# USING MULTIPLY IMPUTED, SYNTHETIC DATA TO FACILITATE DATA SHARING

by

Christine Noelle Kohnen

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Jerome P. Reiter, Supervisor

_____
Dr. David L. Banks

_____
Dr. James O. Berger

_____
Dr. Alan F. Karr

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

2005

# ABSTRACT

(Statistics)

# USING MULTIPLY IMPUTED, SYNTHETIC DATA TO FACILITATE DATA SHARING

by

Christine Noelle Kohnen

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Jerome P. Reiter, Supervisor

_____
Dr. David L. Banks

_____
Dr. James O. Berger

_____
Dr. Alan F. Karr

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

2005

# Abstract

The collection of data by statistical agencies and other statistical organizations for internal use and public release is a complex process. Researchers and policy makers demand high quality public-use data, while agency concerns regarding confidentiality and respondent protection limit the information that can be released. Even the sharing of data between statistical agencies cannot be done without first protecting the data in question. Advances in computer technology pose a threat to data confidentiality because data intruders are equipped with tools and resources that can be used to link public records with released data. Therefore, to limit disclosures, agencies apply disclosure control techniques to their data prior to release to ensure that respondent information is protected. However, the application of such techniques reduces the utility of the released data. The requirements of agencies to safeguard their data from disclosures limit their ability to share and exchange unperturbed data with one another. Even in situations where agencies desire to work in an honest environment and the exchange of data would benefit agencies and the researchers who study public-use data, data sharing is limited.

One approach agencies can use to safely share their data and create public-use data in the process, is to exchange synthetic data rather than real data. If the agencies have mutual interests, then it may be advantageous for them to create a combined data set that is accessible to all contributing agencies. This combined data set would give agencies and public-use data users the ability to incorporate additional records or attributes into their analyses than previously available from the individual data sources. To facilitate the sharing of confidential data between agencies, synthetic data methods are used to create multiply imputed, synthetic data sets that can be shared among participating agencies. Inferential methods for combining data sets

from multiple sources are derived and then validated based on simulation studies that utilize several different analysis models. Implementation of the proposed data sharing methods on real data requires creativity and an inherent understanding of the data to maintain both the overall structure of the data and the underlying relationships.

# Acknowledgements

There are several individuals who have inspired, guided, supported, and helped me through my time in graduate school and the completion of this dissertation, and I want to publicly acknowledge them.

First, I must thank my advisor Dr. Jerome Reiter, for his endless enthusiasm, direction, and patience. Regardless how busy, Jerry always had time to answer my questions, and spend the necessary time to help me through the rough spots of this dissertation. In addition, I would like to thank my other committee members for their help during this process.

I would like to thank Dr. Dalene Stangl for all of the help she gave me during my time at Duke. Specifically, for being there at times I needed someone to talk about graduate school and future plans.

I cannot forget to mention the two who made my first year bearable and introduced me to Women's Basketball: Dr. Jenise Swall for simply being herself — your enthusiasm was contagious; and Dr. Herbie Lee, who not only explained the *sub-50* concept that helped me understand North Carolina weathermen, but also endured my chipper morning attitude in 290.

Thanks to those at Duke who had nothing to do with Statistics, but became a part of my weekly routine, and gave me that much needed outlet — Missy, Oksana, and all the partners (James and Brad, you were the best!) I have had while dancing everything from Waltz and Quickstep, to Salsa and Cha-Cha.

There have been many ups and downs during this process — days where I wanted to give up and do something totally different. However, it was the love and support of my family that saw me through and simply believed in me. Mom, Dad and the rest of my family, you knew this was a dream of mine since I was a child and through

determination and hard work it would be realized.

The completion of this dissertation could not have happened without the endless support of Eric van Gyzen — words cannot fully express your importance in my life.

Finally, for those who truly know me, then I am sure you will understand my reason and underlying motivation for completing this dissertation.

This research was supported by the *Digital Government II Project — Data Confidentiality, Data Quality and Data Integration for Federal Databases: Foundations to Software Prototypes* (NSF Award Number: EIA-0131884).

# Contents

**Abstract** iv

**Acknowledgements** vi

**List of Tables** xii

**List of Figures** xiv

**1 Introduction** 1

1.1 Statistical Disclosure Control . . . . . 3

1.1.1 Disclosure Control Methods for Microdata . . . . . 4

1.1.2 Disclosure Risks for Microdata . . . . . 6

1.2 Synthetic Data Methods . . . . . 9

1.2.1 Fully Synthetic Data . . . . . 10

1.2.2 Partially Synthetic Data . . . . . 11

1.2.3 Limitations of Synthetic Data Methods . . . . . 12

1.2.4 Benefits of Synthetic Data Methods . . . . . 13

1.2.5 Other Synthetic Data Methods . . . . . 13

1.3 Secure Data Sharing Methods . . . . . 14

1.3.1 Secure Summation . . . . . 14

1.3.2 Secure Data Integration . . . . . 15

1.3.3 Secure Matrix Products . . . . . 15

1.3.4 Limitations of Secure Data Sharing Methods . . . . . 16

1.4 Outline of Dissertation . . . . . 16

**2 Horizontally Partitioned Data** **18**

2.1 Data Sharing Notation and Description . . . . . 19

2.2 Bayesian Derivation of the Combining Rules . . . . . 23

2.2.1 Evaluating $f(Q|\mathcal{D}, P^M, d^M)$ . . . . . 24

2.2.2 Evaluating $f(\mathcal{D}|P^M, d^M)$ . . . . . 24

2.2.3 Evaluating $f(P^M|d^M)$ . . . . . 25

2.2.4 Evaluating $f(Q|d^M)$ . . . . . 27

2.3 Extension to Partially Synthetic Data . . . . . 28

2.4 Simulation Studies . . . . . 29

2.4.1 Simulation I: Linear Regression . . . . . 29

2.4.2 Simulation II: Logistic Regression . . . . . 36

**3 Vertically Partitioned Data** **40**

3.1 Data Sharing Notation and Description . . . . . 42

3.2 Bayesian Derivations of the Combining Rules . . . . . 47

3.2.1 Evaluating $f(Q|c^k, B, P^M, d^M)$ . . . . . 47

3.2.2 Evaluating $f(c^k, B|P^M, d^M)$ . . . . . 48

3.2.3 Evaluating $f(P^M|\, d^M)$ . . . . . 50

3.2.4 Evaluating $f(Q|d^M)$ . . . . . 51

3.3 Simulation Studies . . . . . 52

3.3.1 Simulation I: Linear Regression . . . . . 53

3.3.2 Simulation II: Logistic Regression . . . . . 57

3.3.3 Simulation III: Disguising $X$ . . . . . 59

3.3.4 Simulation IV: A Sharing Alternative . . . . . 66

**4 Empirical Investigations of the Data Sharing Approaches 69**

4.1 Generating Disguisers for Real Data . . . . . 71

4.1.1 Disguising `SQFT` . . . . . 75

4.1.2 Disguising `NWKER` . . . . . 77

4.1.3 Disguising `COOLP` . . . . . 79

4.1.4 Disguising `HEATP` and `LTOHRP` . . . . . 82

4.1.5 Evaluation of Disguisers . . . . . 82

4.1.6 Suggestions for Generating Disguisers . . . . . 89

4.2 Investigation of Tree Models for Generating Fully Synthetic Data . . 91

4.2.1 Building the Synthetic Data Model . . . . . 91

4.2.2 Evaluation of the Shared Synthetic Data . . . . . 107

**5 Conclusions and Future Research 113**

**Appendix A 118**

A.1 Derivation of the Approximate Degrees of Freedom for Horizontally Partitioned Data . . . . . 118

A.2 Derivation of the Weighted Estimator (2.5) and Associated Variance (2.6) . . . . . 120

A.3 Extension to Partially Synthetic Data . . . . . 121

A.3.1 Data Sharing Notation and Description . . . . . 121

A.3.2 Bayesian Derivation of the Combining Rules . . . . . 123

A.3.3 Derivation of the Approximate Degrees of Freedom . . . . . 126

**Appendix B 129**

B.1 Derivation of the Approximate Degrees of Freedom for Vertically Partitioned Data . . . . . 129

B.1.1 Asymptotic Comparison of (B.6) to the Partially Synthetic Degrees of Freedom . . . . . 131

**Appendix C** **133**

C.1 Generating Disguisers for Real Data . . . . . 133

C.2 Investigation of Tree Models for Generating Fully Synthetic Data . . 133

C.2.1 Agency 2 . . . . . 134

C.2.2 Agency 3 . . . . . 136

**Bibliography** **139**

**Biography** **143**

# List of Tables


2.1 1,000 simulations combining data from three agencies when $Q$ is the population mean of $Y$. . . . . 32

2.2 1,000 simulations combining data from three agencies when $Q$ is the regression coefficient of $Y$ on $X$. . . . . 35

2.3 1,000 simulations combining data from three agencies when $Q$ is the logistic regression coefficient ($\beta_1$) of $Y$ on $X$. . . . . 38

2.4 1,000 simulations combining data from three agencies when $Q$ is the logistic regression coefficient ($\beta_2$) of $Y$ on $X$. . . . . 39

3.1 1,000 simulations combining data from two agencies when $Q$ is the regression coefficient of $Y$ on $X$. . . . . 55

3.2 1,000 simulations combining data from two agencies when $Q$ is the logistic regression coefficient ($\beta_1$) of $Y$ on $X$. . . . . 59

3.3 1,000 simulations combining data from two agencies when $Q$ is the logistic regression coefficient ($\beta_2$) of $Y$ on $X$. . . . . 60

3.4 Linear regression comparison of the disguising methods for $X$ with $Y$, using a Mahalanobis distance based clustering method. . . . . 63

3.5 Linear regression comparison of the disguising methods for $X$ with $Y$, using directional random noise. . . . . 65

3.6 1,000 simulations combining data from two agencies when $Q$ is the regression coefficient of $Y$ on $X$. . . . . 67

4.1 The 13 variables selected from the 1995 Commercial Building Energy Consumption Survey (CBECS) public-use data and used for analysis. 71

4.2 Principal Building Activity (`PBA`) categories. . . . . 74

4.3 Year Constructed (`YRCON`) categories. . . . . 75

4.4 Linear Regression Model Comparison of $\mathbf{X}$ and $\mathbf{X}_{dis}$ for $\texttt{MFBTU}^{1/2}$. . . 83

4.5 Confidence Interval Overlap based on the Linear Regression Model Comparisons in Table 4.4. . . . . 85

4.6 Logistic Regression Model Comparison of $\mathbf{X}$ and $\mathbf{X}_{dis}$ for $\texttt{NGBTU}^{01}$. . . 87

4.7 Confidence Interval Overlap based on the Logistic Regression Model Comparisons in Table 4.6. . . . . 88

4.8 Variable Ordering for the Synthetic Data Generation Models. . . . . 92

4.9 Joint Distribution of PBA and YRCON for the data designated to agency 1. 95

4.10 Comparison of the Summary Statistics (estimates and associated variances) for the True and Synthetic Data. . . . . 108

4.11 Comparison of the Linear Regression Models for the True and Synthetic Data. . . . . 110

4.12 Comparison of the Logistic Regression Models for the True and Synthetic Data. . . . . 112

C.1 Joint Distribution of PBA and YRCON for the 1995 CBECS public-use data. . . . . 134

C.2 Joint Distribution of PBA and YRCON for the data designated to agency 2.136

C.3 Joint Distribution of PBA and YRCON for the data designated to agency 3.138

# List of Figures


3.1 Vertically Partitioned Synthetic Data Sharing Diagram. . . . . 45

4.1 Histogram of Major Fuel Usage (MFBTU). . . . . 73

4.2 Histogram of Square Root of Major Fuel Usage ($\texttt{MFBTU}^{1/2}$). . . . . 73

4.3 Histogram of Total Floor Space (SQFT). . . . . 76

4.4 Histogram of Square Root Number of Employees ($\texttt{NWKER}^{1/2}$). . . . . 78

4.5 Histogram of Percentage of Floor Space Cooled (COOLP). . . . . 79

4.6 Histogram of Percentage of Floor Space Heated (HEATP). . . . . 81

4.7 Histogram of Percentage of Floor Space Lit (LTOHRP). . . . . 81

4.8 Histogram of Percentage of Floor Space Cooled (COOLP) for Vacant Buildings (PBA = 1). . . . . 84

4.9 Comparison of the Confidence Intervals and Coverage for log(SQFT) and COOLP. . . . . 86

4.10 Comparison of the Confidence Intervals and Coverage for log(SQFT) and HEATP. . . . . 88

4.11 Graphical representation of the relationships in the data designated to agency 1 and used to generate the synthetic data. . . . . 93

4.12 Tree Diagram for Square Root of Number of Employees ($\texttt{NWKER}^{1/2}$) . . 98

4.13 Histogram of the Square Root of Major Fuel Expenditure ($\texttt{MFEXP}^{1/2}$) 101

4.14 Histogram of the Square Root of Electricity Consumption ($\texttt{ELBTU}^{1/2}$) 102

4.15 Histogram of the Square Root of Electricity Expenditure ($\texttt{ELEXP}^{1/2}$) . 103

4.16 Histogram of the Square Root of Natural Gas Consumption ($\texttt{NGBTU}^{1/2}$) 104

4.17 Percentage of ELBTU and NGBTU as compared to MFBTU. . . . . 105

4.18 Histogram of the Square Root of Natural Gas Expenditure ($\texttt{NGEXP}^{1/2}$) 106

C.1 Graphical representation of the relationships in the data designated to agency 2 and used to generate the synthetic data. . . . . 135

C.2 Graphical representation of the relationships in the data designated to agency 3 and used to generate the synthetic data. . . . . 137

# Chapter 1

# Introduction

The collection of data by statistical agencies and other statistical organizations for internal use and public release is a complex process. Researchers and policy makers demand high quality public-use data, while concerns of confidentiality and respondent protection limits the information that can be released. Even the sharing of data between statistical agencies cannot be done without first protecting the data in question. There may be situations where agencies with mutual interests may want to share their respective data sets and create a combined data set accessible to each contributor. For instance, suppose several state or local education agencies want to combine similar data on students within their regions or districts. The expanded coverage of a combined data set may improve the precision of their analyses, as compared to the analyses on their individual data sets (Karr *et al.*, 2004; Sanil *et al.*, 2004b). Alternatively, suppose two statistical agencies individually own income data and education data and want to predict how income is affected by education. In either situation, agencies could benefit from a shared, combined data set created from multiple sources. Benefits of data sharing include the opportunity to incorporate additional records or attributes into analyses, subsequently increasing the precision of inferences or promoting the investigation of new relationships. Furthermore, the

shared data set can be released as public-use data, passing on the benefits to users.

In an ideal world, agencies could share and combine data — sensitive or not — with other agencies. However, due to the risk of disclosures, confidentiality agreements, and the potential misuse of data, agencies may need to safeguard their data and protect the identities or values of respondents. The necessity of statistical disclosure control methods for data protection (Willenborg and de Waal, 1996, 2001) limit the ability of agencies to share unperturbed data. The possibility of sharing and combining data, would require the cooperation of involved agencies to disseminate possibly sensitive information — a request they might not grant. By sharing sensitive respondent information, agencies could reveal actual data values or even identities, thereby violating confidentiality agreements with respondents. The existence of confidentiality agreements, and hence the concern of agencies for disclosure limitation, limits the sharing of data among statistical agencies (Fienberg, 1994).

To reduce disclosure risks and create a shareable data set from multiple sources that can also be released for public use, I propose the use of synthetic data methods. Currently, there are several different variations of synthetic data methods (described in Section 1.2) that agencies can individually use to create public-use data that protects confidentiality. However, there may be situations where multiple agencies own similar data sets and want to create public-use data in a collaborative effort. The implementation of which requires that the current methods be extended to incorporate multiple agencies or entirely new methods be constructed such that agencies have the ability to share data and create public-use data without exposure to additional risks.

Before multiple agencies can begin a collaborative project, they must first determine the overall structure of their respective data sets in order to decide whether or not they should continue with their cooperative efforts and which method best

suites their data and ultimate goals. For instance, agencies may own *horizontally partitioned data* — data collected on the same set of attributes for different sets of respondents — or they may have *vertically partitioned data* — different attributes recorded for the same set of respondents. Regardless of their data type, agencies must decide the circumstances under which the output from their collaborative efforts will be used, to help choose a candidate method that does not expose them to unnecessary disclosure risks. In addition, agencies will need to decide the extent to which information will be shared among the agencies; this may also help determine the method used.

Prior to describing the proposed data sharing methods in Chapters 2 and 3, I will give background information regarding data protection and describe several of the methods commonly used to create public-use data. Synthetic data methods will be introduced as a means of building a foundation from which the data sharing methods have been developed. Finally, I will summarize some of the current data sharing methods available for both horizontally and vertically partitioned data.

## 1.1 Statistical Disclosure Control

Before agencies can release public-use data sets, the data may need to be protected against disclosures of both the actual respondents (identity) and their reported values (attributes). Depending on the form of the data to be released — microdata or tabular data — different disclosure control measures are necessary to ensure sensitive data values remain confidential. The main difference between the two data types is their overall structure; however, the two are related. Microdata are composed of the individual records collected from respondents, usually through the use of a survey (Willenborg and de Waal, 1996, 2001). Each record contains a respondent's answer or value from each of the variables (questions) asked during the survey. In contrast,

tabular data are composed of aggregate information, represented in the form of a table, and often contain frequency count data (Willenborg and de Waal, 1996, 2001). Therefore, the type of disclosure risks associated with each are different, along with the measures used to protect the released data. (Hereafter, any mention of public-use data or data in general will be in reference to microdata, with all presented theory and applications being applicable to microdata only.)

### 1.1.1 Disclosure Control Methods for Microdata

Techniques that restrict the information released as a way to protect sensitive units and values in a data set are *statistical disclosure control* (SDC) methods. Several of the commonly used SDC methods will be described here; for thorough descriptions and illustrations, see Willenborg and de Waal (1996, 2001). The purpose of SDC methods is to protect released data sets from malicious users (often referred to as *data intruders*, *snoopers* or *attackers*) while providing adequate statistical information to *analysts* (who use public-use data for statistical purposes only) (Duncan and Sumitra, 2000). Safeguarding microdata from data intruders, who will use any means available to infer confidential information about individuals in public-use data, is a concern of data disseminators. Data intruders probe public-use data, searching for ways to find and exploit weaknesses, whether trying to link public records with released data or identifying rare combinations of characteristics. Agencies must therefore make a balanced choice to protect respondents and their data, while maintaining the quality of released public-use data to provide analysts with good statistical inferences. (See Fienberg (1994) for more regarding the conflict between confidentiality protection and data access.)

The underlying reason agencies use SDC methods on their data prior to release is to avoid or reduce the risk of disclosures, whether it be *identity disclosures* or *attribute*

*disclosures* (see Section 1.1.2). Of the two, identity disclosures are the more serious, because an individual or establishment has been linked or associated to a specific record in the released data exposing their information. The other type of disclosure occurs when the value of a sensitive variable (attribute) has been determined or closely estimated (a *predictive disclosure*, see Willenborg and de Waal (2001, pp. 42–45) for more detail).

Several commonly used SDC methods are *recoding*, e.g. *global recoding* and *top-coding*; *local suppression*; and *substitution*, e.g. *data swapping* and *jittering* (Willenborg and de Waal, 1996, 2001). The application of SDC methods on data prior to release for public use results in the loss of some information or utility of the data. In several of methods considered, the resulting data are presented at a coarser level than the original data and may require sophisticated analysis methods.

Consider first the recoding techniques, which create new categories containing less detailed information than available in the original data. For instance, global recoding collapses the groupings of a categorical variable, requiring that the variable's definition be redefined in terms of the new, broader categorization. To prevent identity disclosures due to associated extreme values, top-coding is used to recode all values of an attribute whenever they surpass a specified threshold. This process can also be implemented on very small values, in which case it is called bottom-coding. When attribute values are too sensitive for recoding techniques, local suppression is used to remove the values from the data set, replacing them with missing values.

Another broad category of SDC methods is substitution, which encompasses several methods that replace sensitive values with other non-missing values, without necessarily changing the definition of the variable. For instance, the addition of random noise to continuous attributes is called jittering. Another technique is data swapping, where different values of the same attribute are exchanged, therefore not

changing the overall summaries and description of the attribute. Data swapping preserves univariate marginal distributions; however, joint distributions of swapped and unswapped variables will be distorted, with the distortion increasing with swap rates (Gomatam and Karr, 2003). A more extreme SDC method would be to simply not release any real data, rather release completely synthetic data that replicate the properties and relationships of the original data. Several variations of this will be discussed in Section 1.2.

Data sets released for public use should meet the demands for statistical information by researchers — whether government, industry or university-based — providing a good impression or understanding of the population in question. However, because the application of SDC methods can potentially distort relationships in released data, the entire purpose or use of the public-use data may be invalidated. In addition, SDC methods can complicate analyses and if not implemented properly, the resulting inferences may be biased. The agencies performing SDC methods often have information and resources available that users do not and may be creating public-use data that can only be successfully analyzed with specific skills and tools. In addition, not all information regarding the SDC method used is released in conjunction with the released data sets, e.g. swap rates. Therefore, to accurately analyze data after the implementation of SDC methods, users should apply the likelihood-based methods of Little (1993) or the measurement error models in Fuller (1993). In addition, knowledge of the masking technique, method-specific software and new statistical skills may be necessary to obtain valid inferences from the released data.

### 1.1.2 Disclosure Risks for Microdata

Agencies that disseminate public-use data must be cautious regarding the level of information released due to the possibilities of disclosures, because confidentiality

agreements could be breached. The consequences from a publicized disclosure of confidential data would affect the reputation and future ability of agencies to gather high quality statistical data because data collection depends on the goodwill and cooperation of respondents, both individuals and establishments (Fienberg, 1994; Duncan *et al.*, 2004). To achieve zero disclosure risk, agencies would have to refrain from releasing any information, which runs contrary to the dissemination of data for public use and society betterment. Agencies cannot choose to release data to some users and not to others (Zayatz *et al.*, 1999) as a means of controlling disclosure risks. Thus, agencies must try to quantify the amount or level of external information available to intruders for use in conjunction with public-use data in order to choose an appropriate SDC method for disclosure prevention. This task is difficult because the quantity of external information available continues to increase with readily available computing resources and linking techniques, which are continually becoming more sophisticated with advances in technology. In addition, assessing the actual behavior of intruders poses a problem; therefore, agencies generally try to address the risks associated with certain types of attacks.

First, agencies must decide on a method for quantifying the disclosure risks associated with data made for public release. This quantification method does not necessarily have to be a probability model, but can be some type of thresholding model. If an agency uses a thresholding model, then the maximum acceptable risk must be decided and if the risk for a data set falls below, it can be released (Fienberg and Willenborg, 1998).

To access risk, agencies consider different disclosure scenarios, specifically how intruders might try to identify individuals in released data, assuming different levels of knowledge regarding the population in question. Suppose a data intruder wants to exploit vulnerabilities in a released public-use data set to identify specific individu-

als. Under these circumstances their attack would be considered intentional because their goal is to exploit inherent weaknesses, even if the intruder is a legitimate user of the released data set (Keller-McNulty and Unger, 1993). In this case, respondents who have rare combinations of identifying variables have a high risk of disclosure, such as the *sample uniques* — only individuals in a data set with a particular combination of responses. In some situations, a sample unique may also be a *population unique*, but usually there are unsampled individuals in the population with the same characteristics.[1]

Variables or attributes that alone or in combination with others aid in the re-identification of respondents are considered *identifying variables* and a combination of such variables contributes a *key*. Agencies and other similar organizations strip released data sets of all variables that can clearly identify individual respondents. The re-identification concern also increases as the number of variables released increases, because it is more likely for combinations of variables to occur infrequently, increasing risk of disclosures (Fienberg, 1994).

Suppose public-use data are used to closely estimate unknown sensitive values through some type of statistical analyses. If a sensitive value can be inferred, then a *predictive* or *inferential disclosure* has occurred and may result in an attribute disclosure (Keller-McNulty and Unger, 1993; Willenborg and de Waal, 2001, pp. 42–45). This indirect attack does, however, have some level of uncertainty attached to the estimated value. If the level of predictive disclosure is high for a particular record, it may be possible for an intruder to infer the identity of the record; in such case, the re-identification risk of the record would be high (Willenborg and de Waal, 2001) and may result in a breach of confidentiality. However, if the level of uncertainty for an estimated value is large or an intruder's external information is inaccurate then

[1]The risks associated with the unsampled individuals do not have to be considered because the units are outside the released data sample (Willenborg and de Waal, 2001, p. 52).

an *apparent disclosure* may occur (Keller-McNulty and Unger, 1993). In this case, an intruder may assume an incorrect value for a particular attribute, with the actual remaining unknown. Does this disclosure constitute a breach in confidentiality or not?

## 1.2 Synthetic Data Methods

At the extreme end of the disclosure limitation spectrum exist synthetic data methods. Instead of modifying values of a data set to limit disclosures, real data values are replaced with fake values, albeit real-looking ones. There are varying degrees to which synthetic data methods can be applied, of which I will address two — partially synthetic and fully synthetic data methods — along with several variants.

Synthetic data methods were first proposed by Rubin (1993) as a means of creating public-use data sets that honor the confidentiality of respondent values while maintaining simplicity in their analysis. Obtaining valid inferences from synthetic data sets requires knowledge of the relevant combining rules and variance equation for the desired estimand, based on the principles of multiple imputation (Rubin, 1987). The necessary formulas are simple and allow users to obtain valid inferences from synthetic data sets using standard statistical software and methods. Synthetic data methods — along with their predecessor, multiple imputation — put the computation burden on the agencies that generate and release data; end users need only to repeat their analyses on each synthetic data set and combine their results as indicated by the combining rules and variance equation. In addition, confidentiality is preserved because the released synthetic data sets cannot be linked back to actual respondents.

### 1.2.1 Fully Synthetic Data

Since the original proposal, synthetic data concepts have undergone development to include multiple methods and variations thereof. The fully synthetic data method of Raghunathan *et al.* (2003) implements the idea proposed by Rubin (1993), detailing the steps required to create completely synthetic public-use data sets. The resulting data sets are comprised of fake respondents and respondent values modeled after those in the original data and therefore cannot be linked back to actual respondents. This process removes the risks of disclosures inherent in other public-use data sets. Raghunathan *et al.* (2003) derive the necessary combining rules and variance equation needed to obtain valid and unbiased inferences from the fully synthetic data sets.

Suppose $Q$ is the population-based quantity of interest. Let $q_i$ be the estimate of $Q$ from the $m$ completely synthetic data sets and $u_i$, the associated variance of the estimator from the synthetic data sets, such that the combining rules can be written as

$$\bar{q}_m = \frac{1}{m}\sum_{i=1}^{m} q_i \tag{1.1}$$

$$\bar{u}_m = \frac{1}{m}\sum_{i=1}^{m} u_i \tag{1.2}$$

$$b_m = \frac{1}{m-1}\sum_{i=1}^{m}(q_i - \bar{q}_m)^2 \tag{1.3}$$

where $b_m$ is the variance of the $m$ estimates. Therefore, $Q$ can be estimated by $\bar{q}_m$, with associated variance

$$T_F = (1 + 1/m)b_m - \bar{u}_m \tag{1.4}$$

(see Raghunathan *et al.* (2003) for complete details). When $m$ and the size of the synthetic data sets are large, the normal approximation $(Q - \bar{q}_m) \sim N(0, T_F)$ is

appropriate for inferences. However, if these values are moderate, a t-distribution with $\nu_F = (m-1)(1-\bar{u}_m/((1+1/m)b_m))^2$ (Reiter, 2002) degrees of freedom must be used for inferences.

### 1.2.2 Partially Synthetic Data

Similar, but not as extreme in the replacement of real data values, is the partially synthetic method (Little, 1993; Kennickell, 1997; Abowd and Woodcock, 2001; Little and Liu, 2003; Reiter, 2003), which only replaces those values at high risk of disclosure with synthetic ones. The released data sets are then a combination of non-sensitive data values and synthetic replacements of the sensitive values. The method is directly comparable to multiple imputation for missing data (Rubin, 1987): instead of imputing values for missing data, values are imputed for those at high risk of disclosure. The combining rules and variance equation (Reiter, 2003), necessary for valid inferences, are different from those used in multiple imputation for missing data and the fully synthetic data method.

Suppose $Q$ is the population-based quantity of interest and $q_i$ the estimate of it from the $r = m$ partially synthetic data sets. The combining rules given in Section 1.2.1 (see (1.1), (1.2) and (1.3)) can be used to find valid inferences. These rules enable users to obtain an estimate of $Q$, with associated variance

$$T_P = b_r/r + \bar{u}_r. \tag{1.5}$$

Similar to the fully synthetic data method, whenever $r$ and the synthetic data set size are large, the normal approximation $(Q-\bar{q}_r) \sim N(0, T_P)$ is appropriate for inferences. Likewise, a t-distribution with $\nu_P = (r-1)(1+\bar{u}_r/(b_r/r))^2$ degrees of freedom should be used for inferences when the values are moderate.

### 1.2.3 Limitations of Synthetic Data Methods

The combining rules for synthetic data methods and multiple imputation are derived under the assumption of agreement between the imputation models and those used for inferences. In reality, these models may not be in agreement, because the *imputers* — agencies creating and releasing the synthetic data sets — and the *analysts* — users of public-use data — are often different entities with varying resources.

Schafer (1997, pp. 139–143) discusses several such scenarios where the two models are not in agreement due to the assumptions made by either the imputer or analyst, or in cases when the imputer uses additional knowledge while creating the synthetic data sets. For example, suppose the analyst uses a model that is a subset of that used by the imputer. If the assumed model is indeed true, then the resulting inferences will still be valid, although conservative. However, if false, the resulting inferences will be biased and the fault will be with the analyst, not the imputer and imputed data sets. To avoid inconsistencies between the imputation model and those used for inferences, the imputer should create a model that includes as much information possible, given the current computing conditions, so as not to impose unnecessary restrictions unknown by the analysts (see also Meng (1994) for detailed examples). Meng (1994) defines this situation when the imputation model and the analysis procedure are not in agreement as being *uncongenial*.

In some cases, imputers may have auxiliary information available that can be incorporated into the imputation models that cannot be released due to its sensitive nature. Such additions to the imputation models can potentially result in more efficient inferences (Raghunathan *et al.*, 2003). In general, the use of all available information pertaining to the data being imputed should be incorporated into the imputation model. Sensible imputation models will increase the predictive power of the released data and offer greater flexibility in the scope of user analyses (Meng,

1994). When feasible, and confidentiality permitting, agencies should include information regarding the imputation process with the actual public-release data.

### 1.2.4 Benefits of Synthetic Data Methods

A benefit of using synthetic data methods is the simplicity with which valid inferences can be found. Repeated calculations of the desired estimates, used in the combining rules and variance estimate, can easily be obtained with standard complete-data statistical methods and software. Even in situations where the original data were collected using a complex survey design, standard analyses apply due to the simple random sampling of the synthetic samples (in the fully synthetic data method context). For instance, if the data were stratified, unbiased inferences can be obtained with estimates that account for stratum differences and sizes (see Reiter (2002) for more details).

### 1.2.5 Other Synthetic Data Methods

#### Selective Multiple Imputation of Keys (SMIKe)

Selective multiple imputation of keys (SMIKe) by Little and Liu (2003) is a variant of the partially synthetic data method described in Section 1.2.2. Rather than imputing values for sensitive non-categorical values in the data set, values are imputed for the sensitive categorical values in the *key variables* — characteristics of the individual respondents within the data set, such as *age* and *gender*. It is assumed throughout that data intruders can determine values of key variables through the use of publicly available data sources. A drawback of SMIKe is the inability to preserve the original counts of the combinations of key variables, e.g. the cells of a contingency table made with the key variables. Statistical inferences for SMIKed data parallel those for partially synthetic data and use the same combining rules and variance equation

as given in Section 1.2.2.

#### Multiple and Stochastic Swapping of Keys (MASSK)

Based on the same principles as SMIKe, multiple and stochastic swapping of keys (MASSK) by Liu (2003) is similar to the data swapping used for SDC, in the sense that values within a variable are exchanged with one another. However, only key information is swapped between paired cases selected probabilistically. A benefit of swapping between paired cases is that values of sensitive keys are only exchanged with others within the swapping variable and the original key cell counts are preserved.

## 1.3 Secure Data Sharing Methods

Several methods already exist that allow multiple agencies to securely find and share analyses based on disjoint data sets and, in some cases, integrate their individual data sets into a single, shareable data set. Methods for both horizontally and vertically partitioned data sets will be discussed because they are the direct competitors of the data sharing methods presented in this dissertation. Specifically, two methods — secure summation and secure data integration — will be described in the context of horizontally partitioned data (Karr *et al.*, 2004) and secure matrix products for vertically partitioned data (Sanil *et al.*, 2004a).

### 1.3.1 Secure Summation

In Karr *et al.* (2004), the secure summation protocol (Benaloh, 1987) is used to calculate estimates of the sufficient statistics used for constructing linear regression analyses. These quantities are calculated locally by the participating agencies, then combined through the secure summation protocol. The protocol works by only sharing the computed data summaries and not the individual values. For instance, suppose

three or more agencies are cooperating in a semi-honest framework and want to compute the quantity $\kappa = \sum_{j=1}^{J} \kappa_j$, where the $J$ agencies own a value $\kappa_j$, $j = 1, 2, \ldots, J$. The first agency selects a number $m$ from $[0, M)$, where $M$ is very large. The agency then adds $\kappa_1$ to $m$ and passes the sum to the second agency, who adds $\kappa_2$ before passing it on to the next agency. This process continues until the sum is passed back to the first agency, who subtracts $m$ recovering the joint sum $\kappa$. To complete the protocol, the first agency passes the sum $\kappa$ to the participating agencies.

### 1.3.2 Secure Data Integration

Karr *et al.* (2004) also propose a secure data integration method that allows cooperating agencies to integrate their separate data sets into a single, shareable data set. The procedure provides a low level of protection because the identities of data contributors are protected and remain unknown, but the individual data records are left unaltered and may be vulnerable to disclosures. The integrated data set gives contributors the ability to identify their own data records within the shared data set, implying that all remaining records came from the other contributing agencies. However, this knowledge does not allow any one agency to determine the particular source of the records and what portion of the shared data set was contributed by a specific agency.

### 1.3.3 Secure Matrix Products

Sanil *et al.* (2004a) propose a method that allows multiple agencies owning vertically partitioned data the ability to securely find the matrix products of their data sets without actually sharing any data. The method is a linear algebra technique based upon a series of matrix manipulations. Suppose two agencies, $\mathcal{A}$ and $\mathcal{B}$, own data sets $X_{n\times p}$ and $Y_{n\times q}$, respectively (these matrices are assumed to be full rank). To securely

calculate $X'Y$, $\mathcal{A}$ first generates a set of orthonormal vectors $Z_i$, $i = 1, 2, \ldots, g$ such that $Z_i'X_j = 0$ for all $i = 1, 2, \ldots, g$ and $j = 1, 2, \ldots, p$ (see the appendix of Sanil *et al.* (2004a) for details regarding the construction of $Z$). $\mathcal{A}$ then passes $Z$ to $\mathcal{B}$, this matrix does not give $\mathcal{B}$ any information regarding $X$. $\mathcal{B}$ then computes $W = (I - ZZ')Y$, where $I$ is the identity, and sends $W$ to $\mathcal{A}$. $\mathcal{A}$ then calculates $X'W = X'(I - ZZ')Y = X'Y$, yielding the desired product $X'Y$. Once determined, $\mathcal{A}$ sends $X'Y$ to $\mathcal{B}$.

### 1.3.4 Limitations of Secure Data Sharing Methods

The three data sharing methods described above have limitations regarding the types of applicable analyses, but more importantly, lack the ability to create public-use data. When agencies use secure summations to find sufficient statistics for use in regression models, they limit themselves to using linear regression models only. The same is true for agencies who use secure matrix products: only analyses that use covariance matrices can be implemented, e.g. linear regression models. Securely integrated data do not have limitations on the class of applicable models. However, the difference between agencies cannot be studied, because these distinctions are lost during data integration. In addition, the end results of these methods are not in a form that can be released as public-use data.

## 1.4 Outline of Dissertation

This dissertation will consider methods of data sharing among multiple agencies. The methods are classified and described according to the type of data owned by the agencies, whether it be horizontally or vertically partitioned data. The sharing of horizontally partitioned data will be addressed in Chapter 2, with sharing of vertically partitioned data in Chapter 3. In Chapters 2 and 3, methods are presented

for acquiring valid inferences from shared data sets, which require combining rules and variance equations different than those in Section 1.2. The new rules are necessary because they incorporate the additional variability due to multiple agencies generating, sharing and estimating inferences from synthetic data sets. These chapters also contain simulation studies for each of the data sharing methods, considering both linear and logistic regression scenarios. Implementation of components from the data sharing methods on real data are in Chapter 4, with concluding remarks and directions for future research in Chapter 5.

# Chapter 2

# Horizontally Partitioned Data

Consider once again the several state or local education agencies from Chapter 1, who individually own similar data on the students within their region. Given their mutual interest in assessing some aspect of the student population represented by the data, the agencies could potentially improve their analyses' precision by expanding the coverage of the student population, as compared to their individual results. Increased coverage of the population can be achieved if the agencies share their data to create a combined data set, accessible to each of the contributors. In an ideal world without data intruders, agencies could do as suggested and share their data — even sensitive data — without the fear of disclosures or the potential misuse of their students' data.

In this example, agencies own horizontally partitioned data, their observed data sets consist of the same set of attributes for disjoint sets of students. Since the expanded knowledge of the target student population would be mutually beneficial to all cooperating agencies, the use of a data sharing procedure to create a sharable data set is suggested. If agencies only need to protect their identities while integrating their individual data sets, then the protocol of Karr *et al.* (2004), as described in Chapter 1, is ideal. However, if agencies desire the ability to both protect the sampled students' data and create public-use data, then other methods must be explored. I

propose the use of a data sharing method that uses synthetic data principles as the basis for creating data sets that can be shared and disseminated for public use, while protecting the underlying student respondents.

This data sharing method allows agencies to perform all of the synthetic data generation tasks at their own locations, requiring cooperation with the other agencies only during two steps of the procedure. The first of which occurs during the initialization process, when agencies agree upon the attributes in their possession and the protocol to be followed. If agencies agree to cooperate in a semi-honest environment, then no agency will try to exploit the agreed-upon protocol or try to learn additional information about the other agencies' data. If necessary, these exchanges can be performed using some type of secure information exchange, possibly encryption (Schneier, 1996). The other step occurs at the end of the procedure when agencies exchange their synthetic data sets with one another for analysis at their own locations; these data sets can then be released for public use. The proposed data sharing method can be classified as a symmetric process because agencies should learn the same amount of information about the other agencies' data if they follow the agreed-upon protocol.

The data sharing methods described in this chapter extend the synthetic data methods from Sections 1.2.1 and 1.2.2 to incorporate multiple agencies into the procedures. In each case, if only one agency is present, the derived results will be equivalent to those already given in Chapter 1.

## 2.1 Data Sharing Notation and Description

Let $X$ be the $N \times d$ matrix of fully observed background variables known to all $j = 1, 2, \ldots, a \geq 2$ collaborating agencies. Agencies individually own $Y_{obs,j}$, a $n_j \times p$ matrix of fully observed survey variables, a subset of the $n \times p$ matrix $Y_{obs}$ ($n = \sum_{j=1}^{a} n_j$),

the observed portion of the $N \times p$ population matrix $Y$. Ultimately, agencies want to estimate the unknown scalar population-based quantity $Q = Q(X, Y)$, where $Q$ is a function of $X$ and $Y$. They believe it will be mutually beneficial to share their data to find a joint estimate of $Q$. Although agencies can individually estimate $Q$ using their observed data $D_j = (X, Y_{obs,j}, I_j)$, where $I_j$ is the agency-specific survey inclusion indicator, they desire a more precise estimate which can be found by covering a larger subset of $Y$. In addition, agencies want to create a combined data set that can be released for public use.

In most situations, agencies are bound by disclosure agreements that restrict them from sharing sensitive data with other agencies (Fienberg, 1994). However, it is possible for agencies to bypass these restrictions by sharing synthetic replicates of their data instead, because the replicates are not real and therefore not covered by the agreements. By sharing synthetic replicates, agencies can create a shared data set all contributors can use to estimate $Q$, without subjecting themselves to disclosure risks or violating disclosure agreements, and in the process create public-use data. (The creation of synthetic data will be utilized throughout this data sharing method.)

To begin generating the synthetic data sets, agencies first construct completed populations $P_j^{(i)} = (X, Y_{com,j}^{(i)})$ from their observed data, assuming no confidentiality restrictions exist on $X$ and $Y_{com,j}^{(i)} = (Y_{obs,j}, Y_{nobs,j}^{(i)})$, where $Y_{nobs,j}^{(i)}$ are the unobserved units of $Y_{obs}$ for agency $j$. The completed populations are formed by imputing values for the unobserved units[1] based on an agency-specific posterior prediction distribution $P(Y_{nobs,j}^{(i)}|X, Y_{obs,j})$, conditional on their observed data $D_j$ and any underlying design assumptions. Each agency repeats the population completion step $m_j$ times to account for variability introduced by imputing values for the unobserved units. Let $P_j = \{P_j^{(i)}, i = 1, 2, \ldots, m_j\}$ be the set of completed populations for agency $j$

[1]Values could also be imputed for $Y_{obs,j}$, which would eliminate the chance of sampling observed data from $Y_{com,j}^{(i)}$ during the selection of $d_j^{(i)}$.

and $P^M = \{P_j, j = 1, 2, \ldots, a\}$ the set of completed populations for all $a$ agencies. It will be assumed throughout that agencies act independently and have constructed posterior predictive distributions specific to their observed data $D_j$. In addition, let $\mathcal{D} = \{D_j, j = 1, 2, \ldots, a\}$ be the set of observed data for all $a$ agencies, assuming that each $D_j$ has been drawn from the same underlying population.

Next, agencies select random samples $d_j^{(i)} = (X_j^{(i)}, Y_{syn,j}^{(i)})$ from their respective sets of completed populations $P_j$, drawing one sample from each of their $m_j$ completed populations. Let $d_j = \{d_j^{(i)}, i = 1, \ldots, m_j\}$ be the set of synthetic samples agency $j$ selected from $P_j$, and $d^M = \{d_j, j = 1, \ldots, a\}$ the set of synthetic samples for all $a$ agencies. The complete set of synthetic samples $d^M$ can then be shared among agencies and released for public use, from which agencies and analysts can both obtain inferences.

To find valid inferences, agencies or analysts calculate an estimate of the scalar $Q$ and associated measure of uncertainty from each of the $d_j^{(i)}$, these estimates will be denoted by $q_j^{(i)}$ and $u_j^{(i)}$, respectively. These values are found assuming each $d_j^{(i)}$ is a simple random sample from the populations of $X$ and $Y$. The estimates, $q_j^{(i)}$ and $u_j^{(i)}$, are then used in the following combining rules to find valid inferences for scalar $Q$:

$$\bar{q}_M = \frac{1}{a}\sum_{j=1}^{a}\left(\frac{1}{m_j}\sum_{i=1}^{m_j} q_j^{(i)}\right) = \frac{1}{a}\sum_{j=1}^{a}\bar{q}_j \tag{2.1}$$

$$\bar{u}_M = \frac{1}{a}\sum_{j=1}^{a}\left(\frac{1}{m_j}\sum_{i=1}^{m_j} u_j^{(i)}\right) = \frac{1}{a}\sum_{j=1}^{a}\bar{u}_j \tag{2.2}$$

$$\bar{b}_M = \frac{1}{a}\sum_{j=1}^{a}\left(\frac{1}{m_j - 1}\sum_{i=1}^{m_j}(q_j^{(i)} - \bar{q}_j)^2\right) = \frac{1}{a}\sum_{j=1}^{a} b_j. \tag{2.3}$$

In (2.1), the $\bar{q}_j$ is the average estimate of $Q$ from agency $j$, with $\bar{q}_M$ being the averaged

estimate of $Q$ across all agencies. The $b_j$ in (2.3) is the variance of the estimates $q_j^{(i)}$ for agency j, with $\bar{b}_M$ being the average variance of estimates across all agencies. The $\bar{u}_j$ in (2.2) is the average variance estimated for agency $j$ and $\bar{u}_M$ is the average across all agencies. The value $\bar{q}_M$ then estimates $Q$ based on $d^M$, with associated variance

$$T_H = \frac{1}{a^2}\sum_{j=1}^{a}(1+1/m_j)b_j - \bar{u}_j. \tag{2.4}$$

Inferences for the posterior distribution of $Q|d^M$ can then be based on the normal approximation $(Q-\bar{q}_M) \sim N(0, T_H)$ when $n_j$ and $m_j$ are large; when moderate, a t-distribution with approximately $\nu_H = \left(\sum_{j=1}^{a}(m_j-1)\right)\left(1-r_H^{-1}\right)^2$, where $r_H = \frac{1}{\bar{u}_M}\left(\frac{1}{a}\sum_{j=1}^{a}(1+1/m_j)b_j\right)$, degrees of freedom is appropriate. (See Appendix A.1 for derivation of the approximate degrees of freedom.) The negative sign in (2.4) can potentially cause $T_H < 0$ for some values of $b_j$ and $\bar{u}_j$. Such occurrences can be reduced with a larger choice for $m_j$ and $n_{syn}$, and will be discussed further in Section 2.4.1. When $a = 1$, it is important to note that the combining rules and variance equation are equivalent to those in Raghunathan *et al.* (2003) (see Section 1.2.1), as this data sharing method simply extends their work to include multiple agencies.

The combining rules and variance equation already given assume the synthetic samples from the $a$ agencies contribute equally in the estimation of $Q$. To incorporate potential differences between agencies and their observed data sets, the estimates can be weighted with an agency-specific variability measure. For example, (2.1) becomes

$$\bar{q}_M^w = \frac{\sum_{j=1}^{a}\bar{q}_j/w_j}{\sum_{j=1}^{a}1/w_j} \tag{2.5}$$

when weighted by the agency variability measure $w_j = (1+1/m_j)b_j - \bar{u}_j$ (the variance from the fully synthetic data method (1.4)). Likewise, the associated variance

estimate for (2.5) becomes

$$T_H^w = \frac{\sum_{j=1}^a ((1+1/m_j)b_j - \bar{u}_j)/w_j^2}{(\sum_{j=1}^a 1/w_j)^2} = \frac{\sum_{j=1}^a w_j/w_j^2}{(\sum_{j=1}^a 1/w_j)^2}$$

$$= \frac{\sum_{j=1}^a 1/w_j}{(\sum_{j=1}^a 1/w_j)^2} = \frac{1}{\sum_{j=1}^a 1/w_j}. \quad (2.6)$$

See Appendix A.2 for derivations of the weighted estimate and associated variance.

If necessary, agencies can also incorporate difference that occur within their observed data sets into the calculation of $q_j^{(i)}$ and $u_j^{(i)}$. For instance, if agencies own stratified samples, they should use the formulas corresponding to stratified sampling (Särndal *et al.*, 1991, pp. 100–109) to properly estimate $Q$ and the associated variance. (See Reiter (2002) for a discussion and simulation of stratified sampling in the context of synthetic data.)

All further derivations assume an unweighted average of the agencies estimates (based on simple random samples) in the joint calculation of the scalar $Q$. However, both types of averaging will be illustrated in the simulation studies in Section 2.4.

## 2.2 Bayesian Derivation of the Combining Rules

To derive the inferences described in Section 2.1 from a Bayesian perspective, I assume both the imputations and inferences are based on the same model specification. Using the theory developed in Raghunathan *et al.* (2003) as a basis, I extend their work to incorporate multiple agencies whose data have all been drawn from the same underlying population. Therefore, the posterior distribution of $Q|d^M$ can be decomposed as

$$f(Q|d^M) = \iint f(Q|\mathcal{D}, P^M, d^M) f(\mathcal{D}|P^M, d^M) f(P^M|d^M)\, d\mathcal{D}\, dP^M, \quad (2.7)$$

where $\mathcal{D}$, $P^M$ and $d^M$ are as previously defined.

### 2.2.1 Evaluating $f(Q|\mathcal{D}, P^M, d^M)$

To evaluate the first conditional distribution in (2.7), I use only the given observed data $\mathcal{D}$, because $P^M$ and $d^M$ do not provide additional information for determining the distribution of $Q$. Therefore, $f(Q|\mathcal{D}, P^M, d^M) = f(Q|\mathcal{D})$. To find $f(Q|\mathcal{D})$, I assume that Bayesian asymptotics hold, allowing the agency-specific estimand $Q_j$ to be based on the normal distribution

$$Q_j|D_j \sim N(q_{obs,j}, v_{obs,j}), \tag{2.8}$$

where $q_{obs,j}$ and $v_{obs,j}$ are the observed estimate and associated variance computed using $D_j$, the observed data for agency $j$. The distribution of $f(Q|\mathcal{D})$ for all $a$ agencies is found by averaging (2.8) over $j$,

$$Q|\mathcal{D} \sim N(\bar{q}_{obs}, \bar{v}_{obs}/a), \tag{2.9}$$

where $\bar{q}_{obs}$ and $\bar{v}_{obs}$ are the averaged estimate and associated variance for the $a$ agencies.

### 2.2.2 Evaluating $f(\mathcal{D}|P^M, d^M)$

The second conditional distribution in (2.7) is evaluated by first dropping the synthetic samples $d^M$, because they are subsets of the completed populations in $P^M$. Therefore, $f(\mathcal{D}|P^M, d^M) = f(\mathcal{D}|P^M)$. The sufficient statistics (Casella and Berger, 1990, pp. 247–254) for $\mathcal{D}$ are the observed estimate and associated variance, $\bar{q}_{obs}$ and $\bar{v}_{obs}$ respectively. These values are sufficient to described $\mathcal{D}$, because it is based on the normal distribution (see (2.9)). Thus, the evaluation of $f(\mathcal{D}|P^M)$ can proceed by evaluating the sufficient statistics of $\mathcal{D}$, rather than $\mathcal{D}$ itself. So, $f(\mathcal{D}|P^M)$ can be written as $f(\bar{q}_{obs}, \bar{v}_{obs}|P^M) = f(\bar{q}_{obs}|\bar{v}_{obs}, P^M)f(\bar{v}_{obs}|P^M)$. Let $Q_j^{(i)}$ be the estimate of $Q$ from the completed population $P_j^{(i)}$. Since each $P_j^{(i)}$, $i = 1, 2, \ldots, m_j$, was completed from the same $D_j$, each of the $m_j$ estimates should have the same properties

as the estimate $Q_j$ from the observed data $D_j$ (see (2.8)). Therefore,

$$Q_j^{(i)}|D_j \sim N(q_{obs,j}, v_{obs,j}), \tag{2.10}$$

for $i = 1, 2, \ldots, m_j$. Similar to Raghunathan *et al.* (2003), I assume a flat prior distribution on $q_{obs,j}$, and as implied by standard Bayesian theory

$$q_{obs,j}|P_j, v_{obs,j} \sim N(\bar{Q}_j, v_{obs,j}/m_j) \tag{2.11}$$

$$\frac{(m_j - 1)W_j}{v_{obs,j}}|P_j \sim \chi^2_{m_j-1}, \tag{2.12}$$

where $\bar{Q}_j$ is the average of the population-based estimates calculated from the completed populations $P_j$, and with $W_j = \frac{1}{m_j-1}\sum_{i=1}^{m_j}(Q_j^{(i)} - \bar{Q}_j)^2$ being the variance among the estimates. Averaging (2.11) over $j$ yields

$$\bar{q}_{obs}|P^M, \bar{v}_{obs} \sim N\left(\bar{Q}_M, \frac{1}{a^2}\sum_{j=1}^{a} v_{obs,j}/m_j\right). \tag{2.13}$$

### 2.2.3 Evaluating $f(P^M|d^M)$

The last conditional distribution in (2.7) that needs evaluation is $f(P^M|d^M)$. It can also be found in terms of the corresponding sufficient statistics, $\bar{Q}_M$ and $W_j$, which allows it to be written as $f(\bar{Q}_M, W_j|d^M) = f(\bar{Q}_M|W_j, d^M)f(W_j|d^M)$. To find $f(\bar{Q}_M|W_j, d^M)$, I assume the sample estimates $(q_j^{(i)})$ of $Q$ from each $d_j^{(i)}$ are centered at their population estimates $Q_j^{(i)}$, since $d_j^{(i)}$ is a simple random sample from $P_j^{(i)}$. The associated variance for the estimate is calculated based on the sample $d_j^{(i)}$, which gives

$$q_j^{(i)}|P_j^{(i)}, u_j^{(i)} \sim N(Q_j^{(i)}, u_j^{(i)}). \tag{2.14}$$

If $d_j^{(i)}$ were equal to $P_j^{(i)}$, then $q_j^{(i)}$ would be the same as $Q_j^{(i)}$. For simplification purposes, I assume that $u_j^{(i)} = \bar{u}_j$ for all $m_j$ samples created by agency $j$. Since the

estimates have been calculated from synthetic samples randomly selected from completed populations that originated from a single observed data set, each associated variance should be similar with only minor variations (Raghunathan *et al.*, 2003). As in Raghunathan *et al.* (2003), a flat prior distribution is used on $Q_j^{(i)}$, which implies

$$Q_j^{(i)}|q_j^{(i)}, \bar{u}_j \sim N(q_j^{(i)}, \bar{u}_j). \tag{2.15}$$

Averaging (2.15) over $i$ yields

$$\bar{Q}_j|W_j, d_j \sim N(\bar{q}_j, \bar{u}_j/m_j). \tag{2.16}$$

The distribution of $f(W_j|d^M)$ will be written in terms of $d_j$ and will use the simplified $\bar{u}_j$, which results in an approximate distributional form

$$\frac{(m_j-1)b_j}{\bar{u}_j + W_j}|d_j \sim \chi^2_{m_j-1}. \tag{2.17}$$

This approximation was found by first finding an estimate for $b_j$, the sample variance of $q_j^{(i)}$ for $i = 1, 2, \ldots, m_j$. To estimate $b_j$, I condition on the population from which $X$ and $Y$ originate, which yields a conditional variance that can be evaluated by conditioning on the sampled units from $X$ and $Y$

$$\begin{aligned}
\mathbb{V}(q_j^{(i)}|X,Y) &= \mathbb{E}\{\mathbb{V}(q_j^{(i)}|X,Y,I)|X,Y\} + \mathbb{V}\{\mathbb{E}(q_j^{(i)}|X,Y,I)|X,Y\} \\
&= \mathbb{E}\{u_j^{(i)}|X,Y\} + \mathbb{V}\{Q_j^{(i)}|X,Y\} \;=\; \bar{u}_j + W_j.
\end{aligned}$$

This approximation depends on the simplifying assumption used to find (2.15), allowing $u_j^{(i)}$ to be replaced by $\bar{u}_j$. Given that $q_j^{(i)}$ is normally distributed (see (2.14)) and $b_j$ is the sample variance of $q_j^{(i)}$ for $i = 1, 2, \ldots, m_j$, if $(m_j - 1)b_j$ is divided by the associated population-based estimate derived above, a $\chi^2$ distribution with $m_j - 1$ degrees of freedom results (Casella and Berger, 1990, pp. 220–221). This approximation will be sufficient for all subsequent derivations.

### 2.2.4 Evaluating $f(Q|d^M)$

To find the distributional form of $f(Q|d^M)$, multiply (2.9), (2.13), (2.12), the average of (2.16) over $j$, and (2.17), then integrate with respect to $\bar{q}_{obs}$, $v_{obs,j}$, $\bar{Q}_M$ and $W_j$. Combining each of these components in their conditional distributional forms into an integral yields

$$\begin{aligned} f(Q|d^M) &= \int\int\int\int N(Q;\bar{q}_{obs},\bar{v}_{obs}/a)\, N\left(\bar{q}_{obs};\bar{Q}_M,\frac{1}{a^2}\sum_{j=1}^{a} v_{obs,j}/m_j\right) \\ &\times \left(\prod_{j=1}^{a}\text{Inverse }\chi^2(v_{obs,j};m_j-1,W_j)\right)\, N\left(\bar{Q}_M;\bar{q}_M,\frac{1}{a^2}\sum_{j=1}^{a}\bar{u}_j/m_j\right) \\ &\times \left(\prod_{j=1}^{a}\text{Inverse }\chi^2(\bar{u}_j+W_j;m_j-1,b_j)\right) d\bar{q}_{obs}\, dv_{obs,j}\, d\bar{Q}_M\, dW_j. \quad (2.18) \end{aligned}$$

The evaluation of these integrals requires the use of numerical integration because a closed form expression cannot be explicitly found. It is possible to evaluate the integrals from a Bayesian perspective, by simulating each parameter from its respective conditional distribution and averaging accordingly. However, a method that can easily be implemented by all users is desired, thus an approximation to (2.18) is needed. Whenever $m_j$ is large, expectation approximations can be used in the variance terms; for instance, use $\mathbb{E}(W_j|d_j) \approx b_j - \bar{u}_j$ instead of $\mathbb{E}(W_j|d_j) = \frac{m_j}{m_j-2}(b_j-\bar{u}_j)$. These assumptions and approximations give the following for the estimate of $Q$:

$$\begin{aligned} \mathbb{E}(Q|d^M) &= \mathbb{E}\{\mathbb{E}(Q|P^M)|d^M\} = \mathbb{E}\{\mathbb{E}(\mathbb{E}(Q|\mathcal{D})|P^M)|d^M\} \\ &= \mathbb{E}\{\mathbb{E}(\bar{q}_{obs}|P^M)|d^M\} = \mathbb{E}\{\bar{Q}_M|d^M\} = \bar{q}_M. \end{aligned}$$

Similarly, the associated variance becomes

$$\begin{aligned}
\mathbb{V}(Q|d^M) &= \mathbb{E}\{\mathbb{V}(Q|P^M)|d^M\} + \mathbb{V}\{\mathbb{E}(Q|P^M)|d^M\} \\
&= \mathbb{E}\{\mathbb{E}(\mathbb{V}(Q|\mathcal{D})|P^M) + \mathbb{V}(\mathbb{E}(Q|\mathcal{D})|P^M)|d^M\} + \mathbb{V}\{\mathbb{E}(\mathbb{E}(Q|\mathcal{D})|P^M)|d^M\} \\
&= \mathbb{E}\{\mathbb{E}(\bar{v}_{obs}/a|P^M) + \mathbb{V}(\bar{q}_{obs}|P^M)|d^M\} + \mathbb{V}\{\mathbb{E}(\bar{q}_{obs}|P^M)|d^M\} \\
&= \mathbb{E}\left\{\frac{1}{a^2}\sum_{j=1}^{a} W_j + \frac{1}{a^2}\sum_{j=1}^{a} W_j/m_j|d^M\right\} + \mathbb{V}\{\bar{Q}_M|d^M\} \\
&\approx \frac{1}{a^2}\sum_{j=1}^{a}(1+1/m_j)b_j - \bar{u}_j.
\end{aligned}$$

The estimate and associated variance, along with the prescribed integration, suggest the posterior distribution of $Q|d^M$ should be

$$Q|d^M \sim t_{\nu_H}\left(\bar{q}_M, \frac{1}{a^2}\sum_{j=1}^{a}(1+1/m_j)b_j - \bar{u}_j\right) \tag{2.19}$$

with approximately $\nu_H$ degrees of freedom (see Section 2.1). For large $m_j$, the normal approximation $(Q - \bar{q}_M) \sim N(0, T_H)$ can be used, but when moderate, (2.19) should be used for inferences.

## 2.3 Extension to Partially Synthetic Data

Rather than creating fully synthetic data sets, agencies may choose to replace only those values at high risk of disclosure, leaving the rest unaltered (Little, 1993; Kennickell, 1997; Abowd and Woodcock, 2001; Little and Liu, 2003; Reiter, 2003). This type of data replacement strategy yields partially synthetic data, as the resulting data contains a combination of real and synthetic values.

Similar to how the data sharing method of Sections 2.1 and 2.2 extends the fully synthetic data method of Raghunathan *et al.* (2003) to include multiple agencies,

the partially synthetic data method (see Section 1.2.2) can also be extended. The inclusion of multiple agencies, who individually own data sets that originate from the same population, into the process requires modified combining rules and variance equation to obtain valid inferences from the combined synthetic samples. However, as in the previous case, if only one agency is present, the results are the same as those derived in Reiter (2003). For complete details and derivations see Appendix A.3.

## 2.4 Simulation Studies

To show proof of concept, simulation studies were conducted to illustrate horizontally partitioned data sharing using artificial data and the correct posterior distributions for all data generations. The simulations use a three-agency sharing strategy and are implemented with fully synthetic data. All simulations are performed using the software package *R* (R Development Core Team, 2004).

### 2.4.1 Simulation I: Linear Regression

The first simulation I conducted uses a linear regression model to relate $Y$ and $X$, where $X$ and $Y$ are univariate. Infinite populations are assumed to exist for both $X$ and $Y$, which have an underlying multivariate normal distribution with means equal to 0, variances equal to 10 and a common correlation of 0.5. This model specification enables direct sampling of the observed data sets from a multivariate normal distribution. Sampling of the synthetic data will proceed using a Bayesian analysis of the regression of $Y$ on $X$.

In Section 2.1, I assumed that $X$ was fully known to each of the participating agencies. Therefore, the synthetic samples of $X$, denoted as $X_j^{(i)}$, are selected by drawing samples of size $n_{syn,j}$ directly from the known population distribution of $X$, for each of the $i = 1, 2, \ldots, m_j$ replicates I created for agency $j$. Next, the values

of $Y$ corresponding to $X_j^{(i)}$ for each of the $m_j$ samples, are drawn and denoted as $Y_{syn,j}^{(i)}$. The values of $Y$ are generated using an agency-specific posterior predictive distribution and observed data $D_j$:

$$\begin{aligned} f(Y_{syn,j}^{(i)}|X_{obs,j}, Y_{obs,j}, X_j^{(i)}) &= \int f(Y_{syn,j}^{(i)}|X_{obs,j}, Y_{obs,j}, X_j^{(i)}, \theta_j) \\ &\times f(\theta_j|X_{obs,j}, Y_{obs,j}, X_j^{(i)}) d\theta_j. \end{aligned} \tag{2.20}$$

To find the posterior predictive distribution, or rather the conditional distributions that comprise it, a linear regression model is formed with standard non-informative prior distributions on all parameters,

$$Y_{obs,j}|X_{obs,j}, \theta_j \sim N(X_{obs,j}\beta_j, \sigma_j^2 I) \tag{2.21}$$

$$\theta_j|X_{obs,j} \propto 1/\sigma_j^2 \tag{2.22}$$

where $\theta_j = (\beta_j, \sigma_j^2)$ represents the agency-specific parameters. To create the synthetic data, the conditional distributions of $\beta_j$ and $\sigma_j^2$ are found in terms of the observed data, proceeding similarly to $Y_{syn,j}^{(i)}$. (This sampling structure is similar to that in Rubin (1987, pp. 166–167) and Gelman *et al.* (2004, pp. 355–359).) The sampling then proceeds as outlined below

$$\sigma_{j*}^2|X_{obs,j}, Y_{obs,j} = \frac{(n_j - p^*)S_j^2}{\chi^2_{n_j - p^*}} \tag{2.23}$$

$$\beta_{j*}|X_{obs,j}, Y_{obs,j}, \sigma_{j*}^2 \sim N(\hat{\beta}_j, \hat{\Sigma}_j) \tag{2.24}$$

$$Y_{syn,j}^{(i)}|X_{obs,j}, Y_{obs,j}, X_j^{(i)}, \theta_{j*} \sim N(X_j^{(i)}\beta_{j*}, \sigma_{j*}^2 I) \tag{2.25}$$

where $(n_j - p^*)S_j^2 = (Y_{obs,j} - X_{obs,j}\hat{\beta}_j)'(Y_{obs,j} - X_{obs,j}\hat{\beta}_j)$, $\hat{\beta}_j = (X'_{obs,j}X_{obs,j})^{-1}X'_{obs,j}Y_{obs,j}$, $\hat{\Sigma}_j = \sigma_{j*}^2(X'_{obs,j}X_{obs,j})^{-1}$, and $p^*$ includes the intercept.

Estimates of $q_j^{(i)}$ and $u_j^{(i)}$ are then calculated from each $d_j^{(i)}$, which are used to find values for (2.1), (2.2), (2.3), and the associated variance (2.4). Due to the negative

sign in (2.4), negative variances may arise for some values of $b_j$ and $\bar{u}_j$. Therefore, an adjustment measure is needed to force all variance estimates to be non-negative. Using the formula $T^* = \max(0, T) + \delta\left(\frac{n_{syn}}{n_{obs}}\bar{u}_M\right)$ from Reiter (2002) as a basis, the negative agency components which comprise the variance are adjusted with

$$T_H^* = \frac{1}{a^2}\sum_{j=1}^{a}\max(0, w_j) + \delta\left(\frac{n_{syn,j}}{n_j}\bar{u}_j + \frac{b_j}{m_j}\right) \tag{2.26}$$

where $w_j = (1 + 1/m_j)b_j - \bar{u}_j$ such that $\delta = 1$ when $w_j < 0$ and 0 otherwise. The addition of $b_j/m_j$ to the variance adjustment formula helps account for the variability among the synthetic data sets when $m_j$ is small. Generally, it is possible to avoid negative variances by increasing the size of $n_{syn,j}$ or $m_j$, as evident by the simulation results in Tables 2.1 and 2.2. However, additional research is needed concerning (2.26) to fully investigate the properties and determine whether a better alternative exists.

The simulations I conducted had observed data sets with 1,000; 5,000 and 10,000 observations for the three agencies, respectively. Values chosen for $m_j$ ranged from small to large; both constant and non-constant variations of $m_j$ across agencies were considered. I replicated the simulations with $n_{syn,j} = 1,000;\ 10,000;$ 20,000 and 50,000 to check whether the data sharing method was sensitive to the size of the synthetic data sets. In all simulations, $n_{syn,j} = n_{syn}$ for all $j$.

Table 2.1 gives the unweighted and weighted simulation results when $Q$ is the population mean of $Y$. As evident from the table, the estimates of $Q$ derived from the shared synthetic data samples are unbiased for the true value of $Q$, which is known to equal zero. The $\%T_H < 0$ column in the table gives the percentage of simulations with at least one negative variance component and the percentage of negative variance components for all agencies and simulations. (Only the negative components were adjusted prior to use; all others remained unchanged.) As previously stated, the

**Table 2.1**: 1,000 simulations combining data from three agencies when $Q$ is the population mean of $Y$.

| | $\bar{q}^{\star}_{obs}$ | $\bar{v}^{\star}_{obs}$ | $\bar{q}^{\star}_{M}$ | $\mathbb{V}(\bar{q}_M)^{\star}$ | $T_H(T^*_H)^{\star}$ | $\%T_H < 0^{\star\star}$ | 95% $\text{CI}_N(\text{CI}_t)$ |
|---|---|---|---|---|---|---|---|
| $n_{syn} = 1,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 5, 4, 3$ | -0.19 | 1.44 | 1.25 | 2.17 | 2.11 (3.34) | 78.5% (40.1%) | 95% (99.3%) |
| $m = 4$ | -0.43 | 1.44 | 0.94 | 2.16 | 2.19 (3.44) | 78.5% (39.4%) | 94.9% (99.6%) |
| $m = 25$ | -0.21 | 1.44 | -1.07 | 1.24 | 1.31 (1.62) | 57.3% (23.2%) | 95.4% (99.7%) |
| $m = 45, 35, 25$ | -0.01 | 1.45 | -0.92 | 1.22 | 1.24 (1.49) | 55.3% (21.5%) | 95.3% (98.9%) |
| $m = 50$ | 1.59 | 1.45 | 1.23 | 1.15 | 1.17 (1.37) | 51.1% (20%) | 95.5% (98.9%) |
| *weighted average of agencies* | | | | | | | |
| $m = 5, 4, 3$ | -0.25 | 0.62 | 0.86 | 2.69 | 1.70 | 78.5% (40.1%) | 82.6% |
| $m = 4$ | -0.80 | 0.62 | 0.96 | 2.42 | 1.67 | 78.5% (39.4%) | 83.3% |
| $m = 25$ | 1.09 | 0.62 | -1.12 | 0.95 | 0.89 | 57.3% (23.2%) | 88.3% |
| $m = 45, 35, 25$ | -0.82 | 0.63 | -0.89 | 0.86 | 0.85 | 55.3% (21.5%) | 91.6% |
| $m = 50$ | 1.16 | 0.63 | 1.50 | 0.78 | 0.72 | 51.1% (20%) | 89.7% |
| $n_{syn} = 10,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 5, 4, 3$ | -0.14 | 1.45 | 0.17 | 1.33 | 1.43 (1.57) | 48.6% (19%) | 94% (98.6%) |
| $m = 4$ | 0.90 | 1.44 | -0.21 | 1.41 | 1.50 (1.64) | 42.7% (16.7%) | 92% (98.6%) |
| $m = 25$ | -2.32 | 1.44 | -1.78 | 1.11 | 1.14 (1.14) | 3.8% (1.3%) | 93.7% (94.7%) |
| $m = 45, 35, 25$ | -0.43 | 1.45 | -0.01 | 1.15 | 1.12 (1.13) | 4.6% (1.5%) | 94.6% (94.8%) |
| $m = 50$ | 0.96 | 1.44 | 0.58 | 1.16 | 1.10 (1.10) | 0.8% (0.3%) | 94.3% (94.8%) |
| *weighted average of agencies* | | | | | | | |
| $m = 5, 4, 3$ | 0.33 | 0.63 | 0.10 | 1.03 | 0.66 | 48.6% (19%) | 81.3% |
| $m = 4$ | 0.81 | 0.62 | -0.21 | 1.15 | 0.63 | 42.7% (16.7%) | 79.4% |
| $m = 25$ | -0.47 | 0.62 | -0.96 | 0.56 | 0.46 | 3.8% (1.3%) | 88.7% |
| $m = 45, 35, 25$ | 0.49 | 0.63 | 0.76 | 0.55 | 0.46 | 4.6% (1.5%) | 89.1% |
| $m = 50$ | 1.02 | 0.62 | 0.88 | 0.54 | 0.45 | 0.8% (0.3%) | 89.8% |
| $n_{syn} = 20,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | -0.69 | 1.44 | 0.15 | 1.34 | 1.36 (1.43) | 28.1% (10.1%) | 90.5% (95.7%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | -0.76 | 0.62 | 0.10 | 0.97 | 0.51 | 28.1% (10.1%) | 76.2% |
| $n_{syn} = 50,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 1.43 | 1.44 | 1.63 | 1.38 | 1.33 (1.35) | 10.6% (3.6%) | 88.7% (93.9%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 0.79 | 0.63 | 1.20 | 0.95 | 0.44 | 10.6% (3.6%) | 75% |

$\star$ multiply by $10^{-3}$

$\star\star$ % of simulations with at least 1 variance component or $w_j < 0$ (% of total (across all agencies) variance components or $w_j < 0$)

percentage of negative variance components decreases as $m_j$ increases. This decrease is most visible when $n_{syn}$ becomes large with large $m_j$. As verification of (2.4), the variance observed among estimates ($\mathbb{V}(\bar{q}_M)$) is similar to the value calculated for $T_H$.

The coverage of the 95% confidence intervals for the known population mean, calculated using the shared synthetic data sets and a normal approximation, are quite good for all values of $m_j$ and $n_{syn}$ when the agency contributions to $Q$ are unweighted. In most cases, the coverage nearly equals or exceeds the desired level of 95% and is most likely a direct result of the adjustment measure used (see (2.26)) on the negative variance components considering their high occurrence levels. Comparatively, in most cases the coverage of the 95% confidence intervals, as calculated with a t-distribution, is higher than that calculated with the normal approximation; a direct result of the assumptions made during derivation (see Appendix A.1) and calculation of the approximate degrees of freedom (For some values of $b_j$ and $\bar{u}_j$, the degrees of freedom was set equal to one because the calculated value was less than one.) These assumptions inflate the coverage whenever both $m_j$ and $n_{syn}$ are small, and in cases where $m_j$ is small for moderate values of $n_{syn}$ (see the last column of Table 2.1). Further research is required on the degrees of freedom derivation for (2.19) that does not involve underlying assumptions or adjustments.

Differences between agencies, whether observed sample sizes or synthetic data generation techniques, can be incorporated into the estimation of $Q$ as mentioned in Section 2.1. The simulation construction created variability in the agency's estimates due to the differences in their original data sizes. Using the derived formulae (see (2.5) and (2.6)), additional results weighting agency input were calculated using the same shared synthetic data sets as used in the unweighted results. In some cases, the weights used in the estimation were negative, due to the values of $b_j$ and $\bar{u}_j$, and required adjustment. A modified version of the adjustment measure from Reiter

(2002) was used because individually agencies' synthetic samples can be classified as fully synthetic data. As before, I add a component to help measure the variability between samples when $m_j$ is finite,

$$w_j^* = \max(0, w_j) + \delta \left( \frac{n_{syn,j}}{n_j} \bar{u}_j + \frac{b_j}{m_j} \right) \tag{2.27}$$

where $\delta = 1$ when $w_j < 0$ and 0 otherwise. Once again, the percentages in the $\%T_H < 0$ column of Table 2.1 are the percentage of simulations with at least one $w_j < 0$, with the second value representing the percentage of $w_j < 0$ for all agencies and simulations. (Only the negative weights were adjusted prior to use; all others remain unchanged.) As evident by the results in Table 2.1, weighted estimation of $Q$ does not provide the same level of confidence interval coverage as seen for the unweighted average. This may be due to the difference between the weighted $T_H$ and $\mathbb{V}(\bar{q}_M)$, especially in cases when $m_j$ is small.

To illustrate other quantities for $Q$, analogous results were calculated for a $Q$ representing the population regression coefficient of $Y$ on $X$ (see Table 2.2). Once again, unweighted and weighted versions of the estimation are given, with results calculated using the same shared synthetic data sets as in Table 2.1. It is evident from Table 2.2 that the shared data sets give an unbiased estimate for $Q$. In the unweighted case, $\mathbb{V}(\bar{q}_M)$ and $T_H$ are tracking similar amounts of variability, as seen by their similar values. Although, this is not true in the weighted estimation, $T_H$ is consistently lower than $\mathbb{V}(\bar{q}_M)$, but both values are at least the size of the observed variance estimate.

As in the case of the population mean, similar trends are seen with the coverage of the 95% confidence intervals calculated using a normal approximation and t-distribution. Likewise, the percentage of negative variance components or weights decreases with increases in $n_{syn}$ and $m_j$, becoming very small for large $n_{syn}$.

**Table 2.2**: 1,000 simulations combining data from three agencies when $Q$ is the regression coefficient of $Y$ on $X$.

| | $\bar{q}_{obs}$ | $\bar{v}^{\star}_{obs}$ | $\bar{q}_M$ | $\mathbb{V}(\bar{q}_M)^{\star}$ | $T_H(T^*_H)^{\star}$ | $\%T_H < 0^{\star\star}$ | 95% $\text{CI}_N(\text{CI}_t)$ |
|---|---|---|---|---|---|---|---|
| $n_{syn} = 1,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 5, 4, 3$ | 0.5 | 0.11 | 0.5 | 0.21 | 0.20 (0.28) | 76.3% (36.9%) | 94% (100%) |
| $m = 4$ | 0.5 | 0.11 | 0.5 | 0.18 | 0.20 (0.28) | 75.5% (37.8%) | 95.5% (99.9%) |
| $m = 25$ | 0.5 | 0.11 | 0.5 | 0.12 | 0.12 (0.14) | 54.1% (21.3%) | 94.1% (99%) |
| $m = 45, 35, 25$ | 0.5 | 0.11 | 0.5 | 0.12 | 0.12 (0.14) | 49.4% (19.2%) | 94.7% (100%) |
| $m = 50$ | 0.5 | 0.11 | 0.5 | 0.11 | 0.12 (0.13) | 43.3% (16.1%) | 96.2% (100%) |
| *weighted average of agencies* | | | | | | | |
| $m = 5, 4, 3$ | 0.5 | 0.05 | 0.5 | 0.26 | 0.13 | 76.3% (36.9%) | 78.3% |
| $m = 4$ | 0.5 | 0.05 | 0.5 | 0.22 | 0.13 | 75.5% (37.8%) | 82.8% |
| $m = 25$ | 0.5 | 0.05 | 0.5 | 0.10 | 0.08 | 54.1% (21.3%) | 85.8% |
| $m = 45, 35, 25$ | 0.5 | 0.05 | 0.5 | 0.08 | 0.07 | 49.4% (19.2%) | 87.8% |
| $m = 50$ | 0.5 | 0.05 | 0.5 | 0.07 | 0.06 | 43.3% (16.1%) | 88.5% |
| $n_{syn} = 10,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 5, 4, 3$ | 0.5 | 0.11 | 0.5 | 0.13 | 0.15 (0.15) | 41.5% (15.4%) | 92% (100%) |
| $m = 4$ | 0.5 | 0.11 | 0.5 | 0.14 | 0.14 (0.15) | 35.1% (13.2%) | 91.1% (97.4%) |
| $m = 25$ | 0.5 | 0.11 | 0.5 | 0.12 | 0.11 (0.11) | 1.9% (0.6%) | 93.7% (94.2%) |
| $m = 45, 35, 25$ | 0.5 | 0.11 | 0.5 | 0.11 | 0.11 (0.11) | 2.1% (0.7%) | 94.9% (100%) |
| $m = 50$ | 0.5 | 0.11 | 0.5 | 0.11 | 0.11 (0.11) | 0.1% (0.03%) | 94.9% (100%) |
| *weighted average of agencies* | | | | | | | |
| $m = 5, 4, 3$ | 0.5 | 0.05 | 0.5 | 0.11 | 0.06 | 41.5% (15.4%) | 79.7% |
| $m = 4$ | 0.5 | 0.05 | 0.5 | 0.12 | 0.05 | 35.1% (13.2%) | 76.4% |
| $m = 25$ | 0.5 | 0.05 | 0.5 | 0.06 | 0.05 | 1.9% (0.6%) | 89.1% |
| $m = 45, 35, 25$ | 0.5 | 0.05 | 0.5 | 0.05 | 0.05 | 2.1% (0.7%) | 89.2% |
| $m = 50$ | 0.5 | 0.05 | 0.5 | 0.05 | 0.05 | 0.1% (0.03%) | 91.7% |
| $n_{syn} = 20,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 0.5 | 0.11 | 0.5 | 0.14 | 0.14 (0.15) | 22.3% (8.1%) | 89.6% (95.2%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 0.5 | 0.05 | 0.5 | 0.10 | 0.05 | 22.3% (8.1%) | 74.8% |
| $n_{syn} = 50,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 0.5 | 0.11 | 0.5 | 0.13 | 0.14 | 8.8% (3%) | 90.1% (94.3%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 0.5 | 0.05 | 0.5 | 0.09 | 0.04 | 8.8% (3%) | 74.1% |

$\star$ multiply by $10^{-3}$

$\star\star$ % of simulations with at least 1 variance component or $w_j < 0$ (% of total (across all agencies) variance components or $w_j < 0$)

### 2.4.2 Simulation II: Logistic Regression

To illustrate another data sharing scenario, I next consider a logistic regression. As in Section 2.4.1, infinite populations are assumed to exist for both $X$ and $Y$, where $X$ has an underlying normal distribution with mean 0 and variance 10, and $Y$ is a $0-1$ response indicator. Once again a three-agency sharing strategy is implemented.

The method used in Section 2.4.1 to select synthetic samples for $X$ is utilized again. To find the conditional distributions that comprise the posterior predictive distribution (similar to (2.20)), the following logistic model and quadratic link function specifications are used

$$\begin{aligned} Y_{obs,j}|X_{obs,j},\beta &\sim \text{Bernoulli}(\pi_j) && (2.28)\\ \text{logit}(\pi_j)|X_{obs,j} &= \beta_0+\beta_1 X_{obs,j}+\beta_2 X_{obs,j}^2 && (2.29)\end{aligned}$$

where $\beta_0=-3$, $\beta_1=4$ and $\beta_2=1$. The choice of a quadratic link function and values was arbitrary and chosen mainly to illustrate a logistic regression with more than a single variable, without the inclusion of interaction terms.

To create the synthetic data sets, the conditional distributions of $\beta$ and $Y_{syn,j}^{(i)}$ in terms of the observed data are derived. These distributions give a general structure for the sampling process, which proceeds as follows

$$\begin{aligned} \beta_{j*}|X_{obs,j} &\sim N(\hat{\beta}_j,\hat{\Sigma}_j) && (2.30)\\ \text{logit}(\pi_{j*}) &= \beta_{0*}+\beta_{1*}X_j^{(i)}+\beta_{2*}(X_j^{(i)})^2 && (2.31)\\ Y_{syn,j}^{(i)}|X_{obs,j},X_j^{(i)},\beta_{j*} &\sim \text{Bernoulli}(\pi_{j*}) && (2.32)\end{aligned}$$

where $\hat{\beta}$ are the regression coefficients and $\hat{\Sigma}_j$ is the covariance matrix as calculated by the logistic regression model of the observed data (output from the `glm` command in *R*).

The simulations conducted use observed data sets with 1,000; 5,000 and 10,000 observations for the three participating agencies (the same set up as in the linear regression simulation). Values for $m$ ranged from small to large, with $m_j = m$ for all agencies. Likewise, $n_{syn,j} = n_{syn}$ for all agencies and was evaluated at $n_{syn} = 1,000$; 10,000; 20,000 and 50,000 to determine how the synthetic data set size affects output.

The results in Tables 2.3 and 2.4 show trends similar to those observed in the linear regression simulations. The estimates of $Q$ based on the shared synthetic data sets closely estimate the true values of $\beta_1$ and $\beta_2$. The percentage of negative variance components and weights decrease as $m$ and $n_{syn}$ increase. (The same adjustments measures used in Section 2.4.1 were implemented again.) Negatives can be avoided if both $m$ and $n_{syn}$ are large (see Table 2.4), but it is unlikely agencies would want to release large numbers of synthetic data sets.

Coverage of the 95% confidence intervals for the true coefficient values are rather low when calculated with a normal approximation for small $m$, but increase with combinations of large $m$ and $n_{syn}$. Comparable to the previous simulations, coverage of the 95% confidence intervals calculated with a t-distribution overestimate the coverage when $m$ and $n_{syn}$ are small, due to the underlying assumptions and adjustments made to the calculated the approximate degrees of freedom.

**Table 2.3**: 1,000 simulations combining data from three agencies when $Q$ is the logistic regression coefficient ($\beta_1$) of $Y$ on $X$.

| | $\bar{q}_{obs}$ | $\bar{v}^{\star}_{obs}$ | $\bar{q}_M$ | $\mathbb{V}(\bar{q}_M)^{\star}$ | $T_H(T^*_H)^{\star}$ | $\%T_H < 0^{\star\star}$ | 95% $\text{CI}_N(\text{CI}_t)$ |
|---|---|---|---|---|---|---|---|
| $n_{syn} = 1,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 4.04 | 2.16 | 4.14 | 4.49 | 4.84 (6.56) | 75.8% (36.6%) | 93% (99.8%) |
| $m = 25$ | 4.04 | 2.12 | 4.14 | 2.67 | 3.02 (3.37) | 50% (19.3%) | 89.6% (97.3%) |
| $m = 50$ | 4.04 | 2.13 | 4.14 | 2.41 | 2.86 (3.06) | 38.6% (14.2%) | 91.8% (94.7%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 4 | 0.85 | 4.07 | 3.78 | 2.65 | 75.8% (36.6%) | 83.9% |
| $m = 25$ | 4 | 0.85 | 4.09 | 1.86 | 1.62 | 50% (19.3%) | 83.4% |
| $m = 50$ | 4 | 0.85 | 4.09 | 1.52 | 1.45 | 38.6% (14.2%) | 83.2% |
| $n_{syn} = 10,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 4.03 | 2.10 | 4.04 | 2.65 | 2.95 (3.11) | 38.6% (14.1%) | 92.3% (97.3%) |
| $m = 25$ | 4.03 | 2.11 | 4.04 | 2.23 | 2.21 (2.22) | 1.6% (0.5%) | 93.8% (94.5%) |
| $m = 50$ | 4.03 | 2.12 | 4.04 | 2.13 | 2.19 (2.19) | 0.1% (0.03%) | 95.2% (95.5%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 4 | 0.85 | 4.01 | 1.88 | 1.01 | 38.6% (14.1%) | 79.1% |
| $m = 25$ | 4 | 0.85 | 4.01 | 1.03 | 0.82 | 1.6% (0.5%) | 89% |
| $m = 50$ | 4 | 0.85 | 4.01 | 0.95 | 0.85 | 0.1% (0.03%) | 92.3% |
| $n_{syn} = 20,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 4.03 | 2.12 | 4.04 | 3.11 | 2.78 (2.86) | 23.9% (8.7%) | 88.4% (93.9%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 4 | 0.85 | 4 | 2.16 | 0.88 | 23.9% (8.7%) | 78.1% |
| $n_{syn} = 50,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 4.04 | 2.13 | 4.04 | 2.95 | 2.64 (2.67) | 7.6% (2.6%) | 87.8% (93.1%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 4 | 0.85 | 4 | 1.91 | 0.80 | 7.6% (2.6%) | 75.2% |

$\star$ multiple by $10^{-2}$

$\star\star$ % of simulations with at least 1 variance component or $w_j < 0$ (% of total (across all agencies) variance components or $w_j < 0$)

**Table 2.4**: 1,000 simulations combining data from three agencies when $Q$ is the logistic regression coefficient ($\beta_2$) of $Y$ on $X$.

| | $\bar{q}_{obs}$ | $\bar{v}^{\star}_{obs}$ | $\bar{q}_M$ | $\mathbb{V}(\bar{q}_M)^{\star}$ | $T_H(T^*_H)^{\star}$ | $\%T_H < 0^{\star\star}$ | 95% $\text{CI}_N(\text{CI}_t)$ |
|---|---|---|---|---|---|---|---|
| $n_{syn} = 1,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 1.01 | 0.14 | 1.04 | 0.29 | 0.31 (0.42) | 75.7% (36.9%) | 92.5% (99.9%) |
| $m = 25$ | 1.01 | 0.14 | 1.03 | 0.17 | 0.19 (0.21) | 48.7% (18.9%) | 90.8% (96.8%) |
| $m = 50$ | 1.01 | 0.14 | 1.03 | 0.15 | 0.18 (0.19) | 37.1% (13.9%) | 92.7% (95.6%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 1 | 0.05 | 1.02 | 0.24 | 0.17 | 75.7% (36.9%) | 83% |
| $m = 25$ | 1 | 0.05 | 1.02 | 0.12 | 0.10 | 48.7% (18.9%) | 83.1% |
| $m = 50$ | 1 | 0.05 | 1.02 | 0.09 | 0.09 | 37.1% (13.9%) | 84.9% |
| $n_{syn} = 10,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 1.01 | 0.13 | 1.01 | 0.17 | 0.19 (0.20) | 39.8% (14.5%) | 91.1% (97.4%) |
| $m = 25$ | 1.01 | 0.13 | 1.01 | 0.14 | 0.14 (0.14) | 1.9% (0.6%) | 95% (95.5%) |
| $m = 50$ | 1.01 | 0.13 | 1.01 | 0.14 | 0.14 (0.14) | — | 95% (95.2%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 1 | 0.05 | 1 | 0.12 | 0.06 | 39.8% (14.5%) | 77.5% |
| $m = 25$ | 1 | 0.05 | 1 | 0.06 | 0.05 | 1.9% (0.6%) | 89.2% |
| $m = 50$ | 1 | 0.05 | 1 | 0.06 | 0.05 | — | 92.3% |
| $n_{syn} = 20,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 1.01 | 0.13 | 1.01 | 0.20 | 0.18 (0.18) | 22.6% (8.2%) | 88.9% (94%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 1 | 0.05 | 1 | 0.12 | 0.06 | 22.6% (8.2%) | 77.8% |
| $n_{syn} = 50,000$ | | | | | | | |
| *unweighted average of agencies* | | | | | | | |
| $m = 4$ | 1.01 | 0.13 | 1.01 | 0.19 | 0.17 (0.17) | 7.5% (2.5%) | 88.5% (93.7%) |
| *weighted average of agencies* | | | | | | | |
| $m = 4$ | 1 | 0.05 | 1 | 0.12 | 0.05 | 7.5% (2.5%) | 75.2% |

$\star$ multiple by $10^{-2}$

$\star\star$ % of simulations with at least 1 variance component or $w_j < 0$ (% of total (across all agencies) variance components or $w_j < 0$)

# Chapter 3

# Vertically Partitioned Data

The situation briefly described in Chapter 1 — where two statistical agencies individually own income data and education data for the same set of respondents and desire to predict whether education affects income — is the basis for the next data sharing method. Individually, the agencies can study aspects within their data sets, but not how it may interact with other data. However, if the agencies were able to share their data and create a combined data set accessible to both, they could do just that — study how education affects income. The sharing of separate data sets regarding the same individuals enables agencies to potentially gain new perspectives of the sampled respondents and may even allow them to make previously unknown connections due to their expanded knowledge.

Before agencies can share and combine data sets, they must first decide how their data should be categorized. This gives agencies the opportunity to consider available data sharing options, if any, and decide which of the strategies best suites their circumstances, ultimate goals and underlying restrictions. (The agencies' data in the education and income example can be classified as being vertically partitioned, because the agencies have different attributes for the same set of respondents.) If agencies only want to perform linear regression analyses, then the secure matrix

products of Sanil *et al.* (2004a) is ideal. However, if agencies want to perform other types of analyses or create public-use data, then other methods need to be considered. I propose the use of an alternative data sharing method that utilizes synthetic data concepts; it not only allows analytical flexibility, but also creates public-use data as a result of the sharing process.

The utilization of synthetic data as the basis for data sharing enables agencies to reduce disclosure risks and circumvent confidentiality restrictions that limit the dissemination of public-use data. The resulting synthetic data cannot be linked back to the actual respondents because new individuals and values have been generated while still preserving the overall structure and relationships present in the original data sets.

The proposed data sharing method relies on agencies cooperating in a semi-honest environment, such that all agencies follow an agreed-upon protocol to avoid subjecting themselves or others to unnecessary disclosure risks. The method presented will be based only on a two-agency combining strategy, due to the increase in complexity with the inclusion of more than two agencies. (A sharing strategy with more than two agencies is an area of future research.) In addition, the method assumes the data sets are fully observed and one-to-one matching exists between corresponding units across data sets. To satisfy the one-to-one matching assumption, agencies must share unique unit identifiers prior to implementation to guarantee consistency across the data sets. If necessary, these decisions and exchanges between agencies could occur through some type of secure communication (Schneier, 1996). Without the assumptions of fully observed data and one-to-one matching, additional steps are needed to account for the uncertainty due to missing data and record linkage problems. In addition, the possibility of partially overlapping, vertically partitioned data will not be considered (see Reiter *et al.* (2004)).

## 3.1 Data Sharing Notation and Description

The steps two agencies must follow to securely share vertically partitioned data can be separated into a series of data exchanges. (Hereafter, the two agencies will be referred to as $\mathcal{A}$ and $\mathcal{B}$.) Let $\mathcal{A}$ own the $n \times q$ matrix $X$ and $\mathcal{B}$, the $n \times p$ matrix $Y$. Ideally, $\mathcal{A}$ wants to know $Y$ and $\mathcal{B}$ wants to know $X$, yet neither is willing to share their data. Assume $\mathcal{A}$ takes the lead role in the sharing procedure. (The delegation of the lead role is necessary to begin the sharing process and may potentially expose $\mathcal{A}$ to additional risks.)

To begin the sharing process, $\mathcal{A}$ creates a set of $r$ disguisers for $X$, that must capture the overall structure of $X$ while maintaining an ordering of the observations similar to the original data set. If disguisers are not adequately created, the relationship of $X$ and $Y$ will not be preserved, jeopardizing the entire data sharing process. (Methods for creating disguisers of $X$ will be presented in Section 3.3.3.) Suppose $\mathcal{A}$ has created disguisers that both model and protect the observations in $X$; let these disguisers be denoted as $X_{dis}^{(l)}$ for $l = 1, 2, \ldots, r$. If disguisers have been well designed, $\mathcal{A}$ can simply pass the set of $r$ disguisers of $X$ to $\mathcal{B}$ (see Section 3.3.4 for more details). However, the process considered here includes the original $X$ randomly with the $r$ disguisers, thus creating a disguiser set of size $r+1 = r^*$. It is assumed that $\mathcal{B}$ has no knowledge of this inclusion and the location of $X$ within the set, assuming each data set is simply a disguiser of $X$. Including $X$ within the set of disguisers provides the lowest level of protection in terms of disclosure prevention. ($\mathcal{B}$ has at least a $1/r^*$[1] chance of determining the location of the true $X$ within the disguisers.)

Upon receipt of the set $\{X_{dis}^{(l)}\}$, $\mathcal{B}$ models the relationship between each $X_{dis}^{(l)}$ and $Y$. (As already mentioned, the entire process can be undermined if the ordering of

[1]The chance of disclosure will be even higher if the disguisers do not adequately model the relationships between $X$ and $Y$.

the units of $X$ and hence the relationships with $Y$ are not well preserved.) Using the newly formed relationships between $Y$ and each $X_{dis}^{(l)}$ as the basis for a generation scheme, $\mathcal{B}$ creates synthetic copies of $Y$, denoted as $Y_{syn,l}^{(j)}$ for $j = 1, 2, \ldots, k$ and $l = 1, 2, \ldots, r^*$. Once complete, $\mathcal{B}$ passes the set $\{\{Y_{syn,l}^{(j)}\}\}$ to $\mathcal{A}$. Throughout this process it is necessary to preserve the disguisers' labels. Otherwise, $\mathcal{A}$ may not be able to determine which set of the synthetic copies of $Y$ were generated based on the true $X$.

As an alternative to exchanging synthetic data with $\mathcal{A}$, $\mathcal{B}$ could pass the $r^*$ imputation models needed to generate $Y_{syn,l}^{(j)}$ from $\{X_{dis}^{(l)}\}$. The exchange of models between $\mathcal{A}$ and $\mathcal{B}$ may be more efficient if, for instance, the amount of information that can passed is limited. If models are passed, $\mathcal{A}$ can use them to create $Y_{syn}$ while generating $X_{syn}$; thereby constructing fully synthetic data sets. This allows $\mathcal{A}$ and the public-release data user the ability to analyze the synthetic data sets using the methods of Raghunathan *et al.* (2003) and the extension derived in Chapter 2.

Limitations may occur when $\mathcal{B}$ and $\mathcal{A}$ only exchange models, particularly if the models used to generate $Y_{syn}$ are very complex (e.g. a semi-parametric model) or if an agency refuses to share models because they believe it could result in disclosures. In these situations, the exchange of data between agencies is a better option and will be the focus of this data sharing method.

Once $\mathcal{A}$ receives $\{\{Y_{syn,l}^{(j)}\}\}$ from $\mathcal{B}$, all of the synthetic data generated from the $r$ disguisers can be discarded. The remaining steps in the data sharing process only require the use of the $\{Y_{syn}^{(j)}\}$ created specifically from the real $X$ (suppressing the $l$ notation). After these exchanges, $\mathcal{A}$ now has a partially synthetic data set $c_j = (X, Y_{syn}^{(j)})$ for $j = 1, 2, \ldots, k$, because $X$ exists in its original form, while the true values of $Y$ have been replaced with synthetic ones as indicated by Figure 3.1. It is possible for $\mathcal{A}$ to use $c^k = \{c_j, j = 1, 2, \ldots, k\}$ in internal analyses, basing inferences on the

combining rules and variance equation in Reiter (2003). However, any publication of these results would disclose to $\mathcal{B}$ the true $X$ in $\{X_{dis}^{(l)}\}$, if $\mathcal{B}$ is allowed to keep data from intermediate steps. If agencies need to create public-use data, then the data sharing procedure must be continued; that is the assumed situation.

To create synthetic data sets that can be passed back to $\mathcal{B}$ and released as public-use data, $\mathcal{A}$ must apply the fully synthetic data method (Raghunathan *et al.*, 2003) to each partially synthetic data set, $c_j$ for $j = 1, 2, \ldots, k$. The method involves constructing completed populations $P_j^{(i)} = (X_{com}^{(i)}, Y_{com,j}^{(i)})$, where $X_{com}^{(i)} = (X, X_{nobs}^{(i)})$ and $Y_{com,j}^{(i)} = (Y_{syn}^{(j)}, Y_{nobs,j}^{(i)})$ with $X_{nobs}$ and $Y_{nobs}$ being the unobserved values of $X$ and $Y_{syn}$. The procedure is similar to that described in Chapter 2 for horizontally partitioned data, except it is necessary to complete the populations for both $X$ and $Y$.[2] This process results in $M = km$ fully synthetic samples, which are denoted $d^M = \{d_j^{(i)}, i = 1, 2, \ldots, m, j = 1, 2, \ldots, k\}$, where $d_j^{(i)} = (X_{syn,j}^{(i)}, Y_{syn,j}^{(i)})$ and is a simple random sample from the completed population $P_j^{(i)}$. The set of synthetic data sets $d^M$ are then passed back to $\mathcal{B}$ and released for public use. Therefore, $\mathcal{A}$ and $\mathcal{B}$, along with any analysts who may use the public release data can obtained valid inferences from the synthetic data sets (if they use the combining rules and variance estimate to be described).

Since $d_j^{(i)}$ is a random sample from $P_j^{(i)}$, it does not contain any real data and cannot be linked back to actual respondents, but maintains the overall structure and relationships between the original $X$ and $Y$. (This process is illustrated in Figure 3.1 with $k = 2$ and $m = 3$.) The structure of this data sharing process is comparable to both nested multiple imputation (Shen, 2000; Rubin, 2003) and the simultaneous use of multiple imputation and the partially synthetic data method (Reiter, 2004).

[2]Values could also be imputed for $X$, which would eliminate the chance of sampling observed data from $X_{com}^{(i)}$ during the selection of $d_j^{(i)}$.

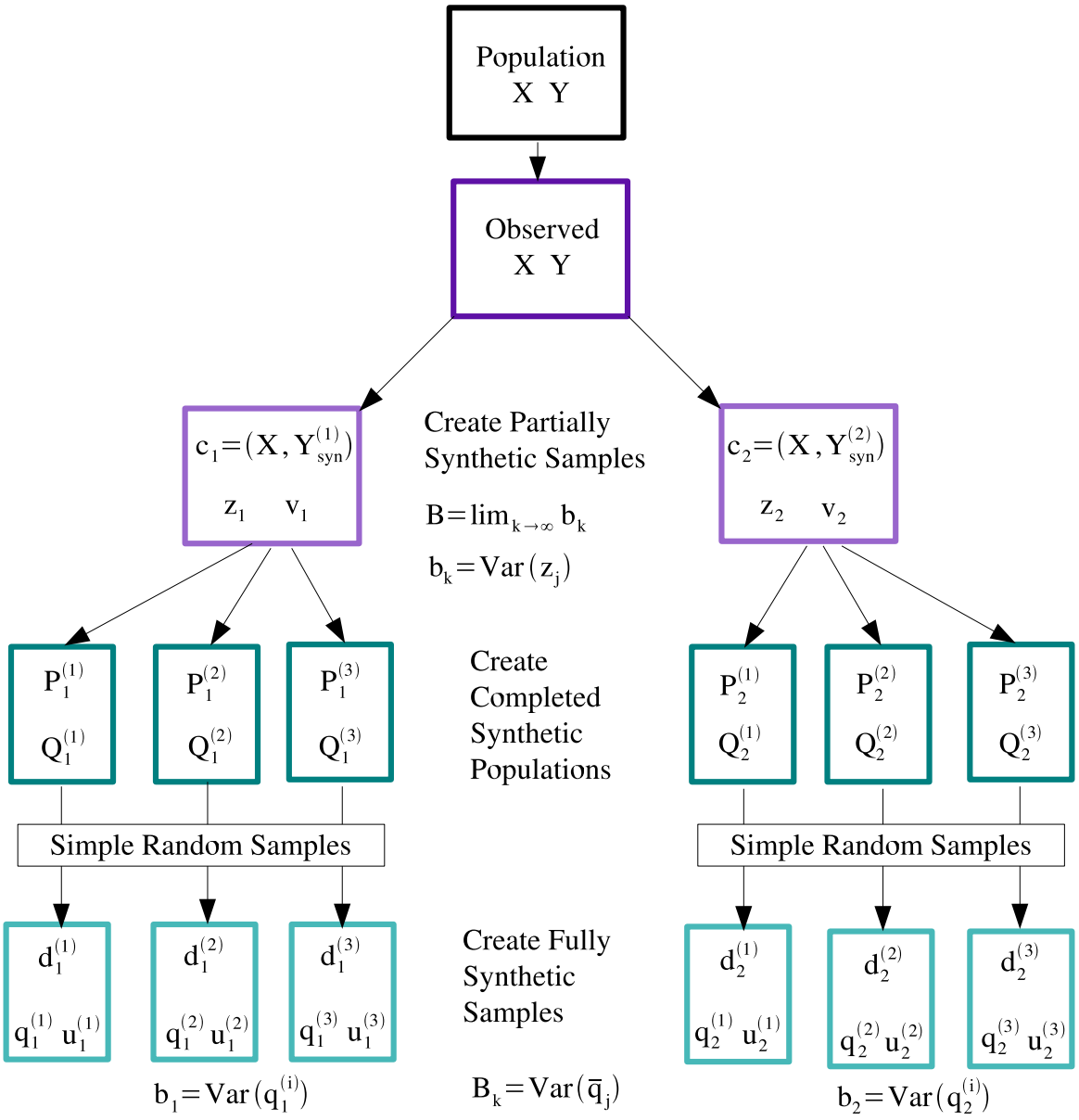


**Figure 3.1**: Vertically Partitioned Synthetic Data Sharing Diagram ($c_j$ is a partially synthetic sample with estimate $z_j$ and variance $v_j$; $P_j^{(i)}$ is a completed population with estimate $Q_j^{(i)}$; $d_j^{(i)}$ is a fully synthetic sample with estimate $q_j^{(i)}$ and associated variance $u_j^{(i)}$).

Both of these methods have a nested structure with some operations occurring first or only implemented on a portion of the data, while others are done multiple times for each data set created in the first step.

To obtain valid inferences for a population-based scalar $Q$, agencies or analysts must calculate estimates of $Q$ and the associated variance from each of the synthetic samples, $d_j^{(i)}$, these estimates will be denoted as $q_j^{(i)}$ and $u_j^{(i)}$, respectively. The calculation of these values assumes each $d_j^{(i)}$ has been randomly selected from the population of $X$ and $Y$ and is a simple random sample from the corresponding $P_j^{(i)}$. Inferences are found by combining the estimates according to the following rules

$$\bar{q}_M = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{m}\sum_{i=1}^{m} q_j^{(i)}\right) = \frac{1}{k}\sum_{j=1}^{k}\bar{q}_j \tag{3.1}$$

$$\bar{b}_M = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{m-1}\sum_{i=1}^{m}(q_j^{(i)}-\bar{q}_j)^2\right) = \frac{1}{k}\sum_{j=1}^{k} b_j \tag{3.2}$$

$$B_k = \frac{1}{k-1}\sum_{j=1}^{k}(\bar{q}_j-\bar{q}_M)^2 \tag{3.3}$$

$$\bar{u}_M = \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{m}\sum_{i=1}^{m} u_j^{(i)}\right) = \frac{1}{k}\sum_{j=1}^{k}\bar{u}_j. \tag{3.4}$$

The $\bar{q}_M$ in (3.1) is the averaged estimate across all synthetic samples. The $b_j$ in (3.2) is the variance of the estimates $q_j^{(i)}$, with $\bar{b}_M$ being the average over the $k$ groupings of synthetic samples. The $B_k$ in (3.3) is the variance of the averaged estimates $\bar{q}_j$, and $\bar{u}_M$ in (3.4) is the averaged estimated variance across all synthetic samples. This allows $\bar{q}_M$ to be used as the estimate of $Q$, with associated uncertainty measure

$$T_V = B_k/k + \bar{b}_M - \bar{u}_M. \tag{3.5}$$

When $n$, $k$, and $m$ are large, inferences for scalar $Q$ can be approximated by the normal distribution $(Q-\bar{q}_M) \sim N(0, T_V)$. However, when these values are moderate,

inferences should be based on a t-distribution with approximately

$$\nu_V = \frac{\left(\frac{B_k/k}{\bar{u}_M} + \frac{\bar{b}_M}{\bar{u}_M} - 1\right)^2}{\frac{(B_k/k)/\bar{u}_M}{k(m-1)} + \frac{\bar{b}_M/\bar{u}_M}{k-1}} \tag{3.6}$$

degrees of freedom (see Appendix B.1 for derivation).

## 3.2 Bayesian Derivations of the Combining Rules

The theory developed by Raghunathan *et al.* (2003), and Reiter (2003, 2004) will be used as a guide to derive the inferences given in Section 3.1 from a Bayesian perspective. I will assume the same model specifications are used for both the imputations and inferences. Therefore, the posterior distribution of $Q|d^M$ can be decomposed as

$$f(Q|d^M) = \iiint f(Q|c^k, B, P^M, d^M) f(c^k, B|P^M, d^M) f(P^M|d^M)\, dc^k\, dB\, dP^M, \tag{3.7}$$

where $B = \mathbb{V}(z_j|X, Y, I)$, $z_j$ is the estimate of $Q$ from $c_j$ and $I$ is the survey inclusion indicator. The $c^k$, $P^M$ and $d^M$ are as previously defined.

### 3.2.1 Evaluating $f(Q|c^k, B, P^M, d^M)$

To evaluate the first conditional distribution in (3.7), I drop $P^M$ and $d^M$ because they are not necessary given $c^k$ and $B$. Therefore, $f(Q|c^k, B, P^M, d^M) = f(Q|c^k, B)$. This expression is the posterior distribution of $Q$, conditional on $B$, for partially synthetic data as shown by Reiter (2003) and is approximately

$$Q|c^k, B \sim N(\bar{z}, B/k + \bar{v}) \tag{3.8}$$

where $\bar{z} = \frac{1}{k}\sum_{j=1}^{k} z_j$ and $\bar{v} = \frac{1}{k}\sum_{j=1}^{k} v_j$, with $v_j$ estimating the associated variance of $z_j$.

If an agency decides to stop at this point and use the set of partially synthetic samples $c^k$ for inferences, it would be necessary to integrate $f(Q|c^k, B)$ with respect to $B$. The distribution of $B|c^k$ can be written as

$$\frac{(k-1)b_k}{B}|c^k \sim \chi^2_{k-1} \tag{3.9}$$

where $b_k = \frac{1}{k-1}\sum_{j=1}^{k}(z_j - \bar{z})^2$. For large $k$, the posterior distribution of $Q|c^k$ can be approximated by the normal distribution

$$Q|c^k \sim N(\bar{z}, b_k/k + \bar{v}) \tag{3.10}$$

as shown by Reiter (2003). However, if an agency wants to create public-use data, the data sharing process must be completed, postponing the above integration step until later in the derivations. This will be the assumed situation.

### 3.2.2 Evaluating $f(c^k, B|P^M, d^M)$

The synthetic samples $d^M$ are not needed given $P^M$, which allows the conditional distribution to be written as $f(c^k, B|P^M, d^M) = f(c^k, B|P^M)$. The expression can then be evaluated in terms of the sufficient statistics $\bar{z}$, $\bar{v}$, and $B$, similar to Section 2.2.2. Therefore, to derive $f(\bar{z}, \bar{v}, B|P^M) = f(\bar{z}|\bar{v}, B, P^M)f(\bar{v}|B, P^M)f(B|P^M)$, the conditional distributions of each sufficient statistic must be found. Given the partially synthetic data set $c_j$, each population estimate of $Q$ should be centered at $z_j$ with variance $v_j$, because each of the $i = 1, 2, \ldots, m$ completed populations have been generated from the same $c_j$. Therefore, the distributional properties of each $Q_j^{(i)}$ should be similar, thus

$$Q_j^{(i)}|c_j, v_j \sim N(z_j, v_j). \tag{3.11}$$

If a flat prior distribution is put on the partially synthetic estimate $z_j$, standard Bayesian theory implies

$$z_j|v_j, B, P_j \sim N(\bar{Q}_j, v_j/m) \tag{3.12}$$

$$\frac{(m-1)W_j}{v_j}|B, P_j \sim \chi^2_{m-1} \tag{3.13}$$

where $W_j = \frac{1}{m-1}\sum_{i=1}^{m}(Q_j^{(i)} - \bar{Q}_j)^2$, the variance for the completed population estimates derived from $c_j$. To find the distribution of $\bar{z}$, average (3.12) over $j$ to yield

$$\bar{z}|\bar{v}, B, P^M \sim N(\bar{Q}_M, \bar{v}/km). \tag{3.14}$$

The distribution of $f(B|P^M)$ cannot be found analytically, but can be approximated by

$$\frac{(k-1)H}{W/m + B}|P^M \sim \chi^2_{k-1} \tag{3.15}$$

where $H = \frac{1}{k-1}\sum_{j=1}^{k}(\bar{Q}_j - \bar{Q}_M)^2$ and assuming $W = W_j$ for all $j = 1, 2, \ldots, k$. Since $B$ is the population-based variance of the estimates from $c^k$, an approximate estimate based on the completed populations is desired. The analogous estimate is $H$, the variance between the averaged completed population estimates. Therefore, by conditioning on the original population of $X$ and $Y$ and the selected sample an approximation can be found based on the following

$$\begin{aligned}\mathbb{V}(\bar{Q}_j|X, Y) &= \mathbb{E}\{\mathbb{V}(\bar{Q}_j|X, Y, I)|X, Y\} + \mathbb{V}\{\mathbb{E}(\bar{Q}_j|X, Y, I)|X, Y\} \\ &= \mathbb{E}\{W_j/m|X, Y\} + \mathbb{V}\{z_j|X, Y\} = W/m + B.\end{aligned}$$

The result is an approximation because $W_j$ is assumed to be the same for all $j$. Similar to (2.17), the resulting distribution will be $\chi^2$ because the sample variance of $\bar{Q}_j$ has been divided by the population-based variance counterpart.

### 3.2.3 Evaluating $f(P^M|\, d^M)$

To evaluate $f(P^M|d^M)$, the last expression in (3.7), replace $P^M$ with the corresponding sufficient statistics. Therefore, $f(P^M|d^M) = f(\bar{Q}_M, W, H|d^M)$. Assume that each $q_j^{(i)}$ (the estimates of $Q$ from each of the synthetic samples) are centered at the corresponding population estimate $Q_j^{(i)}$, with associated variance from $d_j^{(i)}$. Given that $d_j^{(i)}$ has been randomly selected from the $P_j^{(i)}$, the estimates of $Q$ from the synthetic samples should be approximately $Q_j^{(i)}$, but may vary depending on the actual sample drawn from $P_j^{(i)}$. Therefore, the estimates should be distributed as

$$q_j^{(i)}|P_j^{(i)}, u_j^{(i)} \sim N(Q_j^{(i)}, u_j^{(i)}) \tag{3.16}$$

which can be simplified by assuming the variances from each $d_j^{(i)} \in d_j$ are similar. Set $u_j^{(i)} = \bar{u}_j$ for all $i = 1, 2, \ldots, m$ and put a flat prior distribution on the $Q_j^{(i)}$ in (3.16) to find

$$Q_j^{(i)}|q_j^{(i)}, \bar{u}_j \sim N(q_j^{(i)}, \bar{u}_j). \tag{3.17}$$

Average (3.17) over $i$ and $j$, and the distribution of $\bar{Q}_M$ becomes

$$\bar{Q}_M|d^M \sim N(\bar{q}_M, \bar{u}_M/km). \tag{3.18}$$

The distribution of $W_j$ with respect to the synthetic samples derived from $c_j$ is

$$\frac{(m-1)b_j}{\bar{u}_j + W_j}|d_j \sim \chi^2_{m-1}. \tag{3.19}$$

The distribution of $W|d^M$ will not be explicitly found as (3.19) is sufficient for all subsequent derivations and approximations (see the derivation and explanation following (2.17)). The last distribution that needs derivation to complete the evaluation of $f(P^M|d^M)$ is that of $H|d^M$, which can be written as

$$\frac{(k-1)B_k}{\bar{u}_M/m + H}|d^M \sim \chi^2_{k-1}. \tag{3.20}$$

To derive (3.20), a population based estimate of $B_k$, the variance of the averaged estimates from the synthetic samples, is necessary. This variance can be estimated by conditioning on both the original population of $X$ and $Y$ and selected samples, which yields

$$\begin{aligned}\mathbb{V}(\bar{q}_j|X,Y) &= \mathbb{E}\{\mathbb{V}(\bar{q}_j|X,Y,I)|X,Y\} + \mathbb{V}\{\mathbb{E}(\bar{q}_j|X,Y,I)|X,Y\} \\ &= \mathbb{E}\{\bar{u}_j/m|X,Y\} + \mathbb{V}\{\bar{Q}_j|X,Y\} \;=\; \bar{u}_M/m + H.\end{aligned}$$

The resulting distribution will be $\chi^2$ because the sample variance of $\bar{q}_j$ has been divided by its population-based variance.

### 3.2.4 Evaluating $f(Q|d^M)$

To find the posterior distribution of $Q|d^M$, multiply (3.8), (3.13), (3.14), (3.15), (3.18), (3.19) and (3.20), then integrate with respect to $\bar{z}$, the $v_j$, $B$, $\bar{Q}_M$, the $W_j$ and $H$. Writing the components in terms of their conditional distributions gives

$$\begin{aligned}f(Q|d^M) &= \int N(Q;\bar{z},B/k+\bar{v})\,N(\bar{z};\bar{Q}_M,\bar{v}/km) \\ &\times \left(\prod_{j=1}^{k}\text{Inverse}\,\chi^2\,(v_j;m-1,W_j)\right)\text{Inverse}\,\chi^2(W/m+B;k-1,H) \\ &\times N(\bar{Q}_M;\bar{q}_M,\bar{u}_M/km)\left(\prod_{j=1}^{k}\text{Inverse}\,\chi^2\,(\bar{u}_j+W_j;m-1,b_j)\right) \\ &\times \text{Inverse}\,\chi^2\,(\bar{u}_M/m+H;k-1,B_k)\,d\bar{z}\,dv_j\,dB\,d\bar{Q}_M\,dW_j\,dH. \qquad (3.21)\end{aligned}$$

To fully evaluate this integral, numerical integration is required because a closed form expression cannot be explicitly written. An alternative approach would be to use Bayesian principles, such that each parameter is simulated from its conditional distribution and averaging accordingly. However, an approximation is desired that can be implemented by all users without substantial computation. If $m$ is assumed

to be large, expectations and their approximations can be used in the variance expressions for simplification purposes. For example, $E(W_j|d_j) = \frac{m}{m-2}(b_j - \bar{u}_j)$ can be approximated by $E(W_j|d_j) \approx b_j - \bar{u}_j$. Therefore, the estimate of $Q$ can be written as

$$\begin{aligned}
\mathbb{E}(Q|d^M) &= \mathbb{E}\{\mathbb{E}(Q|P^M)|d^M\} = \mathbb{E}\{\mathbb{E}(\mathbb{E}(Q|c^k, B)|P^M)|d^M\} \\
&= \mathbb{E}\{\mathbb{E}(\bar{z}|P^M)|d^M\} = \mathbb{E}\{\bar{Q}_M|d^M\} \\
&= \bar{q}_M.
\end{aligned}$$

Likewise, the associated variance measure becomes

$$\begin{aligned}
\mathbb{V}(Q|d^M) &= \mathbb{E}\{\mathbb{V}(Q|P^M)|d^M\} + \mathbb{V}\{\mathbb{E}(Q|P^M)|d^M\} \\
&= \mathbb{E}\{\mathbb{E}(\mathbb{V}(Q|c^k, B)|P^M) + \mathbb{V}(\mathbb{E}(Q|c^k, B)|P^M)|d^M\} \\
&+ \mathbb{V}\{\mathbb{E}(\mathbb{E}(Q|c^k, B)|P^M)|d^M\} \\
&= \mathbb{E}\{\mathbb{E}(B/k + \bar{v}|P^M) + \mathbb{V}(\bar{z}|P^M)|d^M\} + \mathbb{V}\{\mathbb{E}(\bar{z}|P^M)|d^M\} \\
&= \mathbb{E}\{H/k + W|d^M\} + \mathbb{V}\{\bar{Q}_M|d^M\} \\
&= B_k/k - \bar{u}_M/mk + \bar{b}_M - \bar{u}_M + \bar{u}_M/mk \\
&= \bar{b}_M + B_k/k - \bar{u}_M
\end{aligned}$$

The above estimates and prescribed integration suggest that the posterior distribution of $Q|d^M$ can be approximated by the normal distribution

$$Q|d^M \sim N(\bar{q}_M, B_k/k + \bar{b}_M - \bar{u}_M) \tag{3.22}$$

when $n$, $m$, and $k$ are large; when moderate, the posterior distribution should be a t-distribution with approximately $\nu_V$ degrees of freedom (see Section 3.1 and Appendix B.1).

## 3.3 Simulation Studies

To illustrate the vertically partitioned data sharing formulae and show proof of concept, simulations were conducted using artificial data and the correct posterior distri-

butions for all data generation. The simulations were based on a two-agency sharing strategy and were performed using the software package $R$ (R Development Core Team, 2004).

### 3.3.1 Simulation I: Linear Regression

The multivariate normal model I used to generate the observed data in Section 2.4.1 was utilized again. The use of the multivariate normal distribution creates a known relationship between $X$ and $Y$, which are designated to two agencies $\mathcal{A}$ and $\mathcal{B}$. Therefore, as in the horizontally partitioned data sharing simulations, the observed data set can be directly sampled from the specified distribution, then split into $X$ and $Y$ components. For the purposes of this simulation, the creation of disguisers for $X$ will be skipped. Instead, the simulation will begin with the second agency, $\mathcal{B}$ creating synthetic replicates for $Y$ based only on the true $X$. Strategies for disguiser creation will be addressed in Sections 3.3.3 and 3.3.4.

The process of creating $Y_{syn}^{(j)}$, $j = 1, 2, \ldots, k$, based on the true $X$ follows the steps outlined in Section 2.4.1 (see equations (2.23), (2.24) and (2.25)). Once complete, $\mathcal{B}$ passes $\{Y_{syn}^{(j)}\}$ to $\mathcal{A}$, who then creates $m$ fully synthetic samples of size $n_{syn}$ for each of the $k$ synthetic samples. To create fully synthetic samples for $X$ and $Y_{syn}^{(j)}$, an appropriate posterior predictive distribution for each of the $k$ partially synthetic samples must be derived. Rather than explicitly finding a closed form expression for the posterior predictive distribution, the conditional distributions that comprise it are found based on a multivariate normal model framework. In addition, because $\mathcal{A}$ has only observed a subset of the population of $X$ and synthetic samples for $Y$, the mean and variance of the distribution are assumed to be unknown. Therefore, the following multivariate normal distribution and non-informative Jeffreys prior distribution are

used (see Gelman *et al.* (2004, pp. 86–88)):

$$Y_{syn}^{(j)}|\mu_j, \Sigma_j \sim N(\mu_j, \Sigma_j) \tag{3.23}$$

$$(\mu_j, \Sigma_j) \propto |\Sigma_j|^{(q+1)/2} \tag{3.24}$$

where $\mu_j = X\beta_j$ and $\Sigma_j$ is the covariance matrix of $X$ and $Y_{syn}^{(j)}$. The sampling procedure uses the conditional distributions of $\Sigma_j$ and $\mu_j$, along with the posterior predictive distribution of $X$ and $Y_{syn}^{(j)}$ to create fully synthetic samples. The sampling steps are outlined below:

$$W_{j*}|X, Y_{syn}^{(j)} \sim \text{Wishart}^{-1}(n-1, S_j) \tag{3.25}$$

$$\mu_{j*}|X, Y_{syn}^{(j)}, W_{j*} \sim N(\hat{\mu}_j, W_{j*}/n) \tag{3.26}$$

$$X_{syn,j}^{(i)}, Y_{syn,j}^{(i)}|X, Y_{syn}^{(j)}, W_{j*}, \mu_{j*} \sim N(\mu_{j*}, W_{j*}) \tag{3.27}$$

where $\hat{\mu}_j = (\bar{X}, \bar{Y}_{syn}^{(j)})'$ and $S_j$ is the sum of squares matrix about the sample means for $X$ and $Y_{syn}^{(j)}$.

Next, estimates of $q_j^{(i)}$ and $u_j^{(i)}$ are calculated from the $M = km$ synthetic samples, $d_j^{(i)}$, to find values for (3.1), (3.2), (3.3), (3.4) and (3.5). Due to the presence a negative sign in the variance estimate (see (3.5)), negative estimates can occur for some values of $B_k$, $\bar{b}_M$ and $\bar{u}_M$. In such situations, adjustments are necessary to guarantee all variance estimates are non-negative. Using the formula in Reiter (2002) as a basis, adjusted variance estimates for vertically partitioned shared data are found using

$$T_V^* = \max(0, T_V) + \delta\left(\frac{n_{syn}}{n}\bar{u}_M + \frac{B_k}{k}\right) \tag{3.28}$$

where $\delta = 1$ when $T_V < 0$ and 0 otherwise. Generally, it is possible to avoid negative variance estimates by increasing $n_{syn}$, $k$ or $m$, as indicated by the simulation results in Table 3.1. However, additional research regarding (3.28) is necessary to determine whether it is correctly adjusting negative variances and other properties it may have.

**Table 3.1**: 1,000 simulations combining data from two agencies when $Q$ is the regression coefficient of $Y$ on $X$.

| | $\bar{q}_{obs}$ | $\bar{v}^{\star}_{obs}$ | $\bar{q}_M$ | $\mathbb{V}(\bar{q}_M)^{\star}$ | $T_V(T_V^*)^{\star}$ | % $T_V < 0$ | 95% $\text{CI}_N$ | 95% $\text{CI}_t$ |
|---|---|---|---|---|---|---|---|---|
| $n_{syn} = 1,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 0.5 | 0.75 | 0.5 | 2.11 | 1.74 (3.07) | 38.2% | 90.5% | 99.2% |
| $m = 40$ | 0.5 | 0.75 | 0.5 | 1.33 | 1.32 (1.56) | 14.1% | 89.5% | 100% |
| $k = 30$ | | | | | | | | |
| $m = 4$ | 0.5 | 0.75 | 0.5 | 0.80 | 0.86 (1.25) | 25.4% | 92.8% | 100% |
| $m = 40$ | 0.5 | 0.75 | 0.5 | 0.81 | 0.81 (0.82) | 1.6% | 92.6% | 97.9% |
| $n_{syn} = 10,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 0.5 | 0.75 | 0.5 | 1.41 | 1.37 (1.40) | 2.9% | 88% | 99.4% |
| $m = 40$ | 0.5 | 0.75 | 0.5 | 1.22 | 1.25 | — | 93% | 95.2% |
| $k = 30$ | | | | | | | | |
| $m = 4$ | 0.5 | 0.75 | 0.5 | 0.83 | 0.82 | — | 93.5% | 95.7% |
| $m = 40$ | 0.5 | 0.75 | 0.5 | 0.76 | 0.79 | — | 96.1% | 96.2% |
| $n_{syn} = 20,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 0.5 | 0.75 | 0.5 | 1.48 | 1.39 (1.39) | 0.4% | 89.5% | 97.9% |
| $n_{syn} = 50,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 0.5 | 0.75 | 0.5 | 1.27 | 1.35 | — | 91.3% | 96.2% |

$\star$ multiply by $10^{-4}$

The two agencies in the simulation each had observed data sets with 10,000 observations because I previously had assumed that one-to-one matching exists between data sets. Several synthetic data set sizes were considered with $n_{syn}$ ranging from 1,000 to 50,000. In addition, I used multiple combinations of $k$ and $m$, with values ranging from small to large.

The simulation results in Table 3.1 suggest that estimates of $Q$ based on the fully synthetic data sets are unbiased for the true observed value for the population quantity $Q$. To verify that (3.5) is an appropriate measure of the variance associated

with $\bar{q}_M$, the estimates of $T_V$ are compared to those for $\mathbb{V}(\bar{q}_M)$ (see Table 3.1). In most cases, both values are tracking similar amounts of variability within the synthetic data sets and are always larger than the variability for the observed data.

The simulation results also give insight as to how agencies should choose values for $k$, $m$ and $n_{syn}$. Starting with $n_{syn}$, it is apparent that whenever $n_{syn}$ is smaller than the observed data size $n$, the percentage of negative variances is large, except for large values of $k$ and $m$. When equal, negative variances are greatly reduced and even nonexistent; similar results are true whenever $n_{syn}$ is much larger than $n$. The coverage of the 95% confidence intervals, as calculated with the normal distribution and compared to the true value, increases as $k$ and $m$ increase. In some cases, the coverage is close to the desired rate of 95%. There are some situations where the use of a t-distribution may yield more desirable levels of coverage. However, due to assumptions made during derivation and calculation, coverage is inflated when $n_{syn}$ is small. (Additional research regarding the degrees of freedom used in the t-distribution may result in better coverage if a derivation can be found that does not depend on underlying assumptions made throughout derivation and calculation.) The relationship between the choices of $k$ and $m$ may be attributed to the overall structure of the data sharing mechanism, in particular how the partially synthetic method creates $c^k$ to be the same size as the observed data, $\mathcal{D}$. Thus, results based on fully synthetic data sets the same size as $c^k$ and $\mathcal{D}$ or larger may be more optimal than those based on a smaller sample size.

Other attributes of the output are also affected by the choice of values for $k$, $m$ and $n_{syn}$. Depending on which aspect is most valued — reducing negative variance estimates or increasing coverage — choices may vary. If decreasing the occurrence of negative variances is valued, then $m$ and $n_{syn}$ should be large. If the coverage of the 95% confidence intervals is a high priority, then an increase of $k$ and using a value

of $n_{syn}$ equal to or greater than $n$ will give intervals that are close to the desired coverage levels.

### 3.3.2 Simulation II: Logistic Regression

I next consider a logistic regression scenario to illustrate other models and data types where the sharing of vertically partitioned data can occur. As in Section 3.3.1, infinite populations are assumed to exist for both $X$ and $Y$, where $X$ has an underlying normal distribution with mean zero and variance 10 and $Y$ is a $0-1$ response indicator. Let $X$ be designated to agency $\mathcal{A}$, so that it can be directly sampled from a normal distribution, and $Y$ designated to agency $\mathcal{B}$, such that

$$Y|X,\beta \sim \text{Bernoulli}(\pi) \tag{3.29}$$

$$\text{logit}(\pi)|X = \beta_0 + \beta_1 X + \beta_2 X^2 \tag{3.30}$$

where $\beta_0 = 3$, $\beta_1 = 4$ and $\beta_2 = 1$. The choice of a quadratic link function and values for $\beta$ are arbitrary; any number of link functions and values could have been chosen. However, this link illustrates a model with more than one variable without the inclusion of interaction terms.

As in Section 3.3.1, the creation of disguisers for $X$ is skipped. Therefore, I proceed directly to the imputation of $Y_{syn}^{(j)}$ for $j = 1, 2, \ldots, k$, based on the true $X$, using the method described in (2.30), (2.31) and (2.32). Once the generation of $\{Y_{syn}^{(j)}\}$ is complete, the set of synthetic data is passed to $\mathcal{A}$ who next creates fully synthetic data sets that can be passed back to $\mathcal{B}$ and released for public use.

To create the fully synthetic data sets, $\mathcal{A}$ must first create synthetic samples for $X$, which will be used to generate the associated values for $Y_{syn}^{(j)}$. Since the population of $X$ is unknown, a fully Bayesian approach for generating synthetic samples must

be implemented (Gelman *et al.*, 2004, pp. 74–77):

$$\sigma^2_{X,j*}|X \sim \frac{(n-1)S_X}{\chi^2_{n-1}} \tag{3.31}$$

$$\mu_{X,j*}|X,\sigma^2_{X,j*} \sim N(\bar{X},\sigma^2_{X,j*}/n) \tag{3.32}$$

$$X^{(i)}_{syn,j}|X,\sigma^2_{X,j*},\mu_{X,j*} \sim N(\mu_{X,j*},\sigma^2_{X,j*}) \tag{3.33}$$

where $(n-1)S_X$ is the sum of squares of $X$ about its mean. The sampling process is repeated $M = mk$ times, $m$ times for each of the $k$ $Y^{(j)}_{syn}$. Once complete, $\mathcal{A}$ uses the relationship between $X$ and each $Y^{(j)}_{syn}$, along with the newly generated $X^{(i)}_{syn,j}$ to create $Y^{(i)}_{syn,j}$. This process follows the procedure as detailed in Section 2.4.2 (see (2.30), (2.31) and (2.32)).[3]

Upon completion of the fully synthetic data set, $q^{(i)}_j$ and $u^{(i)}_j$ are calculated from each of the $M = km$ $d^{(i)}_j$ and then used to find estimates for (3.1) to (3.5). As in previous simulations, the variance estimates may require adjustment due to the possibility of negative variances; the same adjustment measure (see (3.28)) was used once again.

Simulation results for both coefficients ($\beta_1$ and $\beta_2$) are given in Tables 3.2 and 3.3. As expected, the percentage of negative variances decreases with increases of $n_{syn}$, $k$, and $m$. Similar to the results in Section 3.3.1, once $n_{syn} \geq n$, the occurrence of negative variances is greatly reduced and in most cases nonexistent.

Coverage of the 95% confidence intervals, as calculated with the normal approximation for the true coefficient values, increases as $n_{syn}$, $k$, and $m$ increase. Even in the case where $n_{syn}$ is small, but $k$ and $m$ are large, coverage is as desired. In situations where the coverage is lower than desired, a t-distribution can be used instead of the normal approximation. Based on the values in Tables 3.2 and 3.3, confidence

[3]This generation method is essentially the same as that used to impute $\{Y^{(j)}_{syn}\}$ from $X$ in the beginning stages of the sharing process.

**Table 3.2**: 1,000 simulations combining data from two agencies when $Q$ is the logistic regression coefficient ($\beta_1$) of $Y$ on $X$.

| | $\bar{q}_{obs}$ | $\bar{v}_{obs}$ | $\bar{q}_M$ | $\mathbb{V}(\bar{q}_M)$ | $T_V(T_V^*)$ | % $T_V < 0$ | 95% $\text{CI}_N$ | 95% $\text{CI}_t$ |
|---|---|---|---|---|---|---|---|---|
| $n_{syn} = 1,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 4 | 0.0136 | 4.10 | 0.0397 | 0.0486 (0.0722) | 36.4% | 90.2% | 99.4% |
| $m = 40$ | 4 | 0.0136 | 4.10 | 0.0249 | 0.0345 (0.0374) | 9.8% | 90.7% | 99.8% |
| $k = 10$ | | | | | | | | |
| $m = 4$ | 4 | 0.0136 | 4.11 | 0.0220 | 0.0328 (0.0459) | 27.9% | 91.8% | 99.3% |
| $m = 40$ | 4.01 | 0.0136 | 4.11 | 0.0180 | 0.0273 (0.0278) | 2.5% | 91.1% | 99.2% |
| $k = 30$ | | | | | | | | |
| $m = 4$ | 4.01 | 0.0136 | 4.11 | 0.0178 | 0.0271 (0.0329) | 18.4% | 89.8% | 99.6% |
| $m = 40$ | 4.01 | 0.0136 | 4.10 | 0.0158 | 0.0248 | — | 94.9% | 97.1% |
| $n_{syn} = 10,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 4.01 | 0.0137 | 4.02 | 0.0265 | 0.0257 (0.0262) | 2.8% | 88.1% | 99.8% |
| $m = 40$ | 4.01 | 0.0137 | 4.03 | 0.0263 | 0.0233 | — | 92.9% | 95% |
| $k = 10$ | | | | | | | | |
| $m = 4$ | 4.01 | 0.0136 | 4.02 | 0.0162 | 0.0174 (0.0174) | 0.1% | 92.7% | 97% |
| $n_{syn} = 20,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 4.01 | 0.0136 | 4.02 | 0.0253 | 0.0247 (0.0247) | 0.2% | 88.3% | 97.9% |
| $k = 10$ | | | | | | | | |
| $m = 4$ | 4.01 | 0.0136 | 4.02 | 0.0176 | 0.0171 | — | 93.6% | 95.2% |
| $n_{syn} = 50,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 4.01 | 0.0137 | 4.02 | 0.0240 | 0.0241 | — | 91.3% | 95.7% |
| $k = 10$ | | | | | | | | |
| $m = 4$ | 4.02 | 0.0137 | 4.03 | 0.0174 | 0.0171 | — | 93.9% | 95% |

interval coverages, using a t-distribution, will be overestimated whenever $n_{syn}$, $k$ and $m$ are small. This can be directly attributed to the assumptions made during derivation and calculation of the approximate degrees of freedom.

### 3.3.3 Simulation III: Disguising $X$

To investigate ways for creating good disguisers of $X$, two different generation methods are considered. The first method utilizes groupings or clusters of $Y$, which re-

**Table 3.3**: 1,000 simulations combining data from two agencies when $Q$ is the logistic regression coefficient ($\beta_2$) of $Y$ on $X$.

| | $\bar{q}_{obs}$ | $\bar{v}_{obs}$ | $\bar{q}_M$ | $\mathbb{V}(\bar{q}_M)$ | $T_V(T_V^*)$ | % $T_V < 0$ | 95% $\text{CI}_N$ | 95% $\text{CI}_t$ |
|---|---|---|---|---|---|---|---|---|
| $n_{syn} = 1,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 1 | 0.0009 | 1.02 | 0.0025 | 0.0030 (0.0045) | 36.2% | 91.4% | 99.3% |
| $m = 40$ | 1 | 0.0009 | 1.03 | 0.0016 | 0.0022 (0.0024) | 9.6% | 89.2% | 99.9% |
| $k = 10$ | | | | | | | | |
| $m = 4$ | 1 | 0.0009 | 1.03 | 0.0014 | 0.0021 (0.0029) | 27.6% | 91.1% | 99% |
| $m = 40$ | 1 | 0.0009 | 1.03 | 0.0011 | 0.0017 (0.0018) | 2.9% | 91.7% | 98.7% |
| $k = 30$ | | | | | | | | |
| $m = 4$ | 1 | 0.0009 | 1.03 | 0.0011 | 0.0017 (0.0021) | 18% | 88.9% | 99.9% |
| $m = 40$ | 1 | 0.0009 | 1.03 | 0.0010 | 0.0016 | — | 95% | 97.3$ |
| $n_{syn} = 10,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 1 | 0.0009 | 1.01 | 0.0017 | 0.0016 (0.0017) | 2.5% | 87.3% | 99.7% |
| $m = 40$ | 1 | 0.0009 | 1.01 | 0.0016 | 0.0015 | — | 92.6% | 94.6% |
| $k = 10$ | | | | | | | | |
| $m = 4$ | 1 | 0.0009 | 1.01 | 0.0010 | 0.0011 (0.0011) | 0.1% | 93% | 97.1% |
| $n_{syn} = 20,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 1 | 0.0009 | 1.01 | 0.0016 | 0.0016 (0.0016) | 0.3% | 89.4% | 98.5% |
| $k = 10$ | | | | | | | | |
| $m = 4$ | 1 | 0.0009 | 1.01 | 0.0011 | 0.0011 | — | 93.8% | 96% |
| $n_{syn} = 50,000$ | | | | | | | | |
| $k = 3$ | | | | | | | | |
| $m = 4$ | 1 | 0.0009 | 1.01 | 0.0015 | 0.0015 (0.0015) | 0.1% | 92.3% | 96.3% |
| $k = 10$ | | | | | | | | |
| $m = 4$ | 1 | 0.0009 | 1.01 | 0.0011 | 0.0011 | — | 94.2% | 95% |

quires the involvement of the agency owning $Y$. I then implement several disguising routines based on these groupings. The second disguising method adds directional random noise to the original $X$, of which several noise levels are considered. In each case, inferences from the disguisers are compared to the true values based on their ability to mask $X$ while maintaining the relationships of $X$ and $Y$. The $X$ and $Y$ used throughout this simulation are generated from a multivariate normal distribution with means equal to zero, variances of 10 and a common correlation of 0.5; a single data set with 10,000 observations was created and utilized throughout.

**Disguising $X$ using groups based on Mahalanobis Distances**

The first disguising method uses a grouping scheme to cluster similar observations together based on their Mahalanobis distances (Weisberg, 1985, p. 112) to the overall mean. (This initialization step must be completed by the owner of $Y$ before the data sharing process can begin.) The groupings or clusters give the agency owning $X$ insight as to which units of $Y$ are similar, a coarse or broad view of its distribution. This step brings balance to the data sharing process, while giving the initiating agency the ability to create the disguisers required for the next set of data exchanges.

To create the groupings, the Mahalanobis distances are calculated for each row, $y_i$ in the data matrix $Y$, $i = 1, 2, \ldots, n$, then compared to the $p \times 1$ vector of sample averages, one for each column in $Y$ using

$$\text{MD}_i = \left[(y_i - \bar{y})' \mathcal{C}^{-1} (y_i - \bar{y})\right]^{\frac{1}{2}} \tag{3.34}$$

where $\mathcal{C} = (\mathcal{Y}'\mathcal{Y})/n - 1$ and $\mathcal{Y}$ is the centered matrix of $Y$. In cases where $Y$ is univariate or uncorrelated, the calculated Mahalanobis distance will be equivalent to the Euclidean distance. However, when the data are correlated, it is desirable to account for the relationships within the data. (Since the $Y$ used in this simulation is univariate, either distance measure will give the same results.) For some cases, the calculation of $\mathcal{C}^{-1}$ in (3.34) may be problematic. In particular, whenever $p$ is large it is likely that some variables may contain redundant information leading to multicollinearity and a singular or near singular matrix that cannot be inverted.

Once the distances are calculated, observations in $Y$ are grouped according to their distances or by a predetermined group size. For the purposes of this simulation, values were first sorted and then grouped into clusters of three observations. The "rule of three" (Willenborg and de Waal, 2001, p. 53) was chosen because it is the minimum group size that still provides protection of the observations. Other group

sizes could be implemented depending on the level of acceptable risk. However, in this simulation, disguisers will only be created with groups of size three.

After creating the groupings, the owner of $Y$ passes them to the owner of $X$, who uses the groupings to create $r$ disguisers of $X$. Five methods of disguiser creation are implemented with the groupings of $Y$, yielding results with various levels of success. To check the quality of the disguisers, the output from a linear regression model with each $X_{dis}^{(l)}$ $(l = 1, 2, \ldots, r)$ and the real $Y$ are compared to the regression model results using the real $X$ and $Y$ (see Table 3.4).

The first set of disguisers were created by re-sampling the observations of $X$ with replacement from those within a particular grouping. This sampling scheme makes it is possible for observations to appear more than once within a grouping. The next method was similar, except sampling was done without replacement, essentially rearranging the ordering of the observations of $X$ within a particular grouping. The third method only re-sampled the values for those observations in $X$ that exceeded some specified threshold. The threshold was set at $\pm 5$, which selected about 12% of the observations in $X$. The high-risk values were re-sampled from all possible values within the group, including those at high risk. The fourth method was similar to the third, except high-risk values were only re-sampled from the low-risk values. The last method randomly swapped a high-and low-risk value within a grouping.

Linear regression model results for the disguisers and true values are given in Table 3.4. It is evident from these values that re-sampling or reordering the values within a group does not adequately preserve the relationship between $X$ and $Y$. The last section of the table combines the results from multiple disguising methods with the original data results to show their disguising abilities. It is obvious that each of these methods reduces the relationship between $X$ and $Y$. Therefore, if the disguisers

**Table 3.4**: Linear regression comparison of the disguising methods for $X$ with $Y$, using a Mahalanobis distance based clustering method.

| | Intercept | SE | t-value | $P(>\|t\|)$ | Slope | SE | t-value | $P(>\|t\|)$ |
|---|---|---|---|---|---|---|---|---|
| *Truth* | 0.023 | 0.027 | 0.885 | 0.376 | 0.492 | 0.008 | 58.2 | *** |
| *Sample values from those in the cluster (with replacement)* | | | | | | | | |
| | 0.008 | 0.031 | 0.270 | 0.787 | 0.175 | 0.010 | 18.2 | *** |
| | 0.009 | 0.031 | 0.305 | 0.760 | 0.159 | 0.010 | 16.3 | *** |
| | 0.004 | 0.031 | 0.123 | 0.902 | 0.174 | 0.010 | 18.1 | *** |
| *Sample values from those in the cluster (without replacement)* | | | | | | | | |
| | 0.005 | 0.031 | 0.167 | 0.867 | 0.151 | 0.010 | 15.6 | *** |
| | 0.007 | 0.031 | 0.239 | 0.811 | 0.190 | 0.010 | 19.8 | *** |
| | 0.007 | 0.031 | 0.216 | 0.829 | 0.178 | 0.010 | 18.5 | *** |
| *Sample high risk values from those in the cluster (with replacement of high risk)* | | | | | | | | |
| | 0.005 | 0.029 | 0.184 | 0.854 | 0.434 | 0.010 | 42.7 | *** |
| | 0.015 | 0.029 | 0.530 | 0.596 | 0.431 | 0.010 | 42.2 | *** |
| | 0.013 | 0.029 | 0.447 | 0.655 | 0.429 | 0.010 | 42.3 | *** |
| *Sample high risk values from those in the cluster (without replacement of high risk)* | | | | | | | | |
| | 0.009 | 0.030 | 0.310 | 0.757 | 0.382 | 0.011 | 33.8 | *** |
| | 0.005 | 0.030 | 0.162 | 0.871 | 0.393 | 0.011 | 34.7 | *** |
| | 0.010 | 0.030 | 0.348 | 0.728 | 0.379 | 0.011 | 33.5 | *** |
| *Swap high risk values with another in the cluster* | | | | | | | | |
| | 0.006 | 0.030 | 0.211 | 0.833 | 0.217 | 0.010 | 22.7 | *** |
| | 0.007 | 0.031 | 0.240 | 0.810 | 0.205 | 0.010 | 21.5 | *** |
| | 0.005 | 0.031 | 0.148 | 0.883 | 0.194 | 0.010 | 20.2 | *** |
| *Combination of methods (includes real data)* | | | | | | | | |
| | 0.005 | 0.029 | 0.184 | 0.854 | 0.434 | 0.010 | 42.7 | *** |
| | 0.009 | 0.030 | 0.310 | 0.757 | 0.382 | 0.011 | 33.8 | *** |
| | 0.023 | 0.027 | 0.885 | 0.376 | 0.492 | 0.008 | 58.2 | *** |

*** $< 0.0001$

were to be passed to the agency owning $Y$, the agency would need only to select the disguiser with the strongest relationship to their data ($Y$) and they would know which of the disguisers is the true $X$ (assuming $X$ was present within the set).

Of the five methods implemented, only the third, re-sampling of the high-risk values, shows potential as a disguising method. To further investigate this methods disguising potential see the simulation in Section 3.3.4.

### Disguising $X$ using directional random noise

The next disguising method uses the addition of directional random noise to the original $X$ as a means of disguising the true values. The addition of random noise from the SDC context (see Section 1.1.1) reduces the relationship between $X_{dis}^{(l)}$ and $Y$ in a manner similar to that seen in the previous attempt at disguiser creation. To effectively hide $X$ within a set of disguisers, it is necessary to both reduce this relationship, but also increase it in such a way that $X$ cannot be easily determined. (The true $X$ had the strongest relationship in previous attempt at disguiser creation). To create disguisers with the ability to both increase and decrease the relationship between $X$ and $Y$, I add directional random noise. For instance, to create the values $x_{dis}^{(l)}(i)$ $(i = 1, 2, \ldots, n)$ that comprise $X_{dis}^{(l)}$, implement the following:

$$x_{dis}^{(l)}(i) \quad = \quad x(i) + \begin{cases} e_i^+ & \text{if } x(i) < \bar{x} \\ e_i^- & \text{if } x(i) > \bar{x} \end{cases} \tag{3.35}$$

where $e_i^+$ represents a random draw from the positive portion of $N(0, s^2)$, $e_i^-$ a random draw from the negative portion of $N(0, s^2)$ and $x(i)$ is the $i^{th}$ unit in $X$. Therefore, noise is subtracted for those values greater than the sample average and added to those less than the sample average. The addition of noise in this manner enables the disguiser to either strengthen or weaken the relationship between $X$ and $Y$.

**Table 3.5**: Linear regression comparison of the disguising methods for $X$ with $Y$, using directional random noise.

| | Intercept | SE | t-value | $P(>\|t\|)$ | Slope | SE | t-value | $P(>\|t\|)$ |
|---|---|---|---|---|---|---|---|---|
| *Truth* | 0.023 | 0.027 | 0.885 | 0.376 | 0.492 | 0.008 | 58.2 | *** |
| *Add directional noise* $N(0, 1.5^2)$ | | | | | | | | |
| | 0.022 | 0.028 | 0.775 | 0.439 | 0.547 | 0.011 | 49.3 | *** |
| | 0.017 | 0.028 | 0.608 | 0.543 | 0.553 | 0.011 | 49.5 | *** |
| | 0.016 | 0.028 | 0.589 | 0.556 | 0.555 | 0.011 | 49.9 | *** |
| *Add directional noise* $N(0, 1.75^2)$ | | | | | | | | |
| | 0.012 | 0.028 | 0.411 | 0.681 | 0.533 | 0.011 | 46.5 | *** |
| | 0.023 | 0.028 | 0.817 | 0.414 | 0.525 | 0.011 | 46.1 | *** |
| | 0.019 | 0.028 | 0.669 | 0.504 | 0.529 | 0.011 | 46.3 | *** |
| *Add directional noise* $N(0, 2^2)$ | | | | | | | | |
| | 0.004 | 0.029 | 0.147 | 0.883 | 0.510 | 0.012 | 44.0 | *** |
| | 0.014 | 0.029 | 0.489 | 0.625 | 0.492 | 0.012 | 42.0 | *** |
| | 0.010 | 0.029 | 0.348 | 0.728 | 0.489 | 0.012 | 41.6 | *** |
| *Add directional noise* $N(0, 2.25^2)$ | | | | | | | | |
| | 0.014 | 0.029 | 0.480 | 0.631 | 0.447 | 0.012 | 37.9 | *** |
| | 0.006 | 0.029 | 0.195 | 0.846 | 0.459 | 0.012 | 38.8 | *** |
| | 0.005 | 0.029 | 0.174 | 0.862 | 0.452 | 0.012 | 38.4 | *** |
| *Combination of methods* | | | | | | | | |
| | 0.022 | 0.028 | 0.775 | 0.439 | 0.547 | 0.011 | 49.3 | *** |
| | 0.019 | 0.028 | 0.669 | 0.504 | 0.529 | 0.011 | 46.3 | *** |
| | 0.023 | 0.027 | 0.885 | 0.376 | 0.492 | 0.008 | 58.2 | *** |
| | 0.006 | 0.029 | 0.195 | 0.846 | 0.459 | 0.012 | 38.8 | *** |

*** $< 0.0001$

The process was implemented on the same $X$ used in the previous disguising simulation and utilizes several different values for $s$. To check the disguisers ability to model $X$, linear regression models were fit to each of the $X_{dis}^{(l)}$ and $Y$, then compared to the truth as determined by $X$ and $Y$. Table 3.5 gives the linear regression results using four variations or levels of random noise ($s$). It is evident from this table, that certain values of $s$ result in regression output that is close to the truth, while others

do not. However, these results suggest that by adding directional random noise to $X$, disguisers can be created with the ability to both increase and decrease the relationship present between $X$ and $Y$, and effectively hide $X$.

A comparison of several disguisers are given in the last section of Table 3.5, suggesting that the truth can be hidden within disguisers without being apparently obvious. Although this method yields promising results, it can only be applied to those variables that have normal or approximately normal distributions.

### 3.3.4 Simulation IV: A Sharing Alternative

Briefly mentioned in Section 3.1 was a reference to an alternative type of disguiser exchange and variation of the data sharing method detailed throughout this chapter. This alternative and data sharing variation will now be considered and implemented with a simulation.

Suppose $\mathcal{A}$ can create a set of $r$ disguisers that accurately model $X$ without disclosing sensitive components. If so, $\mathcal{A}$ can simply pass $\{X_{dis}^{(l)}\}$ to $\mathcal{B}$ without randomly including the true $X$ within the set. In the context of the first disguising method presented in Section 3.3.3, if $\mathcal{A}$ can create good disguisers of $X$, then they have essentially generated disguisers from the distribution of $X_{dis}^{(l)}|X, Y_{grouped}$, if the groupings of $Y$ reasonably model the distribution or similar values of $Y$. Or in terms of the directional noise method, $\mathcal{A}$ has generated disguisers from the distribution of $X_{dis}^{(l)}|X$. Therefore, once $\mathcal{A}$ passes $\{X_{dis}^{(l)}\}$ to $\mathcal{B}$, they now have partially synthetic data sets $c_l = (X_{dis}^{(l)}, Y)$, such that the set $c^r = \{c_l, l = 1, 2, \ldots, r\}$ can be used to generate fully synthetic data sets.

Using the disguising methods from Section 3.3.3 (groupings based on the Mahalanobis Distance and directional random noise), I conducted a simulation with two agencies each having observed data sets of size 10,000. The sharing process is similar

**Table 3.6**: 1,000 simulations combining data from two agencies when $Q$ is the regression coefficient of $Y$ on $X$.

| | $\bar{q}_{obs}$ | $\bar{v}^{\star}_{obs}$ | $\bar{q}_M$ | $\mathbb{V}(\bar{q}_M)^{\star}$ | $T_V(T_V^*)^{\star}$ | % $T_V < 0$ | 95% $\text{CI}_N$ | 95% $\text{CI}_t$ |
|---|---|---|---|---|---|---|---|---|
| $n_{syn} = 1,000$ | | | | | | | | |
| $k = 3$, $m = 4$ | | | | | | | | |
| *High-risk*[1] | 0.5 | 0.75 | 0.43 | 6.07 | 5.41 (6.62) | 21.4% | 26.2% | 91.9% |
| *High-risk*[2] | 0.5 | 0.75 | 0.44 | 3.99 | 4.25 (5.87) | 28.3% | 22.7% | 88.1% |
| *Add noise*[3] | 0.5 | 0.75 | 0.52 | 5.40 | 5.90 (7.83) | 27.3% | 80.4% | 97.5% |
| *Add noise*[4] | 0.5 | 0.75 | 0.50 | 4.52 | 4.71 (6.61) | 28.3% | 89.9% | 99.1% |
| $k = 10$, $m = 4$ | | | | | | | | |
| *High-risk*[1] | 0.5 | 0.75 | 0.43 | 2.34 | 2.52 (3.32) | 21.1% | 3.9% | 76.3% |
| *High-risk*[2] | 0.5 | 0.75 | 0.44 | 1.83 | 2.07 (3.06) | 27.6% | 5.1% | 79.9% |
| *Add noise*[3] | 0.5 | 0.75 | 0.52 | 2.43 | 2.60 (3.81) | 25.6% | 80.6% | 98.7% |
| *Add noise*[4] | 0.5 | 0.75 | 0.50 | 2.22 | 2.51 (3.79) | 28.6% | 92.6% | 99.8% |
| $n_{syn} = 10,000$ | | | | | | | | |
| $k = 3$, $m = 4$ | | | | | | | | |
| *High-risk*[1] | 0.5 | 0.75 | 0.43 | 4.67 | 5.21 (5.22) | 0.5% | 13.1% | 64.1% |
| *High-risk*[2] | 0.5 | 0.75 | 0.44 | 3.19 | 3.35 (3.37) | 1.3% | 5.1% | 42.4% |
| *Add noise*[3] | 0.5 | 0.75 | 0.52 | 4.61 | 4.97 (4.99) | 0.9% | 80% | 94.5% |
| *Add noise*[4] | 0.5 | 0.75 | 0.50 | 3.42 | 3.76 (3.81) | 2.2% | 87.9% | 98.9% |
| $k = 10$, $m = 4$ | | | | | | | | |
| *High-risk*[1] | 0.5 | 0.75 | 0.43 | 2.15 | 2.29 | — | 0.2% | 0.5% |
| *High-risk*[2] | 0.5 | 0.75 | 0.44 | 1.64 | 1.73 | — | — | 0.4% |
| *Add noise*[3] | 0.5 | 0.75 | 0.52 | 2.14 | 2.47 | — | 78% | 84% |
| *Add noise*[4] | 0.5 | 0.75 | 0.50 | 1.67 | 2.14 | — | 94.5% | 97.1% |

$\star$ multiply by $10^{-4}$
[1] High-risk threshold was randomly drawn from the set {4, 5, 6}.
[2] High-risk threshold was randomly drawn from a Uniform(4, 6).
[3] Level of random noise ($s$) was randomly drawn from a Uniform(1.5, 2.25).
[4] Level of random noise ($s$) was randomly drawn from a Uniform(1.75, 2.25).

to that of Section 3.3.1, except that $\mathcal{B}$ creates the fully synthetic data sets and the roles of $X$ and $Y$ are reversed in equations (3.25), (3.26) and (3.27). Two values of $n_{syn}$ were implemented with a small number of combinations of $k$ (number of disguisers in this scenario) and $m$.

The simulation results are given in Table 3.6 and it is evident that the addition of directional random noise does better than re-sampling high-risk values. In both

implementations of high-risk re-sampling, the estimates of $Q$ are biased, while those based on directional random noise were closer to the desired value. As with the simulation in Section 3.3.1, negative variance values were adjusted using (3.28), the occurrence of which was almost non-existent when $n_{syn} = 10,000$. The coverage of the 95% confidence intervals for the high-risk re-sampling simulations are quite low when calculated using the normal approximation, and become increasingly worse as $n_{syn}$ increases. Comparatively, 95% confidence interval coverage reasonable (although still on the low end) for the directional random noise simulations. Based on these results, a combination of random noise levels and re-sampling of high-risk values may create a set of disguiser that accurately model $X$, while maintaining its relationship with $Y$.

# Chapter 4

# Empirical Investigations of the Data Sharing Approaches

To apply the data sharing methods described in Chapters 2 and 3 on a real data set requires a complete understanding of the variables in the data and knowledge of their interactions and dependencies. For some data sets, minimal effort may be sufficient to attain the knowledge necessary to create disguisers and generate synthetic replicates as prescribed by the data sharing methods. However, for data sets that have irregular or highly skewed distributions, this could be a complex process that utilizes multiple methods or combinations of methods and in the end may still yield synthetic data sets that do not model the original data to the desired accuracy.

The real data I used to investigate the data sharing approaches is similar to that described in the latter scenario. The data set contained many highly skewed variables that were also dependent on other highly skewed variables. In addition, several of the variables in the data set were bounded between 0% and 100% (comprised of percentages) which created very irregular distributions.

Due to the complex structures in the data, many challenges were presented throughout the modeling and inference gathering stages that complicated the generation of disguisers and synthetic data used in the sharing procedures. These challenges

resulted in much experimentation with parametric and non-parametric models, which in the end still produced suboptimal results. The results included in this chapter are provided to illustrate the difficulties in implementing the data sharing approaches on real data. The deficiencies in the obtained inferences can be directly linked to the data generation models, not the theory developed in Chapters 2 and 3.

The data I used were from the 1995 Commercial Building Energy Consumption Survey (CBECS), which is conducted quadrennially by the Energy Information Administration of the U.S. Department of Energy and is a national-level sample survey of commercial buildings and their energy supplies[1]. The data used in all investigations pertaining to the data sharing methods were selected from the public-use data file. The data file initially consisted of 6,590 sampled buildings from which 5,766 were successfully interviewed. From those interviewed, a set of 5,655 observations and 13 variables (as used in Dandekar *et al.* (2002)) is used as the basis for the data sharing investigations. Prior to implementation of the sharing methods, I removed another 65 observations from the data set due to the presence of inapplicabilities in one of the variables (see `NWKER` in Table 4.1) leaving 5,590 observations for analysis.

Eleven continuous and two categorical variables are used from the 1995 CBECS public-use data; they are described in Table 4.1 (along with their associated ranges). The continuous variables are electricity consumption (`ELBTU`) and expenditures (`ELEXP`), natural gas consumption (`NGBTU`) and expenditures (`NGEXP`), major fuel consumption (`MFBTU`) and expenditures (`MFEXP`), total floor space (`SQFT`), number of employees (`NWKER`), and percentage of floor space cooled (`COOLP`), heated (`HEATP`), and lit (`LTOHRP`). The categorical variables are year constructed (`YRCON`) and principal building activity (`PBA`). The buildings represented in the CBECS data are all defined as being enclosed roofed and walled structures used predominately for commercial pur-

[1]See `http://www.eia.doe.gov/emeu/cbecs/` for more information.

**Table 4.1**: The 13 variables selected from the 1995 Commercial Building Energy Consumption Survey (CBECS) public-use data and used for analysis.

| Variable | Label | Range |
|---|---|---|
| Electricity Consumption | ELBTU | 0 – 624,203,730 |
| Electricity Expenditures | ELEXP | 0 – 9,573,552 |
| Natural Gas Consumption | NGBTU | 0 – 374,907,582 |
| Natural Gas Expenditures | NGEXP | 0 – 1,042,355 |
| Major Fuel Consumption | MFBTU | 0 – 1,072,285,895 |
| Major Fuel Expenditures | MFEXP | 0 – 10,086,341 |
| Principal Building Activity | PBA | 20 categories, coded 1 – 91 |
| Year Constructed | YRCON | 9 categories, coded 1 – 9 |
| Total Floor Space | SQFT | 1,001 – 1,600,000 |
| Number of Employees | NWKER | 0 – 16,750; 65 were inapplicable |
| Percent of Floor Space Cooled | COOLP | 0 – 100 |
| Percent of Floor Space Heated | HEATP | 0 – 100 |
| Percent of Floor Space Lit | LTOHRP | 0 – 100 |

poses, each having at least 1,001 square feet of total floor space.

## 4.1 Generating Disguisers for Real Data

Before disguisers can be created, the variables in CBECS must be divided into separate components representing the data owned by two separate agencies. Since the CBECS data contain information on building characteristics along with the energy consumption and expenditure levels, a logical division would be to let $\mathbf{X}$ = (PBA, YRCON, SQFT, NWKER, COOLP, HEATP, LTOHRP) and $\mathbf{Y}$ = (ELBTU, ELEXP, NGBTU, NGEXP, MFBTU, MFEXP). If $\mathbf{X}$ can be shared without worries of confidentiality breaches, then disguisers do not need to be created and the sharing process can proceed as described in Chapter 3. However, if concerns exists due to the sensitive nature of $\mathbf{X}$, then disguisers must be generated. This will be the assumed situation.

As discussed in Chapter 3 (see Sections 3.3.3 and 3.3.4), disguisers for $\mathbf{X}$ must accurately model the structure and relationships of the data, while protecting the

sensitive values. Although this may seem like a simple task, disguiser creation is a process of trial and error that involves much experimentation. The creation of good disguisers requires the exploration of the variables to be disguised, specifically their distributions and relationships with other variables. For most real data sets, the methods or combination of methods used to create disguisers will be data dependent; therefore, an in-depth understanding of the data is a necessity. In terms of the CBECS data, most of the disguising methods I explored for $\mathbf{X}$ were based on the general techniques discussed in Sections 3.3.3 and 3.3.4, but modified specifically for the variables being disguised. The process of disguiser creation can become lengthy because it is necessary to check each set of new disguisers to see if they are preserving the overall structure of the data without revealing too much information. Otherwise, modifications to the methods must be made and new disguisers created before the data can be passed to the other agency participating in the data sharing. This was the case for the CBECS data: the investigation of disguising methods required much experimentation and multiple rounds of modifications before the disguisers modeled the data as desired.

Prior to the creation of disguisers, I formed groupings or clusters of $\mathbf{Y}$ using the method described in Section 3.3.3. Before grouping the data, I applied a square root transformation to all of the variables in $\mathbf{Y}$ due to their extremely right skewed distributions. The transformations did not eliminate the right skewness, but did reduce the scale of the distributions, especially the size of the tail values (see Figures 4.1 and 4.2 for a comparison of `MFBTU` and $\texttt{MFBTU}^{1/2}$). Next, I calculated the Mahalanobis distances of the transformed $\mathbf{Y}$ using (3.34), which became the basis for the groupings. (This process was used to help create groups or clusters that had similar units to one another (Jain *et al.*, 1999).) Groupings with three observations (Willenborg and de Waal, 2001, p. 53) were created based on a sequential progression through

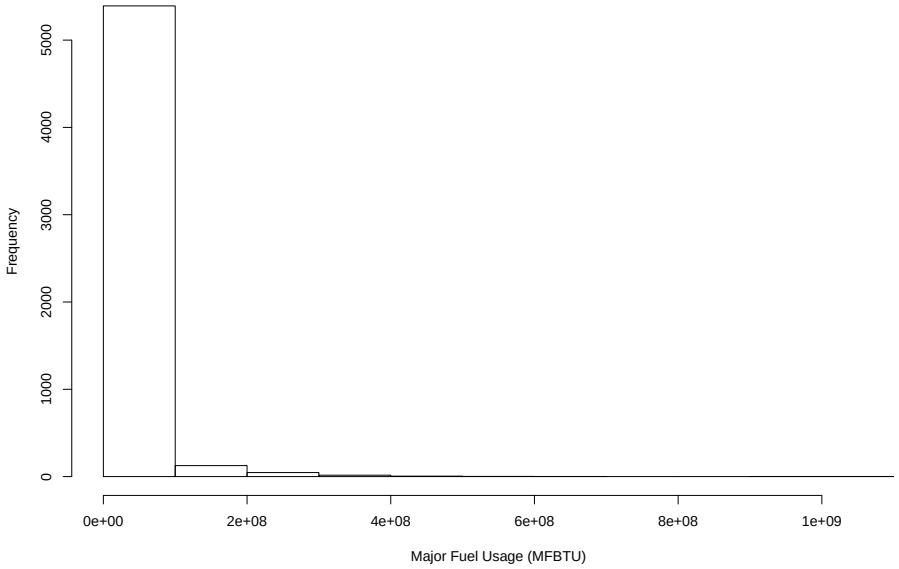


**Figure 4.1**: Histogram of Major Fuel Usage (`MFBTU`).

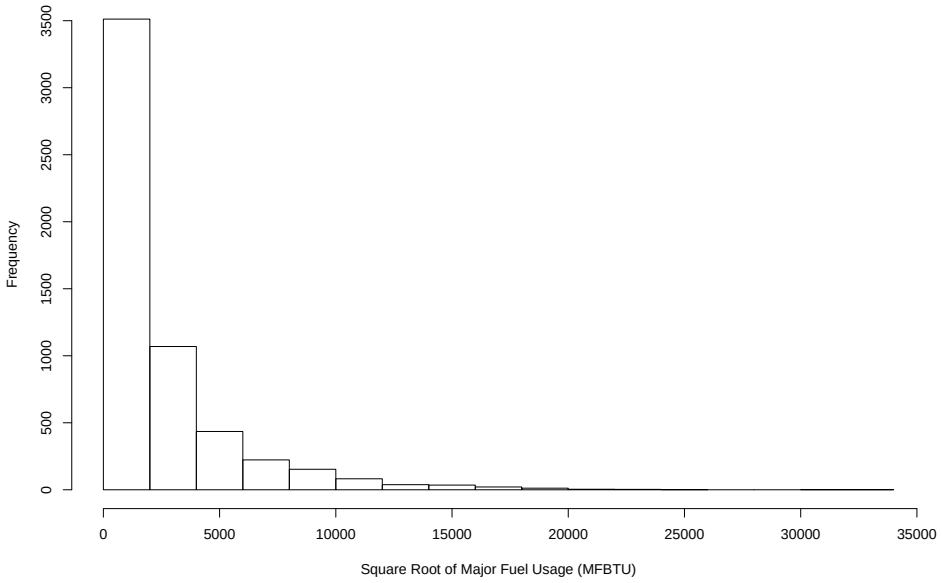


**Figure 4.2**: Histogram of Square Root of Major Fuel Usage ($\texttt{MFBTU}^{1/2}$).

**Table 4.2**: Principal Building Activity (PBA) categories.

| Building Activity | Category Code |
|---|---|
| Vacant | 1 |
| Office/Professional | 2 |
| Laboratory | 4 |
| Warehouse (non-refrigerated) | 5 |
| Food Sales | 6 |
| Public Order and Safety | 7 |
| Health Care (outpatient) | 8 |
| Warehouse (refrigerated) | 11 |
| Religious Worship | 12 |
| Public Assembly | 13 |
| Education | 14 |
| Food Services (restaurants) | 15 |
| Health Care (in-patient) | 16 |
| Nursing Home | 17 |
| Lodging (hotel/motel/dorm) | 18 |
| Strip Shopping | 23 |
| Enclosed Shopping Center/Mall | 24 |
| Retail (exc mall) | 25 |
| Service (exc food) | 26 |
| Other | 91 |

the sorted distances. This process resulted in a total of 1,863 groups of size three, with the last group having four observations.

Three sets of disguisers were created for $\mathbf{X}$ that utilized different parameter values in the implementation of the disguising schemes. I chose to create a small number of disguisers to reduce the chance of the same value appearing in multiple disguiser sets. Repetitions in the disguiser set could help the receiving agency recreate the original data or a closely approximated replicate.

Finally, I assumed that the categorical variables — principal building activity (PBA) and year constructed (YRCON) — were safe to release in their original form. (The variable PBA consists of 20 building categories as defined in Table 4.2. Likewise, the variable YRCON consists of nine categories of varying lengths as shown in Table 4.3.)

**Table 4.3**: Year Constructed (YRCON) categories.

| Construction Year | Category Code |
|---|---|
| 1899 or before | 1 |
| 1900 to 1919 | 2 |
| 1920 to 1945 | 3 |
| 1946 to 1959 | 4 |
| 1960 to 1969 | 5 |
| 1970 to 1979 | 6 |
| 1980 to 1989 | 7 |
| 1990 to 1992 | 8 |
| 1993 to 1995 | 9 |

If the agencies participating in the sharing needed to verify that the records in their respective data sets are one-to-one (as assumed in Chapter 3), they would need to match their data based a specific set of variables. (Data matching is not necessary for the CBECS data because I started with a single data set then split it into two separate components, $\mathbf{X}$ and $\mathbf{Y}$.) By assuming that PBA and YRCON are two of the matching variables, then it is highly likely that both agencies would know the classifications of each building, equivalently the PBA and YRCON category codes. Therefore, assuming that these two variables are safe to release in their current forms is reasonable because they would be common knowledge to both agencies. (The creation of disguisers for categorical variables is an area of future research.)

### 4.1.1 Disguising SQFT

The first variable I chose to disguise was total floor space (SQFT). Exploratory analysis showed that a log transformation changes the distribution of SQFT (see Figure 4.3) from being right skewed to closely approximated by a normal distribution. This transformation allows the use of directional random noise (as described in Section 3.3.3) to be used as the disguising method for log(SQFT). The directional noise was

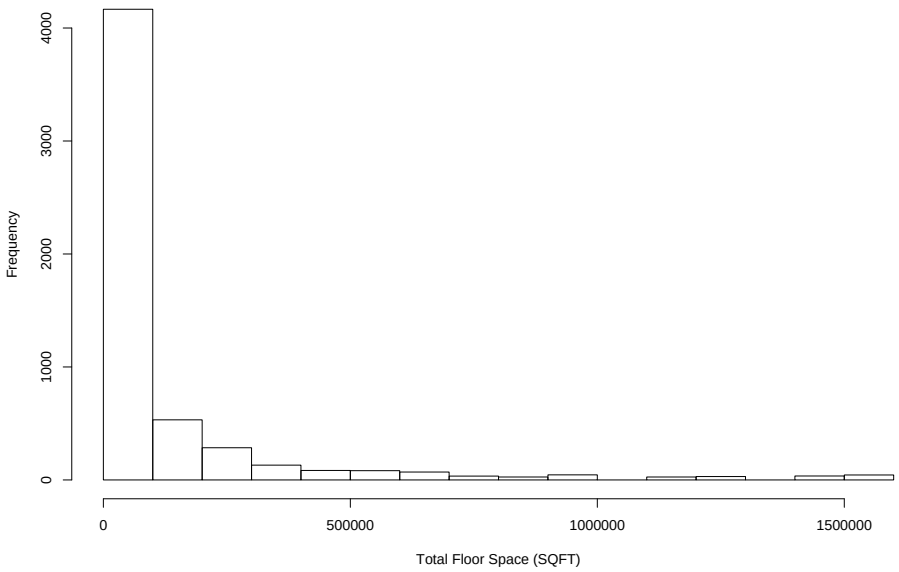


**Figure 4.3**: Histogram of Total Floor Space (`SQFT`).

randomly drawn such that

$$\texttt{log(SQFT)}_{dis}(i) = \texttt{log(SQFT)}(i) + \begin{cases} e_i^+ & \text{if } \texttt{log(SQFT)}(\text{i}) < \overline{\texttt{log(SQFT)}} \\ e_i^- & \text{if } \texttt{log(SQFT)}(\text{i}) > \overline{\texttt{log(SQFT)}} \end{cases} \tag{4.1}$$

where $e_i^+$ is randomly drawn from the positive portion of $N(0, s^2)$, $e_i^-$ is randomly drawn from the negative portion of $N(0, s^2)$, and $s \sim U(0.5, 1)$ for each observation created for the disguiser set. This process adds random noise to the original values of `log(SQFT)` depending on their relation to the sample mean and is limited to situations where the variable to be disguised follows a normal distribution.

The decision to draw $s \sim U(0.5, 1)$ was based on experimentation with $s$ and how changes in the value affected the resulting disguisers as compared to `log(SQFT)`. Since noise was added to the log transformed version of `SQFT`, values for $s$ could not be too large. Otherwise, when the $\texttt{log(SQFT)}_{dis}$ are transformed to the original scale of `SQFT`, the disguised values may fall beyond the range of the original values. If values were allowed outside of this range, extrapolation beyond the data set would

be required, which could reduce the utility of the disguised data.

A potential drawback of the use of directional random noise, specifically when the original data are included in the disguiser set, is the ability to determine the true values among the disguisers. For instance, in cases where the original transformed value (9.529) is smaller than the sample mean, the disguisers will all be larger than it (9.576, 10.395, and 10.415). Likewise, when the value is larger (10.463) the disguisers are all smaller (10.421, 9.866, and 9.696), even though the sample means of the disguisers (10.431, 10.428, 10.434) are similar to the original mean (10.411). These directional discrepancies could become a potential area where disclosures could occur.

### 4.1.2 Disguising NWKER

Next, I decided to create disguisers for number of employees (NWKER). Once again, due to the extreme right skewness of the variable I utilized a square root transformation. However, unlike log(SQFT), the transformed $\texttt{NWKER}^{1/2}$ is still right skewed and therefore cannot be disguised using directional random noise (see Figure 4.4). In addition, the original NWKER data consists of counts of employees or the weighted average of employee counts for those buildings with large numbers of employees. Thus, a different method must be utilized due to the presence of zeros (2.5% of the data) and integer values. To maintain the general distribution of $\texttt{NWKER}^{1/2}$, the disguising method must account for the sparseness in the large values of NWKER. Therefore, to create disguisers, I randomly sampled integers from a preset range, that was dependent on the size of the NWKER values. Next, I added the drawn value to the original and transformed them as done with the original data. The sampling ranges used were found through experimentation on NWKER and its relationship with other variables,

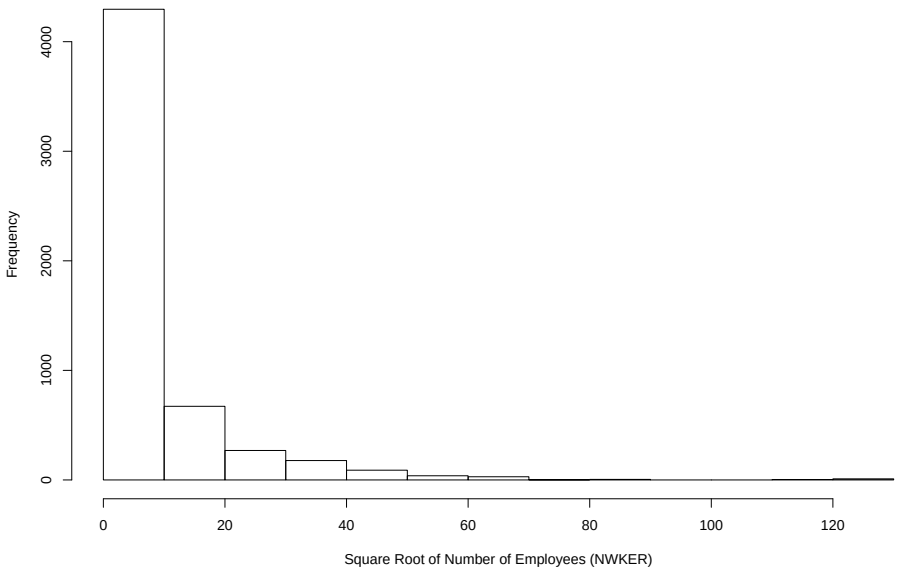


**Figure 4.4**: Histogram of Square Root Number of Employees ($\texttt{NWKER}^{1/2}$).

and are given below:

$$\texttt{NWKER}^{1/2}_{dis}(i) = \texttt{NWKER}^{1/2}(i) + \begin{cases} \text{RS}(0 \text{ to } 2)^{1/2} & \text{if } \texttt{NWKER(i)} = 0 \\ \text{RS}(-1 \text{ to } 2)^{1/2} & \text{if } \texttt{NWKER(i)} = 1 \\ \text{RS}(-2 \text{ to } 2)^{1/2} & \text{if } 1 < \texttt{NWKER(i)} \leq 100 \\ \text{RS}(-5 \text{ to } 5)^{1/2} & \text{if } 100 < \texttt{NWKER(i)} \leq 500 \\ \text{RS}(-10 \text{ to } 10)^{1/2} & \text{if } 500 < \texttt{NWKER(i)} \leq 1,000 \\ \text{RS}(-200 \text{ to } 200)^{1/2} & \text{if } \texttt{NWKER(i)} > 1,000 \end{cases} \tag{4.2}$$

where the notation RS$(\cdot)$ corresponds to a random sample drawn from the range of integers listed in $(\cdot)$. The ranges chosen guarantee that the disguised values are always within the range of the observed data. If values are outside of this range, then (4.2) could be modified accordingly. In addition, the ranges used were chosen to help reduce the risk of disclosures based on how values for NWKER increased. For example, more random noise was added to the large values of NWKER because fewer existed in the data.

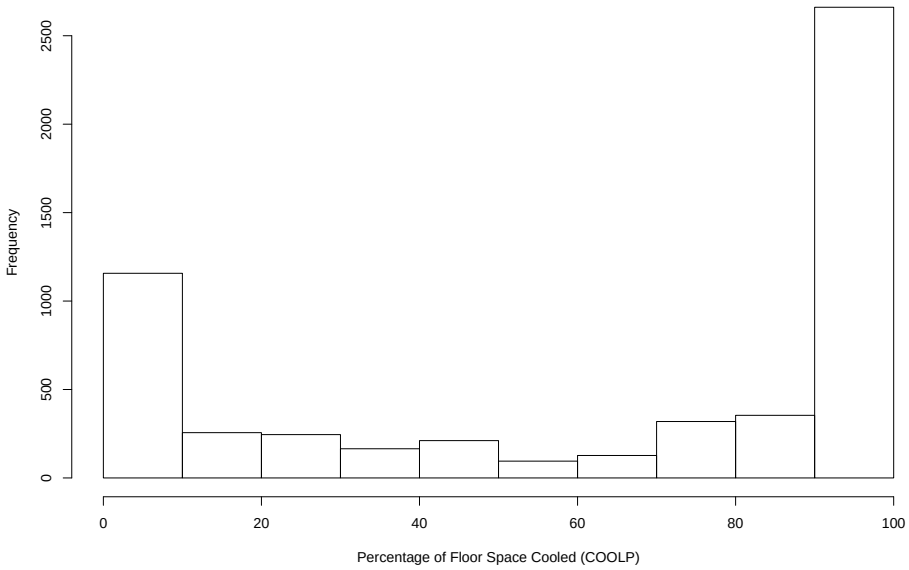


**Figure 4.5**: Histogram of Percentage of Floor Space Cooled (`COOLP`).

### 4.1.3 Disguising `COOLP`

The distribution of percentage of floor space cooled (`COOLP`) as seen in Figure 4.5 is irregular due to its nature as a percentage. Due to the bounding of values between 0% and 100%, the distribution of `COOLP` has peaks at both endpoints and cannot be modified with a transformation.

To disguise `COOLP`, I used a scheme similar to that for `NWKER`, but modified it to incorporate the groupings of $\mathbf{Y}$ and limit the range of values due to the nature of `COOLP` as a percentage. To change the previous disguising scheme to include the groupings of $\mathbf{Y}$, I randomly selected between one and three of the observations within a particular grouping for modification. Then, I randomly sampled a value from the range of values listed in (4.3) based on the value of `COOLP` for the chosen observation. The the ranges of the values from which random noise is sampled are smaller than those for `NWKER`. Since `COOLP` can only take values between 0% and 100%, the addition of at most $\pm 2\%$ noise should yield adequate disguisers because almost every possible

value between 0% and 100% was observed in the original data.

$$\texttt{COOLP}_{dis}(\text{i}) \;=\; \texttt{COOLP}(\text{i}) + \begin{cases} 0 & \text{if } \texttt{COOLP}(\text{i}) \text{ was not sampled} \\ \text{RS}(0 \text{ to } 2)^{1/2} & \text{if } \texttt{COOLP}(\text{i}) = 0 \\ \text{RS}(-1 \text{ to } 2)^{1/2} & \text{if } \texttt{COOLP}(\text{i}) = 1 \\ \text{RS}(-2 \text{ to } 2)^{1/2} & \text{if } 1 < \texttt{COOLP}(\text{i}) \leq 98 \\ \text{RS}(-2 \text{ to } 1)^{1/2} & \text{if } \texttt{COOLP}(\text{i}) = 99 \\ \text{RS}(-2 \text{ to } 0)^{1/2} & \text{if } \texttt{COOLP}(\text{i}) = 100 \end{cases} \tag{4.3}$$

The only exceptions to this disguising scheme are the endpoints, which cannot fall outside the acceptable range. Therefore, in situations where the observed value was close to an endpoint, I reduced the ranges from which random noise could be sampled.

Depending on the level of desired security, the initial sampling of values for adjustment could be modified or removed entirely. If COOLP contained extremely sensitive values, it would be necessary to adjust every value prior to sharing. If only specific values are sensitive, then only these values must be protected. It may also be possible to release COOLP in its original form if the variable does not contain any sensitive observations.

To see how the disguising scheme adjusted COOLP, I compared several different values of COOLP and the corresponding disguisers. For instance, if the original value was 100, under the current scheme, the disguised values would be 99, 100, and 99. If scanning across disguisers for common values in order to find the real observation, there would be a 50% chance of selecting the correct value (if the original data were included with the disguiser set). If the original value was 95, the current method would create 93, 96, and 95 as the disguisers, which means the original value (95) would occur most often and likely be chosen as the truth. These levels of adjustment may or may not add sufficient amounts of uncertainty to the original values. Therefore, if more or less uncertainty is desired, the sampling ranges can be adjusted accordingly.

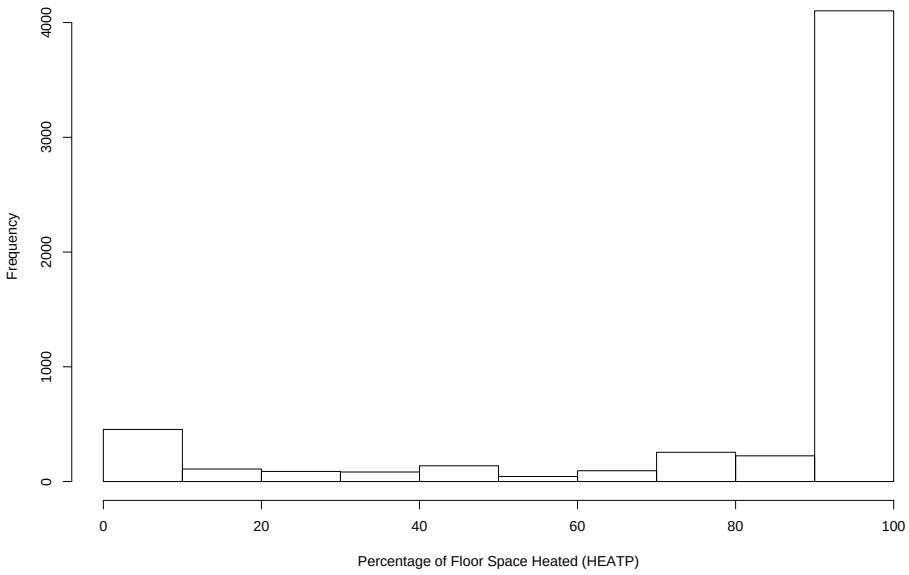


**Figure 4.6**: Histogram of Percentage of Floor Space Heated (`HEATP`).

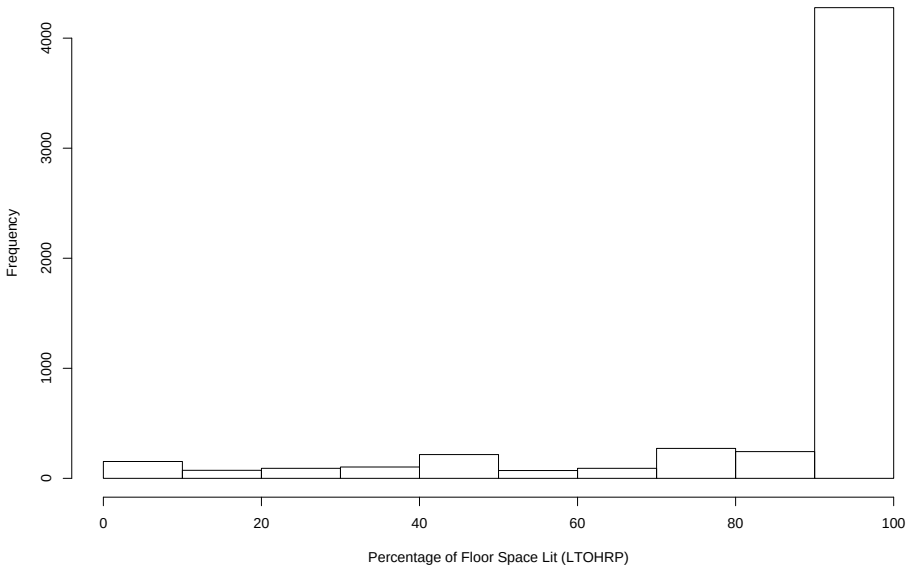


**Figure 4.7**: Histogram of Percentage of Floor Space Lit (`LTOHRP`).

### 4.1.4 Disguising HEATP and LTOHRP

The last two variables that need to be disguised are percentage of floor space heated (HEATP) and percentage of floor space lit (LTOHRP). Since both of these variables are similar to COOLP (see Figures 4.6 and 4.7), I chose to implement the disguising scheme in (4.3) on HEATP and LTOHRP, by replacing each instance of COOLP with that of the variable being disguised. This disguising method should also be appropriate for HEATP and LTOHRP because of the similarities between the variables; however, neither has as large of peak at the left endpoint.

### 4.1.5 Evaluation of Disguisers

To evaluate the effectiveness of the disguisers to both model $\mathbf{X}$ and protect its values, I implemented and compared linear and logistic regression models from both the original and disguised data sets. Since this is an empirical investigation of the disguisers created, lack of model fit or ineffectiveness in disclosure control is directly related to the choice of disguising methods, not the data sharing method that utilizes disguisers. Disguising schemes will be data dependent because each data set will have different circumstances under which it does or does not need protection.

#### Linear Regression Models

The first model implemented to compare the disguiser set with the original data was a linear regression model predicting $\texttt{MFBTU}^{1/2}$ from the variables in $\mathbf{X}$.[2] The model output is given in Table 4.4, with both the parameter estimates and standard errors listed for the included variables. A comparison of the values in the table shows that the disguiser estimates are often smaller or larger than the true estimates, with more estimates being larger.

[2]The use of $\texttt{SQFT}^{1/2}$ instead of log(SQFT) may be a more appropriate choice for this model specification.

**Table 4.4**: Linear Regression Model Comparison of $\mathbf{X}$ and $\mathbf{X}_{dis}$ for $\texttt{MFBTU}^{1/2}$.

| Variables | True Est. (SE) | Disguiser 1 | Disguiser 2 | Disguiser 3 |
|---|---|---|---|---|
| Intercept | -7013.23 (236.07) | -7926.38 (270.59) | -7610.35 (272.66) | -6216.49 (283.98) |
| PBA = 2 | -295.12 (168.15) | -244.25 (172.48) | -390.71 (173.98) | -345.87 (180.92) |
| PBA = 4 | 2029.59 (264.83) | 2191.35 (271.58) | 2040.81 (274.16) | 2259.63 (284.93) |
| PBA = 5 | -310.82 (166.21) | -216.86 (170.45) | -296.27 (172.10) | -130.32 (178.78) |
| PBA = 6 | 786.92 (225.12) | 751.87 (230.90) | 578.18 (232.76) | 527.98 (242.11) |
| PBA = 7 | 207.95 (214.92) | 190.72 (220.43) | 156.84 (222.47) | 196.89 (231.24) |
| PBA = 8 | 178.34 (243.70) | 151.72 (249.94) | 72.42 (252.22) | -3.90 (262.14) |
| PBA = 11 | 422.40 (345.49) | 588.08 (354.33) | 348.65 (357.75) | 571.25 (371.75) |
| PBA = 12 | -273.20 (185.49) | -375.15 (190.15) | -477.39 (191.84) | -537.39 (199.37) |
| PBA = 13 | 399.05 (177.87) | 419.78 (182.44) | 322.16 (184.11) | 408.66 (191.41) |
| PBA = 14 | -243.81 (169.46) | -94.03 (173.79) | -189.68 (175.41) | -49.90 (182.34) |
| PBA = 15 | 843.75 (205.74) | 745.36 (210.87) | 636.12 (212.64) | 410.47 (220.79) |
| PBA = 16 | 4552.12 (196.26) | 4654.80 (201.21) | 4587.27 (203.17) | 4803.04 (210.98) |
| PBA = 17 | 297.13 (249.76) | 371.53 (256.16) | 296.37 (258.56) | 464.52 (268.78) |
| PBA = 18 | 565.80 (180.65) | 679.58 (185.27) | 578.95 (187.02) | 699.06 (194.36) |
| PBA = 23 | -611.58 (183.96) | -528.06 (188.70) | -613.44 (190.44) | -521.49 (197.96) |
| PBA = 24 | -182.45 (223.05) | -2.23 (228.57) | -81.12 (230.95) | 283.90 (239.43) |
| PBA = 25 | 18.44 (174.37) | -43.56 (178.76) | -196.33 (180.23) | -204.71 (187.42) |
| PBA = 26 | 414.61 (178.75) | 347.08 (183.23) | 227.11 (184.75) | 136.46 (192.02) |
| PBA = 91 | 1063.30 (274.14) | 1015.23 (281.15) | 910.40 (283.76) | 1074.39 (295.05) |
| log(SQFT) | 764.97 (17.73) | 832.94 (21.49) | 809.58 (21.82) | 648.31 (22.15) |
| $\texttt{NWKER}^{1/2}$ | 94.92 (2.44) | 103.23 (2.45) | 106.96 (2.44) | 123.35 (2.41) |
| COOLP | 1.18 (0.63) | 1.51 (0.66) | 1.64 (0.66) | 1.66 (0.69) |
| HEATP | 5.36 (0.82) | 5.92 (0.85) | 5.87 (0.86) | 6.58 (0.89) |

For some of the variables included in the regression model, the parameter estimates based on the disguisers are problematic. For instance, the original parameter estimate for PBA = 25 is positive, while for the disguisers it is negative. Since PBA exists in its original form in the disguiser set, the difference in the parameter estimates must be a result of the changes in the other variables due to modification made during disguiser creation. Even though some of the adjustments were small, they reduced the relationships present in the data and its overall utility.

Other problems in the disguiser set that could affect model fit, may be the changes in certain variables for specific values of PBA or YRCON. For the disguised variables, the distributions and basic summaries only experience minor changes for different

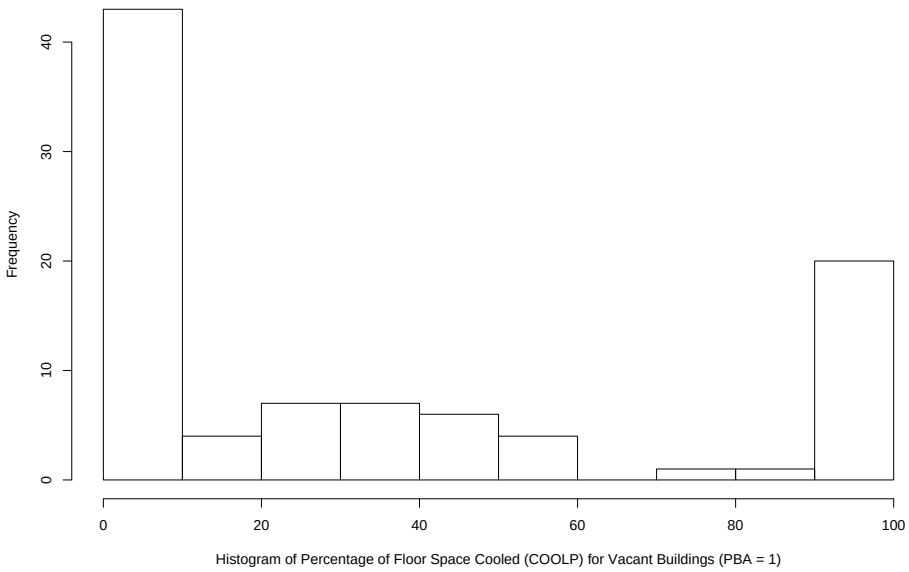


**Figure 4.8**: Histogram of Percentage of Floor Space Cooled (`COOLP`) for Vacant Buildings (`PBA` = 1).

values of `YRCON`. However, for some values of `PBA`, the distributional changes of some variables are large. For instance, Figure 4.8 gives the histogram of `COOLP` for all vacant buildings (`PBA` = 1). A comparison of this histogram and Figure 4.5 show how the shape of certain variables are effected by changes in another variable.

To further compare the parameter estimates in Table 4.4, I calculated the percentage of overlap in the confidence intervals. This percentage compares the confidence intervals based on the disguised data with those from the original data, determining the amount of commonality between the two intervals. The calculation of this percentage is based on the following expression:

$$\mathrm{IO}_l = \frac{1}{2}\left(\frac{U_{over,l} - L_{over,l}}{U_{orig} - L_{orig}} + \frac{U_{over,l} - L_{over,l}}{U_{dis,l} - L_{dis,l}}\right) \tag{4.4}$$

where $L$ and $U$ represent the lower and upper interval bounds, with $(L_{over,l}, U_{over,l})$ being the bounds of the overlap of the original confidence interval with that from the

**Table 4.5**: Confidence Interval Overlap based on the Linear Regression Model Comparisons in Table 4.4.

| Variables | Disguiser 1 | Disguiser 2 | Disguiser 3 |
|---|---|---|---|
| Intercept | 8.08% | 40.32% | 22.02% |
| PBA = 2 | 92.40% | 85.77% | 92.71% |
| PBA = 4 | 84.63% | 98.30% | 78.76% |
| PBA = 5 | 85.77% | 97.84% | 73.40% |
| PBA = 6 | 96.09% | 76.76% | 71.82% |
| PBA = 7 | 98.00% | 94.07% | 96.47% |
| PBA = 8 | 97.26% | 89.13% | 81.73% |
| PBA = 11 | 87.94% | 94.68% | 89.53% |
| PBA = 12 | 86.17% | 72.41% | 65.06% |
| PBA = 13 | 97.08% | 89.19% | 96.46% |
| PBA = 14 | 77.75% | 92.02% | 71.97% |
| PBA = 15 | 87.96% | 74.70% | 48.23% |
| PBA = 16 | 86.83% | 95.54% | 68.65% |
| PBA = 17 | 92.51% | 98.30% | 83.64% |
| PBA = 18 | 84.15% | 98.21% | 81.98% |
| PBA = 23 | 88.58% | 98.30% | 88.08% |
| PBA = 24 | 79.65% | 88.64% | 48.61% |
| PBA = 25 | 91.06% | 69.12% | 68.62% |
| PBA = 26 | 90.49% | 73.70% | 61.80% |
| PBA = 91 | 95.60% | 86.04% | 96.46% |
| log(SQFT) | 11.68% | 42.90% | — |
| $\texttt{NWKER}^{1/2}$ | 13.33% | — | — |
| COOLP | 87.14% | 81.96% | 81.60% |
| HEATP | 82.94% | 84.39% | 63.50% |

$l^{th}$ disguiser (see also Karr *et al.* (2005)).

As seen in Table 4.5, the percentage of interval overlap is high for most parameters. The only exceptions are the instances of variables with 0% coverage (occur in log(SQFT) and $\texttt{NWKER}^{1/2}$). These occurrences of low coverage may simply be due to the choice of disguising method and the random sampling of noise. In the case of log(SQFT), the addition of directional noise could add or subtract large values from the original values when the disguised values are transformed back into the original scale. For $\texttt{NWKER}^{1/2}$, the possibility existed for large values of random noise to be added or subtracted from the original values. This was done primarily to protect the

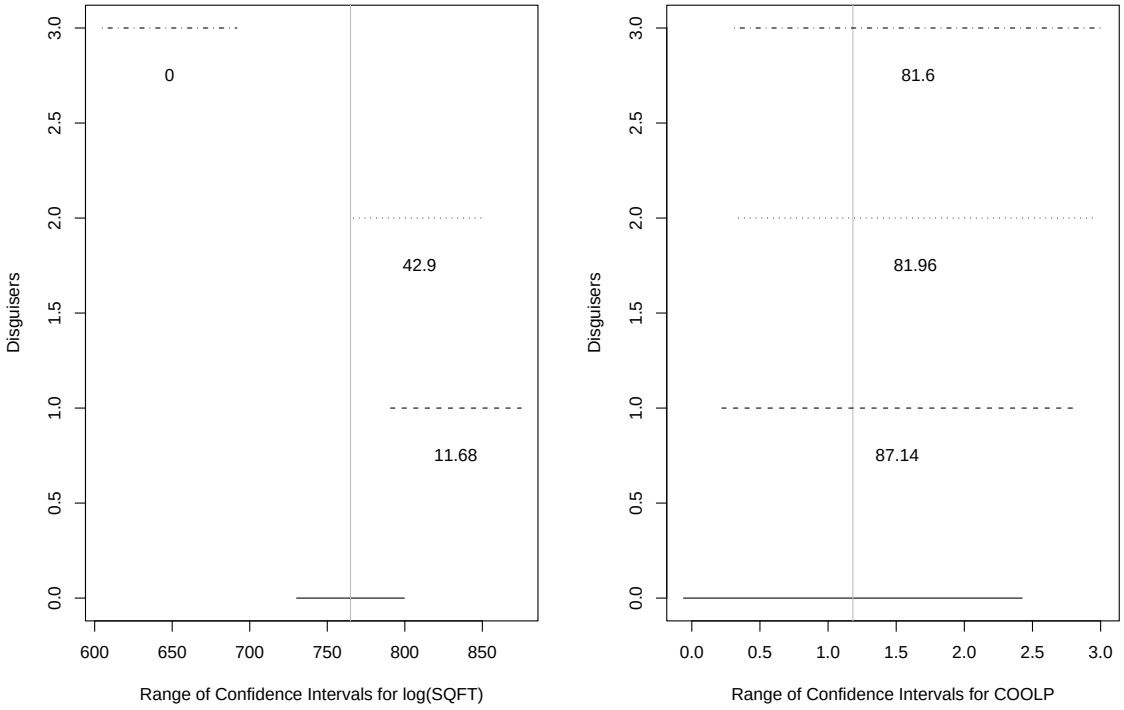


**Figure 4.9**: Comparison of the Confidence Intervals and Coverage for `log(SQFT)` and `COOLP`.

observations with large numbers of employees due to the sparseness in the right tail of the distribution of $\texttt{NWKER}^{1/2}$. In addition, in the original data, all large values of `SQFT` and `NWKER` were the same for a given set of buildings because they had been replaced with a weighted average of the other large values. This feature of the original data does not exist in the disguisers because I added random noise to all values to introduce additional uncertainty in the data.

To examine further the parameter estimates and resulting confidence intervals, I created a plot that graphically compares the confidence intervals for the original data and disguisers. (Figure 4.9 illustrates this comparison for the `log(SQFT)` and `COOLP`.) This gives a visual representation of how the intervals overlap with each other and the original data (the overlap percentage is listed below each interval) and whether

**Table 4.6**: Logistic Regression Model Comparison of $\mathbf{X}$ and $\mathbf{X}_{dis}$ for $\texttt{NGBTU}^{01}$.

| Variables | True Est. (SE) | Disguiser 1 | Disguiser 2 | Disguiser 3 |
|---|---|---|---|---|
| Intercept | -2.093 (0.318) | -2.673 (0.360) | -2.575 (0.362) | -2.345 (0.365) |
| YRCON = 2 | 0.162 (0.271) | 0.184 (0.271) | 0.186 (0.271) | 0.200 (0.271) |
| YRCON = 3 | -0.118 (0.235) | -0.092 (0.235) | -0.090 (0.235) | -0.075 (0.235) |
| YRCON = 4 | -0.195 (0.229) | -0.169 (0.229) | -0.163 (0.229) | -0.154 (0.229) |
| YRCON = 5 | -0.469 (0.225) | -0.441 (0.225) | -0.431 (0.225) | -0.409 (0.225) |
| YRCON = 6 | -0.448 (0.224) | -0.431 (0.224) | -0.420 (0.224) | -0.397 (0.224) |
| YRCON = 7 | -0.917 (0.223) | -0.888 (0.222) | -0.873 (0.222) | -0.862 (0.222) |
| YRCON = 8 | -0.704 (0.249) | -0.667 (0.249) | -0.670 (0.249) | -0.643 (0.248) |
| YRCON = 9 | -0.699 (0.263) | -0.657 (0.263) | -0.643 (0.263) | -0.616 (0.262) |
| log(SQFT) | 0.205 (0.024) | 0.256 (0.029) | 0.245 (0.029) | 0.217 (0.029) |
| $\texttt{NWKER}^{1/2}$ | -0.015 (0.003) | -0.015 (0.003) | -0.014 (0.003) | -0.010 (0.003) |
| HEATP | 0.015 (0.001) | 0.016 (0.001) | 0.016 (0.001) | 0.016 (0.001) |

they cover the observed parameter estimate (vertical line).

### Logistic Regression Models

To further evaluate the disguisers, I compared the output from a logistic regression model based on the original and disguised data. Since the CBECS public-use data file did not contain a dichotomous variable, I created one based on the following:

$$\texttt{NGBTU}^{01} = \begin{cases} 0 & \text{if } \texttt{NGBTU} = 0 \\ 1 & \text{otherwise.} \end{cases} \tag{4.5}$$

The basis of this new variable on NGBTU should provide sufficient results to illustrate logistic regression because the original variable contained 34% zeros.

The output from the logistic regression model based on the original data and the disguisers is shown in Table 4.6. Comparing the parameter estimates in each row of the table suggests that the disguisers yield results similar to those based on the original data. For some variables, the disguiser estimates are smaller than the true estimates, while for others they are larger. The results from this model are better than those from the previous model because the signs of the parameter estimates

**Table 4.7**: Confidence Interval Overlap based on the Logistic Regression Model Comparisons in Table 4.6.

| Variables | Disguiser 1 | Disguiser 2 | Disguiser 3 |
|---|---|---|---|
| Intercept | 56.56% | 64.08% | 81.49% |
| `YRCON` = 2 | 97.88% | 97.74% | 96.40% |
| `YRCON` = 3 | 97.19% | 96.91% | 95.35% |
| `YRCON` = 4 | 97.14% | 96.38% | 95.45% |
| `YRCON` = 5 | 96.81% | 95.61% | 93.14% |
| `YRCON` = 6 | 98.06% | 96.80% | 94.26% |
| `YRCON` = 7 | 96.67% | 94.97% | 93.66% |
| `YRCON` = 8 | 96.21% | 96.57% | 93.82% |
| `YRCON` = 9 | 95.96% | 94.55% | 91.96% |
| `log(SQFT)` | 51.17% | 61.96% | 88.98% |
| $\texttt{NWKER}^{1/2}$ | 98.66% | 89.82% | 54.79% |
| `HEATP` | 92.00% | 93.36% | 92.05% |

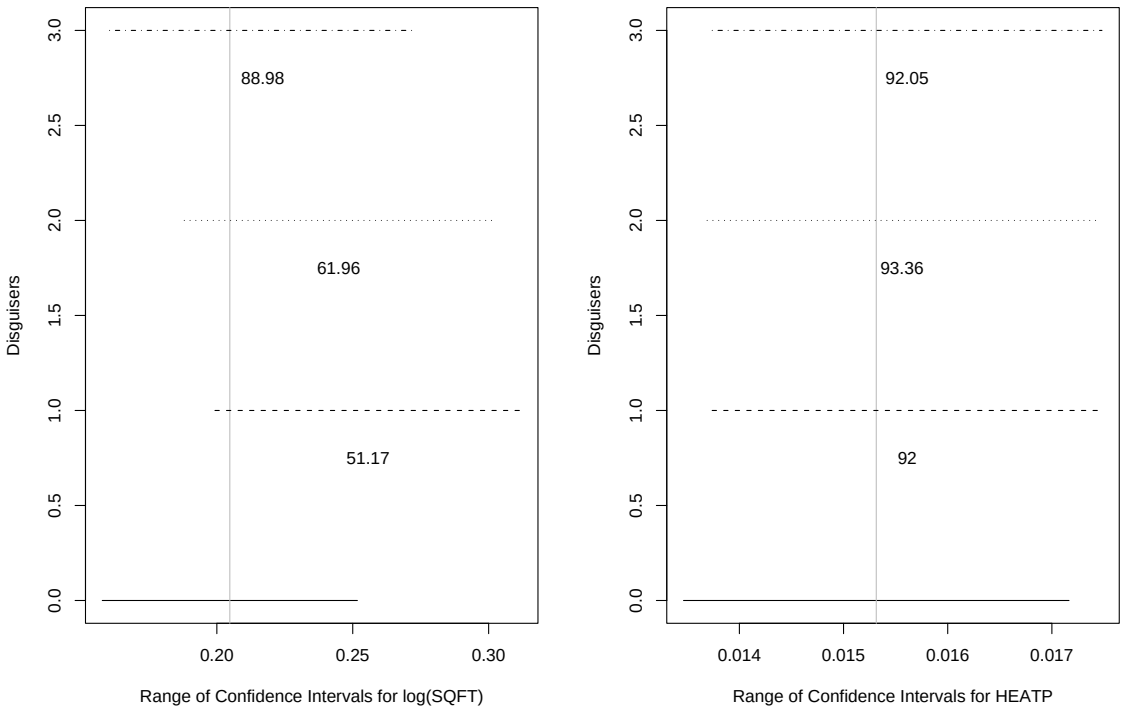


**Figure 4.10**: Comparison of the Confidence Intervals and Coverage for `log(SQFT)` and `HEATP`.

remain the same and estimates are all similar in scale. The results in Table 4.6 are likely a side effect of the chosen disguising methods. Apparently, the relationships used to generate output for this model are not as affected by the adjustments made to create the set of disguisers, as were the relationships in the linear regression model.

To further examine how the parameter estimates compare, I used (4.4) to calculate the percentages of confidence interval overlap. As seen in Table 4.7, the overlap percentages are high for all variables included in the model (`YRCON` is still in its original form). In addition, I created a graphical representation of the confidence intervals for `log(SQFT)` and `HEATP`, to illustrate how the intervals based on the disguisers compare to the original (see Figure 4.10).

### 4.1.6 Suggestions for Generating Disguisers

Additional adjustments could be made to the disguiser set prior to exchange that could increase its similarity to the original data. Although these adjustments only apply to the CBECS data, other disguiser sets may also need additional modifications before exchange. For instance, prior to the release of the CBECS public-use data, SDC methods were performed on `SQFT` and `NWKER`. Specifically, top-coding and rounding were used to protect the identity of the particularly large buildings. (As mentioned in the beginning of this chapter, all buildings in the CBECS data had at least 1,001 square feet of total floor space.) Before release, the area values were rounded to within 5% of the upper limit of the buildings' square footage category. For example, buildings in the range of 5,001 and 10,000 square feet were rounded to the nearest 500 square feet. However, if a building was rounded to 5,000 square feet, it was than coded as having 5,001 square feet. All buildings with over 1,000,000 square feet had their values replaced with a weighted average of the areas of all other large buildings. Rounding was also used on the `NWKER` prior to release. For instance,

all values were rounded to the nearest 250, if the number of employees in the associated building was between 2,500 and 4,999. Buildings with 5,000 or more employees had their values replaced with a weighted average of the number of employees for all buildings with 5,000 or more employees. (For more detailed information on the SDC methods applied to the CBECS data see `http://www.eia.doe.gov/emeu/cbecs/`.) Therefore, to maintain the similarity of the disguised data with the original, similar SDC measures could be applied on the data prior to exchange.

Another aspect of the disguising schemes that could be modified is the way in which noise is added to `log(SQFT)`. It may be beneficial to randomize the noise, such that for some cases noise is added directionally, while in others it is not. This may eliminate the issue described in Section 4.1.1, because it would no longer be easy to determine the original value based on its relationship to the sample average, if known.

Other types of modifications to the disguiser set could change the way in which the data are viewed by the receiving agency. Specifically, variable names could be removed from the data prior to exchange, which would reduce the chance of disclosures because the source of the data would remain unknown. Another possibility would involve the application of linear transformations to the continuous variables prior to exchange. These transformations could be undone by the disguising agency while remaining unknown to the receiving one and creating a further level of confidentiality protection.

To create disguisers for other data sets, these schemes could be modified as dictated by the variables being disguised or new methods could be developed. The adjustment of the sensitive variables in $\mathbf{X}$ is not only dependent on the data to be shared, but also on the level of confidentiality desired.

## 4.2 Investigation of Tree Models for Generating Fully Synthetic Data

In both data sharing methods, fully synthetic data sets must be created during the final stages for dispersement among the participating agencies and public release. Therefore, I will illustrate the creation of fully synthetic data as used in the horizontally partitioned data sharing method of Chapter 2. (Similar data generation techniques could be implemented to create the fully synthetic data used in the vertically partitioned data sharing method of Chapter 3.) First, I randomly partitioned the 5,590 records in the CBECS data among three agencies, such that the distribution of records is 1,500; 2,000; and 2,090. This data will be used to create the fully synthetic data sets that will be shared among the agencies and released for public use. To build each agency's synthetic data generation models, I chose an ordering in which the variables would be generated. This not only simplifies the data generation process, but also creates cohesiveness among the three agencies because I used the same variable orderings to create the synthetic data sets for each agency (similar generation models were implemented for each agency).

Due to the irregular and highly skewed distributions of the variables in the CBECS data, preliminary analyses showed that conventional synthetic data generation methods could not be used for data generation. Therefore, I chose to try other types of models as the basis for the synthetic data generation because this allowed greater flexibility in the modeling of the relationships in the data.

### 4.2.1 Building the Synthetic Data Model

The chosen variable ordering uses the relationships and inter-variable dependencies found within the data during exploratory analyses (see Table 4.8). I chose to begin with the categorical variables (`PBA` and `YRCON`) because they facilitated the generation

**Table 4.8**: Variable Ordering for the Synthetic Data Generation Models.

**1)** PBA

**2)** YRCON | PBA

**3)** log(SQFT) | PBA, YRCON

**4)** $\text{NWKER}^{1/2}$ | PBA, YRCON, log(SQFT)

**5)** COOLP | PBA, YRCON, log(SQFT), $\text{NWKER}^{1/2}$

**6)** HEATP | PBA, YRCON, log(SQFT), $\text{NWKER}^{1/2}$, COOLP

**7)** LTOHRP | PBA, YRCON, log(SQFT), $\text{NWKER}^{1/2}$, COOLP, HEATP

**8)** $\text{MFBTU}^{1/2}$ | **X**

**9)** $\text{MFEXP}^{1/2}$ | **X**, $\text{MFBTU}^{1/2}$

**10)** $\text{ELBTU}^{1/2}$ | **X**, $\text{MFBTU}^{1/2}$, $\text{MFEXP}^{1/2}$

**11)** $\text{ELEXP}^{1/2}$ | **X**, $\text{MFBTU}^{1/2}$, $\text{MFEXP}^{1/2}$, $\text{ELBTU}^{1/2}$

**12)** $\text{NGBTU}^{1/2}$ | **X**, $\text{MFBTU}^{1/2}$, $\text{MFEXP}^{1/2}$, $\text{ELBTU}^{1/2}$, $\text{ELEXP}^{1/2}$

**13)** $\text{NGEXP}^{1/2}$ | **X**, $\text{MFBTU}^{1/2}$, $\text{MFEXP}^{1/2}$, $\text{ELBTU}^{1/2}$, $\text{ELEXP}^{1/2}$, $\text{NGBTU}^{1/2}$

(**X** = PBA, YRCON, log(SQFT), $\text{NWKER}^{1/2}$, COOLP, HEATP, LTOHRP)

of the remaining 11 variables. Exploratory measures suggested that transformations are required to either normalize or decrease the right skewness of some variables (as done in Section 4.1).

The chosen ordering allows the variables to be generated sequentially (Raghunathan *et al.*, 2001), such that the variables with large percentages of zeros (NGBTU and NGEXP) are generated at the end of the data generation process. Many other variable orderings could have been chosen and may have yielded better results. However, I chose this ordering because it allowed all of the building descriptive variables to be generated first, followed by the energy consumption and expenditure variables.

I created a directed graph (see Figure 4.11) representing the relationships used in the generation of synthetic data based on the data designated to agency 1. (In

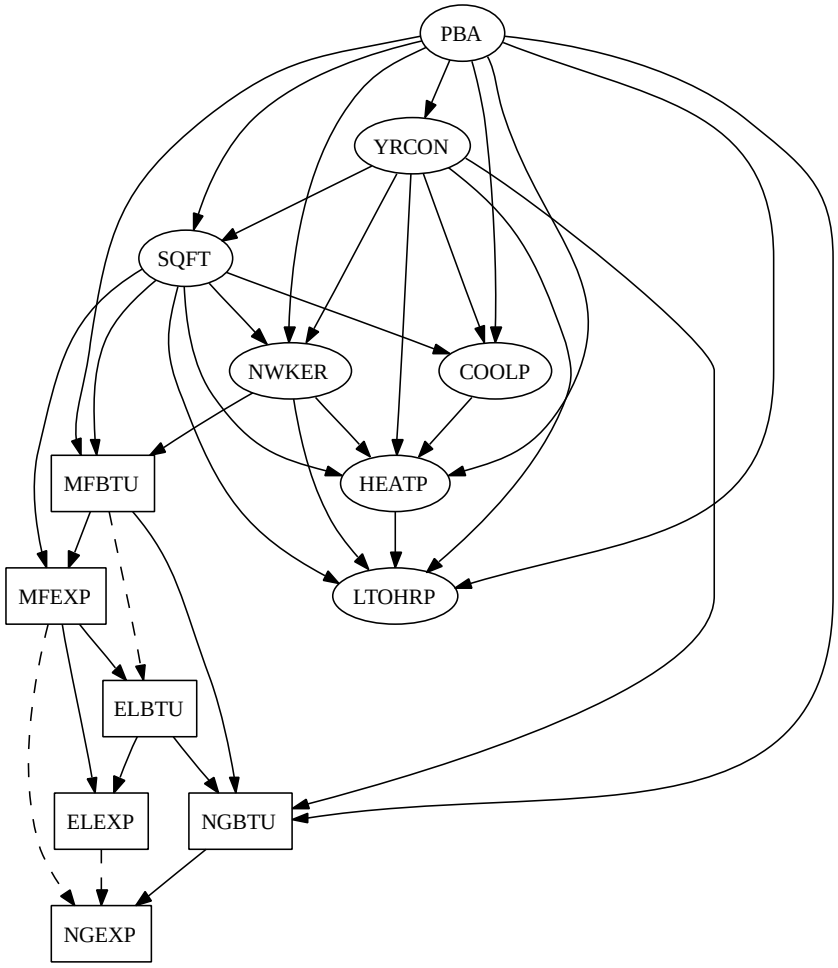


**Figure 4.11**: Graphical representation of the relationships in the data designated to agency 1 and used to generate the synthetic data.

the graph, rectangles represent the energy consumption and expenditure variables, while the ovals correspond to the building descriptive variables. Similar directed graphs were created for the other two agencies and are located in Appendix C.) Even though a sequential ordering was used (see Table 4.8), some of the variables listed ended up not being significant contributors in the data generation models and were not included in Figure 4.11. Several of the connecting links (dashed lines) between the energy consumption and expenditure variables were added because they represent dependencies within the data itself. For instance, for any of the observations in the CBECS data, the constraint `ELBTU` + `NGBTU` $\leq$ `MFBTU` must be satisfied by

the data values. (The same is true for the energy expenditure variables, ELEXP + NGEXP $\leq$ MFEXP.) These dependencies must be preserved in the synthetic data sets; otherwise, the usefulness of the data will be diminished and it would be easy for users to determine that the data were synthetic given available documentation.

This section will focus primarily on the synthetic data generation process for the data designated to agency 1. Brief summaries of the data generation for the remaining two agencies will be given throughout the section (additional information can be found in Appendix C).

### Generating PBA

To generate synthetic replicates for PBA, I used a Bayesian Bootstrap as described in Rubin (1981) and Rubin (1987, p. 44). The Bayesian Bootstrap should draw synthetic values for PBA that yield similar marginal counts (see Table 4.9) because all of the categories of PBA are present within the data designated to agency 1. (The smallest observed count for agency 1 is six, which occurs in category 11.) The remaining two agencies have similar PBA data; therefore, similar synthetic data generation methods were used. To implement the Bayesian Bootstrap (assuming $n$ observations), I used the following steps:

1. Draw $n-1$ uniform random numbers and sort them in ascending order. Label this sequence $0 = a_0, a_1, \ldots, a_{n-1}, a_n = 1$.

2. Calculate the re-sampling probabilities for all $n$ observations in $Y$, $(a_1 - a_0), (a_2 - a_1), \ldots, (a_n - a_{n-1})$.

3. Draw $n$ uniform random numbers, $u_1, u_2, \ldots, u_n$. Impute the value $Y_i$ for $Y_{syn,j}$ if $a_{i-1} < u_j \leq a_i$ for $j = 1, 2, \ldots, n$.

**Table 4.9**: Joint Distribution of PBA and YRCON for the data designated to agency 1.

| | Year Constructed Category | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| PBA | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total |
| 1 | 3 | 4 | 2 | 5 | 3 | 5 | 6 | 0 | 0 | 28 |
| 2 | 9 | 8 | 46 | 33 | 44 | 70 | 102 | 24 | 4 | 340 |
| 4 | 0 | 0 | 1 | 1 | 5 | 4 | 0 | 1 | 1 | 13 |
| 5 | 0 | 4 | 22 | 24 | 39 | 42 | 47 | 12 | 8 | 198 |
| 6 | 0 | 0 | 1 | 3 | 1 | 7 | 4 | 2 | 2 | 20 |
| 7 | 1 | 5 | 2 | 3 | 5 | 9 | 4 | 2 | 2 | 33 |
| 8 | 0 | 0 | 0 | 3 | 5 | 3 | 4 | 2 | 0 | 17 |
| 11 | 0 | 0 | 1 | 1 | 2 | 0 | 1 | 1 | 0 | 6 |
| 12 | 6 | 7 | 4 | 16 | 15 | 11 | 14 | 2 | 0 | 75 |
| 13 | 1 | 5 | 12 | 14 | 11 | 23 | 17 | 5 | 8 | 96 |
| 14 | 3 | 9 | 30 | 38 | 44 | 27 | 17 | 8 | 7 | 183 |
| 15 | 0 | 0 | 5 | 1 | 7 | 10 | 9 | 1 | 1 | 34 |
| 16 | 1 | 0 | 2 | 7 | 7 | 16 | 14 | 3 | 1 | 51 |
| 17 | 0 | 0 | 1 | 2 | 5 | 5 | 5 | 0 | 0 | 18 |
| 18 | 2 | 3 | 2 | 10 | 26 | 12 | 16 | 6 | 2 | 79 |
| 23 | 0 | 1 | 5 | 4 | 8 | 21 | 24 | 3 | 0 | 66 |
| 24 | 0 | 0 | 0 | 0 | 5 | 8 | 15 | 4 | 0 | 32 |
| 25 | 3 | 10 | 13 | 12 | 11 | 20 | 25 | 10 | 6 | 110 |
| 26 | 3 | 4 | 9 | 16 | 14 | 14 | 21 | 6 | 2 | 89 |
| 91 | 0 | 1 | 1 | 2 | 1 | 1 | 4 | 1 | 1 | 12 |
| Total | 32 | 61 | 159 | 195 | 258 | 308 | 349 | 93 | 45 | 1500 |

The steps outlined above are independently repeated $m$ times to create the set of synthetic replicates for agency 1. The same process was used for the remaining two agencies based on their respective data sets. Throughout the generation of synthetic replicates, I will generate the same number of synthetic data sets for each agency.

**Generating** YRCON | PBA

To generate synthetic data for YRCON, I used the synthetic data already generated for PBA and the joint distribution of the two variables as given in Table 4.9. (Marginal counts for the YRCON and PBA categories are listed in the total row and total column of the table.) Similar joint distribution tables of PBA and YRCON for the data designated

to the other agencies are in Appendix C. As seen in Table 4.9, there are a large number of combinations of PBA and YRCON with small numbers of observed occurrences. Comparing this table to that of the entire data set (see Table C.1), it is evident that the designation of records to three different agencies results in a higher number of combinations with small observed counts. The combinations of PBA and YRCON that are singular or occur infrequently may be susceptible to disclosures. However, measures will be taken in future steps to reduce the possibility of such occurrences.

To generate synthetic values for YRCON based upon the PBA values already drawn, a Multinomial distribution is created for each category in PBA. In addition, I place a Dirichlet prior on the sampling probabilities for each category,

$$Y_j|\theta_j \sim \text{Multinomial}(n_j, \theta_{j1}, \theta_{j2}, \ldots, \theta_{jk}) \tag{4.6}$$

$$\theta_j|\alpha_j \sim \text{Dirichlet}(\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jk}) \tag{4.7}$$

for $j = 1, 2, \ldots, J$, such that, $J$ is the number of categories in PBA, $k$ is the number of categories in YRCON, and $\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jk}$ are the observed counts for row $j$ in Table 4.9.

To maintain the structure of the observed data, whenever specific combinations of PBA and YRCON are not present in the data, the corresponding sampling probabilities are set equal to zero. This assumption should be viable because no additional information is known for the unobserved buildings; therefore, if these types of buildings are included in the synthetic data sets, it would be necessary to extrapolate beyond the observed data which could give unreliable inferences.

### Generating log(SQFT) | PBA, YRCON

Prior to modeling SQFT for synthetic replication, I performed a log transformation due to the skewed nature of the original data (similar to Figure 4.3). Transforming SQFT allows the synthetic data to be generated using a normal linear regression model

based on the values already generated for PBA and YRCON. (The transformed SQFT values will be used for all subsequent analyses, but should be transformed back to the original scale prior to any sharing or public release of data.)

The normal linear regression model was set up similar to that described in Gelman *et al.* (2004, pp. 355–356). The model, relating log(SQFT) to PBA and YRCON, was specified as follows:

$$Y|\beta, \sigma^2, X \sim N(X\beta, \sigma^2 I) \tag{4.8}$$

$$(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \tag{4.9}$$

where $Y = \texttt{log(SQFT)}$ and $X = [\texttt{PBA}, \texttt{YRCON}]$. I only used the main effects in this regression model because of the low observed counts for some combination of PBA and YRCON.

Before generating the next variable, I checked the values drawn for log(SQFT) to see whether they were either greater than or equal to the minimum log(SQFT) value of log(1001) or less than or equal to the maximum value of log(1600000). Whenever the values fell outside this range, they were redrawn until acceptable. If values outside the range are allowed into the synthetic data sets, agencies would be able to pick them out as being false, given the documentation provided with the CBECS public-use data file.

### Generating $\texttt{NWKER}^{1/2}$ | PBA, YRCON, log(SQFT)

The distribution of NWKER is similar in shape to the one in Figure 4.4, with a very long right tail and with more than 75% of the buildings having fewer than 100 total employees. Therefore, I chose to apply a variance stabilizing transformation prior to the generation of the synthetic replicates.

To impute synthetic values for $\texttt{NWKER}^{1/2}$ based on those already drawn for PBA,

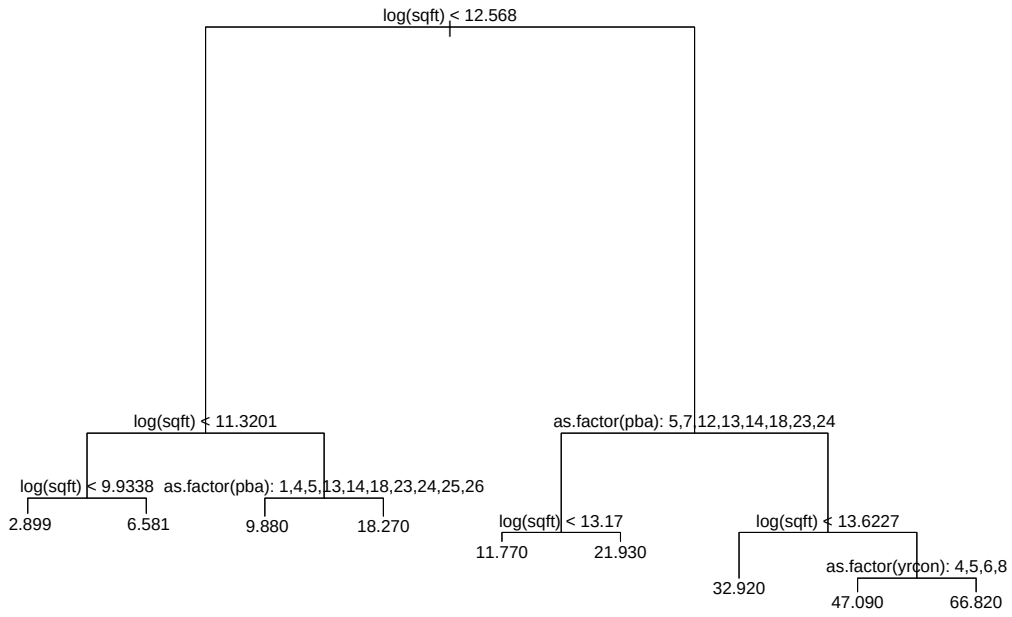


**Figure 4.12**: Tree Diagram for Square Root of Number of Employees ($\texttt{NWKER}^{1/2}$)

YRCON, and log(SQFT), I combined the use of a tree model[3] with a Bayesian Bootstrap (as described in reference to PBA). First, I created a tree model using the original data. This model was then used to separate the synthetic units generated thus far into groups or leaves according to criteria from the original data (see Figure 4.12). Then, I bootstrapped values of $\texttt{NWKER}^{1/2}$ from the leaves of the original data to generate values for the synthetic units corresponding to the data already created. (The leaves in this tree model each had more than 10 observations.) The bootstrapping process was repeated until $m$ sets of replicates were generated.

[3]I used the `tree` function from the *tree* package in R (R Development Core Team, 2004).

**Generating** COOLP | PBA, YRCON, log(SQFT), NWKER$^{1/2}$

The distribution of COOLP for the data designated to agency 1 is similar to the one shown in Figure 4.5. Due to the irregular shape of the distribution, I utilized the same process used to generate NWKER$^{1/2}$ in the previous section to create synthetic replicates for COOLP. The combination of methods is used because it gives flexibility in modeling non-normal distributions like COOLP. (It is possible to independently model COOLP with a Beta distribution, but this does not preserve the relationships to the other variables in the data.) Prior to bootstrapping values from the leaves of the constructed tree model, leaf sizes were checked to see whether they had sufficient numbers for sampling. (The actual tree used in this process will not be shown. However, the overall structure and size is similar to that in Figure 4.12.)

**Generating** HEATP | PBA, YRCON, log(SQFT), NWKER$^{1/2}$, COOLP

HEATP is another percentage variable bounded between 0% and 100%, with a single peak at 100% and a distribution similar to Figure 4.6. As with the previous two variables, methods other than linear regression are required to to create synthetic replicates that both follow the distribution of HEATP and maintain its relationships with other variables. To do so, I constructed a tree model that classified the original observations into groups or leaves. Next, I categorized the synthetic data already generated according to the tree model and bootstrapped synthetic values for HEATP from the original data leaves (the leaves with sufficient numbers for sampling). Therefore, the relationships present in the original data are preserved. This process was repeated $m$ times for each agency and their designated data.

### Generating LTOHRP | PBA, YRCON, log(SQFT), NWKER$^{1/2}$, COOLP, HEATP

I used the same generation scheme implemented for COOLP and HEATP to create the synthetic replicates for LTOHRP. (The shape of its distribution is similar to that in Figure 4.7.) Throughout the generation process, I checked variable summaries and relationships with the other variables to verify that the structure of the original data was actually preserved in the synthetic replicates.

### Generating MFBTU$^{1/2}$ | X

Progressing to the energy consumption and expenditure variables, I used the synthetic values already generated for PBA, YRCON, log(SQFT), NWKER$^{1/2}$, COOLP, HEATP, and LTOHRP, along with the relationships in the original data as the basis of the synthetic generation models. The inter-variable dependencies within the energy consumption variables (MFBTU, ELBTU, and NGBTU) complicate the creation of synthetic replicates. Therefore, I begin by generating major fuel consumption (MFBTU) because the other two variables are subsets of it and are subject to additional constraints.

Before generating synthetic replicates for MFBTU, I applied a variance stabilizing transformation (square root) to reduce the right skewness and tail size (similar to Figures 4.1 and 4.2). Due to the nature of the distribution, I continued using a combination of a tree model and Bayesian Bootstrapping for synthetic data generation. Prior to sampling values from the leaves of the tree model, I checked each leaf for sufficient numbers. Since the synthetic replicates are drawn from the original (transformed) data, all values of MFBTU$^{1/2}$ will be greater than or equal to zero, and will not exceed the maximum MFBTU$^{1/2}$ value.

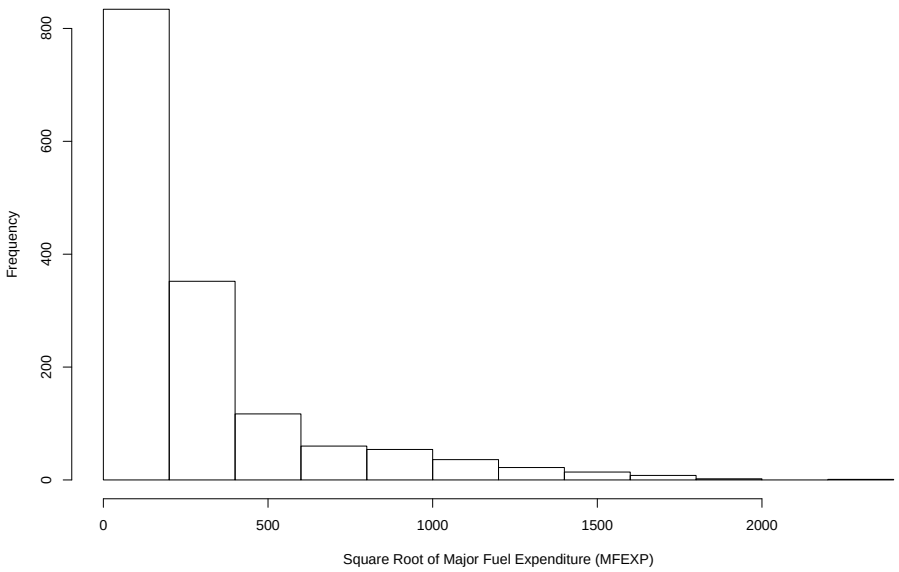


**Figure 4.13**: Histogram of the Square Root of Major Fuel Expenditure ($\texttt{MFEXP}^{1/2}$)

**Generating $\texttt{MFEXP}^{1/2}$ | X, $\texttt{MFBTU}^{1/2}$**

As with MFBTU, the variable major fuel expenditure (MFEXP) was modeled prior to the other energy expenditure variables due to the inter-variable dependencies between it and ELEXP and NGEXP. Similar to MFBTU, I applied a variance stabilizing transformation (square root) on MFEXP prior to modeling due to the extreme right skewness of the variable. Figure 4.13 shows the distribution of $\texttt{MFEXP}^{1/2}$, the transformed version of MFEXP. As with any of the other transformed variables, before sharing the data or releasing the synthetic replicates for public use, the variable should be transformed back to its original form.

The same general process used to generated the synthetic replicates for $\texttt{MFBTU}^{1/2}$ is used once again. The combination of a tree model and Bayesian Bootstrap help preserve the structure of the original data while maintaining the relationships with the other variables. Prior to bootstrapping, all leaves were checked to ensure that

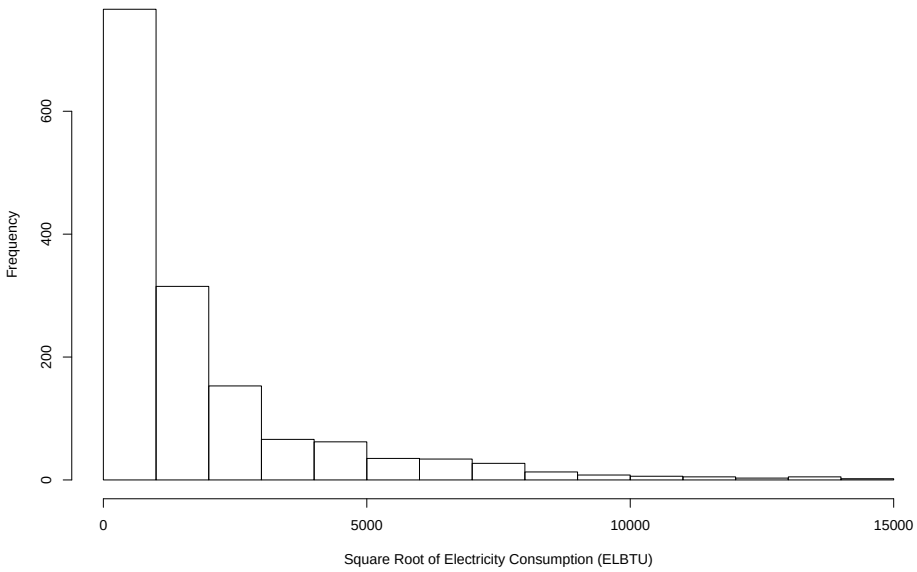


**Figure 4.14**: Histogram of the Square Root of Electricity Consumption ($\texttt{ELBTU}^{1/2}$)

sufficient numbers were present for sampling. In addition, this model structure guarantees that all synthetic values of $\texttt{MFEXP}^{1/2}$ are greater than or equal to zero, but not greater than the largest observed value.

### Generating $\texttt{ELBTU}^{1/2}$ | $\mathbf{X}$, $\texttt{MFBTU}^{1/2}$, $\texttt{MFEXP}^{1/2}$

In order to maintain the structure within the observed data, I proceeded next to the variable electricity consumption (ELBTU) because it is the largest fuel component contained within MFBTU. Exploratory measures on the original data showed that any synthetic value generated for ELBTU must at least zero and at most MFBTU. If the generated values do not follow these constraints, then the synthetic replicates will not have the same properties as the original data and may result in a decrease in the utility of the synthetic data. As with the other energy variables, the right skewness of ELBTU required a transformation (square root) prior to replication. The distribution of the transformed variable ($\texttt{ELBTU}^{1/2}$) is given in Figure 4.14, and although the skewness

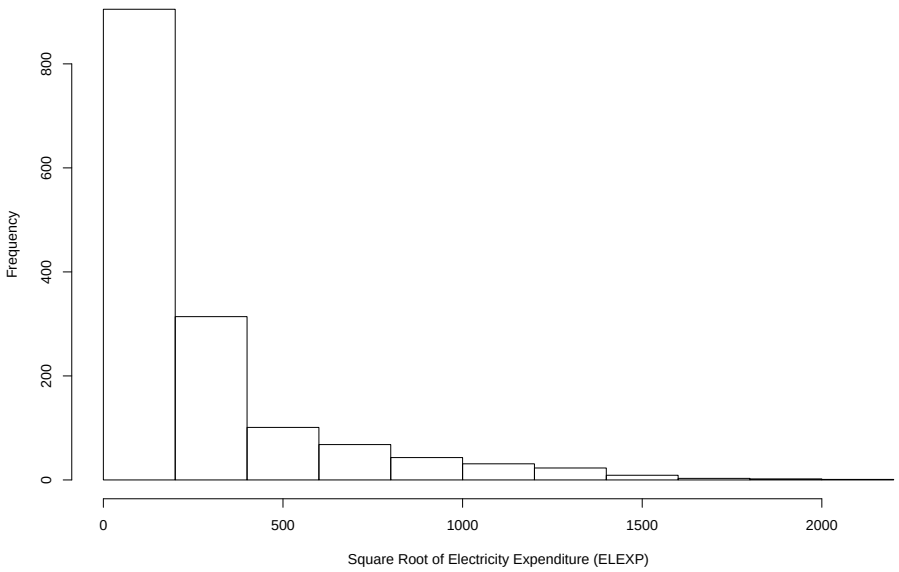


**Figure 4.15**: Histogram of the Square Root of Electricity Expenditure ($\texttt{ELEXP}^{1/2}$)

has been reduced, extreme values still remain. As with the previous variables, I bootstrapped values for the synthetic replicates from the original data based on the leaves determined by a tree model fit to the original data. Once complete, I checked the values of the synthetic replicates to determine whether the constraints had been satisfied. If not, the corresponding values were redrawn until they satisfied all of the constraints.

### Generating $\texttt{ELEXP}^{1/2} \mid \mathbf{X}, \texttt{MFBTU}^{1/2}, \texttt{MFEXP}^{1/2}, \texttt{ELBTU}^{1/2}$

Similar to the dependency between ELBTU and MFBTU, electricity expenditures (ELEXP) has similar properties and relationship with MFEXP. Exploratory measures showed that all values of ELEXP must be at least zero, but at most MFEXP (because it is only one of the fuel source expenditures that comprise MFEXP). If these constraints do not exist in the synthetic replicates, then the usefulness of the synthetic data may decrease because it will not accurately model the data. Prior to synthetic replication,

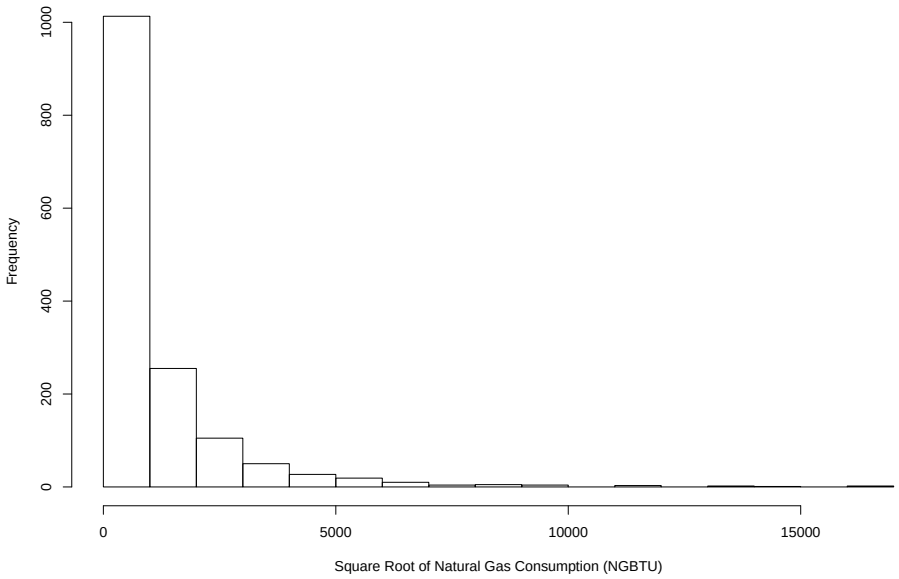


**Figure 4.16**: Histogram of the Square Root of Natural Gas Consumption ($\texttt{NGBTU}^{1/2}$)

I utilized a transformation (square root) on `ELEXP` due to the extreme right skewness of the original distribution (see Figure 4.15). As with the other variables, I used a combination of a tree and bootstrapping sampling procedure. Before completion, the values were checked to determine whether all constraints were satisfied; otherwise, the problematic values were redrawn until the constraints were fulfilled.

### Generating $\texttt{NGBTU}^{1/2} \mid \mathbf{X}, \texttt{MFBTU}^{1/2}, \texttt{MFEXP}^{1/2}, \texttt{ELBTU}^{1/2}, \texttt{ELEXP}^{1/2}$

I chose to generate synthetic replicates for natural gas consumption (`NGBTU`) near the end because, of the three energy consumption variables, it has the largest percentage of zeros — 34.8% of the `NGBTU` records designated to agency 1 are zero. Therefore, the placement of `NGBTU` near the end of the sequential generation scheme is a direct result of the information contained within the variable as compared to the percentage of zeros and other energy consumption variables. As with the other energy variables, I transformed (square root) `NGBTU` prior to synthetic replication due to the shape of

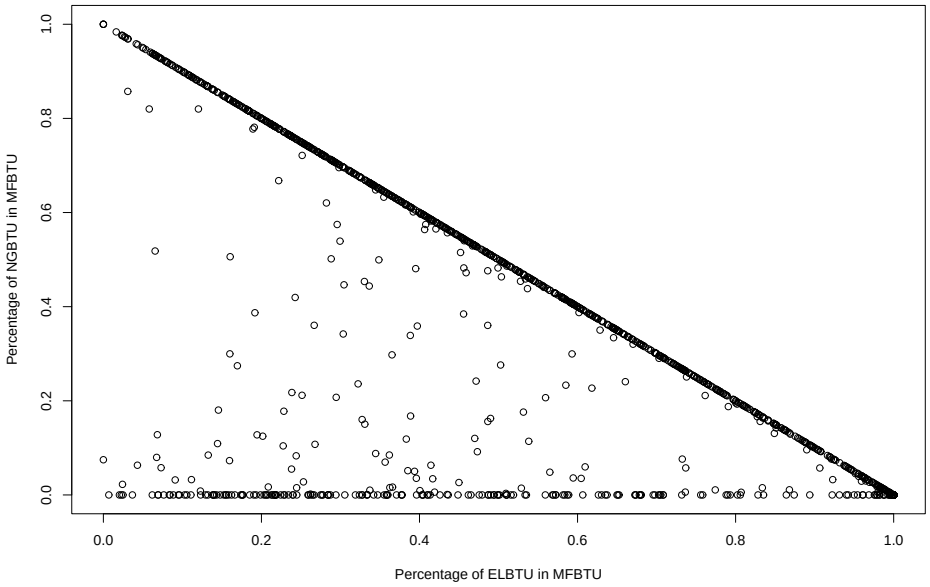


**Figure 4.17**: Percentage of ELBTU and NGBTU as compared to MFBTU.

the distribution (see Figure 4.16).

A combination method was used again to generate synthetic replicates because of the inherent constraints within the original data which can be satisfied by the sampled values, since all values of NGBTU must be at least zero, and be at most the difference of MFBTU and ELBTU. As evident in Figure 4.17, the combined percentage of ELBTU and NGBTU in relation to MFBTU always is 100% or less.[4] Before completion, all bootstrapped values were checked against the constraints, and if they were not satisfied another value was sampled.

### Generating $\texttt{NGEXP}^{1/2} \mid \mathbf{X}$, $\texttt{MFBTU}^{1/2}$, $\texttt{MFEXP}^{1/2}$, $\texttt{ELBTU}^{1/2}$, $\texttt{ELEXP}^{1/2}$, $\texttt{NGBTU}^{1/2}$

The last variable for which I generated synthetic values was natural gas expenditure (NGEXP). This variable was placed at the end of the sampling sequence because of the presence of 34.7% zeros within the variable. Due to the right skewness of

[4]The 1995 CBECS public-use data contains other energy variables that make up the remaining percentages whenever the sum of ELBTU and NGBTU is less than MFBTU.

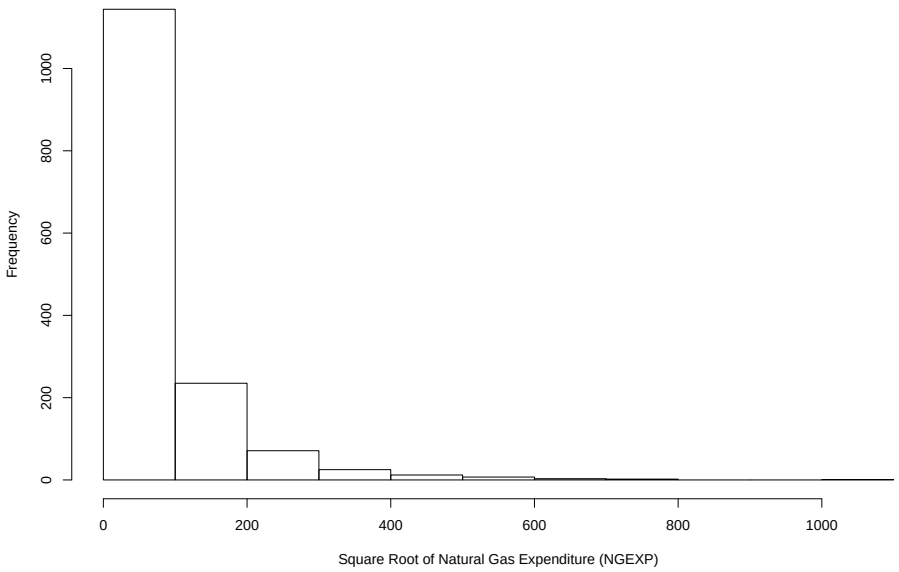


**Figure 4.18**: Histogram of the Square Root of Natural Gas Expenditure ($\texttt{NGEXP}^{1/2}$)

NGEXP, I applied a transformation (square root) to the variable, which resulted in the distribution given in Figure 4.18.

As before, I created a tree model using the original data (see the relationships in Figure 4.11). All leaf sizes were checked prior to sampling to verify that each leaf had sufficient numbers available. The bootstrapping process guaranteed that all synthetic NGEXP values would be greater than or equal to zero. However, to ensure that the sum of NGEXP and ELEXP was always less than or equal to MFEXP, sampled values were checked and, if necessary, re-sampled from the corresponding leaf.

### Other Generation Methods

I also implemented other generation schemes during the creation of the synthetic replicates. Some of these schemes varied the way in which the electricity and natural gas variables were sampled in relation to the previously described constraints. For instance, modeling and sampling the respective sums of electricity and natural gas,

then sampling the values of electricity and finally determining natural gas through subtraction. Other schemes only sampled values from the set of units that satisfied the constraints (the methods described in this section implement a variation of this scheme). Other modifications involved the way that synthetic values were sampled from the leaves of the tree models, including drawing values from a density rather than a Bayesian Bootstrap.

### 4.2.2 Evaluation of the Shared Synthetic Data

To illustrate whether or not the synthetic data accurately model the real data, capturing the data structures and inherent relationships, summary statistics are calculated and regression models (linear and logistic) are constructed. These models may not be the ideal models for this data set, but will be used for comparative purposes only.

#### Summary Statistics

To compare the basic summaries of the combined synthetic data to the combined truth, I calculated estimates of the means and associated variances of each variable in the CBECS data (each of the `PBA` and `YRCON` categories have been listed). As seen in Table 4.10, most of the combined synthetic estimates (using (2.1)) are very close to combined truth, except for the electricity and natural gas consumption and expenditure variables. In addition, most of the associated variances (calculated with (2.4) and adjusted using (2.26)) are larger than the observed values, except for the previous listed variables.

To further compare these estimates, I calculated the percentage of confidence interval overlap using (4.4); these values are listed in the last column of Table 4.10. According to the percentages given, several of the original and synthetic confidence intervals do not overlap because of the large differences in the estimate values or

**Table 4.10**: Comparison of the Summary Statistics (estimates and associated variances) for the True and Synthetic Data.

| Variable | Combined Truth | Combined Synthetic | CI Coverage |
|---|---|---|---|
| % PBA = 1 | 1.68 (0.55) | 1.64 (2.98) | 51.59% |
| % PBA = 2 | 21.96 (5.71) | 21.70 (31.10) | 55.12% |
| % PBA = 4 | 0.91 (0.30) | 0.85 (1.53) | 51.21% |
| % PBA = 5 | 13.24 (3.83) | 13.26 (21.01) | 54.18% |
| % PBA = 6 | 1.59 (0.52) | 1.59 (2.79) | 51.56% |
| % PBA = 7 | 1.97 (0.64) | 1.97 (3.59) | 51.70% |
| % PBA = 8 | 1.21 (0.40) | 1.12 (2.01) | 51.41% |
| % PBA = 11 | 0.43 (0.14) | 0.41 (0.74) | 50.83% |
| % PBA = 12 | 4.54 (1.44) | 4.64 (8.15) | 52.53% |
| % PBA = 13 | 6.33 (1.98) | 6.29 (10.74) | 53.02% |
| % PBA = 14 | 12.39 (3.62) | 12.43 (19.86) | 54.06% |
| % PBA = 15 | 2.50 (0.81) | 2.53 (4.47) | 51.92% |
| % PBA = 16 | 3.64 (1.17) | 3.65 (6.38) | 52.32% |
| % PBA = 17 | 1.12 (0.37) | 1.09 (2.00) | 51.30% |
| % PBA = 18 | 5.60 (1.76) | 5.56 (9.54) | 52.85% |
| % PBA = 24 | 4.98 (1.58) | 5.17 (8.86) | 52.65% |
| % PBA = 24 | 1.82 (0.60) | 1.80 (3.30) | 51.64% |
| % PBA = 25 | 7.53 (2.32) | 7.70 (12.97) | 53.22% |
| % PBA = 26 | 5.75 (1.81) | 5.77 (9.95) | 52.86% |
| % PBA = 91 | 0.81 (0.27) | 0.83 (1.50) | 51.09% |
| % YRCON = 1 | 2.15 (0.70) | 2.09 (3.72) | 51.82% |
| % YRCON = 2 | 3.91 (1.25) | 3.94 (6.95) | 52.38% |
| % YRCON = 3 | 10.18 (3.05) | 10.32 (16.98) | 53.70% |
| % YRCON = 4 | 14.12 (4.04) | 14.21 (22.09) | 54.30% |
| % YRCON = 5 | 18.10 (4.94) | 17.95 (26.80) | 54.77% |
| % YRCON = 6 | 19.94 (5.32) | 19.84 (29.09) | 54.93% |
| % YRCON = 7 | 22.75 (5.86) | 22.91 (32.33) | 55.15% |
| % YRCON = 8 | 5.29 (1.67) | 5.24 (9.21) | 52.75% |
| % YRCON = 9 | 3.56 (1.14) | 3.50 (6.06) | 52.33% |
| log(SQFT) | 10.41 (0.00) | 10.44 (0.00) | 51.11% |
| $\text{NWKER}^{1/2}$ | 9.16 (0.03) | 8.40 (0.02) | — |
| COOLP | 65.39 (0.30) | 65.41 (0.30) | 76.93% |
| HEATP | 84.71 (0.17) | 85.02 (0.17) | 70.81% |
| LTOHRP | 89.27 (0.10) | 90.14 (0.09) | — |
| $\text{MFBTU}^{1/2}$ | 2494.37 (1645.21) | 2378.39 (1756.33) | 51.27% |
| $\text{MFEXP}^{1/2}$ | 284.72 (19.19) | 270.42 (19.16) | 61.41% |
| $\text{ELBTU}^{1/2}$ | 1779.51 (890.41) | 1457.96 (707.12) | 51.49% |
| $\text{ELEXP}^{1/2}$ | 254.65 (16.57) | 209.86 (14.93) | — |
| $\text{NGBTU}^{1/2}$ | 1050.78 (610.22) | 344.94 (62.27) | 50.65% |
| $\text{NGEXP}^{1/2}$ | 69.02 (1.95) | 15.10 (0.05) | — |

extremely small associated variances. For instance, $\texttt{NWKER}^{1/2}$ and LTOHRP have 0% coverage because the estimated values differ and the associated variances are very small; therefore, the calculated confidence intervals do not overlap. (The use of confidence intervals may be questionable for most of the variables in the CBECS data due to their irregular or skewed distributions.) The variables with zero coverage may become the source of additional problems in further inference calculations due to the sequential variable ordering chosen. Misspecification in generation models early in the ordering may cause unforeseen problems in the variables generated near the end and subsequent inference because of unpreserved data relationships. (Variables such as $\texttt{NGBTU}^{1/2}$ and $\texttt{NGEXP}^{1/2}$ may be sources of problems in further inferences.)

### Linear Regression Models

The first model used to evaluate the effectiveness of the synthetic replicates was a linear regression model predicting $\texttt{MFBTU}^{1/2}$ from the descriptive variables (as done in Section 4.1.5). Output (parameter estimates and standard errors) from the regression models based on the original data and then combined across agencies is given in Table 4.11. Similarly, combined synthetic estimates are given for the regression models, calculated using (2.1) and (2.4) (adjusted as shown in (2.26)).

To compare the parameter estimates in Table 4.11, I calculated the percentage of overlap in the confidence intervals using (4.4). A comparison of the confidence interval coverages for the combined synthetic data and the combined truth show that most estimates are reasonably close to the truth. However, for some variables there was no overlap of the intervals, suggesting that the synthetic data does not adequately model certain aspects of the real data. ($\texttt{NWKER}^{1/2}$ was problematic in Table 4.10, so the lack of fit for the associated parameter estimate can be expected.)

The inferential problems seen in Table 4.11 are a direct result of the choice to

**Table 4.11**: Comparison of the Linear Regression Models for the True and Synthetic Data.

| Variables | Combined Truth | Combined Synthetic | CI Coverage |
|---|---|---|---|
| Intercept | -7045.12 (236.52) | -6919.01 (179.76) | 86.15% |
| PBA = 2 | -262.96 (167.67) | -256.14 (113.55) | 83.86% |
| PBA = 4 | 1958.89 (264.94) | 1123.29 (282.75) | 22.18% |
| PBA = 5 | -286.69 (165.86) | -107.34 (113.42) | 69.69% |
| PBA = 6 | 815.38 (230.96) | 311.66 (231.35) | 44.41% |
| PBA = 7 | 229.03 (213.89) | -3.37 (222.98) | 72.89% |
| PBA = 8 | 250.78 (243.89) | 270.01 (180.14) | 86.93% |
| PBA = 11 | 399.47 (349.02) | 338.62 (389.39) | 94.82% |
| PBA = 12 | -243.23 (184.99) | 198.07 (130.39) | 29.50% |
| PBA = 13 | 430.89 (177.46) | 14.99 (179.91) | 40.63% |
| PBA = 14 | -213.82 (169.15) | -92.42 (171.75) | 81.84% |
| PBA = 15 | 857.49 (206.01) | 512.83 (256.89) | 62.77% |
| PBA = 16 | 4501.89 (195.97) | 3178.54 (232.03) | — |
| PBA = 17 | 306.41 (248.79) | 168.94 (289.79) | 87.48% |
| PBA = 18 | 606.36 (180.58) | 5.47 (226.81) | 25.07% |
| PBA = 23 | -597.38 (184.01) | -73.82 (229.17) | 35.78% |
| PBA = 24 | -174.76 (222.06) | 130.33 (91.35) | 60.93% |
| PBA = 25 | 53.41 (174.07) | 74.72 (217.85) | 89.95% |
| PBA = 26 | 436.58 (178.51) | 229.77 (121.07) | 67.25% |
| PBA = 91 | 988.75 (273.47) | 12.11 (214.72) | — |
| log(SQFT) | 765.92 (17.83) | 819.77 (19.61) | 26.68% |
| $\text{NWKER}^{1/2}$ | 95.49 (2.47) | 80.19 (4.00) | — |
| COOLP | 1.16 (0.63) | -0.22 (0.75) | 49.56% |
| HEATP | 5.30 (0.82) | -0.24 (0.98) | — |

combine trees and the Bayesian Bootstrap to generate synthetic data sets. Initially, the decision to use trees as the data generation method was to model the irregular data: ideally, the trees should branch and split as determined by the relationships in the data. However, due to the designation of 5,590 records to three different agencies, the data pool was reduced which limited the overall tree sizes. I was forced to use small trees or else the leaves did not contain sufficient numbers for bootstrapping. Pruning of the trees to increase leaf size reduces the overall tree size and number of splitting relationships.

The variable arrangement may have also played a role in loss of data utility. By

choosing a sequential ordering, the possibility to loose or incorrectly model relationships was great. In addition, the comparison model chosen may not show the true scope of the combined synthetic data because it was for illustrative purposes only. Fitting linear regression models to this data was a difficult task, due to the underlying distributions of the variables.

### Logistic Regression Models

To further evaluate the quality of the shared fully synthetic data sets, I compared the output from a logistic regression model based on the combined original data and that from the shared synthetic replicates. Since the CBECS data does not contain a dichotomous variable, I created one for NGBTU using the data designated to each agency based on the following:

$$\texttt{NGBTU}_j^{01} = \begin{cases} 0 & \text{if } \texttt{NGBTU}_j = 0 \\ 1 & \text{otherwise} \end{cases} \tag{4.10}$$

for agencies $j = 1, 2,$ and 3.

Using the new variable $\texttt{NGBTU}_j^{01}$, I constructed the same logistic model used in Section 4.1.5 which utilizes several of the building descriptive variables to predict whether or not a building used natural gas as a source of energy. The output (parameter estimates and standard errors) from this model is given in Table 4.12, along with the synthetic estimates, associated standard errors and the confidence interval overlap percentages. As seen in the table, only half of the variables in the model have overlapping confidence intervals. The lack of fit for some variables is due to the method of synthetic data generation or simply that the relationships contained within this model were not transferred to the synthetic data.

Problems with the combined results, as compared to the original data, are most likely due to the choice of data generation method. Obviously, the tree models did

**Table 4.12**: Comparison of the Logistic Regression Models for the True and Synthetic Data.

| Variables | Combined Truth | Combined Synthetic | CI Coverage |
|---|---|---|---|
| Intercept | -2.141 (0.324) | -1.085 (0.175) | — |
| YRCON = 2 | 0.218 (0.277) | -0.012 (0.192) | 77.59% |
| YRCON = 3 | -0.085 (0.239) | 0.001 (0.132) | 77.55% |
| YRCON = 4 | -0.182 (0.233) | 0.001 (0.130) | 77.87% |
| YRCON = 5 | -0.432 (0.230) | -0.011 (0.128) | 43.57% |
| YRCON = 6 | -0.404 (0.229) | 0.008 (0.133) | 45.17% |
| YRCON = 7 | -0.877 (0.227) | -0.008 (0.131) | — |
| YRCON = 8 | -0.658 (0.253) | -0.004 (0.144) | 17.23% |
| YRCON = 9 | -0.699 (0.269) | -0.002 (0.058) | — |
| log(SQFT) | 0.205 (0.024) | 0.112 (0.014) | — |
| $\text{NWKER}^{1/2}$ | -0.015 (0.003) | 0.007 (0.003) | — |
| HEATP | 0.015 (0.001) | 0.000 (0.000) | — |

not preserve the original data structure and relationships in the synthetic replicates that are necessary for this logistic regression model. As seen in Table 4.10, the modeling of $\text{NGBTU}^{1/2}$ was problematic and produced suboptimal results. Therefore, using $\text{NGBTU}^{1/2}$ as the basis for the logistic regression only magnifies the problems experienced during the modeling phases. Since $\text{NGBTU}^{1/2}$ was generated near the end of the variables, any modeling problems in the early stages would have been carried through subsequent stages, increasing the lack of fit with each additional variable generated.

In retrospect, the method chosen to model NGBTU (or $\text{NGBTU}^{1/2}$) should have included two separate models. The first model would determine whether or not natural gas was used as a fuel source. If it was used, the second model would then impute a value for NGBTU based on the data already generated and the relationships within the original data. This two-fold process may be able to better maintain the relationships between NGBTU and the other variables, which would result in more optimal inferences.

# Chapter 5

# Conclusions and Future Research

This dissertation described two methods agencies can use to share and combine their data — even sensitive data — with one another based on the overall structure of the data to be shared. The resulting data sets can then be exchanged among participating agencies and released as public-use data. The use of synthetic data as the basis for these methods helps facilitate the data sharing process by allowing agencies to create data that closely resemble their original data without subjecting themselves, the data, and the respective respondents to unwarranted risks.

In either of the data sharing methods described, disclosure limitation is largely dependent on the agencies' ability to generate synthetic replicates that accurately model their data while protecting the original observations. It is imperative that agencies follow the prescribed sharing protocols; otherwise, they may be subjecting themselves and the other participating agencies to unnecessary risks and may even jeopardize the entire sharing process. If an agency were to halt the data sharing because it was exposing them and their data to too much risk, they (along with the other agencies) would lose out on the potential benefits that could result from sharing and combining data from multiple sources. In addition, they would not be able to pass these benefits on to users.

In the horizontally partitioned data sharing method of Chapter 2, disclosure limitation is the sole responsibility of the individual agencies because their resulting synthetic data sets are generated independently. Agencies do not have to rely on the ability of another agency to create viable synthetic data, only their own resources. The synthetic data sets that result from this sharing method can help give both the involved agencies and the users of public-use data a broader insight to the population in question.

By contrast, in the vertically partitioned data sharing method of Chapter 3, disclosure limitation depends largely on the initial generation of the disguisers for $X$ by the agency who begins the sharing process. (This is not the agency who creates groupings or clusters of $Y$.) This method subjects the agency creating disguisers to additional disclosure risks if the agency's true data are included among the set of disguisers that are sent to the other participating agency. The chance of a full disclosure increases if the set of disguisers does not preserve the structure and relationships within the data. However, if an agency can choose an imputation model, disclosure control method or combination thereof that captures both the essence of $X$ while protecting the original values, then the inclusion of the original data is not a necessary step (see Section 3.3.4 for more details). In this situation, the risk of a full disclosure has been reduced and the agency must only be concerned with protecting the sensitive data values from the other agency.

The end benefits are high for the two agencies participating in the vertically partitioned data sharing method. If they follow the protocols as outlined, create synthetic data that closely resemble the original data without revealing the actual values, and do not jeopardize the sharing process, then the end result will be multiple copies of a combined data set that originated from different sources. The possibilities for analyses and the ability to study new relationships should be motivation for the

agencies to attempt data sharing. Not only are they benefiting each other, they are also creating public-use data that can be used by analysts for inferences that ultimately may help improve society.

As with all synthetic data methods, the resulting samples and inferences are only as good as the imputation models used. If the models implemented are not well designed and do not contain the potential for all user analyses, inferences may be invalid and subject to additional biases. This may be avoidable if the distributions of imputations and relationships are compared with the original data for accuracy and modified if necessary. The implementation of these methods on real data is a difficult task and may require many available resources along with inventive methods. However, the resulting data can be analyzed by any user who can repeat their analysis on each synthetic data set and combine the results as described in this dissertation.

Several aspects of the data sharing methods presented in this dissertation could benefit from additional research. There were topics I did not address because they required extensions of the data sharing methods to incorporate other scenarios and variations in the data structure. Several other areas were studied — some only briefly while others to a fuller extent — that showed promising results. These results could be improved or expanded with further research that may simplify some of the steps of the sharing process or remove some of the assumptions on which the methods are based.

The components of this dissertation that could benefit from further research specifically involve the adjustment of negative variances (see (2.26) and (3.28)) and the derivation and calculation of the degrees of freedom used in the derived t-distributions (see Appendix A.1 and Appendix B.1). In terms of adjusting negative variances, alternatives may exist that more accurately quantify the actual variance within and among the synthetic samples. Both data sharing methods would benefit

from such research. The other aspects that deserve further research are the derived degrees of freedom expressions for both data sharing methods. In each case, assumptions had to be made during the actual derivation and subsequent calculation. If expressions could be found that do not involve underlying assumptions, then the inflation of the confidence interval coverages[1] could be reduced and more reliable results could be obtained.

Each data sharing method presented in this dissertation relies on assumptions made during the method description on which the entire method and derivations are based. For instance, each data sharing method assumes the data are fully observed. However, in real-world data collection this is an unattainable assumption because of the high probability of missing data. Therefore, it may be necessary to extend these methods to incorporate missing data scenarios. This would involve the addition of a new step prior to those already given and new combining rules and variance equation derivations.

The next potential research areas are only applicable to the vertically partitioned data sharing method. For instance, how could agencies go about sharing their data when one-to-one matching of the observations across data sets does not exist? Another possibility is the case when the agencies have data that are only partially overlapping. Each of these situations would require the addition of new steps to the sharing process to account for the additional uncertainty present within the data. The last area of research specific to vertically partitioned data is the case where more than two agencies want to participate in the data sharing process, the complexity of which may require inventive methods or routines to incorporate the additional data. Likewise, the security needed to protect the data and safeguard against the plotting of multiple agencies against others may require more extreme measures than seen in

[1]See the simulation output in Chapters 2 and 3.

this dissertation.

Further development and research of the disguising methods used in the first step of the vertically partitioned data sharing method may prove to be beneficial. Additional research could yield other disguising methods that offer greater flexibility in their ability to both disguise the actual data values while maintaining the relationships present within the data. In addition, it is necessary to generalize the data sharing process to incorporate all data types. For instance, development of disguising methods for categorical data are needed and may require the use of aspects commonly used in SDC methods (see Section 1.1).

Finally, an in-depth comparison of the use of tree models and generalized additive models (GAMs) (Hastie and Tibshirani, 1986, 1987, 1990) could result in more efficient methods of synthetic data generation. (Preliminary research suggests GAMs may be affective at modeling data like that in CBECS.) Additional research involving the implementation of the two generation methods could yield guidelines for the specification of parameters. This research may help determine which sampling method provides the best synthetic data when combined with a tree and the degree of smoothing necessary to model the data with GAMs.

# Appendix A

## A.1 Derivation of the Approximate Degrees of Freedom for Horizontally Partitioned Data

Inferences for fully synthetic, horizontally partitioned, shared data sets should be based on a t-distribution whenever $m_j$ and $n_{syn}$ are moderate. To find the approximate degrees of freedom, the mean and variance of

$$\frac{\nu_H T_H}{\frac{1}{a^2}\sum_{j=1}^{a}(1+1/m_j)(W_j+\bar{u}_j)-\frac{1}{a^2}\sum_{j=1}^{a}\bar{u}_j} \,|\, d^M, \tag{A.1}$$

approximated by a $\chi^2$ with $\nu_H$ degrees of freedom, must be matched to a Mean-Square distribution with mean one, variance $2/\omega_H$, and $\omega_H$ degrees of freedom. Let $x = \sum_{j=1}^{a}\frac{W_j+\bar{u}_j}{b_j}$ (see (2.17)) and will be assumed to be equivalent to $\frac{\sum_{j=1}^{a}W_j+\bar{u}_j}{\sum_{j=1}^{a}b_j} = \frac{W+\bar{u_M}}{b_M}$. Therefore, $x^{-1}|d^M$ will have a Mean-Square distribution with $f = \sum_{j=1}^{a}(m_j-1)$

degrees of freedom. Let $g = \frac{\frac{1}{a}\sum_{j=1}^{a}(1+1/m_j)b_j}{\bar{u}_M}$, which allows (A.1) to be written as

$$\begin{aligned}
\frac{T_H}{\frac{1}{a^2}\sum_{j=1}^{a}(1+1/m_j)(W_j+\bar{u}_j)-\frac{\bar{u}_M}{a}} &= \frac{\frac{1}{a^2}\sum_{j=1}^{a}(1+1/m_j)b_j-\frac{\bar{u}_M}{a}}{\frac{1}{a^2}\sum_{j=1}^{a}(1+1/m_j)(W_j+\bar{u}_j)-\frac{\bar{u}_M}{a}} \\
&= \frac{\frac{\bar{u}_M}{a}\left(\frac{\frac{1}{a^2}\sum_{j=1}^{a}(1+1/m_j)b_j}{\bar{u}_M/a}-1\right)}{\frac{\bar{u}_M}{a}\left(\frac{\frac{1}{a^2}\sum_{j=1}^{a}(1+1/m_j)(W_j+\bar{u}_j)}{\bar{u}_M/a}-1\right)} \\
&= \frac{\frac{\frac{1}{a}\sum_{j=1}^{a}(1+1/m_j)b_j}{\bar{u}_M}-1}{\frac{\frac{1}{a}\sum_{j=1}^{a}(1+1/m_j)(W_j+\bar{u}_j)}{\bar{u}_M}-1} \\
&= \frac{g-1}{gx-1} \sim \mathrm{MS}_{\omega_H} \qquad \text{(A.2)}
\end{aligned}$$

where $\omega_H = f(1-1/g)^2$.

To match the moments of the distributions, find the mean and variance of (A.2) using a first-order Taylor Series expansion of $z = x^{-1}$ around its expectation, $\mathbb{E}(z) = 1$

$$\frac{g-1}{g/z-1} = \frac{g-1}{g/1-1}+\frac{g(g-1)}{(g-1)^2}(z-1) = 1+\frac{g(z-1)}{g-1}. \qquad \text{(A.3)}$$

Given (A.3), the expectation of (A.2) is approximately

$$\mathbb{E}\left(\frac{g-1}{gx-1}\,\middle|\,d^M\right) \approx \mathbb{E}\left(1+\frac{g(z-1)}{g-1}\,\middle|\,d^M\right) \approx 1 \qquad \text{(A.4)}$$

as desired. Likewise, the variance of (A.2) is

$$\mathbb{V}\left(\frac{g-1}{gx-1}\,\middle|\,d^M\right) \approx \mathbb{V}\left(1+\frac{g(z-1)}{g-1}\,\middle|\,d^M\right) \approx \frac{2}{f}\left(\frac{g}{g-1}\right)^2. \qquad \text{(A.5)}$$

As desired, the variance of (A.2) is two divided by its associated degrees of freedom. These degrees of freedom correspond to the (approximate) degrees of freedom for the

t-distribution in (2.19)

$$\omega_H \;=\; f(1-1/g)^2 \;=\; \left(\sum_{j=1}^{a}(m_j-1)\right)\left(1-\frac{\bar{u}_M}{\frac{1}{a}\sum_{j=1}^{a}(1+1/m_j)b_j}\right)^2 \;=\; \nu_H. \tag{A.6}$$

## A.2 Derivation of the Weighted Estimator (2.5) and Associated Variance (2.6)

The derivations for the weighted estimator (2.5) and associated variance (2.6) essentially follow the derivations in Section 2.2. The evaluation of $f(Q|d^M)$ is similar to that in Section 2.2.4 and is detailed below

$$\begin{aligned}
\mathbb{E}(Q|d^M) &= \mathbb{E}\{\mathbb{E}(Q|P^M)|d^M\} \;=\; \mathbb{E}\{\mathbb{E}(\mathbb{E}(Q|\mathcal{D})|P^M)|d^M\} \\
&= \mathbb{E}\left\{\mathbb{E}\left(\frac{\sum_{j=1}^{a} q_{obs,j}/w_j}{\sum_{j=1}^{a}1/w_j}|P^M\right)|d^M\right\} \\
&= \mathbb{E}\left\{\frac{\sum_{j=1}^{a}\bar{Q}_j/q_j}{\sum_{j=1}^{a}1/w_j}|d^M\right\} \;=\; \frac{\sum_{j=1}^{a}\bar{q}_j/w_j}{\sum_{j=1}^{a}1/w_j} \;=\; \bar{q}_M^w \\
\mathbb{V}(Q|d^M) &= \mathbb{E}\{\mathbb{V}(Q|P^M)|d^M\}+\mathbb{V}\{\mathbb{E}(Q|P^M)|d^M\} \\
&= \mathbb{E}\left\{\mathbb{E}(\mathbb{V}(Q|\mathcal{D})|P^M)+\mathbb{V}(\mathbb{E}(Q|\mathcal{D})|P^M)|d^M\right\}+\mathbb{V}\left\{\mathbb{E}(\mathbb{E}(Q|\mathcal{D})|P^M)|d^M\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left(\frac{\sum_{j=1}^{a} v_{obs,j}/w_j^2}{(\sum_{j=1}^{a}1/w_j)^2}|P^M\right)+\mathbb{V}\left(\frac{\sum_{j=1}^{a} q_{obs,j}/w_j}{\sum_{j=1}^{a}1/w_j}|P^M\right)|d^M\right\} \\
&+ \mathbb{V}\left\{\mathbb{E}\left(\frac{\sum_{j=1}^{a} q_{obs,j}/w_j}{\sum_{j=1}^{a}1/w_j}|P^M\right)|d^M\right\} \\
&= \mathbb{E}\left\{\frac{\sum_{j=1}^{a} W_j/w_j^2}{(\sum_{j=1}^{a}1/w_j)^2}+\frac{\sum_{j=1}^{a}(W_j/m_j)/w_j^2}{(\sum_{j=1}^{a}1/w_j)^2}|d^M\right\}+\mathbb{V}\left\{\frac{\sum_{j=1}^{a}\bar{Q}_j/w_j}{\sum_{j=1}^{a}1/w_j}|d^M\right\} \\
&= \frac{\sum_{j=1}^{a}((1+1/m_j)b_j-\bar{u}_j)/w_j^2}{(\sum_{j=1}^{a}1/w_j)^2} \;=\; \frac{\sum_{j=1}^{a} w_j/w_j^2}{(\sum_{j=1}^{a}1/w_j)^2} \;=\; T_H^w.
\end{aligned}$$

When $n$ and $m_j$ are large, the posterior distribution of $Q|d^M$ will be approximately $N(\bar{q}_M^w, T_H^w)$.

## A.3 Extension to Partially Synthetic Data

As stated in Section 2.3, agencies may choose to only replace those values at high risk of disclosure, leaving the rest unaltered (Little, 1993; Kennickell, 1997; Abowd and Woodcock, 2001; Little and Liu, 2003; Reiter, 2003). This type of data replacement strategy yields partially synthetic data, as the resulting data contains a combination of real and synthetic values.

### A.3.1 Data Sharing Notation and Description

Let $D_j = (X, Y_{obs,j}, I_j)$ be the observed data for agency $j$, where $I_j$ is the agency-specific survey inclusion indicator, and $\mathcal{D} = \{D_j, j = 1, 2, \ldots, a\}$ the set of observed data for all $a$ agencies. Let $X$ and $Y_{obs,j}$ be as defined in Section 2.1. To create partially synthetic samples, agencies first determine which values in $Y_{obs,j}$ are high disclosure risks, because these values will be replaced with synthetic replicates. Using notation similar to Reiter (2003), let $Z_j = \{Z_j^{(\kappa)}, \kappa = 1, 2, \ldots, n_j\}$ be the set of replacement indicators for agency $j$, such that $Z_j^{(\kappa)} = 1$ if unit $\kappa$ is selected to have any of its observed data values replaced with synthetic ones and $Z_j^{(\kappa)} = 0$ if all unit values are to remain unchanged. Let $Y_{rep,j}^{(i)}$ be the synthetic replacements for agency $j$ generated from the agency-specific posterior predictive distribution $P(Y_{rep,j}^{(i)}|D_j, Z_j = 1)$. To avoid additional biases, this posterior predictive distribution should be based on only those units whose values are to be replaced (Reiter, 2003). Let $Y_{nrep,j}$ be the set of values in $Y_{obs,j}$ that remain constant in each of the partially synthetic samples $d_j^{(i)} = (X, Y_{rep,j}^{(i)}, Y_{nrep,j}, I_j, Z_j)$, $i = 1, 2, \ldots, m_j$, generated by agency $j$. Denote the

set of partially synthetic samples created by agency $j$ as $d_j = \{d_j^{(i)}, i = 1, 2, \ldots, m_j\}$, such that $d^M = \{d_j, j = 1, 2, \ldots, a\}$ is the set of partially synthetic samples for all $a$ agencies. The set $d^M$ can then be shared among the participating agencies and released for public use.

As previously done in Sections 2.1 and 2.2, the combining rules and variance equation for scalar estimates of $Q$ derived by Reiter (2003) must be extended to incorporate multiple agencies into the process. This extension allows valid inferences to be obtained when multiple agencies share horizontally partitioned, partially synthetic data, while also creating public-use data. Using the combining rules (see (2.1), (2.2), and (2.3)) from Section 2.1 allows $\bar{q}_M$ to once again be used as the estimate of $Q$, with

$$T_{HP} = \frac{1}{a^2} \sum_{j=1}^{a} b_j/m_j + \bar{u}_j \tag{A.7}$$

being the associated variance estimate. Whenever $n$ and $m_j$ are large, inferences for $Q$ can be found using the normal approximation $(Q - \bar{q}_M) \sim N(0, T_P)$, but when moderate, inferences for $Q$ should be based on a t-distribution with approximately $\nu_P = \left(\sum_{j=1}^{a}(m_j - 1)\right)\left(1 + r_P^{-1}\right)^2$, where $r_P = \frac{1}{\bar{u}_M}\sum_{j=1}^{a}\frac{b_j}{m_j a}$, degrees of freedom (see Appendix A.3.3 for details).

As in Section 2.1, agencies may want to incorporate potential differences between their observed data sets and generation methods into the joint estimation of $Q$ and $T_P$. To weight estimates with an agency-specific variability measure use (2.5), but let $w_j = b_j/m_j + \bar{u}_j$, which corresponds to the variance expression used in the partially synthetic data method (see Section 1.2.2 and (1.5)). The associated variance for the weighted estimator is the same as (2.6), except use the $w_j$ given above.

All further derivations assume an unweighted average in the joint estimation of the scalar $Q$.

### A.3.2 Bayesian Derivation of the Combining Rules

To derive the previously stated inferences from a Bayesian perspective, I use the theory developed in Reiter (2003) as a basis for the multiple-agency extension, where each observed data set originates from the same underlying population. It will be assumed that the same posterior distribution is used for both the imputations and inferences, allowing the posterior distribution of $Q|d^M$ to be decomposed as

$$f(Q|d^M) \;=\; \iint f(Q|\mathcal{D}, d^M, B_j) f(\mathcal{D}|d^M, B_j) f(B_j|d^M) d\mathcal{D}\, dB_j \qquad \text{(A.8)}$$

where $B_j = \mathbb{V}(q_j^{(i)}|D_j, Z_j = 1)$, and $\mathcal{D}$ and $d^M$ are as defined in Appendix A.3.1.

**Evaluating** $f(Q|\mathcal{D}, d^M, B_j)$

To evaluate the first conditional distribution in (A.8), I drop all references to the synthetic samples, as they are irrelevant given the observed data $\mathcal{D}$; therefore, $f(Q|\mathcal{D}, d^M, B_j) = f(Q|\mathcal{D})$. To find the distribution of $f(Q|\mathcal{D})$, follow the same reasoning outlined in Section 2.2.1, which yields (2.9).

**Evaluating** $f(\mathcal{D}|d^M, B_j)$

The next conditional distribution, $f(\mathcal{D}|d^M, B_j)$, will be evaluated in terms of the sufficient statistics of $\mathcal{D}$, which are $\bar{q}_{obs}$ and $\bar{v}_{obs}$, proceeding in a manner similar to the evaluations in Section 2.2.2. Thus, writing the conditional distribution in terms of the sufficient statistics yields $f(\bar{q}_{obs}, \bar{v}_{obs}|d^M, B_j) = f(\bar{q}_{obs}|d^M, B_j) f(\bar{v}_{obs}|d^M, B_j)$. As indicated by the factorization of the distribution, the sufficient statistics, $\bar{q}_{obs}$ and $\bar{v}_{obs}$, do not dependent on each other, which allows them to be found individually. To evaluate $f(\bar{q}_{obs}|d^M, B_j)$ and $f(\bar{v}_{obs}|d^M, B_j)$, I assume synthetic estimates are drawn

such that

$$q_j^{(i)}|D_j, B_j \sim N(q_{obs,j}, B_j) \tag{A.9}$$

$$u_j^{(i)}|D_j, B_j \sim (v_{obs,j}, << B_j) \tag{A.10}$$

for $i = 1, 2, \ldots, m_j$. (The notation $C \sim (F, << G)$ is used to denote a random variable $C$ that has some distribution with expectation $F$ and variance much less than $G$.) The distributional choice for $q_j^{(i)}$ should be appropriate given the observed data $D_j$ and $B_j$, because each partially synthetic data set should closely resemble the observed data. Likewise, estimates of the variance $u_j^{(i)}$ will have a distributional form centered near the observed estimate, $v_{obs,j}$; however, when $m_j$ is small, the center may be larger than $v_{obs,j}$, because fluctuations in the random draws may be more visible. In addition, the variance associated with $u_j^{(i)}$ will be smaller than that for $q_j^{(i)}$, as less variability occurs among variances than with estimates. Similar to Reiter (2003), a flat prior distribution is used on $q_{obs,j}$, such that standard Bayesian theory implies

$$q_{obs,j}|d_j, B_j \sim N(\bar{q}_j, B_j/m_j) \tag{A.11}$$

$$\frac{(m_j - 1)b_j}{B_j}|d_j \sim \chi^2_{m_j - 1} \tag{A.12}$$

where $b_j = \frac{1}{m_j - 1}\sum_{i=1}^{m_j}(q_j^{(i)} - \bar{q}_j)^2$. Similarly, a flat prior distribution on $v_{obs,j}$ gives

$$v_{obs,j}|d_j, B_j \sim (\bar{u}_j, << B_j/m_j). \tag{A.13}$$

If (A.11) and (A.13) are averaged over $j$, the distributions of $\bar{q}_{obs}$ and $\bar{v}_{obs}$ become

$$\bar{q}_{obs}|d^M, B_j \sim N\left(\bar{q}_M, \frac{1}{a^2}\sum_{j=1}^{a} B_j/m_j\right) \tag{A.14}$$

$$\bar{v}_{obs}|d^M, B_j \sim \left(\bar{u}_M, << \frac{1}{a^2}\sum_{j=1}^{a} B_j/m_j\right). \tag{A.15}$$

**Evaluating $f(Q|d^M)$**

To evaluate $f(Q|d^M)$, multiply (2.9), (A.14), (A.15), and (A.12), then integrate with respect to $\bar{q}_{obs}$, $\bar{v}_{obs}$ and $B_j$. Combining the conditional distributions of each component into an integral yields

$$\begin{aligned} f(Q|d^M) &= \iiint N(Q; \bar{q}_{obs}, \bar{v}_{obs}/a)\, N\left(\bar{q}_{obs}; \bar{q}_M, \frac{1}{a^2}\sum_{j=1}^{a} B_j/m_j\right) \\ &\times \left(\bar{v}_{obs}; \bar{u}_M, << \frac{1}{a^2}\sum_{j=1}^{a} B_j/m_j\right) \\ &\times \left(\prod_{j=1}^{a} \text{Inverse } \chi^2(B_j; m_j - 1, b_j)\right) d\bar{q}_{obs}\, d\bar{v}_{obs}\, dB_j. \end{aligned} \tag{A.16}$$

The evaluation of (A.16) requires numerical integration, because a closed-form expression cannot be explicitly written. Otherwise, the expression can be evaluated from a Bayesian perspective, such that each parameter is simulated from its respective conditional distribution and then averaged accordingly. However, a method is desired that can easily be implemented by all users; therefore, an approximation to the posterior distribution of $Q$ is needed. If $m_j$ is large, expectation approximations can be used in the variance terms, simplifying the posterior evaluation. For instance, $\mathbb{E}(B_j|d_j) \approx b_j$ can be used in place of $\mathbb{E}(B_j|d_j) = \left(\frac{m_j}{m_j-2}\right) b_j$. This allows the approximation of the estimate of $Q$ to be given by

$$\begin{aligned} \mathbb{E}(Q|d^M) &= \mathbb{E}\{\mathbb{E}(Q|d^M, B_j)|d^M\} = \mathbb{E}\{\mathbb{E}(\mathbb{E}(Q|\mathcal{D})|d^M, B_j)|d^M\} \\ &= \mathbb{E}\{\mathbb{E}(\bar{q}_{obs}|d^M, B_j)|d^M\} = \mathbb{E}\{\bar{q}_M|d^M\} = \bar{q}_M. \end{aligned}$$

Similarly, the associated variance estimate becomes

$$\begin{aligned}
\mathbb{V}(Q|d^M) &= \mathbb{E}\{\mathbb{V}(Q|d^M, B_j)|d^M\} + \mathbb{V}\{\mathbb{E}(Q|d^M, B_j)|d^M\} \\
&= \mathbb{E}\{\mathbb{E}(\mathbb{V}(Q|\mathcal{D})|d^M, B_j) + \mathbb{V}(\mathbb{E}(Q|\mathcal{D})|d^M, B_j)|d^M\} \\
&+ \mathbb{V}\{\mathbb{E}(\mathbb{E}(Q|\mathcal{D})|d^M, B_j)|d^M\} \\
&= \mathbb{E}\{\mathbb{E}(\bar{v}_{obs}/a|d^M, B_j) + \mathbb{V}(\bar{q}_{obs}|d^M, B_j)|d^M\} + \mathbb{V}\{\mathbb{E}(\bar{q}_{obs}|d^M, B_j)|d^M\} \\
&= \mathbb{E}\left\{\frac{\bar{u}_M}{a} + \frac{1}{a^2}\sum_{j=1}^{a} B_j/m_j|d^M\right\} + \mathbb{V}\{\bar{q}_M|d^M\} \\
&= \frac{1}{a^2}\sum_{j=1}^{a}\bar{u}_j + \mathbb{E}\{B_j|d^M\}/m_j \quad = \quad \frac{1}{a^2}\sum_{j=1}^{a}\bar{u}_j + b_j/m_j.
\end{aligned}$$

The estimate of $Q$ and associated variance, along with the prescribed integration, suggest the posterior distribution of $Q|d^M$ should be

$$Q|d^M \sim t_{\nu_P}\left(\bar{q}_M, \frac{1}{a^2}\sum_{j=1}^{a}\bar{u}_j + b_j/m_j\right) \tag{A.17}$$

with approximately $\nu_P$ degrees of freedom, as derived in Appendix A.3.3. To find the actual degrees of freedom, fit the mean and variance of $\frac{1}{a^2}\sum_{j=1}^{a}\bar{u}_j + B_j/m_j$ to a Mean-Square random variable as in Rubin (1987, pp. 90–92).

### A.3.3 Derivation of the Approximate Degrees of Freedom

As stated in Appendices A.3.1 and A.3.2, inferences for partially synthetic, horizontally partitioned, shared data sets should be based on a t-distribution when $m_j$ and $n$ are moderate. The approximate degrees of freedom for this t-distribution are found by matching the mean and variance of

$$\frac{\nu_P T_P}{\frac{1}{a^2}\sum_{j=1}^{a} B_j/m_j + \frac{1}{a^2}\sum_{j=1}^{a}\bar{u}_j}\,|\,d^M, \tag{A.18}$$

approximated by a $\chi^2$ with $\nu_P$ degrees of freedom, to a Mean-Square distribution with mean one, variance $2/\omega_P$, and $\omega_P$ degrees of freedom. Let $x = \sum_{j=1}^{a} \frac{B_j}{b_j}$ (see (A.12)) and will be assumed to be equivalent to $\frac{\sum_{j=1}^{a} B_j}{\sum_{j=1}^{a} b_j} = \frac{B}{b_M}$. Therefore, $x^{-1}|d^M$ will have a Mean-Square distribution with $f = \sum_{j=1}^{a}(m_j - 1)$ degrees of freedom. Let, $g = \frac{\frac{1}{a}\sum_{j=1}^{a} b_j/m_j}{\bar{u}_M}$, which allows (A.18) to be written as

$$\begin{aligned}
\frac{T_P}{\frac{1}{a^2}\sum_{j=1}^{a} B_j/m_j + \frac{\bar{u}_M}{a}} &= \frac{\frac{1}{a^2}\sum_{j=1}^{a} b_j/m_j + \frac{\bar{u}_M}{a}}{\frac{1}{a^2}\sum_{j=1}^{a} B_j/m_j + \frac{\bar{u}_M}{a}} \\
&= \frac{\frac{\bar{u}_M}{a}\left(\frac{\frac{1}{a^2}\sum_{j=1}^{a} b_j/m_j}{\bar{u}_M/a} + 1\right)}{\frac{\bar{u}_M}{a}\left(\frac{\frac{1}{a^2}\sum_{j=1}^{a} B_j/m_j}{\bar{u}_M/a} + 1\right)} \\
&= \frac{1 + \frac{\frac{1}{a}\sum_{j=1}^{a} b_j/m_j}{\bar{u}_M}}{1 + \frac{\frac{1}{a}\sum_{j=1}^{a} B_j/m_j}{\bar{u}_M}} \\
&= \frac{1+g}{1+gx} \sim \mathrm{MS}_{\omega_P} \qquad \text{(A.19)}
\end{aligned}$$

where $\omega_P = f(1 + 1/g)^2$.

To match the moments, find the mean and variance of (A.19), using a first-order Taylor Series expansion of $z = x^{-1}$ around its expectation, $\mathbb{E}(z) = 1$

$$\frac{1+g}{1+g/z} = \frac{1+g}{1+g/1} + \frac{g(1+g)}{(1+g)^2}(z-1) = 1 + \frac{g(z-1)}{1+g}. \qquad \text{(A.20)}$$

Given (A.20), the expectation of (A.19) is approximately

$$\mathbb{E}\left(\frac{1+g}{1+gx}\,|d^M\right) \approx \mathbb{E}\left(1 + \frac{g(z-1)}{1+g}\,|d^M\right) \approx 1 \qquad \text{(A.21)}$$

as desired. Likewise, the variance of (A.19) is

$$\mathbb{V}\left(\frac{1+g}{1+gx}\,|d^M\right) \approx \mathbb{V}\left(1 + \frac{g(z-1)}{1+g}\,|d^M\right) \approx \frac{2}{f}\left(\frac{g}{1+g}\right)^2. \qquad \text{(A.22)}$$

As desired, the variance of (A.19) is two divided by its associated degrees of freedom, which are also the approximate degrees of freedom for the t-distribution in (A.17)

$$\omega_P \;=\; f(1+1/g)^2 \;=\; \left(\sum_{j=1}^{a}(m_j-1)\right)\left(1+\frac{\bar{u}_M}{\frac{1}{a}\sum_{j=1}^{a} b_j/m_j}\right)^2 \;=\; \nu_P. \tag{A.23}$$

# Appendix B

## B.1 Derivation of the Approximate Degrees of Freedom for Vertically Partitioned Data

When $k$, $m$, and $n$ are moderate, inferences for $Q$ should be based on a t-distribution with $\nu_V$ degrees of freedom. To derive $\nu_V$, the mean and variance of

$$\frac{\nu_V T_V}{\frac{1}{k}(\bar{u}_M/m + H) + (\bar{u}_M + W) - \bar{u}_M}|d^M, \tag{B.1}$$

approximated by a $\chi^2$ with $\nu_V$ degrees of freedom, must be matched to a Mean Square distribution with mean one, variance $2/\omega_V$, and $\omega_V$ degrees of freedom. Let $\alpha = \frac{\bar{u}_M/m+H}{B_k}$ (see (3.20)) and let $\gamma = \sum_{j=1}^{k} \frac{\bar{u}_j+W_j}{b_j}$ (see (3.19)) which will be approximated by $\gamma^* = \frac{\bar{u}_M+W}{\bar{b}_M}$, such that $\alpha^{-1}|d^M$ and $\gamma^{*-1}|d^M$ have Mean Square distributions with $k-1$ and $k(m-1)$ degrees of freedom, respectively. Let $f = \frac{B_k/k}{\bar{u}_M}$ and $g = \frac{\bar{b}_M}{\bar{u}_M}$, which

allows (B.1) to be written as

$$\begin{aligned}
\frac{T_V}{\frac{1}{k}(\bar{u}_M/m + H) + (\bar{u}_M + W) - \bar{u}_M} &= \frac{B_k/k + \bar{b}_M - \bar{u}_M}{\frac{1}{k}(\bar{u}_M/m + H) + (\bar{u}_M + W) - \bar{u}_M} \\
&= \frac{B_k/k + \bar{b}_M - \bar{u}_M}{\frac{1}{k}(\bar{u}_M/m + H)\frac{B_k}{B_k} + (\bar{u}_M + W) - \bar{u}_M} \\
&= \frac{B_k/k + \bar{b}_M - \bar{u}_M}{\alpha(B_k/k) + (\bar{u}_M + W)\frac{\bar{b}_M}{b_M} - \bar{u}_M} \\
&= \frac{B_k/k + \bar{b}_M - \bar{u}_M}{\alpha(B_k/k) + \gamma^*\bar{b}_M - \bar{u}_M} \\
&= \frac{\bar{u}_M\left(\frac{B_k/k}{\bar{u}_M} + \frac{\bar{b}_M}{\bar{u}_M} - 1\right)}{\bar{u}_M\left(\alpha\left(\frac{B_k/k}{\bar{u}_M}\right) + \gamma^*\left(\frac{\bar{b}_M}{\bar{u}_M}\right) - 1\right)} \\
&= \frac{\frac{B_k/k}{\bar{u}_M} + \frac{\bar{b}_M}{\bar{u}_M} - 1}{\alpha\left(\frac{B_k/k}{\bar{u}_M}\right) + \gamma^*\left(\frac{\bar{b}_M}{\bar{u}_M}\right) - 1} \\
&= \frac{f + g - 1}{\alpha f + \gamma^* g - 1} \quad \sim \quad \mathrm{MS}_{\omega_V}. \qquad \text{(B.2)}
\end{aligned}$$

To match the moments of the distributions, the mean and variance of (B.2) must found. These moments will be calculated using a first order Taylor Series expansion of $z = \alpha^{-1}$ and $x = \gamma^{*-1}$ around their expectations, which both equal one

$$\begin{aligned}
\frac{f + g - 1}{f/z + g/x - 1} &= \frac{f + g - 1}{f/1 + g/1 - 1} + \frac{g(f + g - 1)}{(f + g - 1)^2}(x - 1) + \frac{f(f + g - 1)}{(f + g - 1)^2}(z - 1) \\
&= 1 + \frac{g(x - 1) + f(z - 1)}{f + g - 1}. \qquad \text{(B.3)}
\end{aligned}$$

Using (B.3), the expectation of (B.2) is approximately

$$\mathbb{E}\left(\frac{f + g - 1}{\alpha f + \gamma^* g - 1} \mid d^M\right) \approx \mathbb{E}\left(1 + \frac{g(x - 1) + f(z - 1)}{f + g - 1} \mid d^M\right) \approx 1 \quad \text{(B.4)}$$

as desired. Similarly, the variance of (B.2) is

$$\begin{aligned}\mathbb{V}\left(\frac{f+g-1}{\alpha f+\gamma^* g-1} \mid d^M\right) &\approx \mathbb{V}\left(1+\frac{g(x-1)+f(z-1)}{f+g-1} \mid d^M\right) \\ &\approx \frac{2\left(\frac{g^2}{k(m-1)}+\frac{f^2}{k-1}\right)}{(f+g-1)^2} \qquad \text{(B.5)}\end{aligned}$$

which is of the form $2/\omega_V$. Therefore, the appropriate degrees of freedom for the t-distribution is

$$\omega_V = \frac{(f+g-1)^2}{\frac{g^2}{k(m-1)}+\frac{f^2}{k-1}} = \frac{\left(\frac{B_k/k}{\bar{u}_M}+\frac{\bar{b}_M}{\bar{u}_M}-1\right)^2}{\frac{\left(\frac{\bar{b}_M}{\bar{u}_M}\right)^2}{k(m-1)}+\frac{\left(\frac{B_k/k}{\bar{u}_M}\right)^2}{k-1}} = \nu_V. \qquad \text{(B.6)}$$

### B.1.1 Asymptotic Comparison of (B.6) to the Partially Synthetic Degrees of Freedom

The degrees of freedom for partially synthetic data (Reiter, 2003) given in Section 1.2.2 are $\nu_P = (r-1)\left(1+\frac{\bar{u}_r}{b_r/r}\right)^2$. This expression can also be written as

$$\nu_P = (r-1)\left(\frac{b_r/r+\bar{u}_r}{b_r/r}\right)^2 \qquad \text{(B.7)}$$

which highlights the variance expression for partially synthetic data (numerator of the second quantity of (B.7)). Using this expression for comparative purposes, it is necessary to show that when $m \to \infty$, $\nu_V$ is equivalent to $\nu_P$.

As $m \to \infty$, the first term in the denominator of (B.6) goes to zero, as indicated

below

$$\begin{aligned}
\lim_{m\to\infty} \nu_V &= \frac{\left(\frac{B_k/k}{\bar{u}_M} + \frac{\bar{b}_M}{\bar{u}_M} - 1\right)^2}{\frac{\left(\frac{B_k/k}{\bar{u}_M}\right)^2}{k-1}} \\
&= \frac{(k-1)\left(\frac{B_k/k}{\bar{u}_M} + \frac{\bar{b}_M}{\bar{u}_M} - 1\right)^2}{\frac{(B_k/k)^2}{(\bar{u}_M)^2}} \\
&= \frac{(k-1)(\bar{u}_M)^2\left(B_k/k + \bar{b}_M - \bar{u}_M\right)^2}{\frac{(B_k/k)^2}{(\bar{u}_M)^2}} \\
&= \frac{(k-1)(B_k/k + \bar{b}_M - \bar{u}_M)^2}{(B_k/k)^2} \\
&= (k-1)\left(\frac{B_k/k + \bar{b}_M - \bar{u}_M}{B_k/k}\right)^2 \\
&= (k-1)\left(\frac{b_k/k + \bar{u}_M}{b_k/k}\right)^2. \qquad \text{(B.8)}
\end{aligned}$$

The next four lines in the above expression detail the arrange of terms until a form similar to (B.7) is achieved, with the last completing the limit expression. If $m \to \infty$, then an infinite number of fully synthetic samples are created from a single partially synthetic data set, $c_j$. Thus allowing the variance within and between the partially synthetic samples $c^k$ to be accurately estimated. In particular, when $m \to \infty$, $B_k/k$ is estimating the same quantity as $b_r/r$ in the partially synthetic case. Likewise, $\bar{b}_M - \bar{u}_M$ estimates $\bar{u}_r$, which results in both the variance expression for partially synthetic data generation and its associated degrees of freedom.

# Appendix C

## C.1 Generating Disguisers for Real Data

The joint distribution table of PBA and YRCON for the CBECS data is given in Table C.1. In Section 4.1, it was assumed that PBA and YRCON were safe to release in their present form due to their ability to help match the records in $\mathbf{X}$ with those in $\mathbf{Y}$. Even though some combinations of the two variables have low levels of occurrence, I assumed the building identities were safe given that the values for SQFT, NWKER, COOLP, HEATP, and LTOHRP were modified through various disguising schemes.

## C.2 Investigation of Tree Models for Generating Fully Synthetic Data

The generation of fully synthetic data based on the data designated to agencies 2 and 3, proceeds similar to the methods described in Section 4.2.1 for the data designated to agency 1. Therefore, I will omit the specific generation details due to the similarities between the three data sets. Included in this appendix are the directed graphs illustrating the relationships significant in the generation of the synthetic replicates based on the data designated to agencies 2 and 3 and their respective joint distribution tables of PBA and YRCON.

**Table C.1**: Joint Distribution of PBA and YRCON for the 1995 CBECS public-use data.

| | Year Constructed Category | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| PBA | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total |
| 1 | 6 | 11 | 18 | 19 | 11 | 14 | 11 | 1 | 2 | 93 |
| 2 | 26 | 45 | 120 | 115 | 156 | 239 | 421 | 74 | 28 | 1224 |
| 4 | 0 | 1 | 3 | 12 | 12 | 11 | 7 | 3 | 2 | 51 |
| 5 | 7 | 16 | 79 | 106 | 124 | 155 | 191 | 40 | 22 | 740 |
| 6 | 0 | 1 | 9 | 13 | 17 | 22 | 15 | 6 | 8 | 91 |
| 7 | 5 | 9 | 14 | 13 | 17 | 20 | 17 | 6 | 8 | 109 |
| 8 | 2 | 0 | 2 | 10 | 10 | 16 | 14 | 8 | 6 | 68 |
| 11 | 0 | 0 | 3 | 3 | 6 | 1 | 7 | 2 | 2 | 24 |
| 12 | 16 | 16 | 15 | 59 | 53 | 43 | 39 | 6 | 4 | 251 |
| 13 | 11 | 18 | 53 | 50 | 67 | 63 | 52 | 22 | 18 | 354 |
| 14 | 8 | 27 | 100 | 149 | 169 | 112 | 71 | 21 | 36 | 693 |
| 15 | 5 | 10 | 18 | 15 | 23 | 31 | 30 | 6 | 3 | 141 |
| 16 | 1 | 1 | 10 | 33 | 38 | 75 | 33 | 8 | 6 | 205 |
| 17 | 0 | 1 | 4 | 6 | 19 | 16 | 12 | 0 | 4 | 62 |
| 18 | 6 | 10 | 17 | 41 | 94 | 42 | 78 | 18 | 9 | 315 |
| 23 | 0 | 3 | 12 | 39 | 56 | 66 | 81 | 15 | 10 | 282 |
| 24 | 1 | 1 | 0 | 4 | 20 | 28 | 39 | 5 | 2 | 100 |
| 25 | 16 | 38 | 47 | 58 | 56 | 84 | 75 | 29 | 19 | 422 |
| 26 | 8 | 8 | 37 | 47 | 63 | 65 | 63 | 19 | 11 | 321 |
| 91 | 2 | 2 | 6 | 4 | 5 | 8 | 14 | 2 | 2 | 45 |
| Total | 120 | 218 | 567 | 796 | 1016 | 1111 | 1269 | 291 | 202 | 5590 |

### C.2.1 Agency 2

The directed graph illustrating the significant relationships in the data designated to agency 2, is similar to Figure 4.11 (connecting link arrangement is slightly different). Many of the relationships linking the building description variables with the energy consumption and expenditure variables are still present, but with some small modifications. There are several relationships present in Figure 4.11 that do not appear prominent in this data set and vice versa. As with the other graph, the dashed lines represent those relationships not present in the generation models, but added due to dependencies within the data.

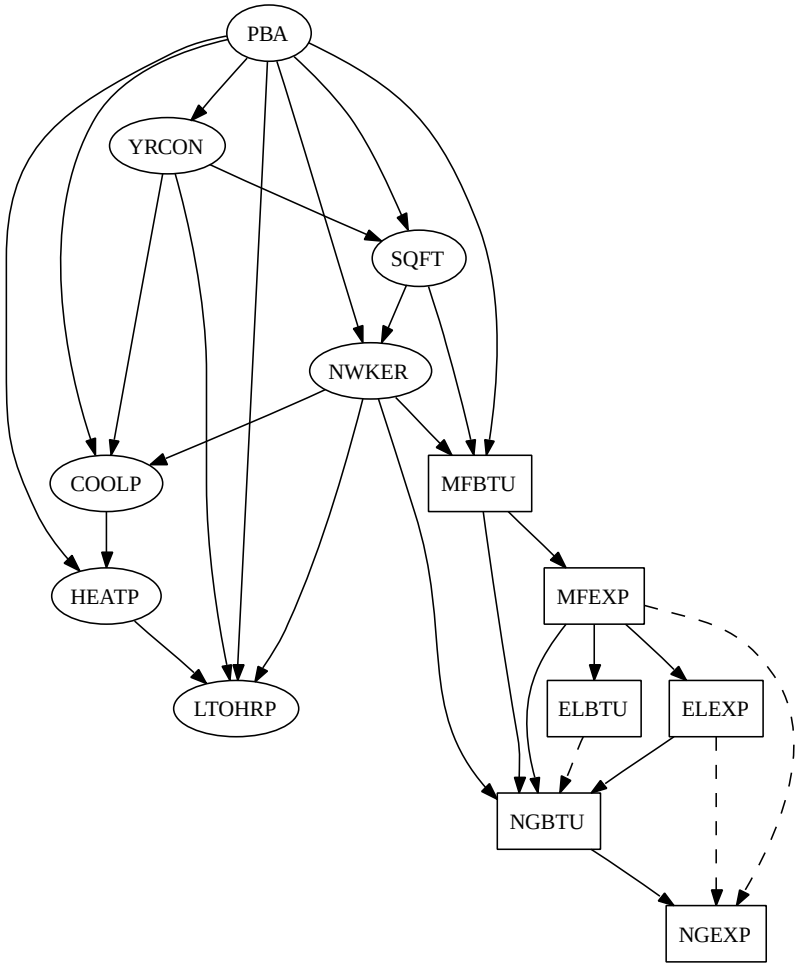


**Figure C.1**: Graphical representation of the relationships in the data designated to agency 2 and used to generate the synthetic data.

Similar to Table 4.9, the joint distribution of `PBA` and `YRCON` for the data designated to agency 2, shows combinations of the two variables that have low levels of occurrences, and in some cases no observations. During the generation of synthetic replicates for `YRCON`, I set the corresponding $\alpha$ value equal to zero to any of the combinations that do not have any observations presents (see Generating `YRCON` | `PBA` in Section 4.2.1). By forcing the probability of these combinations to zero, no unobserved combinations are introduced into the synthetic data sets.

**Table C.2**: Joint Distribution of PBA and YRCON for the data designated to agency 2.

| | Year Constructed Category | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| PBA | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total |
| 1 | 1 | 4 | 6 | 5 | 2 | 7 | 2 | 0 | 1 | 28 |
| 2 | 9 | 13 | 35 | 37 | 61 | 76 | 158 | 19 | 14 | 422 |
| 4 | 0 | 1 | 1 | 5 | 3 | 2 | 4 | 1 | 0 | 17 |
| 5 | 4 | 5 | 29 | 39 | 47 | 58 | 64 | 15 | 7 | 268 |
| 6 | 0 | 0 | 0 | 5 | 4 | 5 | 3 | 2 | 3 | 22 |
| 7 | 3 | 1 | 6 | 7 | 4 | 4 | 5 | 4 | 4 | 38 |
| 8 | 1 | 0 | 2 | 4 | 1 | 9 | 5 | 2 | 3 | 27 |
| 11 | 0 | 0 | 2 | 1 | 1 | 1 | 5 | 0 | 1 | 11 |
| 12 | 6 | 6 | 5 | 21 | 21 | 22 | 10 | 3 | 1 | 95 |
| 13 | 7 | 5 | 18 | 15 | 26 | 16 | 20 | 8 | 5 | 120 |
| 14 | 2 | 7 | 38 | 56 | 68 | 42 | 23 | 7 | 18 | 261 |
| 15 | 3 | 6 | 9 | 6 | 6 | 6 | 10 | 2 | 1 | 49 |
| 16 | 0 | 1 | 2 | 14 | 16 | 28 | 12 | 2 | 2 | 77 |
| 17 | 0 | 1 | 1 | 2 | 8 | 7 | 3 | 0 | 2 | 24 |
| 18 | 3 | 4 | 8 | 12 | 33 | 12 | 29 | 3 | 2 | 106 |
| 23 | 0 | 1 | 3 | 13 | 27 | 25 | 24 | 6 | 3 | 102 |
| 24 | 0 | 1 | 0 | 3 | 4 | 12 | 13 | 1 | 1 | 35 |
| 25 | 7 | 16 | 18 | 17 | 19 | 33 | 30 | 10 | 7 | 157 |
| 26 | 3 | 2 | 13 | 16 | 28 | 28 | 20 | 7 | 7 | 124 |
| 91 | 1 | 0 | 2 | 1 | 1 | 6 | 5 | 1 | 0 | 17 |
| Total | 50 | 74 | 198 | 279 | 380 | 399 | 445 | 93 | 82 | 2000 |

### C.2.2 Agency 3

As with the the data designated to the first two agencies, I created a directed graph from the data designated to agency 3 to give a visual representation of the relationships in the data that were used for synthetic data generation. As seen in Figure C.2, there are very connections between the building description variables and the energy consumption and expenditure variations, but more than in the graphs for the previous two data sets. Once again, I added dashed lines to the graph to indicate the dependencies inherent in the data that may not have appeared significant in the generation of the synthetic data.

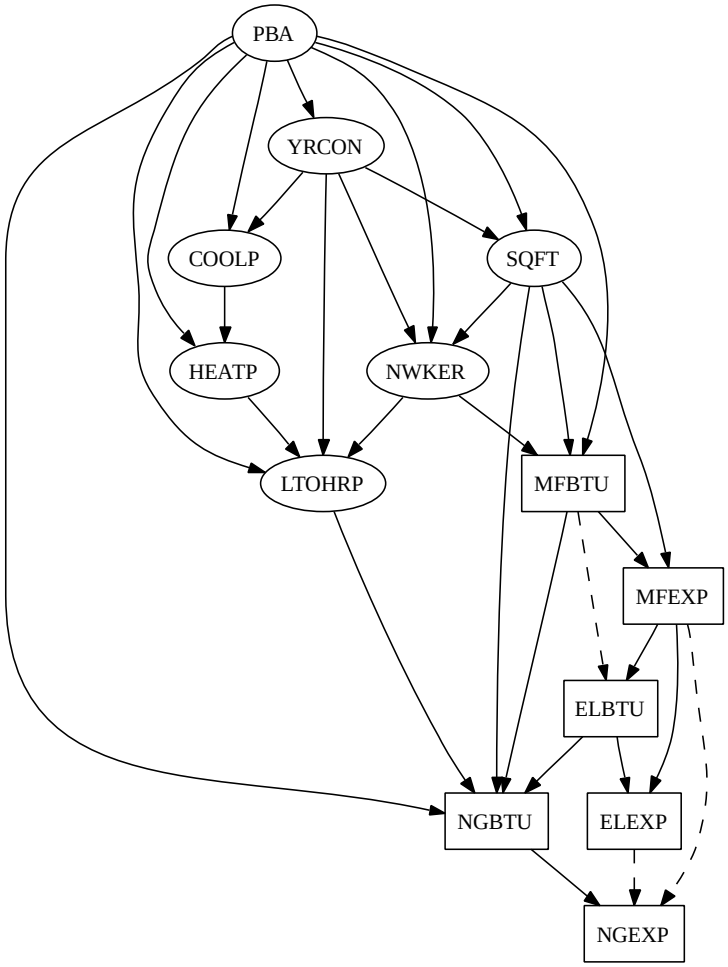


**Figure C.2**: Graphical representation of the relationships in the data designated to agency 3 and used to generate the synthetic data.

Table C.3 gives the joint distribution of `PBA` and `YRCON` for the data designated to agency 3. It is evident from the table that some combinations did not occur in the data, while other occur only with a low frequency. These combinations may be at high risk of disclosure, however, measures were taken during the synthetic data generation to help reduce this risk.

Comparing the joint distribution of `PBA` and `YRCON` (see Table C.3) with the previous constructed tables, fewer variable combinations have low observed frequency counts. This is primarily due to the larger number of records designated to agency 3 from the original data set than were given to the other two agencies.

**Table C.3**: Joint Distribution of PBA and YRCON for the data designated to agency 3.

| | Year Constructed Category | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| PBA | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total |
| 1 | 2 | 3 | 10 | 9 | 6 | 2 | 3 | 1 | 1 | 37 |
| 2 | 8 | 24 | 39 | 45 | 51 | 93 | 161 | 31 | 10 | 462 |
| 4 | 0 | 0 | 1 | 6 | 4 | 5 | 3 | 1 | 1 | 21 |
| 5 | 3 | 7 | 28 | 43 | 38 | 55 | 80 | 13 | 7 | 274 |
| 6 | 0 | 1 | 8 | 5 | 12 | 10 | 8 | 2 | 3 | 49 |
| 7 | 1 | 3 | 6 | 3 | 8 | 7 | 8 | 0 | 2 | 38 |
| 8 | 1 | 0 | 0 | 3 | 4 | 4 | 5 | 4 | 3 | 24 |
| 11 | 0 | 0 | 0 | 1 | 3 | 0 | 1 | 1 | 1 | 7 |
| 12 | 4 | 3 | 6 | 22 | 17 | 10 | 15 | 1 | 3 | 81 |
| 13 | 3 | 8 | 23 | 21 | 30 | 24 | 15 | 9 | 5 | 138 |
| 14 | 3 | 11 | 32 | 55 | 57 | 43 | 31 | 6 | 11 | 249 |
| 15 | 2 | 4 | 4 | 8 | 10 | 15 | 11 | 3 | 1 | 58 |
| 16 | 0 | 0 | 6 | 12 | 15 | 31 | 7 | 3 | 3 | 77 |
| 17 | 0 | 0 | 2 | 2 | 6 | 4 | 4 | 0 | 2 | 20 |
| 18 | 1 | 3 | 7 | 19 | 35 | 18 | 33 | 9 | 5 | 130 |
| 23 | 0 | 1 | 4 | 22 | 21 | 20 | 33 | 6 | 7 | 114 |
| 24 | 1 | 0 | 0 | 1 | 11 | 8 | 11 | 0 | 1 | 33 |
| 25 | 6 | 12 | 16 | 29 | 26 | 31 | 20 | 9 | 6 | 155 |
| 26 | 2 | 2 | 15 | 15 | 21 | 23 | 21 | 6 | 2 | 107 |
| 91 | 1 | 1 | 3 | 1 | 3 | 1 | 5 | 0 | 1 | 16 |
| Total | 38 | 83 | 210 | 322 | 378 | 404 | 475 | 105 | 75 | 2290 |

# Bibliography


Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, 215–277. Amsterdam: North-Holland.

Benaloh, J. (1987). Secret sharing homomorphisms: Keeping shares of a secret sharing. In A. M. Odlyzko, ed., *CRYPTO86, LNCS 263*. Springer-Verlag.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. California: Duxbury Press.

Dandekar, Ramesh, D., Cohen, M., and Kirkendall, N. (2002). Sensitive micro data protection using latin hypercube sampling technique. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases, LNCS 2316*, 117–125. Berlin: Springer-Verlag.

Duncan, G. T., Keller-McNulty, S. A., and Lynne, S. S. (2004). Disclosure risks vs. data utility: The r-u confidentiality map. Tech. rep., National Institute of Statistical Sciences, Technical Report 142.

Duncan, G. T. and Sumitra, M. (2000). Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. *Journal of the American Statistical Association* **95**, 720–729.

Fienberg, Stephen, E. (1994). Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality. *Journal of Official Statistics* **10**, 115–132.

Fienberg, Stephen, E. and Willenborg, Leon, C. R. J. (1998). Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data. *Journal of Official Statistics* **14**, 337–345.

Fuller, W. A. (1993). Masking Procedures for Microdata Disclosure Limitation. *Journal of Official Statistics* **9**, 2, 383–406.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. London: Chapman & Hall.

Gomatam, S. and Karr, A. F. (2003). Distortion measures for categorical data swapping. Tech. rep., National Institute of Statistical Sciences, Technical Report 131.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science* **1**, 297–318.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association* **82**, 371–386.

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

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data cluster: A review. *ACM Computing Surveys* **31**, 3, 264–323.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2005). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. *submitted* .

Karr, A. F., Lin, X., Reiter, J. P., and Sanil, A. P. (2004). Secure regression on distributed databases. *Journal of Computational and Graphical Statistics,* forthcoming .

Keller-McNulty, S. and Unger, E. A. (1993). Database Systems: Inferential Security. *Journal of Official Statistics* **9**, 475–499.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. Washington, D.C.: National Academy Press.

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

Little, R. J. A. and Liu, F. (2003). Selective multiple imputation of keys for statistical disclosure control in microdata. Tech. rep., The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 6.

Liu, F. (2003). *Bayesian Methods for Statistical Disclosure Control in Microdata*. Ph.D. thesis, The University of Michigan Department of Biostatistics.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science* **9**, 538–573.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

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.

Reiter, J. P. (2002). Satisfying Disclosure Restrictions with Synthetic Data Sets. *Journal of Official Statistics* **18**, 531–543.

Reiter, J. P. (2003). Inferences for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology* **29**, 181–188.

Reiter, J. P. (2004). Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. *Survey Methodology* **30**, 2, 235–242.

Reiter, J. P., Kohnen, C. N., Karr, A. F., Lin, X. L., and Sanil, A. (2004). Secure regression for vertically partitioned, partially overlapping data. In *ASA Proceedings of the Joint Statistical Meetings*.

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: John Wiley & Sons.

Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics* **9**, 462–468.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* **57**, 3–18.

Sanil, A. P., Karr, A. F., Lin, X., and Reiter, J. P. (2004a). Privacy preserving analysis of vertically partitioned data using secure matrix products. *Journal of Official Statistics,* submitted for publication .

Sanil, A. P., Karr, A. F., Lin, X., and Reiter, J. P. (2004b). Privacy preserving regression modelling via distributed computations. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Särndal, C., Swensson, B., and Wretman, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

Schneier, B. (1996). *Applied Cryptography: Protocols, Algorithms, and Source Code in C*. New York: John Wiley & Sons.

Shen, Z. (2000). *Nested Multiple Imputation*. Ph.D. thesis, Harvard University, Department of Statistics.

Weisberg, S. (1985). *Applied Linear Regression*. New York: John Wiley & Sons.

Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*. New York: Springer-Verlag.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Zayatz, L., Massell, P., and Steel, P. (1999). Disclosure limitation practices and research at the U.S. Census Bureau. *Special issue on statistical disclosure control. Netherlands Official Statistics* **14**, 26–29.

# Biography

Christine Noelle Kohnen was born the twentieth of December, 1977 in Waconia, Minnesota. She received a B.A. in mathematics with distinction from Saint Olaf College, Northfield, Minnesota in 2000. In 2002, She earned a M.S. in statistics from the Institute of Statistics and Decision Sciences, Duke University. Christine has co-authored the following articles:

1. Kohnen, C. N. and Reiter, J. P. (2004). "Sharing Confidential Data Among Multiple Agencies Using Multiply, Imputed Synthetic Data." *2004 Proceedings of the American Statistical Association.*

2. Reiter, J. P., Kohnen, C. N., Karr, A. F., Lin, X. L. and Sanil A. (2004). "Secure Regression for Vertically Partitioned, Partially Overlapping Data." *2004 Proceedings of the American Statistical Association.*

3. Reiter, J. P. and Kohnen, C. N. (2005). "Categorical data regression diagnostics for remote servers." *Journal of Statistical Computation and Simulation.*