DISCLOSURE RISK ASSESSMENTS AND CONTROL

By

Mandi Yu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in The University of Michigan
2008

Doctoral Committee:

Professor Trivellore E. Raghunathan, Chair
Professor Robert M. Groves
Professor Myron P. Gutmann
Associate Professor Michael R. Elliott

DEDICATION

To Mom, Dad and my lovely son, Zisheng

ACKNOWLEDGEMENTS

I would like to express my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my advisor, Dr. Trivellore Raghunathan. I have been amazingly fortunate to have an advisor who supported and guided every one of my steps towards completing this dissertation. I am thankful to him for holding me to a high research standard and teaching me how to do research.

I am very grateful to Dr. Robert Groves for his insightful comments and constructive criticisms, especially at the early stage of planning my dissertation research. I am indebted to Dr. Michael Elliott for carefully reading and commenting on this manuscript. I thank Dr. Myron Gutmann for serving on my dissertation committee, introducing me to the 1880 decennial census data, which I used in Chapter 4. I also thankful to Dr. John Van Hoewyk for his encouragement and countless help with the data set I used in Chapter 2.

I am also thankful to the former or current fellow students and staff at the Michigan Program in Survey Methodology and the Joint Program in Survey Methodology for their various forms of support during my graduate study. Special thanks to Benmei Liu for her great helps with numerous discussions.

Most importantly, none of this would have been possible without the constant love, support and strength from my mother Kun-ying Tian, father Xi-chen Yu, and my sisters, Manxue Yu and Mandan Yu, my nephew Yu-yao Zhang. Lastly, thank you, my beloved son, Zisheng, for bringing great meaning, peace and joy to my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Recent advances in technology dramatically increase the volume of data that statistical agencies can gather and disseminate. The improved accessibility translates into a higher risk of identifying individuals from public microdata, and therefore increases the importance of the evaluation of disclosure risk and confidentiality control. This dissertation addresses three related but distinct research questions in statistical data confidentiality.

The first study concerns the evaluation of disclosure risk for microdata when an intruder attempts to identify survey respondents by linking data records with a large external commercial data file based on a set of common variables. The dependence of disclosure risk to the commercial data coverage, the accuracy of the common identification information, and the amount of identification information to which an intruder accesses, is discussed theoretically and empirically tested using an experiment. The second study presents a practical implementation of fully-imputed synthetic data approach for a large, complex longitudinal survey as means of protecting confidentiality, following the initial proposal by Rubin (1993) and Little (1993). The imputation uses separate semiparametric algorithms for continuous, binary and categorical variables. A new combining rule of synthetic data inference is proposed to account for the uncertainty due to simultaneously imputing item-missing data and generating synthetic data. The loss of data utility is evaluated via the use of a propensity score approach in addition to three information loss metrics.

The third study extends this fully-synthetic data approach to cope with situations where small area statistics are essential important. This research is the first in the statistical disclosure control literature to consider small area statistics. The goal is to create synthetic data with enough geographical details to permit small area analyses, which otherwise is impossible because such geographical identifiers are usually suppressed due to disclosure control. A Bayesian framework for appropriate small area models is proposed to generate synthetic microdata from the predictive posterior distributions. Two simulation studies and one empirical illustration are used to evaluate this approach.

CHAPTER I

INTRODUCTION

1.1. Objectives

Sample surveys have been a key data source to support research and to inform public policy-making. Statistical agencies are obligated to disseminate high-quality data that are collected using public funds while also fulfilling the pledges of respondent confidentiality that they make to survey participants. One goal of confidentiality protection of such data is to avoid legal action in the case violation or to adhere to ethical mandates. The other important goal is to build public trust, which is a key contributor to survey response and data quality (Singer, 2003; Martin and Straf, 1992).

The inherent tension between data protection and data access imposes a complex set of tasks on the agencies. Research on these tasks arises from a wide spectrum of disciplines, from psychology and sociology to statistics and computer science. Over the past several decades, research topics on statistical disclosure control have specifically included 1) identifying and assessing the risk of disclosure for the original data, 2) developing statistical disclosure control (SDC) methods, 3) evaluating the utility of statistical analysis for a SDC-modified data set, and 4) re-evaluating the risk of disclosure for the modified data. These issues are closely interrelated in such a manner that knowledge about the risk of disclosure allows one to decide what data and how much data should be altered, and which SDC method should be used to achieve an optimum tradeoff between confidentiality protection and data utility loss.

The objectives of this dissertation are 1) to provide a more accurate evaluation of the disclosure risk for a US national survey by incorporating two largely ignored error sources, under-coverage error and measurement error, and 2) to develop robust and novel imputation models to construct fully-synthetic datasets for both a large-scale complex longitudinal survey and for a geographically referenced survey. The synthetic data can be disseminated in place of the real data, thus providing full confidentiality. The research issue is assessing the loss of data utility from such a data set. The synthetic data generation models can be easily adopted to resolve the confidentiality issue in other surveys facing similar disclosure challenges. This line of research is very important for data disseminators at large, and fits with the overarching goal of releasing higher quality data while still fulfilling the pledge of protecting respondents from having identifiable information inadvertently disclosed.

1.2. A brief review on disclosure risk

Breach of confidentiality occurs when a data unit is re-identified and the values of sensitive variables are disclosed. A distinction is made between disclosure risk from the respondents', the intruder's and the agencies' perspectives. From the respondents' perspective, the circumstances in which data, that the survey respondents have provided, may be released as identified data to a third party, and how the data will be used and by whom are concerned. The risk from the intruder's perspective emphasizes the real increase in intruder's knowledge about survey respondents provided that the respondents are believed to be correctly identified. The agencies' perspective speaks to the discredit harm to the statistical agency based on intruder claims of achieving disclosure. These three perspectives have different implications for the practice of risk assessment. The first one corresponds to respondents' perceived risks of disclosure, which leads to the question concerning how such risk influences respondents' willingness to participate in surveys. The latter two collectively concern the actual risk of disclosure, thus the safety of disseminating a particular data set after the data is collected.

More specific to the latter two types of disclosure risks, the literature has discussed several statistical measures. From the intruder's viewpoint, the high per-record risk (Skinner, 1998), which reflects the chance of correctly re-identifying one or more individuals in a public survey data, is desirable. The main practical usage of such a measure is to select records with the highest risk of re-identification in order to modify them, and thereby avoid potential disclosure (Skinner, 1998; Little and Liu, 2003; Reiter, 2005).

Global risk, on the other hand, measures the risk for the entire data file, which concerns the data disseminator in addition to the individual respondent. The global risk is commonly defined as the expected number of correct re-identifications and it can be computed by summing over all per-record risks, or by counting the number of records for which the risk of re-identification exceeds a given threshold (Lambert, 1993; Skinner, 2007). This measure is often used to inform agencies whether releasing one particular data set is safe. It is still up to the agencies to decide what the "safety threshold" should be.

It is worth noting that this definition of individual risk, although widely accepted, is rather generic. The specific measures of disclosure risk and the subsequent assessments are, in fact, highly sensitive to the assumptions of a model, based on how the risk estimates are derived. The key assumptions include those about the searching methods that an intruder may use (Skinner, 2007), the amount of auxiliary information that is available to an intruder, the quality of such auxiliary information (Paass, 1988), sampling design (De Waal and Willenborg, 1995; Benedetti, Capobianchi, and Franconi, 1998), and statistical distribution models for estimating the population frequencies. Examples of such distribution models include commonly used Poisson models, binomial models (Skinner, 2007), multinomial models (McCullagh and Nelder, 1989), and Poisson–gamma models (Bethlehem et al., 1990).

Under different assumptions, the literature includes a large body of risk measures. One class of measures is based on the population uniqueness model (Skinner, 1994; Fienberg and Makov, 1998; Benedetti and Franconi,1998; Franconi and Polettini, 2004). Under this model, by incorporating the sampling information, the risk is evaluated as the probability of being a population unique given being a sample uniqueness under various search methods (Skinner and Holmes, 1998; Skinner, 2007). An individual is population (or sample) unique if this person is unique in a population (or a sample) based on certain survey attributes.

The literatures, specific to microdata, suggest assessing the risk under the framework of record-linkage (Fellegi and Sunter, 1969) as the percent or the probability of correctly linking pairs of records on certain non-unique identifier(s) (Paass, 1988; Duncan and Lambert 1986; Willenborg and de Waal, 2000; Domingo-Ferrer and Torra, 2003). Record-linkage (also called exact matching in contrast to statistical matching), was originally developed to improve data completeness by linking records in separate files that relate to the same individual, allowing an analyst to pool the two sources of information together to develop a more complete statistical picture of each respondent (Winkler, 1997). When used by intruders, this technique poses the threat of a confidentiality breach. From the data producers’ standpoint, taking a record-linkage approach allows statisticians to mimic the intruding behaviors in assessing the risk of disclosure.

Such re-identifications may happen in many situations when the survey respondents are matched with the data from other sources, such as publicly available or privately held files, a different but related survey data file, or data files held by different organizations/businesses (Federal Committee on Statistical Methodology, 2002). The risk of re-identification from matching with external commercial data files particularly concerns statistical agencies because it is more likely than any other case to lead to malicious attacks.

Re-identification is established when the values of the common variables for survey respondents agree to those for the units from the public data. Three important factors contribute to the success of the re-identification: the number of common attribute variables, measurement quality of such common variables in both data sets, and commercial data coverage of the sampling frame population.

Specifically, the more personal identification information that one knows about survey respondents, ceteris paribus, the more likely that one or more particular potential victims are distinguished from the others in the data file, which in turn may lead to a higher risk of disclosure. Risk assessments thus should take into account the uncertainty about how much an intruder knows about the potential victims.

Another factor that may affect the success of linking records between two data files is the measurement quality of the common variables used in the linkage. Measurement error is ubiquitous in surveys. A substantial literature exists on measurement error in sample surveys (Biemer et al., 1991; Lyberg et al., 1997; Biemer and Lyberg, 2003). The measurement properties of commercial data, however, are rarely known. In any case, measurement errors in both data sources, if any, should be incorporated in assessing the correctness of the record-linkage. Suppose, one binary variable used in matching has two outcomes: 1 and 0. The observed values of this variable for this respondent in both data files may deviate from the true values due to measurement error (also called misclassification error specific to categorical variables). Table 1.1 shows the relationship between the occurrence of re-identification for survey respondent and the measurement misclassification errors in a binary key identification variable.

Table 1.1: Re-identification as functions of measurement misclassification in a binary linking variable from both data files.

			Survey Data			
	True Values		*1*		*0*	
		Observed Values	1	0	1	0
Commercial Data	*1*	1	🟧 Yes	🟩 No	🟩 Yes	🟧 No
		0	🟩 No	🟦 Yes	🟦 No	🟩 Yes
	0	1	🟩 Yes	🟦 No	🟦 Yes	🟩 No
		0	🟧 No	🟩 Yes	🟩 No	🟧 Yes

Note: 🟧 : No measurement misclassification errors in neither data files
🟩 : Measurement misclassification errors in only one data files
🟦 : Measurement misclassification errors in both data files

As illustrated in Table 1.1, re-identification occurs, regardless of the true value, as long as the observed values match, which is referred as measurement similarity (or measurement discrepancy). The exact values of measurement errors are irrelevant to re-identification although such information may help evaluate and predict the measurement discrepancy.

The last factor that affects the record-linkage is the coverage property of the commercial data. When the information for a survey respondent is not contained in the commercial data, the correct re-identification would not occur, thus the risk of disclosure is zero. In summary, both under-coverage and the measurement discrepancy are very important factors and should not be ignored in the evaluation of the risk of re-identification.

1.3. The relationship between disclosure risk and disclosure harm

Confidentiality breach involves an intruder gaining new information about the identified individual. Information about survey attributes that would draw exceptional interest to an intruder is usually sensitive in nature, which may lead to nontrivial consequences. For example, knowledge of "sensitive, stigmatizing and even illegal behavior by unauthorized others (family and friends, employers, insurers, or law enforcement agencies, for example) could subject the respondent to loss of reputation or employment, or to civil or criminal penalties" (Singer, 2003).

By taking the correctness of the disclosed sensitive information into account, the unanticipated harm due to the disclosure may or may not occur (Lambert, 1993; Trottini, 2003). The harm may be emotional, financial and physical, and it may damage a person's reputation depending upon the nature of the sensitive attribute. A simple but reasonable illustration on the relation between disclosure risk and disclosure harm is provided as follows. Suppose the intruder is interested in learning information about a sensitive attribute of an individual respondent, for example, a cancer diagnosis result with two possible outcomes: present and not present. Let us also suppose that malicious consequences occur, for example, when a health insurance company intrudes the data and infers that one survey respondent have cancer, and declines his/her insurance application.

Due to survey measurement error, a person's reported value for a survey attribute may appear consistent or inconsistent with his/her true value as shown in Table 1.2. If a person is correctly identified but his/her observed value is "no cancer", then disclosure harm would not occur regardless the underline true value. On the other hand, if a person is falsely identified to be someone who appears to possess the attribute (have cancer in our example), then harm occurs despite incorrect identification and the true value. Therefore, we may conclude that disclosure harm is statistic-specific and is highly related to the correctness of the reported value for the respondent.

Table 1.2: The occurrence of disclosure harm by the correctness of identification and inferred attribute of having cancer.

True Attribute	Have Cancer		No Cancer	
Observed Attribute	Have Cancer	No Cancer	Have Cancer	No Cancer
Correctly Identified	Yes	No	Yes	No
Falsely Identified	Yes	No	Yes	No

A quantitative evaluation of the disclosure harm is needed and is still an open area for research. The outcome of such research may add another aspect in planning SDC procedures. However, it usually requires the knowledge of the true values of the sensitive survey attributes and the magnitude of damage to the respondent from the intruder's knowledge of the attributes. Limited by the data availability, we will not address this issue in this dissertation.

1.4. A review of statistical disclosure control methods

Prior to public release, data are required to satisfy certain disclosure conditions. To achieve such goals, SDC procedures are usually applied to modify data. Common SDC methods include top-coding, data masking, data swapping, noise addition, categorical threshold, geographical thresholds and, most promisingly, synthetic data.

1.4.1. Common SDC methods

Top coding sets top-codes or bottom-codes on continuous variables. A top-code for a variable is an upper limit on all values of that variable. Any values greater than this upper limit are replaced by the top-code. Similarly, a bottom-code is a lower limit on all published values for a variable. Different limits may be applied for different variables, or for different subpopulations. For example, the values for the "self-employment income last year" within the 2006 American Community Survey (ACS) Public Use Microdata Sample (PUMS) files are top-coded at $140,000 and bottom-coded at -$9,999 for the State of Alabama (U.S. Census Bureau, 2007).

A related disclosure control method is data masking (also called data blanking). For example, if the observations in the tails of a distribution reveal the highest risk of disclosure, such as large firms for establish surveys or high-income persons for household surveys, the observations in the higher deciles are top-coded or masked.

Both top coding and data masking methods have the advantage of easy to implement and provide conditionality protection for individuals who have extremely values, which are considered to have the potential to reveal the identities. One common disadvantage, however, is that the data distribution is distorted, which would bias regression estimations and may potentially lead to sample selection problems. Even a data user is capable to apply sophisticated algorithms in analyzing the modified data to such biases; the results are still sensitive to the assumptions about the modified tails of the distribution.

Data swapping (Dalenius and Reiss, 1982; Reiss, 1984) involves swapping the values of variables for records that are statistically “similar”. This technique is usually implemented in such a manner that guarantees (under certain conditions) the maintenance of a set of statistics, such as means, variances and marginal distributions. For example, in the PUMS files of 1990 Census, 2000 Census, and the ACS, a percentage of households are swapped. The swapped households share a few characteristics but residing in different geographic locations. This procedure would not affect the estimation of the marginal totals for these areas and totals that include data from multiple areas. Despite the merit of this statistical maintenance, the joint distributions involving both swapped and un-swapped variables can be distorted.

Another popular method is to add independent random noises to the numerical variables, such as normal noise with the same correlation structure as the actual data, as means of controlling disclosure. The effects of such noises in regressions are well understood and discussed in the literature of measurement error models (Fuller, 1987; Fuller, 1993). Such additive measurement errors would only alter the original values slightly, especially when the original value is high. In addition, such random errors need to be incorporated appropriately into the statistical models that data users would fit to the altered data to ensure inference validity, which increases modeling complexity and the amount of burden on data users.

Categorical Threshold is an often-used approach to detect substantial risks associated with releasing one or more categorical variables. A cell in a table of frequencies formed by cross-classifying multiple categorical variable, is considered to be sensitive if the number of respondents in that cell is less than a certain predetermined number. When this happens, necessary SDC procedures, such as collapsing categories, rounding or suppression, have to be applied to avoid disclosure.

Geographical Threshold can be considered as a special case of categorical threshed method where the geographical identifiers are used together with other categorical variables in forming tables. Two conventional approaches for preventing this type of disclosure are, (1) to withhold reporting information on sensitive attributes in selected geographical combination cells, i.e. local suppression; (2) to aggregate all records within a geographical area so that the population is large enough to ensure any individuals or small groups of individuals can not be re-identified, i.e. geographical threshold, global recoding, or more generally, data aggregation. In either case, geographical details in the suppressed public data may be limited to areas exceeding a certain size. For example, the 5 Percent ACS PUMS do not publish geographic identifiers for geographical areas below a minimum population threshold of 100,000 and the 1 Percent PUMS uses a minimum population threshold of 400,000.

Local suppression and data aggregation are often used in combination and are available in the software program μ-ARGUS (http://neon.vb.cbs.nl/casc/) created by Statistics Netherlands (de Waal and Willenborg 1998). A more concrete description about these two methods is given in μ-ARGUS User's Manual (Statistics Netherlands 2007). These methods provide confidentiality protection but the information loss in data utility can be large, and such loss can not be evaluated systematically. Hurken and Tiourine (1998) constructed a mathematical model for minimizing information loss from global recoding and local suppression, but the heavy computation associated with that model may be impractical for real applications.

Specifically, when applied to geographically referred microdata, these SDC methods preclude the release of information that would otherwise directly provide solutions to address many important concerns in public policy, health or development that increasingly face the state and local governments. In addition, multivariate analysis describing complex interactions among geographical and social segments may also be impossible due to large amount of missing data. Finally, given that these SDC methods are somewhat ad-hoc and model free, the analytic properties using suppressed public data can not be justified (Winkler 2004).

1.4.2. Synthetic data approach

A final SDC method is to synthesize the values of microdata based on a probabilistic model. Initially proposed by Rubin (1993) and further developed by Raghunathan et al. (2003) and Reiter (2005a, Reiter 2005b), releasing multiply-imputed fully-synthetic public-use data in place of the actual data for disclosure control purpose is advantageous over alternative statistical perturbation methods (Winkler 2004, Reiter 2005a). Providing fully-synthetic data limits the risk of disclosing respondents’ identities and sensitive attributes completely since no real information is disseminated. This approach also allows users to analyze data validly using standard statistical packages. Information loss due to SDC procedures can be evaluated in a systematic fashion for pre-specified analyses in which the specific disclosure control procedures are taken into consideration.

The general idea is to treat the unobserved part of the population as missing data to be multiply imputed based on a model fitting with the actual data to complete multiple synthetic populations, then a simple random sample is drawn from each synthetic population which comprise the public-use data files. Valid inferences on a variety of scalar estimands from fully-synthetic data can be made using the methods developed by Raghunathan et al. (2003).

However, it is likely that the model used for synthesis may be not "congenial" to the models used by external data users. The definition of "congeniality" was originally given by Meng (1994) in the context of evaluating the inference of multiple imputation for item-missing data. Uncongeniality happens when the imputation model does not correspond to the analyst model, which in turn may lead to bias in statistical inference. In specific, if one fails to include an important dependent or predictor variable in the imputation model or mis-specify the relationship functions, the estimated coefficients associated with this variable in the analyst model will bias towards zero. In contrast, if one falsely incorporates an unrelated predictor variable into the imputation model, the estimation of coefficients in the correct analyst model is still unbiased although less inefficient.

In the case of fully-synthetic data, this problem is even more severe for two reasons. First, data values for the entire sample are to be imputed, thus imply stronger model dependency than the case of imputing for missing data. Second, synthetic data aims to allow external data analysts plan and test statistical models at will. In another word, the synthetic data is expected to yield valid inference for a large variety of statistical models, which are usually unknown to data imputers when the synthetic data are created. Uncongeniality can be dealt with by (1) incorporating as many variables as possible into the synthesis model to protect against bias, and (2) relaxing the assumptions about the relationships among the variables to avoid model misspecification. These two solutions can also be viewed as responses to the a criticism that is often be made about synthetic data, in which synthetic data only preserve the relationships considered in the model used to create them. However, both aspects would introduce significant extra complexity into the model building.

The SDC literature has taken two routes in response to this challenge by (1) reducing model dependency by imputing a smaller amount of data, thus limiting the damaging features to a smaller portion of the data or (2) by building imputation models with relaxed assumptions about the distributions and the relationships among variables to improve prediction. In the first route, there exist two variants to the fully-synthetic data approach, partial synthetic data and selective synthetic data (Little and Liu, 2002), in which the values for a portion of data records or variables are selectively synthesized at the expenses of losing a fraction of disclosure protection. Such selection is usually guided by disclosure risk assessments, disclosure harm and/or perceptions of disclosure harm suggested by variable sensitivity. The partial-synthetic data approach has been adopted by practice in creating public-use data (Little 1993, Kennickell 1997, Abowd and Woodcock 2001).

The second route involves the use of semi-parametric or non-parametric methods in place of parametric models in generating fully-synthetic data. Semi- and nonparametric models relax the usually strong distributional assumptions as in parametric models, thus potentially improve model fit and protect against model misspecification. These models are being used increasingly to recover the amount of variation in the dependent variable that is not explained by the independent variables under parametric regressions.

Several successful attempts of creating fully synthetic data for a small number of variable in national surveys with a general data structure have used either parametric or semi-parametric methods (Little and Liu 2002, Reiter 2005a). Further exploration of semi- or nonparametric models for the imputation of a large number of survey variables is necessary to make this fully-synthetic data approach really practically feasible for complex surveys in real world applications.

The existing literature in synthetic data (so far) has been mostly concerned with preserving statistics about the entire sample. However, for geographically referred data, statistics about small areas are often most important, and therefore, demanded. Significant theoretical and practical research on model-based small-area estimation has been conducted in the past three decades in an attempt to produce reliable small area estimates (Platek, Rao, Sarndal and Singh 1987, Rao 2003). As such, they contribute to a profound understanding of how small geographic area data can be summarized by statistical models, which can facilitate the building and selection of synthetic data models. Moreover, given the fact that surveys are heavily used to produce small-area statistics, synthetic data method is naturally challenged to support such analysis. The clustering structure due to the small-geographic areas, therefore, needs to be incorporated into the generation of synthetic data to produce valid and comparable statistics at the small-area level.

1.5. Organization of this dissertation

The rest of this dissertation is organized as follows. Chapter 2 evaluates the disclosure risk associated with the situation in which records from two datasets are matched by establishing a correspondence between shared common variables. We also investigate the effects on the risk assessments of the measurement discrepancies in such common variables and the under-coverage of the second dataset relative to the target survey data. We present a theoretical evaluation framework and apply it to an empirical experiment based on a national survey data and a large commercial data file.

In Chapter 3, a semi-parametric, multiply imputed, fully-synthetic data approach is developed to alter the actual survey data to control disclosure. We develop separate semi-parametric regression models for different variables. Synthetic data for ninety-eight variables are constructed based on the sequential regression algorithm. Item-missing data are imputed prior to the generation of the synthetic data. We develop new synthetic data inference rule to incorporate the variations due to simultaneously imputing for missing data and creating synthetic data. The proposed method is applied to the data from the Health and Retirement Study Wave1-4.

Chapter 4 constructs synthetic data for small areas. We develop parametric small area imputation models suitable for variables of different types. We conduct two simulation studies to evaluate the performances of the proposed models theoretically. We also present an empirical illustration of this method using the data from the 1880 US decennial Census. Lastly, Chapter 5 summarizes the main findings from these three studies and describes the directions for further research.

CHAPTER II

THE EFFECTS OF MEASUREMENT DISCREPANCY AND UNDER-COVERAGE RATES ON DISCLOSURE RISK ASSESSMENTS

2.1. Introduction

Statistical agencies pledge to protect the confidentiality when collecting or acquiring information for a statistical purpose. Such confidentiality pledges necessitate that statistical agencies assure that the identity disclosure of survey respondents is prevented. Knowledge about the risk of re-identification is a crucial tool to make informed decisions on the disclosure avoidance rules, the selection of data for modification, and the evaluation of statistical disclosure limitation methods (Little and Liu, 2003; Reiter, 2005).

Increasingly, there exist large external databases containing information on the entire population of the United States, such as commercial credit bureau records, customer transaction records, voting records, property records, employee records, health service records, etc. The availability of such databases with certain identifying information and key demographic variables coupled with powerful record linkage techniques may increase the risk to survey respondent disclosure.

Such disclosure threats, however, may be overstated, as the inaccuracy of the information used in the linking process could decrease the actual success rate of identification. Moreover, the commercial data may not have full coverage over the respondents contained in the public-use data. Thus, the re-identification and disclosure would not occur for those respondents who are not included in the commercial data. In fact, people with certain characteristics, such as youth, poverty or mobility are largely under-covered (Groves and Raghunathan, 2005).

We would expect that the actual re-identification risks would be smaller than those estimated when both measurement and under-coverage errors are ignored (Fellegi and Sunter, 1969; Willenborg and De Waal, 2001; Skinner and Elliot, 2002; Skinner and Carter, 2003; Winkler, 2004; Hawala, Stinson et al., 2005; Skinner, 2007). This argument was also echoed by other researchers to the point that the disclosure risk assessed with the error sources ignored is labeled as “conservative” and “pessimistic” (Skinner and Holmes 1998). Removing the bias towards the “conservativeness” is important in practice to ensure a satisfactory tradeoff between disclosure risk and data utility.

The goal of this chapter is to investigate how data quality, measurement error and under-coverage error, can affect the re-identification risk for a national sample survey. We also investigate the impacts from the various assumptions about the amount of information that is available to an intruder. In Section 2.2 we review the notations and develop a record-linkage framework for measuring the risk of re-identification. Section 2.3 specifies three distance-based measurement discrepancy functions for comparing the common key variables between the survey and commercial data, and a coverage measure of the commercial data. We also provide a summary of estimators for assessing the risk per record as well as for the entire data file under each combination of assumptions. In Section 2.4, we present the results from an empirical illustration, evaluating separately the re-identification risks for households and individuals within a national survey data. Finally, we conclude in Section 2.5 with discussion.

2.2. Disclosure risk under a record-linkage framework

2.2.1. A record-level measure of disclosure risk

Below we develop the measure for the risk of disclosure under the framework of re-identification. We define the risk of disclosure as the probability of one record being correctly linked to the same individual in the external database with respect to the values of certain key variables that are observed in both the survey data and the external data. Skinner (2007) summarizes three commonly used matching rules for re-identification in forensic literature. We focus on the rule in which all key variables are mainly categorical or can be treated as categorical (the variable “age” for example) and a pair of records is said to match only if all the key variables take the same value (i.e. exact matching).

We assume that a potential intruder has access to an external data, B with elements $j, j = 1, 2, ..., N$, which contains the key identifier Y^*. He or she attempts to obtain additional information on one or more survey respondents by examining the released public data, A with element i, where $i = 1, 2, ..., n$, and identifying individuals whose information on the common key identifier X matches Y^*. The typical key variables used to construct X and Y are routinely collected socio-demographic characteristics such as age, gender, race, education and marital status etc. One singular key identifiers X and Y^* are computed as all combinations of the key variables and both have K distinct categories.

We also assume that there exists a common unique identifier in both data files. Typical unique identifiers are a person name, Social Security Number, home address or telephone number, and etc. This type of identification information is almost always withheld from the public survey data for the purpose of confidentiality. Having such information accessible from the survey data, however, is necessary for the evaluation of measurement and coverage properties of the commercial data. The availability and quality of such unique identifier may vary by surveys because of the sampling frames and/or the modes of data collection. We assume that a carefully chosen common unique identifier guarantees that all units in each data file are uniquely identifiable and errors in linking the same individuals between A and B is ignorable. The type of disclosure risk we intend to address in this chapter is only limited to the conditions in which there exists such an external data file.

Suppose that, for an arbitrary respondent $\left(I,x\right)$ in survey data A with unique identifier value I and key identifier value x, an intruder attempts to find an individual $\left(I_j^*, y_j^*\right)$ in the commercial data B, which refers to the same individual as $\left(I,x\right)$. Here I_j^* is the unique identifier value and y_j^* is the key identifier value individual j within B, where $j=1,2,...,N$. Let the disclosure risk for individual $\left(I,x\right)$ be

$$Dr\left(I,x\right)_j = \Pr\left(I = I_j^* \mid y_j^*, i \in A, j \in B\right) \qquad [2.1],$$

which can be interpreted as the probability that $\left(I,x\right)$ and record I_j^* refers to the same person.

Under the assumption that the intruder's strategy is to link records presenting the same value of the key identifier in both data files, we only need to evaluate such disclosure probabilities when both records in a pair have the same value. For any pair of records taking different values, this probability of risk is zero, i.e. $Dr(I,x)_j = \Pr(I_i = I_j^* \mid y_j^* \neq x_i, i \in A, j \in B) = 0$. Therefore, Equation 2.1 reduces to

$$Dr(I,x)_j = \Pr(I = I_j^* \mid y_j^* = x, i \in A, j \in B) \qquad [2.2].$$

The interpretation of the disclosure risk becomes the probability that (I,x) and record I_j^* refers to the same person conditional on the fact that both records take the same value of the key identifier.

Let k be the combination value of the key identifier for individual I, which implies $x = k$. $M_k = \sum_{i \in n} I(x_i = k), k = 1,2,...,K$, where $I(\bullet)$ is the indicator function, $I(\bullet) = 1$ if the condition is met and $I(\bullet) = 0$ if otherwise, and $N_k^* = \sum_{j \in N} I(y_j^* = k), k = 1,2,...,K$ are the frequencies of the same combination, k, in A and B respectively, and $\sum_{k=1}^{K} N_k^* = N$ and $\sum_{k=1}^{K} M_k = n$. When $M_k = 1$, the survey respondent with key identifier value k is normally referred to as sample unique. Similarly, if $N_k^* = 1$, the individual in B with key identifier value k is referred to as population unique provided that we assume the commercial data set is intended to provide full coverage over the entire population of the United States.

If the intruder attempts multiple links between a survey respondent and the N_k^* known individuals in the commercial data, then there is a risk of re-identification for each attempted link. Therefore, we assume each pair of records may be resulted from an attack. For individual (I,x) in A, we get a vector of length N_k^* where each element corresponds to the disclosure probability $\mathrm{P}(I,x)_{N_k^*} = \left[Dr(I,x)_1, ..., Dr(I,x)_{N_k^*}\right]$.

Without seeking extra information to verify the correctness of these match-events, we assume an intruder searches through the external data file until a match is found. This assumption is equivalent to the one made in “Search method r1” of Skinner (2007) and the ‘journalist scenario’ of Paass (1988). We also assume that there is not a systematic order in the external data. Thus, the intruder’s search is equally likely to result in any match in the external data, which implies that the risk of disclosure for individual (I,x) equals the risk from any attack, i.e. $Dr(I,x) = Dr(I,x)_j$, $j = 1,\ldots,N_k^*$. Since the individual subscript j in $Dr(I,x)_j$ becomes irrelevant, in the rest of this chapter, we reduce $Dr(I,x)_j$ in Equation 2.2 to

$$Dr(I,x)_j = Dr(I,x) = \Pr\left(I = I^* \mid y^* = x = k, M_k, N_k^*\right) \qquad [2.3].$$

2.2.2. Under-coverage (UC) in disclosure risk

One important factor that may affect the likelihood for an individual being correctly identified is whether the information about this individual is included in the commercial data. Suppose for a sample unique record (I,x) in A, there exists its true pair record I^* in B, which is also population unique. Both sample uniqueness and population uniqueness are with respect to the key identifier as defined in Section 2.2.1. Ignoring measurement error, individual (I,x) has 100 percent certainty to be re-identified and sensitive survey attributes pertain to (I,x) will be disclosed. On the other hand, if I^* is not included in B, the chance of disclosure is then zero. Therefore, we need to account for the impact of under-coverage in the evaluation of disclosure risk.

We rewrite the disclosure probability $Dr(I,x)$ with respect to whether individual (I,x) is included in B as follows:

$$\begin{aligned} Dr(I,x) &= \Pr\left(I = I^* \mid y^* = x = k, M_k, N_k^*\right) \\ &= \Pr\left(I = I^* \mid y^* = x = k, M_k, N_k^*, I \in B\right)\Pr\left(I \in B \mid y^* = x = k, M_k, N_k^*\right) \\ &+ \Pr\left(I = I^* \mid y^* = x = k, M_k, N_k^*, I \notin B\right)\Pr\left(I \notin B \mid y^* = x = k, M_k, N_k^*\right) \end{aligned} \quad [2.4],$$

where $c(I,x) = \Pr\left(I \in B \mid y^* = x = k, M_k, N_k^*\right)$ is the probability for an individual (I,x) to be included in B, and then the probability that (I,x) is excluded from B, is $\Pr\left(I \notin B \mid y^* = x = k, M_k, N_k^*\right)$, which is the complement of $c(I,x)$, therefore, equals $(1-c(I,x))$. Let $FI(I,x)$ denote the probability for individual (I,x), who is not included in B, being falsely identified, i.e. $FI(I,x) = \Pr\left(I = I^* \mid y^* = x = k, M_k, N_k^*, I \notin B\right)$. We assume the risk of disclosure from a false identification is zero, therefore, $FI(I,x) = 0$ and Equation 2.4 becomes

$$Dr(I,x) = \Pr\left(I = I^* \mid y^* = x = k, M_k, N_k^*, I \in B\right) \times c(I,x) \quad [2.5].$$

Under the assumptions that (1) there is no measurement discrepancy in values of the key identifier between the two data files, and (2) an intruder may infer that individual (I,x), who is included in B, is as likely to be the same person as individual I^* as to any of the other $N_k^* - 1$ individuals in B, who take the same value of key identifier as (I,x), the first part of Equation 2.4 equals

$$\Pr\left(I = I^* \mid y^* = x = k, M_k, N_k^*, I \in B\right) = 1/N_k^* \quad [2.6].$$

Therefore, when there is no measurement error, Equation 2.4 reduces to

$$Dr(I,x) = c(I,x) / N_k^* \qquad [2.7].$$

2.2.3. Measurement Discrepancy (MD) in disclosure risk

In addition to under-coverage, the success of establishing exact linkage may also depend upon the measurement accuracy of the key identifiers between the two data files. The observed values of the key identifiers in both data files may deviate from their true values due to measurement error. In record-linkage, however, measurement error is less important than measurement discrepancy, which is the difference in the observed values between a pair of records that refer to the same entity. The reason is that regardless of the true values, identification occurs when the observed values for two records match, and this may or may not happen if either/both observed values are, in fact, the true values.

To incorporate the uncertainty in risk assessment due to measurement discrepancy, we further rewrite Equation 2.6 as a function of measurement discrepancy in the key identifier according to the Bayes' theorem

$$\begin{aligned}&\Pr\left(I = I^* \mid y^* = x = k, M_k, N_k^*, I \in B\right)\\ &= \frac{\Pr\left(I = I^* \mid M_k, N_k^*, I \in B\right) \Pr\left(y^* = x = k \mid I = I^*, M_k, N_k^*, I \in B\right)}{\Pr(y^* = x = k \mid M_k, N_k^*, I \in B)}\end{aligned} \qquad [2.8].$$

Under the assumptions that (1) an intruder may infer that respondent (I,x) is as likely to be the same person as individual I^* as to any of the other $N-1$ individuals in B when no identification information of any sort is available to the intruder, and (2) the external data contains the record for this respondent, the component $\Pr\left(I = I^* \mid M_k, N_k^*, I \in B\right)$ in Equation 2.8 equals $1/N$.

The second component, $\Pr\left(x = y^* = k \mid I = I^*, M_k, N_k^*, I \in B\right)$, is the measurement discrepancy on the key identifier between the pair of records that refer to the same individual, $\rho(I,x)$.

The third component $\Pr(y^* = x = k \mid M_k, N_k^*, I \in B)$ is the probability that a known individual I^* in the commercial data takes the same key identifier value, k as individual (I,x). Assuming Y^*, which is a vector of length N, follows a multinomial distribution with parameters $\theta = (N, p_k \geq 0, k = 1,2,...,K)$ and $\sum_{k=1}^{K} p_k = 1$, the maximum likelihood estimate for $p_k = \Pr(y_j^* = x = k \mid M_k, N_k^*, i \in B) = \Pr\left(y_j^* = k \mid M_k, N_k^*, i \in B\right)$ is $\hat{p}_k = N_k^* / N$. After substituting all above three parts back into Equation 2.7, we get

$$Dr(I,x) = N_k^{*-1} c(I,x) \rho(I,x) \qquad [2.9].$$

2.3. Evaluating Under-coverage rate and Measurement Discrepancy

2.3.1. Evaluating Under-coverage rate

Let $c_k = \sum_{i=1}^{M_k} I\left(I \in B \mid x = k\right)$ be the frequency of the combination value k of the key identifier for those individuals whose true matches are included in B. Assuming that all survey respondents with key identifier value k are equally likely to be included in the commercial data, the probability for an individual (I,x) to be included in B, $c(I,x) = \Pr\left(I \in B \mid y^* = x = k, M_k, N_k^*\right)$ defined in Section 2.2, equals $c(I,x) = c_k M_k^{-1}$. This equal chance inclusion assumption may not hold if survey respondents are selected with unequal probability or individuals are not randomly chosen to be compiled into the commercial data file. Nevertheless, both conditions are usually difficult to verify.

Furthermore, this assumption may provide a more realistic reflection of the type of information that an intruder may have.

2.3.2. Evaluating Measurement Discrepancy

In previous sections, the measurement discrepancy is expressed as a probability function, which has to be estimated. In this section, we propose three distance-based measures to evaluate measurement discrepancies as follows.

(1) Measure 1

This measure is developed based on the multi-dimensional key identifiers $\vec{X} = \left(X_1, X_2, ..., X_G\right)$ and $\vec{Y} = \left(Y_1^*, Y_2^*, ..., Y_G^*\right)$. Each dimension corresponds to one key variable. Assuming each key variable is equally important and contributes independently to the measurement discrepancy, a Euclidean distance is computed as:

$$\rho^1\left(I, x\right) = 1 - d(\vec{x} - \vec{y}_j^*)_{ij|I_i = I_j^*} = 1 - G^{-1}\sqrt{\sum_{g=1}^{G}\left(x_g - y_{g,j}^*\right)^2} \qquad [2.10],$$

where $\left(x_g - y_{g,j}^*\right)$ takes different forms depending on variable type. Specifically, if X_g is binary, $\left(x_g - y_{g,j}^*\right) = I\left(x_g \neq y_{g,j}^*\right)$; if X_g is nominal, $\left(x_g - y_{g,j}^*\right) = I\left(x_g \neq y_{g,j}^*\right)$, where levels$\left(\bullet\right)$ is the number of distinct levels; and finally if X_g is ordinal, $\left(x_g - y_{g,j}^*\right) = \left(\text{levels}\left(X_g\right) - 1\right)^{-1}\left|x_g - y_{g,j}^*\right|$. $\rho^1\left(I, x\right)$ takes values from zero to one, where one means that the two sets of measures agree perfectly and zero means there is not any similarity.

(2) Measure 2

An alternative crude measure is defined as the ratio of agreed dimensions over the total dimensions in $\vec{X}$. The estimator is

$$\rho^2(I,x) = r(\vec{x} - \vec{y}_j^*)_{j|I=I_j^*} = G^{-1}\sum_{g=1}^{G} I\left(x_g = y_{g,j}^* \mid I = I_j^*\right) \qquad [2.11],$$

where x_g and $y_{g,j}^*$ are the values for the g^{th} variable for individual (I,x) and I_j^* respectively. $\rho^2(I,x) = r(\vec{x} - \vec{y}_j^*)_{j|I=I_j^*}$ can take value within the range $[0,1]$. $\rho^2(I,x) = 0$ if the two vectors $\vec{x}, \vec{y}_j^* \mid I = I_j^*$ don't agree in any dimension of key identifier; and $\rho^2(I,x) = 1$ if the two vectors agree in all dimensions which also means absence of measurement discrepancy.

(3) Measure 3

This measure is the crudest measure defined based on an indicator function:

$$\rho^3(I,x) = \begin{cases} 1 \text{ if } x = y_j^* \mid I = I_j^*, N_k^* \\ 0 \text{ if } x \neq y_j^* \mid I = I_j^*, N_k^* \end{cases} \qquad [2.12],$$

where $\rho^3(I,x) = 1$ means the true pair records agree in their values on composite key identifier, and $\rho^3(I,x) = 0$ indicates measurement discrepancy present in Y_j^*. The information about both the magnitude and direction of the error are ignored.

The three measures vary with respect to the broadness of the measurement difference in the key identifier between two data sets. Unlike the third measure ρ^3, which is the broadest measure of the overall measurement discrepancy, the first and second measure, ρ^1 and ρ^2, considers each key attribute separately, effectively using the marginal distributions of each attribute rather than just their combinations. Therefore, both ρ^1 and ρ^2 distinguish among key identifiers that are more likely to be associated with errors than others. Moreover, ρ^1 differentiates variable types and captures the measurement error accordingly. Given these two distinctions, the respective values of these three measures follow a pattern in that $\rho^1 \geq \rho^2 \geq \rho^3$, although they are all bounded at 0 and 1.

2.3.3. Disclosure risk under assumptions about MD and UCR

Under different combination of assumptions about the measurement and coverage properties, four sets of estimators for the disclosure probability are developed.

(1) Estimator 1: Neither Measurement discrepancy nor Under-coverage is incorporated

Under this assumption, $\rho(I,x)=1$ and $c(I,x)=c_k/M_k=1$, thus $Dr(I,x)=N_k^{*-1}$. As noted by Skinner (2007), this simple form of the disclosure risk has been discussed extensively in SDC literature when the measurement error is ignored and the intruder data set is a complete register of the population or can be treated as providing full coverage for the target individual(s).

(2) Estimator 2: Only Measurement Discrepancy is incorporated

Since full coverage is assumed, $c(I,x)=c_k/M_k=1$. The disclosure risk is $Dr(I,x)=N_k^{*-1}\rho(I,x)$.

(3) Estimator 3: Only Under-coverage is incorporated

Absence of measurement discrepancy implies $\rho(I,x)=1$, thus $Dr(I,x)=N_k^{*-1}c_k M_k^{-1}$.

(4) Estimator 4: Both Measurement and Under-coverage are incorporated

When both error sources are present, the disclosure risk for respondent (I,x) to be correctly identified as subject I^* is estimated by $Dr(I,x)=N_k^{*-1}c_k M_k^{-1}\rho(I,x)$.

2.3.4. Global disclosure risk

The global risk of disclosure for the entire survey data A can be expressed as a simple summary of the individual risks, i.e. $Q = n^{-1}\sum_{i=1}^{n} g\left(Dr\left(I,x\right)\right)$, where $g\left(\bullet\right)$ is a statistical function. Candidate functions for $g\left(\bullet\right)$ can be (1) an indicator function of whether $Dr\left(I,x\right)$ exceeds certain risk threshold, which is considered to be high, thus the global measure can be interpreted as the expected proportion of records with risk of disclosure exceeding this threshold; 2) an identity function which leads to a global risk of the expected proportion of individuals in the survey data who can be correctly re-identified. This measure is very useful to inform agencies whether releasing one particular data is safe or not, although it is an agency's judgment on what the "safety threshold" should be.

2.3.5. Summary of disclosure risk estimators

In this section, we introduce two main factors that may contribute to the uncertainties in disclosure risk assessment: under-coverage and measurement discrepancy. Table 2.1 summarizes the estimators under each combination of assumptions about under-coverage rate and measurement discrepancy. The comparisons among the four sets of estimates under different assumptions will reveal the impacts of each error source alone and altogether on reducing the risk of disclosure.

Table 2.1: The estimators for the individual and global risk of disclosure under different assumptions about measurement quality of key identifier, external data coverage and intruding search methods

	Full Coverage		Under Coverage	
Measurement Discrepancy	Absent	Present	Absent	Present
Per-record Risk $Dr(I,x)$	$\frac{1}{N_k^*}$	$\frac{\rho(I,x)}{N_k^*}$	$\frac{c_k}{N_k^* M_k}$	$\frac{c_k \rho(I,x)}{N_k^* M_k}$
Global Risk[*] Q	$\frac{1}{n}\sum_{k=1}^{K}\frac{M_k}{N_k^*}$	$\frac{1}{n}\sum_{i=1}^{n}\frac{\rho(I,x)}{N_k^*}$	$\frac{1}{n}\sum_{k=1}^{K}\frac{c_k}{N_k^*}$	$\frac{1}{n}\sum_{i=1}^{n}\rho(I,x)\frac{c_k M_k^{-1}}{N_k^*}$

Note: [*] $Q = n^{-1}\sum_{i=1}^{n} g\left(Dr(I,x)\right)$ is evaluated based on the identify function, $g\left(Dr(I,x)\right) = Dr(I,x)$. It can be evaluated with regard to other functions per researchers' will.

2.4. An Empirical Illustration

In order to assess the risk of disclosure, we need to estimate both under-coverage rates and measurement discrepancies. In this section, we present an empirical experiment that involves a national sample survey A and a large commercial data file B. The two data files are linked at household level based on unique identifiers, therefore, they allow the evaluation of both under-coverage rates and measurement discrepancies.

We evaluate the risk of disclosure for both households and individuals in the survey data A from linking records within B based on a set of common key variables. We also evaluate the impacts of measurement and coverage properties on such risk assessments under different assumptions on the amount of information possessed by intruders. Varying the amount of linking information also answers the question of how much risk data disseminators should expect when releasing different sets of key variables.

We are obligated not to disclose any identification information about the data files used in this study. The information such as the name of survey, survey topic, sponsorship and the organizations that collected data, the name of the commercial data vendor, and etc. are held anonymous. Necessary details about sample size, inferential population, key identification variables, and unique identifier are given to allow the evaluation of the scientific value of this research.

2.4.1. Data Description

2.4.1.1. Survey data A

A survey data, A with more than 6,000 subjects was used in this study. The key demographic variables were Age (7 categories), Gender (2 categories), Race (4 categories), Education (5 categories), Marital Status (2 categories), Household Size (6 categories) and Children (short for "whether there are children in the household" and has 2 categories) and Household Income (3 categories). Among these variables, three are household-specific information: Household Size, Children and Household Income, and the remaining variables are individual specific.

2.4.1.2. Commercial data file B

The commercial dataset B is a leased large national population database with approximately 120 million household and 200 million person records. Included in the file are unique identifiers, names and addresses, and several key demographic variables, such as age, gender, education level, race/ethnicity, marital status and the number of children in a household. In addition, the database includes a number of indicators of wealth, purchasing behavior, and leisure and professional activities.

These data were complied from a number of public and private sources and modeled data. Therefore, the data quality may be implicated by modeling uncertainty. For instance, approximately 68% of age values were missing originally and replaced by some model estimates. Twenty-nine percent of Marital Status information has missing values and the remaining observed values are marked with quality confidence levels “extremely likely” (41%) or “likely” (30%) assigned by the data vendor. Moreover, there is only one missing value on Education and 20% values are recorded as “extremely likely” and the other 80% as “likely”.

This type of commercial data file is primarily intended for marketing and market research. We assume a potential intruder has access to such commercial data, and intends to identify households and/or individuals in A by matching the values of the common variables. We simulate the situations in which an intruder has access to different amount of information about the known individuals by attempting linkage based on different subsets of variables. Table 2.2 shows the composition and sample size for each sub dataset. The common variables in the commercial data file are coded consistently with the survey data.

Table 2.2: The collection of datasets simulating different intruding scenarios for household and individual level risk assessments

Variable	Household Level		Individual Level			
(No. of Categories)	Data 1	Data 2	Data 1	Data 2	Data 3	Data 4
Age (7)			×	×	×	×
Race (4)				×	×	×
Gender (2)				×	×	×
Education (5)					×	×
Marital Status (2)					×	×
HH Size (6)	×	×			×	×
Children (2)	×	×			×	×
HH Income (3)		×				×

2.4.2. Analysis Methods

The survey data set A is a file where there is only one individual record in a household. B, on the other hand, is a hierarchical file, i.e. there are multiple individual records within a household. The disclosure of individual or household information requires that the individuals or households be correctly identified. The households from the two data files are matched based on addresses. Due to the non-availability of any unique individual identifiers, the exact match on individuals is unattainable. We conduct a series of analysis evaluating the risk of disclosure for households and individuals respectively.

We first investigate the measurement properties of the key variables at the household-level and the impacts on evaluating the risk of identification for households. These evaluations are conducted under the framework presented in Section 2.4 with the unit of analysis a household. Second, for each pair of matched households, we assume all individuals in B are equally likely to be the correct link to the survey respondent, we evaluate the individual-level risk of disclosure and how the quality of data on individual specific key variables affect the risk assessments.

2.4.2.1. Exact record linkage between the two databases

Exact linkage method was used to match households between A and B based on eight components of house addresses present in both files. The address variables are Zip Code, City Name, Street Name, House Number, Street Suffix, Pre-direction/Post-direction, Unit Designator and Unit Designator Number. A pair of households is a true match if it agrees completely on all eight variables. The record linkage procedure was conducted by another research group and the details about this procedure appear in Raghunathan and Van Hoewyk (2008).

2.4.2.2. Household Level Analysis

Of the total records in A, 50% households were uniquely linked with households in the commercial data B, around 0.8% were multiply linked and the rest of more than 49% were not linked of any sort. The primary reason for multiple links is that the physical address can only be used to identify a building where there are more than one residing units. We treat this small portion of multiple-links as if they were the non-links, which gives around 50% non-linked households. Among the correctly linked HHs, around 6.4% were recorded in the commercial data as empty households with reported household size of zero, so personal level information is unavailable. Therefore, the proportion of uniquely linked non-empty households is 43.6%. The remaining commercial data households are not linked to any records in A.

Table 2.3 shows characteristics on the matched-but-empty, non-matched, and matched households based on the information collected in the survey. Missing data are excluded from this analysis. We also provide the proportions of complete sub-sample for each type of households. For example, the sample size of matched households with the values of Household Size observed is $43.6\% \times 28.3\%$ times the total sample size n. We also incorporate the sampling weights due to unequal probability of sampling selection in estimating subclass means and their corresponding standard errors. We conduct two sets of independent t-tests of comparing the characteristics for empty households and non-matched households against matched households respectively.

The recorded empty households in the commercial data vary widely in size based on the survey data, ranging from one to more than six. On average, they tend to be smaller households. Both the income and the proportion of HHs with children are similar across these two groups. Although race was collected at the individual level, we use the single person race information to infer the whole household considering the fact that household members are usually in the same race category. Empty households tend to be white rather than Black, Hispanic or other races.

Table 2.3: Characteristics from survey data for matched households, matched-but-empty households and non-matched households

	Matched HHs			Empty HHs			Non-Matched HHs		
	Mean	SE	%	Mean	SE	%	Mean	SE	%
HH Size (persons)	3.2	.03	28.3	3.0**	.09	26.3	2.9***	.03	24.3
Children in HH(%)	47.0	1.2	28.3	47.1	3.3	26.3	41.5***	1.1	24.3
HH Income LT 50K (%)	51.5	1.2	25.1	51.6	3.3	24.2	65.8***	1.1	21.5
HH Income 50-74K(%)	21.6	1.1	25.1	20.4	2.8	24.2	17.7***	0.9	21.5
HH Income GT 75K(%)	26.8	1.1	25.1	28.0	2.9	24.2	16.5***	0.9	21.5
White(%)	68.1	1.1	43.5	74.1**	2.7	40.1	55.2***	1.1	37.3
Black(%)	12.9	0.8	43.5	11.1	1.9	40.1	15.6**	0.7	37.3
Hispanic(%)	13.5	0.7	43.5	10.1	1.7	40.1	20.5***	0.9	37.3
Other Race(%)	5.5	0.6	43.5	4.7	1.5	40.1	8.8***	0.6	37.3
Proportion of sample w/ item-missing data:	43.6%			6.4%			50%		

Note: Significance levels for two-sided tests (matched vs. empty and matched vs. non-matched): * $a = 0.10$; ** $a = 0.05$; *** $a = 0.01$.

We also compare non-matched HHs with matched HHs. All these characteristics are different between the match and non-matched households. On average, the matched HHs tend to be larger, have children, and in a higher income category. Non-matched HHs tend to be Black, Hispanic or another race other than white. This is not surprising as income is often found to be associated with race. Such discrepancies suggest that the commercial data tend to omit the financially disadvantaged HHs. It is likely to be explained by the sources that B use to compile the data. If the commercial data rely heavily on credit history or transaction reports, then information about financially inactive and low-income HHs may be less readily available. In terms of the likelihood of having children within a household, there is no difference between these two groups.

As presented in Table 2.2, we simulate situations where an intruder has access to different sets of key information: (1) Data 1: household composition variables only, including Household Size and Children; (2) Data 2: household composition and household income. Table 2.4 shows the summary statistics for the coverage rates estimated when key identifiers are defined by different sets of common household-level attributes. We use the estimator for the coverage probability presented in Section 2.3.1 to calculate the coverage rate for a combination value of key identifier, which is the ratio of the number of survey HHs that are covered in the commercial data over the total number of survey HHs. When only household composition variables are used in constructing the key identifier there are a total of 12 distinct categories in the key identifier and the mean coverage rate across all 12 categories is around 50%. When the information about household income is also used, the number of categories in the key identifier increases to 34, and the mean coverage rate is around 57%. Under either situation, variations exist in the coverage rates across categories.

Table 2.4: Summary statistics of coverage-rate, $c_k M_k^{-1}$, for two sets of household-specific key variables

	No. of Categories	Min.	Mean	Max.	SD
Data 1	12	0.000	0.499	0.609	0.166
Data 2	34	0.000	0.566	0.786	0.135

To shed some light on the data quality for the household-level information, we compare the statistics on matched households from A and B as shown in Table 2.5. Only a quarter of households agree on their HH size. It is worthy noting that the magnitude of measurement discrepancies in Household Size between the two data sources are very large, considering the fact that both measures are top-coded (6 in A and 8 in B). Around 50% of measures on Household Income agree, which is very low considering that both measures are crudely coded using a 3-category ordinal scale. Around 60% of records show consistent reports on whether children are present in a household. 34% of households in A that have children are recorded in B as without children. In sum, there are large measurement discrepancies in household-level statistics between the two data sources.

Table 2.5: Discrepancies on the household-level measures between A and B for matched households

HH Size	%	HH Income	%	Children	%
-5	2.0	-2	8.6	-1	34.0
-4	4.9	-1	16.0	***0***	***60.5***
-3	9.6	***0***	***51.6***	1	5.5
-2	15.3	1	14.0	Total	100.0
-1	21.1	2	9.7		
0	***25.0***	Total	100.0		
1	11.6				
2	5.4				
3	3.1				
4	1.3				
5	0.4				
6	0.1				
7	0.1				
Total	100.0				

To evaluate the risk of disclosure for households in A, we use the framework introduced in earlier sections. Table 2.6 and Table 2.7 show the per-record and the global risk of disclosure estimated under various conditions of measurement discrepancies and under-coverage when different sets of household-level attributes are used in the re-identification.

Table 2.6: Per-record and global risks of disclosure for households in the survey data when the re-identification is based on the first set of household variables.

Data 1: Household Size and Children								
Under-coverage	Incorporated				Not-incorporated			
Measurement Discrepancy	Not-Incor.	ρ^1	ρ^2	ρ^3	Not-incor.	ρ^1	ρ^2	ρ^3
Max per-HH risk ×10^{-6}	1.5	1.5	1.5	1.5	0.78	0.78	0.78	0.78
Expected No. of Identified ×10^{-3}	1.3	0.69	0.31	0.088	0.71	0.38	0.17	0.049
Proportion Identified ×10^{-7}	1.7	0.92	0.41	0.128	0.51	0.065	0.065	0.065
Total Prop. of Zero Prob.(%)	54.31	54.31	69.45	90.38	54.31	54.31	69.45	90.38
Non-matched HH (%)	46.33	46.33	46.33	46.33	46.33	46.33	46.33	46.33
Missing Values in B (%)	6.40	6.40	6.40	6.40	6.40	6.40	6.40	6.40
Zero Freq in B (%)	1.58	1.58	1.58	1.58	1.58	1.58	1.58	1.58
Due to MD or UCR (%)	0.00	0.00	15.14	36.07	0.00	0.00	15.14	36.07
Total proportion of households: 26.2%								

Table 2.7: Per-record and global risks of disclosure for households in the survey data when the re-identification is based on the second set of household variables.

Data 2: Household Size, Children and Household Income								
Under-coverage	Incorporated				Not-incorporated			
Measurement Discrepancy	Not-incor.	ρ^1	ρ^2	ρ^3	Not-incor	ρ^1	ρ^2	ρ^3
Max per HH risk ×10^{-6}	8.4	8.4	8.1	8.1	5.7	5.7	4.4	4.4
Expected No. of Identified ×10^{-3}	3.8	2.7	1.2	0.12	2.2	1.5	0.69	0.07
Proportion Identified ×10^{-7}	5.7	4.0	1.8	0.17	3.3	2.3	1.0	0.1
Total Prop. of Zero Prob.(%)	54.69	54.69	61.34	94.63	54.69	54.69	61.34	94.63
Non-matched HH (%)	46.22	46.22	46.22	46.22	46.22	46.22	46.22	46.22
Missing Values in B (%)	6.75	6.75	6.75	6.75	6.75	6.75	6.75	6.75
Zero Freq in B (%)	1.72	1.72	1.72	1.72	1.72	1.72	1.72	1.72
Due to MD or UCR (%)	0.00	0.00	6.65	39.94	0.00	0.00	6.65	39.94
Total proportion of households: 23.2%								

The first column of Table 2.6 displays the risks of disclosure under the assumption that both under-coverage and measurement discrepancy are ignored. The risks of disclosure shown in the 2nd to 4th columns are estimated with measurement discrepancies calculated based on the three measures ρ^1, ρ^2 and ρ^3 as defined in Section 2.3.2 respectively while assuming full coverage. The 5th column contains the estimated risks under the assumption of under-coverage but no measurement discrepancy. The 6th to 8th columns contain the disclosure risk under the same assumptions about the measurement discrepancy as the 2nd to 4th columns respectively, except that we assume under-coverage.

Three measures of disclosure risk are considered in Table 2.6. The first one is the maximum per-record risk of disclosure. The other two measures are both global measures: 1) the expected number of truly identified HHs, which is calculated as the summation of per-record risk across all HHs within the survey data; 2) the proportion of truly identified HHs, which is the mean per-record risk across all HHs within the survey data. Table 2.7 shows the corresponding risks as in Table 2.6 but estimated when all three household specific variables are used in the re-identification.

As we expected, the risk estimated under alternative measures of measurement discrepancy decreases the broader the measure becomes. For example, in Table 2.7, the maximum per-record risk decreases from 8.4×10^{-6} for ρ^1 to 8.1×10^{-6} for ρ^3 when full coverage is assumed. The corresponding reduction is 5.7×10^{-6} for ρ^1 to 4.4×10^{-6} for ρ^3 when under-coverage is assumed. The expected number of identified HHs also changes from 0.69×10^{-3} for ρ^1, 0.31×10^{-3} for ρ^2, to 0.088×10^{-3} for ρ^3. Among these three measures, ρ^1 leads to the most conservative estimation of the disclosure risk because it discriminates the contributions to the measurement discrepancies from each identification variable and from each type of variable. For simplicity, the rest of the comparisons are based on ρ^1. The risk reduction is considerably larger when the other two measures of measurement discrepancies are considered.

Both per-record risk and global risk decreases when measurement imperfectness and under-coverage are appropriately incorporated. Such reduction diminishes as more attributes are used in the re-identification. When only two key attributes, Household Size and Children, are used, the maximum per-record risk is cut in half from 1.5×10^{-6} to 0.78×10^{-6} (estimated using ρ^1). When all three HH-level key attributes are used, the maximum per-record risk is reduced by 1.5 times from 8.4×10^{-6} to 5.7×10^{-6}.

Similar results are found with the global risk measures. For the expected number of identified HHs, the combined effects of measurement discrepancy and under-coverage on the risk reduction is around 3.5 times (from 1.3×10^{-3} to 0.38×10^{-3}) under the condition of Data 1, whereas it is around 2.5 times (from 3.8×10^{-3} to 1.5×10^{-3}) under Data 2. For the proportion of identified HHs, such combined effect is 26 times (from 1.7×10^{-7} to 0.065×10^{-7}) under Data 1 and it is only 2.5 times (from 5.7×10^{-7} to 2.3×10^{-7}) under Data 2.

Furthermore, the magnitude of reduction due to under-coverage tends to be larger than the measurement discrepancy. For the maximum per-record risk, the risk reduction is solely due to under-coverage under both identification conditions. For expected number of identified HHs, 1.4 times reduction is due to the measurement discrepancy, whereas 1.7 times reduction is due to the under-coverage under Data 2. For the proportions of identified HHs, the magnitude of risk reduction due to measurement discrepancy and under-coverage is 1.85 (1.7/0.92) and 3.33 (1.7/0.51) respectively under Data 1. Under the condition of Data 2, the corresponding magnitude is 1.4 (5.7/4.0) and 1.73 (5.7/3.3) respectively.

There are three portions of households in the survey data to whom the disclosure risk is zero: they are households (1) that are not included in the commercial data; (2) that are included in the commercial data but the values of key identifier are missing in the commercial data; and (3) that have combination value(s) in key identifiers that are not present in the commercial data. In addition to the above three reasons, the risk of disclosure may become zero because of either measurement discrepancy or under-coverage rate, such that when either factor takes the value of zero, then the risk is zero. The more information that is available to an intruder, the larger the proportion of households estimated with zero risk of disclosure because of the measurement discrepancy. This may be because more measurement errors are introduced.

If we consider the measurement discrepancy as whether the two measures from the two data sets match, i.e. ρ^3, then the disclosure risk reduction is substantial when the two factors of data quality are both incorporated. The reduction can be as high as 50 times.

In summary, the risk of disclosure for a household decreases when measurement discrepancy and under-coverage are considered. Such reduction diminishes as more attributes are used in the re-identification. The relative magnitude of such reduction tends to be larger from under-coverage than measurement discrepancy.

2.4.2.3. Individual Level Analysis

More than 9,000 individual records in data file B are linked with records in survey A. Table 2.8 shows the percentages of different numbers of household members in B that are linked to one respondent in A. As the main goal of this study is to enlighten the understanding of the measurement discrepancy between the two data files, we feel reluctant to use any social-demographic personal information to attempt the extension of the record-linkage to the individual level. Therefore, for each pair of matched households, we treat all individual pairs between A and B as correct matches.

Table 2.8: The percentages of household members in the commercial data file B in matched households

Number of household members	Percentage
1	22.4
2	34.1
3	20.8
4	12.6
5	6.3
6	3.8
Total	100.0

To investigate whether all survey individuals have the same propensity to be included in the commercial data, we compare the matched and non-matched respondents as shown in Table 2.9. Weights due to unequal sampling selection are incorporated. Item-missing data were excluded. The percentage columns are the proportions of observed sample for matched and non-matched individuals. Based on the results from the t-tests, compared with matched individuals, non-matched individuals tend to be younger, less-educated, single, and non-whites.

Table 2.9: Individual survey characteristics for matched and non-matched survey respondents

	Matched			Non-Matched		
	Mean	SE	%	Mean	SE	%
Age (years)	29.0	0.3	23.6	28.2*	0.3	20.4
Age LT 19 (%)	17.2	1.0	42.5	14.1*	1.0	36.1
Age 19-23 (%)	16.0	1.4	42.5	19.4	1.7	36.1
Age 24-29 (%)	19.1	1.2	42.5	25.4***	1.8	36.1
Age 30-34 (%)	14.8	0.9	42.5	15.2	1.1	36.1
Age 35-49 (%)	32.9	1.5	42.5	25.8***	1.5	36.1
Race						
White (%)	66.9	1.4	23.6	52.6***	1.8	20.4
Black (%)	12.8	1.0	23.6	15.2*	1.1	20.4
Hispanic (%)	14.3	1.0	23.6	23.0***	1.7	20.4
Other Races (%)	6.0	0.9	23.6	9.2***	0.7	20.4
Male (%)	44.1	1.5	23.6	49.0**	1.8	20.4
Education						
LTHS (%)	30.7	1.4	15.4	29.8	1.4	13.2
HS (%)	15.7	0.9	15.4	21.9***	1.4	13.2
Some College (%)	27.0	1.5	15.4	26.9	1.9	13.2
College (%)	17.7	1.1	15.4	14.4*	1.4	13.2
GRAD/PROF (%)	8.8	1.1	15.4	6.9	1.0	13.2
Married (%)	40.6	1.6	15.4	28.9***	1.5	13.3

Note: Significance levels for two-sided t-tests between matched and non-matched survey respondents: $^{*}\ a = 0.10$; $^{**}\ a = 0.05$; $^{***}\ a = 0.01$.

We simulate four scenarios, namely Data 1-4, where an intruder has access to different amount of identification information as shown in Table 2.2. Table 2.10 describes the estimated coverage rates under these four scenarios. There are five categories in the key identifier when only Age is used in the re-identification. The coverage rates are very similar among each age category. As more variables are used in the re-identification, the number of categories increases and so does the variation in coverage rates across categories. On average, the mean coverage rate is about 50% under all four scenarios.

Table 2.10: Summary statistics of coverage-rate, $c_k M_k^{-1}$, for four sets of individual-specific key variables

	No. of Categories	Min.	Mean	Max.	SD
Data 1	5	0.488	0.530	0.563	0.032
Data 2	40	0.280	0.485	0.639	0.092
Data 3	1574	0.000	0.494	1.000	0.369
Data 4	2385	0.000	0.524	1.000	0.412

According to statistics presented in Table 2.8, 22.4% households in B are one-person households. The one-to-one pair of individuals in a one-person household is more likely to produce a correct match. We use this sub-dataset to illustrate the magnitude of the measurement discrepancy on several key variables as shown in Table 2.11. The most inconsistent measure is Marital Status, where only 16% of reports agree. The consistency rates for Age, Education and Gender are around 19%, 32% and 37% respectively. The most consistent measure is Race with a 68% agreement rate. The discrepancy goes in both directions for all variables.

Table 2.11: Measurement discrepancies on the individual-level attributes between A and B for matched one-person households

Age (7 levels)	%	Race (4 levels)	%	Gender (2 levels)	%	Educ. (5 levels)	%	Marital (2 levels)	%
-3	0.2	-3	2.5	***0***	***37.1***	-4	0.2	***0***	***16.4***
-2	0.9	-2	5.2	1	***62.9***	-3	3.6	1	83.6
-1	3.1	-1	10.2	Total	100.0	-2	9.3	Total	100.0
0	***18.4***	***0***	***67.9***			-1	20.7		
1	10.6	1	4.1			***0***	***32.4***		
2	9.8	2	4.2			1	19.8		
3	7.3	3	6.0			2	10.2		
4	17.5	Total	100.0			3	3.4		
5	17.3					4	0.4		
6	14.7					Total	100.0		
Total	100.0								

Since all matched individuals will be used when we evaluate the disclosure risk, we present the percentage of each discrepancy value calculated based on the entire matched sample for all eight key variables separately in Figure 2.1. A zero value on the x-axis means that there is no measurement discrepancy and the measures from the two data sources agree. Consistent with the results in Table 2.11, the most accurate measure is Race followed by Education and Household Income. The least accurate variable is Age and Marital Status.

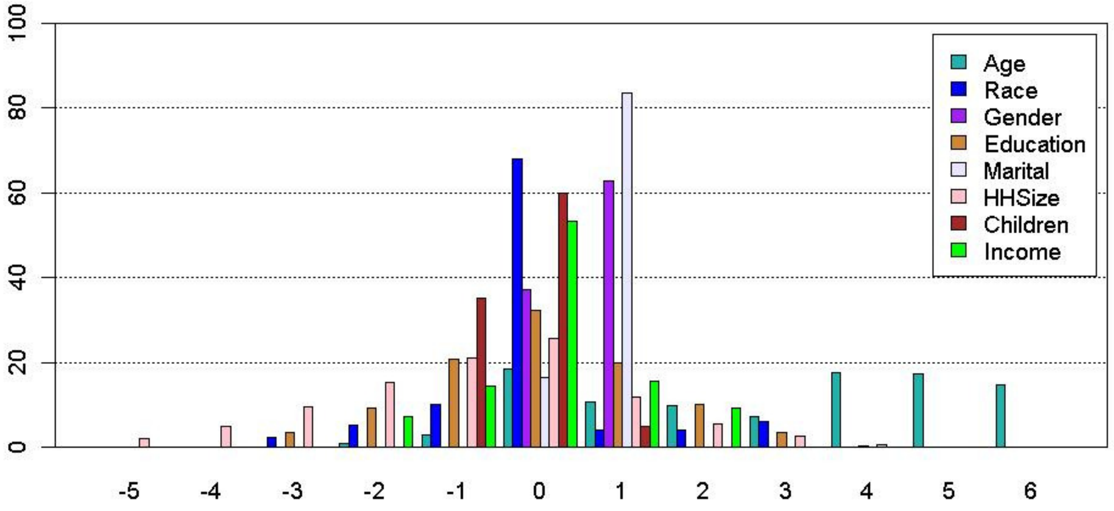


Figure 2.1: Percentages of measurement discrepancies between the values of key variables in the commercial data and the survey data on matched individuals.

The per-individual and global risks of disclosure under each condition of the intruder's information are presented in Table 2.12-15 respectively. The eight columns in these tables are defined in the same way as in Table 2.6 of the household-level analysis in Section 2.4.2.2. The three measures of disclosure risk are still 1) the maximum per-record risk of disclosure, 2) the expected number of disclosed individuals, and 3) the proportion of disclosed individuals.

A total of five groups of individuals are free of disclosure risk. In addition to the four groups due to the same reasons as those we discussed in household-level risk assessments, such as 1) under-covered individuals; 2) covered individuals but with missing values in key identifiers; 3) individuals with combination values in key identifiers that are not present in the commercial data, and 4) individuals for whom the measurement discrepancy probability and/or the coverage probability are estimated as zero, teenagers who are 19 years old or younger are also assigned zero risks. The reason is that the commercial data only compiles information about adults.

Table 2.12: Individual and global risks of disclosure for individuals in the survey data based on the first set of identification variables

Data 1: Age								
Under-coverage	Incorporated				Not-incorporated			
Measurement Discrepancy	Not-incor.	ρ^1	ρ^2	ρ^3	Not-incor.	ρ^1	ρ^2	ρ^3
Max per Ind. risk $\times 10^{-7}$	1.0	1.0	1.0	1.0	0.49	0.49	0.49	0.49
Exp. No. of Identified $\times 10^{-4}$	4.6	3.3	1.2	1.2	2.4	1.7	0.62	0.62
Proportion Identified $\times 10^{-8}$	1.5	1.1	0.39	0.39	0.77	0.55	0.2	0.20
Total Prop. of Zero Prob.(%)	70.9	70.9	88.1	88.1	70.9	70.9	88.1	88.1
Non-matched Ind. (%)	31.1	31.1	31.1	31.1	31.1	31.1	31.1	31.1
Missing Values in B (%)	4.2	4.2	4.2	4.2	4.2	4.2	4.2	4.2
Teens under-covered (%)	35.6	35.6	35.6	35.6	35.6	35.6	35.6	35.6
Zero Freq in B (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Due to MD or UCR (%)	0.0	0.0	17.2	17.2	0.0	0.0	17.2	17.2
Total number of Categories: 5; Sample Unique Categories: 0								

Table 2.13: Individual and global risks of disclosure for individuals in the survey data based on the second set of identification variables

Data 2: Age, Race and Gender								
Under-coverage	Incorporated				Not-incorporated			
Measurement Discrepancy	Not-incor.	ρ^1	ρ^2	ρ^3	Not-incor.	ρ^1	ρ^2	ρ^3
Max per Ind. risk $\times 10^{-6}$	4.5	4.5	4.5	4.5	2.3	2.3	2.3	2.3
Exp. No. of Identified $\times 10^{-3}$	5.2	3.7	2.4	0.44	2.4	1.7	1.1	0.22
Proportion Identified $\times 10^{-7}$	3.1	2.2	1.4	0.26	1.4	1.0	0.67	0.13
Total Prop. of Zero Prob.(%)	49.4	49.4	52.9	90.2	49.4	49.4	52.9	90.2
Non-matched HH (%)	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6
Missing Values in B (%)	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3
Teens under-covered (%)	11.8	11.8	11.8	11.8	11.8	11.8	11.8	11.8
Zero Freq in B (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Due to MD or UCR (%)	1.7	1.7	5.2	42.5	1.7	1.7	5.2	42.5
Total number of Categories: 40; Sample Unique Categories: 0								

Table 2.14: Individual and global risks of disclosure for individuals in the survey data based on the third set of identification variables

Data 3: Age, Race, Gender, Education, Marital, HH Size and HH Children								
Under-coverage	Incorporated				Not-incorporated			
Measurement Discrepancy	Not-incor.	ρ^1	ρ^2	ρ^3	Not-incor.	ρ^1	ρ^2	ρ^3
Max per Ind. risk $\times10^{-4}$	9.8	9.0	5.6	0.49	8.3	7.0	4.7	0.44
Exp. No. of Identified $\times10^{-2}$	12.0	9.6	5.6	0.045	8.0	6.4	3.7	0.027
Proportion Identified $\times10^{-6}$	11.0	8.7	5.1	0.041	7.3	5.8	3.4	0.025
Total Prop. of Zero Prob.(%)	81.4	81.4	81.5	99.5	81.4	81.4	81.5	99.5
Non-matched HH (%)	31.5	31.5	31.5	31.5	31.5	31.5	31.5	31.5
Missing Values in B (%)	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4
Teens under-covered (%)	13.3	13.3	13.3	13.3	13.3	13.3	13.3	13.3
Zero Freq in B (%)	22.8	22.8	22.8	22.8	22.8	22.8	22.8	22.8
Due to MD or UCR (%)	9.4	9.4	9.5	27.5	9.4	9.4	9.5	27.5
Total number of Categories: 1574; Sample Unique Categories: 539								

Table 2.15: Individual and global risks of disclosure for individuals in the survey data based on the fourth set of identification variables

Data 4: Age, Race, Gender, Education, Marital, HH Size, HH Children and HH Income								
Under-coverage	Incorporated				Not-incorporated			
Measurement Discrepancy	Not-incor.	ρ^1	ρ^2	ρ^3	Not-incor.	ρ^1	ρ^2	ρ^3
Max per Ind. risk $\times10^{-4}$	16	14	10	0.62	11	9.1	6.7	0.57
Exp. No. of Identified $\times10^{-2}$	12	9.9	6.0	0.041	8.9	7.3	4.4	0.026
Proportion Identified $\times10^{-6}$	12	10	6.1	0.042	9.1	7.5	4.5	0.027
Total Prop. of Zero Prob.(%)	84.0	84.0	84.0	99.7	84.0	84.0	84.0	99.7
Non-matched HH (%)	31.6	31.6	31.6	31.6	31.6	31.6	31.6	31.6
Missing Values in B (%)	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
Teens under-covered (%)	10.7	10.7	10.7	10.7	10.7	10.7	10.7	10.7
Zero Freq in B (%)	29.2	29.2	29.2	29.2	29.2	29.2	29.2	29.2
Due to MD or UCR (%)	8.0	8.0	8.0	23.7	8.0	8.0	8.0	23.7
Total number of Categories: 2385; Sample Unique Categories: 1227								

Similar results about the impacts of measurement discrepancy, under-coverage, and amount of information that an intruder has about the potential victims, on the risk of disclosure are found with the risk of disclosure for survey individuals as for survey households.

In specific, the risk of disclosure increases as more information is available to an intruder. When both measurement discrepancy and under-coverage are ignored, the maximum per-individual risk increases from 1.0×10^{-7} under Data 1, 4.5×10^{-6} under Data 2, 9.8×10^{-4} under Data 3, to 1.6×10^{-3} under Data 4. Similarly, the expected number of identified individuals rises from 4.6×10^{-4} under Data 1, 5.2×10^{-3} under Data 2, to 1.2×10^{-1} under both Data 3 and Data 4. Same pattern is found with the proportion of identified individuals.

Similar to the household-level analysis, both the per-individual risk and the global risk diminish when the measurement discrepancy and under-coverage are incorporated. The collective effect on risk reduction becomes less significant as more individual attributes are used in the re-identification. When only Age is used, the maximum per-individual risk is reduced by 2 times, which becomes 1.9, 1.4, and 1.76 when more variables are involved. For both global measures, the expected number of disclosed individuals and the proportions of disclosed individuals, the risk reduction decreases from 2.7 (Data 1), 3.5 (Data 2), 1.8 (Data 3), to 1.6 (Data 4).

However, for the maximum per-individual risk, the change in risk reduction from Data 3 (1.4) to Data 4 (1.76) turns out to be increasing instead of decreasing. Similarly, for both global risks, the change in risk reduction from Data 1 (2.7) to Data 2 (3.5) is also increasing. A speculation about this irregularity can be that there may exist a systematic change in the data composition between the sub data sets because each data set is composed of different survey individuals due to item missing data, thus the results are distorted by the potential “selection bias”. Another possible reason is that our assumption that each household member is equally likely to be the true match of the survey respondent in matched households may not hold. Therefore, the measurement discrepancy may be overly overestimated, thus confounding the results. However, these two speculations are unverifiable in the current study.

Figure 2.2-5 are the plots of the average estimated record-level disclosure risk when both under-coverage and measurement discrepancy are present, by the combination values of the key identifier, with each figure reporting a key identifier defined from a different set of key variables. A close examination of these plots suggests the record-level disclosure risks vary greatly across different combination values of key identifiers, and this finding holds for all conditions on the amount of common key variables that an intruder may have access to. Such variation is not solely due to the unbalanced population distribution in key identifiers, but also partially because of variations in measurement discrepancy and/or under-coverage rate across the key identifier categories.

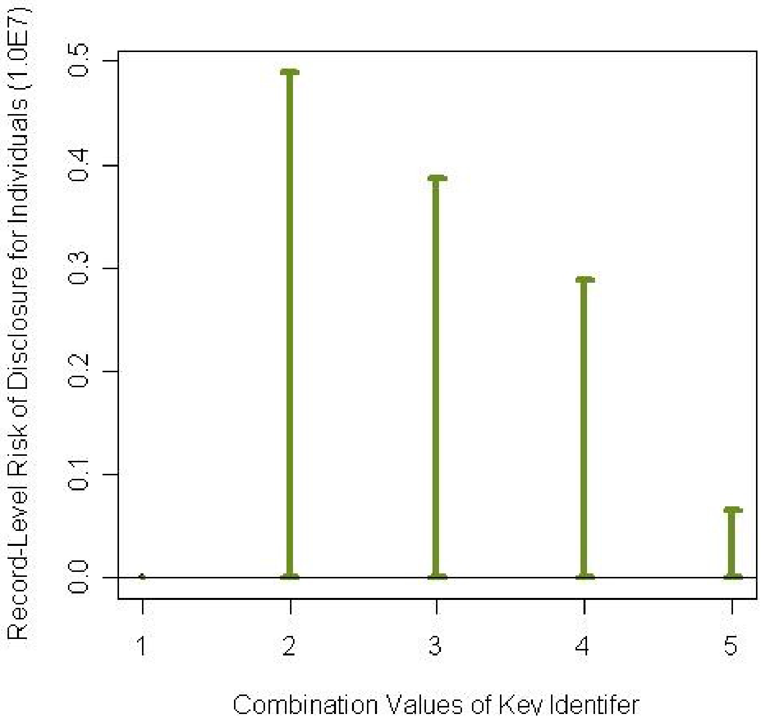


Figure 2.2: Histogram of record-level disclosure risk for individuals based on key identifier in Data 1.

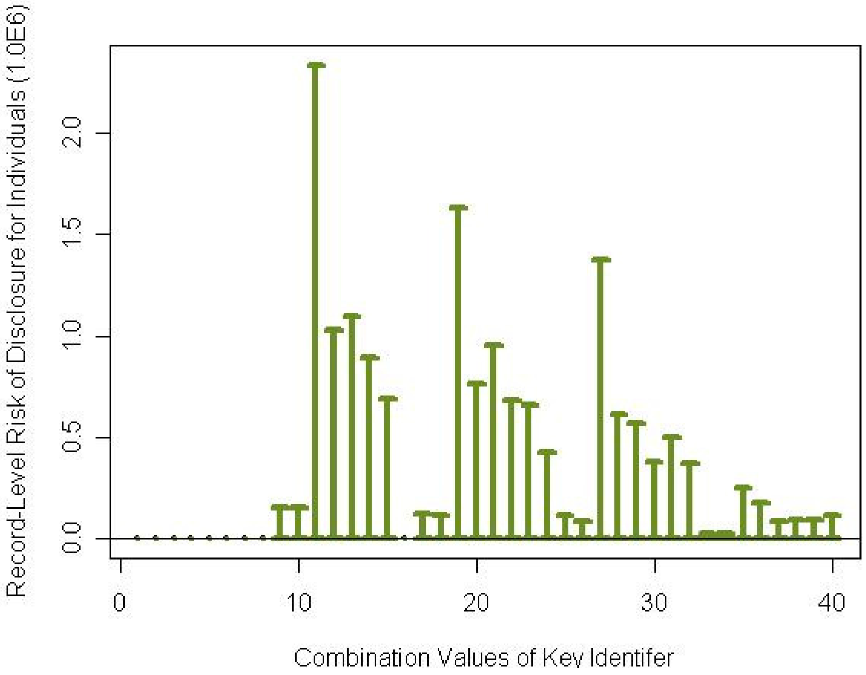


Figure 2.3: Histogram of record-level disclosure risk for individuals based on key identifier in Data 2.

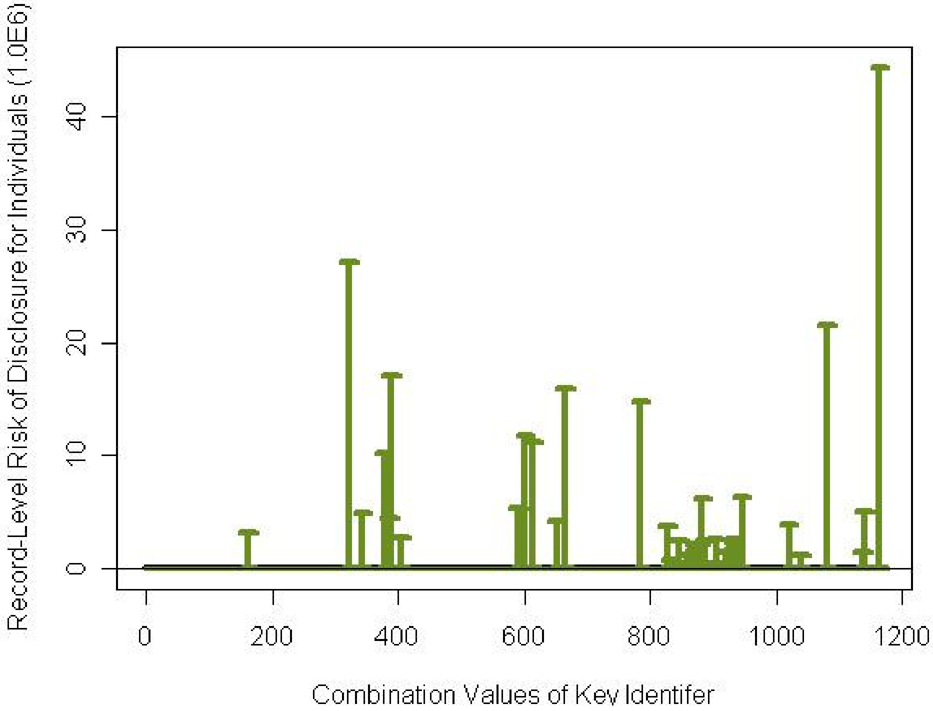


Figure 2.4: Histogram of record-level disclosure risk for individuals based on key identifier in Data 3.

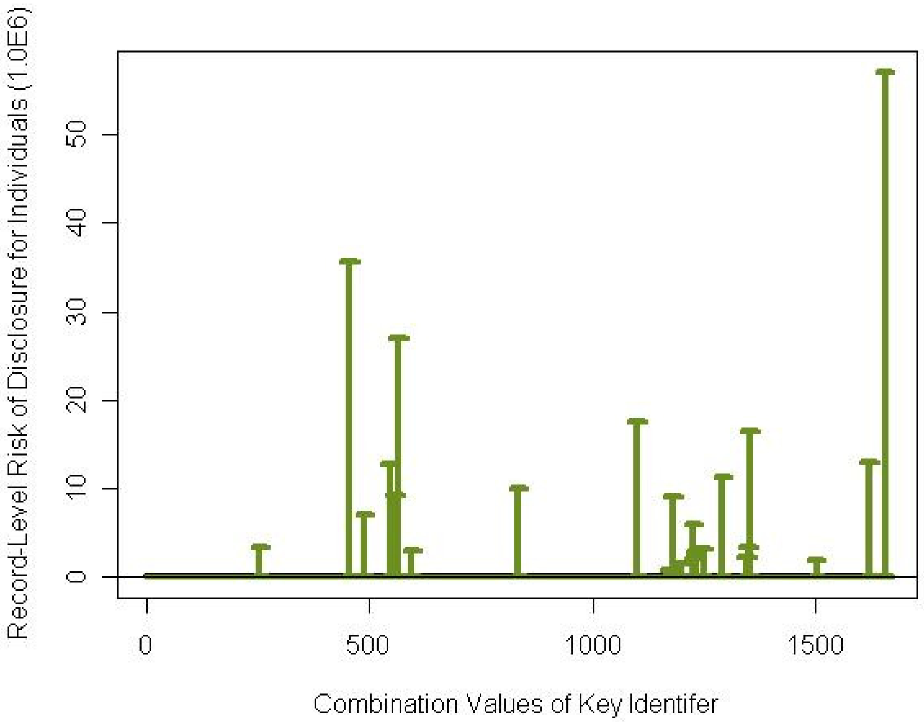


Figure 2.5: Histogram of record-level disclosure risk for individuals based on key identifier in Data 4.

2.5. Conclusion and Discussion

This study addresses and tests three aspects of uncertainty about measuring and assessing the risk of disclosure when an intruder attempts to re-identify survey respondents in microdata by linking with records in an external data file based on common key variables.

The two main aspects are the measurement discrepancy in the values of common key variables between a pair of records that refer to the same entity within the target survey data and external commercial data respectively, and the under-coverage of the commercial data. Both aspects are largely ignored in the SDC literature. We have discussed and illustrated, theoretically and numerically, that ignoring these errors results in disclosure risk being over estimated. The magnitude of this reduction has shown to be as little as 2 times or as large as more than 40 times. In either case is too large to be treated as ignorable. We also find that although the majority of such influence comes from under-coverage, the impact from imperfect measurement is still enormous. This line of research result is consistent with our hypotheses and provides evidence about important assumptions that future research on risk assessments should take into account.

This finding is very important to federal statistical agencies and other data disseminators for several reasons. On the one hand, overstatement of disclosure risk may deter agencies from disseminating the data set that can be of great benefit to the public. On the other hand, if agencies have decided to publish the information contained in the data set, they may choose to disseminate the information which they consider to be “safe” or use a dissemination product that ensures that safety. For example, the agencies may preclude the risky information from the microdata, or publish summary statistics, such as tables, instead of the microdata. Such a decision process will delay the dissemination at the very least and data utility may be sacrificed as well. Lastly, when the over-estimated per-record risk is used to decide the records on which the information needs to be modified, a significantly larger portion of records will be selected, which leads to more information loss due to the modification procedure. In sum, these two factors should be incorporated in risk assessments.

Specific for the effect from the measurement discrepancies, extensive efforts in the literature of survey methodology have been made to minimize measurement error on survey attributes from social and psychological viewpoints during data collection period. The main goal of such research is to produce high quality data to support research and policymaking. Similarly, the commercial data are compiled with the goals of achieving high data quality and complete coverage. However, the presence of such errors lead to reduced risk of disclosing information about survey respondents, which helps relieve the concerns about data confidentiality from the data disseminators. The two roles that measurement error play are both important but contradictory to each other. The contradiction between statistical inference validity and disclosure control can be resolved by merging well-understood measurement error into actual data. Properly designed and implemented, measurement error with known statistical properties can be created during the data collection procedure. Alternatively and most realistically, we can add random error or noise to the original values of identification information in a systematic manner during the stage of post-data collection, as in most statistical disclosure control methods.

The other aspects of intruding assumptions involve the amount of information that an intruder possesses about the known victims. It is not surprising that the more information that is known about the potential victims, the more likely they are to be identified, as suggested by the results. We also found that the effect of data quality on disclosure risk diminishes, as more information is available to an intruder.

Unfortunately, this study has two limitations. First, the definition of household may be different across the two data files, which may lead to the large discrepancy in the household size. Furthermore, there is no unique identifier at the individual level, such as social security numbers. Therefore, our assumption of equal chances for all household members in the commercial data to be the correct match for the survey respondent may not hold. Because an intruder may use such information to make subjective judgments about which household member is the survey respondent, further exploration of the usage of household decomposition information should be conducted. Second, there may be a time gap between when the survey is implemented and when the commercial data is compiled. Therefore, the time-sensitive variables such as age and household composition may not be comparable. We may be over estimating the impact from measurement discrepancy if such inconsistency is not resolved.

CHAPTER III

SEMI-PARAMETRIC MULTIPLE IMPUTATIONS OF FULLY-SYNTHETIC DATA

3.1. Introduction

Statistical agencies face the inherent tension between the increasing demand for publicly accessible data and growing concerns about confidentiality. Common disclosure avoidance practices involve perturbing the values in the actual data set. One such approach, proposed by Rubin (1993) and Little (1993), further developed by Raghunathan et al. (2003) and Reiter (2005a), is releasing the multiple fully-imputed synthetic data in place of the actual data. This approach has several advantages over alternative statistical perturbation methods (Winkler 2004, Reiter 2005a). First, it allows the researchers to completely eliminate the risk of disclosing respondents' identities or sensitive attributes as no real information is disseminated. Second, valid statistical inference can be made by analyzing the synthetic data using standard statistical packages and similar statistical inference can be achieved using the synthetic data and the actual data.

In general, the synthetic data are generated from models of the actual data and can be viewed as imputations (or predictions) of data values for a new sample drawn from the same population as the actual sample. In order to preserve statistical properties of actual data, the imputation models have to be carefully chosen after a thorough analysis of the complex relationships among survey variables. This task is challenging because typically surveys collect data on hundreds of variables whose distributions often cannot be adequately captured with standard parametric models. Therefore, it is preferable to use semi-parametric methods, which relax the distributional assumptions of the parametric models, thus protecting against model misspecification. Two alternative synthetic data approaches, partial (Reiter 2005b) and selective synthesis (Little and Liu 2002), in which values of selected variables and/or selected individuals are replaced by imputations, have been shown to be considerably less model-dependent but offer weaker confidentiality protection.

Multivariate relationships are preserved when imputations are drawn from the joint posterior predictive distributions of the unobserved data. However, it is difficult to specify such a joint distribution model for a large number of variables of different types. In response to this challenge, a multivariate sequential regression approach is first proposed by Raghunathan et al. (2001) and widely used in missing data imputation. A regression model, for instance, a linear, logistic, Poisson or polynomial regression depending on the type of the variable, is fitted for one variable at a time conditional on the remaining variables. The joint distribution is then approximated by sequentially fitting the conditional models for multiple rounds. The imputations created using this approach are statistically comparable to those obtained by the Bayesian method, in which joint distributions are modeled explicitly (Raghunathan, Lepkowski, van Hoewyk and Solenberger 2001) and the computation may be intense when the number of variables gets large.

This chapter presents and evaluates a sequential regression-based semi-parametric approach for constructing fully-synthetic data for a large national survey. We use different semi-parametric models to create synthetic data for each type of variables; specifically, Alternating Conditional Expectations (ACE) models (Breiman and

Friedman 1985) for continuous variables, Ridge-penalized logistic regression models (Schaefer 1986, Cessie and Houwelingen 1992) for binary variables, and sequential nested ridge-penalized logistic regression models for multinomial categorical variables with more than two levels.

Most existing imputation models for continuous variables are built on the normality assumption. However, the distributions of variables we encounter in real-data applications often deviate from normality. Transformations can be used to improve model fit in the presence of non-normality. A class of parametric transformation techniques with an emphasis on transforming the response variable has been suggested (Box and Cox 1964). However, such parametric transformation methods are still susceptible to failing to meet *a priori* assumptions about the functional forms that relate the response variable and covariates, thus they may lead to model misspecification.

The ACE method (Breiman, et al. 1985) makes minimal assumptions about the data distribution and functional forms by estimating the optimal nonlinear transformations for both the response and predictor variables in a multiple regression model. This algorithm aims to maximize the correlations between variables and normalize the unexplained errors. This method also allows the complex relationships among variables to be revealed more accurately, thus improving predictions.

For binary data, the predictions based on logistic regressions can become very unstable due to sparse data. Such data sparseness is very common in surveys and can occur in any of the three situations: (1) when the binary response variable is skew-distributed, for example, measures with low population prevalence, such as rare disease or severe health condition, (2) when the number of covariates is relatively large, and (3) when the covariates are highly correlated. Ridge-penalized logistic regression resolves this ill-conditioning problem by estimating the regression parameters using a shrinkage technique. The shrinkage estimators are consistent, and asymptotically more efficient than traditional alternatives, thus potentially improving the predictions. The amount of shrinkage can be determined by the data, which precludes the uncertainty associated with the estimation due to subjectivity.

The success of this fully-synthetic data approach critically depends upon demonstrating whether the synthetic data inferences are valid. Raghunathan et al. (2003) evaluates such “validity” from a repeated sampling perspective. However, such evaluations are too computational burdensome to be practical or realistic. Furthermore, it is more meaningful to assess whether a particular set of synthetic data sets produce similar results as the actual data set, because eventually only a small number of synthetic data are disseminated for public use. Therefore, diagnostic tools are needed to assess such similarities in terms of the distributional properties and statistical inferences for a set of pre-specified analyses.

In this chapter, we evaluate this fully-synthetic data approach using the data from the Health and Retirement Study conducted by the University of Michigan. The rest of this chapter is organized into six sections. In Section 3.2, we review the notations and the inference method for multiple fully-synthetic data. Then we describe the imputation algorithms for generating the synthetic data based on the sequential regression technique in Section 3.3. Separate regression models, specifically, ACE for continuous variables, Ridge logistic for binary data and hierarchical ridge logistic for polynomial variables, are used to model the conditional distributions. We introduce a diagnostic tool for synthetic data in Section 3.4. After the description of the data structure in Section 3.5, we introduce three information loss matrices, and compare the statistical inferences from the synthetic data with those from the actual data in Section 3.6. Finally, Section 3.7 concludes this chapter with discussion.

3.2. Methods

3.2.1. Synthetic data

Suppose the actual data is a sample of size n from a finite population $\mathrm{P} = \left(Z, Y\right)$ of size N, by a given sample design and recruitment protocol. $Z = \left(Z_i, i = 1, 2, ..., N\right)$ denotes design variables available on all population units. $Y = \left(Y_i, i = 1, 2, ..., N\right)$ represents survey variables of interests, which can be decomposed into $Y = \left(Y_{inc}, Y_{exc}\right)$, where $Y_{inc} = \left(Y_i, i = 1, 2, ..., n\right)$ is the sampled portion of Y, and $Y_{exc} = \left(Y_i, i = n+1, ..., N\right)$ is the non-sampled portion of Y. Y_{inc} may include data values on units who (1) respond to survey requests and provide answers to questions that measure Y, (2) respond to survey requests but fail to answer this particular question that Y pertains to, and the resulting loss of data is called Item-nonresponse, and (3) refuse to participate in the survey, and the subsequent loss of cases is called Unit-nonresponse. For simplicity, we assume that there is no unit-nonresponse, and all missing data are due to item-nonresponse and they are missing at random (MAR). We first describe the formulations of generating synthetic data as if there were no item-missing data, and then we extend this framework to cope with item-missing imputation and full synthesis simultaneously.

We study the case where statistical agencies seek to release multiple fully-synthetic data $D_l = \left\{(Z, Y_{l.inc}), l = 1, ..., L\right\}$ built on the actual survey data $D = (Z, Y_{inc})$. Each synthetic data $D_l = (Z, Y_{l.inc})$ of size n, is achieved by (1) filling in the unobserved values of Y by draws from the posterior predictive distribution of $\left(Y \mid Z, Y_{inc}\right)$, thus completing a synthetic population, and then (2) drawing a simple random sample from this synthetic population. We repeat this two-step process independently L times to obtain L synthetic samples, which are released for public use. In practice, the first step of generating complete synthetic population data is unnecessary and we only need to generate values of Y for synthetic samples.

When item-missing data is present, the synthesis approach then involves a two-stage imputation. The sampled portion of data Y_{inc} breaks down into two components $Y_{inc} = (Y_{mis}, Y_{obs})$, where Y_{mis} is the portion of item-missing data. In the first stage, the missing data Y_{mis} is multiply-imputed with draws from the posterior predictive distribution $(Y_{mis} \mid Z, Y_{obs})$ (Raghunathan, et al. 2001) to complete the sampled rectangular data. We repeat these draws independently M times to obtain M complete data sets $D^M = \left\{D^{(m)}, m = 1, 2, ..., M\right\}$, where $D^{(m)} = \left(Z, Y_{inc}^{(m)}\right)$ and $Y_{inc}^{(m)} = \left(Y_{mis}^{(m)}, Y_{obs}\right)$.

In the second-stage, conditional on each complete data $D^{(m)}$, L synthetic samples $D_l^{(m)} = \left\{Z, Y_{l.inc}^{(m)}, l = 1, 2, ..., L\right\}$ are generated independently as there were no item-missing data. Therefore, we obtain a total of $M \times L$ fully-synthetic datasets which are to be released for public use. We assume there is no confidentiality concern over releasing information about Z. If Z is also subject to confidentiality constraint, the above setup can be conveniently adapted to synthesize Z.

The number of multiple imputations (or syntheses), (i.e. M and L), is chosen based on the fraction of missing data (Rubin, 1987; Little and Rubin, 2002), the desired accuracy for synthetic data inference and the risk of disclosing identities and/or sensitive attributes of survey respondents (Reiter and Drechsler, 2007). The last factor becomes irrelevant under the assumption that the claim from statistical agencies that the public data contains no real information about any survey respondents is effective in deterring any potential intruders.

To account for uncertainty for a small fraction of missing information due to the imputation for item-nonresponses, $M = 5$ is usually sufficient to achieve satisfactory precision in statistical inference (Rubin 1987). Theoretically, a larger L is needed for creating fully-synthetic samples because the fraction of missing information is one-hundred percent. A modest number of fully-synthetic data sets, for instance, 5, 10 or 20, is still sufficient to ensure the inference validity (Raghunathan et. al, 2003).

3.2.2. Multiple fully-synthetic data inference

Item-missing data are ubiquitous in surveys. It is logical to impute for item-missing data and build fully-synthetic data simultaneously. However, neither the multiple fully-synthetic data inference (Raghunathan et. al, 2003) nor the standard multiple imputation inference (Rubin 1987) by themselves will result in valid inference because both inference methods ignore the fact that there are two separate sources of variability due to imputing item-missing data and synthesizing the entire data set. Reiter (2004) developed a combination rule to ensure valid inference for partial synthetic data where item-missing data are handled prior to the synthesis procedure. However, this inference rule is unsatisfactory when fully-synthetic data are created. Therefore, a new combination rule is needed. Before presenting this new rule, we first review multiple imputation inference for missing data and multiple fully-synthetic data inference.

3.2.2.1. Single-stage multiple imputation inference

For each imputed complete data $D^{(m)}, m = 1, 2, ..., M$, a scalar estimand Q which may be a function of (Z, Y) is estimated with a point estimate $q^{(m)}$ and associated variance estimate $v^{(m)}$. Under the assumptions described in Rubin (1987), Q can be estimated by

$$\bar{q}_m = M^{-1} \sum\nolimits_{m=1}^{M} q^{(m)} \qquad [3.1]$$

with variance

$$T_m = \left(1 + M^{-1}\right) b_m + \bar{v}_m \qquad [3.2],$$

where $b_m = (M-1)^{-1} \sum\nolimits_{m=1}^{M} \left(q^{(m)} - \bar{q}_m\right)^2$ is the between imputation variance and $\bar{v}_m = M^{-1} \sum\nolimits_{m=1}^{M} v^{(m)}$ is the within imputation variance.

Developed under a Bayesian framework, when M is small to modest, for example 5 to 10, the posterior distribution of Q can be approximated by a Student's t-distribution with degrees of freedom $\gamma_m = (M-1)\left(\left(1+M^{-1}\right) b_m T_m^{-1}\right)^{-2}$. The degrees of freedom, γ_m, reflect the statistical uncertainty due to missing data. When a quantity being estimated is strongly influenced by missing data, its inference then is based on a smaller value of γ_m, thus a wider confidence interval.

3.2.2.2. Single-stage full synthesis inference

Raghunathan et. al (2003) presented the combining rule for fully-synthetic data. Suppose there are no missing data and fully-synthetic data $D_l, l = 1, 2, ..., L$ are generated conditional on D. For each D_l, Q is estimated with a point estimate q_l with variance v_l. The multiple synthetic data estimate of Q can be expressed as

$$\bar{q}_L = \sum\nolimits_{l=1}^{L} q_l \big/ L \qquad [3.3]$$

with variance

$$T_L = \left(1 + L^{-1}\right) b_L - \bar{v}_L \qquad [3.4],$$

where $b_L = \sum_{l=1}^{L} (q_l - \bar{q}_L)^2 \big/ (L-1)$ is the between synthesis variance and $\bar{v}_L = \sum_{l=1}^{L} v_l \big/ L$ is the within synthesis variance. Because there is an additional level of sampling of drawing synthetic samples from each synthetic population, the between synthesis variance already reflects the within synthesis variability. Therefore, T_L is defined as the between imputation variance minus within imputation variance, which is different from the one in standard multiple imputation inference. The posterior distribution of Q given fully-synthetic samples is approximated by a normal distribution with mean $\bar{q}_L$ and variance T_L. The Bayesian confidence intervals are shown to be identical to the large sample frequentist intervals (Raghunathan et. al, 2003).

3.2.2.3. Two-stage item-missing data imputed fully-synthetic data inference

In presence of item-missing data, we apply a two-stage fully-synthetic data approach. The goal becomes making inference of Q based on D_L^M . Let $q_l^{(m)}$ and $v_l^{(m)}$ be the point estimates and associated variance estimates based on $D_l^{(m)}$, the l_{th} synthetic data generated from the m_{th} complete data, where $l = 1,...,L$ and $m = 1,...,M$. Assuming Bayesian asymptotic conditions meet, Q can be estimated by

$$\bar{q}_M = (ML)^{-1} \sum_{m=1}^{M} \sum_{l=1}^{L} q_l^{(m)} \qquad [3.5]$$

with variance

$$T_M = \left(1 + M^{-1}\right) B_M + \left(1 + L^{-1}\right) \bar{b}_M - \bar{v}_M \qquad [3.6],$$

where $\bar{b}_M = \left[M (L-1) \right]^{-1} \sum_{m=1}^{M} \sum_{l=1}^{L} \left(q_l^{(m)} - \bar{q}^{(m)} \right)^2$, $B_M = (M-1)^{-1} \sum_{m=1}^{M} \left(\bar{q}^{(m)} - \bar{q}_M \right)^2$, and $\bar{v}_M = (ML)^{-1} \sum_{m=1}^{M} \sum_{l=1}^{L} v_l^{(m)}$. $\bar{q}_M$ and $\bar{v}_M$ are the overall means of point estimates and estimated variances across all synthetic data. B_M is the variance of $\bar{q}^{(m)}$ across all complete data. When n, M and L are large, the inference can be approximated by normal distribution, thus the 95% confidence interval can be computed as $\left[\bar{q}_M - z_{0.975}\sqrt{T_M}, \bar{q}_M + z_{0.975}\sqrt{T_M}\right]$. For small to modest $n,\ M$ and L, the inference about Q is made by approximating its posterior distribution by a t distribution with degrees of freedom $\gamma_M = (M-1)\left(\left(1+M^{-1}\right)B_M T_M^{-1}\right)^{-2}$.

One drawback about this variance estimator is that the variance estimates can be negative, which is shared by the inference method about single stage fully-synthetic data. This problem can be avoided by increasing $n,\ M$ and/or L, however, the computational requirements soon became practically unrealistic. We provide a modified variance estimator to accommodate this situation in that

$$T_M^* = T_M + \lambda \bar{v}_M \qquad [3.7],$$

where $\lambda = I\left(T_M \le 0\right)$. Note that negative T_M happens when within synthetic data variation dominates the total variance. What this modification does is replace the total variance with the summation of the estimated between missing data imputation variance and between synthesis variance. Note that the within synthetic data variance is already included in the between synthetic data variance. This new estimator ensures the estimated variance is always positive and somewhat slightly conservative.

In addition, the estimated degrees of freedom, γ_M, can be very small when within synthetic data variation is large. As a result, the confidence intervals tend to be excessively wide, which responds to overly conservative inferences. We use the adjusted estimates for the degrees of freedom similarly as described in Reiter and Drechsler (2007) as

$$\gamma_M^* = \max\left(M-1, \gamma_M\right) \qquad [3.8]$$

3.2.3. Theoretical justification for the two-stage inference

We built this new combining rule based on the two existing rules developed by Rubin (1987) and Raghunathan, Reiter, and Rubin (2003). We also assume all the conditions specified in both articles meet.

3.2.3.1. First-stage inference: $f\left(Q \mid D^M\right)$

The first stage of this synthetic data approach creates multiply-imputed item-missing data $D^M = \left(D^{(m)}, m = 1, 2, ..., M\right)$. Let $\bar{q}_m$, b_m, $\bar{v}_m$, γ_m and $\left\{\left(q^{(m)}, v^{(m)}\right), m = 1, 2, ..., M\right\}$ be defined as in Section 3.2.2.1. By Rubin (1987), the inference of Q based on D^M can be approximated by a t-distribution with posterior mean $\bar{q}_m$, posterior variance T_m and degrees of freedom γ_m.

3.2.3.2. Second-stage approximations

In the second stage, L fully-synthetic data $D_L^{(m)} = \left(D_l^{(m)}, l = 1, 2, ..., L\right)$ are generated from each complete data $D^{(m)}$ separately. We treat each imputed sample $D^{(m)}, m = 1, 2, ..., M$ as if it was an actual sample from the population. Then for each $D^{(m)}$, a frequentist could make inference of Q on $D^{(m)}$ with unbiased estimate $q^{(m)}$ and sampling variance $v^{(m)}$. Using fully-synthetic data inference results in Section 3.2.2.2, the equivalent Bayesian inference of Q based on $D_L^{(m)}$ can be approximated by the normal distribution with the posterior mean $\bar{q}^{(m)} = L^{-1}\sum_{l=1}^{L} q_l^{(m)}$ and posterior variance $T^{(m)} = \left(1 + L^{-1}\right) b_L^{(m)} - \bar{v}_L^{(m)}$, where $b_L^{(m)} = \sum_{l=1}^{L}\left(q_l^{(m)} - \bar{q}^{(m)}\right)^2 \Big/ \left(L - 1\right)$ and $\bar{v}_L^{(m)} = \sum_{l=1}^{L} v_l^{(m)} \Big/ L$. From randomization validation point of view, $\bar{q}^{(m)}$ and $T^{(m)}$ are unbiased estimates of $q^{(m)}$ and $v^{(m)}$ respectively (Raghunathan et al, 2003).

3.2.3.3. Approximate $f\left(Q \mid D^M\right)$ with $f\left(Q \mid D_L^M\right)$

In deriving $f\left(Q \mid D^M\right)$, $q^{(m)}$ and $v^{(m)}$ are viewed as the sufficient summaries of $D^{(m)}$. In Section 3.2.3.2, unbiased estimates for $q^{(m)}$ and $v^{(m)}$ are constructed based on the set of synthetic data nested within $D^{(m)}$. The next step, therefore, is to approximate $f\left(Q \mid D^M\right)$ with $f\left(Q \mid D_L^M\right)$ by substituting $\left\{\left(q^{(m)}, v^{(m)}\right), m = 1, 2, ..., M\right\}$ with their estimates $\left\{\left(\overline{q}^{(m)}, T^{(m)}\right), m = 1, 2, ..., M\right\}$ respectively in the first stage t-distribution inference equations. The posterior mean of Q based on D_L^M is

$\hat{\hat{q}}_m = \frac{\sum_{m=1}^{M} \hat{q}_m}{M} = \frac{\sum_{m=1}^{M} \overline{q}^{(m)}}{M} = \frac{\sum_{m=1}^{M} \sum_{l=1}^{L} q_l^{(m)}}{M \times L} = \overline{q}_M$, and its posterior variance is

$$\begin{aligned}
\hat{T}_m &= \left(1 + M^{-1}\right) \hat{b}_m + \hat{\overline{v}}_m \\
&= \left(1 + M^{-1}\right)\left(M - 1\right)^{-1} \sum_{m=1}^{M} \left(\hat{q}^{(m)} - \hat{\overline{q}}_m\right)^2 + M^{-1} \sum_{m=1}^{M} \hat{v}^{(m)} \\
&= \left(1 + M^{-1}\right) B_M + M^{-1} \sum_{m=1}^{M} \left[\left(1 + L^{-1}\right) b_L^{(m)} - \overline{v}_L^{(m)}\right] \\
&= \left(1 + M^{-1}\right) B_M + \left(1 + L^{-1}\right) \overline{b}_M - \overline{v}_M = T_M
\end{aligned} \qquad [3.9],$$

where $\overline{q}_M$, $\overline{v}_M$, B_M and $\overline{b}_M$ are defined as in Section 3.2.2.3.

To estimate the degrees of freedom, we assume a t-distribution for first-stage item imputation and normal distribution for second-stage full synthesis. These assumptions are very reasonable as they are consistent with those we impose in deriving the mean and variance estimators. After substituting the between Item-missing imputation variation b_m and total variance T_m with their estimates B_M and T_M respectively in the degrees of freedom estimator in Section 3.2.2.1, the estimated degrees of freedom based on D_L^M is, therefore, $\gamma_M = \left(M - 1\right)\left(\left(1 + M^{-1}\right) B_M T_M^{-1}\right)^{-2}$.

3.2.4. Simulation validation

This section presents a simulation study for the two-stage full synthesis. We evaluate this new combining rule, the degrees of freedom estimator as well as the adjustments for negative variance and degrees of freedom.

We first generate a population of size $N = 1000$ from a 5-variate normal distribution with means equal to zero, variances equal to one and a common covariance of 0.5. Then we draw 500 independent actual samples of size $n = 100$ from this population by simple random sampling. For each actual sample, we create general pattern item-missing data on all five variables under the assumption of missing at random (MAR). Each incomplete actual sample is considered to be one observed data set. The algorithms for generating the missing data are described below. Under the assumptions of multivariate normality and MAR, for each observed data, we impute for the missing data independently $M = 5$ times using Markov chain Monte Carlo (MCMC) method (Schafer, 1997), which gives five imputed complete samples. For each complete sample, $L = 5$ synthetic populations of size 1000 are created. Finally, from each synthetic population, a random sample of size $k = 250$ is drawn. Thus, we obtain $500 \times 5 \times 5 = 12,500$ synthetic samples, which are considered the public data.

3.2.4.1. Generate item-missing values

MAR assumption suggests that the missing-data mechanism depends only on the observed data. To simulate the arbitrary missing data pattern, we develop four logistic models of creating item-missing data on only one variable, jointly on two variables, jointly three variables, and jointly four variables respectively. Every unit has at least one observed value across all five variables.

Specifically, for each actual sample, we divide the entire sample equally into four groups. The missing-data-generation models are then randomly assigned to the groups, such that each model is applied to one and only one group in an actual sample. For each data group, after the model is assigned, we 1) randomly select one or more variables to be subject to missing data, in which the number of variables to select is informed by the model; 2) fit the deletion model on the rest of variables, and finally 3) create missing values on selected variable(s) as suggested by the model. We repeat this three-step process for all four groups to complete the data deletion procedure for one actual sample. We repeat the above procedure for all 500 actual samples independently to create 500 observed samples with missing values.

We next illustrate this procedure using the following example. Let X_1, X_2, X_3, X_4, and X_5 be the five normally distributed variables respectively. The missing-data-generation models have this general form: $\text{logit}\left(\text{Prob}\left(R=1 \mid X\right)\right)=X^T\beta$, where R is the missing indictor and X is one subset of the data matrix, which varies by models. Suppose for the first group, one-variable model is assigned and the selected item-missing-data variable is X_1. The model creating around 50% missing data is $\text{logit}\left(\text{Prob}\left(R_i=1 \mid X_2, X_3, X_4, X_5\right)\right)=0.5+X_2+X_3+X_4+X_5;\ i=1,2,...,n/4$.

In specific, the missing data are created using the following steps:

a. Estimate the predictive probabilities P for X_1 to be missing from the above model, where P is a vector of size $n/4$.

b. Generate a Uniform random deviate vector of size $n/4$, $U \sim \text{Uniform}(0,1)$. Calculate the missing indictor $R=I\left(P>=U\right)$ where $I(\bullet)$ is indictor function.

c. Create missing data on X_1. For observation i, $X_{1i} =$ missing if $R_i = 1$ and $X_{1i} = X_{1i}$ if $R_i = 0$.

If a three-variable model is assigned and a random selection of variables that are subject to item missing data is (X_1, X_2, X_4), then the model of creating around 50% missing data is $\text{logit}\left(\text{Prob}\left(R_i = 1 \mid X_3, X_5\right)\right) = 0.5 + X_3 + X_5;\ i = 1, 2, ..., n/4$. The estimated predictive probabilities are for (X_1, X_2, X_4) to be jointly missing. R is computed in the same way as shown above. Then, for individual i, we assign X_{1i}, X_{i2} and X_{i4} missing values if $R_i = 1$. Similar steps apply when other missing-data-generation models are selected.

3.2.4.2. Generate synthetic data

We use the same synthetic data model as described in the Simulation Study 1 in Raghunathan et. al (2003). Specifically, the synthesis model assumes multivariate normal distribution with unknown mean and covariance matrix. Non-information Jeffrey priors are applied (Jeffreys, 1961). Because the synthesis model matches the true model, the synthetic data are created under the best situation. This setup allows for the evaluation of our inference method without unnecessary implications from other factors.

3.2.4.3. Simulation results

The estimands of interest are the marginal means for all five variables and the regression coefficients of X_1 on (X_2, X_3, X_4, X_5). Almost all variance estimates are positive with only one negative estimate out of 5000. 670 out of 5000 estimated degrees of freedoms are smaller than $M - 1 = 4$, and adjustment rule for the degrees of freedom applies. Small estimates for the degrees of freedom occur mostly on analytic statistics with only one on descriptive mean estimate. The sampling distribution of the actual sample and synthetic sample estimates of the regression coefficients are almost the same. Table 3.1 compares the inferences of descriptive and analytic statistics from the synthetic data sets and the actual data. The point estimates for both types of statistics are very similar across the synthetic and the actual data. The synthetic data standard errors are larger than the actual data suggesting some loss of precision. Consistent results with respect to this precision loss are found with the coverage rates as the synthetic data provides an over nominal coverage, 96.08% than the actual sample, 95.52%. The intervals from synthetic samples are wider than those from the actual samples as the average interval lengths are 0.38 and 0.66 respectively. Based on this new combining rule, the repeated sampling properties of the inference from the actual and synthetic sample are almost identical as predicted. In conclusion, the combining rule and degrees of freedom estimator yield valid synthetic inference when item-missing data is imputed prior to creating fully-synthetic data.

Table 3.1: Descriptive and analytic statistics estimated from the synthetic data sets and the actual data in the simulation evaluation of combining rule.

	Synthetic			Actual			No. of
Type	Estimate	S.E.	Coverage (%)	Estimate	S.E.	Coverage (%)	estimates
Mean	0.04	0.14	97.6	0.04	0.10	95.7	2500
Intercept	0.05	0.12	97.6	0.05	0.08	95.7	500
Slope	0.20	0.13	95.1	0.20	0.10	96.0	2000

3.3. Imputation models

3.3.1. Sequential regression

Sequential regression was originally motivated by the widely recognized difficulty of generating imputes for a large number of variables of different types from a full Bayesian model (Raghunathan, et al. 2001). For complex data structure, such a realistic joint model is difficult to formulate, although it is theoretically appealing to impute for the missing values from the joint posterior predictive distribution.

Multivariate sequential regression imputation approach is a flexible alternative that only requires the specification of a series of multivariate conditional models.

Suppose that a dataset without missing data is comprised of survey variables $\left(Y_1, Y_2, ..., Y_p\right)$ and a design variable Z. Fully-synthetic approach replaces the observed values of $\left(Y_1, Y_2, ..., Y_p\right)$ for all respondents with imputations performed multiple times. Under the sequential regression approach, each set of imputes is created by draws from the posterior predictive distribution generated from a series of conditional regressions, $f_k\left(Y_k \mid Z, \text{ all } Y_{j \neq k}, \theta_k\right)$, $k = 1, 2, ..., p$, where f_k are the conditional distribution function with parameter θ_k. We assume a diffuse non-informative prior for θ_k. The sequence of imputation is continued in a cyclical manner, each time using updated predictor sets and overwriting previously drawn values. This procedure is repeated independently multiple times to complete multiple synthetic-data. More details on the sequential regression imputation procedure for missing data appear in Raghunathan et. al (2001).

The variables in the dataset are assumed to be one of the following three types: continuous, binary or categorical. Separate regression models are used for different type variables. The specific conditional functions suitable for each variable type are described in the next section.

3.3.2. ACE model for continuous data

The Alternating Conditional Expectation (ACE) model (Breiman, et al. 1985) is a semi-parametric regression technique aiming to fully explore and explain the effect of covariates on a continuous response variable in multiple regression while making few assumptions about the regression function. This motivation is facilitated by estimating the transformations of the response and a set of covariates that produce the maximum linear effect between the (transformed) covariates and the (transformed) response variable. Because the ACE algorithm does not require defining the relation structure *a priori*, it is superior to the standard parametric tools in modeling the complex and irregular relationships among survey variables.

An ACE regression model has the following general form:

$$\psi(Y) = a + \sum_{j=1}^{p} \phi_j(X_j) + \varepsilon \qquad [3.10],$$

where $\varepsilon \sim N(0,1)$, ψ is a function of the response variable, Y and ϕ_j are functions of the covariate variable X_j, $j = 1, 2, ..., p$. The optimal transformation functions ψ and ϕ_j are estimated using an iterative method by minimizing the unexplained variance, ε^2, of a linear relationship between the transformed response variable and the sum of transformed covariates: $\varepsilon^2(\psi, \phi_1, ..., \phi_p) = E\left\{\left[\psi(Y) - \sum_{j=1}^{p} \phi_j(X_j)\right]\right\}^2$.

If a covariate is a categorical variable (either ordinal or nominal), the final optimal function ϕ_j assigns a real valued score to each level. The algorithm can be implemented using the ace function in the R statistical package. For details of the iterative fitting algorithm, see Breiman and Friedman (1985).

To demonstrate how the ACE algorithm can be used to identify the functional relationship between response and covariates and improve prediction precision, we use the data from the Health and Retirement Study (n=12319) as an example. Suppose that the response variable is the household assets. The covariates include age (years), school years (years), height (cm), weight (pounds), household income (1000 dollars), gender (men and women), self-rated health status (5 points Likert-scale), smoking status (never smoked, Past and current smoker). Figure 3.1 shows the transformations for the response variable and the four covariates estimated by ACE. From these figures, it can be seen that the transformation functions are rather tailored to the irregular empirical distributions that are observed in a real data set. This flexibility of simultaneously estimating such transformation functions for multiple variables makes the ACE superior to parametric approaches.

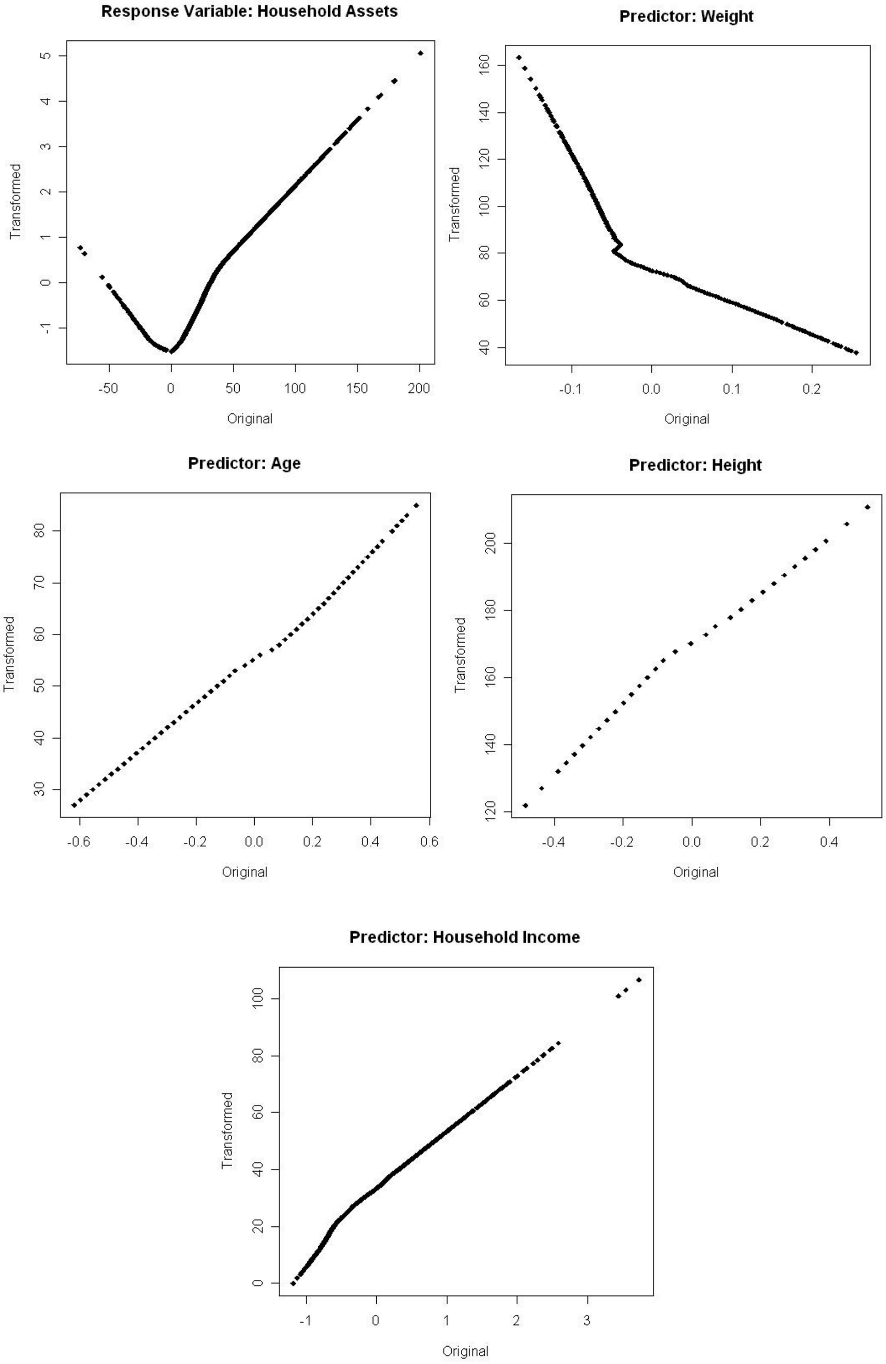
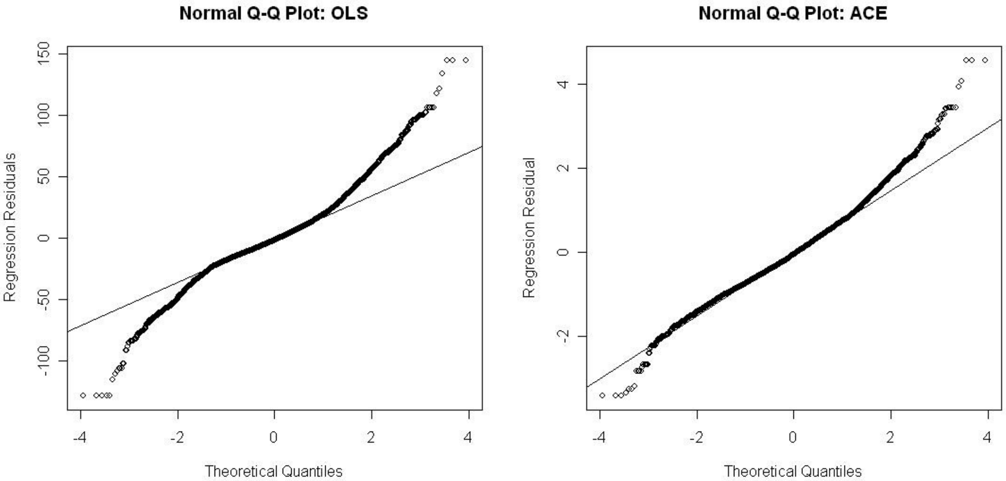


Figure 3.1: ACE optimal transformations for continuous variables in the HRS dataset

Figure 3.2 displays the normality Q-Q plots for the residuals from the simple linear regression and ACE regression. From comparing these two plots, the model fit is improved for ACE as the error is more distributed more towards normality.


Normal Q-Q Plot: OLS

Normal Q-Q Plot: ACE

Figure 3.2: Q-Q Plots for the residuals from ACE and standard linear regression

A regression of the transformed response variable on all transformed covariates results in all parameter coefficients of the predictors as positive and close to one as shown in the following equation:

$$\begin{aligned}\hat{\psi}(Assets) = .99\hat{\phi}_1(Hgt) + \hat{\phi}_2(Income) + .96\hat{\phi}_3(Wgt) + 1.04\hat{\phi}_4(Age) + \\ 1.02\hat{\phi}_5(Gender) + .98\hat{\phi}_6(Smoke) + \hat{\phi}_7(Health)\end{aligned}.$$

The adjusted R^2 for the ACE model, 0.37, is larger than that for the Ordinary Least Squares (OLS), 0.31. Thus, more variation in response variable is explained by the independent variables in ACE. Therefore, ACE improves model fit and increases the correlations between the response and predictor variables. ACE is exceptionally powerful in predictions. However, when the sole goal in statistical analysis is estimation, ACE should be used with caution because the interpretations are complicated by the transformation functions.

In addition, the prediction performance using the ACE approach depends on the order in which the covariates are entered into the model (Hastie and Tibshirani, 1990, Breiman and Friedman, 1985). To capture the prediction uncertainty, random permutation on the order of the continuous covariates is built within each imputation iteration in constructing the synthetic data.

3.3.3. Ridge-logistic regression model for binary data

Standard parametric logistic regression is often used to impute values for a binary data by random draws from the posterior predictive distribution. However, with the presence of sparse or highly correlated binary covariates, the maximum likelihood estimates, though unbiased, are very unstable. Therefore, the imputation may perform poorly because of large prediction errors. One commonly used technique to obtain more stable estimates is to drop non-significant covariates. However, it comes with a cost of losing the full conditionality, on which we build the imputation framework. An alternative approach, which maintains full conditionality, is to shrink the parameters and permit a slight bias in the estimates but with smaller variances. These type of estimators are usually achieved by minimizing the mean square errors (MSE) instead of bias as in standard logistic models.

One such useful shrinkage method is based on ridge-penalized function, which was originally developed under standard linear regression (Duffy and Santner, 1989), then extended to logistic regression by Le Cessie and Van Houwelingen (1992).

The ridge-penalized logistic estimator is derived as the restricted maximum likelihood estimator, which can be obtained by Newton-Raphson algorithm with ridge parameter defined *a priori*. Choosing the optimum ridge parameter to minimize the prediction error is very important. A cross-validation method is usually used to evaluate several selected ridge parameter values. However, it is very time consuming and computation intensive to build this procedure into each iteration of imputation. Alternatively, the appropriate ridge parameter can be estimated from the data (Schaefer et al., 1984). Two ridge estimators with different definitions are proposed in the literature. These two approaches have been shown to be equivalent asymptotically (Le Cessie and Van Houwelingen, 1992).

Let us consider the probability function p for a standard logistic regression model:

$$p(X_i) = \exp(\alpha + X_i\beta) \big/ \{1 + \exp(\alpha + X_i\beta)\} \qquad [3.11],$$

where β is a p-dimensional parameter vector excluding the intercept, and the ridge estimator $(\alpha^{\lambda}, \beta^{\lambda})$ is obtained by maximizing the following penalized log-likelihood $l^{\lambda}(\alpha, \beta)$ (Le Cessie and Van Houwelingen, 1992):

$$l^{\lambda}(\alpha, \beta) = \sum_i \left[Y_i \log p(X_i) + (1 - Y_i) \log\{1 - p(X_i)\} \right] - \lambda \|\beta\|^2,$$

where $\lambda = \frac{p}{2} \hat{\beta}^{MLE\prime} \hat{\beta}^{MLE}$ (Schaefer et al., 1984). The first component is the unrestricted log-likelihood as in standard logistic regression and the second part is the penalty as a function of maximum likelihood estimator for β. The penalizations only apply on β because adding a constant to the base of odds ratio would not result in a shift of the prediction by a constant.

Computationally, this estimation is achieved in two steps. In the first step, the point estimates $\hat{\beta}^{mle}$ is obtained by Newton-Raphson algorithm by maximizing the unrestricted log-likelihood. The ridge parameter λ is computed as a function of $\hat{\beta}^{mle}$. Then another Newton-Raphson procedure is carried out by maximizing the penalized log-likelihood initialized with $(\hat{a}^{mle}, \hat{\beta}^{mle})$ to obtain the point and interval ridge estimates.

We next present a simple example to show how ridge regression helps improve prediction. The binary response variable is whether a respondent was working for pay at the time the interview was carried out. The marginal percentage estimated from the HRS (n=12652) is 33.5%. Ninety-two covariates measuring social-demographic, health characteristics and design variables are included. For categorical variables with more than two levels a set of dummy variables are created, so that the final number of covariates is 116. Some categorical predictors are highly correlated. The proportions of dichotomized predictors range from 0.7% to 80%, which suggests a sparse data condition.

The first step of Newton-Rapson algorithm of maximizing the unrestricted likelihood takes 16 iterations. The penalty parameter based on $\lambda = \frac{p}{2}\hat{\beta}^{MLE\prime}\hat{\beta}^{MLE}$ is 4.55. When $\lambda = 0$, the ridge estimator then reduces to standard maximum likelihood estimator. Three model fit statistics are defined as follows:

a. Mean Square Error (MSE): $MSE = n^{-1}\sum_{i=1}^{n}\left(y-\hat{p}\right)^2$,

b. Mean Classification Error (MCE):

$MCE = n^{-1}\sum_{i=1}^{n}\left[y \times I\left(\hat{p} < 0.5\right) + \left(1-y\right)I\left(\hat{p} > 0.5\right) + 0.5\left(\hat{p} = 0.5\right)\right]$, and

c. Akaike Information Criterion (AIC):

$AIC = -2\sum_{i=1}^{n}\left[y \times \log\left(\hat{p}\right) + \left(1-y\right)\log\left(1-\hat{p}\right)\right] + 2 \times \mathrm{df}$, where we substitute df with the effective version, $\mathrm{df}_{\mathrm{eff}}$, for penalized regression as suggested by Cessie and Houwelingen (1992).

Table 3.2 shows the comparisons on these statistics, between the standard and ridge-penalized logistic regressions. Both AIC and MCE provide support that more precise predictions can be expected to obtain based on ridge regression. The MSE is almost identical between the two models. Considering the extreme complexity of these models, the model fitting for the ridge-logistic model is significantly improved as suggested by much lower values associated with AIC and MCE.

Table 3.2: Model fit statistics for unrestricted logistic and ridge logistic regressions

	AIC	MCE	MSE
Unrestricted Logistic	8997.8	.1434	.1044
Ridge Logistic	8985.8	.1400	.1046

3.3.4. Hierarchical ridge-logistic regression model

A multinomial categorical variable of total k levels can break down into $k-1$ nested binary variables. Then sequential ridge penalized logistic regression is employed to impute for each binary variable. For example, suppose Y has three levels, which are sorted in a decreasing order by frequency and each takes a value: 1, 2 and 3. Let n_1, n_2 and n_3 be the cell sizes respectively. Y is redefined using $k-1=2$ binary variables, $Z_1 = I\left(Y=1\right)$ and $Z_2 = I\left(Y=2 \mid Y \geq 2\right)$. Ridge-logistic regression is used to model Z_1 and Z_2 separately. First, the mode on Z_1 is fit on all sample units of size $\left(n_1+n_2+n_3\right)$ to generate imputations for $\hat{Y}=1$ versus $\hat{Y}=2 \text{ or } 3$ collectively. Second, we estimate the parameters of the model on Z_2 using the portion of sample defined by $\left(n_2+n_3\right)$, and then create imputations for $\hat{Y}=2$ versus $\hat{Y}=3$ for those sample units who are predicted to take values 2 or 3 from the first step. For a categorical variable with a more general number of levels, say k levels, the nested imputation approach then involves $k-1$ steps with each subsequent step used to generate imputations for one level versus the collection of all later levels. Prior to imputation, we order all the k levels decreasingly by the frequencies. The reason is to ensure the resulted nested binary variables are optimally balanced in such a manner that their means are mostly close to 0.5, thus precluding one potential contributor to the ill-conditioning problem in logistic regressions due to the skewness of the response variable. The gains in predictions are expected to be more prominent for a later nested variable as the sample size can be very small and the numerical ill-conditioning problems are more likely to occur (Clogg et al., 1991).

3.4. Diagnostic tool for synthetic data inference

3.4.1. Propensity score balance check

As discussed in Raghunathan (2008), the success of this synthetic data approach depends on establishing the validity of the synthetic data generation process, and achieving inferential validity of analyzing particular synthetic datasets. Both model inadequacy and random errors may contribute to the statistical discrepancies between analyzing one synthetic data and the actual data. Even if created from a carefully tested model, one synthetic dataset may appear very different from the actual data solely due to random errors because there are several levels of randomizations during the whole process of synthetic data generation.

One useful approach of evaluating the statistical similarities between a particular synthetic sample and the actual data is the propensity score analysis (Raghunathan, 2008). We use this approach as a diagnostic tool to detect potential data imbalances and reject imbalanced synthetic data.

The propensity score model (Rosenbaum and Rubin 1983) was originally developed to reduce bias when making casual inference about certain treatments based on observational data. The imbalance in the distributions of a set of covariates between treatment and control groups is summarized through a singular measure, the propensities for individuals of belonging to the treatment group. Bias reduction then is achieved by weighted analysis inversely proportional to the propensities.

To evaluate the dissimilarity between a synthetic data of size m and actual data of size n, we construct a propensity score model to estimate the propensity for a unit to belong to the synthetic data. To achieve this goal, we append a synthetic data matrix, S, to the actual data, C, from which S is generated. An indicator variable R is created as $R = I(i \in S)$ if unit i belongs to the synthetic data and $R = 0$ otherwise. We fit a ridge-penalized logistic model for R on all synthesized variables. The predicted probabilities are then sorted and grouped into deciles (or quintile, if sample size is limited). If the two datasets are equivalent in the distribution of perturbed variables, the proportion of units belonging to the synthetic data within each decile is very close to m/n. In this study, all synthetic data are of the same size as the actual data, therefore, $m/n = 0.5$. Simple t-tests for the discrepancy between the observed proportion and the expected rate, i.e. 0.5, are conducted for each decile group. In addition, a chi-square test with 10 degrees of freedom (if deciles are used, otherwise the degrees of freedom equals the number of propensity groups) provides summary statistics for the overall balance of the synthetic data. One must be cautious when trying to draw conclusions based on these tests because of 1) overpowered tests due to large sample size and 2) violated independence conditions between the two datasets (Raghunathan, 2008). Despite these two limitations, the descriptive evaluation based on this diagnostic tool is still very informative. Empirical evidence in later sections show that proportions falling within the range of $(0.4, 0.6)$ in the propensity balance checking procedure are often associated with smaller information loss in the analytic estimates from substantive models.

3.4.2. Information loss functions

The extent of information loss in statistical inference when analyzing the synthetic data compared to the actual microdata is a very important criterion for evaluating a statistical disclosure control method. Data utility (Willenborg and De Waal 2001), the accuracy of inferences obtained from publicly released data, can be evaluated by judging the closeness of the confidence intervals based inference using the release synthetic data with the confidence intervals obtained using the actual data.

Our first measure is adopted from Karr et al. (2006). By approximating the posterior distribution of estimand Q by normal distribution, the 95% confidence interval for the multiple synthetic data estimate $\overline{q}_M$ is $\left(L_{syn,\overline{q}_M}, U_{syn,\overline{q}_M}\right)$. Let $\left(L_{act,\hat{q}}, U_{act,\hat{q}}\right)$ be the corresponding interval for point estimate $\hat{q}$ obtained using the actual data which also follows t distribution with $(n-p)$ degrees of freedom where n and p are sample size and the number of parameters fitted in one particular analysis model respectively, which are different from those defined in the imputation model in later sections. Let $f_{syn,\overline{q}_M}$ and $f_{act,\hat{q}}$ be the estimated posterior distributions of Q computed using synthetic and actual data respectively. So the probability overlap in the confidence intervals for Q equals

$$I_Q = \frac{1}{2}\left[\int_{L_{syn,\hat{p}}}^{U_{syn,\overline{p}_M}} f_{act,\hat{p}}(t)dt + \int_{L_{act,\hat{p}}}^{U_{act,\hat{p}}} f_{syn,\overline{p}_M} dz\right] \qquad [3.12].$$

I_Q may take value $[0, 0.95]$. If there is no overlap, then $I_Q = 0$ and if the two interval overlap perfectly then $I_Q = 0.95$. Therefore, a large value for I_Q implies better data utility in estimating this scalar estimand Q.

We define the second measure as the relative overlap in the interval lengths. Let $\left(L_{over,q}, U_{over,q}\right)$ be the overlap of the two intervals, then the relative overlap in the interval length is

$$J_Q = \left(U_{over,q} - L_{over,q}\right) / \left(U_{act,\hat{q}} - L_{act,\hat{q}}\right) \qquad [3.13].$$

J_Q can take any value between 0 and 1. Zero means that there is no overlap between the two intervals and one means the synthetic interval completely covers the actual interval. This measure is different from the analogue measure used by Karr et al. (2006) in that they took the average of the relative overlap in the interval lengths for actual data and for synthetic data. Our measure is more realistic as the interval from actual data is what is to be compared against and we want to evaluate what proportion of the actual interval length is overlapped with the synthetic interval without being contaminated by the length of the synthetic interval.

A third measure is to evaluate whether point estimate $\overline{q}_W$ falls within the actual confidence interval:

$$K_Q = I\left(\overline{q}_W \in \left[L_{act,\hat{q}}, U_{act,\hat{q}}\right]\right) \qquad [3.14],$$

where $I(\bullet)$ is an indicator function. $K_Q = 1$ if $L_{act,\hat{q}} \leq \overline{q}_M \leq U_{act,\hat{q}}$ and $K_Q = 0$ otherwise. This measure allows evaluation of the bias of the inference drawn based on synthetic data.

We also compute the Z score. The Z score for a scalar statistics Q is calculated by

$$Z_Q = \left(q_{syn} - \hat{q}_{act}\right) / se_{act}\left(\hat{q}_{act}\right) \qquad [3.15],$$

which merely considers the closeness of the two point estimates. Small absolute Z value suggests low information loss.

Lastly, we evaluate whether similar statistical hypothesis testing conclusions can be drawn using the synthetic and actual data at some significance level. The statistical validity is met if either of the following two conditions are satisfied. The first condition is satisfied if given a significant result based on actual data, a significant testing result in the same direction using synthetic data is achieved. The second condition is given a non-significant result based on actual data; a non-significant result is achieved regardless of the directions. This measure allows one to summarize how sensitive the actual statistical conclusions are to the synthetic data.

3.5. Data description

In this section, the semi-parametric fully-synthetic algorithm described in earlier sections is applied to the HRS data. HRS is a longitudinal survey with a two-year interval starting in 1992. The data comes from a United States national multistage area probability sample of elder adults aged 51-61 as of 1992. This survey provides extensive information on physical and mental health, insurance, financial status, family support systems, labor market status and retirement planning. The microdata are publicly accessible and have been heavily used by researchers in health, public policy and social-economic areas. Thousands of papers or book chapters based on HRS data have been published (Institute for Social Research, 2008). Three articles are selected among those identified with a Medline search with keyword HRS. These articles encompass the main types of heavily used statistical models. Table 3.3 shows the citations and types of analytic models used in each article.

Table 3.3: Analysis types and article citations of three articles using HRS data

	Citations	Analysis Models	Cited by*
1	Buckley, C. B., Angel, J. L. and Donahue, D. (2000). Nativity and Older Women's Health: Constructed Reliance in the Health and Retirement Study, Journal of Women and Aging Vol.12, 21-37	One sample t-test Two sample t-test Logistic-Regression	4
2	He, X. and Baker, D. W., (2004). Changes in Weight among a Nationally Representative Cohort of Adults Aged 51 to 61, 1992 to 2000, American Journal of Preventive education, Vol. 27(1), 8-15	One sample t-test Two sample t-test Linear Regression	4
3	Siegel, M. J., Bradley, E. H., Gallo, W. T. and Kasl, S. V. (2004). The Effect of Spousal Mental and Physical Health on Husbands' and Wives ´ Depressive Symptoms, Among Older Adults: Longitudinal Evidence From the Health and Retirement Survey, Journal of Aging and Health Vol. 16(3), 398-425	One sample t-test Two sample t-test Seemly Unrelated Regression	11

Note: * statistics dated March 2008

This empirical study is based on a subset of the public release data selected by and including all the units and variables used in these articles. The data comprise 98 variables across four waves measured on 12,652 respondents. Due to panel attrition, there are some missing data due to wave unit nonresponse as shown in Table 3.4.

Table 3.4: Sample size and realization of panel attrition

Design/Common Variables	Wave 1 1992	Wave 2 1994	Wave 3 1996	Wave 5 2000
HHID PN SECU, Stratum Sampling Weights Age Gender Race Hispanic School year	n=12652	n=11492	n=10964	n=8896

Note:
🟨: Observed
🟦: Missing data due to panel Attrition

The synthetic data procedure is applied on all 98 variables involved in the statistical analysis described in these articles. Five complete data are generated by imputing the item-missing data caused by either item nonresponse or wave unit nonresponses. Based on each complete data, ten fully-synthetic data are generated for public dissemination. The data values in the synthetic data that are originally missing due to panel attrition are reset to be missing to ensure the same amount of information are used in the actual data analysis as published and synthetic data analysis. The inferential models of interest are those statistical analyses conducted in these articles.

Table 3.5 shows the descriptions of variables used in this study. CES-D, Center for Epidemiologic Studies Depression Scale, comprises eight items measuring depressive symptoms: felt depressed, felt everything she or he did was an effort, experienced restless sleep, could not get going, felt lonely, felt sad, enjoyed life and was happy. All eight items are dummy coded as one, if depressive symptom present and zero otherwise. Health status is an index of self-reported presence of physician-diagnosed health conditions including high blood pressure, diabetes, cancer, chronic lung disease, heart problems, stroke and arthritis. ADLs are comprised of 17 items measuring self-reported limitations on functional ability, including ability to run or jog a mile; walk several blocks; walk one block; walk across a room; get up from a chair after sitting for long periods; get in and out of bed without help; climb several flights of stairs without resting; climb one flight of stairs without resting; lift or carry weights more than 10lb; stoop, kneel or crunch; pick up a dime from a table; bathe or shower without help; reach or extend arms above shoulder level; pull or push large objects such as a living room chair; eat without help and dress without help. Items within each index are mostly correlated with each other as some concepts are nested. Furthermore, the reports for very severe conditions are usually associated with sparse distribution, such as less than 0.1% respondents report to difficult to eat without help.

Table 3.5: Description of variables used in the empirical simulation

	Variables	Units	Range/Coding
Continuous	Age	Years	[23, 85]
	HH Income	$	(0, 1378,750)
	HH Assets	$	(-743,677, 8,096,385)
	Non-housing Assets	$	(-733.871, 8,230,173)
	School Year	year	[0, 17]
	Height 92	cm	(94, 211)
	Weight 92	kg	(36, 181)
	Weight 96	kg	(27, 177)
	Weight 2000	kg	(31, 181)
Binary	Gender	2	Men; Women
	Foreign born	2	Foreign born; Native born
	Neighbor chat	2	Chat w/t neighbor, Not chat
	Volunteer	2	Volunteer 100hrs plus, Less or no volunteer
	Attend church	2	Attend church 1+/month, Less or No attend
	Satisfy w/t family life	2	Good+Fair+Poor; Excel+Vgood
	Self Rated Health 94	2	Excel+VGood+Good, Fair+Poor
	Religious	2	Have religion, No religion
	Work for pay 92	2	Work for pay, Not work for pay 92
	Work for pay 94	2	Work for pay, Not work for pay 94
	No income 92	2	No income, Have Income
	Proxy Interview 92	2	Yes, No
	Proxy Interview 94	2	Yes, No
	Primary respondent 92	2	Primary Respondent, Secondary Respondent
	HS degree	2	HS/GED, less than HS/GED
	CES-D 92	2	8 binary variables
	CES-D 94	2	8 binary variables
	Health Status 92	2	7 binary variables
	Health Status 94	2	7 binary variables
	ADLs 92	2	17 binary variables
	ADLs 94	2	17 binary variables
Multinomial	Work involves labor 92	4	None, Some, Most and All of the time
	Work involves labor 94	4	None, Some, Most and All of the time
	Race	5	White, African American, Mexico-Hispanic, Other-Hispanic and Other races
	Marriage status	4	Never Married, Div/Separated, Widowed, and Married
	Smoking status	3	Never, Past smoker and Current smoker
	Alcohol usage	3	Abstainer, Moderate and Heavy
	Heavy housework	5	Never, <0.25/wk,0.25-0.74/wk, 0.75-2/wk, and >=3/wk
	Light activity 92	5	Never, <0.25/wk,0.25-0.74/wk, 0.75-2/wk, and >=3/wk
	Vigor activity 94	5	Never, <0.25/wk,0.25-0.74/wk, 0.75-2/wk, and >=3/wk
	Self Rated Health 92	5	Excellent, Very good, Good, Fair and Poor

3.6. Evaluations Inference from Synthetic Data

In this section, we use the HRS data comprised of all variables previously identified to compare the properties of inferences for a variety of descriptive and analytic statistics from the multiply-imputed fully-synthetic data and the actual data.

3.6.1. Generating synthetic data

The generation of multiple synthetic data involves accomplishing two tasks. We first impute the item-missing data $M = 5$ times to complete the actual data and then for the second task create $L = 10$ fully-synthetic data based on each complete data. The multiple synthetic data produced from completing the second task are released for the public use. Both tasks are fulfilled in two steps in a similar fashion. The only difference is the amount of data that replaced by imputes. Item-missing data are imputed in the first task and values on all units are imputed in the second task. The specific steps for generating the synthetic data are as follows. We also show the modifications to these steps for completing the first task.

The first step involves constructing synthetic population $P_l^{(m)} = \left(Z, Y_l^{(m)}\right), m = 1, 2, ..., M$ from a rectangular actual data $D^{(m)}$ of size n, which is achieved by semi-parametric approach using Bayesian bootstrap (Rubin, 1981, Raghunathan et al, 2003). Specifically,

a. Draw $(n-1)$ uniform random numbers. Sort those numbers in ascending order. We label this ordered sequence as $\alpha_0 = 0, \alpha_1, \alpha_2, ..., a_{n-1}, a_n = 1$.

b. Draw n uniform random numbers $u_1, u_2, ..., u_N$. Select unit j (row j) if $a_{j-1} < u_r \le a_j$ where $r = 1, 2, ..., n$. The resulting $n \times p$ matrix $Y_l^{(m)}$ together with Z, is a synthetic population $P^{(m)}$, where p is the dimension of survey variables to be imputed. Repeat $a-b$ M times independently giving multiple synthetic samples.

Raghunathan et al. (2001) shows comparable inferences can be obtained by analyzing these bootstrap synthetic samples and the actual data. However, imputations from bootstraps might still cause concern about confidentiality due to two reasons: (1) the released data still contains the actual values, and (2) this type of de-identification is reversible. Thus, further data perturbation based on imputation models is necessary.

In the second step, the conditional density function for the k^{th} variable $p\left(Y_{(k)}^{(m)} \mid Z, Y_{(-k)}^{(m)}, k=1,2,\ldots,p\right)$ is estimated based on synthetic population $P^{(m)}$. The impute, $Y_{i,(k)}^{*(m)}$ is drawn from the posterior predictive distribution

$$p\left(Y_{i,(k)} \mid X, Y_{(-k)}^{*(m)}, P^{(m)}\right) = \int p\left(Y_{i,(k)} \mid X, Y_{(-k)}^{*(m)}, P^{(m)}, \theta\right) p\left(\theta \mid P^{|m|}\right) d\theta \text{ for } i=1,2,\ldots,n,$$

which can be accomplished by drawing $\theta^{(m)}$ from the posterior distribution of θ given $P^{(m)}$, where θ is the regression parameter. Then, we update the actual values of $Y_{i,(k)}$ by $Y_{i,(k)}^{*(m)}$ to complete the imputation. We repeat step 2 in a cycling manner for all $Y_{(k)}$, each time conditional on updated predictors, and overwrite previously imputed values multiple iterations gives synthetic data $D_l^{(m)} = \left(X, Y_l^{*(m)}\right)$. We repeat the above procedure independently to create L multiple synthetic data nested within a complete data. For each complete data, repeat the above synthetic steps to account for synthesis uncertainty.

Different conditional models are used to draw imputations for various variable types. Computationally, binary and multinomial data are imputed identically.

a. If $Y_{(k)}$ is continuous,

   I. The estimated conditional distribution based on $P^{(m)}$ is

$$Y^*_{i,(k)} \mid X_i, Y^*_{i,(-k)}, P^{(m)} \sim \psi^{-1}\left[a^{(m)} + \phi_x^{(m)}(X_i) + \sum_{j \neq k}^{p} \phi_j^{(m)}(Y_{i,j}^{*(m)}) \right],$$

where θ^{-1} is achieved by linear interpolation on θ.

II. Then, the posterior distribution of parameter θ is given by

$$\theta \mid X_i, Y^*_{i,(-k)} \sim N\left[\left(a^{(m)} + \phi_x^{(m)}(X_i) + \sum_{j \neq k}^{p} \phi_j^{(m)}(Y_{i,j}^{*(m)}) \right), 1\right].$$

III. Draw a random value, $\theta^{(m)}$ from its posterior distribution in II. It can be accomplished by drawing a random residual r from $N(0,1)$ or from $\underset{\rightarrow}{\hat{r}}$, observed residual vector obtained from I, then

$$\begin{aligned} \theta^{(m)} &= \left[a^{(m)} + \phi_x^{(m)}(X_i) + \sum_{j \neq k}^{p} \phi_j^{(m)}(Y_{i,j}^{*(m)}) \right] + r \\ Y_{(k)}^{*(m)} &= \psi^{-1}\left[a^{(m)} + \phi_x^{(m)}(X_i) + \sum_{j \neq k}^{p} \phi_j^{(m)}(Y_{i,j}^{*(m)}) + r \right] \end{aligned}.$$

b. If $Y_{(k)}$ is binary,

I. The estimated conditional distribution is

$$\begin{aligned} &Y^*_{i,(k)} \mid X_i, Y^*_{i,(-k)}, P^{(m)} \\ &\quad \sim \text{Bernoulli}\left\{ \exp\left(\alpha^{\lambda(m)} + X_i^{T(m)} \beta^{\lambda(m)} \right) \Big/ \left[1 + \exp\left(\alpha^{\lambda(m)} + X_i^{T(m)} \beta^{\lambda(m)} \right) \right] \right\} \end{aligned}.$$

II. Then, the posterior distribution of parameter θ is

$$\theta \mid X_i, Y^*_{i,(-k)}, P^{(m)} \sim \text{Uniform}(0,1).$$

III. Draw a random value, $\theta^{(m)}$ from its posterior distribution in II, then

$$Y_{(k)}^{*(l)} = I\left(p_{i(k)}^{*(l)} >= \theta^{(m)} \right), \text{ where } p_{i(k)}^{*(l)} = \frac{\exp\left(\alpha^{\lambda(l)} + X_i^{T(l)} \beta^{\lambda(l)} \right)}{\left[1 + \exp\left(\alpha^{\lambda(l)} + X_i^{T(l)} \beta^{\lambda(l)} \right) \right]}.$$

To impute for item-missing data, let $M_{i,(k)} = 1$ if unit i for the k^{th} variable is missing and hence need to be replaced with imputed value. In addition, let $M_{i,(k)} = 0$

if it is observed and left unchanged. This $n \times K$ missing indictor matrix of $M = \left(M_{(1)},...,M_{(K)}\right)$ is constant for creating all complete data $D^{(m)}$, $m = 1,2,...,M$. Step 1 is unchanged as in full synthesis task. We replace Step 2 with drawing values $Y_{i,(k)}^{*(m)}$ for only those records with $M_{i,(k)} = 1$ from the estimated posterior predictive distribution with this general form:

$$p\left(Y_{i,(k)} \mid X, M_{i,(k)} = 1, Y_{(-k)}^{*(m)}, P^{(m)}\right) = \int p\left(Y_{i,(k)} \mid X, M_{i,(k)} = 1, Y_{(-k)}^{*(m)}, P^{(m)}, \theta\right) p\left(\theta \mid P^{(m)}\right) d\theta .$$

The rest of procedures in Step 2 are unchanged, which include 1) sequentially imputing for missing values for each variable multiple rounds to produce a complete data $D^{(m)}$, and 2) repeating Step 1 and this modified Step 2 M times independently to produce multiple complete data.

3.6.2. Results

We first evaluate the similarity of the synthetic data as a whole with the actual data based on the propensity scoring method, and then we examine the information loss due to the disclosure control procedures on a series of pre-defined statistics.

3.6.2.1. Propensity scoring balance

We create a total of 50 synthetic samples, 10 synthetic samples based on each of five imputed data. As explained in Section 3.4.1, we append the synthetic data to the imputed actual data and estimate the propensity probability for each observation to belong to the actual data. Because the synthetic data size and the actual data size are the same, the estimated probabilities are expected to equal or be close to 0.5 if the synthetic data is similar to the actual data in terms of distributions of covariates contained in the propensity-scoring model aggregately. Figure 3.3 shows the side-by-side histograms of estimated propensity probabilities for observations in the synthetic and the actual data sets across all 50 synthetic data. The distributions for the two sets of probabilities are almost indistinguishable with the mass largely concentrated around their peaks at 0.5. Therefore, the synthetic data is very comparable to the actual data as a whole with regard to the distribution of estimated propensities.

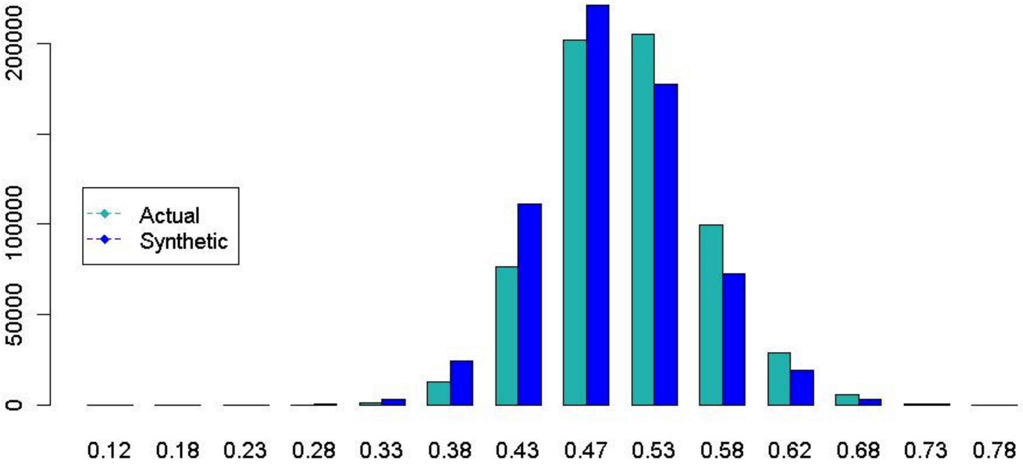


Figure 3.3: Overlap of estimated propensity probabilities for observations in the synthetic data and the actual data

We further group the observations into deciles according to their order in the estimated propensities. Table 3.6 shows the average estimated proportions of observations that belong to the actual data within each decile, and the chi-square statistics with degrees of freedom of 10 for all 50 synthetic data. Although both the t-tests and chi-square test appear significant because of overpowering by a large sample size, the mean proportions are very close to 0.5.

Table 3.6: Propensity score balance statistics across all 50 synthetic data

	Mean	Min.	Max.
Estimated Prob. $\hat{p}$	0.50	0.39	0.62
Chi-square Statistics	462.10	310.06	588.13
Sample Size	25304		

We also test the similarities in the distributions of all 98 synthesized covariates within each propensity deciles (or propensity strata). For continuous variables, we test the differences in means between two groups: synthetic data group and actual data group using one-way ANOVA tests. Non-significant F test results suggest that there is no difference between the two data groups with regards to the distributions of the variable of interest. For both binary and categorical variables, we test whether the proportions of observations with certain variable attributes in the synthetic data is about the same as the proportion in the actual data across all propensity strata using Cochran-Mantel-Haenszel Chi-Squared test. Table 3.7 summarizes the test statistics for all 98 variables by variable types. All test values are very small for all three types of variables. The corresponding p values for such tests are very large on average. These non-significant test results indicate that the synthetic data and the actual data are very similar with regard to the distributions of the synthesized variables. The balancing test statistics for each variable are provided in Appendix 3.A.

Table 3.7: Distributional homogeneity test for all 98 synthesized variables across all 50 synthetic data sets

Variable Type	No. of Variables	Test Values			P-Values		
		Mean	Min.	Max.	Mean	Min.	Max.
Binary†	80	0.255	0.000	12.347	0.784	0.000	1.000
Categorical†	9	0.400	0.000	3.320	0.953	0.506	1.000
Continuous*	9	4.723	0.000	120.138	0.379	0.000	0.999

Note: * For continuous variables, the test values are F-values with numerator degrees of freedom of 1 and denominator degrees of freedom of 25293;
† For both binary and categorical variables, the test values are Chi-square test values.

3.6.2.2. Information loss in statistics

These synthetic data are analyzed as described in the published articles as if they were the actual data. A total of 572 statistics, which include 203 means, 197 mean differences and 172 regression coefficients, are estimated from the synthetic data and the actual data. The variance estimates are almost positive. Only 3 out of all 572 variance estimates are negative. For these variance estimates, we use adjusted variance estimators as shown in Equation 3.7. They are the overall sample means for CES-D 92, ADLs 92 and ADLs 94. All three statistics are computed as the sums of several highly positively correlated binary indictors, thus we speculate that sampling variance for these composite variables is large. Therefore, within-synthesis variance may dominate the total variance, which leads to negative variance estimates. Only two estimated degrees of freedom are less than $M-1=4$ and are set to be 4 following Equation 3.8.

Figure 3.4 shows the scatter plots of the point estimates obtained using the synthetic data and the actual data. The fact that the points are almost clustered around the 45-degree line provides evidence that synthetic data yield similar point estimates as those from the actual data.

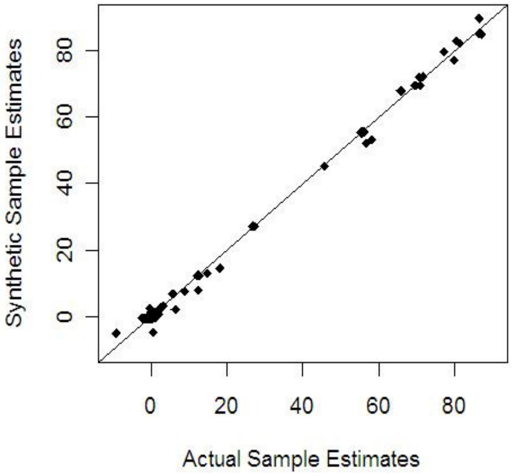


Figure 3.4: Scatter plots of point estimates in means, subgroup mean differences and regression coefficients from synthetic and actual data

We also fit a simple linear regression of the synthetic estimates on the actual data estimates. If the two sets of estimates are similar, we expect an intercept of 0 and a slope of 1. The estimated intercept and slope of this regression with 570 degrees of freedom are -0.033 ± 0.56 and 0.996 ± 0.004 respectively. Neither estimate is significantly different from its respective expected value according to results from the Wald tests.

We compare the point and interval estimates for all statistics from the actual and synthetic data as shown in Table 3.8. We first examine the loss of synthetic data utility with respect to the estimation of means/proportions. Then we discuss such information loss in statistics that involve sub domain comparisons and regression models.

Table 3.8: Information loss for various descriptive and analytic statistics

	Actual		Synthetic		I_Q	J_Q	K_Q	Z	Test
	Est.	SE	Est.	SE					
Article 1: Buckley et al., 2000									
Means									
Binary	0.31	0.02	0.29	0.02	0.67	0.64	0.96	-0.39	56/56
Continuous	55.68	0.13	55.39	0.35	0.67	0.95	1.00	-3.51	7/7
Mean Differences									
Binary	0.10	0.03	0.09	0.03	0.69	0.65	1.00	0.03	19/24
Continuous	0.07	0.22	0.15	0.20	0.84	0.74	1.00	0.43	3/3
Logistic Coefs.	-0.16	0.12	-0.11	0.08	0.65	0.53	1.00	0.31	37/52
Article 2: He and Baker, 2004									
Means									
Binary	0.23	0.01	0.23	0.01	0.68	0.69	0.93	0.55	85/86
Continuous	12.48	0.31	12.21	0.37	0.69	0.76	0.95	-0.70	27/38
Mean Differences									
Binary	0.24	0.07	0.23	0.10	0.57	0.59	0.80	-0.38	115/133
Continuous	26.12	0.57	25.84	0.84	0.47	0.53	0.78	-0.72	27/37
Linear Coefs.	-0.06	0.35	0.00	0.35	0.79	0.78	1.00	0.11	42/54
Article 3: Siegel et al., 2004									
Means									
Binary	0.62	0.01	0.60	0.01	0.37	0.37	1.00	-1.44	6/6
Continuous	15.71	0.28	14.82	0.23	0.57	0.54	0.80	-2.01	10/10
SUR Coefs.	0.08	0.01	0.07	0.02	0.73	0.82	1.00	-0.52	54/66
TOTAL									488/572

On average, the point estimates for estimating the means or proportions are very similar between the actual and synthetic data across all three articles. The standard errors obtained from synthetic estimates are larger than those from the actual data. The confidence interval overlap measures, I_Q and J_Q, yield almost identical results. For article 1 and 2, the average probability for both I_Q and J_Q are nearly 0.70. For article 3, these probabilities are lower with 0.37 for proportions and 0.57 for means.

From the perspective of K_Q, on the other hand, the agreement in the confidence intervals is higher with almost perfect coverage across all articles. The reason is that K_Q is defined as whether the actual point estimates are covered by the synthetic data intervals and the information about the synthetic data variance is ignored, thus it is a lower overlap criterion than I_Q and J_Q. Another "similarity" measure is the Z score, which estimates the biasness of the synthetic data estimates. The average Z scores are all close to zero except for the continuous variables in article 1, which is -3.51. A close examination of these statistics with larger Z scores indicates that they are mostly statistics about small domains, such as the mean age for individuals who are born abroad and self-reported to have poor health. Therefore, the imputations may be less stable because of small sample sizes, thus leading to larger discrepancies in point estimates relative to the variance. The last column in Table 3.8 is the number of hypothesis tests and the number of tests about which the synthetic and the actual data yield the same inference conclusions. For means or proportions, the null hypotheses for these t-tests are whether the means are equal to zero. Almost all such t-tests yield consistent conclusions.

For the estimation of the mean differences between domains, similar findings are found as in means or proportions. On average, the overlap probabilities range from around 60% to 80% according to I_Q and J_Q. Z scores are all very close to zero, which suggests the synthetic point estimates are all close to the center of the actual confidence intervals. In terms of the consistency in hypothesis testing, the tests conclusions on domain differences are less consistent, in which around 20% synthetic data inferences conclude differently from the actual data inferences.

Three types of regression models are fitted in each article. The point estimates for the synthetic data and the actual data are very close for all three types of regression coefficients. Synthetic data analyses are less efficient than the actual data analysis for both linear and SUR regressions. However, for logistic regression, the synthetic standard errors (0.08) are smaller than their actual data counterparts (0.12). One possible reason is that some of these substantive logistic models are included in the imputation model, thus the relationships among variables may be strengthened due to the imputation procedures. The Z scores are almost all zero, suggesting nearly unbiased point estimates for all coefficients are obtained from the synthetic data. However, only 133 pairs of coefficients tests, out of 172, yield consistent inference conclusions. This may be partly due to the large sample size, which gives great statistical power in detecting small differences. Given the extreme complex data structure and large number of synthetic variables, the achieved levels of similarity are very satisfactory.

Figure 3.5-7 shows the side-by-side confidence intervals of the estimated regression coefficients from the synthetic and actual data for each article. The two sets of confidence intervals are very close for all three articles. Generally, the synthetic data intervals are wider than the actual data intervals, except for some logistic regression coefficients in Buckley et al. (2000), where the synthetic data inference is more efficient. This may be due to the fact that some of the substantive models are originally poorly fitted and improved because of the synthesis procedures. Therefore, the synthetic data estimates of the relationship among variables are more efficient as more information is included.

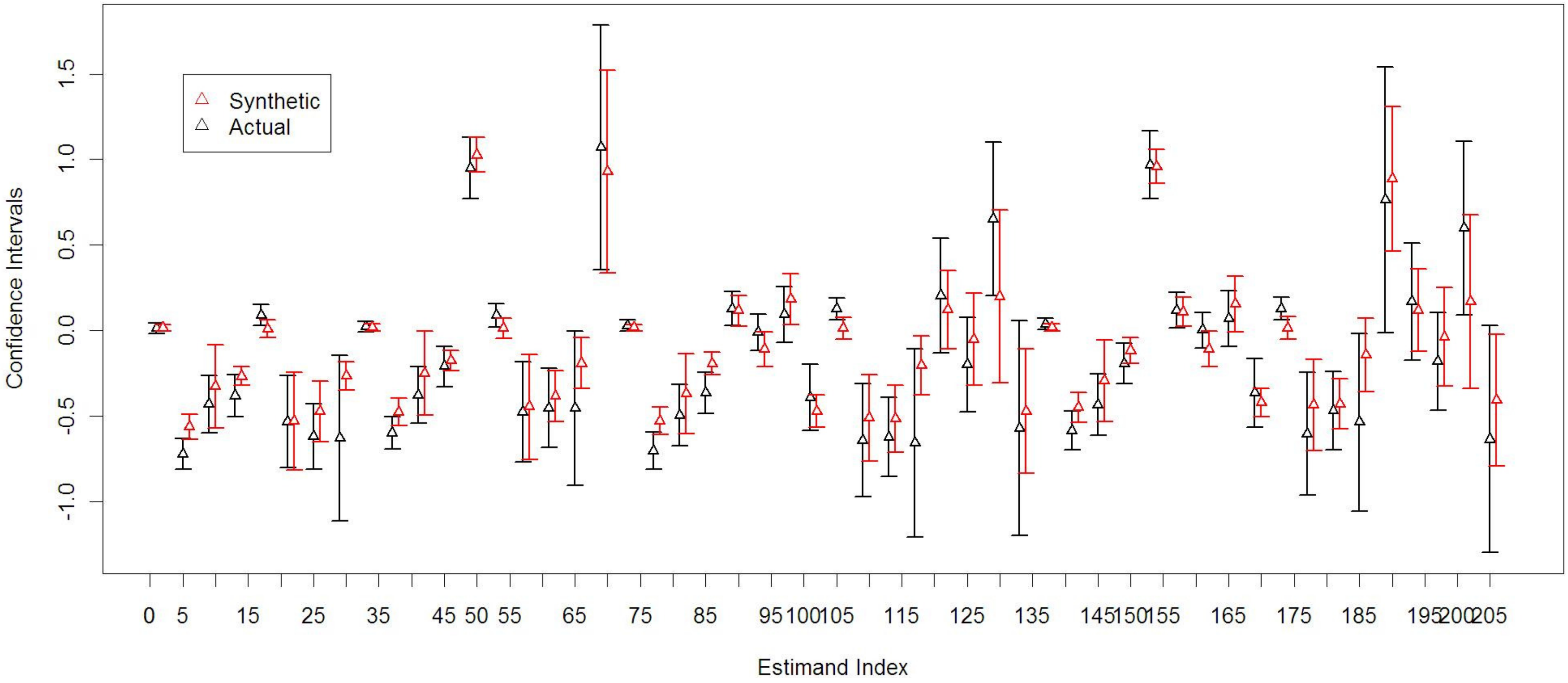


Figure 3.5: Side-by-side confidence interval plots for the logistic regression coefficients estimates in Buckley et al, 2000

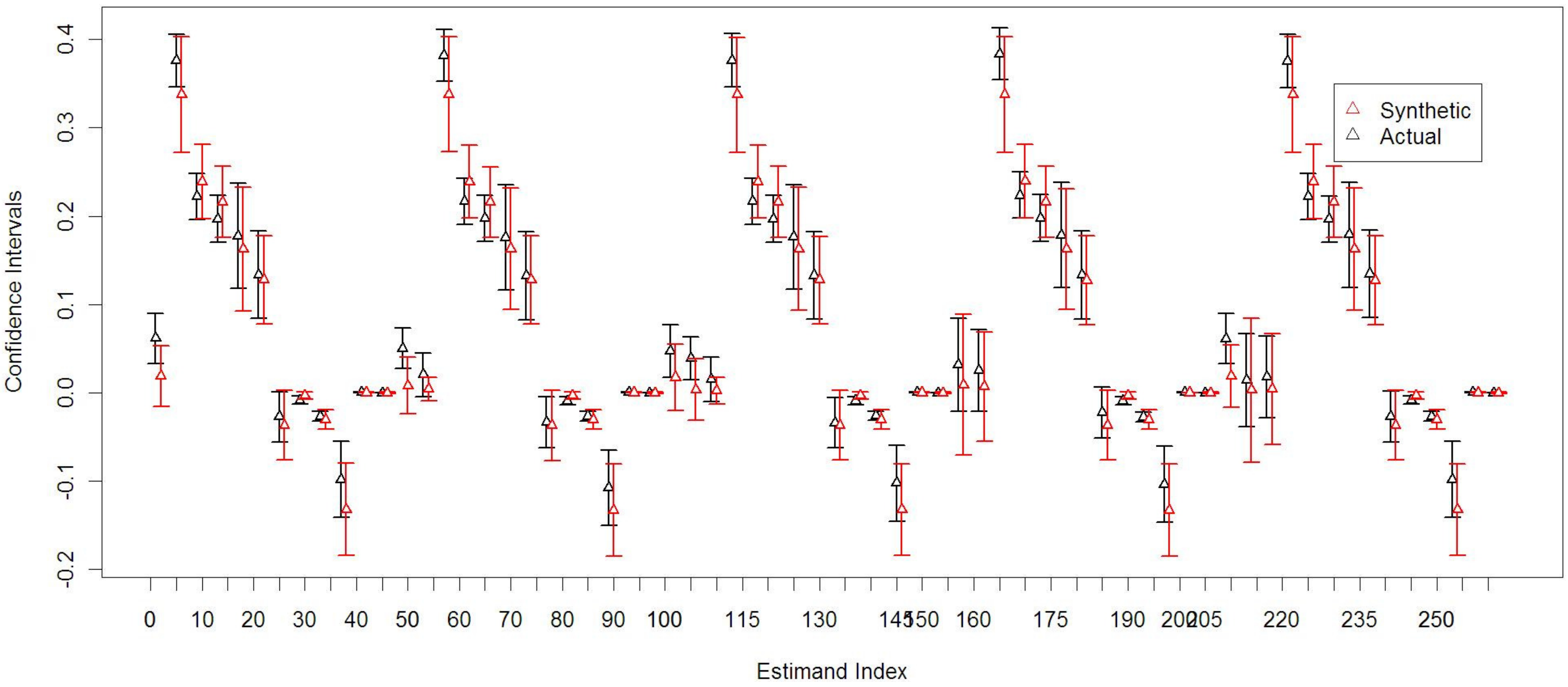


Figure 3.6: Side-by-side confidence interval plots for the Seemly-Unrelated regression coefficients estimates in Siegel et al, 2004

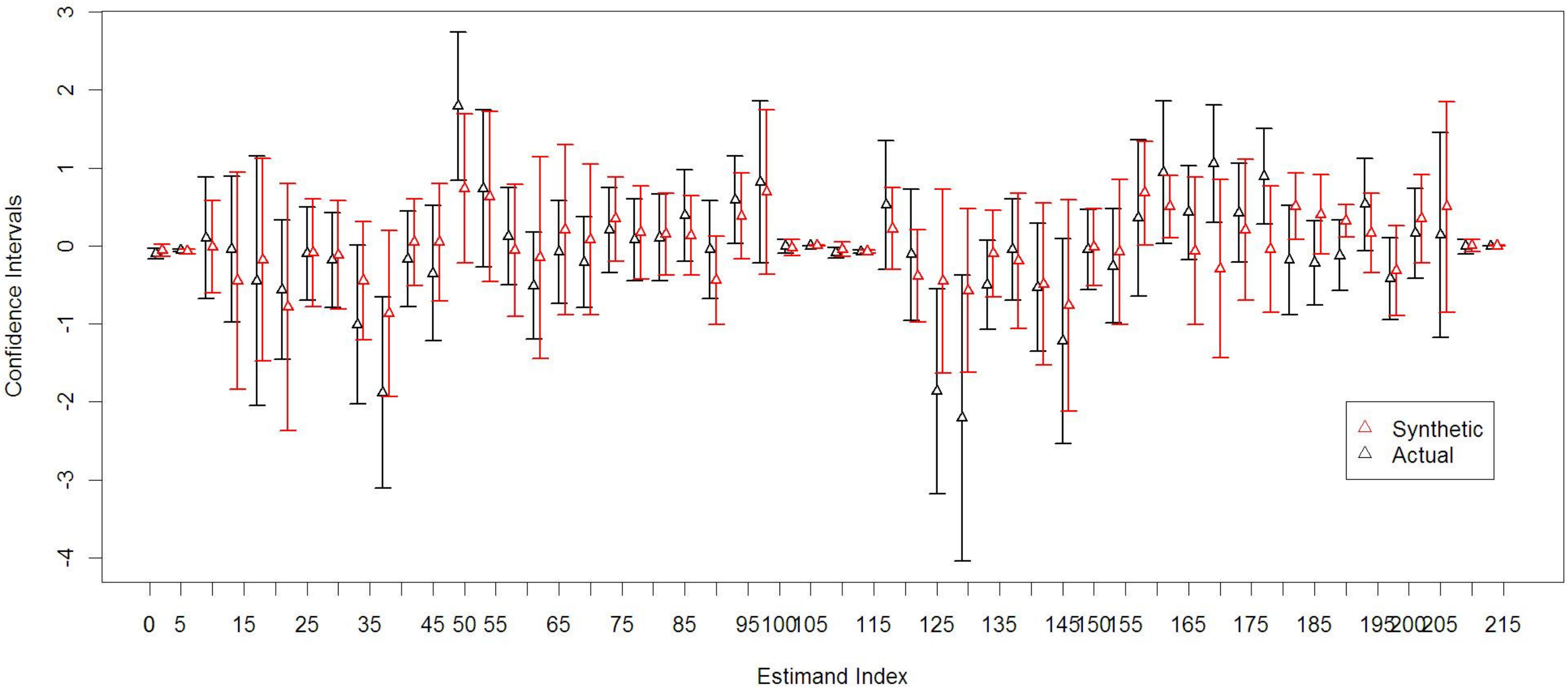


Figure 3.7: Side-by-side confidence interval plots for the linear regression coefficients estimates in He et al, 2004

3.7. Conclusion and discussion

In this chapter, we develop and evaluate a multiple-imputation approach of creating fully-synthetic datasets to be disseminated in place of the actual data to avoid inadvertently disclosing information about survey respondents. We derive a new combining rule for making inferences about a scalar estimand using multiple synthetic data when item-missing data are imputed prior to the synthetic data generation. We validate this new rule via a simulation study.

We investigate the use of a series of semi-parametric regression models for generating imputations for different types of variables under the framework of sequential regressions. We tested this approach using a national longitudinal complex survey sample, which is comprised of a large number of variables of different types and the distributions of most variables are irregular. We also showed the successful use of the propensity score balance check as a diagnostic tool for assessing the similarity between a particular synthetic data and the actual data, and as selection criteria to select “best” synthetic data. Based on several information loss functions, the descriptive and analytic statistics estimated from the synthetic data are very similar to those from the actual data.

As all data values contained in any data record are to be modified, it no longer makes sense to refer the resulted synthetic data record as it pertains to any individual. This lack of correspondence between data values and individuals can be used by agencies to deter intruders from exploiting the published data.

The statistical modeling challenges presented in the HRS dataset, such as nonnormality, sparseness, multi-linearity, correlation due to longitudinal data, complex sample designs etc., are commonly shared with other large-scale complex national surveys. Therefore, it is very promising to evaluate the techniques on this data set to speculate about its value for other surveys.

This research provides solid evidence for filling the gap between the fully-synthetic data approach being considered as only conceptually attractive to highly practically feasible. This approach is particularly suitable and practicable for surveys, which collect information on a small to modest number of variables, such as the short-form United States Decennial Census with only about only ten variables, because of reduced modeling complexity.

Appendix 3.A. Distributional homogeneity test for all 98 synthesized variables

		Test Value				P Value		
Type	Variable	Mean	Min.	Max	DF	Mean	Min.	Max
Continuous[*]	Age	17.886	0.001	120.138	1	0.162	0.000	0.980
	Height 92	9.857	0.006	87.207	1	0.203	0.000	0.936
	Non-housing assets	1.009	0.000	6.238	1	0.501	0.013	0.990
	School years	5.922	0.004	32.453	1	0.208	0.000	0.951
	HH assets	0.926	0.000	5.699	1	0.555	0.017	0.999
	Non zero HH income	2.238	0.000	14.174	1	0.394	0.000	0.999
	Weight 2000	1.683	0.002	11.963	1	0.466	0.001	0.968
	Weight 92	1.566	0.001	13.929	1	0.469	0.000	0.981
	Weight 96	1.420	0.002	11.664	1	0.455	0.001	0.968
Binary[†]	Church Asked	0.036	0.000	0.206	1	0.890	0.650	0.997
	HS degree	0.751	0.000	9.806	1	0.596	0.002	0.994
	Family	0.023	0.000	0.225	1	0.913	0.635	0.996
	Foreign born	0.073	0.000	0.439	1	0.836	0.507	0.999
	Gender	0.066	0.000	0.386	1	0.846	0.535	0.997
	Health rating 94	0.252	0.000	2.448	1	0.705	0.118	0.991
	Zero income	2.026	0.000	9.535	1	0.427	0.002	0.998
	Marriage status	0.639	0.006	3.663	3	0.882	0.300	1.000
	Neighbor chat	0.015	0.000	0.079	1	0.926	0.779	0.999
	Primary respondent	1.489	0.006	4.716	1	0.346	0.030	0.941
	Proxy interview 92	0.140	0.000	0.863	1	0.760	0.353	0.989
	Proxy interview 94	0.496	0.000	2.896	1	0.571	0.089	0.987
	Race	1.682	0.166	7.437	4	0.798	0.115	0.997
	Run or jog 92	0.019	0.000	0.142	1	0.917	0.706	0.998
	Walk blocks 92	0.115	0.000	1.094	1	0.794	0.296	0.999
	Walk one block 92	0.336	0.000	2.673	1	0.664	0.102	0.992
	Walk across room 92	0.170	0.000	1.916	1	0.776	0.166	0.996
	Sit 2 hours 92	0.026	0.000	0.283	1	0.907	0.595	0.999
	Get up chair 92	0.043	0.000	0.289	1	0.875	0.591	1.000
	In and out bed 92	0.135	0.000	1.078	1	0.790	0.299	0.994
	Climb flights 92	0.060	0.000	0.345	1	0.849	0.557	1.000
	Climb one flight 92	0.127	0.000	1.475	1	0.795	0.225	0.995
	Carry 10 lb more 92	0.054	0.000	0.491	1	0.861	0.484	0.992
	Stoop, kneel 92	0.046	0.000	0.435	1	0.877	0.510	0.999
	Dime 92	0.093	0.000	0.559	1	0.814	0.455	0.999

Binary†	Bathe 92	0.231	0.000	1.920	1	0.736	0.166	0.997
	Shoulder 92	0.126	0.000	1.608	1	0.793	0.205	0.999
	Large object 92	0.079	0.000	0.766	1	0.840	0.382	0.997
	Eat 92	1.400	0.000	5.897	1	0.385	0.015	0.998
	Dress 92	0.425	0.000	1.928	1	0.596	0.165	0.986
	High blood pressure 92	0.042	0.000	0.300	1	0.874	0.584	1.000
	Diabetes 92	0.043	0.000	0.382	1	0.879	0.536	0.998
	Cancer 92	0.022	0.000	0.192	1	0.914	0.661	0.998
	Lung 92	0.067	0.000	0.834	1	0.847	0.361	0.999
	Heart 92	0.031	0.000	0.386	1	0.901	0.535	1.000
	Stroke 92	0.232	0.000	1.529	1	0.725	0.216	0.999
	Arthritis 92	0.028	0.000	0.146	1	0.902	0.702	0.998
	Depressed 92	0.128	0.000	0.769	1	0.784	0.381	0.999
	Effort 92	0.102	0.000	0.653	1	0.805	0.419	0.999
	Restless sleep 92	0.046	0.000	0.388	1	0.866	0.533	0.997
	Happy 92	0.058	0.000	0.864	1	0.871	0.353	0.999
	Lonely 92	0.104	0.000	1.217	1	0.818	0.270	0.997
	Enjoy life 92	0.051	0.000	0.277	1	0.867	0.599	1.000
	Sad 92	0.119	0.000	0.631	1	0.786	0.427	0.997
	Not going 92	0.067	0.000	0.576	1	0.845	0.448	0.999
	Volunteer	0.047	0.000	0.276	1	0.865	0.600	0.999
	Run or jog 94	0.082	0.000	0.391	1	0.813	0.532	0.996
	Walk blocks 94	0.089	0.000	0.753	1	0.827	0.386	1.000
	Walk one block 94	0.187	0.000	1.083	1	0.740	0.298	1.000
	Walk across room 94	2.498	0.000	12.347	1	0.348	0.000	0.998
	Sit 2 hours 94	0.033	0.000	0.246	1	0.892	0.620	0.999
	Get up chair 94	0.031	0.000	0.289	1	0.903	0.591	0.995
	In and out bed 94	0.576	0.001	4.830	1	0.582	0.028	0.980
	Climb flights 94	0.076	0.000	1.180	1	0.839	0.277	0.998
	Climb one flight 94	0.148	0.000	1.258	1	0.779	0.262	0.992
	Carry 20 lb more 94	0.106	0.000	0.829	1	0.799	0.363	0.997
	Stoop, kneel 94	0.057	0.000	0.605	1	0.866	0.437	0.999
	Dime 94	0.079	0.000	0.706	1	0.827	0.401	0.996
	Bathe 94	0.225	0.000	2.005	1	0.745	0.157	1.000
	Shoulder 94	0.087	0.000	0.772	1	0.832	0.380	0.998
	Large objects 94	0.085	0.000	0.814	1	0.835	0.367	1.000
	Eat 94	0.633	0.001	3.748	1	0.586	0.053	0.980

Binary[†]	Dress 94	0.167	0.000	1.037	1	0.764	0.309	0.997
	High blood pressure 94	0.046	0.000	0.503	1	0.883	0.478	0.998
	Diabetes 94	0.143	0.000	0.997	1	0.772	0.318	0.995
	Cancer 94 if cancer 92	0.198	0.000	1.189	1	0.745	0.276	0.994
	Lung 94	0.147	0.000	1.423	1	0.794	0.233	1.000
	Heart 94 if heart 92	0.122	0.000	0.937	1	0.787	0.333	0.990
	Stroke 94	0.799	0.002	3.475	1	0.509	0.062	0.967
	Arthritis 94	0.011	0.000	0.093	1	0.943	0.761	1.000
	Depressed 94	0.187	0.000	1.424	1	0.753	0.233	0.995
	Effort 94	0.143	0.000	2.220	1	0.810	0.136	0.998
	Restless sleep 94	0.050	0.000	0.496	1	0.884	0.481	1.000
	Not going 94	0.096	0.000	1.323	1	0.831	0.250	0.999
	Lonely 94	0.086	0.000	0.717	1	0.843	0.397	0.999
	Sad 94	0.154	0.000	1.160	1	0.764	0.282	0.998
	Enjoy life 94	0.153	0.000	1.443	1	0.804	0.230	0.999
	Happy 94	0.099	0.000	1.719	1	0.848	0.190	0.999
	Work for pay 92	0.055	0.000	0.345	1	0.852	0.557	0.999
	Work for pay 94	0.388	0.000	2.062	1	0.622	0.151	0.998
Categorical[†]	Alcohol usage	0.128	0.001	0.698	2	0.940	0.705	1.000
	Church	0.077	0.001	0.314	2	0.963	0.855	1.000
	Heavy house work	0.616	0.020	2.098	4	0.945	0.718	1.000
	Health rating 92	0.762	0.036	2.993	4	0.928	0.559	1.000
	Light activity 92	0.174	0.011	0.979	4	0.994	0.913	1.000
	Smoke status	0.075	0.000	0.603	2	0.964	0.740	1.000
	Vigor activity 92	0.250	0.015	1.242	4	0.988	0.871	1.000
	Work involves labor 92	0.474	0.005	1.972	4	0.968	0.741	1.000
	Work involves labor 94	1.042	0.043	3.320	4	0.887	0.506	1.000

Note: [*] For continuous variables, the test values are F-values with numerator degrees of freedom shown in the column of DF. The denominator degrees of freedom all equal 25293;

[†] For both binary and categorical variables, the test values are Chi-square values.

[†] For both binary and categorical variables, the test values are Chi-square values.

CHAPTER IV

SYNTHETIC DATA FOR SMALL AREA ESTIMATION

4.1. Introduction

The need for detailed statistics on small geographical areas is constantly increasing. Such information is an indispensable tool in policy and decision-making at all levels of government and business. For example, the American Community Survey (ACS) is a key part of the 2010 Decennial Census Program and is designed to replace the long form sample used in past censuses to produce timely population estimates on demographic, social, housing and economic characteristics for a broad spectrum of geographic areas in the United States (2006). However, releasing small geographic details, for instance, for counties, regions, or even small areas, may increase the risk of identification of the ACS respondents when such geographic information is combined with some distinct personal-level characteristics. This issue is even more critical when there is good background knowledge of particular areas.

Current disclosure avoidance practices involve suppressing geographical details or making the data available only through data enclaves or Research Data Centers (RDC). For example, the Department of Energy precludes geographic identifiers for areas below the Census Division level (The U.S. Department of Energy, 2004). The Census Bureau and the National Center for Health Statistics release information for only areas with more than 100,000 residents (Census Bureau, 2007). As an alternative, the researchers are forced to use one or more research data centers and have to go through several hurdles to obtain permission to perform analyses. In addition, computer software environments in these RDCs may not permit implementing user defined specialized software for advanced analysis.

The key research question of interest, therefore, is how to release more geographical details without increasing the risk of disclosure. This chapter, built on a basic method described in Rubin (1993), creates and releases multiple fully-synthetic data sets with geographical details, and hence, permits small area estimation. Under this proposed methodology, no real data on the survey attributes are disseminated, thus, it offers full protection against disclosure. The data can be analyzed using the statistical package of a user's choice.

The general idea is to treat the unobserved part of the population as missing data, which includes all non-sampled individuals within sampled areas and all individuals within non-sampled geographic areas. The unobserved portion of the population is multiply imputed to create synthetic population datasets. A simple random sample (SRS) for each area is drawn from each synthetic population. The collection of SRS samples then comprises the public use data file, which contains statistical information on not only sampled areas but also non-sampled areas.

The fully-synthetic sample size does not have to be the same as the actual data. In fact, we can increase the sample size to the extent of allowing direct estimates of small areas statistics. Small area analyses then are simplified, which otherwise usually requires fit complex random effect models. For state-level statistics, each synthetic data set is analyzed as if it were the actual data. Valid inferences are obtained by combining the point estimates and the associated variances from each synthetic sample using the method described by Raghunathan, Reiter and Rubin (2003).

The proposed method developed in Section 4.2 is evaluated via two simulation studies in Section 4.3. In Section 4.4, this method is applied to the microdata from the 1880 Census. We evaluate whether such synthetic data sets yield valid inferences for statistics at subnational and state-levels. Finally, in Section 4.5, we provide conclusions with discussions.

4.2. Methods

4.2.1. Overview

Suppose that a population of size N consists of C small geographical areas for which we desire estimates, and $B_i, i = 1,2,...,C$ are the population area sizes. The survey data of size n consists of data from c areas drawn using certain sample design. Let $b_i, i = 1,2,...,c$ be the sample size for the sampled area i, and $\sum_{i=1}^{c} b_i = n$. Let $y_{ij}, i = 1,2,...,c; j = 1,2,...,b_i$ be a survey variable of interest (possibly a vector, but consider a scalar case for simplicity, for now). Let $\pi_{ij}, i = 1,2,...,C; j = 1,2,...,B_i$ be the inclusion probability for subject j in area i. Let $z_{ij}, i = 1,2,...,C; j = 1,2,...,B_i$ be a vector of subject specific auxiliary information available for all population subjects.

Conceptually, we propose to create a synthetic population $p_{syn} = \left\{ y_{ij}^*, i = 1,2,...,C; j = 1,2,...,B_i \right\}$, from which we select a synthetic sample $d_{syn} = \left\{ y_{ij}^*, i = 1,2,...,C; j = 1,2,...,b_i^* \right\}$ by probability proportional to size (PPS) design. In specific, we draw b_i^* units from each of C areas in p_{syn}. The area sample size b_i^* equals

$f \times B_i$, for some fixed value of f. We repeat this procedure independently M times and create multiple synthetic data sets $D_{syn} = \left\{ d_{syn}^{l}, l = 1, \ldots, M \right\}$.

The following general three-stage model, for example, could be used to create synthetic data sets. In the first-stage sampling model, the distribution of y_{ij} (discrete or continuous) is assumed as $y_{ij} \sim f\left(\theta_{ij}, \psi\right), i = 1, 2, \ldots, C;\ j = 1, 2, \ldots, b_i$, where the density function $f(\bullet)$ is parameterized with individual specific mean θ_{ij} and scale parameter ψ. Then θ_{ij} is modeled using $\theta_{ij} = z_{ij}^{T}\beta + v_{ij}$ in the second-stage, where z_{ij} are known auxiliary vectors of size p for individual j in area i, β is the unknown regression coefficients also of size p, and v_{ij} is the scalar error. It is assumed β is random and follows a normal distribution. We also assume that v_{ij} is distributed normally with mean zero and variance σ_v^2, i.e. $v_{ij} \sim N\left(0, \sigma_v^2\right)$. We are interested to obtain the joint posterior distribution of all unknown parameters $\left(\beta, \psi, \sigma_v^2\right)$ under a fully Bayesian framework, from which we generate predictions for y_{ij} within a synthetic sample. We assume mutually independent prior distributions for the model parameters $\left(\beta, \psi, \sigma_v^2\right)$.

Given the model assumptions, we can use Markov chain Monte Carlo (MCMC) methods such as Gibbs sampling or Metropolitan-Hasting approach to simulate the joint posterior distribution of the model parameters. M synthetic data sets are drawn independently from the simulated posterior predictive distribution. Synthetic data inference is then obtained based on the method presented in Section 4.2.2. To illustrate the basic idea we present two simulation studies and one empirical example. The first study simulates the situation when the sample is drawn by simple random sampling from a scalar normal population. The second study extends this scalar model to a mixed-type bivariate model comprised of one binary and one normal scalar variable. The last study presents a real data solution to synthesize values on variables of three different types: scalar, binary and semicontinuous. The synthesis models for these studies are described in Section 4.2.3, 4.2.4 and 4.2.5 respectively.

4.2.2. Analysis of synthetic data sets

Suppose the statistic of interest is a scalar estimand Q, which may be a function of the population. Let q_l and v_l be the point estimate and associated variance estimate of Q based on d_{syn}^l, the l_{th} synthetic data generated from the actual survey data. Then Q can be estimated by $\overline{q}$ with variance T,

$$\begin{aligned} \overline{q} &= M^{-1}\sum\nolimits_{l=1}^{M} q_l \\ T &= (1+M^{-1})B-\overline{v} \end{aligned} \qquad [3.16],$$

where $B=\left(M-1\right)^{-1}\sum\nolimits_{l=1}^{M}\left(q_l-\overline{q}\right)^2$ and $\overline{v}=M^{-1}\sum\nolimits_{l=1}^{M}v_l$.

One disadvantage of this variance estimator is that it may be negative, although negative values generally can be avoided by making M and/or n large. To accommodate this possibility, we use the adjusted variance estimator

$$T^* = I\left(T>0\right)\times T + I\left(T\le 0\right)\times\overline{v} \qquad [3.17],$$

where $I\left(\bullet\right)$ is the indicator function. When M is small or modest, inferences for scalar Q can be obtained based on a t-distribution with $M-1$ degrees of freedom, thus, a 95% synthetic confidence interval for Q is approximated by $\overline{q}\mp t_{0.975,df=M-1}\sqrt{T^*}$. For large M, the inference can be made based on a normal distribution (Raghunathan, et al. 2003).

For inferences on state-level statistics, each synthetic data set is analyzed as if it was the actual data set as in multiple imputation inference (Rubin, 1987) or fully-synthetic data inference (Raghunathan et al, 2003).

However, for inferences on small area statistics, the synthetic data are analyzed differently from the actual data. For example, suppose an analyst seeks inferences about the county-level means on a scalar variable Y, $\theta_i, i = 1, 2, ..., C$, where θ_i is the population mean for county i, and C is the total number of counties. Synthetic estimate for θ_i from d^l_{syn} is based on the direct estimator $\hat{\theta}^l_i = \bar{y}^{*l}_i$ with variance $\hat{\text{var}}\left(\hat{\theta}^l_i\right)$, where $\bar{y}^{*l}_i = \left(b^*_i\right)^{-1} \sum_{j=1}^{b^*_i} y^{*l}_{ij}$ and $\hat{\text{var}}\left(\hat{\theta}^l_i\right) = \left[b^*_i\left(b^*_i - 1\right)\right]^{-1} \sum_{j=1}^{b^*_i}\left(y^{*l}_{ij} - \bar{y}^{*l}_i\right)^2$. The synthetic data inference on θ_i is obtained by combining $\left\{\left(\hat{\theta}^l_i, \hat{\text{var}}\left(\hat{\theta}^l_i\right)\right), l = 1, ..., M\right\}$ using the rule presented earlier in this section.

In contrast, the point and interval estimates of θ_i from the actual data are usually based on a small area model. One possible such model for a scalar variable can be in the following form:

$$\begin{array}{l} \text{1. Sampling model: } y_{ij} \sim N\left(\theta_i, \sigma^2_e\right),\ i = 1, ..., C;\ j = 1, ..., b_i \\ \text{2. Linking model: } \theta_i \sim N\left(\mu, \sigma^2_v\right) \\ \text{3. Independent diffuse priors on } \left(\mu, \sigma^2_v, \sigma^2_e\right) \end{array} \qquad [3.18].$$

Then the actual data inference on θ_i can be obtained by a normal distribution with mean and variance approximated by the posterior mean and posterior variance respectively.

4.2.3. Model for a scalar variable

Under this setup, the general sampling and linking models presented in Section 4.2.1 reduce to

$$\begin{aligned}&\text{Stage 1. } y_{ij}=\theta_i+e_{ij},\ e_{ij}\sim N\left(0,\sigma_e^2\right)\\&\text{Stage 2. } \theta_i=\mu+v_i,\ v_i\sim N\left(0,\sigma_v^2\right)\end{aligned} \qquad [3.19],$$

where a constant variance, σ_v^2, is assumed across areas. We assume a flat prior for μ, i.e. $\pi(\mu)\propto 1$, and vague inverse gammar distributions for σ_e^2 and σ_v^2.

4.2.4. Bivariate model for a scalar and a binary variables

For a bivariate situation with one scalar variable and one binary variable, the general sampling and linking models can be reduced as below.

$$\begin{aligned}&\text{Stage 1. Sampling Model}\\&\quad x_{ij}\sim Bernoulli\left(\theta_i\right), i=1,2,\ldots,C, j=1,2,\ldots,b_i\\&\quad y_{ij}\mid x_{ij}\sim N\left(\beta_{0i}+\beta_{1i}x_{ij},\sigma_e^2\right)\\&\text{Stage 2. LinkingModel}\\&\quad\left[\operatorname{logit}\left(\theta_i\right),\beta_{0i},\beta_{1i}\right]\sim MVN(\mu,\Sigma)\end{aligned} \qquad [3.20],$$

where x_{ij} is the value of the binary variable for individual j in area i, and y_{ij} is the corresponding value of the scalar variable. μ is a 3×1 mean vector and Σ is a 3×3 variance-covariance matrix for the joint prior distribution of the mean function of (X,Y). We again assume a flat prior for μ, a diffuse inverse-Wishart distribution for Σ, i.e. $\Sigma\sim IW(\Psi,v)$, where Ψ is an identity matrix of size 3 with degrees of freedom $v=3$, and a diffuse inverse gammar distribution for σ_e^2.

4.2.5. Trivariate model for a scalar, a binary and a semicontinuous variables

In this section, we further extend the models to include another common type of variable, a semicontinuous variable. A variable is classified as semicontinuous if a large portion of data values is zero and the rest of the values that are greater than zero are continuously distributed. We adopt a two-step approach to model this type of variable.

The value of zero is treated as a discrete category. A Bernoulli distribution is assumed to model this zero versus non-zero status. Then, conditional on being non-zero, a normal distribution is used to capture the non-zero values.

Let X, Y and Z be the binary, semicontinuous, and scalar variables respectively. To explain the mixed distributional structure in Y, we create a zero status indicator such that $\mathrm{Y.d} = I(Y=1)$. The synthesis model is as follows:

$$\begin{array}{l}
\text{Stage 1. Sampling model:} \\
\quad \mathrm{X}_{ij} \sim \mathrm{Bernoulli}(\theta_i);\ i=1,2,\ldots,C;\ j=1,2,\ldots,b_i. \\
\quad \mathrm{Y.d}_{ij} \mid \mathrm{X}_{ij} \sim \mathrm{Bernoulli}\left(\alpha_{1i}+\beta_{1i}X_{ij}\right) \\
\quad \mathrm{Y}_{ij} \mid \mathrm{X}_{ij} \sim \mathrm{N}\left(\alpha_{2i}+\beta_{2i}\mathrm{X}_{ij},\sigma_1^2\right) \\
\quad \mathrm{Y}_{ij} = \mathrm{Y}_{ij} \times I\left(\mathrm{Y.d}_{ij}=1\right)+\mathrm{Y.d}_{ij} \times I\left(\mathrm{Y.d}_{ij}=0\right) \qquad\qquad\qquad [3.21]. \\
\quad \mathrm{Z}_{ij} \mid \mathrm{X}_{ij},\mathrm{Y}_{ij} \sim \mathrm{N}\left(\alpha_{3i}+\beta_{3i}\mathrm{X}_{ij}+\beta_{4i}\mathrm{Y}_{ij},\sigma_2^2\right) \\
\text{Stage 2. Linking model:} \\
\quad \left[\mathrm{logit}(\theta_{\mathrm{i}}),\alpha_{1i},\beta_{1i},\alpha_{2i},\beta_{2i},\alpha_{3i},\beta_{3i},\beta_{4i}\right] \sim MVN_8(\mu,\Sigma)
\end{array}$$

Similar to the Bivariate model, we assume a flat prior for μ, a vague inverse Wishart distribution for $\sum$, and vague inverse Gamma distributions for σ_1^2 and σ_2^2.

4.3. Simulation studies

In this section, we describe two simulation studies to compare the properties of inferences from multiple fully-synthetic data and the actual data. In both simulations, we create the actual survey sample data from a model that matches the synthetic data model. Thus, the comparison of inferences is optimal where the imputer's assumed model is indeed the true model.

4.3.1. Simulation study 1: a scalar variable

4.3.1.1. Generate actual survey sample

The survey data are generated as below:

1. Draw c areas from a total of C population areas.
2. Generate a vector of area-specific effect $v = \left(v_i, i = 1, 2, ..., c\right)$ from $N\left(\mu, \sigma_v^2\right)$.
3. Within each selected area, b_i subjects are randomly drawn from its population. $b_i, i = 1, 2, ..., c$ is calculated by proportional allocation.
4. Individual values are generated on the variable of interest $y_{ij}, j = 1, 2, ..., b_i$ from $N\left(v_i, \sigma_e^2\right)$. The observed sample is then denoted by $D = \left\{\left(I_i, y_{ij}\right), i = 1, 2, ..., c, j = 1, 2, ..., b_i\right\}$ where I is the area identifier.

4.3.1.2. Generate synthetic data

We use the Gibbs sampler to generate synthetic data sets as follows:

a. To draw $\left\{\left(\theta_i^*, i = 1, 2, ..., c\right), \mu^*, \sigma_e^{2*}, \sigma_v^{2*}\right\}$, we use the standard Gibbs sampler to draw $\left\{\left(\theta_i, i = 1, 2, ..., c\right), \mu, \sigma_e^2, \sigma_v^2\right\}$ from their joint posteriror distribution (we used WinBUGS for this step). For nonsampled areas, draw $\theta_i^*, i = c+1, c+2, ..., C$ from the normal distribution $N\left(\mu^*, \sigma_v^{2*}\right)$.

b. Define area sample size for both sampled and nonsampled areas as $b_i^* = f \times B_i, i = 1, 2, ..., C$.

c. Finally, draw values $y_{ij}^*, i = 1, 2, ..., C, j = 1, 2, ..., b_i^* \mid b_i^*, \theta_i^*, \sigma_e^{2*} \sim N\left(\theta_i^*, \sigma_e^{2*}\right)$. The synthetic data set is then denoted by $d_{syn} = \left\{\left(I_i, y_{ij}^*\right), i = 1, 2, ..., C, j = 1, 2, ..., b_i^*\right\}$.

d. Repeat steps 1 to 3 a total of M times to get M synthetic data $D_{syn} = \left(d_{syn}^1, d_{syn}^2, ..., d_{syn}^M \right)$.

4.3.1.3. Results

This simulation was carried out using R and WinBUGS, which is developed based on BUGS (Bayesian inference Using Gibbs Sampling). We used the estimated population for a total of 100 counties from the state of North Carolina published on the American FactFinder (The Census Bureau, 2006) as the measure of size for our hypothetical population. The sampling fraction λ is chosen to ensure the county sample sizes range from single digit to hundreds, which mimics the common situation in most national surveys.

Specifically, the model parameters used in this simulation are: $\mu = 0.5$; $\sigma_e^2 = \sigma_v^2 = 1$; $\lambda = 0.367$; $c = 50$; $C = 100$; $B_i, i = 1, 2, ..., C = (14, 2800)$; $b_i, i = 1, 2, ..., c = (5, 307)$ and $n = \sum_{i=1}^{c} b_i = 3000$. 250 replicates and $M = 20$ synthetic data sets from each replicate are generated. We increase the sampling proportion by a factor of ten, i.e. $b_i = 10 \times \lambda \times B_i$, when generating the synthetic data to allow the production of direct county-level estimates. All 25250 synthetic variance estimates are positive. Indirect county-level estimates on the actual data are obtained based on the same model as the one used to generate synthetic data. We compare the county-level and state-level means of Y as well as the corresponding variance and confidence intervals, which are obtained by analyzing the synthetic data against the same set of estimates obtained from the actual data.

Figure 4.1 displays the scatter plot of the posterior means of all sampled counties from the actual data and the direct means from multiple synthetic datasets across all replicates.

The fact that the points are clustered tightly around the 45-degree line provides evidence that synthetic data yield almost identical small area estimates. We fit a simple linear regression of a vector of the synthetic data estimates on the vector of the actual data estimates. Both vectors are of length 12500 $\left(= c \times \text{No.of replicates}\right)$. If the two sets of estimates are similar, we expect an intercept of 0 and a slope of 1. The estimated intercept and slope of this regression with 12498 degrees of freedom are -0.0005 ∓ 0.0008 and 1.0003 ∓ 0.0008 respectively. Neither estimate is significantly different from its respective expected value according to the Wald tests.

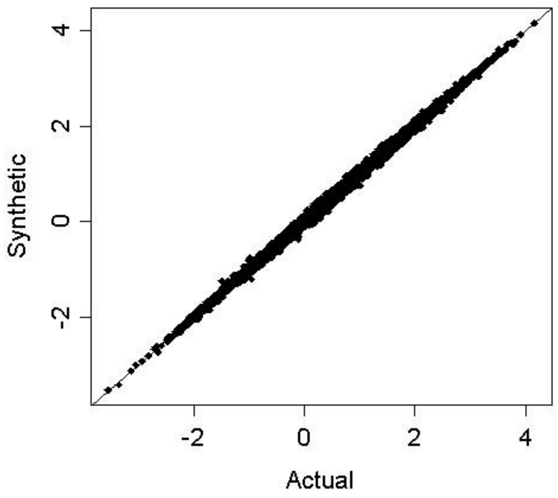


Figure 4.1: County-level estimates for sampled counties

Figure 4.2 shows the histogram of small area estimates from multiple synthetic data sets and the actual data for all nonsampled counties across all replicates. The estimates are distributed symmetrically around the population mean, $\mu = 0.5$, which indicates the synthetic estimates for nonsampled counties are similar to those based on observed data. There is a slight loss in precision for synthetic data estimates, as the distribution is flatter compared with those for the actual data estimates. On average, the synthetic data yield a confidence interval width of 0.707 compared with 0.655 for actual data based county-level estimates for sampled counties, which also indicates a slight loss in efficiency in estimating area means based on synthetic data compared with the actual data. The level of efficiency loss is indeed a decreasing function of county sample size as shown in Figure 4.3. For large counties, of size 50 or 100, the ratio of synthetic data variance and actual data variance is relatively small, around 1 or slightly more than 1. For small counties, of size 10 or less, the relative efficiency can be as high as 3 times, which means the synthetic data confidence interval may be 70% wider than the confidence interval from the actual data.

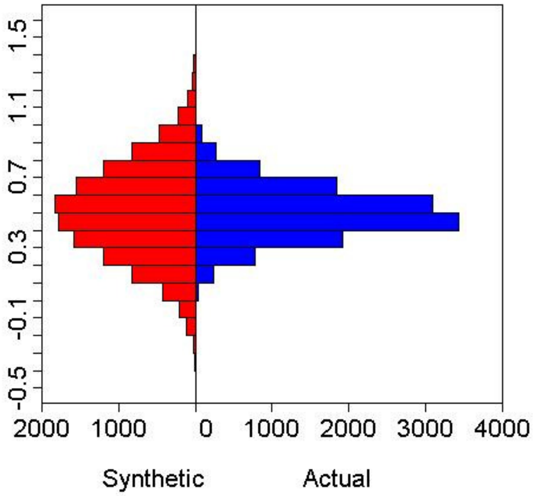


Figure 4.2: County-level estimates for nonsampled counties

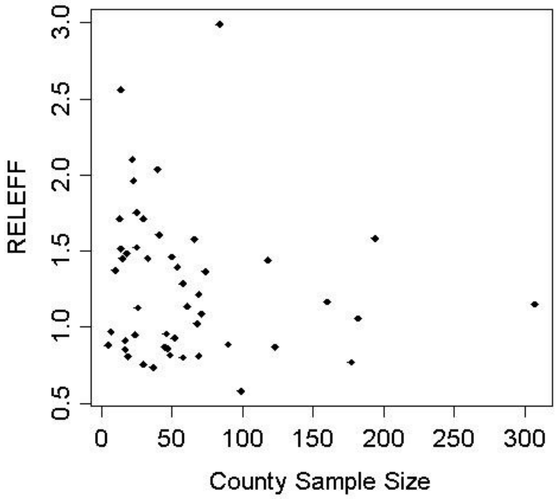


Figure 4.3: Efficiency loss of county-level estimates by size of sampled county

We also evaluate the repeated sampling properties of these estimates. The coverage rates for all 100 county-level estimates from the synthetic data range from 90% to 99%, and the average coverage rate is 95%. In contrast, the coverage from the actual data ranges from 92% to 99% with average value of 95%. The two sets of coverage rates are similar. The average biases on these estimates from synthetic and actual data are almost identical and both very close to zero (both are -.0015). The average Mean Square Error (MSE) is 0.555 and 0.533 from the synthetic and actual data respectively. Considering the fact that these estimates are unbiased, this relative larger MSE value from analyzing the synthetic data suggests again a slight loss in efficiency. Consistent with the finding on relative efficiency, all of the three statistics, coverage rate, bias and MSE, on sampled county-level estimates present decreasing relationships with county sample size as shown in Figure 4.4-4.6. These relationships are similar across synthetic and actual data.

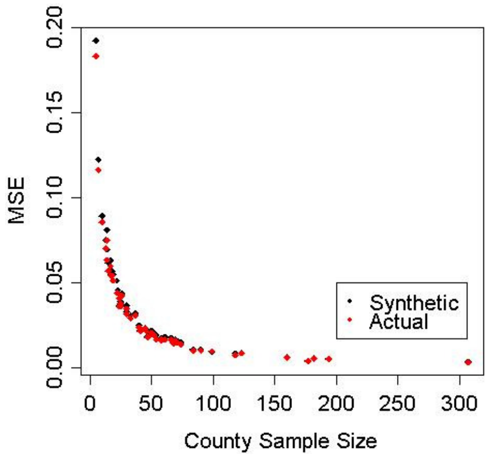


Figure 4.4: Mean Square Error (MSE) as functions of county sample size

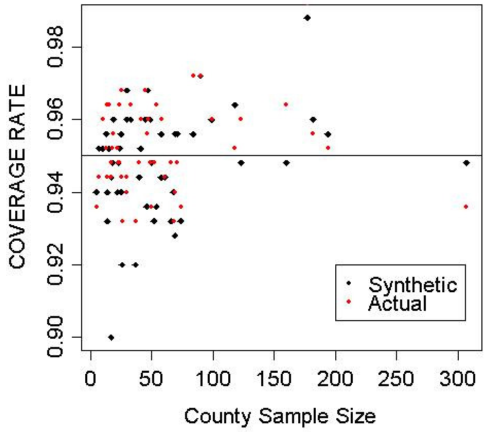


Figure 4.5: Coverage-rate as functions of county sample size

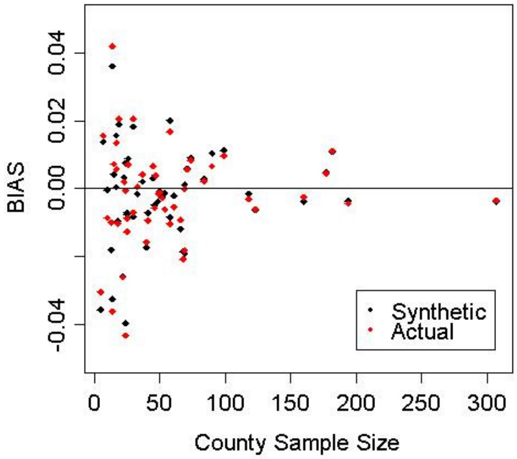


Figure 4.6: Bias as functions of county sample size

Statistical inferences for statistics at state-level are also compared. The sampling expectations for estimating the direct state mean are almost identical between synthetic, 0.507 and actual data, 0.518. The bias (0.007) is almost zero. The MSE from synthetic data, 0.023, is slightly smaller than that from the actual data, 0.035, which is again due to the strength from synthesizing values of nonsampled areas and subjects.

In sum, in this simulation, both the imputer's model and the analyst model agree with the true model, which is a univariate normal random effect model. The synthetic data are optimal in terms of preserving the distributional properties of the actual data. The synthetic data inferences are valid for both small area and state-level statistics from the repeated sampling point of view, except the fact of being slightly less efficient than the actual data.

4.3.2. Simulation study 2: a mixed-type bivariate situation

4.3.2.1. Generate actual survey sample

The survey data are generated as below:

1. Draw c counties from a total of C population counties.
2. Generate $v_i = \left(\text{logit}\left(\theta_i\right), \beta_{0i}, \beta_{1i}\right), i = 1, 2, ..., c$, from a 3-variate normal distribution with mean μ and variance-covariance matrix Σ.
3. Within each selected county, b_i subjects are randomly drawn from its population. $b_i, i = 1, 2, ..., c$ is calculated by proportional allocation.
4. Individual values of $x_{ij}, j = 1, 2, ..., b_i$ are drawn from $\text{Bernoulli}\left(\theta_{\text{i}}\right)$.

5. Conditional on the values of x_{ij}, draw values y_{ij} from $N\left(\beta_{0i}+\beta_{1i}x_{ij},\sigma_e^2\right)$, where $\sigma_e^2=1$. The observed sample is then denoted by $D=\left(I_i,x_{ij},y_{ij},i=1,2,...,c;j=1,2,...,b_i\right)$.

4.3.2.2. Generate synthetic data

We again use the Gibbs sampler to generate synthetic data sets as follows:

1. To draw $\left\{\Sigma^*,\mu^*=\left(\operatorname{logit}\left(\theta_i\right)^*,\beta_{0i}^*,\beta_{1i}^*;i=1,2,...,c\right),\sigma_e^{2*}\right\}$ for the sampled areas, we use the Gibbs sampler to draw $\left\{\Sigma,\mu=\left(\operatorname{logit}\left(\theta_i\right),\beta_{0i},\beta_{1i};i=1,2,...,c\right),\sigma_e^2\right\}$ from their joint posterior distribution. For nonsampled areas, draw $\left\{\Sigma^*,\mu^*=\left(\operatorname{logit}\left(\theta_i\right)^*,\beta_{0i}^*,\beta_{1i}^*;i=a+1,a+2,...,C\right)\right\}$ from $MVN\left(\mu^*,\Sigma^*\right)$.
2. Define area sample sizes for both sampled and nonsampled clusters as $b_i^*=f\times B_i,i=1,2,...,C$.
3. Draw values $x_{ij}^*,j=1,2,...,b_i^* \mid i,b_i^*,\operatorname{logit}\left(\theta_i\right)^* \sim \text{Bernoulli}\left(\theta_i^*\right)$.
4. Finally, draw values $y_{ij}^*,j=1,2,...,b_i^* \mid i,b_i^*,x_{ij}^*,\beta_{0i}^*,\beta_{1i}^* \sim N\left(\beta_{0i}^*+\beta_{1i}^*x_{ij}^*,\sigma_e^{2*}\right)$. The synthetic data set is then denoted by $d_{syn}=\left\{\left(I_i,x_{ij}^*,y_{ij}^*\right),i=1,2,...,C,j=1,2,...,b_i^*\right\}$.
5. Repeat steps 1 to 4 a total of M times to get M synthetic data $D_{syn}=\left(d_{syn}^1,d_{syn}^2,...,d_{syn}^M\right)$.

4.3.2.3. Results

The model parameters specific to this simulation are

$$\sigma_e^2 = 1,\ \mu = [0.5, 0, 0] \text{ and } \Sigma = \begin{vmatrix} 0.5 & 0.1 & 0.1 \\ 0.1 & 1 & 0.5 \\ 0.1 & 0.5 & 1 \end{vmatrix}.$$

We generated 250 replicates and $M = 20$ synthetic data sets for each replicate. We evaluated small area estimates, state-level descriptive mean estimates and two types of state-level regression coefficients estimates, i.e. Ordinary linear regression (OLS) and Logistic regression. The total number of estimands is 206. They include 50 sampled small area means, 50 nonsampled small area means and one state-level direct mean estimate of Y (continuous data); the same set of 101 estimands for the binary variable X; 1 intercept and 1 slope for the linear regression of Y on X; and another 1 intercept and 1 slope for the logistic regression of X on Y.

All 51500 synthetic variance estimates are positive. To ensure fair comparison between synthetic data estimates and actual data estimates, we assume one data user is interested in inference on either variable but not both at the same time[1]. Therefore, actual data county-level inferences are based on marginal models with only one variable involved. The point estimates of small area statistics are unaffected by this assumption about the analyst model. The only possible implication, however, is that the estimation is generally more efficient when the correlation between the two variables is considered.

Figure 4.7 and Figure 4.8 show the scatter plots of the direct county means for all sampled counties from the synthetic data and corresponding posterior means from the actual data on the normal variable Y and binary variable X across all replicates

[1] We also simulated the situation where one data user is interested in making county-level inference on both variables, and the results are similar to what we present under the current assumption

respectively. The fact that the two sets of small area mean estimates on both variables line up very well provides evidence that synthetic data yield almost identical estimates.

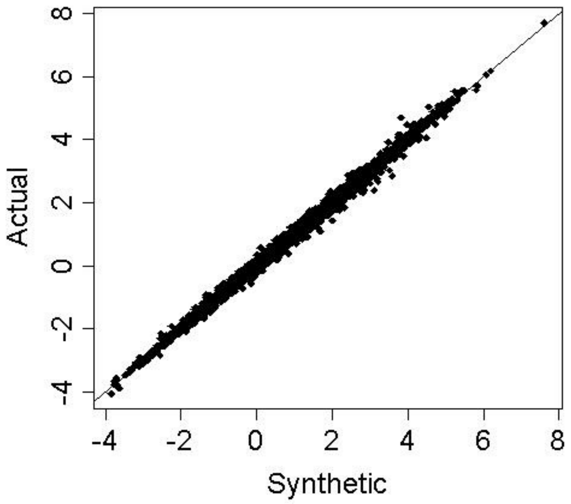


Figure 4.7: Area means for continuous variable Y from synthetic and actual data.

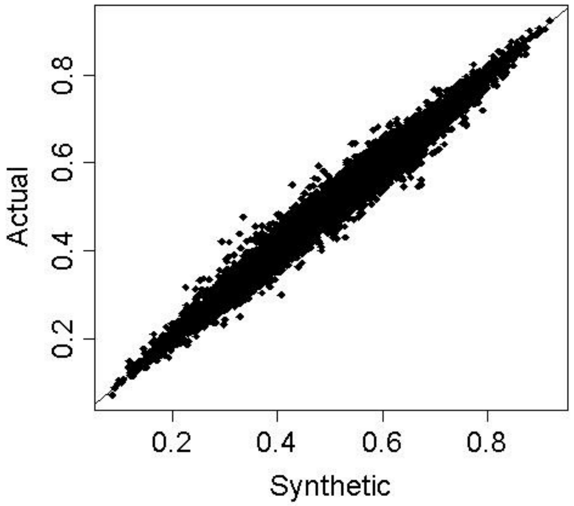


Figure 4.8: Area means for binary variable X from synthetic and actual data

Similarly as in simulation study 1, we fit separate simple linear regressions of the actual data estimates on the synthetic data estimates of sampled areas for X and Y as shown in Table 4.1. The fact that both intercepts and slopes are not significantly different from their respective expected values suggests that the synthetic estimates are very similar to the actual estimates.

Table 4.1: The correlation statistics of the synthetic and actual estimates for sampled counties

	Expected values	*X*		*Y*	
		Estimate	S.E.	Estimate	S.E.
Intercept	0	0.0009	0.0006	0.0018	0.0007
Slope	1	0.9838	0.0012	0.9981	0.0005

Next we evaluate the simple mean statistics, the linear and logistic regression coefficients based on the entire state. As shown in Figure 4.9, for state-level means, means of the continuous variable and proportions of the binary variable, the synthetic data estimates are very similar to the corresponding actual data estimates as these two sets of estimates line up around the 45 degree line. Table 4.2 shows the regression intercept and slope coefficients of the actual data estimates on the synthetic data estimates for these six sets of state-level estimates respectively. The fact that all intercepts and slopes are not significantly different from their expected values, 0 and 1, provides further evidence for such similarities.

Table 4.2: The correlation statistics of the synthetic and actual estimates for state-level statistics

		Expected value	*X*		*Y*	
			Estimate	S.E.	Estimate	S.E.
Mean	Intercept	0	-0.016	0.016	-0.009	0.034
	Slope	1	1.029	0.033	1.004	0.041
Regression Intercept	Intercept	0	-0.003	0.018	0.009	0.009
	Slope	1	0.999	0.035	1.023	0.036
Regression Slope	Intercept	0	-0.011	0.028	-0.004	0.009
	Slope	1	1.012	0.041	1.016	0.037

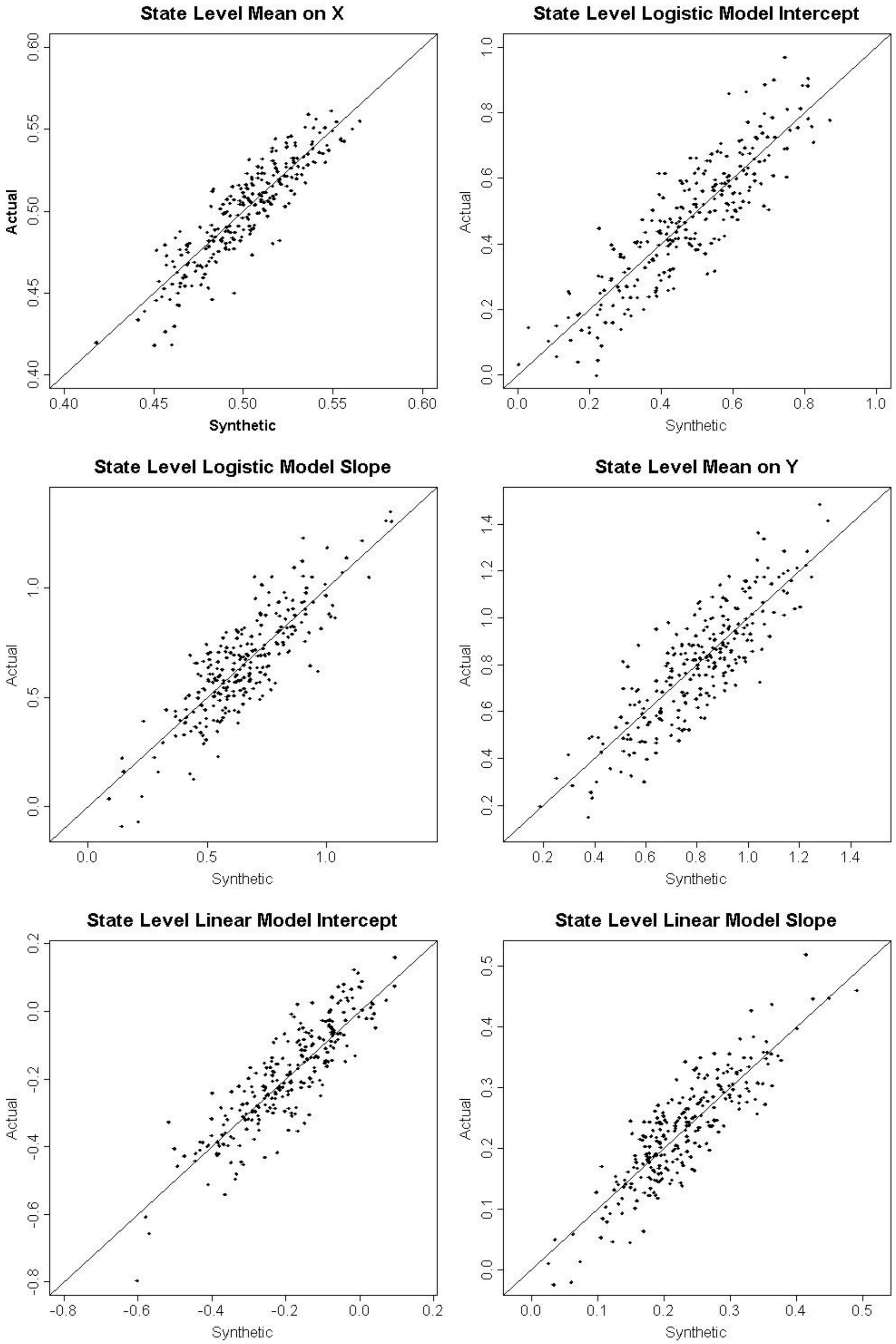


Figure 4.9: Scatter plots of state-level means for X and Y, as well as state-level linear and logistic regression intercepts and slopes estimates involve X and Y.

We next compare the inferences from a repeated sampling perspective. Table 4.3 shows the sampling expectation and variance over repeated samples, as well as the coverage, bias and mean square error for the small area statistics. All expectations are similar between synthetic and actual data, which suggests the point estimates are unbiased from the repeated sampling viewpoint. On the other hand, the fact that all synthetic data variances are larger than their corresponding actual data variance suggests there is a slight loss in estimation precision. Both average coverage rates for small area estimates are below the nominal level, though the average coverage rate for small area estimates on Y is slightly better than that on X. The biases of synthetic data estimates are almost zero, which suggest these estimates are unbiased on average. In addition, consistent with the findings from the first study, the coverage rate, bias and MSE are all negatively correlated with county sizes.

Table 4.3: Repeated sampling properties for county-level statistics

		Expectation	Variance	Coverage	MSE	Bias
Y	Synthetic	0.805	1.036	0.947	0.991	0.010
	Actual	0.804	0.991	0.948	0.945	0.008
X	Synthetic	0.501	0.017	0.952	0.016	0.000
	Actual	0.502	0.016	0.954	0.016	0.000

From Table 4.4, the expectations for all state-level statistics, including both descriptive and analytic statistics are almost identical across synthetic and actual estimates. The synthetic data variances larger than those from actual data are due to this disclosure limitation procedure. The magnitude of variance inflation is more significant for the state-level statistics than for the small area statistics. The relative efficiency, in expectation, ranges from 8, for estimating the slope coefficient for the linear model of Y on X, to 75 for estimating the state-level mean of Y.

Table 4.4: Repeated sampling properties for state-level statistics

			Expectation	Variance
Y	Mean	Synthetic	0.805	0.075
		Actual	0.799	0.001
X	Mean	Synthetic	0.501	0.001
		Actual	0.500	0.000
Linear Regression Y~X	Intercept	Synthetic	-0.196	0.028
		Actual	-0.192	0.002
	Slope	Synthetic	0.228	0.008
		Actual	0.228	0.001
Logistic Regression X~Y	Intercept	Synthetic	0.475	0.046
		Actual	0.472	0.002
	Slope	Synthetic	0.655	0.074
		Actual	0.651	0.004

In sum, these results provide evidence that the inference from synthetic data is valid for both county-level and state-level inferences. The geographic area random effect imputation model not only reproduces the small area data structure, but also ensures the statistical integrity for inferences involving larger areas. Based on the above results, we can conclude that the inference from synthetic data sets is valid from the Frequentist point of view. The information loss in estimation precision is very small. Thus releasing multiple synthetic data is a very promising method for protecting confidentiality while allowing high quality research on small geographic areas.

4.4. An empirical study

4.4.1. Data source

The 1880 Census Integrated Public Use Microdata Series (IPUMS) is a 1% national random, representative sample of the United States population. It is comprised of approximately 107,000 household records and 503,000 person records. The small geographic areas of interest in this study are counties. To reduce the computational burden, we randomly chose a subset of 1880 IPUMS, which includes 2877 households for a total of 11,408 people from the region of South Pacific. This data set covers all 3 states within this region, i.e. California, Washington and Oregon, and all 97 counties within these states. The county sample sizes vary from 4 to 2425. To illustrate the application of this approach in real data, we consider fully synthesized values of 3 variables: Age, Gender and SEI (Duncan's Socioeconomic Index).

We select these three variables not only to present several generic modeling challenges, but also due to the fact that they are routinely collected survey variables, therefore, are most likely to be used by an intruder in the re-identification of survey respondents. Among these three, SEI is a widely used measure of occupational standings and it plays an important role in studies of stratification in the United States (Duncan et al., 1972). The types of variables presented in this data set include numerical, binary, and semicontinuous. The empirical distribution of some numerical variables deviates largely from normality and/or is zero inflated. The data structure presented here is sufficiently complex to permit a wide variety of analyses and can be used to speculate the creation of synthetic data for the ACS. The description of these variables is presented in Table 4.5.

Table 4.5: Description of the variables used in this empirical study

Variable	Type	Data Range
Age	Continuous	1-110 years
Gender	Binary	0: Male; 1: Female
SEI	Semicontinuous	0-96

4.4.2. Generate synthetic data

To capture the zero-inflated probability mass for SEI, we create a zero-status indicator, SEI.d. For a respondent, SEI.d takes the value 1 if his SEI is greater than zero, and 0 if it equals zero. We normalize Age and non-zero SEI by Box-Cox power transformation (Box and Cox, 1964). For simplicity, we round the estimated power parameters for the

Box-Cox transformation for Age and non-zero SEI to 0.5 and -0.5 respectively, which are used in creating the transformed variables. Table 4.6 shows the estimated Box-Cox transformation parameters, the corresponding standard errors and the rounded parameters. Figure 4.10 shows the quantile-quantile (q-q) plots for Age and Non-zero SEI in their original scale and Box-Cox transformed scale based on the rounded power parameters respectively. If the plots approximate straight lines, the distributions for the variables are close to normality. Comparisons of the plots between the orignal and transformed scale suggest Box-cox transformations significantly improve normality for both Age and SEI.

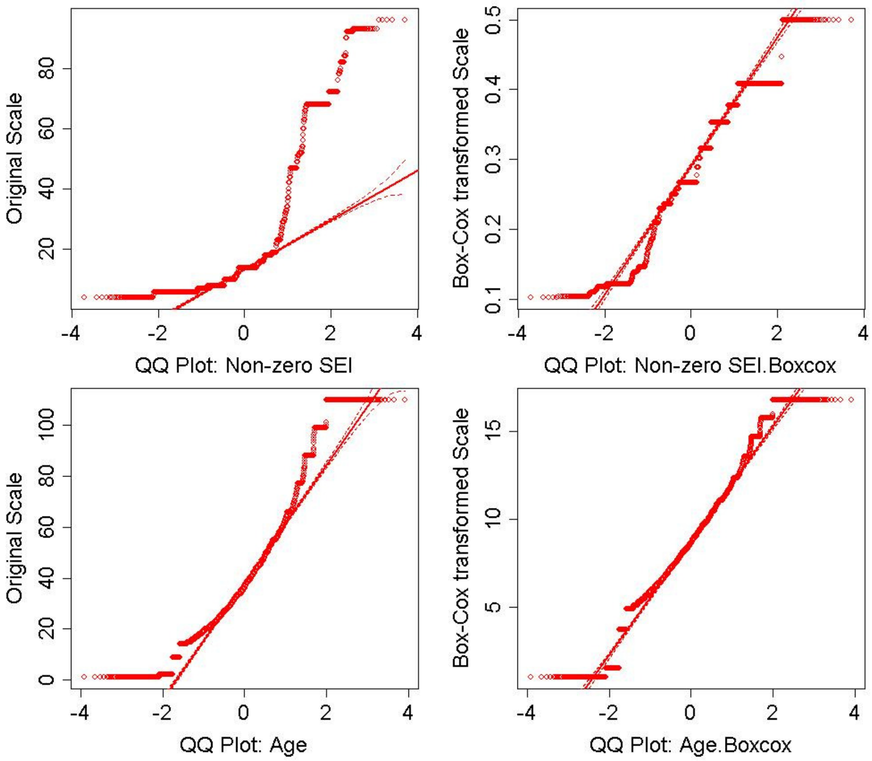


Figure 4.10: Q-Q plots for Age and Non-zero SEI in the original and Box-Cox transformed scales

Table 4.6: Rounded and estimated Box-Cox transformation power parameters

	Rounded Power Parameters	Estimated Power Parameters (SE)
Age	0.5	0.55 (0.0093)
SEI	-0.5	-0.4751 (0.0187)

To ensure the synthetic data distributes within the same range as the actual data, we use truncated normal distributions for both Age and SEI. We have attempted the truncated t-distributions to capture the outliers. We treat the degrees of freedom as unknown parameters and assigned non-informative Gamma prior distributions (Xie, Raghunathan and Lepkowski, 2007). Since the estimated degrees of freedom for both Age (estimated degrees of freedom equals 15) and SEI (estimated degrees of freedom equals 30 plus) are quite large, this means the distributions are very close to the normality. Thus, we consider normal distributions for our final imputation model for simplicity. The synthetic data generation model is as below.

Stage 1. Sampling Model

$\text{Gender}_{ij} \sim \text{Bernoulli}\left(\theta_i\right);$

where, $i = 1,2,...,C;\ j = 1,2,...,b_i;\ C = 97;\ b_i \in \left[4, 2425\right]$

$\text{SEI.d}_{ij} \mid \text{Gender}_{ij} \sim \text{Bernoulli}\left(\alpha_{1i} + \beta_{1i}\text{Gender}_{ij}\right)$

$\text{SEI}_{ij} \mid \text{Gender,SEI.d} = 1 \sim \text{TN}\left(\alpha_{2i} + \beta_{2i}\text{Gender}_{ij}, \sigma_1^2, a^1, b^1\right)$

$\text{Age}_{ij} \mid \text{Gender,SEI} \sim \text{TN}\left(\alpha_{3i} + \beta_{3i}\text{Gender}_{ij} + \beta_{4i}\text{SEI}, \sigma_2^2, a^2, b^2\right)$ [3.22].

where, $\text{TN}\left(\mu, \sigma^2, a, b\right)$ denotes the truncated normal distribution with mean μ, variance σ^2, truncation points a and b.

Stage 2. LinkingModel

$\left[\text{logit}\left(\theta_{\text{i}}\right), \alpha_{1i}, \beta_{1i}, \alpha_{2i}, \beta_{2i}, \alpha_{3i}, \beta_{3i}, \beta_{4i}\right] \sim MVN_8\left(\mu, \Sigma\right)$

Stage 3. Independent diffuse priors for the model parameters $\left(\mu, \Sigma, \sigma_1^2, \sigma_2^2\right)$

The synthetic transformed data are generated from the posterior predictive distributions estimated based on this following model fitted using WinBUGs. Then they are reversely transformed into their original scales. Both Age and SEI are discrete quantitative variables, which only take integer values. Therefore, it is desirable to round the imputed and reversely transformed noninteger values to the discrete scales. The county size is calculated proportional to the corresponding population sizes at the rate of 10% to allow the direct estimation of small area statistics from the synthetic data.

We again use the Gibbs sampler to generate synthetic data sets as follows. Since all counties are covered in this Census data, we only need to generate imputations for nonsampled individuals within each county to complete the synthetic population, and then draw random samples from each synthetic population to obtain the synthetic samples. Specifically, we generate the synthetic samples following the steps below.

1. To draw $\left\{\Sigma^*, \mu^* = \left(\log it\left(\theta_i^*\right), \alpha_{1i}^*, \beta_{1i}^*, \alpha_{2i}^*, \beta_{2i}^*, \alpha_{3i}^*, \beta_{3i}^*, \beta_{4i}^*; i = 1, 2, ..., C\right), \sigma_1^{2*}, \sigma_2^{2*}\right\}$, where C=97 , we use the Gibbs sampler to draw $\left\{\Sigma, \mu = \left(\log it\left(\theta_i\right), \alpha_{1i}, \beta_{1i}, \alpha_{2i}, \beta_{2i}, \alpha_{3i}, \beta_{3i}, \beta_{4i}\right), \sigma_1^2, \sigma_2^2\right\}$ from their joint posterior distribution.
2. Define county sample sizes as $b_i^* = f \times B_i, i = 1, 2, ..., C,$ where $f = 0.1$.
3. Draw values of $Gender_{ij}^*, j = 1, 2, ..., b_i^* \mid i, b_i^*, \text{logit}\left(\theta_i^*\right) \sim \text{Bernoulli}\left(\theta_i^*\right)$.
4. Draw values of $SEI.d_{ij}^*, j = 1, .., b_i^* \mid b_i^*, Gender_{ij}^*, \alpha_{1i}^*, \beta_{ij}^* \sim Bernoulli\left(\alpha_{1i}^* + \beta_{ij}^* Gender_{ij}^*\right)$. For an individual j in area i, whose value of $SEI.d_{ij}^* = 1$, assign $SEI_{ij}^* = 0$. For individuals whose value of $SEI.d_{ij}^* = 0$, then draw values of

$SEI_{ij}^{*} \mid SEI.d_{ij}^{*} = 0 \sim TN\left(\alpha_{2i}^{*} + \beta_{2i}^{*} \times Gender_{ij}^{*}, \sigma_{1}^{2*}, a^{1}, b^{1} \mid SEI.d_{ij}^{*} = 0\right)$, and transform the values back to the original scale by taking squares and finally round them to the nearest integers to be consistent with the actual data.

5. Finally, draw values $Age_{ij}^{*} \sim TN\left(\alpha_{3i}^{*} + \beta_{3i}^{*} Gender_{ij}^{*} + \beta_{4i}^{*} SEI_{ij}^{*}, \sigma_{2}^{2*}, a^{2}, b^{2}\right)$, and transform them back into its original scale by taking the inverse of square, and finally round them to the nearest integers. The resulted synthetic data set is then denoted by $d_{syn} = \left\{\left(Gender_{ij}^{*}, SEI.d_{ij}^{*}, SEI_{ij}^{*}, Age_{ij}^{*}\right), i = 1,...,C; j = 1,...,b_{i}^{*}\right\}$.

6. Repeat Step 1 to Step 6 M times to get the synthetic data, $D_{syn} = \left(d_{syn}^{1},...,d_{syn}^{M}\right)$.

4.4.3. Results

We evaluate the synthetic data utility by comparing the descriptive statistics, at both county-level and region level, and two types of regression coefficients, Ordinary linear regression (OLS) and Logistic regression, estimated from the synthetic data with those from the actual data. The total number of estimands is 599. They include six sets of 97 county means for Gender, Age, transformed Age, SEI.d, SEI and transformed SEI , 6 state-level direct mean estimates and 2 median estimates for these variables, and 9 linear and logistic regression coefficients of each variable on the remaining variables. The synthetic variance estimates are mostly positive, only about 2 of a possible 403 variance estimates are negative (i.e. approximately 0.5%).

Table 4.7 shows the region-level means, proportions or medians, as well as three sets of regression coefficients estimated from synthetic and actual data. The mean estimates are similar across synthetic and actual data except for Non-zero SEI. However, the mean is a good measure of central tendency for roughly symmetric distributions. As for a skewed distribution, it is sensible to be summarized using a typical value, in which median would serve as a good measure for such typicality. For both Age and SEI, the medians and intraquartile range estimated from the synthetic data match perfectly with the ones from the actual data.

Table 4.7: Comparisons on region-level descriptive and analytic statistics estimates from the synthetic and the actual data set.

Variable Name		Type	Synthetic		Actual		Z Score	Overlap Prob.
			Est.	SE	Est.	SE		
Descriptive Statistics								
	Sex	Mean	0.60	0.00	0.60	0.00	-0.31	0.93
	Age	Mean	40.01	0.21	40.82	0.22	-3.66	0.05
	Age	Median	36.00	5.57^*	36.00	5.57^*	0.00	0.95
	I(SEI=0)	Mean	0.56	0.01	0.56	0.00	0.32	0.94
	Non-zero SEI	Mean	17.35	0.14	20.09	0.27	-10.12	0.00
	Non-zero SEI	Median	14.00	3.32^*	14.00	3.32^*	0.00	0.95
	Transformed Age	Mean	6.05	0.02	6.09	0.02	-2.09	0.45
	Transformed Non-zero SEI	Mean	0.28	0.00	0.28	0.00	2.62	0.18
Logistic Regression								
Sex~ Age+SEI		Intercept	-0.55	0.05	-0.43	0.04	-3.00	0.24
	Age	Slope	0.00	0.00	0.00	0.00	1.18	0.82
	SEI	Slope	0.19	0.01	0.15	0.00	7.36	0.01
Linear Regressions								
Age~ Sex+SEI		Intercept	39.00	0.28	40.33	0.35	-3.78	0.02
	Sex	Slope	1.74	0.49	1.07	0.49	1.38	0.74
	SEI	Slope	0.00	0.01	-0.02	0.01	0.89	0.77
SEI~ Sex+Age		Intercept	1.51	0.14	2.07	0.33	-1.71	0.46
	Sex	Slope	10.38	0.20	11.89	0.29	-5.18	0.00
	Age	Slope	0.00	0.00	-0.01	0.01	0.98	0.68

Note: * denotes the Intra-quartile range (the range between 25th and 75th quartile).

The Z score in Table 4.7 is computed as the difference in point estimates from the synthetic and the actual data relative to the actual data standard error. A Z score with very small absolute value, for example, less than 2 or 3, suggests that the synthetic data and the actual data produce similar point estimates. Examining the absolute values of Z scores in Table 4.5, larger values tend to associate with statistics about SEI and Age in their original scales. The largest one is 10.12 for the mean of SEI and the second largest is for the slope coefficient for Sex when the dependent variable is SEI. Because both are measures about the central tendency of the distribution of SEI, they are very sensitive to extreme values and can be seriously contaminated even by one observation. Extreme values are very likely when we transform the imputed values of SEI back into its original scale by a power of -2, thus distorting the means.

The last column in Table 4.7 is the probability overlap in the confidence intervals for a scalar estimand Q (Karr, et al. 2006). By approximating the posterior distribution of estimand Q by normal distribution, the 95% confidence interval for the multiple synthetic data estimate $\overline{q}$ is $\left(L_{syn,\overline{q}}, U_{syn,\overline{q}}\right)$. Let $\left(L_{act,\hat{q}}, U_{act,\hat{q}}\right)$ be the corresponding interval for point estimate $\hat{q}$ obtained using the actual data which follows a t-distribution with $(n-p)$ degrees of freedom, where n and p are sample size and the number of parameters in an analyst model. Let $f_{syn,\overline{q}}$ and $f_{act,\hat{q}}$ be the estimated posterior distributions of Q computed using synthetic and actual data respectively. Therefore, the probability overlap in the confidence intervals for Q equals $I_Q = 2^{-1}\left[\int_{L_{syn,\overline{p}}}^{U_{syn,\overline{p}}} f_{act,\hat{p}}(t)\,dt + \int_{L_{act,\hat{p}}}^{U_{act,\hat{p}}} f_{syn,\overline{p}} dz\right]$. I_Q may take value $[0, 0.95]$. $I_Q = 0$ if there is no overlap and $I_Q = 0.95$ if the two intervals overlap perfectly. A large value on I_Q implies better data utility in estimating Q.

Almost all descriptive statistics have perfect confidence interval overlaps. Somewhat high overlap probabilities, in the range of 0.50 to 0.80, are found with the regression coefficients describing the bivariate relationships among Age, Sex and SEI. The only exception is the slope coefficient of Sex on SEI, where the overlap probability is only 0.1, and it is also detected by Z scores.

Figure 4.11 shows the histograms of SEI in the transformed and original scales. To ensure the cell frequencies of the synthetic data and the actual data are in a comparable scale, we choose one synthetic data randomly and compare its distribution with the actual data. The shape of the distribution is preserved although the outliers at the high end of distribution in actual data are not reflected perfectly.

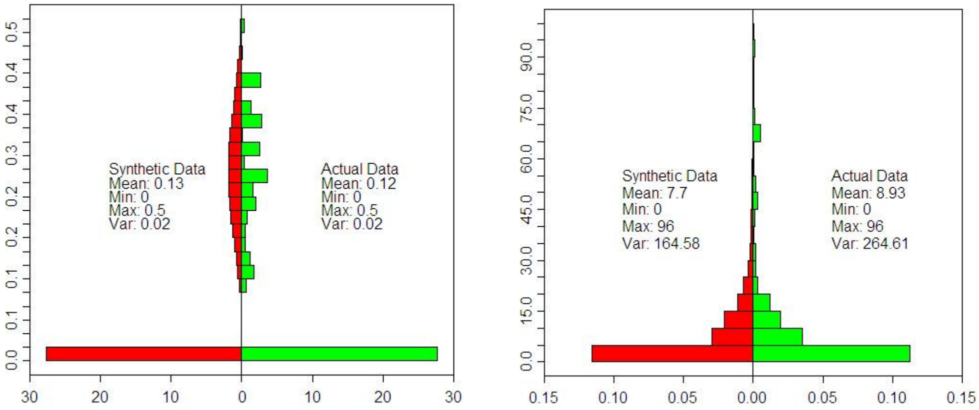


Figure 4.11: Histogram of SEI from one randomly chosen synthetic data and the actual data on transformed scale (left panel) and original scale (right panel)

It is interesting that for some regression slopes, the synthetic data estimates are larger than the actual data estimates in their absolute values. For example, the slope coefficient for SEI when we regress Sex on Age and SEI (synthetic 0.19 vs. actual 0.15), the coefficient for SEI in the regression of Age on Sex and SEI (synthetic 0.00 vs. actual -0.02), and the coefficient for Age in the regression of SEI on Age and Sex (synthetic 0.00 vs. -0.01). A close look at these statistics gives us the similarity among these estimates, which is that they all are highly associated with the variable SEI. Such contradiction may be due to the inadequacy of replicating the actual sampling distribution of SEI in the synthetic data, which leads to the distortion of estimating the regional level statistics.

Suppose that data users are interested in the county-level means or proportions on Age, Sex and SEI. Given that the county sample size is too small to permit the direct estimation, indirect estimators based on a small area model are appropriate. Under the parametric approach, one option of normalizing both Age and SEI to meet the model assumption on normality is Box-Cox transformation. Then indirect means on Sex and the transformed Age and SEI are the posterior county means from the small area model on the transformed data. Because data users may be more interested in inference on the original scale, we also obtain the posterior means in the original scales by simulating the corresponding posterior distributions.

Table 4.8 summarizes the comparison of county means on 2 binary variables, Sex and SEI.d, 2 continuous variables in the original scales, Age and SEI, and 2 in their transformed scales. On average, the synthetic and actual data produce similar county-level means. The average overlap probabilities are very close to 0.95, which suggest the confidence intervals constructed based on the synthetic and actual data are very close. The fact that all z-scores are close to zero provides evidence that the point estimates are very similar.

Table 4.8: County-level means estimated from synthetic and actual data

Variable	Synthetic Data		Actual Data		Overlap Prob.	Z score
	Estimate	SE	Estimate	SE		
Sex	0.62	0.05	0.62	0.04	0.93	-0.04
SEI.d	0.55	0.05	0.55	0.05	0.93	0.03
Trans.Age	6.05	0.14	6.05	0.15	0.92	-0.02
Age	39.99	1.67	40.01	1.86	0.92	-0.01
Trans.SEI	0.13	0.02	0.13	0.02	0.93	-0.03
SEI	7.13	1.23	7.17	1.32	0.93	-0.04

As shown in Figure 4.12, the estimates from synthetic and actual data are clustered around the 45-degree line for all six variables, which indicates all county means are similar regardless of their county sample sizes. The correlation statistics of evaluating the closeness of two sets of estimates from the synthetic data and the actual data respectively are shown in Table 4.9. If a simple regression of the synthetic data estimates on the actual data estimates fit the line of identity (intercept = 0, slope = 1), then the estimates are identical. The non-significant Wald-tests suggest the synthetic data estimates on county means are very similar to the actual data counterparts for all variables involved in this study.

Table 4.9: The correlation statistics of the synthetic and actual estimates for region-level statistics

Variable		Expected Value	Estimate	S.E.
Sex	Intercept	0	0.017	0.014
	Slope	1	0.970	0.022
Age	Intercept	0	1.629	1.819
	Slope	1	0.959	0.045
Age boxcox	Intercept	0	0.221	0.273
	Slope	1	0.963	0.045
SEI.d	Intercept	0	0.015	0.009
	Slope	1	0.976	0.016
SEI	Intercept	0	-0.069	0.173
	Slope	1	1.004	0.024
SEI boxcox	Intercept	0	0.002	0.002
	Slope	1	0.978	0.015

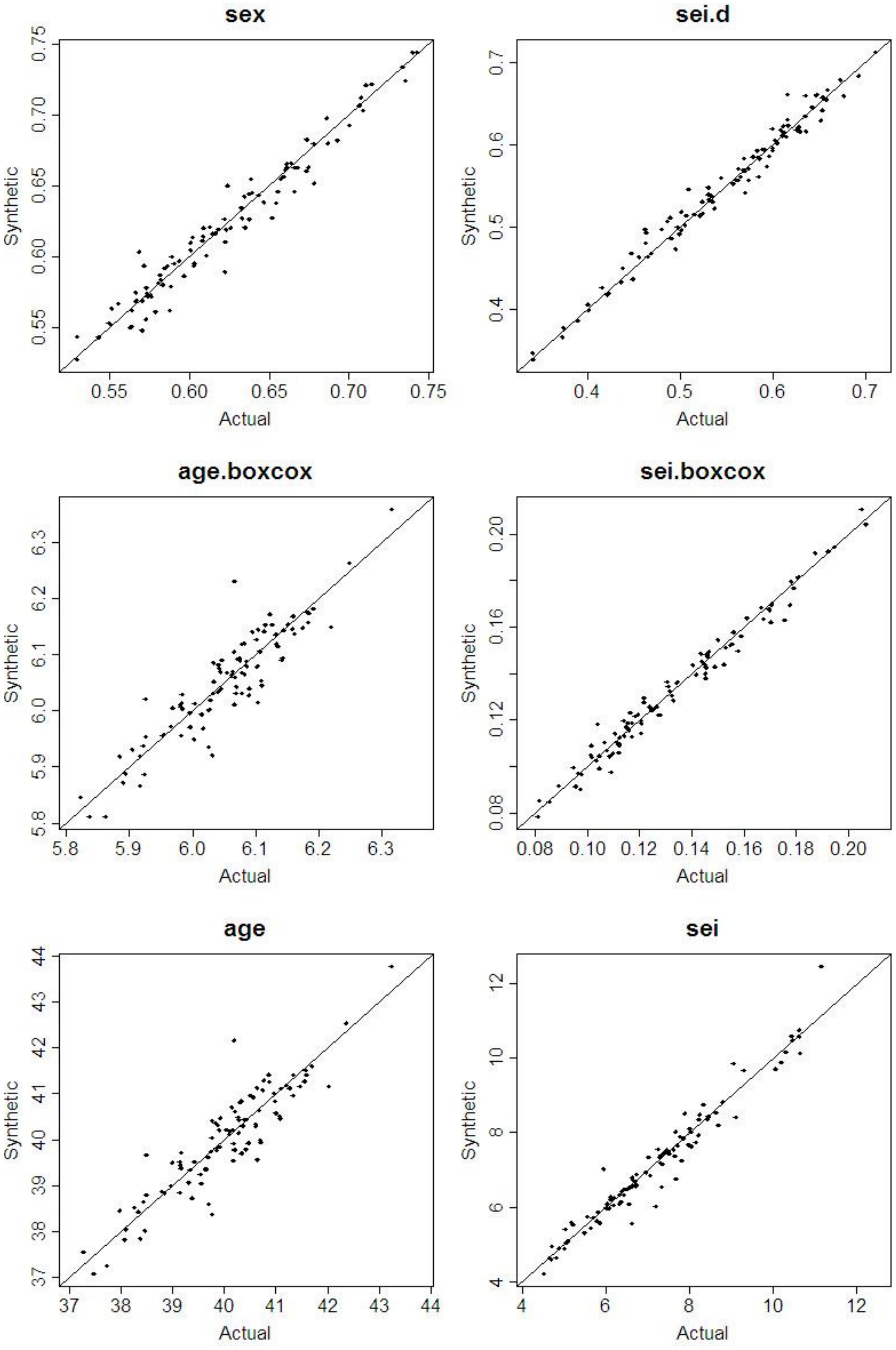


Figure 4.12: County-level mean estimates from synthetic and actual data

In conclusion, this empirical illustration shows this synthetic data approach provides similar statistical inferences on both small area levels as well as region-level as the actual data. The techniques that we demonstrate in dealing with non-normality, outliers and zero-inflated data can potentially be applied to a situation that involves a larger data set.

4.5. Conclusion

In this chapter, we evaluate this method of using multiple imputation techniques to create fully-synthetic data for small area estimation both theoretically and empirically. Compared with the current practices of using geographic threshold and restricted data access, the synthetic data method not only offers the data users full flexibility to conduct customized geographical analysis, but also has the potential to extend the scope of analysis to the non-sampled areas, which are not contained in the actual data.

Synthetic microdata generated from the posterior predictive distributions built on properly specified small area models yield similar statistical inference on small area level statistics with the ones on the actual data. Via this 1880 Census empirical study, we successfully demonstrated the solutions for several common modeling challenges, such as non-normality, outliers, bounds and zero-inflation. These techniques can be adapted easily for synthesis projects that involve large-scale survey data such as the American Community Survey, which will have significant impacts in a broad-spectrum of areas such as demographics, sociology and economics.

CHAPTER V

CONCLUSIONS AND DISCUSSIONS

5.1. Summary of this dissertation

The issue of data confidentiality has become increasingly important because of the improvements in record linkage technology and ease of access to electronic databases. It concerns both federal statistical agencies, who constantly face the pressure to publish high quality data, and researchers, who depend upon the collaboration of individuals and businesses in data sharing.

This dissertation addresses two topics in statistical data confidentiality: disclosure risk assessment and the synthetic data method for disclosure control. The two topics are closely related in two ways. The risk of disclosure is usually used to inform whether a statistical disclosure avoidance procedure should be implemented to the actual data, and if such procedures should be applied, then to which data records. In addition, the comparison of estimated disclosure risks for a survey data set prior to and after the modifications offers another way to evaluate an SDC method. Under this framework, the best SDC procedure is the one that achieves the optimal trade-off between disclosure risk reduction and data utility loss.

This dissertation, rather than attempting to offer an "end-to-end" solution by using disclosure risk as a guidance and/or checkup for the SDC procedures, provides answers to three separate questions that have puzzled the field. The reason for lack of connections between the chapters is that such linkage between disclosure risk and SDC methods is meaningless in the case of this dissertation. Both SDC methods we proposed involve a full synthesis of the values of all survey variables in all subjects, thus, each record of the resulting microdata can no longer be interpreted as having originated from a given individual, which leads to no grounds for evaluating the risk of being re-identified.

Specifically, Chapter 2, 3 and 4 identified and provided answers to three separate research questions. Chapter 2 is motivated by the prevailing concern of increased risk of disclosure of survey respondents due to the availability of commercial databases with identifying information and key demographic variables coupled with powerful record linkage techniques. This study illustrated, theoretically and empirically using an empirical experiment, the significant impacts of four sources of uncertainties on disclosure risk assessments. The uncertainties include assumptions about the amount of identification information to which an intruder has access, the accuracy of such identification information and the under-coverage of the commercial data. The latter two factors, both largely ignored in the literature, worked collectively in the direction of reducing the risk of disclosure. This finding, when used to inform SDC procedures, is very reassuring to the field as it suggests that fewer data modifications are needed to achieve a desired amount of disclosure prevention.

Chapter 3 demonstrates a successful practical implementation of fully-imputed synthetic data for a large complex longitudinal survey as means to protect confidentiality. Ever since this approach was initially proposal by Rubin (1993) and Little (1993), skepticism about its feasibility has abounded. This research filled this feasibility gap by synthesizing the values of about one hundred variables of different types derived from more than 12,000 cases, from a national longitudinal survey. Separate semiparametric imputation algorithms for continuous, binary and categorical variables were developed and tested. We also developed a new combining rule for synthetic data inference to account for the imputation variation due to both item-missing data and synthetic data. Data utility of the synthetic data is comparable to that of the original data. The imputation models illustrated in this chapter can be easily adapted to resolve confidentiality issues for other large-complex surveys.

The third study, discussed in Chapter 4, extends this fully-synthetic data approach to cope with situations where small area statistics are of vital importance. This study is the first in the SDC literature to respond to this ever-increasing demand for small-area microdata. The goal was to create synthetic microdata with enough geographical detail to permit small area analyses, which otherwise is not permitted, because such geographical identifiers are usually suppressed due to disclosure control. Small area models developed under a Bayesian framework are used to generate synthetic data. This approach is evaluated by the use of an empirical example in addition to a series of simulations. Both small-area statistics and national level statistics based on synthetic data are similar to those obtained from the original data. Moreover, the modeling burden is reduced, especially when public data users attempt to produce small-area statistics from the synthetic data.

5.2. Future research

5.2.1. Generalization of results from the risk assessment experiment

In Chapter 2, the evaluation of disclosure risk was facilitated by the per-record measures of measurement discrepancies and the disaggregated measures of under-coverage rates for a large number of subclasses. However, in real practice, assessment of measurement discrepancies between the survey data and the commercial data are rarely attainable. Therefore, despite the importance of these two factors in risk assessments, the direct estimation of such factors is not possible. Methods for incorporating the uncertainties due to these two types of errors into the assessments of the risk of disclosure when the exact estimates are not available, are needed.

5.2.2. Synthetic data utility assessments

In both Chapter 3 and Chapter 4, we evaluated the validity of synthetic data inference separately for a singular estimand by assessing closeness of the two confidence intervals obtained from the actual data and synthetic data. We assumed that estimands are independent from each other, thus the correlation among estimands are ignored in the evaluation of the inference validity. The current approach has the advantages of simplicity and easy interpretation. However, it is very unlikely that the assumptions would hold. For instance, the estimated regression coefficients from a multivariate regression model are usually correlated, and statistics for one small area, especially the ones obtained based on "indirect estimators[2]", by definition, are related to those for other small areas. Furthermore, the evaluation of data utility on multi-dimensional statistics, such as the joint confidence intervals of two or more estimands, which may be of interest

[2] "Indirect estimators have been characterized in the empirical Bayes literature as estimators that "borrow strength" by incorporating values of the variable of interest from units in domains other than the domain of interest" (U.S. Office of Management and Budget, 1993).

to public data-users, is not permitted. Future research should include the evaluation of information loss for multivariate statistics.

In addition, synthetic data utility is evaluated by comparing the estimates with those from the actual data. Information loss is considered negligible if similar estimates are obtained from the two data files. However, a disadvantage of this evaluation method is the lack of definition for what constitutes similarity (or dissimilarity). Diagnostic tools are needed for assessing how sensitive the current data evaluation methods are to different definitions of "similarity".

5.2.3. Future research associated with synthetic data for small-area estimation

In Chapter 4, we provide a small area synthetic data model incorporating three types of variables: continuous, binary and semicontinuous. Another common type of survey variable is a categorical variable with more than two levels. Model ill-conditioning is a problem that is very likely to occur in a small-area model if an unequally distributed high-dimensional categorical variable correlates with the small areas for which statistics are desired. Further research is needed to deal with this situation.

Another modeling challenge may occur when we plan to create synthetic data for a large number of variables of different types. Specifying a proper joint-distribution model, as the data set dimensions increase, may become difficult or even impossible. To deal with this situation, one may adopt the idea of sequential conditional regression (Raghunathan, et al. 2001). We demonstrated this approach in Chapter 3 when we generated synthetic data for microdata without geographical area structures. In specific, the joint-distribution is replaced with a series of conditional distributions of one outcome variable conditional on the rest of the outcome and auxiliary variables. One iterates the process of drawing imputed values from each conditional posterior predictive distribution so that the final values converge to draws from the multivariate distribution.

We model the semi-continuous variable via assuming the non-zero portion of data follows a normal distribution after the Box-Cox transformation. However, the several peaks of density at the right tail of this distribution were not fully captured which distorts the validity of regional level analysis. This type of distribution is very common in survey data. Two common causes are 1) rounding to the nearest integer when survey respondents report their answers and 2) aggregating several closely related variables in creating another summary variable. Possible variables may be Age, Income, Education (in years) etc. Potentially, such model inadequacy can be overcome by further relaxing the distributional assumption. One can employ a more non-parametric approach by breaking the distribution into more segments and fit separate models for each segment of data. Another option is to fit a more complex parametric model, such as a Tukey's gh distribution model, with additional model parameters to capture such irregularity in the empirical distribution.

Most survey data are realizations of the population from complex sample designs. Without losing generalizability, this method can be adapted to solve the disclosure problem for such sample data. A complex multistage design is usually well-approximated by a two-stage stratified clustered sampling design with strata formed in the first stage, clusters selected within stratum, as well as weighting to account for unequal sampling probabilities, nonresponse, and postratification. Among these three factors, stratification can be incorporated by building an imputation model within each stratum, and clusters can instead be treated as random effects. The most challenging task is how to deal with survey weights. The general principles of modeling weights imply that models should be constructed incorporating all variables that affect each component of weighting. However, such models can quickly become overwhelmingly complicated as the number of the adjustment cells increase (Gelman 2007). Therefore, we propose to treat survey weights as a scalar summary of the variables related with unequal selection, nonresponse and postratification, and to then incorporate it as a covariate in the regression imputation model.

To improve the model efficiency, a set of area-specific auxiliary variables such as the number of households, per capita income (PCI), value of housing, geographic size, social demographic decompositions, etc., can be extracted from external sources and used as covariates. Incorporating such information into the synthesis model offers additional protection against model failure.

Lastly, unbiasedness of small area estimates with respect to survey sampling design is often desired. Future research on synthetic data models should consider ensuring that small-area estimates are calibrated with external design-based benchmarks.

BIBLIOGRAPHY

Abowd, J., and Woodcock, S. D. (2001), "Disclosure Limitation in Longitudinal Linked Data," in *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies*, eds. P. Doyle, J. Lane, L. Zayatz and J. Theeuwes, Amsterdam: North-Holland, pp. 215-277.

Armstrong, M. P., Rushton, G., and Zimmerman, D. L. (1999), "Geographically Masking Health Data to Preserve Confidentiality," *Stat Med*, 18, 497-525.

Atchley, W. R., Nordheim, E. V., Gunsett, F. C., and Crump, P. L. (1982), "Geometric and Probabilistic Aspects of Statistical Distance Functions," *Systematic Zoology*, 31, 445-460.

Balding, D. J. (2002), "The Dn a Database Search Controversy," *Biometrics*, 58, 241-244.

Benedetti, R., Capobianchi, A., and Franconi, L. (1998), "Individual Risk of Disclosure Using Sampling Design Information," *Contributi Istat 1412003*.

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), "Disclosure Control of Microdata," *Journal of american statistics association*, 85, 38-45.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (eds.) (1991), *Measurement Errors in Surveys*, New York: John Wiley & Sons, Inc.

Biemer, P. P., and Lyberg, L. E. (2003), *Introduction to Survey Quality*, New York: John Wiley &Sons, Inc.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, Series B (Methodological)*, 26, 211-252.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80.

Buckley, C. B., Angel, J. L., and Donahue, D. (2000), "Nativity and Older Women's Health: Constructed Reliance in the Health and Retirement Study," *Journal of Women and Aging*, 12, 21-37.

Cessie, S. L., and Houwelingen, J. C. V. (1992), "Ridge Estimators in Logistic Regression " *Applied Statistics*, 41, 191-201.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991), "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association*, 86, 68-78.

Cochran, W. G. (1977), *Sampling Techniques* (3rd ed.), New York: Wiley.

Cox, L. H. (1996), "Protecting Confidentiality in Small Population Health and Environmental Statistics," *Stat Med*, 15, 1895-1905.

Dalenius, T., and Reiss., S. P. (1982), "Data-Swapping: A Technique for Disclosure Limitation," *Journal of Statistical Planning and Inference*, 6, 73-85.

Dawid, A. P. (2001), "Comment on Stockmarr's "Likelihood Ratios for Evaluating DNA Evidence When the Suspect Is Found through a Database Search"," *Biometrics*, 57, 976-978.

de Waal , A. G., and Willenborg, L. C. R. J. (1995), "Statistical Disclosure Control and Sampling Weights.," Technical.

de Waal , A. G., and Willenborg, L. C. R. J. (1998), "Optimal Local Suppression in Microdata," *Journal of Official Statistics*, 14, 421-435.

Domingo-Ferrer, J., and Torra, V. (2003), "On the Connections between Statistical Disclosure Control for Microdata and Some Artificial Intelligence Tools," *Information Sciences*, 151, 153-170.

Duffy, D. E., and Santner, T. J. (1989), "On the Small Sample Properties of Normrestricted Maximum Likelihood Estimators for Logistic Regression Numbers " *Communications in Statistics - Theory and Methods*, 18, 959-980.

Duncan, G., and Lambert, D. (1989), "The Risk of Disclosure for Microdata," *Journal of Business & Economic Statistics*, 7, 207-217.

Duncan, O. D., Featherman, D. L., and Duncan, B. (1972), *Socioeconomic Background and Achievement*, New York: seminar press.

Elliot, M. J., Skinner, C. J., and Dale, A. (1998), "Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk," *Research in Official Statistics*, 1, 53-67.

Federal Committee on Statistical Methodology. (2002), "Identifiability in Microdata Files."

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, 1183-1210.

Fienberg, S. E., and Makov, U. E. (1998), "Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data," *Journal of Official Statistics*, 14.

Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998), "Disclosure Limitation Using Perturbation and Related Methods for Categorical Data," *Journal of Official Statistics*, 14, 485-502.

Franconi, L., and Polettini, S. (2004), "Individual Risk Estimation in M-Argus: A Review," in *Privacy in Statistical Databases*, eds. J. Domingo-Ferrer and V. Torra, Berlin: Springer, pp. 262-272.

Fuller, W. A. (1987), *Measurement Error Models*, New York: John Wiley & Sons.

Fuller, W. A. (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, 9, 383-406.

Gelman, A. (2007), "Struggles with Survey Weighting and Regression Modeling (with Discussion)," *Statistical Science*, 22, 153-164.

Groves, R. M., and Raghunathan, T. E. (2005), "Mixed Mode Methods in a World of Social Isolates, Pervasive Surveillance, and Ubiquitous Transaction Records: A Modest Proposal," *Conference on Mixed Mode Methods in Comparative Social Surveys*.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman&Hall/CRC.

Hawala, S., Stinson, M., and Abowd, J. (2005), "Disclosure Risk Assessment through Record Linkage," in *Joint UNECE/Eurostat work session on statistical data confidentiality*, Geneva, Switzerland.

He, X. Z., and Baker, D. W. (2004), "Changes in Weight among a Nationally Representative Cohort of Adults Aged 51 to 61, 1992 to 2000.," *American Journal of Preventive Medicine*, 27, 8-15.

Hurkens, C. A. J., and Tiourine, S. R. (1998), "Models and Methods for the Microdata Protection Problem," *Journal of Official Statistics*, 14, 437-447.

Institute for Social Research. (2008), "Papers and Publications: Health and Retirement Study."

Jeffreys, H. (1961), *Theory of Probability* (3 ed.), New York: Oxford University Press.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006), "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality," *The American Statistician*, 60, 224-232.

Kennickell, A. B. (1997), "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances," in *Record Linkage Techniques*, eds. W. Alvey and B. Jamerson, Washington D.C.: National Academy Press, pp. 248-267.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm," *Journal of Official Statistics*, 9, 313-331.

Le Cessie, S., and Houwelingen, J. C. V. (1992), "Ridge Estimators in Logistic Regression," *Applied Statistics*, 41, 191-201.

Little, R. J. A. (1993), "Statistical Analysis of Masked Data," *Journal of Official Statistics*, 9, 407-426.

Little, R. J. A., and Liu, F. (2002), "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata," in *Joint Statistical Meeting*, New York City, New York

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis with Missing Data* (2nd ed.), New York: Wiley-Interscience.

Lohr, S. L., and Prasad, N. G. N. (2003), "Small Area Estimation with Auxiliary Survey Data," *Canadian Journal of Statistics-Revue Canadienne De Statistique*, 31, 383-396.

Lyberg, L., et al. (eds.) (1997), *Survey Measurement and Process Quality*, New York: : Wiley-Interscience.

Martin, M. E., and Straf, M. L. (eds.) (1992), *Principles and Practices for a Federal Statistical Agency* (3 ed.), ed. C. o. N. Statistics, Washington, D.C. : NATIONAL ACADEMY PRESS.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2 ed.), London: Chapman and Hall.

Meng, X. (1994), "Multiple-Imputation Inferences with Uncongenial Sources of Input," *Statistical Science*, 9, 538-558.

Paass, G. (1988), "Disclosure Risk and Disclosure Avoidance for Microdata," *Journal of Business & Economic Statistics*, 6, 487-500.

Platek, R., Rao, J. N. K., Sarndal, C., and Singh, M. (eds.) (1987), *Small Area Statistics*, New york: John Wiley and Sons.

Raghunathan, T. E. (2008), "Diagnostic Tools for Assessing the Validity of Synthetic Data Inferences."

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27, 85-95.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003), "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T. E., and Van Hoewyk, J. (2008), "Disclosure Risk Assessment for Survey Microdata."

Rao, J. N. K. (2003), *Small Area Estimation*, New York: Wiley.

Reiss, S. P. (1984), "Practical Data-Swapping: The First Steps," *ACM Transactions on Database Systems*, 9, 20-37.

Reiter, J. P. (2004), "Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation," *Survey Methodology*, 30.

Reiter, J. P. (2005), "Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study," *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 168, 185-205.

Reiter, J. P. (2005), "Using Cart to Generate Partially Synthetic, Public Use Microdata," *Journal of Official Statistics*, 21, 441-462.

Reiter, J. P., and Drechsler, J. (2007), "Releasing Multiply-Imputed Synthetic Data Generated in Two Stages to Protect Confidentiality," *IABDiscussionPaper*, 20.

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

Rubin, D. B. (1981), "The Bayesian Bootstrap," *The Annals of Statistics*, 9, 130-134.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley & Sons.

Rubin, D. B. (1993), "Discussion: Statistical Disclosure," *Journal of Official Statistics*, 9, 461-468.

Schaefer, R., Roi, L., and Wolfe, R. (1984), "A Ridge Logistic Estimator," *Communications in Statistics – Theory and Methods* 13, 99-113.

Schaefer, R. L. (1986), "Alternative Estimators in Logistic Regression When the Data Are Collinear," *Journal of Statistical Computation and Simulation*, 25, 75 - 91.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data: Algorithms and Examples*, London: Chapman & Hall.

Siegel, M. J., Bradley, E. H., Gallo, W. T., and Kasl, S. V. (2004), "The Effect of Spousal Mental and Physical Health on Husbands' and Wives' Depressive Symptoms, among

Older Adults: Longitudinal Evidence from the Health and Retirement Survey," *Journal of Aging & Health*, 16, 398-425.

Singer, E. (2003), "Exploring the Meaning of Consent: Participation in Research and Beliefs About Risks and Benefits," *Journal of Official Statistics*, 19, 273-285.

Skinner, C. J. (2007), "The Probability of Identification: Applying Ideas from Forensic Statistics to Disclosure Risk Assessment," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 195-212.

Skinner, C. J., and Carter, R. G. (2003), "Estimation of a Measure of Disclosure Risk for Survey Microdata under Unequal Probability Sampling," *Survey Methodology*, 29, 197-201.

Skinner, C. J., and Elliot, M. J. (2002), "A Measure of Disclosure Risk for Microdata," *Journal Of The Royal Statistical Society Series B*, 64, 855-867.

Skinner, C. J., and Holmes, D. J. (1998), "Estimating the Re-Identification Risk Per Record in Microdata," *Journal of Official Statistics*, 14, 361-372.

Spitzer, D. L. (2005), "Engendering Health Disparities," *Canadian Journal of Public Health-Revue Canadienne De Sante Publique*, 96, S78-S96.

Subcommittee on Disclosure Limitation Methodology, F. C. o. S. M. (revised 2005), "Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology."

The Census Bureau. (2006), "Design and Methodology, the American Community Survey," Technical.

The Energy Information Administration of the U.S. Department of Energy. (2004), "2001 Public Use Data Files ".

Tranmer, M., et al. (2005), "The Case for Small Area Microdata," *Journal of the Royal Statistical Society Series A*, 168, 29-49.

Trevor Hastie, R. T. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

Trottini, M. (2003), "Decision Models for Data Disclosure Limitation. Unpublished Doctoral Dissertation."

U.S. Census Bureau. (2007), "Public Use Microdata Sample File (Pums): 2006 Pums Top Coded and Bottom Coded Values," 2008.

Willenborg, L., and De Waal, T. (2000), *Elements of Statistical Disclosure Control* (Vol. 155), New York: Springer.

Winkler, W. E. (1997), "Matching and Record Linkage," in *Federal Committee on Statistical Methodology*, Arlington, VA.

Winkler, W. E. (2004), "Masking and Re-Identification Methods for Public-Use Microdata: Overview and Research Problems," Technical Report Statistics #2004-06, U.S. Bureau of the Census.

Xie, D., Raghunathan, T., and Lepkowski, J. M. (2007), "Estimation of the Proportion of Overweight Individuals in Small Areas - a Robust Extension of the Fay-Herriot Model," *Statistics in Medicine*, 26, 2699-2715.

Zaslavsky, A. M., and Horton, N. J. (1998), "Balancing Disclosure Risk against the Loss of Nonpublication," *Journal of Official Statistics*, 14, 411-419

Zayatz, L. (2005), "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update," Technical Report Statistics #2005-2006, Statistical Research Division, U.S. Census Bureau.