should be conducted by including more observations in the sample, in addition to more variables as mentioned before.

Due to the uncertainty with the current sample size and variables, it is not clear whether the two-phase model can be applied to other datasets. The study focuses on synthesizing only income, resulting in a partially synthetic data. However, we have not implemented our model to a fully synthetic data. We could continue using a Bayesian logistic regression to synthesize the binary income values, then implement a sequential synthesis for the second phase. Next, we can sequentially synthesize each variable at a time, given the previously synthesized variable. It would be beneficial to also implement our two-phase approach to various datasets and synthesize variables other than income.

A possible future work could involve the assessment of attribute disclosure risks. Attribute disclosure is when the intruder correctly infers the true values of synthesized variables in the publicly released synthetic datasets. In the paper, we only focused on three summaries of identification disclosure risks (expected match risk, true match rate, and false match rate). The addition of attribute disclosure risks can provide more insight on the effect of the two-phase model compared to the single-phase model.

Furthermore, research could be conducted to examine a way to synthesize values that are top-coded. A data point that is top-coded refers to an upper limit on data points that surpass an upper boundary. This is applied to datasets that may include outliers. The practice of top-coding can adversely censor high income points, which will make it more difficult to assess the distribution of income. However, it will protect the privacy of the individual's income. Developing methods for integrating top-coding measures, as well as attribute disclosure risk measures, is a key focus for further research.

References

[1] J. Drechsler and J. P. Reiter. Disclosure risk and data utility for partially synthetic data: An empirical study using the german iab establishment survey. *Journal of Official Statistics*, 25:589–603, 2009.

[2] J. Hu. Bayesian estimation of attribute and identification disclosure risks in synthetic data. *Transactions on Data Privacy*, 12:61–89, 2019.

[3] Mi-Ja Woo, Jerome P Reiter, Anna Oganian, and Alan F Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1), 2009.