Lecture Notes in Statistics 201

Synthetic Datasets for Statistical Disclosure Control

Theory and Implementation

Springer

Lecture Notes in Statistics **201**

Edited by P. Bickel, P.J. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

Jörg Drechsler

Synthetic Datasets for Statistical Disclosure Control

Theory and Implementation

Springer

Jörg Drechsler
Department for Statistical Methods
Institute for Employment Research
Regensburger Straße 104
90478 Nürnberg
Germany
Joerg.Drechsler@iab.de

ISSN 0930-0325
ISBN 978-1-4614-0325-8 e-ISBN 978-1-4614-0326-5
DOI 10.1007/978-1-4614-0326-5
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2011931290

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my mother and my father (in loving memory) for their love and support

Foreword

The topic of Jörg Drechsler's work is, in my view, extremely important because it addresses two conflicting demands that are becoming ever more important and complex with the increasing sophistication of our society. First, there is the demand to have access to the vast amounts of publicly supported data collected on all of us. Second, there is the demand to preserve the confidentiality of critical information about individuals in the data being released.

For a specific example of the first demand, in the United States there is the recent call to use the vast collection of medical data, routinely collected on patients from hospitals, pharmacies, etc., to conduct "comparative effectiveness research" in order to find the best combination of medical treatments for individuals. The search for answers to such questions, and therefore the request for publicly available microdata, i.e., data on individuals, is legitimate. Nevertheless, the release of such data threatens the privacy of patients.

The second demand, therefore, is for any released data to preserve confidential information from the individuals whose data are being released, whether because of explicit or implicit guarantees made to them. Even the release of one piece of confidential information can have relatively dire consequences when combined with publicly available information. For another U.S. example, with a person's name and birth date, both of which are available essentially to anyone, all an "intruder" needs is a social security number (taxpayer number) to open credit card accounts, obtain loans, charge hospital bills, open Internet and cell phone accounts, etc. – with all records and debts attached to that social security number. The result is that the holder of that social security number can have a disastrous credit rating that is essentially impossible to correct, even after thousands of dollars in expenses and many years of trying. This "stolen identity" problem is just one example of the untoward effects of the release of confidential information, which may include life-altering consequences, such as being denied mortgages on home purchases.

The work that Jörg Drechsler is pursuing in this book addresses both demands by trying to find ways to benefit society by releasing microdata, here multiply imputed synthetic microdata, that simultaneously preserve individuals' confidential information and yet allow valid inferences at some level of detail through the use of

specialized methods for combining the analyses of the resulting multiply imputed datasets. The topic is a statistically challenging one that needs much development, and I'm sure that this book will be a critical stimulus to this development. Jörg is to be congratulated for this great contribution.

Cambridge, Massachusetts, March 2011 *D. B. Rubin*

Acknowledgements

This book would never have been possible without the help of many colleagues and friends, and I am very grateful for their wonderful support. First, I want to thank my Ph.D. advisor, Susanne Rässler, for introducing me to the world of multiple imputation and suggesting I join a research project on synthetic data at the Institute for Employment Research (IAB) that finally became the cornerstone of my thesis and eventually of this book. Her remarkable enthusiasm helped me pass some of the local minima of my dissertation function, and without her I would never have met and eventually worked with some of the greatest researchers in the field.

I am very grateful to Trivellore Raghunathan for joining my dissertation committee and providing helpful suggestions for the revision of my thesis for this book. Although I only had two weeks during a visit at the University of Michigan to benefit from his expertise, I learned a lot in that short period of time and I am still deeply impressed by his ability to grasp complex research problems within seconds but even more importantly by his capacity to instantly come up with often simple and straightforward solutions for seemingly (at least for me) unsolvable problems.

I also want to thank John Abowd for inviting me to participate in weekly videoconferences with the leading experts on synthetic data in the United States. When I started my research, I was the only one involved in that topic in Europe and following the discussions and learning from the experience of these experts during these weekly meetings was extremely helpful for my endeavor. To Don Rubin, one of the founding fathers of synthetic data for data confidentiality, I am thankful for his invitation to present my work at Harvard and for fruitful discussions on some of my papers on the topic that later found their way into my thesis. I feel especially honored that he accepted writing the foreword for this book. Bill Winkler deserves my gratitude for providing the extensive list of references on microdata confidentiality included in the appendix of this book. John Kimmel and Marc Strauss at Springer provided great support while I worked on turning my thesis into an acceptable contribution for the Springer Lecture Notes in Statistics Series. Anne-Sophie Charest contributed very helpful comments on an earlier version of this book.

At the IAB I am especially thankful to Hans Kiesl, Thomas Büttner, and Stefan Bender. Hans always helped me out when my lack of background in survey statis-

tics once again became too obvious. Thomas joined me in the dissertation journey. It was a great relief to have a fellow sufferer. And both of them provided helpful discussions on the details of multiple imputation and unforgettable road trips framing JSMs and other conferences around the world. Stefan was very supportive of my research from the very beginning. He stood up for my work when others were still merely laughing at the idea of generating synthetic datasets, even though he was and probably still is skeptical about the idea himself. He helped me find my way in the jungle of official statistics and assisted me in any way he could.

My deepest gratitude is to Jerry Reiter, with whom I had the pleasure to work on several projects that later became part of my thesis. Chapters 6, 7, and especially Chapter 9 in this book borrow heavily from joint papers that were a direct result of these projects. Almost everything I know on the theoretical concepts behind synthetic datasets I owe to Jerry. He has been and continues to be a great mentor and friend.

Finally, I want to thank my mother, Ursula Drechsler, her partner Jochen Paschedag, and the rest of my family for their wonderful support and care. Even though spending three years developing fake data must have seemed bizarre to them, they were always interested in the progress of my work and helped me whenever they could. Most importantly, I would never have survived this trip without the constant love of my fiancee, Veronika. There is no way I can thank her enough for all her patience and understanding for numerous weekends and evenings I spent in front of the computer. She always cheered me up when deadlines were approaching surprisingly fast and the simulations still didn't provide the results they were supposed to show. I thank her for bringing more colors to my life.

Nürnberg, April 2011 *Jörg Drechsler*

Contents

Acronyms

FCS	fully conditional specification
GSSD	German Social Security Data
MAR	missing at random
MCAR	missing completely at random
MISD	multiply imputed synthetic datasets
MNAR	missing not at random
PUMS	public use microdata samples
RMSE	root mean squared error
SDC	statistical disclosure control
SDL	statistical disclosure limitation
SMIKE	selective multiple imputation of keys
SRMI	sequential regression multivariate imputation

List of Figures

List of Tables

Chapter 1
Introduction

National statistical institutes (NSIs) such as the U.S. Census Bureau or the German Federal Statistical Office gather valuable information on many different aspects of society. Broad access to this information is desirable to stimulate research in official statistics. However, most data obtained by the institutes are collected under the pledge of privacy, and thus the natural interest in enabling as much research as possible with the collected data has to take a back seat to the confidentiality guaranteed to the survey respondent. But not only legal aspects are relevant when considering disseminating data to the public. Respondents who feel their privacy is at risk might be less willing to provide sensitive information, might give incorrect answers, or might even refuse to participate completely – with devastating consequences for the quality of the data collected (Lane, 2007). Traditionally, this meant that access to the data was strictly limited to researchers working for the NSI. With the increasing demand for access to the data on the micro-level from external researchers, accelerated by the improvements in computer technology, agencies started looking for possibilities to disseminate data that provide a high level of data quality while still guaranteeing confidentiality for the participating units.

Over the years, a broad body of literature on statistical disclosure limitation (SDL) techniques for microdata has evolved (see Bill Winkler's famous list of microdata confidentiality references in the Appendix A). These techniques can be divided into two main categories: approaches that protect the data by reducing the amount of information contained in the released file through coarsening of the data and approaches classified as data perturbation methods that try to maintain most of the originally collected information but protect the data by changing some of the values on the micro-level. Information-reducing approaches protect the data by

- *categorizing continuous variables*: building categories from the underlying continuous variables and reporting only the category in which the unit falls; for example, building age groups in five-year intervals.
- *top coding*: setting values above a certain threshold equal to the threshold; for example, reporting the income for all individuals with income above $100,000 as "100,000+."

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_1,

- *coarsening categorical variables*: coarsening to a reduced number of categories; for example, instead of providing information on the state level, only reporting whether a respondent lives in West or East Germany.
- *dropping variables*: dropping some variables that are considered too sensitive (e.g., HIV status) or are not protected enough by any of the methods above.

There is much literature on data perturbation methods, and discussing all approaches, including possible modifications, is beyond the scope of this introduction. A detailed overview is given in the *Handbook on Statistical Disclosure Control* (Center of Excellence for Statistical Disclosure Control, 2009), issued by members of the CENEX-SDC project funded by Eurostat. Good references for recent developments are the proceedings from the biannual conference *Privacy in Statistical Databases* (Springer LNCS 3050, 4302, 5262, 6344).

While the first methods developed in the 1980s, such as swapping and adding noise, mainly focused on disclosure protection and preserved only some univariate statistics such as the population mean and the variance of a single variable, more sophisticated methods have emerged in recent years. But these sophisticated methods often require different complicated adjustments for each estimate to get unbiased results, preserve only certain statistics, such as the vector of the means or the variance-covariance matrix, or are valid only under specific distributional assumptions, such as multivariate normality, that are unrealistic for real datasets. Besides, most statistical agencies still only apply standard methods, mainly because of their ease of implementation. Winkler (2007b) shows the devastating consequences on data quality for many of these easy-to-implement procedures and the remaining procedures often fail to achieve their primary goal: protecting the data adequately.

Since many of the proposed data perturbation methods significantly reduce data quality and it is often impossible for the researcher using the perturbed data to judge whether the results are still at least approximately valid, there is a common mistrust of these methods among researchers. Still, strict legal requirements in many countries often force agencies to perturb their data before release, even though they know that data quality can be heavily affected.

The situation is a little different in Germany, where the required disclosure protection for datasets used only for scientific purposes (so-called *scientific use files*) is lower than for datasets that are available to anybody (*public use files*). For scientific use files, the German Federal Law on Statistics enables the release of de facto anonymous microdata. “Factual anonymity means that the data can be allocated to the respondent or party concerned only by employing an excessive amount of time, expenses and manpower” (Knoche, 1993). The concept of factual anonymity takes into account a rational thinking intruder who calculates the costs and benefits of the reidentification of the data. Because factual anonymity depends on several conditions and is not further defined by law, it is necessary to estimate the costs and benefits of a reidentification for every dataset with a realistic scenario. Disseminating scientific use files under this law is much easier than under the usual requirement that a reidentification of a single unit should be impossible under any circumstances. For this reason, the scientific use files available in Germany traditionally are protected using only a mixture of the nonperturbative methods described above. Nevertheless,

there is common agreement that the dissemination of microdata on businesses is not possible using only nonperturbative methods since the risk of disclosure is much higher for these data than it is for microdata on individuals for several reasons:

- The underlying population is much smaller for businesses than it is for individuals.
- Variables such as turnover or establishment size have very skewed distributions that make the identification of single units in the dataset very easy.
- There is a lot of information about businesses in the public domain already. This information can be used to identify records in the released dataset.
- The benefit from identifying a unit in an establishment survey might be higher for a potential attacker than the benefit of identifying a unit in a household survey.
- In most business surveys, the probability of inclusion is very high for large businesses (often close to 1), so there is no additional privacy protection from sampling for these units.

Since only a few variables such as turnover, region, and industry code, are necessary to identify many businesses, no data on enterprises were disseminated for many years. In 2002, a joint project of the German Federal Statistical Office, several Statistical Offices of the Länder, and the Institute for Applied Economic Research started investigating the possibility of generating scientific use files for these data by applying perturbative methods for the first time in Germany. They came to the conclusion that a release is possible using these methods and disseminated several survey datasets protected by either adding multiplicative noise or microaggregation (Statistisches Bundesamt, 2005). With the long history of releasing only unperturbed data, it is not surprising that acceptance of these datasets was rather limited in the following years. Many users of these data tend to believe the collected data is the direct truth and ignore all the additional uncertainty and possible bias introduced at the collection stage by measurement errors, coding mistakes, bad sampling design and especially steadily increasing nonresponse rates. The additional bias introduced by the perturbation method might be dwarfed by the bias already inherent in the data due to these facts. But also the selected perturbation methods might be a reason for the limited acceptance. Winkler (2007b) illustrates the negative consequences of univariate microaggregation, namely on correlations, and although correction factors for estimations based on data perturbed by multiplicative noise are illustrated in the German *Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten* (Statistisches Bundesamt, 2005) for the linear model and the SIMEX method (Lechner and Pohlmeier, 2005) can be used for nonlinear models, both are difficult to compute and are applicable only under some additional assumptions. The *Handbuch* shows that the SIMEX method produces biased results for a probit regression using simulated data. A further disadvantage the two methods share with most data perturbation methods is that logical constraints between variables are not preserved.

This illustrates the common dilemma for data disseminating agencies: fulfilling only one goal – no risk of disclosure or high data quality – is straightforward; release data generated completely at random or release the original unchanged data. In both cases, at least one party will be unhappy with the results, but balancing the two goals

is extremely difficult. A dataset that guarantees the confidentiality of the respondent but is not accepted by the research community due to data quality concerns is of little value, and the question arises whether the high costs in time and money to produce these datasets are justified.

A new approach to address the trade-off between data utility and disclosure risk that overcomes the problems discussed above was proposed by Rubin (1993): the release of multiply imputed synthetic datasets (MISDs). Specifically, he proposed that agencies (i) randomly and independently sample units from the sampling frame to comprise each synthetic dataset, (ii) impute unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) release multiple versions of these datasets to the public. These are called fully synthetic datasets.

However, the quality of this method strongly depends on the accuracy of the model used to impute the "missing" values. If the model doesn't include all the relationships between the variables that are of interest to the analyst or if the joint distribution of the variables is misspecified, results from the synthetic datasets can be biased. Furthermore, specifying a model that considers all the skip patterns and constraints between the variables in a large dataset can be cumbersome, if not impossible. To overcome these problems, a related approach suggested by Little (1993) replaces observed values with imputed values only for variables that bear a high risk of disclosure or for variables that contain especially sensitive information, leaving the rest of the data unchanged. This approach, discussed as generating partially synthetic datasets in the literature, has been adopted for some datasets in the United States (Abowd and Woodcock, 2001, 2004; Kennickell, 1997; Abowd et al., 2006).

The aim of this book is to give the reader a detailed introduction to the different approaches to generating multiply imputed synthetic datasets by combining the theory with illustrative examples using a real dataset, the German IAB Establishment Panel. I start by giving an overview of the history of synthetic datasets and discussing the major advantages of this approach compared with other perturbation methods. Since the method is based on the ideas of multiple imputation (Rubin, 1978), the next chapter recapitulates its basic concepts originally proposed to impute values missing due to nonresponse. Advantages and disadvantages of the two major imputation strategies (joint modeling and fully conditional specification (FCS)) are also addressed.

The Chapters 5–9 on different multiple imputation approaches for nonresponse and synthetic data generation, are all organized in the same manner. First the general ideas of the specific approach are discussed and then the point and variance estimates that provide valid inferences in this context are presented. Each section concludes with an extensive application to a real dataset. Since all applications are based on the German IAB Establishment Panel, this dataset is introduced in a separate chapter at the beginning of the main part of the book (Chapter 4). The multiple imputation approaches discussed include imputation for nonresponse (Chapter 5), generating fully synthetic datasets (Chapter 6), generating partially synthetic datasets (Chapter 7), and generating synthetic datasets when the original data are subject to nonresponse (Chapter 8).

Chapter 9 contains an extension to the standard synthetic data generation to better address the trade-off between data utility and disclosure risk, imputation in two stages, where variables that drive the disclosure risk are imputed less often than others. Since, in general, data quality and disclosure risk both increase with the number of imputations, defining a different number of imputations for different variables can lead to datasets that maintain the desired data quality with reduced risk of disclosure. In this chapter, the new combining procedures that are necessary for the point and variance estimates are presented for fully and partially synthetic datasets, and the IAB Establishment Panel is used to illustrate the impact of the number of imputations on the data quality and the disclosure risk and to show the possible advantage of using a two stage imputation approach. The book concludes with a glimpse into the future of synthetic datasets, discussing the potentials and possible obstacles of the approach and ways to address the concerns of data users and their understandable discomfort with using data that don't consist only of the originally collected values.

Chapter 2
Background on Multiply Imputed Synthetic Datasets

2.1 The history of multiply imputed synthetic datasets

In 1993, the *Journal of Official Statistics* published a special issue on data confidentiality. Two articles in this volume laid the foundation for the development of multiply imputed synthetic datasets (MISDs). In his discussion "Statistical Disclosure Limitation," Rubin (1993) for the first time suggested generating synthetic datasets based on his ideas of multiple imputation for missing values (Rubin, 1987). He proposed to treat all the observations from the sampling frame that are not part of the sample as missing data and to impute them according to the multiple imputation framework. Afterwards, simple random samples from these fully imputed datasets should be released to the public. Because the released dataset does not contain any real data, disclosure of sensitive information is very difficult. On the other hand, if the imputation models are selected carefully and the predictive power of the models is high, most of the information contained in the original data will be preserved. This approach is now called generating fully synthetic datasets in the literature.

In the same issue, Little (1993) suggested a closely related approach that is also based on the idea of replacing sensitive information by multiple imputation. The major difference is that only part of the data are replaced. The replaced data could either be some sensitive variables, such as income or turnover, or key variables such as age, place of birth, and sex that could be jointly used to identify a single unit in the dataset. With this approach, now called generating partially synthetic datasets, it is not mandatory to replace all units for one variable. The replacement can be tailored only to the records at risk. It might be sufficient for example to replace the income only for units with a yearly income above 100,000 euros to protect the data. This method guarantees that only those records that need to be protected are altered. Leaving unchanged values in the dataset will generally lead to higher data quality, but releasing unchanged values obviously poses a higher risk of disclosure.

Fienberg (1994) proposed a related approach for data confidentiality. He suggested generating synthetic datasets by bootstrapping from a "smoothed" estimate

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_2,

of the empirical cumulative density function of the survey data. This approach was further developed for categorical data in Fienberg et al. (1998).

Ten years after the initial proposal, the complete theory for deriving valid inferences from multiply imputed synthetic datasets was presented for the first time. Raghunathan et al. (2003) illustrated why the standard combining procedures for multiple imputation (Rubin, 1987) are not valid in this context and developed the correct procedures for fully synthetic datasets. The procedures for partially synthetic datasets were presented by Reiter (2003). One year earlier, Liu and Little (2002) had suggested the selective multiple imputation of key variables (SMIKe), replacing a set of sensitive and nonsensitive cases by multiple draws from their posterior predictive distribution under a general location model.

Reiter also demonstrated the validity of the fully synthetic combining procedures under different sampling scenarios (Reiter, 2002), derived the combining procedures when using multiple imputation for missing data and for disclosure avoidance simultaneously (Reiter, 2004), developed significance tests for multi-component estimands in the synthetic data context (Reiter, 2005c; Kinney and Reiter, 2010), provided an empirical example for fully synthetic datasets (Reiter, 2005b), and presented a nonparametric imputation method based on CART models to generate synthetic data (Reiter, 2005d). Recently he compared CART models with imputation models based on random forests (Caiola and Reiter, 2010). Further work includes suggestions for the adjustment of survey weights (Mitra and Reiter, 2006), selecting the number of imputations when using multiple imputation for missing data and disclosure control (Reiter, 2008b), measuring the risk of identity disclosure for partially synthetic datasets (Reiter and Mitra, 2009; Drechsler and Reiter, 2008), a two-stage imputation strategy to better address the trade-off between data utility and disclosure risk (Reiter and Drechsler, 2010), and an alternative approach for generating public use microdata samples (PUMS) from Census data called sampling with synthesis (Drechsler and Reiter, 2010).

A new imputation strategy based on kernel density estimation for variables with very skewed or even multimodal distributions has been suggested by Woodcock and Benedetto (2009), while Winkler (2007a) proposed the use of different EM algorithms to generate synthetic data subject to convex constraints. The attractive features of synthetic datasets are further discussed by Fienberg and Makov (1998); Abowd and Lane (2004); Little et al. (2004); An and Little (2007), and Domingo-Ferrer et al. (2009).

It took several years before the groundbreaking ideas proposed in 1993 were ever applied to any real dataset. The U.S. Federal Reserve Board was the first agency to protect data in its Survey of Consumer Finances by replacing monetary values at high risk of disclosure with multiple imputations, releasing a mixture of these imputed values and the unreplaced, collected values (Kennickell, 1997). Abowd and Woodcock (2001) illustrated the possibilities of protecting longitudinal, linked datasets with data from the French National Institute of Statistics and Economic Studies (INSEE). A very successful implementation of a partially synthetic dataset is the data behind *On the Map*, illustrating commuting patterns (i.e., where people live and work) for the entire United States via maps available to the public on the

Web (http://lehdmap.did.census.gov/). Since the point of origin (where people live) is already in the public domain, only the destination points are synthesized. Machanavajjhala et al. (2008) developed a sophisticated synthesizer that maximizes the level of data protection based on the ideas of differential privacy (Dwork, 2006) while still guaranteeing a very high level of data utility. The most ambitious synthetic data project to date is the generation of a public use file for the Survey of Income and Program Participation (SIPP) funded by the U.S. Census Bureau and the Social Security Administration (SSA). The variables from the SIPP are combined with selected variables from the Internal Revenue Service's (IRS) lifetime earnings data and the SSA's individual benefit data. Almost all of the approximately 625 variables contained in this longitudinal, linked dataset were synthesized. In 2007, four years after the start of the project, a beta version of the file was released to the public (www.sipp.census.gov/sipp/synthdata.html). Abowd et al. (2006) summarize the steps involved in creating this public use file and provide a detailed disclosure risk and data utility evaluation that indicates that confidentiality is guaranteed while data utility is high for many estimates of interest.

The Census Bureau also protects the identities of people in group quarters (e.g., prisons, shelters) in the public use files of the American Community Survey by replacing demographic data for people at high disclosure risk with imputations. The latest release of a synthetic data product by the Census Bureau is a synthetic version of the Longitudinal Business Database (Kinney et al., 2011) that is available as a public use dataset through the VirtualRDC's Synthetic Data Server located at Cornell University (http://www.vrdc.cornell.edu/news/data/lbd-synthetic-data/). Partially synthetic, public use datasets are in the development stage in the U.S. for the Longitudinal Employer–Household Dynamics survey and the American Community Survey veterans and full sample data. Recently, a statement by the American Statistical Association on data access and personal privacy explicitly mentioned distributing synthetic datasets as an appropriate method of disclosure control (http://www.amstat.org/news/statementondataaccess.cfm).

Outside the U.S., the ideas for generating multiply imputed synthetic datasets were ignored for many years, except for some small simulation studies at ISTAT in Italy (Polettini, 2003; Franconi and Stander, 2002, 2003; Polettini et al., 2002). They suggest generating model-based synthetic datasets. The main difference from the methods described in this book is that they do not propose multiple imputation and therefore do not correct for the additional variance from imputation. In 2006, the German Institute for Employment Research launched a research project to generate synthetic datasets of its longitudinal establishment survey for release as a scientific use file. In the first phase of the project, the fully and partially synthetic approaches were tested on a subset of the data (Drechsler et al., 2008b,a; Drechsler and Reiter, 2009). Drechsler et al. (2008a) also discuss the advantages and disadvantages of the two approaches in terms of data utility and disclosure risk. Since the evaluations during the first stage of the project indicated that the dataset could be sufficiently protected by the partial synthetic approach, the second stage of the project focused on the generation of a partially synthetic dataset for the complete 2007 wave of the

survey. This dataset, the first outside the U.S., was released in 2011. The growing interest in synthetic datasets in Europe is also documented by the report on synthetic data files requested by Eurostat 2008 and published by Domingo-Ferrer et al. (2009). Outside Europe, statistical agencies in Australia, Canada, and New Zealand (Graham and Penny, 2005; Graham et al., 2009) also are investigating the approach.

2.2 Advantages of multiply imputed synthetic datasets compared with other SDC methods

MISDs provide a number of advantages over other methods that are discussed in the statistical disclosure control (SDC) literature

First, the aim of any SDC method should be to preserve the joint distribution of the data. But most data perturbation methods either preserve only univariate statistics or some predefined multivariate statistics such as the mean and the variance-covariance matrix in previously defined subgroups. However, SDC methods are used to generate datasets for public release on the microdata level, and it is impossible to anticipate all analyses potential users will perform with the data. For example, one analyst might remove some outliers before running her regressions, and it is completely unclear what the effects of SDC methods that only preserve statistics in predefined subsets of the data will be for this reduced dataset. Besides, for some analyses it might be desirable to preserve more than just the first two moments of the distribution (e.g., maintain interaction and nonlinear effects).

Second, many SDC methods are only applicable either to categorical variables or continuous variables. This means that often a combination of different techniques is required to fully protect a dataset before release. Methods based on multiple imputation, on the other hand, can be applied to categorical and continuous variables likewise, rendering the use of different methods that might require different adjustments by the data analyst unnecessary.

Third, most of the data collected by agencies are subject to nonresponse, and besides the fact that missing data can lead to biased estimates if not treated correctly by the analyst, many SDC methods cannot be applied to datasets containing missing values. Since generating multiply imputed synthetic datasets is based on the ideas of multiple imputation for handling item nonresponse in surveys, it is straightforward to impute missing values before generating synthetic datasets. Reiter (2004) developed methods for simultaneous use of multiple imputation for missing data and disclosure limitation.

Fourth, model-based imputation procedures offer more flexibility if certain constraints need to be preserved in the data. For example, non-negativity constraints and linear constraints such as *total number of employees* $\geq$ *number of part-time employees* can be directly incorporated at the model-building stage. Almost all SDC methods fail to preserve linear constraints unless the exact same perturbation is applied to all variables for one unit, which in turn significantly increases the risk of disclosure.

Fifth, skip patterns (e.g. a battery of questions are only asked if they are applicable) are very common in surveys. Especially if the skip patterns are hierarchical, it is very difficult to guarantee that perturbed values are consistent with these patterns. With the fully conditional specification approach (see also Section 3.1.2) that sequentially imputes one variable at a time by defining conditional distributions to draw from, it is possible to generate synthetic datasets that are consistent with all these rules.

Lastly, as Reiter (2008a) points out, the MI approach can be relatively transparent to the public analyst. Metadata about the imputation models can be released, and the analyst can judge based on this information whether the analysis he or she seeks to perform will give valid results with the synthetic data. For other SDC approaches, it is very difficult to decide how much a particular analysis has been distorted.

On the other hand, as with any perturbation method, limited data utility is a problem of synthetic data. Only the statistical properties explicitly captured by the model used by the data protector are preserved. A logical question at this point is, why not directly publish the statistics one wants to preserve rather than release a synthetic micro-dataset? Possible defenses against this argument are:

- Synthetic data are normally generated by using more information on the original data than is specified in the model whose preservation is guaranteed by the data protector releasing the synthetic data.
- As a consequence of the above, synthetic data may offer utility beyond the models they explicitly preserve.
- Not all users of a public use file will have a sound background in statistics. Some of the users might only be interested in some descriptive statistics and won't be able to generate the results if only the parameters are provided.
- The imputation models in most applications can be very complex because different models are fitted for every variable and often for different subsets of the dataset. This might lead to hundreds of parameters just for one variable. Thus, it is much more convenient even for the skilled user of the data to have the synthesized dataset available.
- The most important reason for not releasing the parameters is that the parameters themselves could be disclosive on some occasions. For that reason, only some general statements about the generation of the public use file should be released. For example, these general statements could provide information about which variables were included in the imputation model but not the exact parameters. So the user can judge whether his analysis would be covered by the imputation model, but he will not be able to use the parameters to disclose any confidential information.

Chapter 3
Background on Multiple Imputation[1]

Multiple imputation, introduced by Rubin (1978) and discussed in detail in Rubin (1987; 2004), is an approach that retains the advantages of imputation while allowing the uncertainty due to imputation to be directly assessed. Originally developed for the imputation of missing values in surveys, the approach can also be applied to generate synthetic datasets (Rubin, 1993; Little, 1993) for high-quality data dissemination without compromising the confidentiality of the survey respondents. With multiple imputation, the missing or sensitive values in a dataset are replaced by $m > 1$ simulated versions, generated according to a probability distribution for new values given the observed data. Thus the general aim is to generate replacement values by multiply drawing from $P(Y_{new}|Y_{obs})$, where Y_{new} represents either values that are initially not observed (in the missing-data context) or values that should be replaced by imputed values (in the data confidentiality context).

Each of the imputed datasets is first analyzed by standard methods designed for complete data, and the results of the m analyses are then combined to produce estimates, confidence intervals, and test statistics that properly reflect the uncertainty from imputation. As pointed out in Chapter 1, the combining rules for the final estimates differ depending on the context for which multiple imputation is used, and each of the following chapters in this book will start by presenting the correct combining rules for the relevant setting.

The main aim of this chapter is to discuss some of the issues that are generally important for multiple imputation regardless of whether the aim is to handle nonresponse or to address confidentiality concerns. I start by introducing the two main approaches for multiple imputation, joint modeling and sequential regression, with a discussion of their advantages and disadvantages. Then I present some adjustments for standard multiple imputation routines to handle problems that often arise with real data.

[1] Parts of this chapter are taken from Drechsler and Rässler (2008) and Drechsler (2011a).

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_3,

3.1 Two general approaches to generate multiple imputations

Over the years, two different methods have emerged to generate draws from $P(Y_{new}|Y_{obs})$: joint modeling and fully conditional specification (FCS), often also referred to as sequential regression multivariate imputation (SRMI) or chained equations. The first assumes that the data follow a specific distribution (e.g., a multivariate normal distribution). Under this assumption, a parametric multivariate density $P(Y|\theta)$ can be specified with θ representing parameters from the assumed underlying distribution. Within the Bayesian framework, this distribution can be used to generate draws from $(Y_{new}|Y_{obs})$. Methods to create multivariate imputations using this approach have been described in detail by Schafer (1997) (e.g., for the multivariate normal, the log-linear, and the general location model).

Fully conditional specification (van Buuren and Oudshoorn, 2000; Raghunathan et al., 2001), on the other hand, does not require an explicit assumption for the joint distribution of the dataset. Instead, conditional distributions $P(Y_j|Y_{-j},\theta_j)$ are specified for each variable separately. Thus imputations are based on univariate distributions allowing for different models for each variable. Values in Y_j can be imputed, for example, by a linear or a logistic regression of Y_j on Y_{-j}, depending on the scales of measurement of Y_j, where Y_{-j} denotes all columns of Y excluding Y_j. The process of iteratively drawing from the conditional distributions can be viewed as a Gibbs sampler that will converge to draws from the theoretical joint distribution of the data if this joint distribution exists. Detailed descriptions of the approach can be found in Raghunathan et al. (2001).

3.1.1 Joint modeling

In general, it will not be possible to specify $P(Y_{new}|Y_{obs})$ directly. Note, however, that we can write

$$P(Y_{new}|Y_{obs}) = \int P(Y_{new},\theta|Y_{obs})d\theta = \int P(Y_{new}|Y_{obs},\theta)P(\theta|Y_{obs})d\theta. \qquad (3.1)$$

Given this equation, imputations can be generated in two steps:

1. Generate random draws for the parameter θ from its observed-data posterior distribution $P(\theta|Y_{obs})$ given the observed values.
2. Generate random draws for Y_{new} from its conditional predictive distribution $P(Y_{new}|Y_{obs},\theta)$ given the actual parameter θ from step 1.

With joint modeling, the second step usually is straightforward. The distribution of $(Y_{new}|Y_{obs},\theta)$ can be obtained from the underlying model. For example, a multivariate normal density can be assumed for the complete data. But the first step usually requires Markov Chain Monte Carlo techniques since the observed-data posterior distribution for $(\theta|Y_{obs})$ seldom follows standard distributions, especially if the missing pattern is not monotone (see Section 3.1.2). This means that, even for

joint modeling, convergence of the Markov Chain has to be monitored and it is not guaranteed that it will ever converge. Discussing the advantages and limits of the many different methods to monitor the convergence of Markov Chain Monte Carlo algorithms is beyond the scope of this book. One way of monitoring convergence if imputations are drawn from multiple chains is introduced in the following section. A more detailed discussion can be found in Schafer (1997, Chapter 4.4).

3.1.2 Fully conditional specification (FCS)

With FCS, the problem of drawing from a k-variate distribution is replaced by drawing k times from much easier to derive univariate distributions. Every variable in the dataset is treated separately using a regression model suitable for that specific variable. Thus, continuous variables can be imputed using a normal model, binary variables can be imputed with a logit model and so on.[2] Here, we can specify $P(\theta|Y_{obs})$ directly and no iterations are necessary, because we don't have to draw from possibly awkward multivariate distributions. To give an example, let us assume we want to impute values for a continuous variable Y. We can assume $Y|X \sim N(\mu, \sigma^2)$, where X denotes all variables that are used as explanatory variables for the imputation. The two-step imputation approach described above can now be applied as follows. Let n be the number of observations in the dataset. Let k be the number of regressors to be included in the regression. Finally, let $\hat{\sigma}^2$ and $\hat{\beta}$ be the variance and the beta-coefficient estimates obtained from ordinary least squares regressions. In the missing-data context, we assume that plausible starting values for the missing part of Y have been filled in or have been imputed in previous imputation rounds. Starting values can be obtained for example by using the predicted values from a linear regression of Y on X. Imputed values for Y_{new} can be generated using the following algorithm:

Step 1: Draw new values for $\theta = (\sigma^2, \beta)$ from $P(\theta|Y)$; i.e.,

- draw $\sigma^2|X \sim (Y - X\hat{\beta})'(Y - X\hat{\beta})\chi^{-2}_{n-k}$,
- draw $\beta|\sigma^2, X \sim N(\hat{\beta}, (X'X)^{-1}\sigma^2)$.

Step 2: Draw new values for Y_{new} from $P(Y_{new}|Y, \theta)$; i.e.,

- draw $Y_{new}|\beta, \sigma^2, X \sim N(X\beta, \sigma^2)$.

Note that we are drawing new values for the parameters directly from the observed-data posterior distributions. This means we do not need Markov Chain Monte Carlo techniques to obtain new values from the complete-data posterior distribution of the parameters. However, there might be more than one variable with missing data. Thus, we generate new values for Y_{new} by drawing from $P(Y_{new}|\beta, \sigma^2, X)$,

[2] An alternative nonparametric imputation approach based on CART models was suggested by Reiter (2005d) for the synthetic data context and was recently applied in the nonresponse context by Burgette and Reiter (2010). I discuss this approach in detail in Section 8.3.2.

and the matrix of regressors X might contain imputed values from an earlier imputation step. These values have to be updated now, based on the new information in our recently imputed variable Y. Hence, we have to sample iteratively from the fully conditional distribution for every variable in the dataset. This iterative procedure essentially can be seen as a Gibbs sampler for which the iterative draws will converge to draws from the joint distribution, if the joint distribution exists.

In a more detailed notation, for multivariate Y, let $Y_j|Y_{-j}$ be the distribution of Y_j conditional on all columns of Y except Y_j and θ_j be the parameter specifying the distribution of $Y_j|Y_{-j}$. If Y consists of p columns and each Y_j is univariate, then the tth iteration of the method consists of the following successive draws:[3]

$$\begin{aligned}
\theta_1^{(t)} &\sim P(\theta_1|Y_1^{(t-1)}, Y_2^{(t-1)}, ..., Y_p^{(t-1)}) \\
Y_1^{(t)} &\sim P(Y_1^{new}|Y_2^{(t-1)}, ..., Y_p^{(t-1)}, \theta_1^{(t)}) \\
&\vdots \\
\theta_p^{(t)} &\sim P(\theta_p|Y_p^{(t-1)}, Y_1^{(t)}, Y_2^{(t)}, ..., Y_{p-1}^{(t)}) \\
Y_p^{(t)} &\sim P(Y_p^{new}|Y_1^{(t)}, ..., Y_{p-1}^{(t)}, \theta_p^{(t)}).
\end{aligned} \tag{3.2}$$

Since imputations are generated sequentially variable by variable, this approach is also called sequential regression multivariate imputation (SRMI; see Raghunathan et al., 2001). The sampler will converge to the desired joint distribution of $(Y_{new}|Y_{obs})$, but only if this joint distribution really exists. In practice, it is often impossible to verify this; thus its existence is implicitly assumed. This is problematic since it will always be possible to draw from the conditional distributions and we will not get any hint that the Gibbs sampler actually never converges.

A simple way to detect problems with the iterative imputation procedure is to store the mean of every imputed variable for every iteration of the Gibbs sampler. A plot of the means from the imputed variables over the iterations can indicate if there is only the expected random variation between the iterations or if there is a trend between the iterations indicating problems with the model. Of course, no observable trend over the iterations does not guarantee convergence since the monitored estimates can stay stable for hundreds of iterations before drifting off to infinity. Nevertheless, this is a straightforward method to identify flawed imputation models. If different imputation chains are run to generate the m imputations, convergence can be monitored by calculating the variance of a given estimate of interest ψ (e.g., the mean and the standard deviation of each variable) within and between different imputation chains. Let ψ_{ij} denote the estimate obtained at iteration i, $i = 1, ..., T$, in chain j, $j = 1, ..., m$. The between-sequence variance B and the average within-sequence variance W can be calculated as

[3] For notational convenience, I assume that the conditional distributions of θ_j and Y_j are independent of the remaining θs, an assumption that is often met in practice, especially when imputing missing or synthetic data. Formally, we should also condition on all θs in each draw.

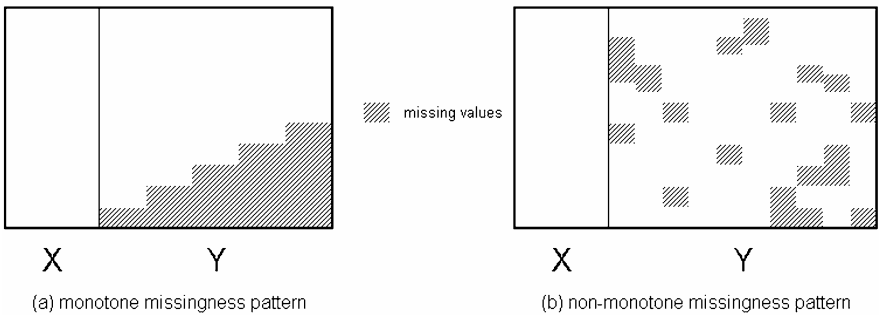

Fig. 3.1 Two missing-data patterns.

$$B = \frac{T}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{.j} - \bar{\psi}_{..})^2, \quad \text{where} \quad \bar{\psi}_{.j} = \frac{1}{T}\sum_{i=1}^{T}\psi_{ij}, \quad \bar{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{.j},$$

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \quad \text{where} \quad s_j^2 = \frac{1}{T-1}\sum_{i=1}^{T}(\psi_{ij} - \bar{\psi}_{.j})^2.$$

Gelman et al. (2004), p. 297 suggest that convergence can be assumed if

$$\hat{R} = \sqrt{\frac{(1-1/T)W + B/T}{W}} < 1.1 \tag{3.3}$$

However, it should be noted at this point that iterating between the imputations is not always necessary. If we can reorder the data in such a way that Y_j is fully observed whenever Y_{j+1} is observed, we can use a slightly different sequential regression algorithm that renders iterations between the imputations unnecessary. Let X be all the variables in the dataset that are fully observed and $Y_1,...,Y_p$ be the variables with missing values ordered by the amount of missingness. Figure 3.1 depicts two different types of missing-data patterns. The left pattern is called a *monotone missingness pattern* since the number of missing cases increases monotonically from Y_1 to Y_p. The right pattern is not monotone since there are values that are observed for Y_{j+1} but not for Y_j.

Now, remember that we can always write the joint probability as a product of conditional probabilities:

$$P(Y_1,...,Y_p|X) = P(Y_1|X)P(Y_2|Y_1,X)...P(Y_p|Y_1,...,Y_{p-1},X). \tag{3.4}$$

If the missingness pattern is monotone, $Y_1,...,Y_{j-1}$ will also be fully observed whenever Y_j is observed, so the conditional distributions do not change if we impute the missing values in $Y_1,...,Y_{j-1}$. Consequently, we do not need to update the parameters every time we impute these variables. Each draw will be a direct draw from the posterior distribution, and we do not have to wait for convergence.

Unfortunately, for most collected data, the missingness pattern will not be monotone unless we have some sort of missing by design; for example, if a follow-up study is conducted only with a subset from the original survey respondents. However, the situation is different when using multiple imputation to generate synthetic datasets. In this case, it is common for the same number of records to be replaced with synthetic values for each sensitive variable (i.e. the decision whether a value is at risk is based on the combined attributes of the record and not on the variable). This implies that for generating synthetic datasets we can often use the simplified algorithm, significantly reducing the amount of time required to generate the datasets while at the same time rendering the monitoring of convergence unnecessary.

3.1.3 Pros and cons of joint modeling and FCS

In general, empirical data will seldom follow a standard multivariate distribution, especially if they consist of a mix of numerical and categorical variables. Furthermore, FCS provides a flexible tool to account for bounds, interactions, skip patterns, or constraints between different variables (see Section 3.2). It will be very difficult to handle these restrictions which are very common in survey data, by joint modeling. In practice, the imputation task is often centralized at the methodological department of the statistical agency, and imputation experts will impute values for all the surveys conducted by the agency. Imputed datasets that don't fulfill simple restrictions such as non-negativity or other logical constraints will never be accepted by subject matter analysts from other departments. Thus, preserving these constraints is a central element of the imputation task. For this reason, most applications of multiple imputation are based on FCS. Van Buuren and Groothuis-Oudshoorn (2010) provide a vast list of applied papers that rely on this approach.

Overall, joint modeling will be preferable if only a limited number of variables need to be imputed, no restrictions have to be maintained, and the joint distribution can be approximated reasonably well with a standard multivariate distribution. For more complex imputation tasks, only fully conditional specification will enable the imputer to preserve constraints inherent in the data. In this case, convergence of the Gibbs sampler should be monitored carefully.

3.2 Real data problems and possible ways to handle them

The basic concept of multiple imputation is straightforward to apply, and multiple imputation tools available for most statistical software packages further reduce the modeling burden for the imputer. For example, the fully conditional approach is implemented in IVEware for SAS (Raghunathan et al., 2002), in the packages `mice` (van Buuren and Groothuis-Oudshoorn, 2010) and `mi` (Su et al., 2009) for R, and in a set of ado-files called `ice` for Stata (Royston, 2005, 2007, 2009). The latest

version of the Missing Values add-on (MVA) module for SPSS 17.0 also includes a multiple-imputation feature based on this approach. Joint modeling is implemented in the stand-alone packages NORM, CAT, MIX, and PAN (Schafer, 1997), the R package `AMELIA II` (Honaker et al., 2010), and `INORM` (Galati and Carlin, 2009) and the new multiple imputation system, also called `mi` (StataCorp, 2009), in Stata. However, simply applying standard imputation procedures to real data can lead to biased or inconsistent imputations. Several additional aspects have to be considered in practice when imputing real data. Unfortunately, at present most of the standard software can only handle some of these aspects, which will be discussed below.

3.2.1 Imputation of semi-continuous variables

A problem with modeling continuous variables that often arises in surveys is the fact that many of these variables in fact are semi-continuous (i.e., they have a spike at one point of the distribution, but the remaining distribution can be seen as a continuous variable). For most variables, this spike will occur at zero. To give an example, in our dataset (see Chapter 4), the establishments are asked how many of their employees obtained a college degree. Most of the small establishments do not require such highly skilled workers. In this case, I suggest adopting the two step imputation approach proposed by Raghunathan et al. (2001). In the first step, we impute whether the missing value is zero or not. For that, missing values are imputed using a logit model with outcome 1 for all units with a positive value for that variable. In the second step, a standard linear model is applied only to the units with observed positive values to predict the actual value for the units with a predicted positive outcome in step one. All values for units with outcome zero in step one are set to zero.

3.2.2 Bracketed imputation

Often, imputed values are required to fall into certain bounds. These bounds might be defined by the outcome of another variable (e.g., when the survey respondent refused to report his or her exact income but reported that it was between 80,000 and 90,000 euros). But imputation bounds might also be necessary because the outcome space for a variable is limited. For example, many survey variables can never be negative in reality. This has to be considered during the imputation process. A simple way to achieve this goal is to redraw from the imputation model for those units with imputed values that are outside the defined bounds until all values fulfill the constraints. In practice, usually an upper bound z has to be defined for the number of redraws for one unit since it is possible that the probability of drawing a value inside the bounds for this unit from the defined model is very low. The value for this unit is set to the closest boundary if z draws from the model never produced

a plausible value. However, there is a caveat with this approach. Redrawing from the model for implausible values is equivalent to drawing from a truncated distribution. If the truncation points are not at the very far end of the distribution (i.e., the model is misspecified), even simple descriptive analyses such as the mean of the imputed variable will differ significantly from the true value of the complete data. For this reason, this approach should only be applied if the probability of drawing implausible values from the specified model is very low and we only want to prevent some very unlikely unrealistic values from being imputed. If the fraction of units that would have to be corrected with this approach is too high, the model needs to be revised. Usually it is helpful to define different models for different subgroups of the data. For example, to overcome the problem of generating too many negative values, a separate model for the units with small values should be defined. An alternative for the special case of non-negativity constraints is to log transform the variable before imputation. This will guarantee that all imputed values are positive after backtransformation.

3.2.3 Imputation under linear constraints

In many surveys, the outcome of one variable by definition has to be equal to or above the outcome of another variable. For example, the total number of employees always has to be at least as high as the number of part-time employees. When imputing values in this situation, Schenker et al. (2006) suggest the following approach. Variables that define a subgroup of another variable are always expressed as a proportion (i.e., all values for the subgroup variable are divided by the total before the imputation and thus are bounded between zero and one). A logit transformation of the variables guarantees that they will have values in the full range $]-\infty,\infty[$ again. Values for these transformed variables can be imputed with a standard imputation approach based on linear regressions. After the imputation, all values are transformed back to get proportions again, and finally all values are multiplied with the totals to get back the absolute values. To avoid problems on the bounds of the proportions, I suggest setting proportions greater than 0.999999 to 0.999999 before the logit transformation and using the two-step imputation approach described in Section 3.2.1 to determine zero values.

3.2.4 Skip patterns

Skip patterns (e.g., a battery of questions are only asked if they are applicable) are very common in surveys. Although it is obvious that they are necessary and can significantly reduce the response burden for the survey participant, they are a nightmare for anybody involved in data editing and imputation or statistical disclosure control. Especially if the skip patterns are hierarchical, it is very difficult to guar-

antee that imputed values are consistent with these patterns. With fully conditional specification, it is straightforward to generate imputed datasets that are consistent with all these rules. The two-step approach described in Section 3.2.1 can be applied to decide whether the questions under consideration are applicable. Values are imputed only for the units selected in step one. Nevertheless, correctly implementing all filtering rules is a labor-intensive task that can be more cumbersome than defining good imputation models. Furthermore, skip patterns can lead to variables that are answered by only a small fraction of the respondents, and it can be difficult to develop good models based on a small number of observations.

Chapter 4
The IAB Establishment Panel

Since the establishment survey of the German Institute for Employment Research (IAB) is used throughout this book to illustrate the different aspects of multiple imputation, a short introduction to this dataset should prelude the body of this book. The IAB Establishment Panel[1] is based on the German employment register aggregated via the establishment number as of June 30 of each year. The basis of the register, the German Social Security Data (GSSD), is the integrated notification procedure for health, pension, and unemployment insurance, which was introduced in January 1973. This procedure requires employers to notify the social security agencies about all employees covered by social security. As by definition the German Social Security Data only include employees covered by social security – civil servants and unpaid family workers, for example, are not included – approximately 80% of the German workforce are represented. However, the degree of coverage varies considerably across occupations and industries.

Since the register only contains information on employees covered by social security, the panel includes establishments with at least one employee covered by social security. The sample is drawn using a stratified sampling design. The stratification cells are defined by ten classes for the size of the establishment, 16 classes for the region,[2] and 17 classes for the industry.[3] These cells are also used for weighting and extrapolation of the sample. The survey is conducted by interviewers from TNS Infratest Sozialforschung. For the first wave, 4,265 establishments were interviewed in West Germany in the third quarter of 1993. Since then, the Establishment Panel has been conducted annually – since 1996 with over 4,700 establishments in East Germany in addition. In the 2008 wave more than 16,000 establishments participated in the survey. The response rate of units that have been interviewed repeatedly is over 80%. Each year, the panel is accompanied by supplementary samples and

[1] The approach and structure of the establishment panel are described, for example, by Fischer et al. (2008) and Kölling (2000).

[2] Before 2006, the stratification by region contained 17 classes since two separate classes were used for East and West Berlin.

[3] Between 2000 and 2004, 20 industry classes were used, and before 2000 the sample was stratified by 16 industry classes.

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_4,

follow-up samples to include new or reviving establishments and to compensate for panel mortality. The questionnaire contains a set of core questions that are asked annually with detailed information about employment development, business policy, vocational training, personnel structure and personnel movements, investments, wages and salaries, and adherence to collective agreements. Information on further training, working time, public funding, and innovations is asked every other year. Additional changing questions relevant for the current political debate complete the survey.

Considered one of the most important business surveys in Germany, there is high demand for access to these data from external researchers. Because of the sensitive nature of the data, researchers desiring direct access to the data have to work onsite at the IAB. Alternatively, researchers can submit code for statistical analyses to the IAB research data center, whose staff run the code on the data and send the results to the researchers. To help researchers develop code, the IAB provides access to a publicly available "dummy dataset" with the same structure as the Establishment Panel. For all analyses done with the genuine data, researchers can publicize their analyses only after IAB staff check for potential violations of confidentiality.

Releasing scientific use files of the Establishment Panel will allow more researchers to access the data with fewer burdens, stimulating research on German business data. It also will free up staff time from running code and conducting confidentiality checks. Because there are so many sensitive variables in the dataset, standard disclosure limitation methods such as swapping or microaggregation would have to be applied with high intensity, which would severely compromise the utility of the released data. Therefore, the IAB decided to develop synthetic data. The first synthetic dataset generated for the 2007 wave of the panel was released in January 2011.

To evaluate the quality of the different synthetic datasets that are used throughout this book to illustrate the different MISD approaches, I always compare analytic results achieved with the original data with results from the synthetic data. For most datasets, comparisons are based on an analysis by Thomas Zwick, "Continuing Vocational Training Forms and Establishment Productivity in Germany" published in the *German Economic Review*, Vol. 6, No. 2, pp. 155–184, in 2005. Since this analysis is used for validity evaluations in several chapters of the book, I provide a detailed description here.

Zwick analyzes the productivity effects of different continuing vocational training forms in Germany. He argues that vocational training is one of the most important measures to gain and keep productivity in a firm. For his analysis, he uses the 1997 to 2001 waves from the IAB Establishment Panel.

In 1997 and 1999, the Establishment Panel included the following additional question that was asked if the establishment did support continuous vocational training in the first part of 1997 or 1999, respectively: "For which of the following internal or external measures were employees exempted from work or were costs completely or partly taken over by the establishment?" Possible answers were: formal internal training, formal external training, seminars and talks, training on the job, participation at seminars and talks, job rotation, self-induced learning, quality

circles, and additional continuous vocational training. Zwick examines the productivity effects of these training forms and demonstrates that formal external training, formal internal training, and quality circles do have a positive impact on productivity. Especially for formal external courses, the productivity effect can be measured even two years after the training.

To detect why some firms offer vocational training and others do not, Zwick runs a probit regression using the 1997 wave of the establishment panel. In the regression, Zwick uses two variables (*investment in IT* and the *codetermination of the employees*) that are only included in the 1998 wave of the establishment panel. Moreover, he excludes some observations based on information from other years. As I use only the 1997 wave for the illustrations in the following chapters, the two variables from the 1998 wave are dropped from the regression and all results presented are based on the full sample.

For his analysis, Zwick runs the regression only on units with no missing values for the regression variables, losing all the information on establishments that did not respond to all variables used. This might lead to biased estimates if the assumption of a missing pattern that is completely at random (see Rubin (1987) or Section 5.2 in this book) does not hold. For that reason, I will always compare the regression results from the synthetic datasets, which by definition have no missing values, with the results Zwick would have achieved if he had run his regression on a dataset with all the missing values multiply imputed.

Chapter 5
Multiple Imputation for Nonresponse[1]

For many datasets, especially for nonmandatory surveys, missing data are a common problem. Deleting units that are not fully observed and using only the remaining units is a popular, easy-to-implement approach in this case. However, using only fully observed observations will generally lead to reduced efficiency for the estimates. But even more problematic, this approach can possibly lead to severe bias if the strong assumption of a missing pattern that is missing completely at random (MCAR; see Section 5.2) is not fulfilled. Imputing missing values can help handle this problem. However, imputing missing values only once (single imputation) generally doesn't account for the fact that the imputed values are only estimates for the true values. After the imputation process, they are often treated like originally observed values, leading to an underestimation of the variance in the data and from this to p values that are too significant. Multiple imputation was suggested by Rubin (1978) to overcome these problems.

5.1 Inference for datasets multiply imputed to address nonresponse

5.1.1 Univariate estimands

To understand the procedure of analyzing multiply imputed datasets, think of an analyst interested in an unknown scalar parameter Q, where Q could be, for example, the population mean or a regression coefficient in a linear regression. Inferences for this parameter for datasets with no missing values usually are based on a point estimate q, a variance estimate u, and a normal or Student's t reference distribution. For analysis of the imputed datasets, let $q^{(i)}$ and $u^{(i)}$ for $i = 1, 2, ...m$ be the point and variance estimates achieved from each of the m completed datasets. To get a final estimate over all imputations, these estimates have to be combined using the com-

[1] Most of this chapter is taken from Drechsler and Rässler (2008) and Drechsler (2011a).

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_5,

bining rules first described by Rubin (1978). The following quantities are needed for inferences for scalar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q^{(i)}/m, \tag{5.1}$$

$$b_m = \sum_{i=1}^{m} (q^{(i)} - \bar{q}_m)^2/(m-1), \tag{5.2}$$

$$\bar{u}_m = \sum_{i=1}^{m} u^{(i)}/m. \tag{5.3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_m = \bar{u}_m + (1+m^{-1})b_m \tag{5.4}$$

to estimate the variance of $\bar{q}_m$.

For the point estimate, the final estimate is simply the average of the m point estimates. Its variance is estimated by combining the "within-imputation" variance $\bar{u}_m$ and the "between-imputation" variance b_m. We will see in the following chapters that all combining rules for the different multiple-imputation settings more or less rely on these three quantities. The factor $(1+m^{-1})$ reflects the fact that only a finite number of completed-data estimates $q^{(i)}$ are averaged together to obtain the final point estimate.

The quantity $r = (1+m^{-1})b_m/T_m$ estimates the fraction of information about Q that is missing due to nonresponse.

Inferences from multiply imputed data are based on $\bar{q}_m$, T_m, and a Student's t reference distribution. Thus, for example, interval estimates for Q have the form $\bar{q}_m \pm t(1-\alpha/2)\sqrt{T_m}$, where $t(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of the t distribution. Rubin and Schenker (1986) provide the approximate value $\nu_{RS} = (m-1)r^{-2}$ for the degrees of freedom of the t distribution under the assumption that with complete data a normal reference distribution would have been appropriate. Barnard and Rubin (1999) relax the assumption of Rubin and Schenker (1986) to allow for a t reference distribution with complete data and suggest the value $\nu_{BR} = (\nu_{RS}^{-1} + \hat{\nu}_{obs}^{-1})^{-1}$ for the degrees of freedom in the multiple-imputation analysis, where $\hat{\nu}_{obs} = (1-r)(\nu_{com})(\nu_{com}+1)/(\nu_{com}+3)$ and ν_{com} denotes the complete-data degrees of freedom.

5.1.2 Multivariate estimands

Often, researchers will be interested in testing a null hypothesis of the form $\mathbf{Q} = \mathbf{Q}_0$ for some k-component estimand $\mathbf{Q}$, for example when testing the null hypothesis that some regression coefficients in a standard regression model equal 0. Following the notation in Reiter and Raghunathan (2007), let $\bar{\mathbf{q}}_m$, $\mathbf{b}_m$, and $\bar{\mathbf{u}}_m$ be

the multivariate analogs to $\bar{q}_m$, b_m, and $\bar{u}_m$ defined in (5.1) to (5.3). For the multivariate case, the quantities are based on the k-dimensional estimates $\mathbf{q}^{(i)}$ and $k \times k$ covariance matrices $\mathbf{u}^{(i)}$. Unfortunately, the standard Wald test with statistic $(\bar{\mathbf{q}}_m - \mathbf{Q}_0)^T T_m^{-1} (\bar{\mathbf{q}}_m - \mathbf{Q}_0)$ provides unreliable results, when $k > m$ and m is moderate, because of the potentially large variability in $\mathbf{b}_m$ (Rubin, 1987; Li et al., 1991).

Two alternatives have been proposed in the literature that provide more stable results. Rubin (1987) suggests using the following test statistic under the assumption that the fraction of missing information r is equal for all components of $\mathbf{Q}$:

$$S_m = (\bar{\mathbf{q}}_m - \mathbf{Q}_0)^T \bar{\mathbf{u}}_m^{-1} (\bar{\mathbf{q}}_m - \mathbf{Q}_0)/(k(1+r_m)), \tag{5.5}$$

where $r_m = (1+1/m)tr(\mathbf{b}_m \bar{\mathbf{u}}_m^{-1})/k$ is the average relative increase in variance due to nonresponse across the components of $\mathbf{Q}$. Inference is based on an approximate F distribution, F_{k,ν_w}, with $\nu_w = 4 + (t-4)(1+(1-2/t)/r_m)^2$ and $t = k(m-1) > 4$. When $t \leq 4$, $\nu_w = t(1+1/k)(1+1/r_m)^2/2$. The p value for testing $\mathbf{Q} = \mathbf{Q}_0$ is $Pr(F_{k,\nu_w} > S_m)$.

If $\mathbf{Q}$ contains a large number of components k, using $\bar{\mathbf{u}}_m$ can be cumbersome. Meng and Rubin (1992) suggest a different approach based on the log-likelihood ratio test that avoids calculating $\bar{\mathbf{u}}_m$. Again following the notation given in Reiter and Raghunathan (2007), let ψ be the vector of parameters in the analyst's model, and let $\psi^{(i)}$ be the maximum likelihood estimate of ψ computed from $D^{(i)}$, where $D^{(i)}$ is the ith imputed dataset and $i = 1, \ldots, m$. The analyst is interested in testing the hypothesis that $\mathbf{Q}(\psi) = \mathbf{Q}_0$, where $\mathbf{Q}(\psi)$ is a k-dimensional function of ψ. Let $\psi_0^{(i)}$ be the maximum likelihood estimate of ψ obtained from $D^{(i)}$ subject to $\mathbf{Q}(\psi) = \mathbf{Q}_0$. The log-likelihood ratio test statistic associated with $D^{(i)}$ is $L^{(i)} = 2\log f(D^{(i)}|\psi^{(i)}) - 2\log f(D^{(i)}|\psi_0^{(i)})$. Let $\bar{L} = \sum_{i=1}^m L^{(i)}/m$, $\bar{\psi} = \sum_{i=1}^m \psi^{(i)}/m$, and $\bar{\psi}_0 = \sum_{i=1}^m \psi_0^{(i)}/m$. Finally, let $\bar{L}_0 = (1/m)\sum_{i=1}^m (2\log f(D^{(i)}|\bar{\psi}) - 2\log f(D^{(i)}|\bar{\psi}_0))$, the average of the log-likelihood ratio test statistics evaluated at $\bar{\psi}$ and $\bar{\psi}_0$. The likelihood ratio test statistic is given by

$$\hat{S}_m = \bar{L}_0/(k(1+\hat{r}_m)), \tag{5.6}$$

where $\hat{r}_m = ((m+1)/t)(\bar{L} - \bar{L}_0)$. The reference distribution for $\hat{S}_m$ is $F_{k,\hat{\nu}}$, where $\hat{\nu}$ is defined as ν but using $\hat{r}_m$ instead of r_m.

For small sample sizes, Reiter (2007) presents an alternative estimator for the denominator degrees of freedom in the reference distribution for S_m. The derivation is basically an extension of the methods developed in Barnard and Rubin (1999) to the multivariate case. A simplified approximation to these degrees of freedom is given by

$$\nu_{fapp} = 4 + \left(\frac{1}{\nu_{com}^* - 4(1+a)} + \frac{1}{t-4} \left(\frac{a^2(\nu_{com}^* - 2(1+a))}{(1+a)^2(\nu_{com}^* - 4(1+a))} \right) \right)^{-1}, \tag{5.7}$$

where $\nu^*_{com} = \nu_{com}(\nu_{com}+1)/(\nu_{com}+3)$ and $a = r_m t/(t-2)$. Reiter (2007) also presents a more complicated expression using higher-order terms in the formula for the degrees of freedom.

5.2 Analytical validity for datasets multiply imputed to address nonresponse

It is difficult to evaluate the quality of the imputations for missing values, since information about the missing values usually is not available by definition, and the assumption that the response mechanism is ignorable (Rubin, 1987), necessary for obtaining valid imputations if the response mechanism is not modeled directly, cannot be tested with the observed data. A response mechanism is considered ignorable if, given that the sampling mechanism is ignorable, the response probability only depends on the observed information.[2] If these conditions are fulfilled, the missing data are said to be *missing at random* (MAR) and imputation models only need to be based on the observed information. As a special case, the missing data are said to be *missing completely at random* (MCAR), if the response mechanism does not depend on the data (observed or unobserved), which implies that the distribution of the observed data and the distribution of the missing data are identical. If the requirements above are not fulfilled, the missing data are said to be *missing not at random* (MNAR) and the response mechanism needs to be modeled explicitly. Little and Rubin (2002) provide examples for nonignorable missing-data models.

As noted before, it is not possible to check whether the missing data are MAR with the observed data. But even if the MAR assumption cannot be tested, this does not mean the imputer cannot test the quality of his imputations at all. Abayomi et al. (2008) suggest several ways of evaluating model-based imputation procedures. Basically their ideas can be divided into two categories. On the one hand, the imputed data can be checked for reasonability. Simple distributional and outlier checks can be evaluated by subject matter experts for each variable to avoid implausible imputed values like a turnover of $100 million for a small establishment in the social sector. On the other hand, since imputations usually are model-based, the fit of these models can and indeed should be tested. Abayomi et al. (2008) label the former as *external* diagnostic techniques since the imputations are evaluated using outside knowledge and the latter *internal* diagnostic techniques since they evaluate the modeling based on model fit without the need of external information.

To automate the external diagnostics to some extent, Abayomi et al. (2008) suggest using the Kolmogorov–Smirnoff test to flag any imputations for which the distribution of the imputed values significantly differs from the distribution of the ob-

[2] The additional requirement that the sampling mechanism also be ignorable (Rubin, 1987) (i.e., the sampling probability only depends on observed data) is usually fulfilled in scientific surveys. The stratified sampling design of the IAB Establishment Panel also satisfies this requirement since the sampling probabilities are defined solely by the stratification cells derived from the German Social Security Data (see Chapter 4).

served values. Of course, a significant difference in the distributions does not necessarily indicate problems with the imputation. Indeed, if the missing-data mechanism is MAR but not MCAR, we would expect the two distributions to differ. The test is only intended to decrease the number of variables that need to be checked manually, implicitly assuming that no significant difference between the original and the imputed data indicates no problem with the imputation model.

However, I am skeptical about this automated selection method since the test is sensitive to the sample size and thus the chance of rejecting the null hypothesis will be lower for variables with lower missing rates and variables that are answered only by a subset of the respondents. Furthermore, it is unclear what significance level to choose and, as noted above, rejection of the null hypothesis does not necessarily indicate an imputation problem, but not rejecting the null hypothesis is not a guarantee that we found a good imputation model either. However, this is implicitly assumed by this procedure.

5.3 Multiple imputation of the missing values in the IAB Establishment Panel

In this section, I illustrate how multiple imputation for nonresponse could be implemented in practice. I discuss the extensive imputation task required to impute all missing values in the 2007 wave of the IAB Establishment Panel and describe the methods I used to evaluate the quality of the imputations.

5.3.1 The imputation task

Most of the 284 variables included in the 2007 wave of the Panel are subject to nonresponse. Only 26 variables are fully observed. However, missing rates vary considerably between variables and are modest for most variables: 65.8% of the variables have missing rates below 1%, 20.4% of the variables have missing rates between 1% and 2%, 15.1% have rates between 2% and 5%, and only 12 variables have missing rates above 5%. The five variables with missing rates above 10% are *subsidies for investment and material expenses* (13.6%), *payroll* (14.4%), *intermediate inputs as proportion of turnover* (17.4%), *turnover in the last fiscal year* (18.6%), and *number of workers who left the establishment due to restructuring measures* (37.5%). Obviously, the variables with the highest missing rates contain information that is either difficult to provide, such as *number of workers who left the establishment due to restructuring measures*, or considered sensitive, such as *turnover in the last fiscal year*. The variable *number of workers who left the establishment due to restructuring measures* is only applicable to the 626 establishments in the dataset that declared they had restructuring measures in the last year. Of these 626, only 391 establishments provided information on the number of workers that left the estab-

lishment due to these measures. Clearly, it is often difficult to tailor exactly which workers left as a result of the measures and which left for other reasons. This might be the reason for the high missing rates. The low number of observed values is also problematic for the modeling task, so this variable should be used with caution in the imputed dataset.

5.3.2 Imputation models

Since the dataset contains a mixture of categorical variables and continuous variables with skewed distributions and a variety of often hierarchical skip patterns and logical constraints, it is impossible to apply the joint modeling approach described in Section 3.1.1. I apply the fully conditional specification approach described in Section 3.1.2, iteratively imputing one variable at a time, conditioning on the other variables available in the dataset. For the imputation, I basically rely on three different imputation models: the linear model for the continuous variables, the logit model for binary variables, and the multinomial logit for categorical variables with more than two categories. Multiple-imputation procedures for these models are described in Raghunathan et al. (2001). In general, all variables that don't contain any structural missings are used as predictors in the imputation models in hopes of reducing problems from uncongeniality (Meng, 1994). Uncongeniality refers to the situation where the model used by the analyst of the data differs from the model used for the imputation. This can lead to biased results if the analyst's model is more complex than the imputation model and the imputation model omitted important relationships present in the original data. Since the true data-generating model is usually unknown and an imputation model that is more complex than the true model only causes some loss in efficiency, the standard imputation strategy should be to include as many variables as possible in the imputation model (Little and Raghunathan, 1997).

In the multinomial logit model for the categorical variables, the number of explanatory variables is limited to 30 variables found by stepwise regression. This reduction is necessary since the full model never converges for most categorical variables due to multicollinearity. But even if the model eventually converges, the rate of convergence is so slow that finding the maximum likelihood estimates could easily take more than 12 hours. Because I generate $m = 5$ imputed datasets running 100 iterations of the Gibbs sampler before storing the next dataset to avoid dependencies between the imputed values, the imputation would take several months to finish. Thus, I generally reduce the number of explanatory variables to 30 for the multinomial imputation model, although this might increase the risk of uncongeniality discussed above. The stepwise regression procedure should limit this risk because the variables with the highest influence on the dependent variable are always included in the imputation model.

To improve the quality of the imputation, I define several separate models for the variables with high missing rates, such as *turnover* or *payroll*. Independent mod-

els are fit for East and West Germany and for different establishment size classes. Ideally, different imputation models should be defined for every stratification cell to correctly account for the stratified sampling design. Since most of the stratification cells would be too small to allow useful modeling, we follow the advice in Reiter et al. (2006) and always include the survey weights as predictors in every imputation model instead.

All continuous variables are subject to non-negativity constraints, and the outcome of many variables is further restricted by linear constraints. The imputation process is further complicated by the fact that most variables have huge spikes at zero and, as mentioned before, the skip patterns are often hierarchical. I therefore have to rely on a mixture of the adjustments presented in Section 3.2. To control for the skewness, I transform each continuous variable by taking the cubic root before the imputation. I prefer the cubic root transformation over the log transformation which is often used in the economics literature to model skewed variables such as turnover, because the cubic root transformation is less sensitive to deviations between the imputed and original values in the right tail of the distribution. Since the slope of the exponential function increases exponentially, whereas the slope of $f(x) = x^3$ increases only quadratically, a small deviation in the right tail of the imputed transformed variable has more severe consequences after backtransformation for the log-transformed variable than for the variable transformed by taking the cubic root.

5.3.3 Evaluating the quality of the imputations

Following Abayomi et al. (2008), I searched for possible flaws in the imputations by plotting the distributions for the original and imputed values for every continuous variable. I checked whether any notable differences between these distributions could be justified by differences in the distributions of the covariates. Figure 5.1 displays the distributions for two representative variables based on kernel density estimation. Original values are represented with a solid line, imputed values with a dashed line. Both variables are reported on the log scale. The left variable (*payroll*) represents a candidate that I did not investigate further since the distributions match almost exactly. The right variable (*number of participants in further education (NB.PFE)*) is an example of a variable for which I tried to understand the difference between the distribution of the observed values and the distribution of the imputed values before accepting the imputation model.

Obviously, most of the imputed values for the variable *NB.PFE* are larger than the observed values for this variable. To understand this difference, I examined the dependence between the missing rate and the establishment size. In Table 5.1, I present the percentage of missing units in ten establishment size classes defined by quantiles and the mean of *NB.PFE* within these quantiles. The missing rates are low up to the sixth establishment size class. Beyond that point, the missing rates increase substantially with every class. The average number of further education participants

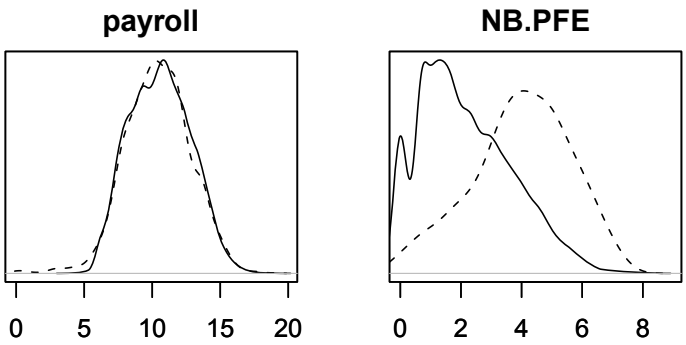

Fig. 5.1 Observed (solid line) and imputed (dashed line) data for *payroll* and *number of participants in further education (NB.PFE)*. Both variables are reported on the log scale.

increases steadily with every establishment size class, with the largest increases in the second half of the table. With these results in mind, it is not surprising that the imputed values for that variable are often larger than the observed values.

I inspected several continuous variables by comparing the distributions of the observed and imputed values in the dataset and did not find any differences in the distributions that could not be explained by the missingness pattern. I also investigated whether any weighted imputed value for any variable was above the maximum weighted observed value for that variable. Again, this would not necessarily be problematic, but I did not want to impute any unrealistic influential outliers. However, I did not find any weighted imputed value that was higher than the maximum of its weighted observed counterpart.

Following Su et al. (2009), I used three graphics as internal diagnostics to evaluate the model fit: a normal Q-Q plot, a plot of the residuals from the regression against the fitted values, and a binned residual plot (Gelman and Hill, 2006). The

Table 5.1 Missing rates and means per quantile for *NB.PFE*.

Est. size quantile	Missing rate in %	$mean(NB.PFE)$ per quantile
1	0.09	1.61
2	0.00	2.49
3	0.57	3.02
4	0.36	4.48
5	0.44	6.09
6	0.37	9.53
7	0.85	15.48
8	1.16	26.44
9	3.18	56.39
10	6.66	194.09

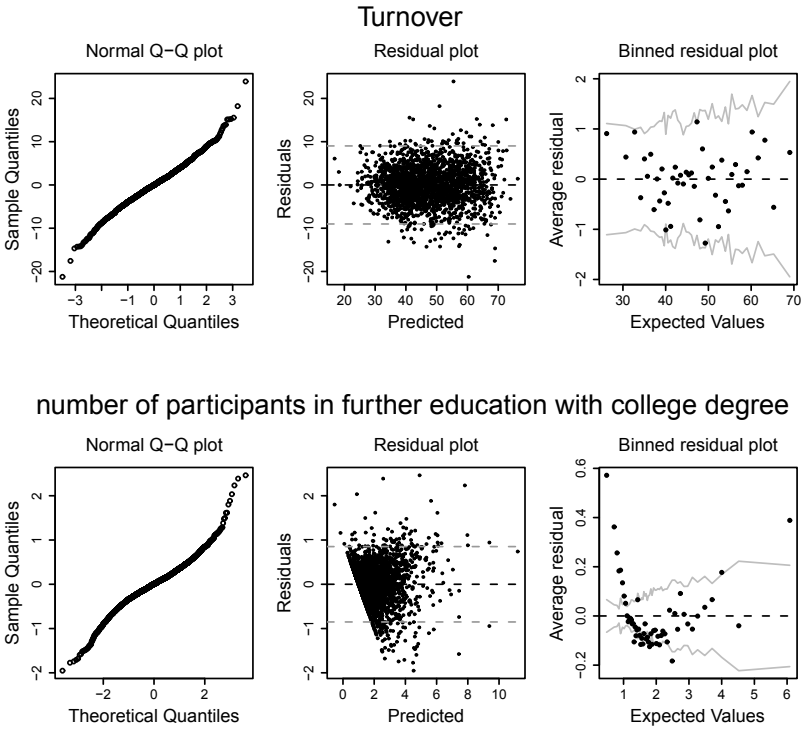

Fig. 5.2 Model checks for *turnover* and *number of participants in further education with college degree*.

normal Q-Q plot indicates whether the assumption of a normal distribution for the residuals is justified by plotting the theoretical quantiles of a normal distribution against the empirical quantiles of the residuals. The residual plot visualizes any unwanted dependencies between the fitted values and the residuals. For the binned residual plot, the average of the fitted values is calculated within several predefined bins and plotted against the average of the residuals within these bins. This is especially helpful for categorical variables since the output of a simple residual plot is difficult to interpret if the outcome is discrete.

Figure 5.2 again provides an example of one model (one of the models for the variable *turnover*) that I did not inspect any further and one model (for the variable *number of participants in further education with college degree (NB.PFE.COL)*) for which I checked the model for necessary adjustments.

For both variables, the assumption that the residuals are more or less normally distributed seems to be justified. For the variable *turnover*, the two residual plots further confirm the quality of the model. Only a small amount of residuals fall outside of the grey dotted 95% confidence bands for the residual plot, and none of the averaged residuals fall outside the grey 95% confidence bands for the binned resid-

uals. This is different for *NB.PFE.COL*. Although most of the points are still inside the 95% confidence bands, we see a clear relationship between the fitted values and the residuals for the small values, and the binned residuals for these small values all fall outside the confidence bands. However, this phenomenon can be explained if we inspect the variable further. Most establishments do not have any participants in further training with a college degree, and I fitted the model only to the 3,426 units reported to have at least one participant.Of these units, 648 reported that they had only one participant, leading to a spike at 1 in the original data. Since I simply fit a linear model to the observed data, the almost vertical line in the residual plot is not surprising. It contains all the residuals for all the units with only 1 participant in the original data. The binned residual plot indicates that the small fitted values sometimes severely underestimate the original values. The reason for this is again the fact that the original data are truncated at 1, whereas the fitted values are predictions from a standard linear model that would even allow negative fitted values since I computed the fitted values before the adjustments for non-negativity described in Section 3.2.2. The consequence is a slight overestimation for the larger fitted values.

I found similar patterns in some other variables that had huge spikes at 1. I could have tried to model the data with a truncated distribution or applied the semi-continuous approach described in Section 3.2.1 to model the spike at 1 separately, but since I expect that the non-negativity adjustments reduce this effect, I decided to avoid making the already complex modeling task even more difficult.

Missing rates are substantially lower for the categorical variables. Only 59 out of the close to 200 categorical variables in the dataset have missing rates above 1%, and I limited my evaluation to these variables. I compared the percentage of responses in each category for the observed and the imputed values and flagged a variable for closer inspection if the percentage of responses in one imputed category differed by more than 20% from the relative number in the observed category. I further limited my search to categories that contained at least 25 units since small changes in categories with fewer units would lead to significant changes in the relative differences for these categories. All 15 variables that were flagged by this procedure had a missing rate below 5%, and the differences between the imputed and original response rates could be explained by the missingness pattern for all of them. I select one variable here to illustrate the significant differences between observed and imputed values that can arise from a missingness pattern that is definitely not missing completely at random. The variable under consideration asks for the expectations about the investment in 2007 compared with 2006. Table 5.2 provides some summary statistics for this variable. There is a substantial difference for the second and the third categories if we simply compare the observed response rates (column 1) with the imputed response rates (column 2). But the missing rate is only 0.2% for this variable for units with investments in 2006 but soars to 10.5% for units without investments in 2006. Thus, the response rates across categories for the imputed values will be influenced by the expectations for those units that had no investments in 2006 (column 4) even though only 12.9% of the participants who planned investments for 2007 reported no investments in 2006. These response rates differ completely from the response rates for units that reported investments in 2006 (column 3). Thus the

Table 5.2 Expectations for the investments in 2007 (response rates in % for each category).

Category	Obs. data	Imp. data	Observed units with investment 2006	Observed units without investment 2006
Will stay the same	36.57	37.96	41.33	0.59
Increase expected	38.79	57.66	30.74	99.41
Decrease expected	20.33	0.73	23.05	0.00
Don't know yet	4.31	3.65	4.88	0.00

percentage of establishments that expect an increase in investments is significantly larger in the imputed data than in the original data.

For categorical data, the normal Q-Q plot is not appropriate as an internal diagnostic tool, and the residual plot is difficult to interpret if the outcome is discrete. Therefore, I only examined the binned residual plots for the 59 categorical variables with missing rates above 1%. All plots indicate a good model fit. I moved all graphics to the Appendix B for brevity.

To check for possible problems with the iterative imputation procedure, I stored the mean for several continuous variables after every imputation iteration. I did not find any inherent trend for the imputed means for any of the variables.

Chapter 6
Fully Synthetic Datasets[1]

In 1993, Rubin suggested creating fully synthetic datasets based on the multiple-imputation framework. His idea was to treat all units in the population that have not been selected in the sample as missing data, impute them according to the multiple-imputation approach, and draw simple random samples from these imputed populations for release to the public. Most surveys are conducted using complex sampling designs. Releasing simple random samples simplifies research for the potential user of the data since the design doesn't have to be incorporated in the model. It is not necessary, however, to release simple random samples. If a complex design is used, the analyst accounts for the design in the within-variance $u^{(i)}$, $i = 1, ..., m$.

As an illustration, think of a dataset of size n sampled from a population of size N. Suppose further that the imputer has information about some variables X for the whole population, for example from census records, and only the information from the survey respondents for the remaining variables Y. Let Y_{inc} and Y_{exc} be the observed units and the nonsampled units of the population respectively. For simplicity, assume that there are no data with items missing in the observed dataset. Generating fully synthetic datasets if the original data are subject to nonresponse is discussed in Chapter 8. The synthetic datasets can be generated in two steps. First, construct m imputed synthetic populations by drawing Y_{exc} m times independently from the posterior predictive distribution $f(Y_{exc}|X, Y_{inc})$ for the $N-n$ unobserved values of Y. If the released data should contain no real data for Y, all N values can be drawn from this distribution. Second, take simple random samples from these populations and release them to the public. The second step is necessary, as it might not be feasible to release m whole populations due to the simple matter of data size. In practice, it is not mandatory to generate complete populations. The imputer can make random draws from X in the first step and only impute values of Y for the drawn X. The analysis of the m simulated datasets follows the same lines as the analysis after multiple imputation for missing values in regular datasets, as described in Section 5.1.

[1] Most of this chapter is taken from Drechsler et al. (2008b) and Drechsler and Reiter (2009).

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_6,

6.1 Inference for fully synthetic datasets

6.1.1 Univariate estimands

To understand the procedure of analyzing fully synthetic datasets, think of an analyst interested in an unknown scalar parameter Q, where Q could be, for example, the mean of a variable, the correlation coefficient between two variables, or a regression coefficient in a linear regression. Inferences for this parameter derived from the original datasets usually are based on a point estimate q, an estimate for the variance of q, u, and a normal or Student's t reference distribution. For analysis of the imputed datasets, let $q^{(i)}$ and $u^{(i)}$ for $i = 1, ..., m$ be the point and variance estimates for each of the m synthetic datasets. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q^{(i)}/m, \tag{6.1}$$

$$b_m = \sum_{i=1}^{m} (q^{(i)} - \bar{q}_m)^2/(m-1), \tag{6.2}$$

$$\bar{u}_m = \sum_{i=1}^{m} u^{(i)}/m. \tag{6.3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_f = (1 + m^{-1})b_m - \bar{u}_m \tag{6.4}$$

to estimate the variance of $\bar{q}_m$. The difference in this variance estimate compared with the variance estimate for standard multiple imputation (see Section 5.1) is due to the additional sampling from the synthetic units for fully synthetic datasets. Hence, the variance b_m between the datasets already reflects the variance within each imputation. When n is large, inferences for scalar Q can be based on t distributions with degrees of freedom $\nu_f = (m-1)(1 - \bar{u}_m/((1 + m^{-1})b_m))^2$. Derivations of these methods are presented in Raghunathan et al. (2003).

A disadvantage of this variance estimate is that it can become negative. For that reason, Reiter (2002) suggests a slightly modified variance estimator that is always positive, $T_f^* = \max(0, T_f) + \delta(\frac{n_{syn}}{n}\bar{u}_m)$, where $\delta = 1$ if $T_f < 0$ and $\delta = 0$ otherwise. Here, n_{syn} is the number of observations in the released datasets sampled from the synthetic population.

6.1.2 Multivariate estimands

Significance tests for multicomponent estimands are presented in Reiter (2005c). The derivations are based on the same ideas as those described in Section 5.1.2. Let

$\bar{\mathbf{q}}_m$, $\mathbf{b}_m$, and $\bar{\mathbf{u}}_m$ be the multivariate analogs to $\bar{q}_m$, b_m, and $\bar{u}_m$ defined in (6.1) to (6.3). Let us assume the user is interested in testing a null hypothesis of the form $\mathbf{Q} = \mathbf{Q}_0$ for a multivariate estimand with k components. Following the notation in Reiter and Raghunathan (2007), the Wald statistic for this test is given by

$$S_f = (\bar{\mathbf{q}}_m - \mathbf{Q}_0)^T \bar{\mathbf{u}}_m^{-1} (\bar{\mathbf{q}}_m - \mathbf{Q}_0)/(k(r_f - 1)), \tag{6.5}$$

where $r_f = (1 + 1/m)tr(\mathbf{B}_m \bar{\mathbf{U}}_m^{-1})/k$. The reference distribution for S_f is an F distribution, F_{k,ν_f}, with $\nu_f = 4 + (t-4)(1-(1-2/t)/r_f)^2$, where $t = k(m-1)$. Fully synthetic datasets generally require a larger number of imputations m than standard multiple imputation for nonresponse since the fraction of "missing" information is large (Reiter, 2005b). Thus, generating less than $m = 4$ fully synthetic datasets is not recommended, and I do not consider alternative degrees of freedom for $t \leq 4$ as I did in Section 5.1.2.

If $\mathbf{Q}$ contains a large number of components k, using $\bar{\mathbf{u}}_m$ can be cumbersome. As pointed out by Meng and Rubin (1992), it might be more convenient to use a likelihood ratio test in this case. Reiter (2005c) also presents the derivations for this test for fully synthetic datasets.

Again following the notation given in Reiter and Raghunathan (2007), let ψ be the vector of parameters in the analyst's model, and let $\psi^{(i)}$ be the maximum likelihood estimate of ψ computed from $D^{(i)}$, where $D^{(i)}$ is the ith imputed dataset and $i = 1, ..., m$. The analyst is interested in testing the hypothesis that $\mathbf{Q}(\psi) = \mathbf{Q}_0$, where $\mathbf{Q}(\psi)$ is a k-dimensional function of ψ. Let $\psi_0^{(i)}$ be the maximum likelihood estimate of ψ obtained from $D^{(i)}$ subject to $\mathbf{Q}(\psi) = \mathbf{Q}_0$. The log-likelihood ratio test statistic associated with $D^{(i)}$ is $L^{(i)} = 2\log f(D^{(i)}|\psi^{(i)}) - 2\log f(D^{(i)}|\psi_0^{(i)})$. Let $\bar{L} = \sum_{i=1}^m L^{(i)}/m$, $\bar{\psi} = \sum_{i=1}^m \psi^{(i)}/m$, and $\bar{\psi}_0 = \sum_{i=1}^m \psi_0^{(i)}/m$. Finally, let $\bar{L}_0 = (1/m)\sum_{i=1}^m (2\log f(D^{(i)}|\bar{\psi}) - 2\log f(D^{(i)}|\bar{\psi}_0))$, the average of the log-likelihood ratio test statistics evaluated at ψ and ψ_0. The likelihood ratio test statistic is given by

$$\hat{S}_f = \bar{L}_0/(k(\hat{r}_f - 1)), \tag{6.6}$$

where $\hat{r}_f = ((m+1)/t)(\bar{L} - \bar{L}_0)$. The reference distribution for $\hat{S}_f$ is $F_{k,\hat{\nu}_f}$, where $\hat{\nu}_f$ is defined as for ν_f using $\hat{r}_f$ instead of r_f.

6.2 Analytical validity for fully synthetic datasets

It is important to quantify the analytic usefulness of the synthetic datasets. Existing utility measures are of two types: (i) comparisons of broad differences between the original and released data and (ii) comparisons of differences in specific models between the original and released data. Broad difference measures essentially quantify some statistical distance between the distributions of the original and released data,

for example, a Kullback-Leibler or Hellinger distance. As the distance between the distributions grows, the overall quality of the released data generally drops.

A very useful measure for specific estimands is the interval overlap measure of Karr et al. (2006). For any estimand, we first compute the 95% confidence intervals for the estimand from the synthetic data, (L_s, U_s), and from the collected data, (L_o, U_o). Then, we compute the intersection of these two intervals, (L_i, U_i). The utility measure is

$$I = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_s - L_s)}. \tag{6.7}$$

When the intervals are nearly identical, corresponding to high utility, $I \approx 1$. When the intervals do not overlap, corresponding to low utility, $I = 0$. The second term in (6.7) is included to differentiate between intervals with $(U_i - L_i)/(U_o - L_o) = 1$ but different lengths. For example, for two synthetic data intervals that fully contain the collected data interval, the measure I favors the shorter interval. The synthesis is successful if we obtain large values of I for many estimands. To compute one-number summaries of utility, we can average the values of I over all estimands. This utility measure provides more information than a simple comparison of the two point estimates from the different datasets because it also considers the standard error of the estimate. Estimates with large standard errors might still have a high confidence interval overlap and from this a high data utility even if their point estimates differ considerably from each other because the confidence intervals will increase with the standard error of the estimate. For more details on this method, see Karr et al. (2006).

There do not exist published broad utility measures that account for all m synthetic datasets. The U.S. Census Bureau has adapted an approach described by Woo et al. (2009) that is based on how well one can discriminate between the original and disclosure-protected data. In this approach, the agency stacks the original and synthetic datasets in one file and estimates probabilities of being "assigned" to the original data conditional on all variables in the dataset. When the probabilities are close to 0.5 for all records in the original and synthetic data, the distributions of the variables are similar–this fact comes from the literature on propensity scores (Rosenbaum and Rubin, 1983)–so that the synthetic data have high utility. This approach is especially useful as a diagnostic for deficiencies in the synthesis methods (variables with significant coefficients in the logistic regression have different distributions in the original and synthetic data).

6.3 Disclosure risk for fully synthetic datasets

In general, the disclosure risk for fully synthetic datasets is very low since all values are synthetic values. Still, it is not necessarily zero. For example, in most establishment surveys, the probability of inclusion depends on the size of the establishment

and sometimes can be close to 1 for the largest establishments. Since the released synthetic samples will have to be stratified, too, to take advantage of the efficiency gained by stratification, the additional protection offered with the fully synthetic approach by drawing new samples from the sampling frame can be very modest for larger establishments. A possible intruder can be confident that large establishments in the released synthetic data represent establishments that were also included in the original survey. The same argument holds for the release of synthetic census data.

Besides this actual risk of disclosure, the perceived risk of disclosure also needs to be considered. The released data might look like the data from a potential survey respondent an intruder was looking for. And once the intruder thinks he identified a single respondent and the estimates are reasonably close to the true values for that unit, it is no longer important that the data are all made up. The potential respondent will feel that her privacy is at risk. Nevertheless, the disclosure risk in general will be very low since the imputation models would have to be almost perfect and the intruder faces the problem that he never knows (i) if the imputed values are anywhere near the true values and (ii) if the target record is included in one of the different synthetic samples.

For this reason, the theory on disclosure risk for fully synthetic datasets is far less developed than the theory for partially synthetic datasets (see Section 7.3). Only recently Abowd and Vilhuber (2008) have proposed some measures based on the ideas of differential privacy from the computer science literature. To understand the concept of differential privacy, we need some further definitions. Let D_{rel} be the released dataset. Let N be the hypothetical population – unknown to the intruder – from which D_{rel} was supposedly generated. According to Dwork (2006), ε-differential privacy is fulfilled if

$$\max \left| \ln \left(\frac{Pr(D_{rel}|N^1)}{Pr(D_{rel}|N^2)} \right) \right| \leq \varepsilon, \tag{6.8}$$

where ε is a predefined threshold and the maximum is taken over all N^1, N^2 that differ only in a single row. The basic idea is that if the ratio is too large, the intruder gains too much information from the released data since it is far more likely that D_{rel} was generated from N^1 and not from N^2. The data-releasing agency can decide which level of ε it is willing to accept. Abowd and Vilhuber (2008) show that this definition of disclosure risk is closely related to the risk of inferential disclosure from the SDC literature, which measures the risk by the information gain about a single respondent from the released data compared with the a priori information before the release. The paper also illustrates that synthesizing categorical variables under a multinomial/Dirichlet model can fulfill the requirements of ε-differential privacy. However, informative priors need to be incorporated in the imputation models to guarantee this strict privacy definition. All the multiple imputation combining rules developed so far are based on the assumption that noninformative priors are used in the imputation models. Charest (2010) illustrates that applying the standard combining rules for fully synthetic datasets described in Section 6.1 will lead to

biased results under the multinomial/Dirichlet model proposed by Abowd and Vilhuber (2008).

Still, the definition of ε-differential privacy is very appealing since it is the only concept that guarantees a formal level of privacy independent of the actual data – the agency only needs to select an SDC method that can guarantee ε-differential privacy and knows directly how protected the generated datasets are. Furthermore, the agency can also select the level of privacy guaranteed by defining ε. But the measure is based on the very strong assumption that the intruder knows all records in the dataset except one and measures how much information the intruder can reveal about this one record. To keep this information low, strong requirements for the SDC method are necessary, namely that the transition matrix between the observed and the released data doesn't contain any zeros (i.e., any point in the outcome space of a variable must be reachable with positive probability from any given observed value through the transition function between the original and the disclosure-protected data implicitly specified by the SDC method). For many datasets, this would mean that some very unlikely or even unrealistic events must be reachable with positive probability. Thus, the gain in data protection can come at a very high price in terms of data quality. For this reason, Machanavajjhala et al. (2008) defined (ε, δ)-probabilistic differential privacy, where $1 - \delta$ is the probability that (6.8) holds. This measure has been developed for the multinomial/Dirichlet model. Further research is necessary to investigate whether it is possible to either adjust the combining rules to allow for informative priors or to develop synthesis models that fulfill ε-differential privacy without the need to define informative priors.

6.4 Application of the fully synthetic approach to the IAB Establishment Panel

To generate fully synthetic datasets for the IAB Establishment Panel, information from the sampling frame of the Establishment Panel is necessary. I obtain this information by aggregating the German Social Security Data (GSSD) to the establishment level. From this aggregated dataset, I can sample new records that provide the basis for the generation of the synthetic datasets. As noted earlier, the German Social Security Data contain information on all employees covered by social security. The notifications of the GSSD include for every employee, among other things, the workplace and the establishment identification number. By aggregating records with the same establishment identification number, it is possible to generate establishment information from the GSSD. As I use the 1997 wave of the IAB Establishment Panel for the analysis, data are taken and aggregated from the GSSD for June 30, 1997 (see Figure 6.1 for all characteristics used). I use the establishment identification number again to match the aggregated establishment characteristics from the GSSD with the IAB Establishment Panel.

In this simulation, I only impute values for a set of variables from the 1997 wave of the IAB Establishment Panel. As it is not feasible to impute values for the millions

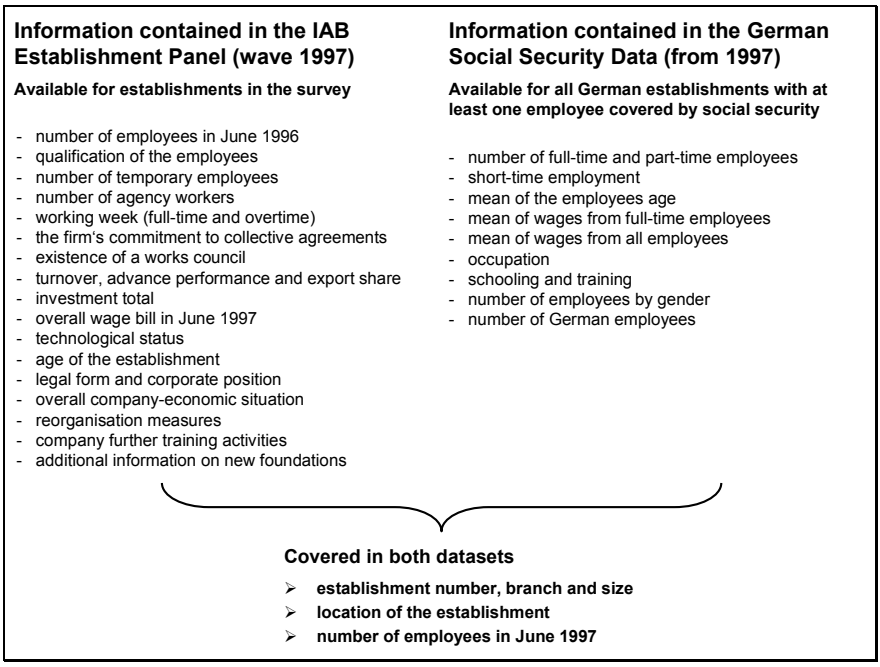

Fig. 6.1 Included variables from the IAB Establishment Panel and the German Social Security Data.

of establishments contained in the German Social Security Data for 1997, I sample from this frame using the same sampling design as for the IAB Establishment Panel: stratification by establishment size, region, and industry. Every stratum contains the same number of units as the observed data from the 1997 wave of the Establishment Panel.

Cross-tabulation of the stratum parameters for the 7,332 observations in our sample provides a matrix containing the number of observations for each stratum. A new dataset can be generated easily by drawing establishments from the German Social Security Data according to this matrix.

After matching, every dataset is structured as follows. Let N be the total number of units in the newly generated dataset; that is, the number of units in the new sample n_s plus the number of units in the panel n_p, $N = n_s + n_p$. Let X be the matrix of variables with information for all observations in N. Then X consists of the variables *establishment size* (from the GSSD), *region* and *industry*, and the other variables added from the German Social Security Data. Note that the variable *establishment size* is included in both the GSSD and the Establishment Panel. These two variables need not necessarily be identical since they are reported at different points in time. However, I use the establishment size from the GSSD as a very strong predictor when synthesizing the establishment size in the Establishment Panel. Let Y be the

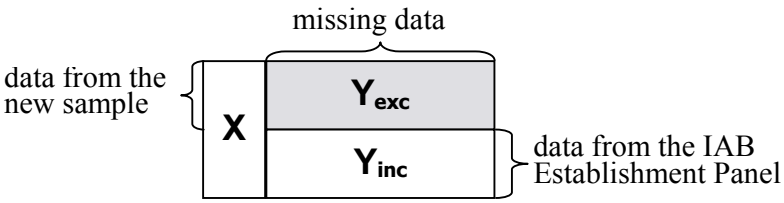

Fig. 6.2 The fully synthetic approach for the IAB Establishment Panel.

selected variables from the Establishment Panel, with $Y = (Y_{inc}, Y_{exc})$, where Y_{inc} are the observed values from the Establishment Panel and Y_{exc} are the hypothetical missing data for the newly drawn values in X (see Figure 6.2).

Now, values for the missing data can be imputed as outlined in Chapter 3 by drawing Y_{exc} from the posterior predictive distribution $f(Y_{exc}|X, Y_{inc})$ for the $N - n_p$ unobserved values of Y. After the imputation procedure, all observations from the GSSD and all originally observed values from the Establishment Panel are omitted and only the imputed values for the panel are released. Results from an analysis of these released data can be compared with the results achieved with the real data.

6.4.1 The imputation procedure

For this simulation, I only generate ten synthetic datasets. I deliberately selected a small number of imputations to allow a direct comparison of the results with the results of the partially synthetic approach described in Section 7.4. A larger number of imputations is recommended in practice. Previous research has shown that releasing large numbers of fully synthetic datasets improves synthetic data inferences (Reiter, 2005b). The usual advice for multiple imputation for missing data – release five multiply imputed datasets – tends not to work well for fully synthetic data because the fractions of "missing" information are large. Drechsler et al. (2008b) obtain higher analytic validity by generating 100 fully synthetic datasets using the two-stage imputation approach described in Chapter 9.

To generate the synthetic datasets, I use the FCS approach (see Section 3.1.2) as implemented in the software IVEware (Raghunathan et al., 2002). Since most of the continuous variables, such as *establishment size*, are heavily skewed, these variables are transformed by taking the cubic root before imputation to get rid of the skewness. In general, all variables are used as predictors in the imputation models in hopes of reducing problems from uncongeniality (Meng, 1994;, see also Section 5.3.2). In the multinomial logit model for the categorical variables, some explanatory variables are dropped for multicollinearity reasons. For the imputation procedure, I use 26 variables from the GSSD and reduce the number of panel variables to be imputed to 48 (Figure 6.1 provides a broad description of the information contained in these variables).

6.4.2 Measuring the analytical validity

To evaluate the quality of the synthetic data, I use the analysis by Zwick (2005) described in detail in Chapter 4. Comparison results from Zwick's regression run on the original data and synthetic data are presented in Table 6.1. The last column of the table measures data utility by looking at the overlap between the confidence intervals for the estimates from the original data and the confidence intervals for the estimates from the synthetic data as described in Section 6.2. All variables in the regression except for the industry dummies that are part of the sampling design are synthesized. Since all imputation models (except for some categorical variables) are based on all variables in the dataset, the imputation model for the vocational training variable contains all the variables that are used in the regression.

All estimates are close to the estimates from the real data and except for the variable *high number of maternity leaves expected*, which is not significant at any given significance level in the synthetic data, remain significant at the same level when using the synthetic data. The confidence interval overlap is high for most estimates, but it drops below 50% for four of the 13 variables. Only for the dummy variable that indicates establishments with 200 to 499 employees and the dummy variable for establishments with more than 1,000 employees are the absolute deviations between the estimates from the two datasets higher than 0.1 (0.138 and 0.202, respectively). Obviously Zwick would have come to nearly the same conclusions in his analysis if he had used the synthetic data instead of the real data. See Drechsler et al. (2008b) for a two-stage imputation approach that could further improve the quality of the synthetic data. These results indicate that valid statistical inferences

Table 6.1 Results from the vocational training regression for full synthesis.

	Original data	Synthetic data	CI overlap
Redundancies expected	0.253***	0.293***	0.848
Many employees expected on maternity leave	0.262**	0.240	0.770
High qualification need expected	0.646***	0.601***	0.227
Appren. train. reaction on skill shortages	0.113*	0.149*	0.930
Training reaction on skill shortages	0.540***	0.532***	0.620
Establishment size 20–199	0.684***	0.649***	0.857
Establishment size 200–499	1.352***	1.215***	0.457
Establishment size 500–999	1.346***	1.404***	0.382
Establishment size 1,000 +	1.955***	1.753***	0.932
Share of qualified employees	0.787***	0.812***	0.437
State-of-the-art tech. equipment	0.171***	0.186***	0.712
Collective wage agreement	0.255***	0.293***	0.901
Apprenticeship training	0.490***	0.423***	0.534
Industry, East Germany dummies	Yes		

Notes: * Significant at the 5% level,** significant at the 1% level,
*** significant at the 0.1% level.

Source: IAB Establishment Panel 1997 without newly founded establishments and establishments not represented in the GSSD; regression according to Zwick (2005).

can be achieved using the synthetic datasets, but is the confidentiality of the survey respondents guaranteed? Disclosure of potentially sensitive information could be possible, when the following two conditions are fulfilled:

1. An establishment is included in the original dataset and in at least one of the newly drawn samples.[2]
2. The original values and the imputed values for this establishment are nearly the same.

6.4.3 Assessing the disclosure risk

To determine the disclosure risk in this setting, I assume that the intruder would search for records that appear in more than one of the ten new samples. Since the intruder doesn't know if any establishment in the synthetic datasets is also included in the original dataset, she may use the probability of inclusion in the synthetic datasets as an estimator for the probability that this record is also included in the original survey.

Identifying records that were also included in the original survey is rational for the intruder since their data were used to fit the imputation models. Thus, the chance that their imputed values are close to the original values arguably is higher than for records that were not included in the original survey. Table 6.2 displays how often

Table 6.2 How many records are sampled how often in the new samples?

Occurrence in sample(s)	Number of records	Percentage
1	45,553	82.75%
2	5,600	10.17%
3	1,805	3.28%
4	873	1.59%
5	507	0.92%
6	320	0.58%
7	164	0.30%
8	99	0.18%
9	45	0.08%
10	86	0.16%
Total	55,052	100%

[2] In theory, it is possible that even units that did not participate in the survey will face an increased risk of disclosure if they are included in the released synthetic datasets and their generated values are close to the true values. Since I do not have any information regarding the survey questions for these units, I can only compare the synthetic values with the true values for the survey respondents. Arguably, the risk of imputing values that are too close to the true values is higher for the survey respondents since their data were used to fit the models. Thus, if the risks for these units are low, they should be low for those units that did not participate in the survey, too.

Table 6.3 Establishments from the IAB Establishment Panel that also occur in at least one of the new samples.

Occurrence in sample(s)	number of records	percentage
None	4,469	61.0%
1	1,091	14.9%
2	535	7.3%
3	362	4.9%
4	275	3.8%
5	199	2.7%
6	144	2.0%
7	89	1.2%
8	53	0.7%
9	32	0.4%
10	83	1.1%
Total	7,332	100%

different records occur in the synthetic samples. Overall, 55,052 establishments are sampled in the synthetic datasets. The vast majority are sampled only once or twice. Only roughly 7% of the establishments are sampled at least three times, and less than 1% are sampled more than six times. But even if the intruder is able to identify records that are sampled more than once, which in itself is a difficult task, since almost all values are imputed and thus differ from sample to sample, he cannot be sure whether this record really is included in the original survey. Table 6.3 displays how often the records from the original survey actually occur in the synthetic samples. Of the establishments included in the original survey, 61.0% do not occur in any of the ten new drawn samples, 14.9% are contained in one of the ten samples, and only 5.5% can be found more than five times. Larger establishments have a higher probability of inclusion in the original survey (for some of the cells of the stratification matrix, this probability is close to one). Since I use the same sampling design for drawing new establishments for our synthetic datasets, this means that larger establishments also have a higher probability of being included in the original survey and in at least one of the new samples. Keeping that in mind, having only 25% of establishments with 200 to 999 employees and 49% of establishments with 1000+ employees in at least one of the new samples is a very good result in terms of data confidentiality (see Figure 6.3).

Comparing Tables 6.2 and 6.3, it is obvious that only for the records that occur in all ten datasets the probability that these records are also included in the original survey is very high. Of these establishments, 96.5% (83 of the 86 records) are contained in the original survey. But this probability decreases quickly. It is 71.1%, 53.5%, and 54.3% for establishments that occur in nine, eight and seven samples, respectively. For establishments that occur less than seven times, the probability is always lower than 50%.

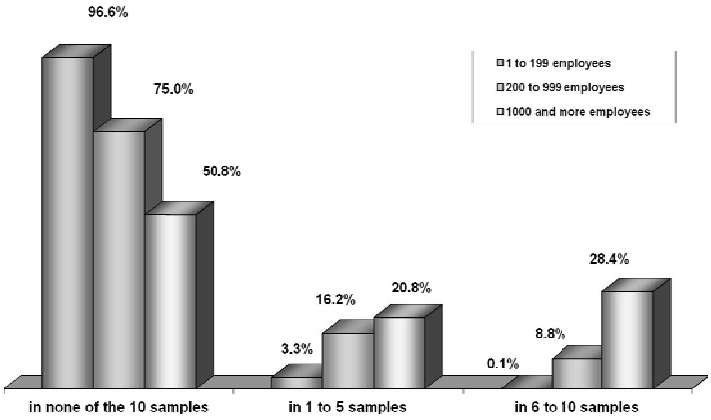

Fig. 6.3 Occurrence of establishments already included in the original survey by establishment size.

But even if a record is correctly identified, the intruder will only benefit from the identification if the imputed values of these establishments are close to the original ones. The second step of my evaluation therefore takes a closer look at the establishments from the survey that appear at least once in the newly drawn samples. Using only these establishments, the differences between original and imputed values can be detected. For each synthetic record that is also included in the original survey, I compare the imputed value with the true value. Binary variables tend to have a matching rate between 60% and 90% (i.e., for 60% to 90% of these synthetic records, the imputed binary value is the same as the true value from the survey). Multiple-response questions with few categories show a high rate of identical answers in the total item block, too. But with an increase in the number of categories, this rate decreases rapidly. For example, for an imputed multiple-response variable consisting of four categories, the probability of having the same values for all four categories is about 57%. This probability decreases to about six percent if the number of categories increases to 13.

Imputed numeric variables always differ more or less from the original value. To evaluate the uncertainty for an intruder wanting to identify an establishment using the imputed data, I examine the variable *establishment size* for the 83 establishments that appear in all ten datasets. The average relative difference between the imputed and the original values is 21%. A plot of the distribution of the relative difference for each record in each synthetic dataset shows that there are outliers for which the imputed values are two, three, or even four times higher than the original ones (see Figure 6.4). Thus, for an intruder who wants to identify an establishment using his knowledge of the true size of the establishment, the imputed variable *establishment size* will hardly be of any use.

Summing up the second step, I find that for establishments that are represented in both datasets, up to 90% of some imputed binary variables are identical to the

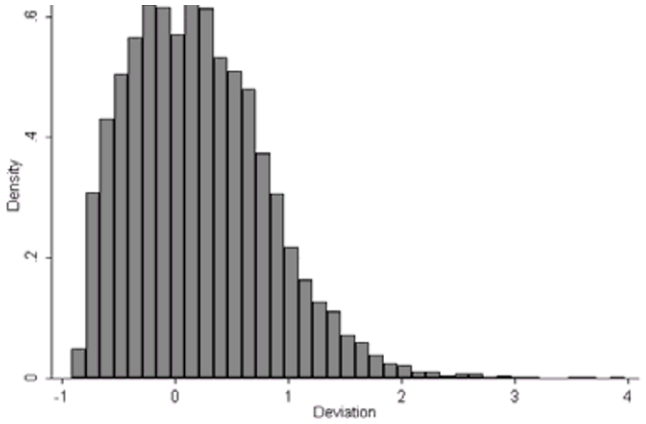

Fig. 6.4 Histogram of the relative difference between original and imputed values for the variable *establishment size*.

original values. But just one binary variable won't be sufficient to identify a single establishment. Using more binary variables, the risk of identical values will decrease quickly. If, for example, we assume the intruder needs five binary variables for identification and the variables are independently distributed, the risk will be $0.9^5 = 0.59$. Normally an intruder needs variables with more information than just two categories for a successful reidentification. But as shown for the variable *establishment size*, the chance of identifying an establishment by combining information from numeric and categorical variables is very low.

These results together with the results for the data utility in Section 6.4.2 indicate that a release of the described subset of the data would be possible. Of course, the data utility for different estimates should be evaluated in detail for different kinds of estimates before an actual release.

Chapter 7
Partially Synthetic Datasets[1]

As of this writing, no agency has adopted the fully synthetic approach discussed in the previous chapter, but some agencies have adopted a variant of Rubin's original approach suggested by Little (1993): release datasets comprising the units originally surveyed with some collected values, such as sensitive values or values of key identifiers, replaced with multiple imputations. These are called *partially synthetic datasets*. For example, the U.S. Federal Reserve Board protects data in the Survey of Consumer Finances by replacing large monetary values with multiple imputations (Kennickell, 1997). In 2007, the U.S. Census Bureau released a partially synthetic, public use file for the Survey of Income and Program Participation (SIPP) that includes imputed values of social security benefits information and dozens of other highly sensitive variables (`http://www.census.gov/sipp/synth_data.html`). The Census Bureau also protects the identities of people in group quarters (e.g., prisons, shelters) in the public use files of the American Community Survey by replacing demographic data for people at high disclosure risk with imputations. The latest release of a synthetic data product by the Census Bureau is a synthetic version of the Longitudinal Business Database (Kinney et al., 2011) that is available as a public use dataset through the VirtualRDC's Synthetic Data Server located at Cornell University (`http://www.vrdc.cornell.edu/news/data/lbd-synthetic-data/`). Partially synthetic, public use datasets are in the development stage in the U.S. for the Longitudinal Employer–Household Dynamics survey and the American Community Survey veterans and full sample data.

7.1 Inference for partially synthetic datasets

Following Reiter (2003, 2004), let $Z_j = 1$ if unit j is selected to have any of its observed data replaced, and let $Z_j = 0$ otherwise. Let $Z = (Z_1, \ldots, Z_s)$, where s is the number of records in the observed data. Let $Y = (Y_{rep}, Y_{nrep})$ be the data collected

[1] Most of this chapter is taken from Drechsler et al. (2008a) and Drechsler and Reiter (2008).

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_7,

in the original survey, where Y_{rep} includes all values to be replaced with multiple imputations and Y_{nrep} includes all values not replaced with imputations. Let $Y_{rep}^{(i)}$ be the replacement values for Y_{rep} in synthetic dataset i. Each $Y_{rep}^{(i)}$ is generated by simulating values from the posterior predictive distribution $f(Y_{rep}^{(i)}|Y,Z)$, or some close approximation to the distribution such as those of Raghunathan et al. (2001). The agency repeats the process m times, creating $D^{(i)} = (Y_{nrep}, Y_{rep}^{(i)})$ for $i = 1,\ldots,m$, and releases $\mathbf{D} = \{D^{(1)},\ldots,D^{(m)}\}$ to the public.

7.1.1 Univariate estimands

To get valid inferences, secondary data users can use the combining rules presented by Reiter (2003). Let Q be an estimand, such as a population mean or regression coefficient. Suppose that, given the original data, the analyst estimates Q with some point estimator q and the variance of q with some estimator u. For simplicity, assume that there are no data with items missing in the observed dataset. Generating partially synthetic datasets if the original data are subject to nonresponse is discussed in Chapter 8. Let $q^{(i)}$ and $u^{(i)}$ be the values of q and u in synthetic dataset $D^{(i)}$ for $i = 1,...,m$. The analyst computes $q^{(i)}$ and $u^{(i)}$ by acting as if each $D^{(i)}$ is the genuine data. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q^{(i)}/m, \tag{7.1}$$

$$b_m = \sum_{i=1}^{m} (q^{(i)} - \bar{q}_m)^2/(m-1), \tag{7.2}$$

$$\bar{u}_m = \sum_{i=1}^{m} u^{(i)}/m. \tag{7.3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_p = b_m/m + \bar{u}_m \tag{7.4}$$

to estimate the variance of $\bar{q}_m$.

Similar to the variance estimator for multiple imputation of missing data, b_m/m is the correction factor for the additional variance due to using a finite number of imputations. However, the additional b_m necessary in the missing-data context is not necessary here since $\bar{u}_m$ already captures the variance of Q given the observed data. This is different in the missing-data case, where $\bar{u}_m$ is the variance of Q given the completed data and $\bar{u} + b_m$ is the variance of Q given the observed data.

When n is large, inferences for scalar Q can be based on t distributions with degrees of freedom $\nu_p = (m-1)(1+\bar{u}_m/(b_m/m))^2$. Note that the variance estimate

T_p can never be negative, so no adjustments are necessary for partially synthetic datasets.

7.1.2 Multivariate estimands

Significance tests for multicomponent estimands are presented by Reiter (2005c). The derivations are based on the same ideas as those described in Section 5.1.2. Let $\bar{\mathbf{q}}_m$, $\mathbf{b}_m$, and $\bar{\mathbf{u}}_m$ be the multivariate analogs to $\bar{q}_m$, b_m, and $\bar{u}_m$ defined in (7.1) to (7.3). Let us assume the user is interested in testing a null hypothesis of the form $\mathbf{Q} = \mathbf{Q}_0$ for a multivariate estimand with k components. Following the notation in Reiter and Raghunathan (2007), the Wald statistic for this test is given by

$$S_p = (\bar{\mathbf{q}}_m - \mathbf{Q}_0)^T \bar{\mathbf{u}}_m^{-1} (\bar{\mathbf{q}}_m - \mathbf{Q}_0)/(k(1+r_p)), \tag{7.5}$$

where $r_p = (1/m)tr(\mathbf{B}_m \bar{\mathbf{U}}_m^{-1})/k$. The reference distribution for S_p is an F distribution, F_{k,ν_p}, with $\nu_p = 4 + (t-4)(1+(1-2/t)/r_p)^2$, where $t = k(m-1)$. Synthetic datasets generally require a larger number of imputations m than standard multiple imputation for nonresponse since the fractions of "missing" information tend to be large. Thus, generating less than $m = 4$ synthetic datasets is not recommended, and I do not consider alternative degrees of freedom for $t \leq 4$ as I did in Section 5.1.2.

If $\mathbf{Q}$ contains a large number of components k, using $\bar{\mathbf{u}}_m$ can be cumbersome. As pointed out by Meng and Rubin (1992), it might be more convenient to use a likelihood ratio test in this case. Reiter (2005c) also presents the derivations for this test for partially synthetic datasets.

Again following the notation given in Reiter and Raghunathan (2007), let ψ be the vector of parameters in the analyst's model, and let $\psi^{(i)}$ be the maximum likelihood estimate of ψ computed from $D^{(i)}$, where $D^{(i)}$ is the ith imputed dataset and $i = 1, ..., m$. The analyst is interested in testing the hypothesis that $\mathbf{Q}(\psi) = \mathbf{Q}_0$, where $\mathbf{Q}(\psi)$ is a k-dimensional function of ψ. Let $\psi_0^{(i)}$ be the maximum likelihood estimate of ψ obtained from $D^{(i)}$ subject to $\mathbf{Q}(\psi) = \mathbf{Q}_0$. The log-likelihood ratio test statistic associated with $D^{(i)}$ is $L^{(i)} = 2\log f(D^{(i)}|\psi^{(i)}) - 2\log f(D^{(i)}|\psi_0^{(i)})$. Let $\bar{L} = \sum_{i=1}^m L^{(i)}/m$, $\bar{\psi} = \sum_{i=1}^m \psi^{(i)}/m$, and $\bar{\psi}_0 = \sum_{i=1}^m \psi_0^{(i)}/m$. Finally, let $\bar{L}_0 = (1/m)\sum_{i=1}^m (2\log f(D^{(i)}|\bar{\psi}) - 2\log f(D^{(i)}|\bar{\psi}_0))$, the average of the log-likelihood ratio test statistics evaluated at ψ and ψ_0. The likelihood ratio test statistic is given by

$$\hat{S}_p = \bar{L}_0/(k(1+\hat{r}_p)), \tag{7.6}$$

where $\hat{r}_p = (1/t)(\bar{L} - \bar{L}_0)$. The reference distribution for $\hat{S}_p$ is $F_{k,\hat{\nu}_p}$, where $\hat{\nu}_p$ is defined as for ν_p using $\hat{r}_p$ instead of r_p.

7.2 Analytical validity for partially synthetic datasets

To evaluate the analytical validity of partially synthetic datasets, we can use the same methods as for fully synthetic datasets, namely measuring the confidence interval overlap between confidence intervals obtained from the synthetic data and confidence intervals obtained from the original data or measuring how well one can discriminate between the original and the synthetic data based on the ideas of propensity score matching. See Section 6.2 for details.

7.3 Disclosure risk for partially synthetic datasets

The disclosure risk is higher for partially synthetic datasets than it is for fully synthetic datasets, especially if the intruder knows that some unit participated in the survey, since true values remain in the dataset and imputed values are generated only for the survey participants and not for the whole population. Thus, for partially synthetic datasets, assessing the risk of disclosure is as important an evaluation step as assessing the data utility. It is essential that the agency identifies and synthesizes all variables that bear a risk of disclosure. A conservative approach would be to also impute all variables that contain the most sensitive information. Once the synthetic data are generated, careful checks are necessary to evaluate the disclosure risk for these datasets. Only if the datasets prove to be useful both in terms of data utility and in terms of disclosure protection should a release be considered.

As noted above, the risk of disclosure significantly increases if the intruder knows who participated in a survey. Thus, it is important to distinguish between a scenario in which the intruder knows that the target he is looking for is in the data and a scenario in which the intruder has some external information but does not know whether any of the targets he is looking for are actually included in the survey. For most surveys, the latter case will be a more realistic assumption, but there might be situations in which it is publicly known who participated in a survey or the agency might want to release a complete synthetic population. I therefore start by presenting methods to evaluate the disclosure risk under the conservative assumption that the intruder has full information about survey participation and afterwards discuss necessary extensions to account for the additional sampling uncertainty if the intruder does not have any response knowledge. Both methods only evaluate the risk of *identification disclosure* (i.e., the risk that a unit is correctly identified in the released data). Methods to evaluate the risk of *inferential disclosure* (i.e., the amount of additional information an intruder might obtain about a unit for which she already knows that it participated in the survey) still need to be developed for partially synthetic datasets.

7.3.1 Ignoring the uncertainty from sampling

To evaluate disclosure risks if the intruder knows which units are included in the released data, we can compute probabilities of identification by following the approach of Reiter and Mitra (2009). Related approaches are described by Duncan and Lambert (1989), Fienberg et al. (1997), and Reiter (2005a). Roughly, in this approach we mimic the behavior of an ill-intentioned user of the released data who possesses the true values of the quasi-identifiers for selected target records (or even the entire database). To illustrate, suppose the malicious user has a vector of information, $\mathbf{t}$, on a particular target unit in the population corresponding to a unit in the m released simulated datasets, $\mathbf{D} = \{D^{(1)}, \ldots, D^{(m)}\}$. Let t_0 be the unique identifier (e.g., establishment name) of the target, and let d_{j0} be the (not released) unique identifier for record j in $\mathbf{D}$, where $j = 1, \ldots, s$. Let M be any information released about the simulation models.

The malicious user's goal is to match unit j in $\mathbf{D}$ to the target when $d_{j0} = t_0$ and not to match when $d_{j0} \neq t_0$ for any $j \in \mathbf{D}$. Let J be a random variable that equals j when $d_{j0} = t_0$ for $j \in \mathbf{D}$ and equals $s+1$ when $d_{j0} = t_0$ for some $j \notin \mathbf{D}$. The malicious user thus seeks to calculate the $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$ for $j = 1, \ldots, s+1$. He then would decide whether or not any of the identification probabilities for $j = 1, \ldots, s$ are large enough to declare an identification. Note that in this scenario $Pr(J = s+1|\mathbf{t}, \mathbf{D}, M) = 0$ because the intruder knows that the target record he is looking for is included in the released data. Because the malicious user does not know the actual values in Y_{rep}, he should integrate over its possible values when computing the match probabilities. Hence, for each record in $\mathbf{D}$, we compute

$$Pr(J = j|\mathbf{t}, \mathbf{D}, M) = \int Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{rep}, M) Pr(Y_{rep}|\mathbf{t}, \mathbf{D}, M) dY_{rep}. \quad (7.7)$$

This construction suggests a Monte Carlo approach to estimating each $Pr(J = j|\mathbf{t}, \mathbf{D}, M)$. First, sample a value of Y_{rep} from $Pr(Y_{rep}|\mathbf{t}, \mathbf{D}, M)$. Let Y^{new} represent one set of simulated values. Second, compute $Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{rep} = Y^{new}, M)$ using exact or, for continuous synthesized variables, distance-based matching assuming Y^{new} are collected values. This two-step process is iterated R times, where ideally R is large, and (1) is estimated as the average of the resultant R values of $Pr(J = j|\mathbf{t}, \mathbf{D}, Y_{rep} = Y^{new}, M)$. When M has no information, the malicious user can treat the simulated values as plausible draws of Y_{rep}.

To illustrate, suppose that region and employee size are the only quasi-identifiers in a survey of establishments. A malicious user seeks to identify an establishment in a particular region of the country with 125 employees. The malicious user knows that this establishment is in the sample. Suppose that the agency releases m datasets after simulating only employment size, without releasing information about the imputation model. In each $D^{(i)}$, the malicious user would search for all establishments matching the target on region and having synthetic employee size within some interval around 125, say 110 to 140. The agency selects the intervals for employment size based on its best guess of the amount of uncertainty that intruders would be willing

to tolerate when estimating true employee sizes. Let $N^{(i)}$ be the number of records in $D^{(i)}$ that meet these criteria. When no establishments with all of those characteristics are in $D^{(i)}$, set $N^{(i)}$ equal to the number of establishments in the region (i.e., match on all non-simulated quasi-identifiers). For any j,

$$Pr(J = j|\mathbf{t},\mathbf{D},M) = (1/m)\sum_i (1/N^{(i)})(Y_j^{new,i} = \mathbf{t}), \tag{7.8}$$

where $(Y_j^{new,i} = \mathbf{t}) = 1$ when record j is among the $N^{(i)}$ matches in $D^{(i)}$ and equals zero otherwise. Similar computations arise when simulating region and employee size: the malicious user exactly matches on the simulated values of region and distance-based matches on employee size to compute the probabilities.

Following Reiter (2005a) and Drechsler and Reiter (2008), I quantify disclosure risk with summaries of these identification probabilities. It is reasonable to assume that the malicious user selects as a match for $\mathbf{t}$ the record j with the highest value of $Pr(J = j|\mathbf{t},\mathbf{D},M)$, if a unique maximum exists. I consider three risk measures: the *expected match risk*, the *true match risk*, and the *false match rate*. To calculate them, we need some further definitions. Let c_j be the number of records in the dataset with the highest match probability for the target t_j for $j = 1,...,s$; let $I_j = 1$ if the true match is among the c_j units and $I_j = 0$ otherwise. Let $K_j = 1$ when $c_j I_j = 1$ and $K_j = 0$ otherwise. The *expected match risk* can now be defined as $\sum_j (1/c_j)I_j$. When $I_j = 1$ and $c_j > 1$, the contribution of unit j to the expected match risk reflects the intruder randomly guessing at the correct match from the c_j candidates. The *true match risk* equals $\sum_j K_j$. Finally, let $F_j = 1$ when $c_j(1-I_j) = 1$ and $F_j = 0$ otherwise, and let s equal the number of records with $c_j = 1$. The *false match rate* equals $\sum F_j/s$. It is important to note that these summary statistics are helpful to summarize the overall disclosure risk for the complete data, but the real advantage of the suggested measures is the fact that the identification probabilities are calculated on the record level. This enables disclosure risk evaluations for specified subgroups of the data. In some situations, only a few records in the dataset might be correctly identified but all identified records belong to the same subgroup. In this case, an overall measure that indicates a low disclosure risk might be misleading since the risk of disclosure, for example for the largest establishments in the dataset, might still be very high.

7.3.2 Accounting for the uncertainty from sampling

If the intruder does not know if the target she is looking for participated in the survey, the fact that the survey usually only comprises a sample of the population adds an additional layer of protection to the released data. In this case we can use the extensions to the measures described above suggested by Drechsler and Reiter (2008). We simply have to replace $N_{\mathbf{t},i}$ in (7.8) with $F_{\mathbf{t}}$, the number of records in the population that match the target on region and establishment size in the example above. When the intruder and the agency do not know $F_{\mathbf{t}}$, it can be estimated using the

approach in Elamir and Skinner (2006), which assumes that the population counts follow an all-two-way-interactions log-linear model. The agency can determine the estimated counts, $\hat{F}_{\mathbf{t}}$, by fitting this log-linear model with D_{obs}. Alternatively, since D_{obs} is in general not available to intruders, the agency can fit a log-linear model with each D_i, resulting in the estimates $\hat{F}_{\mathbf{t},i}$ for $i = 1,\ldots,m$. Note that in this scenario $Pr(J = s+1|\mathbf{t},\mathbf{D},M) = 1 - \sum_{j=1}^{s} Pr(J = j|\mathbf{t},\mathbf{D},M)$.

For some target records, the value of $N_{\mathbf{t},i}$ might exceed $F_{\mathbf{t}}$ (or $\hat{F}_{\mathbf{t}}$ if it is used). It should not exceed $\hat{F}_{\mathbf{t},i}$ since $\hat{F}_{\mathbf{t},i}$ is required to be at least as large as $N_{\mathbf{t},i}$. For such cases, we presume that the intruder sets $Pr(J = s+1|\mathbf{t},\mathbf{D},M) = 0$ and picks one of the matching records at random. To account for this case, we can rewrite (7.8) for $j = 1,\ldots,s$ as

$$Pr(J = j|\mathbf{t},\mathbf{D},M) = (1/m)\sum_{i} \min\left(1/F_{\mathbf{t}}, 1/N_{\mathbf{t},i}\right)\left(Y_{ij}^{\text{new}} = \mathbf{t}\right). \tag{7.9}$$

We can use the three summary statistics of the identification probabilities described in Section 7.3.1, with the important difference that we also have to consider $Pr(J = s+1|\mathbf{t},\mathbf{D},M)$, the probability for a match outside the sample. In many cases, this will be the highest match rate. It is reasonable to assume that the intruder does not match whenever $Pr(J = s+1|\mathbf{t},\mathbf{D},M)$ is the maximum probability for the target. If this assumption is considered too strong, the data-disseminating agency can define a threshold γ and assume that the intruder matches to the released data only when $Pr(J = s+1|\mathbf{t},\mathbf{D},M) \leq \gamma$, where $0 \leq \gamma \leq 1$.

7.4 Application of the partially synthetic approach to the IAB Establishment Panel

To achieve results that can be compared with the results in Section 6.4, I use the same subset of variables from the 1997 wave as in the fully synthetic application (see Section 6.4 for a description of the variables selected).

For the partially synthetic datasets, I replace only two variables (the *number of employees* and the *industry*, coded in 16 categories) with synthetic values. If the data should actually be released to the public, some other variables would have to be synthesized, too. Identifying all the variables that provide a potential disclosure risk is an important and labor-intensive task. Nevertheless, the two variables mentioned above definitely impose a high risk of disclosure since they are easily available in public databases and especially large firms can be identified without difficulty using only these two variables. I define a multinomial logit model for the imputation of the industry code and a linear model stratified by four establishment size classes defined by quartiles for the number of employees. For the partially synthetic datasets, I use the same number of variables in the imputation model as in the fully synthetic data example (26 from the German Social Security Data (GSSD), 48 from the establishment panel), but the original sample is used and no additional samples are

drawn from the GSSD. As in the fully synthetic data example, I generate ten synthetic datasets to allow a direct comparison of the results with the results in Section 6.4.2.

7.4.1 Measuring the analytical validity

For an evaluation of the utility of the partially synthetic data, I compare analytic results achieved with the original data with results from the synthetic data. The regression results in Table 7.1 are again based on the analysis by Zwick (2005) described in detail in Chapter 4.[2]

All estimates are very close to the estimates from the real data, and except for the variables *many employees expected on maternity leave* and *apprenticeship training reaction on skill shortages*, for which the significance level increases from 1% to 0.1% and from 5% to 1% respectively, remain significant at the same level when using the synthetic data. With an average of 0.925 over all 13 estimates, the confidence interval overlap is very high. Only the effect of the largest establishment size class is slightly underestimated, leading to a reduced overlap of 0.685. For all other estimates, the overlap is above 0.85, indicating very high quality for the synthetic

Table 7.1 Results from the vocational training regression for partial synthesis.

	Original data	Synthetic data	CI overlap
Redundancies expected	0.250***	0.259***	0.956
Many employees expected on maternity leave	0.267**	0.316***	0.869
High qualification need expected	0.648***	0.653***	0.982
Appren. train. reaction on skill shortages	0.115*	0.121**	0.969
Training reaction on skill shortages	0.539***	0.547***	0.962
Establishment size 20–199	0.682***	0.695***	0.920
Establishment size 200–499	1.350***	1.335***	0.936
Establishment size 500–999	1.344***	1.344***	0.994
Establishment size 1,000 +	1.956***	1.754***	0.685
Share of qualified employees	0.789***	0.803***	0.948
State-of-the-art tech. equipment	0.170***	0.175***	0.962
Collective wage agreement	0.257***	0.275***	0.894
Apprenticeship training	0.488***	0.496***	0.953
Industry, East Germany dummies	Yes		

Notes: * Significant at the 5% level,** significant at the 1% level,
*** significant at the 0.1% level.

Source: IAB Establishment Panel 1997 without newly founded establishments and establishments not represented in the GSSD; regression according to Zwick (2005).

[2] For simplicity, I impute all missing values first and treat one fully imputed dataset as the original data. Since missing rates are low for all variables used in the regression, results for the original data only change in the third digit compared with the results in Table 6.1. See Chapter 8 on how to correctly generate synthetic datasets from data that are subject to nonresponse.

data. Obviously, Zwick would have come to the same conclusions in his analysis if he had used the partially synthetic data instead of the real data.

7.4.2 Assessing the disclosure risk

To evaluate the risk of disclosure, I apply the disclosure risk measures described in Section 7.3.1 (i.e. I assume, the intruder knows, who participated in the survey). I further assume the intruder knows the true values for the number of employees and industry. This is a conservative scenario but gives, in some sense, an upper bound on the risk for this level of intruder knowledge. For an application of the disclosure risk measures without response knowledge, see Section 8.3.5. Intruders might also know other variables in the file, in which case the agency may need to synthesize them as well. The intruder computes probabilities using the approach outlined in Section 7.3.1. I assume that the agency does not reveal the synthesis model to the public, so that the only information in M is that establishment size and industry were synthesized. For a given target t, records from each $D^{(i)}$ must meet two criteria to be possible matches. First, the record's synthetic industry code must exactly match the target's true industry code. Second, the record's synthetic number of employees lies within an agency-defined interval around the target's true number of employees. Acting as the agency, I define the interval as follows. I divide the true number of employees (transformed by taking the cubic root) into 20 quantiles and calculate the standard deviation of the number of employees within each quantile. The interval is $t_e \pm sd_s$, where t_e is the target's true value and sd_s is the standard deviation of the quantile in which the true value falls. When there are no synthetic records that fulfill both matching criteria, the intruder matches only on the industry code. I use 20 quantiles because this is the largest number of groups that guarantees at least some variation within each group. Using a larger number of quantiles results in groups with only one value of employment, which forces exact matching for targets in those quantiles. On the other hand, using a small number of quantiles does not differentiate adequately between small and large establishments. For small establishments, potential matches should deviate only slightly from the original values. For large establishments, higher deviations are acceptable.

Given this matching scenario, both, the expected match risk and the true match risk would be 139 (i.e. the intruder would get 139 true correct single matches from the 7,332 records in her target file). The false match rate would be 98.1%. There is no obvious common pattern for the identified records. Neither for the region nor for the industry does the distribution of the identified records differ significantly from the distribution in the underlying data. The identified records consist of very small and very large establishments. However, as one might expect, the actual risk of disclosure depends on establishment size. While only 1.38% of the establishments with less than 100 employees are identified, this rate increases to 1.87% for establishments with 100–1,000 employees and to 5.21% for establishments with more than 1,000 employees. Considering the fact that the intruder matches on 7,332

records and never knows which of the 7,330 single matches she obtains are actually correct matches, the risk is very moderate, especially since these measures are based on the very conservative assumptions that (i) the intruder knows who participated in the survey and (ii) has exact information on the industry code and the establishment size for all the survey participants. If the agency deems the risk of disclosure still too high, it might broaden the industry codes or suppress this information completely in the released file. Another possibility would be to use less detailed models for the large establishments to ensure a higher level of perturbation for these records. As an alternative, the agency might consider releasing fully synthetic datasets instead.

7.5 Pros and cons of fully and partially synthetic datasets

Obviously there are advantages and disadvantages for the partially and the fully synthetic approach. The fully synthetic approach provides a very high level of disclosure protection, rendering the identification of single units in the released data almost impossible. Partially synthetic datasets cannot offer such a high level of protection per se since true values remain in the data and synthetic values are only generated for units that participated in the survey. This means that evaluating the disclosure risk is as important a step as evaluating the data quality for partially synthetic datasets.

Nevertheless, partially synthetic datasets have the important advantage that in general the data utility will be higher since only for some variables do the true values have to be replaced with imputed values, so by definition the joint distribution for all the unchanged variables will be exactly the same as in the original dataset. The quality of the synthetic datasets will depend highly on the quality of the underlying models, and for some variables it will be very hard to define good models, especially if logical constraints and skip patterns should be preserved. But if these variables do not contain any sensitive information or information that might help identify single respondents, why bother to find these models? Why bother to perturb these variables in the first place? Furthermore, the risk of biased imputations will increase with the number of variables that are imputed if the SRMI approach (see Section 3.1.2) is used for imputations. If one of the variables is imputed based on a *bad* model, the biased imputed values for that variable could be the basis for the imputation of another variable and this variable again could be used for the imputation of another one and so on. Thus, a small bias could increase to a really problematic bias over the imputation process.

A comparison of the results in Sections 6.4.2 and 7.4.1 underlines these thoughts. The partially synthetic datasets provide higher data quality in terms of lower deviation from the true estimates and higher confidence interval overlap between estimates from the original data and estimates from the synthetic data for almost all estimates. Still, this increase in data utility comes at the price of an increase in the risk of disclosure. Although the disclosure risk for fully synthetic datasets might not be zero, the disclosure risk will definitely be higher if true values remain in the

dataset and the released data are based only on survey participants. Thus, it is important to make sure that all variables that might lead to disclosure are imputed in such a way that confidentiality is guaranteed. This means that a variety of disclosure risk checks are necessary before the data can be released, but this is a problem common to all perturbation methods that are based only on the information from the survey respondents. Agencies willing to release synthetic public use files will have to consider carefully which approach best suits their datasets. If the data consist only of a small number of variables and imputation models are easy to set up, the agencies might consider releasing fully synthetic datasets since these datasets will provide the highest confidentiality protection, but if there are many variables in the data considered for release and the data contain a lot of skip patterns, logical constraints, and questions that are asked of only a small subgroup of survey respondents, the agencies might be better off releasing partially synthetic datasets and include a detailed disclosure risk study in their evaluation of the quality of the datasets considered for release.

Chapter 8
Multiple Imputation for Nonresponse and Statistical Disclosure Control[1]

Most if not all surveys are subject to item nonresponse, and even registers can contain missing values, if implausible values are set to missing during the data-editing process. Since the generation of synthetic datasets is based on the ideas of multiple imputation, it is reasonable to use the approach to impute missing values and generate synthetic values simultaneously. At first glance, it seems logical to impute missing values and generate synthetic values in one step using the same model as for the originally observed values. However, as Reiter (2004) points out, this can lead to biased imputations if only a subset of the data (e.g., the income for units with income above $100,000) should be replaced with synthetic values but the imputation model for the missing values is based on the entire dataset. To allow for different models, Reiter (2004) suggests imputation in two stages. In the first stage, all missing values are imputed m times using the standard multiple-imputation approach for nonresponse (see Chapter 5). In the second stage, all values that need to be replaced are synthesized r times in every first-stage nest, leading to a total of $M = m * r$ datasets that are released to the public. Each released dataset includes a label indicating the first-stage imputed dataset from which it was generated. As of this writing, only the combining rules for partially synthetic datasets have been derived. Developing the correct combining rules for fully synthetic datasets if the original data are subject to nonresponse is an area for future research.

8.1 Inference for partially synthetic datasets when the original data are subject to nonresponse

The two-stage imputation described above generates two sources of variability: first, when missing values are imputed, and second, when sensitive or identifying variables are replaced with synthetic values. Neither the combining rules for the imputation of missing values described in Section 5.1 nor those for synthetic datasets

[1] Most of this chapter is taken from Drechsler (2011b).

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_8,

described in Sections 6.1 and 7.1 correctly reflect these two sources of variability. Reiter (2004) derived the combining rules necessary to obtain valid inferences in this two-stage setting for partially synthetic datasets.

8.1.1 Univariate estimands

Again, let Q be an estimand, such as a population mean or regression coefficient. Suppose that, given the original data, the analyst estimates Q with some point estimator q and the variance of q with some estimator u. Let $q^{(i,j)}$ and $u^{(i,j)}$ be the values of q and u in synthetic dataset $D^{(i,j)}$ for $i = 1,...,m$ and $j = 1,...,r$. The analyst computes $q^{(i,j)}$ and $u^{(i,j)}$ by acting as if each $D^{(i,j)}$ is the genuine data. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_M = \sum_{i=1}^{m}\sum_{j=1}^{r} q^{(i,j)}/(mr) = \sum_{i=1}^{m} \bar{q}^{(i)}/m, \tag{8.1}$$

$$\bar{b}_M = \sum_{i=1}^{m}\sum_{j=1}^{r} (q^{(i,j)} - \bar{q}^{(i)})^2/m(r-1) = \sum_{i=1}^{m} b^{(i)}/m, \tag{8.2}$$

$$B_M = \sum_{i=1}^{m} (\bar{q}^{(i)} - \bar{q}_M)^2/(m-1), \tag{8.3}$$

$$\bar{u}_M = \sum_{i=1}^{m}\sum_{j=1}^{r} u^{(i,j)}/(mr). \tag{8.4}$$

The analyst then can use $\bar{q}_M$ to estimate Q and

$$T_P = (1+1/m)B_M - \bar{b}_M/r + \bar{u}_M \tag{8.5}$$

to estimate the variance of $\bar{q}_M$.

When n is large, inferences for scalar Q can be based on t distributions with degrees of freedom

$$\nu_P = \left(\frac{((1+1/m)B_M)^2}{(m-1)T_M^2} + \frac{(\bar{b}_M/r)^2}{m(r-1)T_M^2} \right)^{-1}. \tag{8.6}$$

Similar to the variance estimate for fully synthetic datasets, T_P can become negative since $\bar{b}_M/r$ is subtracted. In this case, Reiter (2008b) suggests using the conservative variance estimator $T_P^{adj} = (1+1/m)B_m + \bar{u}_M$. This estimator is equivalent to the variance estimator for multiple imputation for missing data. Consequently, the degrees of freedom are given by

$$\nu_P^{adj} = (m-1)(1 + m\bar{u}_M/((m+1)B_M))^2. \tag{8.7}$$

Generally, negative variances can be avoided by increasing m and r.

8.1.2 Multivariate estimands

Significance tests for multicomponent estimands for partially synthetic datasets are presented in Kinney and Reiter (2010). The derivations are based on the same ideas as those described in Section 5.1.2. Let $\bar{\mathbf{q}}_M$, $\bar{\mathbf{b}}_M$, $\mathbf{B}_M$, and $\bar{\mathbf{u}}_M$ be the multivariate analogs to $\bar{q}_M$, $\bar{b}_M$, B_M, and $\bar{u}_m$ defined in (8.1) to (8.4). Let us assume the user is interested in testing a null hypothesis of the form $\mathbf{Q} = \mathbf{Q}_0$ for a multivariate estimand with k components. Following the notation in Kinney and Reiter (2010), the Wald statistic for this test is given by

$$S_M = (\mathbf{Q_0} - \bar{\mathbf{q}}_M)^T \bar{\mathbf{u}}_M^{-1} (\mathbf{Q}_0 - \bar{\mathbf{q}}_M)/(k(1 + r_B - r_b)), \tag{8.8}$$

where $r_B = (1+1/m)tr(\mathbf{B}_M \bar{\mathbf{u}}_M^{-1})/k$ and $r_b = (1/r)tr(\bar{\mathbf{b}}_M \bar{\mathbf{u}}_M^{-1})/k$. The reference distribution for S_M is an F distribution, F_{k,ν_M}, with

$$\nu_M = 4 + \left(1 + \frac{r_B \nu_B}{\nu_B - 2} - \frac{r_b \nu_b}{\nu_b - 2}\right)^2 \Bigg/ \left(\frac{(r_B \nu_B)^2}{(\nu_B - 2)^2(\nu_B - 4)} + \frac{(r_b \nu_b)^2}{(\nu_b - 2)^2(\nu_b - 4)}\right), \tag{8.9}$$

for $\nu_B > 4$ and $\nu_b > 4$, and $\nu_B = k(m-1)$ and $\nu_b = km(r-1)$. When $m \leq 3$ and k is small, ν_M is not defined. Generally, generating a small number of imputed datasets is not recommended, especially since this would lead to a high probability that $T_M < 0$. Alternative degrees of freedom for $\nu_B \leq 4$ and $\nu_b \leq 4$ are provided in Kinney and Reiter (2010).

If $\mathbf{Q}$ contains a large number of components k, using $\bar{\mathbf{u}}_m$ can be cumbersome. As pointed out by Meng and Rubin (1992), it might be more convenient to use a likelihood ratio test in this case. Kinney and Reiter (2010) also present the derivations for this test for partially synthetic datasets when the original data are subject to nonresponse.

Again following the notation given in Kinney and Reiter (2010), let ψ be the vector of parameters in the analyst's model, and let $\psi^{(i,j)}$ be the maximum likelihood estimate of ψ computed from $D^{(i,j)}$, where $D^{(i,j)}$ is the jth synthetic dataset generated from the ith imputed dataset and $i = 1, ..., m$, $j = 1, ..., r$. The analyst is interested in testing the hypothesis that $\mathbf{Q}(\psi) = \mathbf{Q}_0$, where $\mathbf{Q}(\psi)$ is a k-dimensional function of ψ. Let $\psi_0^{(i,j)}$ be the maximum likelihood estimate of ψ obtained from $D^{(i,j)}$ subject to $\mathbf{Q}(\psi) = \mathbf{Q}_0$. The log-likelihood ratio test statistic associated with $D^{(i,j)}$ is $L^{(i,j)} = 2\log f(D^{(i,j)}|\psi_0^{(i,j)}) - 2\log f(D^{(i,j)}|\psi^{(i,j)})$. Let $\bar{L} = \sum_{i=1}^m \sum_{j=1}^r L^{(i,j)}/(mr)$, $\bar{\psi}^{(i)} = \sum_{j=1}^r \psi^{(i,j)}/r$, $\bar{\psi}_0^{(i)} = \sum_{j=1}^r \psi_0^{(i,j)}/r$, $\bar{\psi} = \sum_{i=1}^m \bar{\psi}^{(i)}/m$, and $\bar{\psi}_0 = \sum_{i=1}^m \bar{\psi}_0^{(i)}/m$. Furthermore, let $\bar{L}_0 = (1/(mr)) \sum_{i=1}^m \sum_{j=1}^r (2\log f(D^{(i)}|\bar{\psi}_0) - 2\log f(D^{(i)}|\bar{\psi}))$, the average of the log-likelihood ratio test statistics evaluated at ψ and ψ_0. Similarly, let $\bar{L}_M = (1/(mr)) \sum_{i=1}^m \sum_{j=1}^r (2\log f(D^{(i)}|\bar{\psi}_0^{(i)}) - 2\log f(D^{(i)}|\bar{\psi}^{(i)}))$, the average of the log-likelihood ratio test statistics evaluated at $\psi^{(i)}$ and $\psi_0^{(i)}$. The likelihood ratio test statistic is given by

$$\hat{S}_M = \bar{L}_0/(k(1+\hat{r}_B-\hat{r}_b)), \tag{8.10}$$

where $\hat{r}_B = (m+1)(\bar{L}_M - \bar{L}_0)/(k(m-1))$ and $\hat{r}_b = (\bar{L} - \bar{L}_M)/(k(r-1))$. The reference distribution for $\hat{S}_M$ is $F_{k,\hat{\nu}_M}$, where $\hat{\nu}_M$ is defined as for ν_M using $\hat{r}_B$ and $\hat{r}_b$ instead of r_B and r_b.

8.2 Analytical validity and disclosure risk

To evaluate the data utility in this setting, we can use the same measures as for fully synthetic or partially synthetic datasets, namely measuring the confidence interval overlap between confidence intervals obtained from the synthetic data and confidence intervals obtained from the original data or measuring how well one can discriminate between the original and the synthetic data based on the ideas of propensity score matching (see Section 6.2). The only difference from the standard partially synthetic data approach is that we compare the synthetic datasets with the datasets for which all missing values have been multiply imputed.

For disclosure risk evaluations, the disclosure risk measures described in Section 7.3 can be used for partially synthetic datasets. Depending on the scenario, measures that assume the intruder knows who participated in a survey (see Section 7.3.1) or measures that consider the additional uncertainty from sampling (see Section 7.3.2) can be applied. Useful disclosure risk measures for fully synthetic datasets still need to be developed (see Section 6.3).

8.3 Generating synthetic datasets from the multiply imputed IAB Establishment Panel

In the remainder of this chapter, I describe all the steps that were necessary to generate a scientific use file of the 2007 wave of the IAB Establishment Panel, which was released in January 2011. I start by briefly discussing how I selected the variables to be synthesized. I also describe the synthesis process and the models I implemented for the synthesis. Finally, I present results from the data utility and disclosure risk evaluations that I performed before the actual release. I refer to Section 5.3 for a discussion of the extensive imputation task required to impute all missing values in the dataset.

8.3.1 Selecting the variables to be synthesized

Once all missing values in the original data have been imputed (see Section 5.3), we can begin with the actual synthesis. The first and crucial step in the synthesis pro-

cess is to decide which variables need to be synthesized and whether it is necessary to synthesize all records in the dataset. In general, agencies can decide whether they only want to select key variables for synthesis or whether they also want to synthesize some of the sensitive variables. Key variables are those variables that could be used for reidentification purposes (i.e., variables for which the intruder knows the true values for some target records from external databases such as business or credit information databases). Sensitive variables are all those variables that contain information that a survey respondent would not be willing to provide to the general public.

In theory, there is often no need to synthesize sensitive variables that are not considered key variables. If all key variables are sufficiently protected, it will not be possible to link any record in the dataset to a specific respondent. Synthesizing sensitive variables is a conservative approach that might be justified since the amount of data available in external databases might increase over time and records that are considered safe now might be at risk later. It also helps convince survey respondents that their information is sufficiently protected.

For the IAB Establishment Panel, I decided to synthesize a combination of both variable types. Obviously, key variables such as *establishment size*, *region*, and *industry code* need to be protected since a combination of the three variables would enable the intruder to identify most of the larger establishments, but I also synthesized the most sensitive variables in the dataset such as *turnover* or *amount of subsidies received from the government*. Almost all numerical variables and some of the categorical variables are synthesized.

In many datasets, it is sufficient to alter only the subset of records that are actually at risk. These records can be found by cross-tabulating the key variables. Only those records in cross-tabulation cells with cell counts below an agency-defined threshold might need protection. The selective multiple imputation of keys (SMIKE) (Liu and Little, 2002) approach aims in that direction. In this application, it might have been sufficient to synthesize values only for the larger establishments since the sampling uncertainty and the similarities of the small establishments will make reidentification very difficult. Besides, arguably intruders will only be interested in identifying some larger establishments. However, I decided to synthesize all records since, given the large amount of information contained in the dataset (close to 300 variables), all records are sampling uniques arguably even population uniques. Of course, only a few variables in the dataset can be considered key variables, but once the dataset is released, a survey respondent might try to identify himself in the released dataset. Since the respondent knows all the answers he provided, it will be easy for him to find himself in the dataset. If he realizes that his record is included completely unchanged, he will feel that his privacy is at risk, even if an intruder who does not have the same background information will never be able to identify this respondent. To drive down this perceived risk, I decided to synthesize all 15,644 records in the dataset.

8.3.2 The synthesis task

For the synthesis, I use the sequential regression multivariate imputation approach (SRMI) (Raghunathan et al., 2001) with linear regression models for the continuous variables and logit models for the binary variables (see Section 3.2 for details on how to adjust these methods for skip patterns and logical constraints). Since I always replace all records with synthetic values for the variables at risk, the imputation task is comparable to imputation under a monotone missingness pattern, and thus I do not have to iterate between the imputations (see Section 3.1.2 for details).

But replacing all records with imputed values means that developing good models is essential. All variables that don't contain any structural missings are used as predictors in the imputation models in hopes of reducing problems from uncongeniality (Meng, 1994). For the synthesis, I use several imputation models for every variable whenever possible. Different models are defined for West and East Germany and for different establishment size classes defined by quantiles. Depending on the number of observations that could be used for the modeling, I define up to eight different regression models. I do not use the multinomial logit model for the synthesis of the polytomous variables since I already experienced problems with this approach when imputing the missing values in the dataset (see Section 5.3). For the synthesis, I do not want to limit the imputation models to some 30 explanatory variables. Furthermore, I also have to synthesize variables with a large number of categories such as region (16 categories) and industry code (41 categories). The multinomial model would hardly ever converge for these variables.

The standard approach for a model-based imputation of categorical variables with many categories is the multinomial/Dirichlet approach (see, for example, Abowd et al., 2006). The disadvantage of this approach is that covariates cannot be incorporated into the model directly. In general, a different model is fit for a large number of subcategories of the data, defined by cross-classifying some of the covariates to preserve the conditional distributions in the defined classes. This approach is impractical if the number of observations in a survey is low, because the number of observations will be too low to define suitable models in every subclass for which the marginal distribution should be preserved. For this reason, I follow a different strategy when synthesizing the categorical variables in the dataset. I generate synthetic values using CART models, as suggested by Reiter (2005d).

CART models are a flexible tool for estimating the conditional distribution of a univariate outcome given multivariate predictors. Essentially, the CART model partitions the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes. The partitions are found by recursive binary splits of the predictors. The series of splits can be effectively represented by a tree structure, with leaves corresponding to the subsets of units.

CART models can also be used to generate partially synthetic data (Reiter, 2005d). To illustrate the approach, let us assume that we only want to synthesize three categorical variables: region, industry code, and legal form. To generate synthetic datasets for these three variables, we proceed as follows. Using the original data, D_{obs}, we fit a tree of region on all other variables that don't contain any struc-

tural missings except industry code and legal form.[2] Label this tree $\mathscr{Y}_{(R)}$. We require a minimum of five records in each leaf of the tree and do not prune it; see Reiter (2005d) for a discussion of pruning and minimum leaf size. Let L_{Rw} be the wth leaf in $\mathscr{Y}_{(R)}$, and let $Y_{(R)}^{L_{Rw}}$ be the $n_{L_{Rw}}$ values of $Y_{(R)}$ in leaf L_{Rw}. In each L_{Rw} in the tree, we generate a new set of values by drawing from $Y_{(R)}^{L_{Rw}}$ using the Bayesian bootstrap (Rubin, 1981). These sampled values are the replacement imputations for the $n_{L_{Rw}}$ units that belong to L_{Rw}. Repeating the Bayesian bootstrap in each leaf of the region tree results in the ith set of synthetic regions, $Y_{(R)\text{rep},i}$.

Next, imputations are made for the industry code. Using D_{obs}, we fit the tree, $\mathscr{Y}_{(I)}$, with all variables except legal form as predictors. To maintain consistency with $Y_{(R)\text{rep},i}$, units' leaves in $\mathscr{Y}_{(I)}$ are located using $Y_{(R)\text{rep},i}$. Occasionally, some units may have combinations of values that do not belong to one of the leaves of $\mathscr{Y}_{(I)}$. For these units, we search up the tree until we find a node that contains the combination, then treat that node as if it were the unit's leaf. Once each unit's leaf is located, values of $Y_{(I)\text{rep},i}$ are generated using the Bayesian bootstrap. Imputing legal form follows the same process: we fit the tree $\mathscr{Y}_{(L)}$ using all variables that don't contain any structural missings as predictors, place each unit in the leaves of $\mathscr{Y}_{(L)}$ based on their synthesized values of region and industry code, and sample new legal forms using the Bayesian bootstrap.

I generate $r = 5$ datasets for every imputed dataset (i.e., $m * r = 25$ synthetic datasets will be released). Reiter (2008b) elaborates on the number of imputations at stages one and two when using multiple imputation for nonresponse and disclosure control simultaneously. He suggests setting $m > r$, especially if the fraction of missing information is large, to reduce variance from estimating missing values. But this approach will increase the risk of negative variance estimates since $\bar{b}_M$ will increase relative to B_M.

In the IAB Establishment Panel, only 12 variables (out of more than 300) have missing rates above 5%. On the other hand, I always synthesize 100% of the records. In his simulations, Reiter (2008b) does not find a significant reduction in variance with increasing m compared with r for 100% synthesis paired with low missing rates. On the other hand, the risk of negative variance estimates increases significantly. From these results, I conclude that it is preferable to set $m = r$ in this case.

8.3.3 Measuring the analytical validity

I evaluate the analytical validity of the generated datasets by comparing analytic results achieved with the original (fully imputed) data[3] with results from the synthetic

[2] To improve the data quality, I actually grow several trees for different subsets of the data. The subsets are defined by West and East Germany and by up to 25 different establishment size classes, defined by quantiles. To simplify the notation, I illustrate the approach assuming that only one tree is fit for each variable.

[3] For convenience, I will refer to the dataset with all missing values multiply imputed as the original data from here on.

Table 8.1 Regression results from a probit regression of *part time-employees (yes/no)* on 19 explanatory variables in West Germany. For the CI length ratio, the CI length of the original datasets is in the denominator.

	Original data	Synthetic data	CI overlap	z-score orig.	z-score syn.	CI length ratio
Intercept	−0.809	−0.752	0.87	−7.23	−6.85	0.99
5–10 employees	0.443	0.437	0.97	8.52	7.99	1.06
10–20 employees	0.658	0.636	0.90	11.03	10.88	0.98
20–50 employees	0.797	0.785	0.95	13.02	12.36	1.04
100–200 employees	0.892	0.908	0.96	9.23	9.48	0.99
200–500 employees	1.131	1.125	0.99	9.99	9.87	1.01
>500 employees	1.668	1.641	0.97	8.22	8.33	0.97
Growth in employment expected	0.010	0.006	0.98	0.18	0.12	0.99
Decrease in employment expected	0.087	0.100	0.96	1.11	1.27	1.00
Share of female workers	1.449	1.366	0.73	17.63	18.71	0.89
Share of employees with univ. degree	0.319	0.368	0.91	2.18	2.59	0.97
Share of low qualified workers	1.123	1.148	0.93	12.17	11.87	1.05
Share of temporary employees	−0.327	−0.138	0.75	−1.74	−0.71	1.05
Share of agency workers	−0.746	−0.856	0.88	−3.09	−4.24	0.84
Employed in last 6 months	0.394	0.369	0.87	8.33	7.82	1.00
Dismissal in the last 6 months	0.294	0.279	0.92	6.38	6.03	1.00
Foreign ownership	−0.113	−0.117	0.99	−1.33	−1.38	0.99
Good/very good profitability	0.029	0.033	0.98	0.72	0.82	0.99
Salary above coll. wage agreement	0.020	0.031	0.95	0.35	0.54	0.99
Collective wage agreement	0.016	0.007	0.95	0.31	0.13	0.97

data. The probit regression displayed in Tables 8.1 and 8.2 is adapted from a regression originally based on a different wave of the Establishment Panel. The dependent variable indicates whether an establishment employs part-time employees. The 19 explanatory variables include, among others, dummies for the establishment size, whether the establishment expects changes in the number of employees, and information on the personnel structure. Since there are still differences within Germany, the results are computed for West Germany (Table 8.1) and East Germany (Table 8.2) separately.

Both regressions clearly demonstrate the good data quality. All point estimates from the synthetic data are close to the point estimates from the original data, and the confidence interval overlap (see Section 6.2) is higher than 90% for most estimates, with an average of 90% for West Germany and 93% for East Germany. Some researchers are concerned that synthetic datasets will provide valid results for the significant variables but might provide less accurate results for variables with lower z-scores. The results indicate that this is not true at least for this analysis. The z-scores from the synthetic data are very close to the z-scores from the original data. This is an important result since model selections are often based on significance levels. The last column reports the 95% confidence interval length ratio with the confidence interval length of the original data in the denominator.

Table 8.2 Regression results from a probit regression of *part time-employees (yes/no)* on 19 explanatory variables in East Germany. For the CI length ratio, the CI length of the original datasets is in the denominator.

	Original data	Synthetic data	CI overlap	z-score orig.	z-score syn.	CI length ratio
Intercept	−0.712	−0.742	0.93	−6.42	−7.21	0.93
5–10 employees	0.266	0.257	0.96	4.81	4.53	1.03
10–20 employees	0.416	0.399	0.93	6.94	6.76	0.99
20–50 employees	0.542	0.532	0.96	9.18	8.72	1.04
100–200 employees	0.757	0.808	0.86	8.02	8.47	1.01
200–500 employees	0.971	1.013	0.91	8.25	8.57	1.00
>500 employees	1.401	1.422	0.98	5.69	5.66	1.02
Growth in employment expected	−0.041	−0.040	1.00	−0.73	−0.73	1.00
Decrease in employment expected	0.035	0.040	0.98	0.44	0.50	1.00
Share of female workers	1.006	1.041	0.88	12.63	14.93	0.88
Share of employees with univ. degree	0.221	0.197	0.95	1.86	1.76	0.95
Share of low qualified workers	0.976	1.042	0.87	8.44	7.84	1.19
Share of temporary employees	−0.049	0.049	0.84	−0.31	0.34	0.91
Share of agency workers	−0.176	−0.232	0.94	−0.73	−1.08	0.89
Employed in last 6 months	0.230	0.210	0.89	4.95	4.55	1.00
Dismissal in the last 6 months	0.301	0.295	0.97	6.43	6.35	0.99
Foreign ownership	−0.176	−0.176	1.00	−1.83	−1.84	1.00
Good/very good profitability	0.097	0.097	1.00	2.35	2.37	1.00
Salary above coll. wage agreement	0.080	0.086	0.98	1.04	1.10	1.01
Collective wage agreement	0.097	0.069	0.86	1.87	1.36	0.98

Since the multiple-imputation procedure for generating synthetic datasets correctly reflects the uncertainty in the imputation models, it can happen that the confidence intervals from the synthetic datasets are much wider and thus less efficient than the confidence intervals from the original data. For the variable *share of low qualified workers* in Table 8.2, the confidence interval length is increased by 19%. For all other estimands, the intervals are never increased more than 7%.

The second regression is an ordered probit regression with the expected employment trend in three categories (increase, no change, decrease) as the dependent variable. In this regression, I use 39 explanatory variables and the industry dummies as covariates. Again the analysis is computed for West Germany and East Germany separately. Figure 8.1 contains a plot of the original point estimates against the synthetic point estimates and a boxplot for the confidence interval overlap and the confidence interval length ratio. All graphs are based on the 78 estimates from the two regressions. Most of the point estimates in the first graph are close to the 45 degree line, indicating that the point estimates from the synthetic data are very close to the point estimates from the original data. But even if the point estimates differ, the data utility measured by the confidence interval overlap is high. The measure never drops below 61%, and the median overlap is 92.7%. Thus, even though some estimates are a little off the 45 degree line, the results are close to the orig-

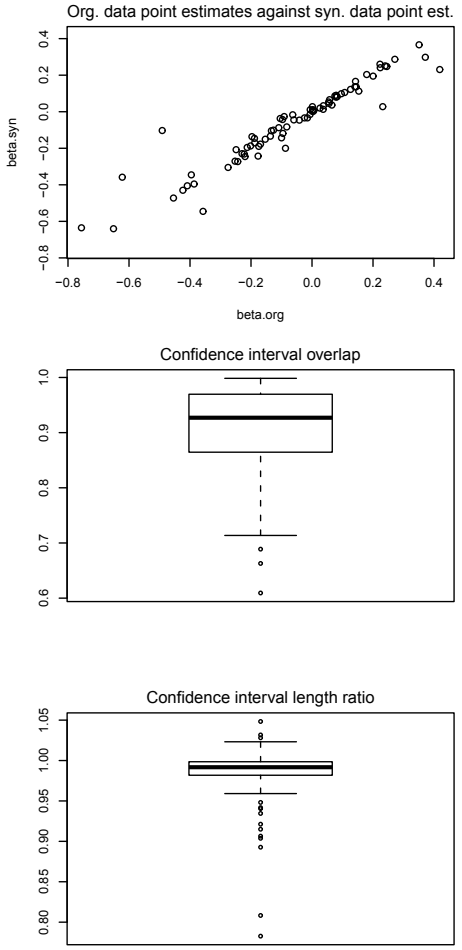

Fig. 8.1 Ordered probit regression of *expected employment trend* on 39 explanatory variables and industry dummies.

inal results since these coefficients are estimated with a high standard error. The boxplot of the confidence interval length ratio indicates that we do not lose much efficiency by using the synthetic data instead of the original data. The confidence interval never increases by more than 5% compared with the original data. Not all users of the data will be interested in multivariate regression analysis. For this reason, I also included an evaluation of the data utility for univariate statistics. For this, I compare the weighted overall mean and the weighted mean in different subgroups for all continuous variables in the dataset. The subgroups are defined by establishment size (ten categories, defined by quantiles), industry code (17 categories), and

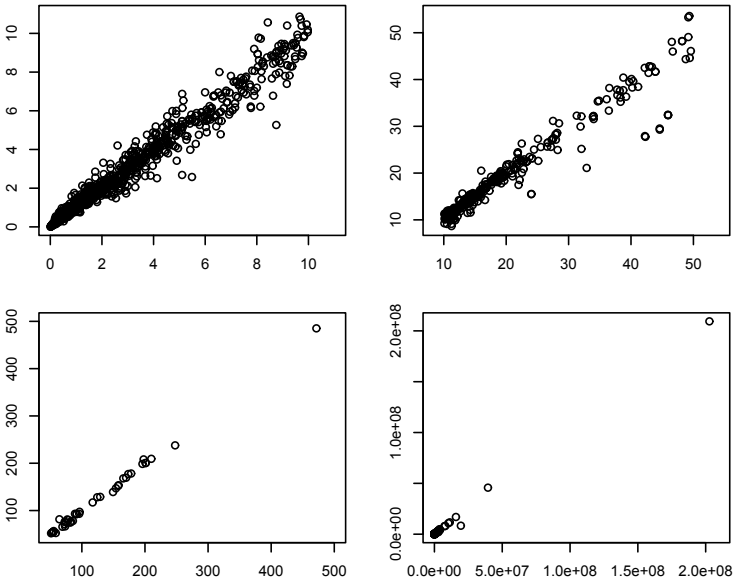

Fig. 8.2 Original point estimates against synthetic point estimates for the overall mean and the means in subgroups defined by establishment size class, industry code, and region.

region (16 categories). I do not investigate any cross-classifications since the cell sizes would be too small to obtain meaningful results. I also limit the evaluation to cells with at least 200 observations above zero for the same reason. This leads to a final number of 2,170 estimates. Figure 8.2 again presents the plots of the estimates from the original fully imputed datasets against the synthetic estimates. For readability, the plots are divided into four parts depending on the original value of the mean ($[0;10], (10;50], (50;500], (500;\infty)$). Most of the synthetic estimates are close to their original counterparts. Only a few estimates differ substantially from the original values. Figure 8.3 contains boxplots for the confidence interval overlap. The results for each stratifying variable and the overall mean are reported separately. The means within different establishment size classes and those across regions provide the best results, with median overlaps of 81.2% and 84.1%, respectively. For only 3.3% of the means within size classes and for less than 0.8% of the means within regions is there no overlap between the 95% confidence intervals from the original data and the 95% confidence intervals from the synthetic data. The results for the overall means and the means for different industry codes are good for most of the estimates, with median overlaps of 70.5% and 61.1%, respectively, but for some of the estimates (14.6% and 12.7%) the overlap is actually zero.

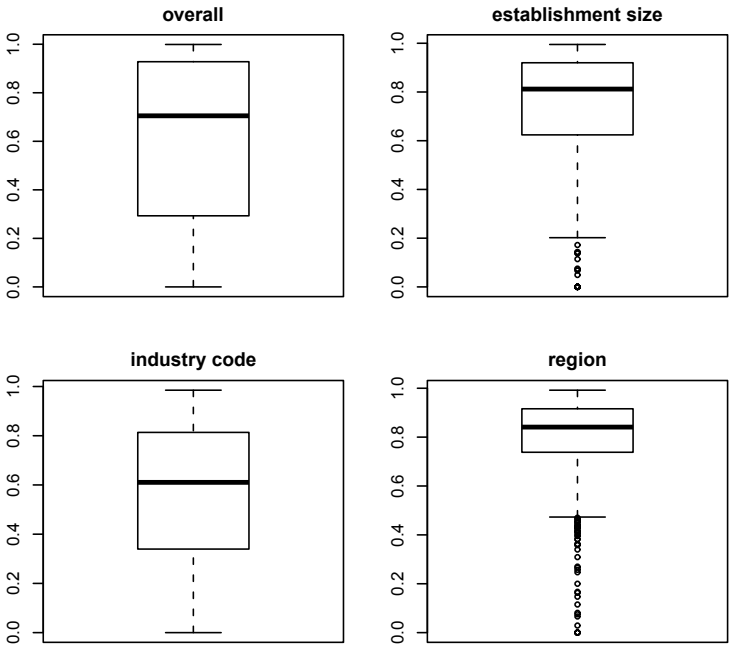

Fig. 8.3 Boxplots of CI overlaps for all continuous variables for the overall mean and the means in all subgroups defined by different stratifying variables.

8.3.4 Caveats in the use of synthetic datasets

Despite these mostly promising results, it would be overly optimistic to assume that synthetic datasets will provide results of similar quality for any potential analysis. It is crucial that the potential user of the data knows which analysis might provide valid results and for which analysis she might have to apply for direct access to the data at the research data center. To enable the user to make these decisions, it is very important that additional information about the imputation models be released in combination with the synthetic data. For example, the IAB will release information about which explanatory variables were used in the imputation models for each variable.

To give an example for which the synthetic data would not give valid results, I run a probit regression with the same explanatory variables as in Table 8.1, but I replace the dependent variable with an employment trend variable that equals 1 if the number of employees covered by social security increases between 2006 and 2007 and is 0 otherwise. I don't claim that this is a useful applied analysis; it only helps to illustrate that users should be careful when fitting models with dependent variables derived from two or more variables.

Table 8.3 Regression results from a probit regression of *employment trend (increase/no increase)* on 19 explanatory variables in West Germany. For the CI length ratio, the CI length of the original datasets is in the denominator.

	Original data	Synthetic data	CI overlap	z-score orig.	z-score syn.	CI length ratio
Intercept	−1.396	−0.978	0.05	−11.99	−9.28	0.92
5–10 employees	0.130	0.354	0.00	2.61	7.75	0.92
10–20 employees	0.316	0.495	0.05	6.19	11.19	0.87
20–50 employees	0.355	0.541	0.05	7.33	10.93	1.06
100–200 employees	0.366	0.351	0.94	5.69	6.09	0.91
200–500 employees	0.475	0.347	0.48	7.29	5.80	0.92
>500 employees	0.375	0.472	0.66	5.06	6.58	0.99
Growth in employment expected	0.374	0.148	0.00	9.29	3.59	1.05
Decrease in employment expected	−0.376	−0.020	0.00	−6.16	−0.38	0.86
Share of female workers	−0.140	−0.054	0.67	−2.09	−0.84	1.00
Share of employees with univ. degree	0.229	0.199	0.91	1.94	2.05	0.83
Share of low qualified workers	−0.043	−0.004	0.84	−0.68	−0.07	0.97
Share of temporary employees	0.434	0.226	0.62	3.25	1.60	1.07
Share of agency workers	0.058	0.013	0.69	0.94	0.08	2.61
Employed in last 6 months	0.948	0.368	0.00	24.94	11.60	0.84
Dismissal in the last 6 months	−0.172	−0.030	0.00	−4.42	−0.97	0.81
Foreign ownership	−0.165	−0.113	0.79	−2.60	−1.90	0.98
Good/very good profitability	0.248	0.100	0.00	7.69	3.35	0.93
Salary above coll. wage agreement	0.039	0.033	0.96	0.87	0.81	0.91
Collective wage agreement	0.003	0.063	0.62	0.06	1.72	0.85

Table 8.3 provides the results for this regression, and it is obvious that they are by no means close to the results given above in terms of data quality. Six of the 20 estimates actually have no confidence interval overlap at all, and the point estimates and z-scores often differ substantially from the original estimates. So the question arises, what is the reason for the poor performance of the synthetic datasets for this regression? To understand the problem, I first compare the original data and the synthetic data for the number of employees covered by social security 2006 and 2007. Figure 8.4 presents Q-Q plots of the original values against the synthetic values. The first two graphs present the plots for the two variables, and the last plot depicts the Q-Q plot for the difference in the number of employees between 2006 and 2007. The synthesis model did a very good job in capturing the distribution of the variables for 2006 and 2007; the quantiles are more or less identical. The distribution of the difference between the number of employees covered by social security in 2006 and 2007 is also well preserved. If I were to run a simple linear regression with the same covariates but with the difference in employment as the dependent variable, the average confidence interval overlap would be 75%, a significant improvement compared with 42% for the results in Table 8.3.

The actual problem stems from the fact that there is not much variation between the employment numbers for 2006 and 2007. In the original dataset, 5,376 of the

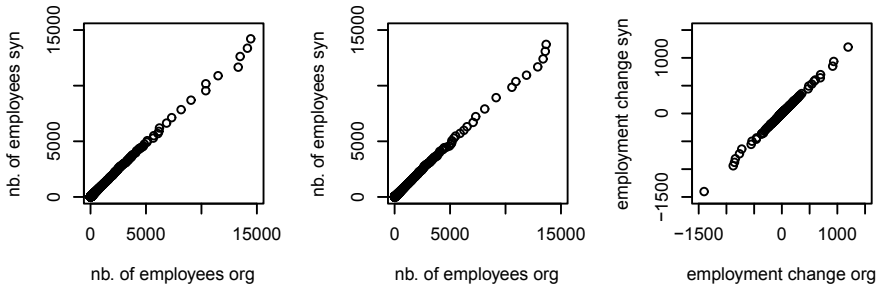

Fig. 8.4 Q-Q plots for the *number of employees covered by social security* in 2006 and 2007 and the employment trend between the two years.

15,644 establishments report no change in employment numbers, and more than 90% of the establishments report change rates of $\pm 5\%$. It can easily happen that in the original data an establishment reported an increase from 30 to 31 employees but in the synthetic data this establishment might have imputed values of 30 in both years or maybe 29 in the second year. Thus, the actual number is estimated very well and even the predicted difference is very close, but this record will change from an establishment with a positive employment trend to an establishment with no change or even a negative employment trend. The opposite is likely to occur as well: a record with a small negative employment trend might end up with a positive employment trend. If this happens for many records, which is to be expected since changes are very small for most records in the original dataset, the binary variable for employment trend will assign ones to a completely different subset of records. It is not surprising that results from the synthetic data will be different from the results in the original data in this case. It is important that users be made aware of this problem, which is likely to occur if the user derives his variable of interest from two or more variables in the dataset, and small changes in the underlying variables can have huge impacts on the derived variable. As a side note, this problem is not limited to multiply imputed synthetic datasets. In fact, most if not all standard perturbative SDC methods, such as swapping, adding noise, or micro aggregation will lead to similar problems.

8.3.5 Assessing the disclosure risk

It is unlikely that an intruder has detailed information about who participated in the survey; thus, using the actual data from the survey for the disclosure risk calculations is an unrealistic conservative scenario. For this reason, I apply the disclosure risk measures described in Section 7.3.2 that account for the additional uncertainty from sampling.

Table 8.4 Probabilities of being included in the target sample and in the original sample depending on establishment size.

Establishment size class	Probability(%)
1–4 employees	0.91
5–9 employees	1.62
10–19 employees	2.87
20–49 employees	4.10
50–99 employees	6.55
100–199 employees	11.39
200–499 employees	16.69
500–999 employees	20.48
1000–4,999 employees	31.89
>=5,000 employees	39.39

To obtain a set of target records for which the intruder has some knowledge from external databases that she uses to identify units in the survey, I sample new records from the sampling frame of the survey, the German Social Security Data (GSSD). I sample from this frame using the same sampling design as for the IAB Establishment Panel: stratification by establishment size, region, and industry code.

I find that 917 records from the target sample are also included in the original sample. Table 8.4 displays the percentage of records from the original dataset that are also included in the target sample for different establishment size classes. As expected, this probability increases with the establishment size. For establishments with less than 100 employees, the probability is always less than 10%, whereas large establishments with more than 5,000 employees are included in both samples with a probability close to 40%.

For the disclosure scenario I assume, the intruder has information on region, industry code (in 17 categories), and establishment size (measured by the number of employees covered by social security) for his target records and uses this information to identify units in the survey. I further assume that he would consider any record in the synthetic datasets a potential match for a specific target record if it fulfills two criteria: first, that the record's synthetic industry code and region exactly matches the target's true industry code and region; and second, that the record's synthetic number of employees lies within a defined interval around the target's number of employees. To define these intervals, I divide the number of employees by the ten stratification classes for establishment size and calculate the standard deviation within each size class. The interval is $t_e \pm \sqrt{sd_s}$, where t_e is the target's true value and sd_s is the standard deviation of the size class in which the true value falls. I investigated several other intervals (e.g., using the standard deviation directly or defining the intervals by 10-20 establishment size classes as I did in the example in Section 7.4.2 instead of using the stratification classes). However, I found that the criteria above led to the highest risk of disclosure.

8.3.5.1 Log-linear modeling to estimate the number of matches in the population

In general, the intruder will not know the number of records $F_{\mathbf{t}}$ that fulfill the matching criteria in the population to estimate the matching probabilities given in (7.9). One way to estimate the population counts from the released samples was suggested by Elamir and Skinner (2006). I apply this approach to the data assuming that the population counts follow an all-two-way-interactions log-linear model. To simplify the computation, I use the original sample to fit the log-linear model instead of fitting a log-linear model to each synthetic dataset separately. Arguably, this will increase the estimated risk, but the results should differ only slightly.

To fit the log-linear model, the three matching dimensions region, industry code, and establishment size are cross-tabulated in the sample. To obtain the correct number of establishment size matches, it is necessary to identify all records that fulfill the establishment size match criterion in the survey sample for each integer value of establishment size in the target sample. This leads to a $16 \times 17 \times 1102$-dimensional table to which I fit an all-two-way-interactions log-linear model. To calculate $\hat{F}_{\mathbf{t}}$, I need the sampling probabilities for each entry in this table. I obtain these probabilities by dividing the stratification matrix from the original sample by the stratification matrix from the GSSD. I assign the same probability to all establishment size values that fall into the same stratification cell. Again, an intruder will not know the exact sampling probabilities because he can only estimate the stratification matrix of the original sample from the synthetic samples, but arguably it is possible to obtain information about the number of establishments in Germany by region times industry times establishment size class. Since the stratification matrix from the synthetic samples will not differ very much from the matrix of the original sample, the estimated sampling probabilities might be reasonably close to the true sampling estimates. In any case, using the true sampling probabilities provides an upper bound for the disclosure risk.

Table 8.5 Average F_t and $\hat{F}_t$ for different establishment size classes.

Establishment size class	$mean(\hat{F}_{\mathbf{t}})$	$mean(F_{\mathbf{t}})$
1–4 employees	6467.66	6685.90
5–9 employees	1661.49	1737.89
10–19 employees	408.78	440.85
20–49 employees	161.98	179.01
50–99 employees	47.07	52.60
100–199 employees	17.91	22.89
200–499 employees	8.06	9.23
500–999 employees	2.17	2.88
1000–4,999 employees	1.51	2.03
>=5,000 employees	1.00	1.11

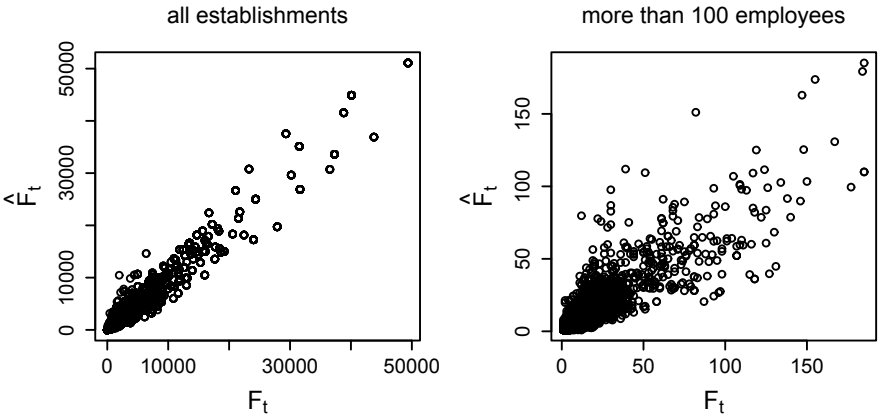

Fig. 8.5 Plots of F_t against $\hat{F}_t$ for all establishments and for establishments with more than 100 employees.

Since I can actually compute the true $F_{\mathbf{t}}$ from the GSSD, I am able to evaluate how well the true population counts can be estimated with the log-linear modeling approach. In Table 8.5 and Figure 8.5, I compare the estimated $\hat{F}_{\mathbf{t}}$ with the true F_t. In Table 8.5, I compute the average $\hat{F}_t$ and $F_{\mathbf{t}}$ for the target records in the ten establishment size stratification classes. The average estimated population count slightly underestimates the true counts but nevertheless is always very close to the average true population count. In Figure 8.5, I plot $\hat{F}_{\mathbf{t}}$ against F_t for each record in the target sample. The left graph presents the results for all establishments, and the right graph is limited to establishments with more than 100 employees. The figure shows that the log-linear modeling approach performs very well even at the record level.

8.3.5.2 Results from the disclosure risk evaluations

To estimate the actual risk of disclosure, I use the summary statistics presented in Section 7.3.1, accounting for the uncertainty from sampling as described in Section 7.3.2. These statistics are presented in Table 8.6. Notice that using $\hat{F}_{\mathbf{t}}$ instead of $F_{\mathbf{t}}$ gives almost similar results. In both cases, the disclosure risk is very low. Overall, only about 150 of the 15,624 records in the target sample are matched correctly, and the false match rate is 98.8%. I evaluated the disclosure risk in different establishment size classes and found that the percentage of true matches increases with the establishment size but never exceeds 7%. I also investigated whether the risks increase if the intruder only matches, when the average match probability exceeds a predefined threshold γ. Table 8.7 lists the false match rate and the number of true matches for different threshold values using $F_{\mathbf{t}}$ (there is almost no difference in the results if I use $\hat{F}_{\mathbf{t}}$ instead). The false match rates continually decrease to almost 80%

Table 8.6 Disclosure risk summaries for the synthetic Establishment Panel 2007 wave.

	$mean(\hat{F}_{\mathbf{t}})$	$mean(F_{\mathbf{t}})$
Expected match risk	162.34	160.92
True match risk	152	150
False match rate (%)	98.75	98.76

at $\gamma \leq 0.5$. Further reducing γ leads to no improvements in terms of the false match rate. Only for $\gamma \leq 0.1$ the rate drops to 66.7%. At the same time, the number of true matches continuously decreases until no true match is found at a threshold of $\gamma = 0$. Since the intruder never knows which matches actually are true matches, these results indicate that the data seem to be well protected at least under the given assumptions about the information an intruder can gather in her target data.

8.3.5.3 Disclosure risk for large establishments

Even though the results in the last section indicate a low risk of disclosure, large establishments might still be at risk because these establishments might be identifiable by matching on establishment size alone. Since a potential intruder will know that region and industry code have been synthesized, she might match only on establishment size for large establishments and ignore the fact that region and industry code are different between the target record and the match found in the synthetic data.

To quantify the risk from this approach, I evaluate two disclosure risk scenarios. In the first scenario, the intruder ranks all synthetic datasets by establishment size and considers the mode of the ranks for one unit across the synthetic datasets as the true rank of this unit. She then links that unit to the unit with the same rank in her target dataset. The second scenario assumes that the intruder performs a simple

Table 8.7 False match rate and true match risk for different levels of γ.

γ	False match rate	True match risk
1	98.76	150
0.9	94.42	97
0.8	91.47	59
0.7	88.72	38
0.6	84.57	27
0.5	81.91	17
0.4	84.62	8
0.3	82.14	5
0.2	85.71	2
0.1	66.67	1
0.0	-	0

nearest neighbor match between the records in her target data and the records in the synthetic samples using the establishment size variable.

Since the largest establishments are sampled with high probability, I treat the original sample as the target sample from which the intruder knows the true reported establishment size. This is still conservative since the reported establishment size might differ from the size reported in other databases, but it is not unlikely that the intruder well get reasonably close estimates of the true establishment size for large establishments in Germany.

Table 8.8 provides the results for the 25 largest establishments. The average match rate in column three is the percentage of times the declared match from the nearest neighbor matching approach actually is the true match across the 25 synthetic datasets. Obviously, the largest establishments face a very high risk of disclosure in both scenarios. The mode of the ranks in the synthetic datasets is almost always the same as the rank in the original sample, and the nearest neighbor matching approach will lead to correct matches for most of the datasets. If the intruder were also to pick the mode of declared matches as the correct match, she would be

Table 8.8 Mode of the establishment size rank and average match rate for large establishments.

Original rank	Mode of synthetic ranks	Average match rate
1	1	0.96
2	2	0.72
3	3	1.00
4	4	1.00
5	5	1.00
6	6	0.88
7	7	0.64
8	8	0.56
9	9	0.44
10	10	0.32
11	11	0.84
12	12	0.56
13	13	0.56
14	14	0.68
15	15	0.76
16	17	0.56
17	18	0.48
18	16	0.00
19	19	0.56
20	20	0.04
21	22	0.44
22	23	0.72
23	21	0.00
24	24	0.40
25	25	0.28

right for 21 of the 25 establishments. Clearly, there is a need to further protect the largest establishments in the dataset.

8.3.5.4 Additional protection for the largest establishments in the survey

A simple strategy to better protect large establishments would be to reduce the quality of the imputation model for establishment size, for example by dropping explanatory variables from the imputation model until a predefined criterion of variability between the imputations is met. However, since it would be necessary to drop the variables with the highest explanatory power to considerably increase the variability, important relationships between the variables would not be reflected in the released data, leading to uncongeniality problems if the analyst's model differs from the imputation model. It is also not an option to use other SDL techniques since methods such as noise addition would have to be applied at a very high level and other methods like data swapping and microaggregation, are well known to have severe negative consequences for data quality in the upper tail of the distribution. I therefore decided to inflate the variance of the beta coefficients in the imputation model instead. Remember that the imputation process always consists of two steps. In the first step, new parameters for the imputation model are drawn from their posterior distributions given the observed data. In the second step, new values for the variable to be imputed are drawn from the posterior predictive distribution given the parameters drawn in step one. For the standard linear model, this means that step one consists of drawing new values of σ^2 and β from their posterior distributions. I decided to protect records at risk by inflating the variance of β in the underlying imputation models. I inflate the variance by drawing new values of β from

$$\beta|\sigma^2 \sim N(\hat{\beta}, \alpha\sigma^2(X'X)^{-1}), \tag{8.11}$$

where α is the variance inflation factor, $\hat{\beta}$ and X are the regression coefficients and the explanatory variables from the underlying imputation model, respectively, and σ^2 is the new value of the variance drawn from its posterior distribution. Of course, imputation under this variance-inflated model is not proper in Rubin's sense (see Rubin, 1987, pp. 118–119), so I conducted a small simulation study to evaluate the impact of different levels of α on the validity of the results from a frequentist perspective. In the simulation, reported in the Appendix C, I found almost no impact on coverage rates. Even when synthesizing all records with α=1,000, the coverage rate for the 95% confidence interval never dropped below 90% and was close to the nominal 95% for most of the estimates of interest. The most notable consequence is that we lose efficiency since the between-imputation variance increases linearly with α. But since only some records at risk need to be replaced, I am not concerned that this will have huge impacts on data utility. The utility evaluations in Section 8.3.3 that were actually performed on the final dataset after applying the additional protection step described here seem to support this.

To apply the variance inflation approach, it is necessary to define which records are considered to be at risk. I define a record to be at risk if one of the following two conditions is fulfilled:

1. The standard deviation of the establishment size rank across the synthetic datasets for the record is less than 2.
2. The mode of the declared matches in the nearest neighbor matching scenario is the correct match.

The threshold value for the standard deviation of the ranks is chosen somewhat arbitrarily. Defining justifiable threshold rules is an area for future research.

To keep the negative impacts of this procedure at a minimum, I developed an iterative replacement algorithm. For a given level of α, all records that fulfill one of the criteria above are replaced by new draws from the variance-inflated imputation model. Records that still are at risk after ten rounds of repeatedly drawing from this model are replaced by draws from a model with the next higher level of α. In this application, I set the levels arbitrarily to α=(10;100;1,000). Developing methods to derive useful levels of α is an area for future research. Overall, I replace 79 records in the dataset by this procedure. Less than ten are replaced by draws from imputation models with $\alpha \geq 100$. Evaluating the disclosure risk for large establishments again, I find that the mode of the establishment size rank in the synthetic datasets is equal to the rank in the original data for only 12 of the 100 largest establishments. Since the intruder never knows if her match is correct and it is also unlikely that the intruder will know the original rank beyond the 20 largest establishments in the survey, the data are well protected from these kinds of attacks. For the nearest neighbor matching scenario, I guaranteed that the mode of the declared matches is never the correct match. I also find that no record is identified correctly in more than five of the 25 datasets. These results together with the results in Section 8.3.5 and the promising results on data utility in Section 8.3.3 demonstrate that the dataset is ready for release.

Chapter 9
A Two-Stage Imputation Procedure to Balance the Risk–Utility Trade-Off[1]

There has been little discussion in the literature on how many multiply imputed datasets an agency should release. From the perspective of the secondary data analyst, a large number of datasets is desirable. The additional variance introduced by the imputation decreases with the number of released datasets. For example, Reiter (2003) finds nearly a 100% increase in the variance of regression coefficients when going from 50 to two partially synthetic datasets. From the perspective of the agency, a small number of datasets is desirable. The information available to ill-intentioned users seeking to identify individuals in the released datasets increases with the number of released datasets. Thus, agencies considering the release of partially synthetic data generally are confronted with a trade-off between disclosure risk and data utility.

The empirical investigations presented in Section 9.3 indicate that increasing m results in both higher data utility and higher risk of disclosures. In this chapter, I present an alternative synthesis approach that can maintain high utility while reducing disclosure risks. The basic idea behind this approach is to impute variables that drive the disclosure risk only a few times and other variables many times. This can be accomplished by generating data in two stages, as described by Reiter and Drechsler (2010). In general, two-stage and one-stage approaches require similar amounts of modeling effort; however, in some settings, the two-stage approach can reduce computational burdens associated with generating synthetic data and thereby speed up the process; see Reiter and Drechsler (2010) for further discussion of this point. The two-stage imputation procedure is applicable to both, partially and fully synthetic datasets. In the following sections, I present the combining rules for univariate estimands for both approaches and provide an application of the two-stage partially synthetic approach to illustrate the potential benefits of this procedure. Deriving the combining rules for multivariate estimands is an area for future research.

[1] Most of this chapter is taken from Drechsler and Reiter (2009) and Reiter and Drechsler (2010).

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_9,

9.1 Inference for synthetic datasets generated in two stages

For a finite population of size N, let $I_l = 1$ if unit l is included in the survey and $I_l = 0$ otherwise, where $l = 1,\ldots,N$. Let $I = (I_1,\ldots,I_N)$, and let the sample size $s = \sum I_l$. Let X be the $N \times d$ matrix of sampling design variables (e.g., stratum or cluster indicators or size measures). I assume that X is known approximately for the entire population; for example, from census records or the sampling frame(s). Let Y be the $N \times p$ matrix of survey data for the population. Let Y_{inc} be the $s \times p$ submatrix of Y for all units with $I_l = 1$. I assume that there are no missing data. The observed data are thus $D_{obs} = (X, Y_{inc}, I)$. Methods for handling missing data and one stage of partial synthesis simultaneously are presented in Chapter 8. Developing two-stage imputation methods for data that are subject to nonresponse is an area for future research.

9.1.1 Fully synthetic data

Let Y_a be the values simulated in stage one, and let Y_b be the values simulated in stage two. The agency seeks to release fewer replications of Y_a than of Y_b, yet do so in a way that enables the analyst of the data to obtain valid inferences with standard complete-data methods. To do so, the agency generates synthetic datasets in a three-step process. First, the agency fills in the unobserved values of Y_a by drawing values from $f(Y_a \mid D_{obs})$, creating a partially completed population. This is repeated independently m times to obtain $Y_a^{(i)}$ for $i = 1,\ldots,m$. Second, in each partially completed population defined by nest i, the agency generates the unobserved values of Y_b by drawing from $f(Y_b \mid D_{obs}, Y_a^{(i)})$, thus completing the rest of the population values. This is repeated independently r times for each nest to obtain $Y_b^{(i,j)}$ for $i = 1,\ldots,m$ and $j = 1,\ldots,r$. The result is $M = mr$ completed populations, $P^{(i,j)} = (D_{obs}, Y_a^{(i)}, Y_b^{(i,j)})$, where $i = 1,\ldots,m$ and $j = 1,\ldots,r$. Third, the agency takes a simple random sample of size n_{syn} from each completed population $P^{(i,j)}$ to obtain $D^{(i,j)}$. These M samples, $D_{syn} = \{D^{(i,j)} : i = 1,\ldots,m; j = 1,\ldots,r\}$, are released to the public. Each released $D^{(i,j)}$ includes a label indicating its value of i (i.e., an indicator for its nest).

The agency can sample from each $P^{(i,j)}$ using designs other than simple random samples, such as the stratified sampling in the IAB Establishment Panel synthesis. A complex design can improve efficiency and ensure adequate representation of important subpopulations for analyses. When synthetic data are generated using complex samples, analysts should account for the design in inferences, for example with survey-weighted estimates. One advantage of simple random samples is that analysts can make inferences with techniques appropriate for simple random samples.

The agency could simulate Y for all N units, thereby avoiding the release of actual values of Y. In practice, it is not necessary to generate completed-data populations

for constructing the $D^{(i,j)}$; the agency only needs to generate values of Y for units in the synthetic samples. The formulation of completing the population and then sampling from it aids in deriving inferential methods (see Reiter and Drechsler, 2010).

Let Q be the estimand of interest, such as a population mean or a regression coefficient. For all (i, j), let $q^{(i,j)}$ be the estimate of Q, and let $u^{(i,j)}$ be the estimate of the variance associated with $q^{(i,j)}$. The $q^{(i,j)}$ and $u^{(i,j)}$ are computed based on the design used to sample from $P^{(i,j)}$. Note that when $n_{syn} = N$, the $u^{(i,j)} = 0$. The following quantities are necessary for inferences:

$$\bar{q}_r^{(i)} = \sum_{j=1}^{r} q^{(i,j)}/r, \tag{9.1}$$

$$\bar{q}_M = \sum_{i=1}^{m} \bar{q}_r^{(i)}/m = \sum_{j=1}^{r}\sum_{i=1}^{m} q^{(i,j)}/mr, \tag{9.2}$$

$$b_M = \sum_{i=1}^{m} (\bar{q}_r^{(i)} - \bar{q}_M)^2/(m-1), \tag{9.3}$$

$$w_r^{(i)} = \sum_{j=1}^{r} (q^{(i,j)} - \bar{q}_r^{(i)})^2/(r-1), \tag{9.4}$$

$$\bar{u}_M = \sum_{j=1}^{r}\sum_{i=1}^{m} u^{(i,j)}/mr. \tag{9.5}$$

The analyst then can use $\bar{q}_M$ to estimate Q and

$$T_{2st,f} = (1+m^{-1})b_M + (1-1/r)\bar{w}_M - \bar{u}_M \tag{9.6}$$

to estimate the variance of $\bar{q}_M$, where $\bar{w}_M = \sum_{i=1}^{m} w_r^{(i)}/m$. Inferences can be based on a t distribution with degrees of freedom

$$\nu_{2st,f} = \left(\frac{((1+1/m)b_M)^2}{(m-1)T_{2st,f}^2} + \frac{((1-1/r)\bar{w}_M)^2}{(m(r-1))T_{2st,f}^2} \right)^{-1}.$$

Derivations of these methods are presented in Reiter and Drechsler (2010). It is possible that $T_{2st,f} < 0$, particularly for small values of m and r. Generally, negative values of $T_{2st,f}$ can be avoided by making n_{syn} or m and r large. To adjust for negative variances, one approach is to use the always positive variance estimator $T_{2st,f}^* = T_{2st,f} + \lambda\bar{u}_M$, where $\lambda = 1$ when $T_{2st,f} \leq 0$ and $\lambda = 0$ when $T_{2st,f} > 0$. When $T_{2st,f} < 0$, using $\nu_{2st,f}$ is overly conservative since $T_{2st,f}^*$ tends to be conservative when $\lambda = 1$. To avoid excessively wide intervals, analysts can base inferences on t distributions with degrees of freedom $\nu_{2st,f}^* = \nu_{2st,f} + \lambda\infty$.

9.1.2 Partially synthetic data

The agency generates the partially synthetic data in two stages. Let $Y_a^{(i)}$ be the values imputed in the first stage in nest i, where $i = 1, \ldots, m$. Let $Y_b^{(i,j)}$ be the values imputed in the second stage in dataset j in nest i, where $j = 1, \ldots, r$. Let Y_{nrep} be the values of Y_{inc} that are not replaced with synthetic data and hence are released as is. Let $Z_{a,l} = 1$ if unit l, for $l = 1, \ldots, s$, is selected to have any of its first-stage data replaced, and let $Z_{a,l} = 0$ otherwise. Let $Z_{b,l}$ be defined similarly for the second-stage values. Let $Z = (Z_{a,1}, \ldots, Z_{a,s}, Z_{b,1}, \ldots, Z_{b,s})$.

To create $Y_a^{(i)}$ for those records with $Z_{a,l} = 1$, first the agency draws from $f(Y_a \mid D_{obs}, Z)$, conditioning only on values not in Y_b. Second, in each nest, the agency generates $Y_b^{(i,j)}$ for those records with $Z_{b,l} = 1$ by drawing from $f(Y_b^{(i,j)} \mid D_{obs}, Z, Y_a^{(i)})$. Each synthetic dataset $D^{(i,j)} = (X, Y_a^{(i)}, Y_b^{(i,j)}, Y_{nrep}, I, Z)$. The entire collection of $M = mr$ datasets, $D_{syn} = \{D^{(i,j)}, i = 1, \ldots, m; j = 1, \ldots, r\}$, with labels indicating the nests, is released to the public.

To obtain inferences from nested partially synthetic data, I assume the analyst acts as if each $D^{(i,j)}$ is a sample according to the original design. Unlike in fully synthetic data, there is no intermediate step of completing populations. The analyst again can use $\bar{q}_M$ to estimate Q and

$$T_{2st,p} = \bar{u}_M + b_M/m \tag{9.7}$$

to estimate the variance of $\bar{q}_M$. Inferences can be based on a t distribution with degrees of freedom $\nu_{2st,p} = (m-1)(1 + m\bar{u}_M/b_M)^2$. Derivations of these methods are presented in Reiter and Drechsler (2010). Note that $T_{2st,p} > 0$ always holds, so that negative variance estimates do not arise in two-stage partial synthesis.

9.2 Analytical validity and disclosure risk

To evaluate the analytical validity and disclosure risk, the same methods as with standard one-stage synthesis can be applied. I refer to Section 6.2 for possible data utility measures and to Section 6.3 and Section 7.3 for possible disclosure risk evaluations.

9.3 Application of the two-stage approach to the IAB Establishment Panel

To assess the impact of different numbers of imputations, I first evaluate the trade-off between risk and utility as a function of m for standard one-stage imputation. I then

compare the results with results achievable with the proposed two-stage imputation approach.

For this simulation study, I synthesize two variables in the IAB Establishment Panel for 1997: the number of employees and the industry coded in 16 categories. For both variables, all 7,332 observations are replaced by imputed values. Employment size and industry code are high-risk variables since (i) they are easily available in other databases and (ii) the distribution for the number of employees is heavily skewed. Imputations are based on linear models with more than 100 explanatory variables for the number of employees and on a multinomial logit model with more than 80 explanatory variables for the industry. Some variables for the multinomial logit model are dropped for multicollinearity reasons.

9.3.1 Analytical validity for the panel from one-stage synthesis

I investigate data utility for some descriptive statistics and a probit regression. The descriptive statistics are the (unweighted) average number of employees by industry; they are based solely on the two variables I synthesized. The probit regression, which originally appeared in an article by Zwick (2005), is used in various places throughout the book; see Section 6.4.2 for a detailed description.

Tables 9.1–9.4 display point estimates and the interval overlap measures for different values of m. For most parameters, increasing m moves point estimates closer to their values in the original data and increases the overlaps in the confidence intervals. Increasing $m = 3$ to $m = 10$ results in the largest increase in data utility, as

Table 9.1 Average number of employees by industry for one-stage synthesis.

	Original data	m=3	m=10	m=50	m=100
Industry 1	71.5	84.2	84.2	82.6	82.4
Industry 2	839.1	919.4	851.2	870.2	852.9
Industry 3	681.1	557.7	574.5	594.4	593.1
Industry 4	642.9	639.9	644.8	643.5	649.6
Industry 5	174.5	179.8	176.0	183.5	187.4
Industry 6	108.9	132.4	121.8	120.8	120.7
Industry 7	117.1	111.6	112.9	117.1	119.6
Industry 8	548.7	455.3	504.3	514.2	513.0
Industry 9	700.7	676.9	689.4	711.8	713.4
Industry 10	547.0	402.4	490.3	499.3	487.7
Industry 11	118.6	142.7	130.2	132.1	131.0
Industry 12	424.3	405.6	414.9	424.5	425.2
Industry 13	516.7	526.1	549.1	550.2	551.9
Industry 14	128.1	185.8	167.1	160.0	159.0
Industry 15	162.0	292.8	233.4	221.9	238.1
Industry 16	510.8	452.8	449.9	441.5	439.3

Table 9.2 Results from the vocational training regression for one-stage partial synthesis revisited.

	Original data	m=3	m=10	m=50	m=100
Intercept	−1.319	−1.323	−1.322	−1.323	−1.324
Redundancies expected	0.253	0.268	0.262	0.264	0.264
Many employees expected on maternity leave	0.262	0.334	0.316	0.312	0.314
High qualification need expected	0.646	0.636	0.640	0.640	0.639
Appren. training reaction on skill shortage	0.113	0.098	0.106	0.110	0.112
Training reaction on skill shortage	0.540	0.529	0.538	0.542	0.543
Establishment size 20–199	0.684	0.718	0.709	0.705	0.701
Establishment size 200–499	1.352	1.363	1.333	1.339	1.343
Establishment size 500–999	1.346	1.315	1.386	1.377	1.367
Establishment size 1,000 +	1.955	1.782	1.800	1.778	1.776
Share of qualified employees	0.787	0.787	0.788	0.784	0.785
State-of-the-art technical equipment	0.171	0.183	0.178	0.174	0.174
Collective wage agreement	0.255	0.268	0.264	0.267	0.268
Apprenticeship training	0.490	0.501	0.510	0.507	0.507
East Germany	0.058	0.038	0.033	0.042	0.044

the average confidence interval overlap over all 31 parameters in Table 9.3 and Table 9.4 increases from 0.828 to 0.867. Increasing $m = 50$ to $m = 100$ does not have much impact on data utility.

Table 9.3 Confidence interval overlap for the average number of employees for one-stage synthesis.

	m=3	m=10	m=50	m=100
Industry 1	0.778	0.770	0.777	0.782
Industry 2	0.844	0.893	0.853	0.874
Industry 3	0.730	0.776	0.797	0.800
Industry 4	0.983	0.992	0.995	0.971
Industry 5	0.920	0.935	0.863	0.817
Industry 6	0.605	0.749	0.764	0.767
Industry 7	0.809	0.820	0.863	0.876
Industry 8	0.692	0.862	0.894	0.890
Industry 9	0.926	0.966	0.968	0.963
Industry 10	0.660	0.876	0.897	0.871
Industry 11	0.609	0.804	0.773	0.792
Industry 12	0.903	0.912	0.916	0.918
Industry 13	0.946	0.814	0.809	0.799
Industry 14	0.408	0.589	0.655	0.664
Industry 15	0.586	0.639	0.654	0.638
Industry 16	0.666	0.645	0.583	0.566
Average	0.754	0.815	0.816	0.812

Each entry in Table 9.1–9.4 results from one replication of a partially synthetic data-release strategy. To evaluate the variability across different replications, I repeated each simulation ten times. Table 9.5 presents the average confidence interval overlap over all 31 estimands for the ten simulations. The variation in the overlap measures decreases with m. This is because the variability in $\bar{q}_m$ and T_m decreases with m, so that results stabilize as m gets large. I believe most analysts would prefer to have stable results across different realizations of the synthesis and hence favor large values of m.

9.3.2 Disclosure risk for the panel from one-stage synthesis

To assess the disclosure risk, I assume that the intruder knows which establishments are included in the survey and the true values for the number of employees and industry (i.e., I assume the intruder scenario described in Section 7.3.1). This is a conservative scenario but gives, in some sense, an upper bound on the risk for this level of intruder knowledge. Intruders might also know other variables in the file, in which case the agency may need to synthesize them as well.

The intruder computes probabilities using the approach outlined in Section 7.3.1. I assume that the agency does not reveal the synthesis model to the public, so that the only information in M is that employee size and industry were synthesized. For a given target $\mathbf{t}$, records from each $D^{(i)}$ must meet two criteria to be possible matches. First, the record's synthetic industry code exactly matches the target's true

Table 9.4 Confidence interval overlap for the vocational training regression for one-stage synthesis.

	m=3	m=10	m=50	m=100
Intercept	0.987	0.989	0.986	0.984
Redundancies expected	0.931	0.958	0.946	0.948
Many emp. exp. on maternity leave	0.808	0.856	0.867	0.861
High qualification need exp.	0.965	0.977	0.978	0.976
Appren. train. react. on skill shortages	0.928	0.964	0.984	0.996
Training react. on skill shortages	0.946	0.989	0.989	0.982
Establishment size 20–199	0.802	0.856	0.879	0.902
Establishment size 200–499	0.934	0.939	0.935	0.933
Establishment size 500–999	0.926	0.907	0.928	0.953
Establishment size 1,000 +	0.731	0.763	0.727	0.723
Share of qualified employees	0.995	0.997	0.989	0.993
State-of-the-art tech. equipment	0.919	0.953	0.976	0.977
Collective wage agreement	0.926	0.952	0.934	0.927
Apprenticeship training	0.937	0.883	0.899	0.899
East Germany	0.872	0.840	0.899	0.912
Average	0.907	0.922	0.928	0.931

Table 9.5 Average confidence interval overlap for all 31 estimands for ten independent simulations of one-stage synthesis.

	m=3	m=10	m=50	m=100
Simulation 1	0.828	0.867	0.870	0.870
Simulation 2	0.864	0.869	0.869	0.874
Simulation 3	0.858	0.866	0.873	0.868
Simulation 4	0.881	0.861	0.874	0.871
Simulation 5	0.872	0.865	0.866	0.875
Simulation 6	0.845	0.862	0.869	0.865
Simulation 7	0.849	0.851	0.871	0.873
Simulation 8	0.841	0.862	0.871	0.873
Simulation 9	0.841	0.877	0.873	0.872
Simulation 10	0.861	0.865	0.874	0.867
Average	0.854	0.865	0.871	0.871

industry code. Second, the record's synthetic number of employees lies within an agency-defined interval around the target's true number of employees. Acting as the agency, I define the interval as follows. I divide the cubic root of the true number of employees into 20 quantiles and calculate the standard deviation of the number of employees within each quantile. The interval is $t_e \pm sd_s$, where t_e is the target's true value and sd_s is the standard deviation of the quantile in which the true value falls. When there are no synthetic records that fulfill both matching criteria, the intruder matches only on the industry code.

I use 20 quantiles because this is the largest number of groups that guarantees some variation within each group. Using more than 20 quantiles results in groups with only one value of employment, which forces exact matching for targets in those quantiles. On the other hand, using a small number of quantiles does not differentiate adequately between small and large establishments. For small establishments, I want the potential matches to deviate only slightly from the original values. For large establishments, I accept higher deviations.

I studied the impact of using different numbers of groups for $m = 50$. I found a substantial increase in the risks of identification, especially for the small establishments, when going from exact matching to five quantiles. Between five and 20 quantiles, the disclosure risk doesn't change dramatically. For more than 20 quantiles, the number of identifications starts to decline again.

Table 9.6 displays the average true matching risk and expected matching risk over the ten simulation runs used in Table 9.5. Since the largest establishments are usually considered as the records most at risk of identification, I also include the risk measures for the largest largest establishments in parentheses. There is clear evidence that a higher number of imputations leads to a higher risk of disclosure, especially for the largest establishments. This is because, as m increases, the intruder has more information to estimate the distribution that generated the synthetic data.

Table 9.6 Averages of disclosure risk measures over ten simulations of one-stage synthesis. Measures for the 25 largest establishments are reported in parentheses.

	m=3	m=10	m=50	m=100
Expected match risk	67.8 (3.2)	94.8 (5.2)	126.9 (6.9)	142.5 (7.1)
True match risk	35.2 (2.0)	82.5 (4.9)	126.1 (6.8)	142.4 (7.1)

It is arguable that the gains in utility, at least for these estimands, are not worth the increases in disclosure risks.

The establishments that are correctly identified vary across the ten replicates. For example, for $m = 50$, the total number of identified records over all ten replicates is 614. Of these records, 319 are identified in only one simulation, 45 are identified in more than five simulations, and only ten records are identified in all ten replications. For $m = 10$, no records are identified more than seven times.

The risks are not large on an absolute scale. For example, with $m = 10$, I anticipate that the intruder could identify only 83 establishments out of 7,332. This assumes that the intruder already knows the establishment size and industry classification code and also has response knowledge; i.e., he knows which establishments participated in the survey. Furthermore, the intruder will not know how many of the unique matches (i.e., $c_j = 1$) actually are true matches.

I also investigated the disclosure risk for different subdomains for $m = 50$. Four of the 16 industry categories had less than 200 units in the survey. For these categories, the percentage of identified records ranged between 5% and almost 10%. For the remaining categories, the percentage of correct identifications never went beyond 2.3%. If these risks are too high, the agency could collapse some of the industry categories.

The percentage of identified establishments was close to 5% for the largest decile of establishment size and never went beyond 2.5% for all the other deciles. The identification risk is higher for the top 25 establishments but still moderate. When $m = 3$, only two of these establishments are correctly identified; this increases to seven establishments with $m = 100$. The intruder also makes many errors when declaring matches for these establishments. In fact, the false match rate for these top establishments is 87% for $m = 3$, 77% for $m = 10$, and approximately 70% for $m = 50$ and $m = 100$. None of the top ten establishments are identified in all ten simulations.

The largest establishment's size is reduced by at least 10% in all synthetic datasets. This can be viewed as reduction in data utility since the tail is not accurate at extreme values. It may be possible to improve tail behavior with more tailored synthesis models, such as CART approaches (Reiter, 2005d; see also Section 8.3.2).

As noted previously, these risk computations are in some ways conservative. First, they presume that the intruder knows which records are in the survey. This is not likely to be true since most establishments are sampled with probability less than one. However, large establishments are sampled with certainty, so that the risk

calculations presented here apply for those records. Second, the risk measurements presume that the intruder has precise information on establishment size and industry code. In Germany, it is not likely that intruders will know the sizes of all establishments in the survey, because there is no public information on small establishments. However, intruders can obtain size and industry type for large companies from public databases. They also can purchase large private databases on establishments, although the quality of these databases for record linkage on employee size is uncertain. Thus, except possibly for the largest establishments, the risk measures here could overstate the probabilities of identification.

9.3.3 Results for the two-stage imputation approach

For the two-stage imputation, I impute the industry in stage one and the number of employees in stage two. Exchanging the order of the imputation does not materially impact the results. I consider different values of m and r. I run ten simulations for each setting and present the average estimates over these ten simulations.

Table 9.7 displays the average confidence interval overlap for all 31 parameters and the two disclosure risk measures for the different settings. As with one-stage synthesis, there is not much difference in the data utility measures for different M, although there is a slight increase when going from $M = 9$ to $M \approx 50$. The two-stage results with $M = 9$ (average overlap of .867) are slightly better than the one-stage results with $m = 10$ (average overlap of .865). The two-stage results with $M \approx 50$ are approximately on the same level or slightly above the one-stage results for $m = 50$ (average overlap of .871).

The improvements in data utility when using the two-stage approach are arguably minor, but the reduction in disclosure risks is more noticeable. The measures are always substantially lower for the two-stage approach compared with the one-stage approach with approximately the same number of synthetic datasets. For example, releasing two-stage synthetic data with $M = 9$ carries an average true match risk of 67 (3.4 for the top 25 establishments), whereas releasing one-stage synthetic data

Table 9.7 Average CI overlap and match risk for two-stage synthesis based on ten simulations. Match risk for the 25 largest establishments is in parentheses.

m,r	Avg. overlap	Expected match risk	True match risk
m=3,r=3	0.867	83.1 (4.0)	67.6 (3.4)
m=3,r=16	0.868	98.0 (4.1)	91.8 (4.0)
m=3,r=33	0.870	99.8 (3.8)	96.3 (3.8)
m=5,r=10	0.869	106.1 (4.6)	101.2 (4.4)
m=10,r=5	0.875	113.8 (5.0)	109.4 (5.0)
m=16,r=3	0.874	119.9 (5.2)	116.4 (5.2)

with $m = 10$ has a true match risk of 82 (4.9). Risks are lower for $M \approx 50$ compared with one-stage synthetic data with $m = 50$ as well. Additionally, for the top 25 establishments, the percentage of unique matches that are true matches is lower for the two-stage approach. When $M = 9$, this percentage is 17% for the two-stage approach, compared with around 23% for one-stage synthetic data with $m = 10$. When $M \approx 50$, this percentage varied between 18% and 22%, whereas it is around 30% for one-stage synthetic data with $m = 50$.

The two-stage methods have lower disclosure risks at any given total number of released datasets because they provide fewer pieces of data about industry codes. This effect is evident in the two-stage results with $M \approx 50$. The risks increase monotonically with the number of imputations dedicated to the first stage.

Chapter 10
Chances and Obstacles for Multiply Imputed Synthetic Datasets

The main focus of the first statistical disclosure limitation (SDL) techniques proposed in the literature was on providing sufficient disclosure protection. At that time, agencies paid only little attention to the negative impacts of these approaches on data utility. Over the years, more and more sophisticated methods evolved. However, these methods also became more complicated to implement and often required correction methods difficult to apply for nonstandard analysis. For these reasons, most agencies still tend to rely on standard, easy-to-implement SDL techniques such as data swapping or noise addition, although it has been shown repeatedly that these methods can have severe negative consequences on data utility and may even fail to fulfill their primary goal – to protect the data sufficiently (see, for example, Winkler (2007b)).

Generating multiply imputed synthetic datasets is a promising alternative. With this approach, the user doesn't have to learn complicated adjustments that might differ depending on the kind of analysis the user wants to perform. Instead, she can use the combining rules presented in this book, which are simple and straightforward to calculate. With any synthetic data approach that is based on multiple imputation, the point estimate is simply the average of the point estimates calculated for every dataset, and its variance is estimated by a simple combination of the estimated variance within each dataset and the variance of the point estimates between the datasets. Furthermore, it is possible with synthetic datasets to account for many real data problems such as skip patterns and logical constraints (see Section 3.2 for details). Most standard SDL techniques cannot deal with these problems. Besides, it is very easy to address missing-data problems and confidentiality problems at the same time when generating partially synthetic datasets. Since both problems can be handled by multiple imputation, it is reasonable to impute missing values first and then generate synthetic datasets from the multiply imputed datasets as described in Chapter 8. This will actually increase the value of the generated datasets since the fully imputed, nonsynthesized datasets could be used by other researchers inside the agency who otherwise might not be able to adjust their analyses to account for the missing values properly.

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_10,

However, most research on generating synthetic data, especially with real data applications, dates back no more than five years, so it is not surprising that at the current stage there are some obstacles to this approach that still need to be addressed. First and foremost, many agencies complain that developing synthetic datasets for complex surveys is too labor intensive, takes too long, and requires experts who are familiar with the data on the one hand but also need detailed knowledge of Bayesian statistics and excellent modeling skills to generate synthetic data with a high level of data utility. Many small agencies cannot afford to fund research on synthetic data for several months or even years. Other agencies are reluctant to invest in a new data-dissemination strategy before the usefulness of this strategy has been clearly demonstrated. This may change with the release of high-quality synthetic data in the United States and in Germany. Besides, a new version of the multiple-imputation software IVEware (Raghunathan et al., 2002) for generating synthetic datasets is under development at the University of Michigan. This software will allow researchers without a sound background in modeling and Bayesian statistics to develop synthetic data. Another promising approach that might speed up the synthetic data generation is the use of nonparametric imputation methods such as CART (Reiter, 2005d) or random forests (Caiola and Reiter, 2010). With these approaches the modeling task can be simplified significantly. Evaluating to what extent the synthesis can be automated and testing the feasibility of these approaches for complex datasets with skip patterns and logical constraints is an area for future research.

But it is not only the agencies that are concerned about this new data-dissemination strategy. Many potential users of the released data are skeptical about the approach, too. They insist that they would only work with the original data, ignoring the fact that unrestricted access to the original data is not an option in many cases. It is important that users understand that they should focus on the potential benefits of this approach relative to other SDC methods instead of comparing the approach with unrestricted access. They also tend to see the original data as the true data, ignoring other sources of uncertainty and potential bias such as nonresponse, undercoverage, reporting or coding errors, etc., that might dwarf the additional bias potentially introduced by the synthesis. Furthermore, a common misconception is that the synthetic data will only provide valid results if the imputation model and the analyst's model match exactly. This is not true. If the imputation model contains more variables than the analyst's model, the results will still be valid, albeit with a reduced efficiency. But even if the imputation model does not contain all the variables that are included in the analyst's model, this does not necessarily mean that the results will be biased. In fact, if one variable is omitted from the imputation, the model implicitly assumes conditional independence between the dependent variable and this variable. Now, if the imputation model is already based on hundreds of variables, the assumption of conditional independence given all the other variables might be appropriate. In this case, the analyst would obtain valid results with the released data, even if some of the information in her model was not included in the imputation model.

Still, it would be misleading to praise the synthetic data approach as the panacea for data dissemination. It is simply impossible to generate a dataset with any kind of statistical disclosure limitation technique that provides valid results for any potential

analysis while at the same time guaranteeing 100% disclosure protection. The synthetic data reflect only those relationships included in the data generation models. When the models fail to reflect certain relationships accurately, analysts' inferences also will not reflect those relationships. Similarly, incorrect distributional assumptions built into the models will be passed on to the users' analyses. In practice, this dependence means that some analyses cannot be performed accurately and that agencies need to release information that helps analysts decide whether or not the synthetic data are reliable for their analyses. For example, agencies might include summaries of the posterior distributions of parameters in the data-generation models as attachments to public releases of data. Or, they might include generic statements that describe the imputation models, such as "Main effects for age, sex, and race are included in the imputation models for education." This transparency also is a benefit of the synthetic data approach: analysts are given indications of which analyses can be reliably performed with the synthetic data. Analysts who desire finer detail than afforded by the imputations may have to apply for special access to the observed data.

To overcome the skepticism against synthetic data, agencies can also offer some incentives to work with the synthetic data. For example the research teams at Cornell University and the IAB independently decided to offer the guarantee that for an initial phase any research that is performed on the synthetic data will also be run on the original data and the results from the original data will be sent back to the researcher after checks for potential confidentiality violations. This is a very strong incentive since researchers do not have to apply for access to the research data center but still can be sure that they will finally get the results from the original data. At the same time, they can compare the results from the original data with the results from the synthetic data, and if they repeatedly find that the results actually do not differ very much, they hopefully will give up some of their reservations against the use of synthetic data over time.

Finally, researchers tend to be reluctant to use new methods until they are implemented in standard statistical software and results are easily obtainable using standard commands. For example, the use of multiple imputation has significantly increased since routines to multiply impute missing values and analyze the imputed data became readily available in all major statistical software packages, such as Stata, SAS, or R. I suggest that agencies work with academic researchers and software developers to write software routines that implement the combining rules necessary to obtain valid results for the different synthetic data approaches.

The interest in synthetic data is ever-growing, and many seemingly insurmountable obstacles have been overcome in the last few years. There are still some efforts necessary to make the concept a universal, widely accepted, and easy-to-implement approach, but the first releases of partially synthetic datasets in the United States and Germany demonstrate that the approach successfully managed the critical step from a pure theoretical concept to practical implementation. Nevertheless, plenty of room remains for future research in this area that will further improve the feasibility of this approach. With the continuous proliferation of publicly available databases and improvements in record linkage technologies, releasing synthetic datasets might

soon be the only reasonable strategy to balance the trade-off between disclosure risk and data utility when disseminating data collected under the pledge of privacy to the public.

Appendix A
Bill Winkler's Microdata Confidentiality References (August 1, 2009)

Abowd, J. M., and Vilhuber, L. (2008). "How Protective are Synthetic Data?" in (J. Domingo-Ferrer and V. Yucel, eds.) *Privacy in Statistical Databases*, New York, N.Y.: Springer, 239-246.

Abowd, J. M., and Woodcock, S. D. (2002), "Disclosure Limitation in Longitudinal Linked Data," in (P. Doyle et al, eds.) *Confidentiality, Disclosure, and Data Access*, Amsterdam, The Netherlands: North Holland.

Abowd, J. M., and Woodcock, S. D. (2004), "Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data, in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases*, New York: Springer.

Adams, N. R., and Wortmann, J. C., (1989), "Security-control Methods for Statistical Databases, A Comparative Study," ACM Computing Surveys, 21, 515-556.

Aggarwal, C. C., (2005), "On k-Anonymity and the Curse of Dimensionality," *Proceedings of the 31st VLDB Conference*, Trondheim, Norway, `http://www.vldb2005.org/program/paper/fri/p901-aggarwal.pdf`.

Aggarwal, C. C., and Parthasarathy, S. (2001), "Mining Massively Incomplete Data Sets through Conceptual Reconstruction," *Proceedings of the ACM KDD Conference*, 227-232.

Aggarwal, C. C., and Yu, P. (2004), "A Condensation Approach to Privacy Preserving Data Mining," *Proceedings of the EBDT Conference*, 183-199.

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A. (2005), "Anonymizing Tables," *International Conference on Database Theory*.

Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S. Panigrahy, R., Thomas, D., and Zhu, A. (2006), "Achieving Anonymity via Clustering," *ACM PODS '06*.

Agrawal, D., and Aggarwal, C. C. (2001), "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," Association of Computing Machinery, *Proceedings of PODS 2001*, 247-255.

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4419-9554-4,

Agrawal, R., and Srikant, R. (2000), Privacy Preserving Data Mining, *Proceedings of the ACM SIGMOD 2000*, 439-450.

Agrawal, R., Srikant, R., and Thomas, D. (2005), "Privacy Preserving OLAP," *ACM SIGMOD Conference*.

Agrawal, R., Kiernan, J., Srikant, R., and Xu, Y. (2002), "Hippocratic Databases," *Very Large Databases 2002*.

Bacher, J., Bender, S., and Brand, R. (2001), "Re-identifying Register Data by Survey Data: An Empirical Study," presented at the *UNECE Workshop On Statistical Data Editing*, Skopje, Macedonia, May 2001.

Bacher, J., Brand, R., and Bender, S. (2002) Re-identifying Register Data by Survey Data using Cluster Analysis: An Empirical Study, *International Journal of Uncertainty, Fuzziness, Knowledge-Based Systems*, 10 (5) 589-608.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., and Talwar, K. (2007), "Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release," *PODS '07*, Beijing, China.

Bayardo, R. J., and Agrawal, R. (2005), "Data Privacy Through Optimal K-Anonymization," *IEEE 2005 International Conference on Data Engineering*.

Bethlehem, J. A., Keller, W. J., and Pannekoek, J., (1990), "Disclosure Control of Microdata," *Journal of the American Statistical Association*, 85, 38-45.

Blien, U., Wirth, U., and Muller, M. (1992), "Disclosure Risk for Microdata Stemming from Official Statistics," *Statistica Neerlandica*, 46, 69-82.

Blum, A., Dwork, C., McSherry, F., and Nissim, K. (2005), "Practical Privacy: The SuLQ Framework," *ACM SIGMOD Conference* (also `http://research.microsoft.com/research/sv/DatabasePrivacy/bdmn.pdf`).

Brand, R. (2002), "Microdata Protection Through Noise Addition," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, New York: Springer, 97-116.

Castro, J. (2004), "Computational Experience with Minimum-Distance Controlled Perturbation Methods," in (J. Domingo-Ferrer, ed.), *Privacy in Statistical Databses 2004*, Springer: New York.

Chawla, S., Dwork, C., McSherry, F., Smith, A., and Wee, H. (2004), "Toward Privacy in Public Databases," Microsoft Research Technical Report, Theory of Cryptography Conference.

Chawla, S., Dwork, C., McSherry, F., and Talwar, K. (2005), "On the Utility of Privacy-Preserving Histograms," `http://research.microsoft.com/research/sv/DatabasePrivacy/cdmt.pdf`.

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

Dandekar, R. A. (2004), Maximum Utility Minimum Information Loss Table Server Design of Statistical Disclosure Control of Tabular Data, in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases*, New York: Springer, 121-135.

Dandekar, R. A., Domingo-Ferrer, J., and Sebe, F. (2002), "LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, New York: Springer, 175-186.

Dandekar, R., Cohen, M., and Kirkendal, N. (2002), "Sensitive Microdata Protection Using Latin Hypercube Sampling Technique," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, New York: Springer, 117-125.

Defays, D., and Anwar, M. N. (1998), "Masking Microdata Using Microaggregation," *Journal of Official Statistics*, 14, 449-461.

Defays, D., and Nanopolis, P. (1993), "Panels of Enterprises and Confidentiality: the Small Aggregates Method," in *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, 195-204.

De Waal, A. G., and Willenborg, L.C.R.J. (1995), "Global Recodings and Local Suppressions in Microdata Sets," *Proceedings of Statistics Canada Symposium* 95, 121-132.

De Waal, A. G., and Willenborg, L.C.R.J. (1996), "A View of Statistical Disclosure Control for Microdata," *Survey Methodology*, 22, 95-103. De Wolf, P.-P. (2007), "Risk, Utility and PRAM: A Comparison of Proximity Swap and Data Shuffle for Numeric Data," in (J. Domingo-Ferrer, ed.) *Statistical Data Protection 2006*, Springer: New York, N.Y.

Dinur, I., and Nissim, K. (2003), "Revealing Information while Preserving Privacy," *ACM PODS Conference*, 202-210.

Domingo-Ferrer, J. (2001), "On the Complexity of Microaggregation," presented at the *UNECE Workshop On Statistical Data Editing*, Skopje, Macedonia, May 2001.

Domingo-Ferrer, J. (ed.) (2002) *Inference Control in Statistical Databases*, New York: Springer

Domingo-Ferrer, J., and Mateo-Sanz, J. M. (2001), "An Empirical Comparison of SDC Methods for Continuous Microdata in Terms of Information Loss And Re-Identification Risk," presented at the *UNECE Workshop On Statistical Data Editing*, Skopje, Macedonia, May 2001.

Domingo-Ferrer, J., and Mateo-Sanz, J. M. (2002), "Practical Data-Oriented Microaggregation for Statistical Disclosure Control," *IEEE Transactions on Knowledge and Data Engineering*, 14 (1), 189-201.

Domingo-Ferrer, J., Mateo-Sanz, J., Oganian, A., and Torres, A. (2002), "On the Security of Microaggregation with Individual Ranking: Analytic Attacks," *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 477-492.

Domingo-Ferrer, J., Sebé, F., and Castellà-Roca, J. (2004), "On the Security of Noise Addition for Privacy In Statistical Databases, in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases*, New York: Springer, 149-161.

Domingo-Ferrer, J., and Torra, V. (2001) A Quantitative Comparison of Disclosure Control Methods for Microdata in (P. Doyle, J. Lane, J. Theeuwes, and L.

Zayatz, eds.) *Confidentiality, Disclosure Control and Data Access: Theory and Practical Applications*, Amsterdam, The Netherlands: North Holland, 111-134.

Domingo-Ferrer, J., and Torra, V. (2003), "Statistical Data Protection in Statistical Microdata Protection Via Advanced Record Linkage," *Statistics and Computing*, 13 (4), 343-354.

Du., W., Han, Y. S., and Chen, S. (2004), "Privacy Preserving Multivariate Statistical Analysis: Linear Regression and Classification," *SIAM International Conference on Data Mining 2004*.

Du, W., and Zhan, Z. (2003), "Using Randomized Response Techniques for Privacy Preserving Data Mining," *ACM Knowledge Discovery and Data Mining Conference 2003*, 505-510.

DuMouchel, W., Volinsky, C., Johnson, T., Cortes, C., and Pregibon, D. (1999), "Squashing Flat Files Flatter," *Proceedings of the ACM Knowledge Discovery and Data Mining Conference*, 6-15.

Duncan, G. T., Keller-McNulty, S. A., and Stokes, S. L. (2001), "Disclosure Risk vs. Data Utility: The R-U Confidentiality Map," Los Alamos National Laboratory Technical Report LA-UR-01-6428.

Dwork, C. (2006), "Differential Privacy," *33rd International Colloquium on Automata, Languages and Programming - ICALP 2006*, Part II, 1-12.

Dwork, C. (2008), "Differential Privacy: A Survey of Results," in (M. Agrawal et al., eds.) *TAMC 2008*, LNCS 4978, 1-19.

Dwork, C., and Lei, J. (2009), "Differential Privacy and Robust Statistics," *STOC '09*.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006), "Calibrating Noise to Sensitivity in Private Data Analysis," *3rd Conference on Cryptography - TCC 2006*, 365-384.

Dwork, C., McSherry, F., and Talwar, K. (2007a), "The Price of Privacy and the Limits of LP Decoding," *STOC '07*, San Diego, CA.

Dwork, C., McSherry, F., and Talwar, K. (2007b), "Differentially Private Marginals Release with Mutual Consistency and Error Independent Sample Size," *UNECE Worksession on Statistical Data Confidentiality*, Manchester, UK,at `http://www.unece.org/stats/documents/2007/12/confidentiality/wp.19.e.pdf`.

Dwork, C. and Yekhanin, S. (2008), "New Efficient Attacks on Statistical Disclosure Control Mechanisms," Advances in *Cryptology-CRYPTO 2008*, to appear, also at http://research.microsoft.com/research/sv/DatabasePrivacy/dy08.pdf .

Dwork, C. and Nissim, K. (2004), "Privacy-Preserving Datamining on Vertically Partitioned Databases," Microsoft Research Technical Report.

Elamir, E. A. H. (2004), "Analysis of Re-identification Risk based on Log-Linear Model," in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases*, Springer: New York, 273-281.

Elamir, E. A. H. and Skinner, C. J. (2006), "Record Level Measures of Disclosure Risk for Survey Microdata," *Journal of Official Statistics*, 22, 525-539.

Elliott, M. A., Manning, A. M., and Ford, R. W. (2002), "A Computational Algorithm for Handling the Special Uniques Problem," *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 10 (5), 493-510.

Elliot, M.J., Skinner, C. A., and Dale, A. (1998), "Special Uniques, Random Uniques, and Sticky Populations: Some Counterintuitive Effects of Geographic Detail in Disclosure Risk," *Research in Official Statistics*, 1, 53-68.

Evfimievski, A. (2004), "Privacy Preserving Information Sharing," Ph.D. Dissertation, Cornell University, `http://www.cs.cornell.edu/aevf/` .

Evfimievski, A., Gehrke, J., and Srikant, R. (2003), "Limiting Privacy Breaches in Privacy Preserving Data Mining," *ACM PODS Conference*, 211-222.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. (2002), "Privacy Preserving Mining of Association Rules," Association of Computing Machinery, Special Interest Group on Knowledge Discovery and Datamining 2002.

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

Fellegi, I. P. (1972), "On the Question of Statistical Confidentiality," *Journal of the American Statistical Association*, 67, 7-18.

Fellegi, I. P. (1999), "Record Linkage and Public Policy - A Dynamic Evolution," *Proceedings of the Record Linkage Workshop 1997*, Washington, DC: National Academy Press, 3-12.

Fienberg, S. E. (1997), "Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, commissioned by Committee on National Statistics of the National Academy of Sciences.

Fienberg, S. E. and MacIntyre, J. (2005), "Data Swapping: Variations on a Theme of Dalenius and Reiss," *Journal of Official Statistics*, 21 (2), 309-323.

Fienberg, S. E., Makov, E. U., and Sanil, A. P., (1997), "A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data," *Journal of Official Statistics*, 14, 75-89.

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

Frakes, W., and Baeza-Yates, R. (1992), *Information Retrieval - Data Structures and Algorithms*, Upper Saddle River, NJ:.Prentice-Hall.

Franconi, L., Capobianchi, A., Polletini, S., and Seri, G. (2001), "Experiences in Model-Based Disclosure Protection," presented at the *UNECE Workshop on Statistical Data Confidentiality*, Skopje, Macedonia, May 2001.

Fuller, W. A. (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, 9, 383-406 (`http://www.jos.nu/Articles/abstract.asp?article=92383`).

Fung, B. C. M., Wang, K., and Yu, P. S. (2005), Top-Down Specialization for Information and Privacy Preservation," *IEEE International Conference on Data Engineering*, 205-216.

Ganta, S., Prasad, S., and Smith, A. (2008), Compositional Attacks and Auxiliary Information in Data Privacy," *ACM KDD* '08, 265-273.

Gopal, R., Goes, P. and Garfinkel, R. (1998) "Confidentiality Via Camouflage: The CVC Approach to Database Query Management," in *Statistical Data Protection '98*, Eurostat, Brussels, Belgium, 1-8. (also (2002) *Operations Research*, 50 (3)).

Gilburd, B., Schuster, A., and Wolff, R. (2004b), "k-TTP: A New Privacy Model for Large-Scale Distributed Environments," *ACM Knowledge Discovery and Data Mining Conference 2004*, 599-604.

Gill, L. (1999), "OX-LINK: The Oxford Medical Record Linkage System," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 15-33.

Gomatam, S. V., and Karr, A. (2003), "On Data Swapping of Categorical Data," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, CD-ROM.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L. C. R. J., and De Wolf, P.-P. (1998), "Post Randomisation For Statistical Disclosure Control: Theory and Implementation," *Journal of Official Statistics*, 14, 463-478.

Graham, P., Young, J., and Penny, R. (2009), "Multiply Imputed Synthetic Data: Evaluation of Hierarchical Bayesian Imputation Models," *Journal of Official Statistics*, 25 (2), 245-268.

Grim, J., Bocek, P., and Pudil, P. (2001), "Safe Dissemination of Census Results by Means of Interactive Probabilistic Models," *Proceedings of 2001 NTTS and ETK*, Luxembourg: Eurostat, 849-856.

Huang, Z., Du, W., and Chen, B. (2005), "Deriving Private Information from Randomized Data," *ACM SIGMOD 2005 Conference*, 37-48.

Huckett, J. C. (2008), "Synthetic Data Methods for Disclosure Limitation," Ph.D. Thesis, Department of Statistics, Iowa State University.

Huckett, J. C., and Larsen, M. D. (2007), "Microdata Simulation for Confidentiality Protection Using Regression Quantiles and Hot Deck," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, CD-ROM, 3053-3060.

Hwang, J. T. (1986), "Multiplicative Error-in-Variables Models with Applications to Recent Data Released by the U.S. Department of Energy," *Journal of the American Statistical Association*, 81 (395), 680-688.

Iyengar, V. (2002), "Transforming Data to Satisfy Privacy Constraints," Association of Computing Machinery, *Knowledge Discovery and Datamining Conference 2002*, 279-288.

Kantarcioglu, M., and Clifton, C. (2004a), "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," *IEEE Transactions on Knowledge and Data Engineering*, 16 (9), 1026-1037.

Kantarcioglu, M., and Clifton, C. (2004b), "When Do Data Mining Results Violate Privacy?" Association of Computing Machinery, *Knowledge Discovery and Data Mining Conference 2004*, 599-604.

Kargupta, H., Datta, S., Wang, Q., and Ravikumar, K. (2003) "Random Data Perturbation Techniques and Privacy Preserving Data Mining," Expanded version of best paper awarded paper from the *IEEE International Conference on Data Mining*, November, 2003, Orlando, FL, (also version to appear in Knowledge and Information Systems Journal, http://www.cs.umbc.edu/~hillol/PUBS/kargupta_privacy03a.pdf).

Kaufmann, S., Seastrom, M., and Roey, S. (2005), "Do Disclosure Controls to Protect Confidentiality Degrade the Quality of the Data? Data from the 2003 Trends in Mathematics and Science Study," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, CD-ROM.

Keller-McNulty, S., and Unger, E. (2003), "Database Systems: Inferential Security," *Journal of Official Statistics*, 9 (2), 475-499.

Kennickell, A. B. (1999), "Multiple Imputation and Disclosure Control: The Case of the 1995 Survey of Consumer Finances," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 248-267 (available at http://www.fcsm.govunderMethodologyreports).

Kifer, D. (2009), "Attacks on Privacy and deFinetti's Theorem," *ACM SIGMOD Conference*.

Kifer, D. and Gehrke, J. (2006), "Injecting Utility into Anonymized Data Sets," *ACM SIGMOD*.

Kim, J. J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 370-374 (http://www.amstat.org/sections/SRMS/Proceedings/papers/1986_069.pdf).

Kim, J. J. (1990), "Subdomain Estimation for the Masked Data," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 456-461 (http://www.amstat.org/sections/SRMS/Proceedings/papers/1990_075.pdf).

Kim, J. J., and Winkler, W. E. (1995), "Masking Microdata Files," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 114-119 (http://www.amstat.org/sections/SRMS/Proceedings/papers/1995_017.pdf , longer report http://www.census.gov/srd/papers/pdf/rr97-3.pdf).

Kim, J. J., and Winkler, W. E. (2001), "Multiplicative Noise for Masking Continuous Data," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, CD-ROM.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm," *Journal of Official Statistics*, 9, 313-331 (http://www.jos.nu/Articles/abstract.asp?article=92313).

Lane, J. (2007), "Optimizing the Use of Microdata: An Overview of the Issues," *Journal of Official Statistics*, 23 (3), 299-317 (http://www.jos.nu/Articles/abstract.asp?article=233299).

Lawrence, C., Zhou, J.L., and Tits, A. L. (1997), "User's Guide for CFSZP Version 2.5: A C Code for Solving (Large Scale) Constrained Nonlinear Inequality Constraints," Unpublished, Electrical Engineering Dept. and Institute for Systems Research, University of Maryland.

Lakshmanan, L., Ng, R., and Ramesh, G. (2005), "To Do or Not To Do - The Dilemma of Disclosing Anonymized Data," *ACM SIGMOD Conference*.

Lakshmanan, L. K. S., Ng, R., Ramesh, G. (2008), "On Disclosure Risk Analysis of Anonymized Itemsets in the Presences of Prior Knowledge," *ACM Transactions on Knowledge Discovery from Data*, 2 (3), 13.1-13.44.

LeFevre, K., DeWitt, D. and Ramakrishnan, R. (2005), "Incognito: Efficient Full-Domain K-Anonymity," *ACM SIGMOD Conference*.

Li, X.-B., and Sarkar, S. (2007), "A Tree-Based Data Perturbation Technique for Privacy-Preserving Data Mining," *IEEE Transactions on Knowledge and Data Engineering*, 18 (9), 1278-1283.

Liew, C. K., Choi, U. J., and Liew, C. J. (1991), "A Data Distortion by Probability Distribution," *ACM Transactions on Database Systems*, 10, 395-411.

Little, R. J. A. (1993), "Statistical Analysis of Masked Data," *Journal of Official Statistics*, 9, 407-426 (http://www.jos.nu/Articles/abstract.asp?article=92407).

Little, R. J. A., and Liu, F. (2002), "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, CD-ROM.

Little, R. J. A., and Liu, F. (2003), "Comparison of SMIKe with Data-Swapping and PRAM for Statistical Disclosure Control of Simulated Microdata," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, CD-ROM.

Lindell, Y. and Pinkas, B. (2002), "Privacy Preserving Data Mining," *Proceedings of Crypto 2000*, Springer LNCS 1880, 20-24.

Liu, H., Kargupta, H., and Ryan, J. (2007), "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining," *IEEE Transactions of Knowledge and Data Engineering*, 18 (1), 92-106.

Machanavajjhala, A., Gehrke, and M., Goetz. (2009), "Data Publishing against Realistic Adversaries," *VLDB 2009*.

Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramamiam, M. (2005), "l-Diversity: Privacy Beyond k-Anonymity," Cornell CS Dept. technical report, ht

tp://www.cs.cornell.edu/johannes/papers/2005/publishing-icde-final.pdf.

Machanavajjhala, A., Kifer, D., Abowd, J. Gehrke, J., and Vilhuber, L. (2008), "Privacy: Theory meets Practice on the Map," *ICDE 2008*.

Malin, B. Sweeney, L., and Newton, E. (2003), "Trail Re-identification: Learning Who You are from Where You have Been, *Workshop on Privacy in Data*, Carnegie-Mellon University, March 2003.

McSherry, F. (2009), "Privacy Integrated Queries," *SIGMOD 2009*.

McSherry, F., and Talwar, K. (2007), "Mechanism design via differential privacy," *Proceedings of the 48th Symposium of the Foundations of Computer Science*.

Mera, R. (1998), "Matrix Masking Methods That Preserve Moments," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 445-450.

Moore, R. (1995), "Controlled Data Swapping Techniques For Masking Public Use Data Sets," U.S. Bureau of the Census, Statistical Research Division Report rr96/04, (available at http://www.census.gov/srd/www/byyear.html).

Moore, A. W., and Lee, M. S. (1998), "Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets," *Journal of Artificial Intelligence Research*, 8, 67-91.

Motwani, R., and Xu, Y. (2007), "Efficient Algorithms for Masking and Finding Quasi-Identifiers," *VLDB '07*.

Müller, W., Blien, U., and Wirth, H. (1995), "Identification Risks of Micro Data," Evidence from Experimental Studies. *Sociological Methods and Research*, 24, 131-157.

Muralidhar, K., Batrah, D., and Kirs, P.J. (1995), "Accessibility, Security, and Accuracy in Statistical Databases : The Case for the Multiplicative Fixed Data Perturbation Approach," *Management Science*, 41 (9), 1549-1584

Muralidhar, K., Parsa, R., and Sarathy, R. (1999), "A General Additive Data Perturbation Method for Database Security," *Management Science*, 45 (10), 1399-1415.

Muralidhar, K., and Sarathy, R. (1999) "Security of Random Data Perturbation Methods," *ACM Transactions on Database Systems*, 24 (4), 487-493.

Muralidhar, K., and Sarathy, R. (2003), "A Theoretical Basis for Perturbation Methods," *Statistics and Computing*, 13 (4), 329-335.

Muralidhar, K., and Sarathy, R. (2006a), "Data Shuffling - A New Masking Approach to Numerical Data," *Management Science*, 52 (5), 658-670.

Muralidhar, K., and Sarathy, R. (2006b), "A Theoretical Basis for Perturbation Methods,"*Journal of Statistics*, 22 (3), 507-524.

Muralidhar, K. and Sarathy, R. (2007), " 'Easy to Implement' is Putting the Cart before the Horse: Effective Techniques for Masking Numerical Data," Federal

Committee on Statistical Methodology Research Conference, to appear on CD-ROM.

Muralidhar, K., Sarathy, R., and Dandekar, R. (2006), "Why Swap When You Can Shuffle? A Comparison of Proximity Swap and Data Shuffle for Numeric Data," in (J. Domingo-Ferrer, ed.) *Statistical Data Protection 2006*, Springer: New York, N.Y.

Muralidhar, K., Sarathy, R., and Parsa, R. (2001) "An Improved Security Requirement for Data Perturbation with Implications for E-Commerce," *Decision Sciences*, 32 (4), 683-698.

Nissim, K., Raskhodnikova, S., and Smith, A. (2007), "Smooth Sensitivity and Sampling in Private Data Analysis," *STOC'07*, June 11-13, 2007, San Diego, California, USA.

Onn, S. (2007), "Entry Uniqueness in Margined Tables," in (J. Domingo-Ferrer, ed.) *Statistical Data Protection 2006*, Springer: New York, N.Y.

Owen, A. (2003), "Data Squashing by Empirical Likelihood," Data Mining and Knowledge Discovery, 7 (1), 101-113.

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

Palley, M. A., and Simonoff, J. S. (1987), "The Use of Regression Methodology for the Compromise of Confidential Information in Statistical Databases," *ACM Transactions on Database Systems*, 12 (4), 593-608.

Polletini, S. (2003), "Maximum Entropy Simulation for Microdata Protection," *Statistics and Computing*, 13 (4), 307-320.

Polletini, S., Franconi, L., and Stander, J. (2002), "Model Based Disclosure Protection," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*: New York: Springer.

Polletini, S., and Stander, J. (2004), "A Bayesian Hierarchical Model Approach to Risk Estimation in Statistical Disclosure Limitation," in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases*, New York: Springer, 247-261

Raghunathan, T. E. (2003), Evaluation of Inferences from Multiple Synthetic Data Sets Created Using Semiparametric Approach, Panel on Confidential Data Access for Research Purposes, Committee On National Statistics, October 2003.

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., and Sollenberger, P. (1998), "A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models," Survey Research Center, University of Michigan.

Raghunathan, T. E., and Reiter, J. P. (2007), "The Multiple Adaptations of Multiple Imputation," *Journal of the American Statistical Association*, 102, 1462-1471.

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

Raghunathan, T.E., and Rubin, D.R. (2000), "Multiple Imputation for Disclosure Limitation" University of Michigan, Department of Biostatistics technical report

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

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

Reiter, J.P. (2003a), "Inference for Partially Synthetic, Public Use Data Sets," *Survey Methodology*, 181-189.

Reiter, J.P. (2003b), Estimating Probabilities of Identification for Microdata, Panel on Confidential Data Access for Research Purposes, Committee On National Statistics, October 2003.

Reiter, J.P. (2005a), "Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study, *Journal of the Royal Statistical Society*, A, 168 (1), 185-205.

Reiter, J. P. (2005b), "Estimating Risk of Identify Disclosure in Microdata," *Journal of the American Statistical Association*, 100, 1103-1112.

Reiter, J. P. (2008), "Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure protection," *Statistics and Probability Letters*, 78, 15-20.

Reiter, J. P. and Mitra, R. (2009), "Estimating risks of identification disclosure in partially synthetic data," *Journal of Privacy and Confidentiality*, 01 (01), 99-110.

Roque, G. M. (2000), "Masking Microdata Files with Mixtures of Multivariate Normal Distributions," Ph.D.Dissertation, Department of Statistics, University of California at Riverside.

Rubin, D. B. (1993), "Satisfying Confidentiality Constraints through the Use of Synthetic Multiply-imputed Microdata," *Journal of Official Statistics*, 91, 461-468.

Samarati, P. (2001), "Protecting Respondents' Identity in Microdata Release," *IEEE Transactions on Knowledge and Data Engineering*, 13 (6), 1010-1027.

Samarati, P., and Sweeney, L. (1998), "Protecting Privacy when Disclosing Information: k-anonymity and Its Enforcement through Generalization and Cell Suppression," Technical Report, SRI International.

Sarathy, R., and Muralidhar, K. (2002), "The Security of Confidential Numerical Data in Databases," *Information Systems Research*, 48 (12), 1613-1627.

Sarathy, R., Muralidhar, K., and Parsa, R. (2002), Perturbing Non-Normal Attributes: The Copula Approach, *Management Science*, 48 (12), 1613-1627

Scheuren, F., and Winkler, W. E. (1993), "Recursive Merging and Analysis of Administration Lists," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 123-128 (presently available on `http://www.amstat.orgintheSectiononGovernmentStatistics`).

Scheuren, F., and Winkler, W. E. (1997), "Regression Analysis of Data Files that are Computer Matched - Part II," *Survey Methodology*, 157-165).

Schlörer, J. (1981), "Security of Statistical Databases: Multidimensional Transformation," *ACM Transactions on Database Systems*, 6, 91-112.

Skinner, C. J., and Elliot, M. A. (2001), "A Measure of Disclosure Risk for Microdata," *Journal of the Royal Statistical Society*, 64 (4), 855-867.

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

Skinner, C. J. and Shlomo, N. (2007), "Assessing Identification Risk in Survey Data Using Loglinear Models," *Journal of the American Statistical Association*, 103 (483), 989-1001, also at `http://eprints.soton.ac.uk/48103/`.

Stander, J., and Franconi, L. (2001), "A Model-Based Disclosure Limitation Method for Business Microdata," presented at the *UNECE Workshop On Statistical Data Editing*, Skopje, Macedonia, May 2001.

Sullivan, G., and Fuller, W. A. (1989), "The Use of Measurement Error to Avoid Disclosure," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 802-807.

Sullivan, G., and Fuller, W. A. (1990), "Construction of Masking Error for Categorical Variables," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 435-439.

Sweeney, L. (1999), "Computational Disclosure Control for Medical Microdata: The Datafly System" in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 442-453.

Sweeney, L. (2002), "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 10 (5), 571-588.

Sweeney, L. (2004), "Optimal Anonymity using K-similar, a New Clustering Algorithm," manuscript.

Takemura, A. (2002), "Local Recoding and Record Swapping by Maximum Weight Matching for Disclosure Control of Microdata Sets," *Journal of Official Statistics*, 18 (2), 275-289.

Tendick, P., and Matloff, N. (1994), "A Modified Random Perturbation Method for Database Security," *ACM Transactions on Database Systems*, 19, 47-63.

Thibaudeau, Y. (2004), "An Algorithm for Computing Full Rank Minimal Sufficient Statistics with Application to Confidentiality Protection, *UNECE Statistical Journal*, to appear.

Thibaudeau, Y., and Winkler, W.E. (2002), "Bayesian Networks Representations, Generalized Imputation, and Synthetic Microdata Satisfying Analytic Restraints," Statistical Research Division Report RR 2002/09 at `http://www.census.gov/srd/www/byyear.html`.

Thibaudeau, Y., and Winkler, W.E. (2004), "Full Rank Minimal Statistics for Disclosure Limitation and Variance Estimation: A Practical Way to Release Count Information," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, CD-ROM.

Torra, V. (2004), "OWA Operators in Data Modeling and Re-identification," *IEEE Transactions on Fuzzy Systems*, 12 (5), 652-660.

Torra, V., Domingo-Ferrer, J., and Abowd, J. (2007), "Using Mahalanobis-Distance Record Linkage for Disclosure Risk Assessment," in (J. Domingo-Ferrer, ed.) *Statistical Data Protection 2006*, Springer: New York, N.Y.

Torra, V., and Miyamoto, S. (2004), "Evaluating Fuzzy Clustering Algorithms for Microdata Protection," in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases*, New York: Springer, 175-186.

Trottini, M., and Fienberg, S. E. (2002), "Modelling User Uncertainty for Disclosure Risk and Data Utility,"*International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 511-528.

Van Den Hout, A., and Van Der Heijden, P. G. M. (2002), "Randomized Response, Statistical Disclosure Control, and Misclassification: A Review," *International Statistical Review*, 70 (2), 269-288.

Van Gewerden, L., Wessels, A., and Hundepol, A. (1997), "Mu-Argus Users Manual, Version 2," Statistics Netherlands, Document TM-1/D.

Willenborg, L., and De Waal, T. (1996), *Statistical Disclosure Control in Practice*, Vol. 111, Lecture Notes in Statistics, New York: Springer.

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

Willenborg, L. and Van Den Hout, A. (2006), "Peruco: A Method for Producing Safe and Consistent Microdata," *International Statistical Review*, 74 (2), 271-284.

Winglee, M., Valliant, R., Clark, J., Lim, Y., Weber, M., and Strudler, M. (2002), "Assessing Disclosure Protection for the SOI Public Use File," American Statistical Association, *Proceedings of the Section on Survey Research Methods*, CD-ROM.

Winkler, W. E. (1994), "Advanced Methods for Record Linkage, American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 467-472 (`http://www.census.gov/srd/papers/pdf/rr94-5.pdf`).

Winkler, W. E. (1995), "Matching and Record Linkage," in (B. G. Cox et al, ed.) *Business Survey Methods*, New York: J. Wiley, 355-384 (also `http://www.fcsm.gov/working-papers/wwinkler.pdf`).

Winkler, W. E. (1997), "Views on the Production and Use of Confidential Microdata," Statistical Research Division report RR 97/01 at `http://www.census.gov/srd/www/byyear.html`.

Winkler, W. E. (1998), "Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata," *Research in Official Statistics*, 1, 87-104, `http://www.census.gov/srd/papers/pdf/rrs2005-09.pdf`.

Winkler, W. E. (2002a), "Using Simulated Annealing for k-anonymity," Statistical Research Division report RR 2002/07 at `http://www.census.gov/srd/www/byyear.html`.

Winkler, W. E. (2002b), "Single Ranking Microaggregation and Re-identification," Statistical Research Division report RR 2002/08 at http://www.census.gov/srd/www/byyear.html.

Winkler, W. E. (2004a), Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems, in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases*, Springer: New York, 231-247, also http://www.census.gov/srd/papers/pdf/rrs2004-06.pdf.

Winkler, W. E. (2004b), Re-identification Methods for Masked Microdata, in (J. Domingo-Ferrer and V. Torra, eds.), *Privacy in Statistical Databases*, Springer: New York, 216-230, also http://www.census.gov/srd/papers/pdf/rrs2004-03.pdf.

Winkler, W. E. (2005), "Modeling and Quality of Masked Microdata," American Statistical Association, *Proceedings of the Section on Survey Research Method*, CD-ROM, also http://www.census.gov/srd/papers/pdf/rrs2006-01.pdf.

Winkler, W. E. (2007a), "Analytically Valid Discrete Microdata and Re-identification," http://www.census.gov/srd/papers/pdf/rrs2007-19.pdf.

Winkler, W. E. (2007b), "Examples of Easy-to-implement, Widely Used Masking Methods for which Analytic Properties are not Justified," http://www.census.gov/srd/papers/pdf/rrs2007-21.pdf.

Winkler, W. E. (2008), "General Discrete-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties," *IAB Workshop on Confidentiality and Disclosure*, http://fdz.iab.de/en/FDZ_Events/SDC-Workshop.aspx, Nuremberg, Germany, November 20-21, 2008.

Winkler, W. E. (2009), Should Social Security numbers be replaced by modern, more secure identifiers?" *Proceedings of the National Academy of Science*.

Woo, M., Reiter, J. P., Oganian, A., Karr, A. F. (2009) "Global measures of data utility for microdata masked for disclosure limitation," *Journal of Privacy and Confidentiality*, 01 (01), 111-124.

Xiao, X., and Tao, Y. (2007a), "m-Invariance: Towards Privacy Preserving Republication of Dynamic Datasets," *ACM SIGMOD*, 689-700.

Xiao, X., and Tao, Y. (2007b), "Anatomy: Simple and Effective Privacy Preservation," *VLDB*, 139-150.

Xiao, X., and Tao, Y. (2008a), "Dynamic Anonymization: Accurate Statistical Analysis with Privacy Preservation," *ACM SIGMOD*, 107-120.

Xiao, X., and Tao, Y. (2008b), "Output Perturbation with Query Relaxation," *VLDB*, 857-869.

Yancey, W.E., Winkler, W.E., and Creecy, R. H. (2002) "Disclosure Risk Assessment in Perturbative Microdata Protection," in (J. Domingo-Ferrer, ed.) *Inference Control in Statistical Databases*, New York: Springer, 135-151, (also http://www.census.gov/srd/papers/pdf/rrs2002-01.pdf).

Yang, Z., Zhong, S., and Wright, R. (2005), "Anonymity Preserving Data Collection," *ACM KDD Conference*, 334-343.

Zhang, N., Wang, S., and Zhao, W. (2005), "A New Scheme on Privacy-Preserving Data Classification," *ACM KDD Conference*, 374-383.

Zhong, S., Yang, Z., and Wright, R. (2005), "Privacy-Enhancing k-Anonymization of Customer Data," *ACM Principals of Database Systems Conference 2005*, 139-147.

Zhu, Y., and Liu, L. (2004), "Optimal Randomization for Privacy Preserving Data Mining," *ACM Knowledge Discovery and Data Mining Conference 2004*, 761-765.

Appendix B
Binned Residual Plots to Evaluate the Imputations for the Categorical Variables

Figure B.1 presents the binned residual plots for all 59 categorical variables with missing rates $\geq 1\%$. For variables with more than two categories, I present a graph for each category (the first category is always defined as the reference category in the multinomial imputation model).[1]

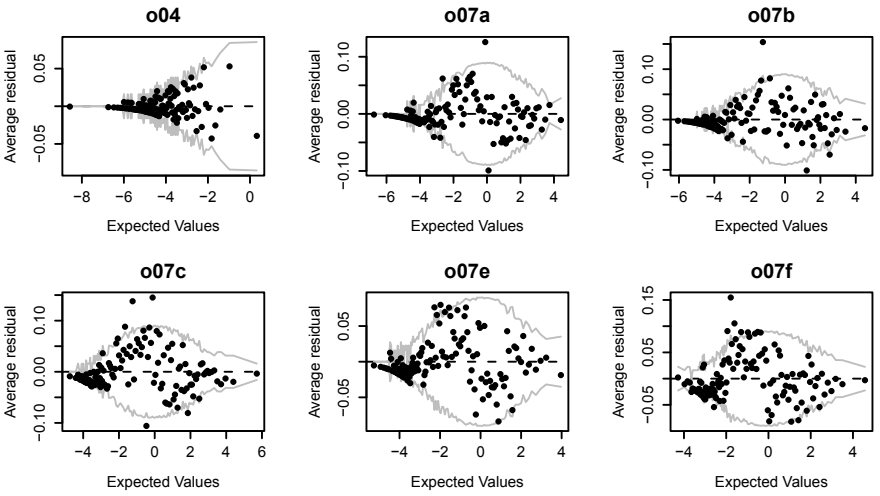

Fig. B.1 Binned residual plots for the categorical variables with missing rates above 1%.

[1] For readability, I use the internal labeling for the variables. A detailed description of all variables can be obtained from the author upon request.

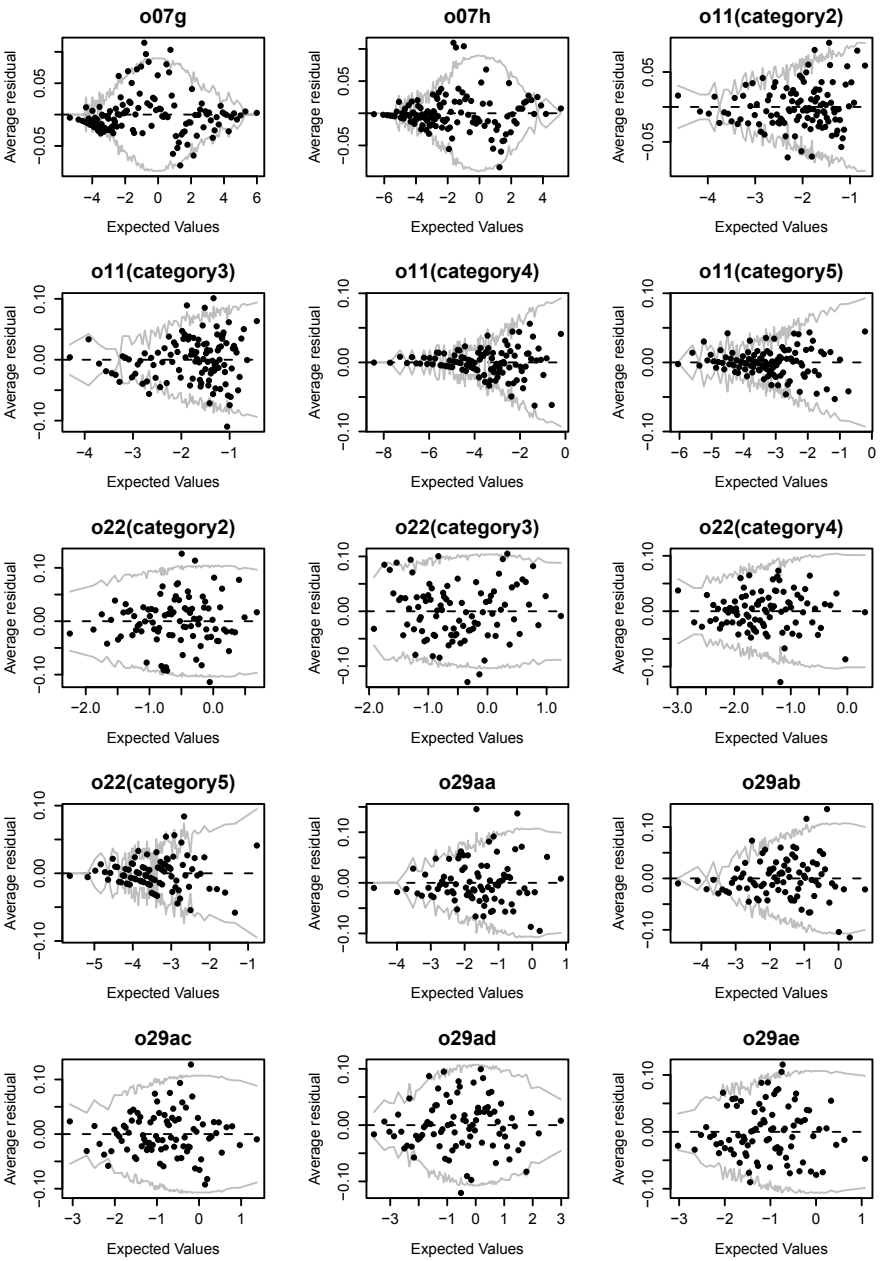

Fig. B.1 Continued.

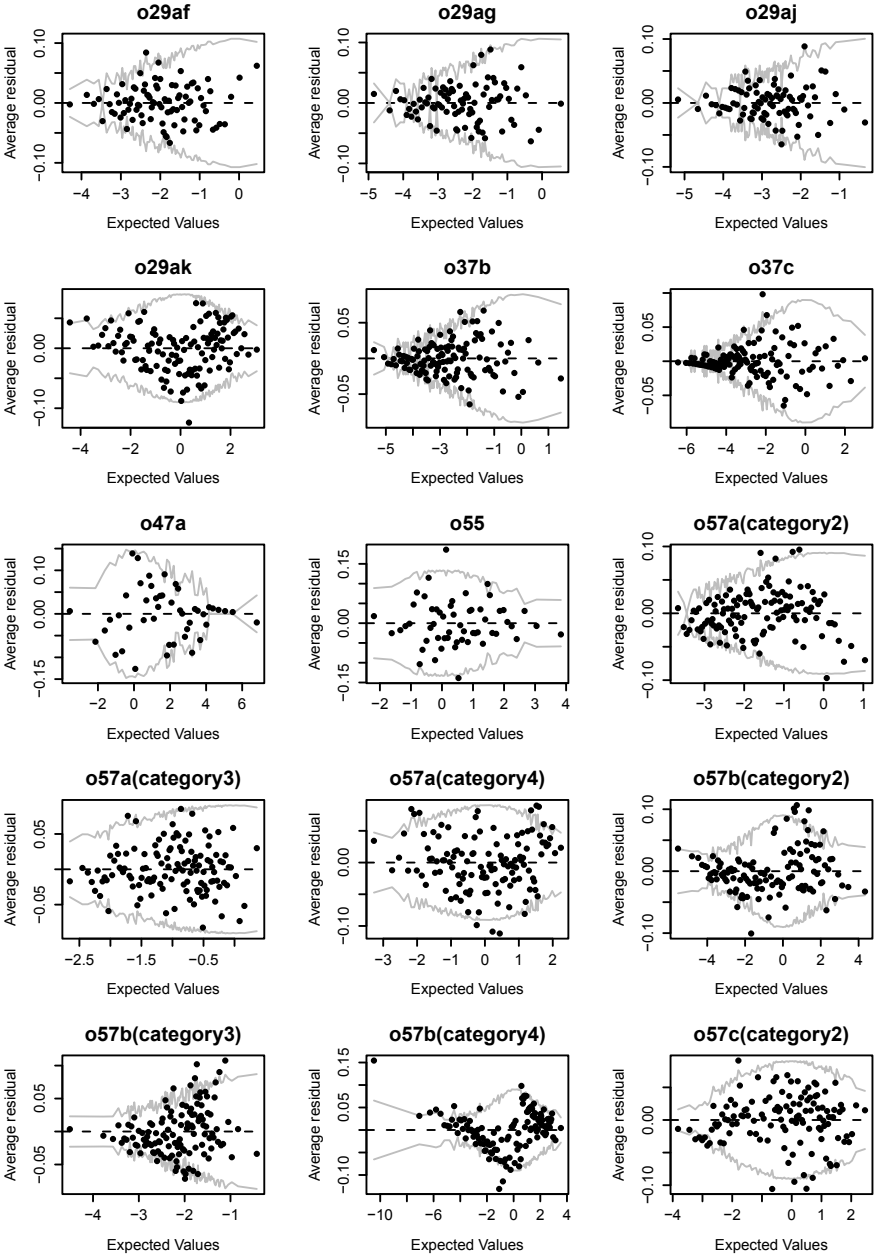

Fig. B.1 Continued.

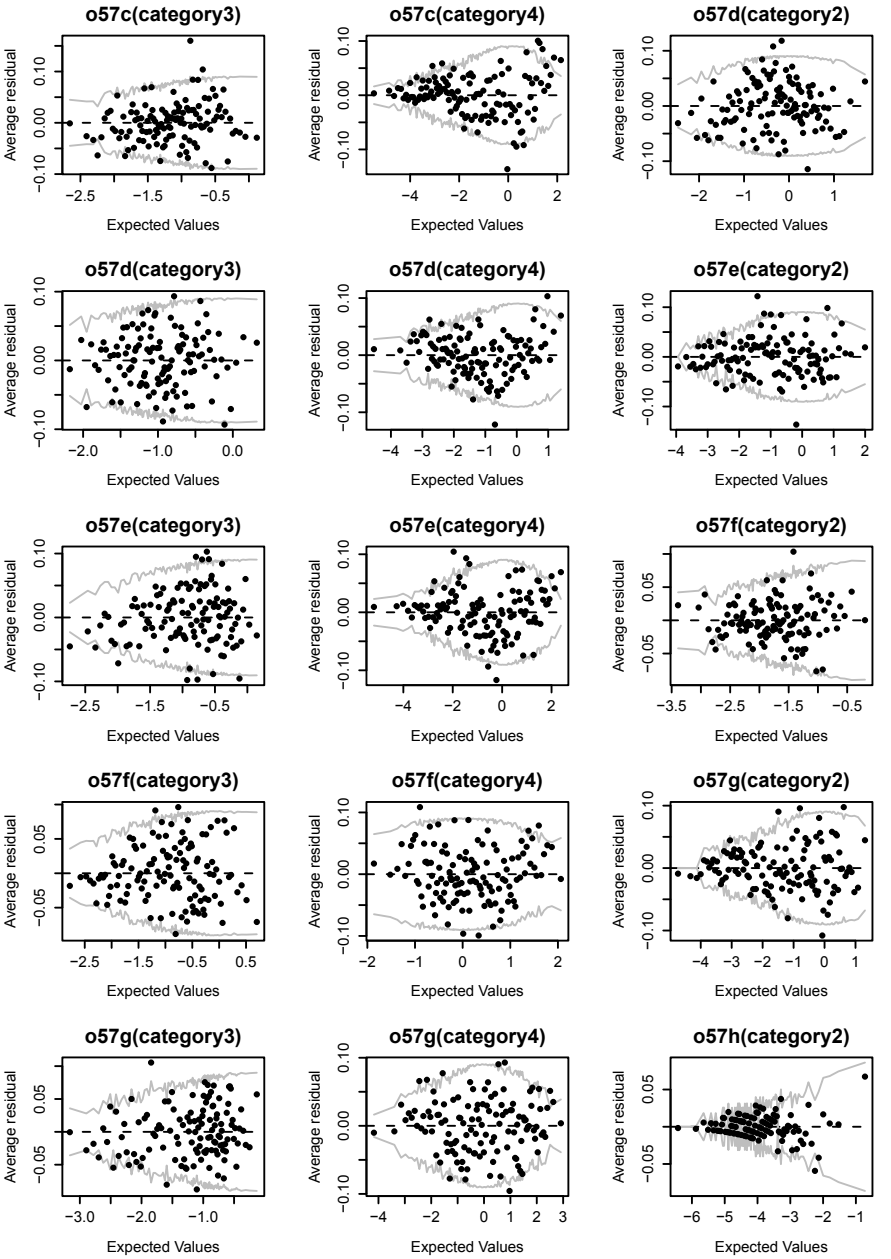

Fig. B.1 Continued.

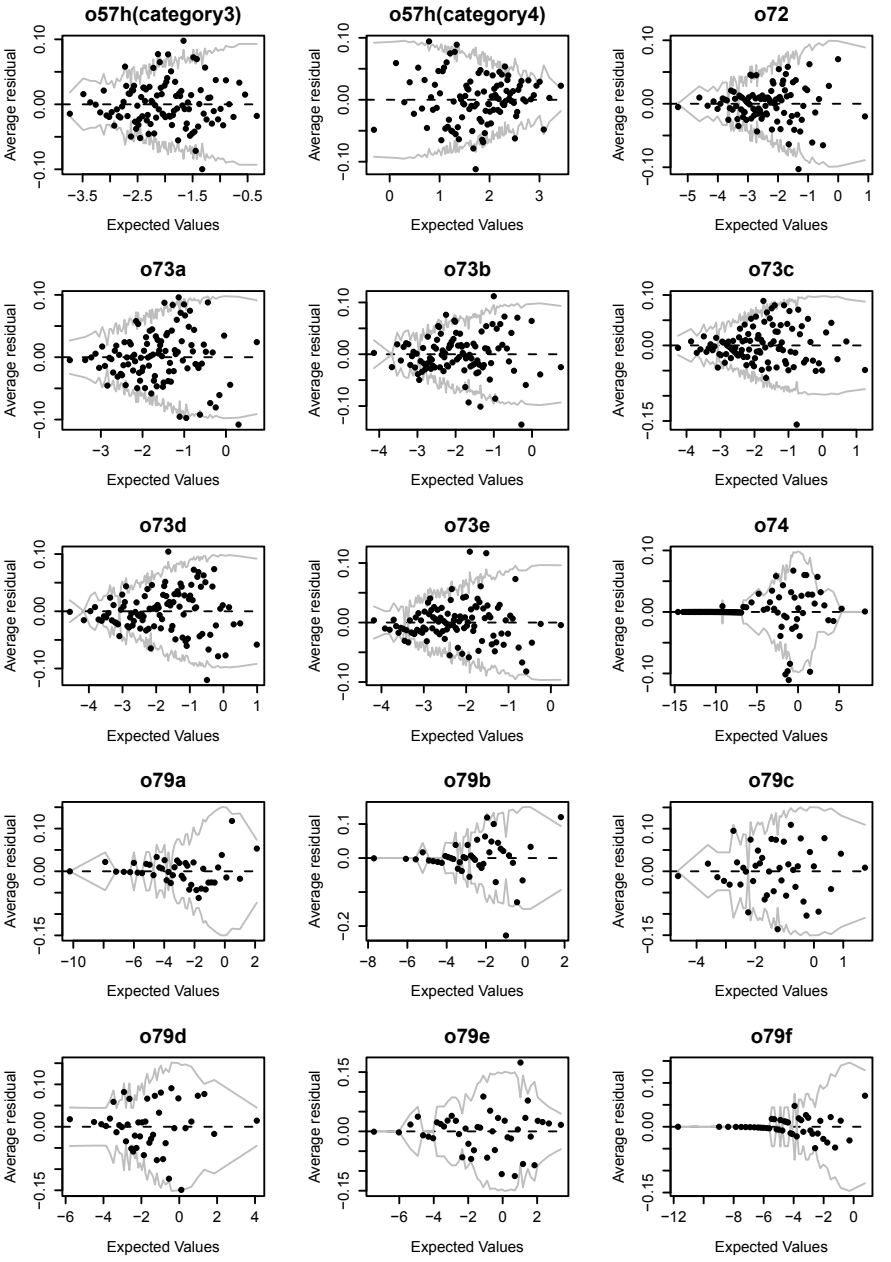

Fig. B.1 Continued.

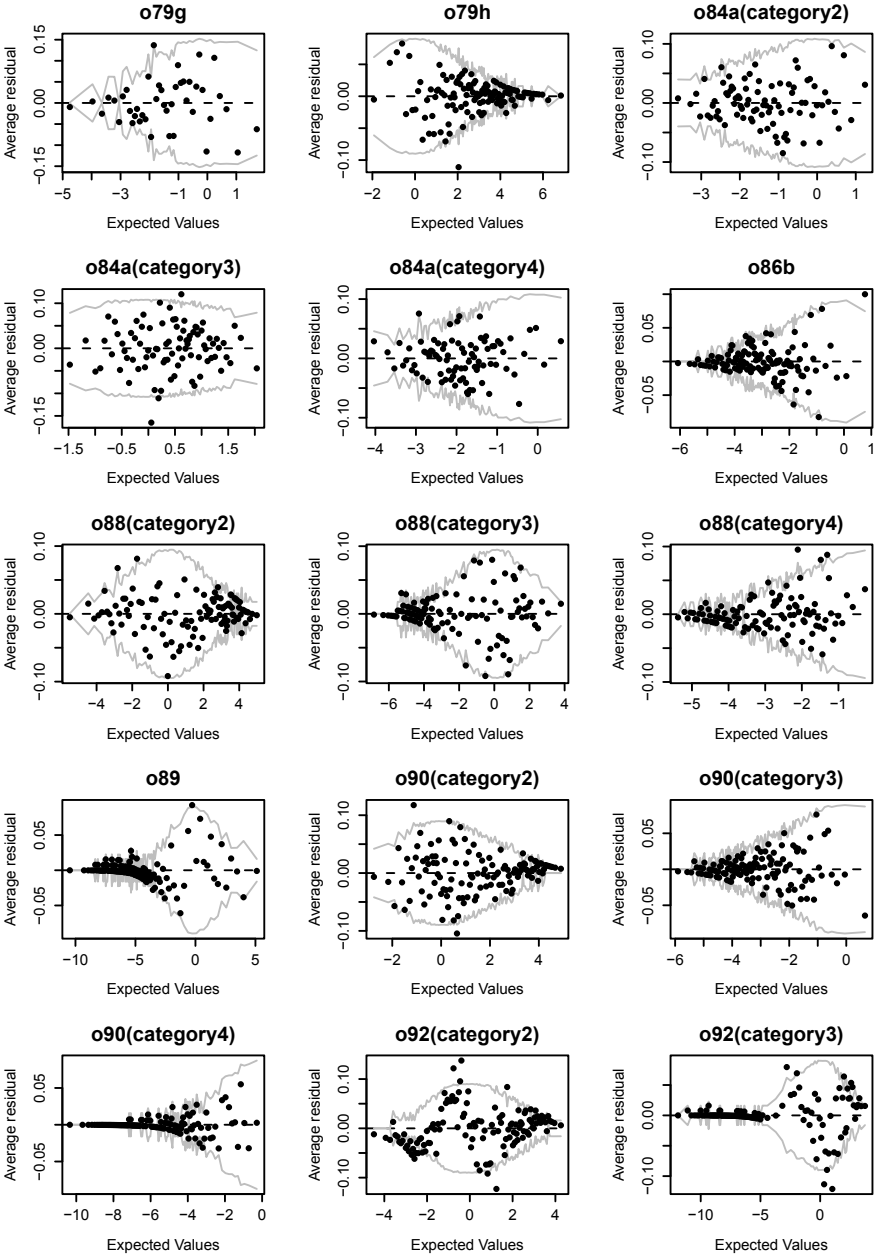

Fig. B.1 Continued.

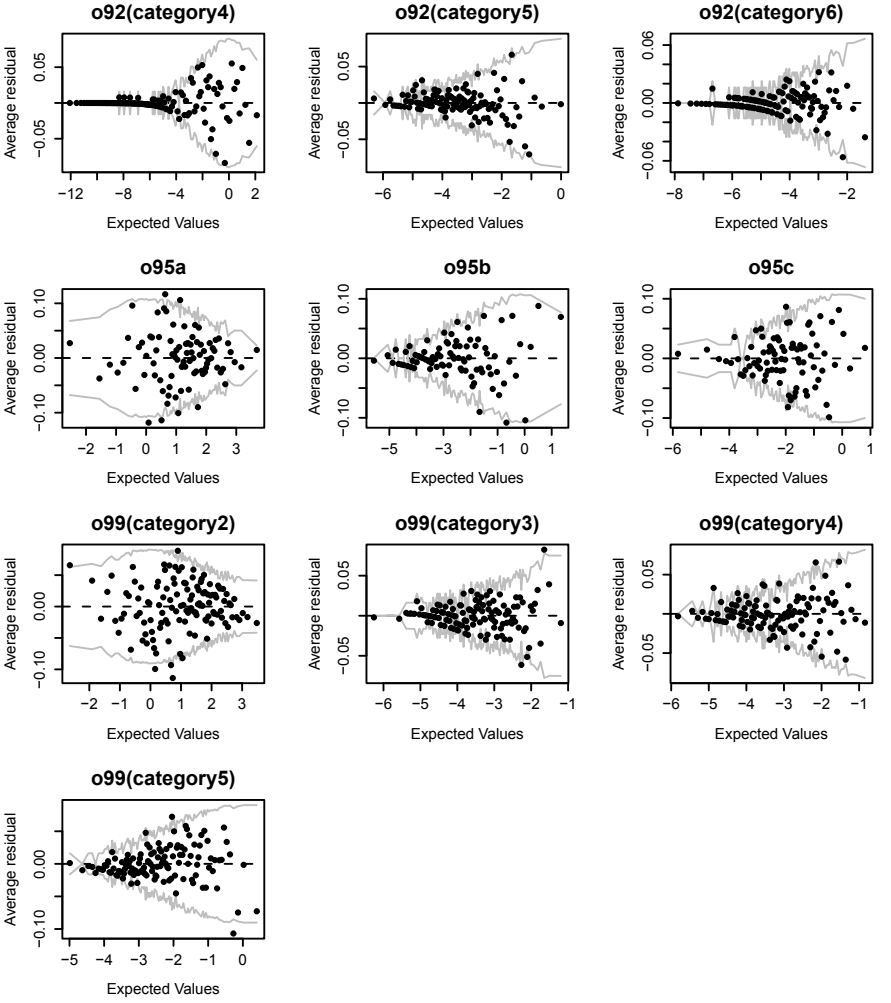

Fig. B.1 Continued.

Appendix C
Simulation Study for the Variance-inflated Imputation Model

Here I present results from a small simulation study that I conducted to evaluate the impact on data quality for the variance-inflated imputation model described in Section 8.3.5.4. For the simulation, I generate a population of $N = 1{,}000{,}000$ records comprising three variables, $Y_1, \ldots, Y_3$, drawn from $N(0, \Sigma)$, where Σ has variances equal to one and correlations ranging from 0.3 to 0.7. From this population, I repeatedly draw simple random samples of size $s = 10{,}000$ and treat these samples as the originally observed samples D_{obs}. For the synthesis, I replace values of Y_3 for all records in D_{obs}. I generate replacement values by sampling from the posterior predictive distribution, $f(Y_3|D_{obs})$, using parameter values drawn from the variance-inflated posterior distribution given in (8.11) with different levels of the variance inflation factor α. For comparison, I also generate synthetic datasets with Y_1 omitted from the imputation model to illustrate the negative consequences of dropping explanatory variables from the models to obtain a higher level of data protection. In analogy with the real data application, I generate $m = 5$ synthetic datasets for any one iteration of the simulation design. I obtain inferences for four quantities in each simulation run, including the population mean and the intercept and regression coefficients of Y_2 (β_1) and Y_3 (β_2) in a regression of Y_1 on Y_2 and Y_3. I repeat each simulation 5,000 times.

Table C.1 displays key results from the simulations. The average $\bar{q}_m$ across the 5,000 simulation runs is always very close to the average q_{obs} for $\alpha \leq 100$. For $\alpha = 1{,}000$, I find small biases for all point estimates. The variance estimator T_p (column four) correctly estimates the true variance of $\bar{q}_m$ (column three) for any given level of α. Columns six and seven summarize the percentages of the 5,000 synthetic 95% confidence intervals that cover their corresponding Q for the original sample and the synthetic samples, respectively. The coverage rates from the synthetic samples are always close to the expected nominal coverage of 95% for $\alpha \leq 100$. Only for $\alpha = 100$ is there a slight undercoverage for the regression coefficient β_2 compared with the coverage rate of β_2 in the original sample. The undercoverage increases for $\alpha = 1{,}000$. All estimates slightly undercover, and for β_2 the coverage rate actually drops to 90.8%. The ninth column reports the ratio of the confidence interval length from the synthetic datasets over the confidence interval length from

Table C.1 Simulation results for the variance-inflated imputation model. The denominators of the confidence interval length ratios and the RMSE ratios are based on the point estimates from the sample without synthesis.

α	Q	q_{obs}	$\bar{q}_m$	$var(q_m)$	T_p	CI cov. org.	CI cov. syn.	CI length ratio	RMSE ratio
1	$mean(Y_3)$	$-3.27 * 10^{-3}$	$-3.41 * 10^{-3}$	$3.65 * 10^{-3}$	$3.58 * 10^{-3}$	94.92	95.04	1.11	1.10
	Intercept	$3.84 * 10^{-3}$	$3.93 * 10^{-3}$	$1.74 * 10^{-3}$	$1.76 * 10^{-3}$	95.54	95.26	1.10	1.10
	β_1	$9.93 * 10^{-2}$	$9.93 * 10^{-2}$	$6.50 * 10^{-5}$	$6.40 * 10^{-5}$	94.78	94.96	1.10	1.08
	β_2	$6.70 * 10^{-1}$	$6.70 * 10^{-1}$	$6.60 * 10^{-5}$	$6.50 * 10^{-5}$	95.26	94.96	1.11	1.11
10	$mean(Y_3)$	$-2.91 * 10^{-3}$	$-2.69 * 10^{-3}$	$6.23 * 10^{-3}$	$6.31 * 10^{-3}$	94.90	95.18	1.57	1.45
	Intercept	$3.79 * 10^{-3}$	$3.64 * 10^{-3}$	$2.97 * 10^{-3}$	$2.99 * 10^{-3}$	95.42	94.64	1.52	1.43
	β_1	$9.93 * 10^{-2}$	$9.94 * 10^{-2}$	$1.08 * 10^{-4}$	$1.05 * 10^{-4}$	94.78	94.30	1.48	1.39
	β_2	$6.70 * 10^{-1}$	$6.70 * 10^{-1}$	$6.80 * 10^{-5}$	$6.70 * 10^{-5}$	94.74	94.82	1.12	1.11
100	$mean(Y_3)$	$-1.18 * 10^{-03}$	$-1.20 * 10^{-03}$	$3.29 * 10^{-02}$	$3.34 * 10^{-02}$	94.32	94.58	4.19	3.29
	Intercept	$3.39 * 10^{-03}$	$3.41 * 10^{-03}$	$1.50 * 10^{-02}$	$1.50 * 10^{-02}$	95.08	94.44	3.94	3.17
	β_1	$9.93 * 10^{-02}$	$1.01 * 10^{-01}$	$4.87 * 10^{-04}$	$5.07 * 10^{-04}$	94.90	94.44	3.78	2.98
	β_2	$6.70 * 10^{-01}$	$6.65 * 10^{-01}$	$8.90 * 10^{-05}$	$9.10 * 10^{-05}$	95.58	94.52	1.33	1.45
1000	$mean(Y_3)$	$-8.98 * 10^{-04}$	$9.97 * 10^{-03}$	$3.11 * 10^{-01}$	$3.02 * 10^{-01}$	95.00	94.40	13.29	10.22
	Intercept	$3.14 * 10^{-03}$	$-3.20 * 10^{-03}$	$1.23 * 10^{-01}$	$1.20 * 10^{-01}$	95.18	94.62	11.78	9.21
	β_1	$9.93 * 10^{-02}$	$1.14 * 10^{-01}$	$4.16 * 10^{-03}$	$4.20 * 10^{-03}$	95.00	94.08	11.53	8.92
	β_2	$6.70 * 10^{-01}$	$6.24 * 10^{-01}$	$1.66 * 10^{-03}$	$1.68 * 10^{-03}$	95.06	90.82	5.65	8.35

the original samples. Not surprisingly, the ratio increases with increasing α since the variance-inflated imputation model increases the between-imputation variance b_m and thus the variance of $\bar{q}_m$. Comparing the confidence interval length ratio with the root mean squared error (RMSE) ratio in the last column, it is obvious that the RMSE ratio is always smaller than or equal to the confidence interval length ratio, indicating that the increased RMSE in the synthetic datasets is likely due to the increased variance from the variance-inflated imputation model. Only for the regression coefficient β_2 and $\alpha \geq 100$ is there an increased RMSE ratio compared with the confidence interval length ratio. Overall levels of $\alpha \leq 100$ lead to reduced efficiency in the estimation but no noticeable bias, at least for this simulation. For $\alpha = 1{,}000$, there is a small bias that leads to slight undercoverage, but note that I replaced all records with variance-inflated imputations in these simulations. In practice, agencies will only replace some records that are specifically at risk with draws from the variance-inflated imputation model. I expect that the bias will be small in this context.

The results for the data generation that drops Y_1 from the imputation model to obtain a higher level of data protection are presented in Table C.2. $\bar{Y}_3$ and the intercept from the regression are not affected, but the two regression coefficients are completely biased, leading to a 0% coverage rate for both estimates and a substantially increased RMSE ratio. It is obvious that the variance-inflated imputation model provides far better results in terms of data validity. Dropping variables from the imputation models should only be considered an option if the data-disseminating agency knows that the data user will never evaluate the relationship between the dropped variable and the variable to be imputed.

Table C.2 Simulation results if Y_1 is excluded from the imputation model. The denominators of the confidence interval length ratios and the RMSE ratios are based on the point estimates from the sample without synthesis.

Q	q_{obs}	$\bar{q}_m$	$var(q_m)$	T_p	CI cov. org.	CI cov. syn.	CI length ratio	RMSE ratio
$mean(Y_3)$	$-2.07*10^{-3}$	$-1.87*10^{-3}$	$4.12*10^{-3}$	$4.07*10^{-3}$	95.40	94.92	1.20	1.18
Intercept	$3.86*10^{-3}$	$2.51*10^{-3}$	$2.68*10^{-3}$	$2.70*10^{-3}$	94.88	95.08	1.35	1.34
β_1	$9.93*10^{-2}$	$3.00*10^{-1}$	$8.90*10^{-5}$	$1.01*10^{-4}$	95.10	0.00	1.36	27.15
β_2	$6.70*10^{-1}$	$-7.60*10^{-5}$	$2.00*10^{-5}$	$1.19*10^{-4}$	94.72	0.00	1.49	90.70

References

Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society, Series C* **57**, 273–291.

Abowd, J. and Vilhuber, L. (2008). How protective are synthetic data? In J. Domingo-Ferrer and Y. Saygin, eds., *Privacy in Statistical Databases*, 239–246. New York: Springer.

Abowd, J. M. and Lane, J. I. (2004). New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, 282–289. New York: Springer.

Abowd, J. M., Stinson, M., and Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS public use file project. Technical report, Longitudinal Employer–Household Dynamics Program, U.S. Bureau of the Census, Washington, DC.

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.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, 290–297. New York: Springer.

An, D. and Little, R. J. A. (2007). Multiple imputation: An alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A* **170**, 923–940.

Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple-imputation. *Biometrika* **86**, 948–955.

Burgette, L. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology* **170**, 1070–1076.

Caiola, G. and Reiter, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy* **3**, 27–42.

Center of Excellence for Statistical Disclosure Control (2009). *Handbook on Statistical Disclosure Control.* Available at `http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf`.

Charest, A.-S. (2010). How can we analyze differentially-private synthetic datasets. *Journal of Privacy and Confidentiality* **2**, 21–33.

Domingo-Ferrer, J., Drechsler, J., and Polettini, S. (2009). Report on synthetic data files. Technical report, Eurostat, The Hague.

Drechsler, J. (2011a). Multiple imputation in practice – a case study using a complex German establishment survey. *Advances in Statistical Analysis* 1–26.

Drechsler, J. (2011b). New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey. *Journal of Applied Statistics* (online first).

Drechsler, J., Bender, S., and Rässler, S. (2008a). Comparing fully and partially synthetic data sets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy* **1**, 105–130.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008b). A new approach for disclosure control in the IAB Establishment Panel – multiple imputation for a better data access. *Advances in Statistical Analysis* **92**, 439–458.

Drechsler, J. and Rässler, S. (2008). Does convergence really matter? In Shalabh and C. Heumann, eds., *Recent Advances in Linear Models and Related Areas*, 341–355. Heidelberg: Physica.

Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin, eds., *Privacy in Statistical Databases*, 227–238. New York: Springer.

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

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association* **105**, 1347–1357.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics* **7**, 207–217.

Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, eds., *ICALP 2006*, 1–12. New York: Springer.

Elamir, E. and Skinner, C. J. (2006). Record level measures of disclosure risk for survey microdata. *Journal of Official Statistics* **22**, 525–539.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical report, Department of Statistics, Carnegie-Mellon University.

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics* **14**, 361–372.

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *Journal of Official Statistics* **13**, 75–89.

Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics* **14**, 485–502.

Fischer, G., Janik, F., Müller, D., and Schmucker, A. (2008). The IAB Establishment Panel – from sample to survey to projection. Technical report, FDZ-Methodenreport, No. 1, Institute for Employment Research, Nuremberg.

Franconi, L. and Stander, J. (2002). A model based method for disclosure limitation of business microdata. *The Statistician* **51**, 1–11.

Franconi, L. and Stander, J. (2003). Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistics and Computing* **13**, 295–305.

Galati, J. C. and Carlin, J. B. (2009). Inorm: Stata module to perform multiple imputation using Schafer's method. Statistical Software Components, Department of Economics, Boston College.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. London: Chapman and Hall, 2nd edn.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambrdige: Cambridge University Press.

Graham, P. and Penny, R. (2005). Multiply imputed synthetic data files. Technical report, University of Otago. `http://www.uoc.otago.ac.nz/departments/pubhealth/pgrahpub.htm`.

Graham, P., Young, J., and Penny, R. (2009). Multiply imputed synthetic data: Evaluation of hierarchical Bayesian imputation models. *Journal of Official Statistics* **25**, 407–426.

Honaker, J., King, G., and Blackwell, M. (2010). *AMELIA II: A Program for Missing Data* Available at `http://gking.harvard.edu/amelia`.

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

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

Kinney, S. K. and Reiter, J. P. (2010). Tests of multivariate hypotheses when using multiple imputation for missing data and disclosure limitation. *Journal of Official Statistics* **26**, 301–315.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. Technical report, Center for Economic Studies (CES), CES-WP-11-04.

Knoche, P. (1993). Factual anonymity of microdata from household and person-related surveys – the release of microdata files for scientific purposes. In *Proceedings of the International Symposium on Statistical Confidentiality*, 407–413. Eurostat, Dublin.

Kölling, A. (2000). The IAB-Establishment Panel. *Journal of Applied Social Science Studies* **120**, 291–300.

Lane, J. I. (2007). Optimizing the use of microdata: An overview of the issues. *Journal of Official Statistics* **23**, 299–317.

Lechner, S. and Pohlmeier, W. (2005). Data masking by noise addition and the estimation of nonparametric regression models. *Journal of Economics and Statistics* **225**, 517–528.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991). Large sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution. *Journal of the American Statistical Association* **86**, 1065–1073.

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

Little, R. J. A., Liu, F., and Raghunathan, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In A. Gelman and X. L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 141–152. New York: John Wiley and Sons.

Little, R. J. A. and Raghunathan, T. E. (1997). Should imputation of missing data condition on all observed variables? In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 617–622. Alexandria, VA: American Statistical Association.

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

Liu, F. and Little, R. J. A. (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In *ASA Proceedings of the Joint Statistical Meetings*, 2133–2138.

Machanavajjhala, A., Kifer, D., Abowd, J. M., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering*, 277–286.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science* **9**, 538–558.

Meng, X. L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103–111.

Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In J. Domingo-Ferrer and L. Franconi, eds., *Privacy in Statistical Databases*, 177–188. New York: Springer-Verlag.

Polettini, S. (2003). Maximum entropy simulation for microdata protection. *Statistics and Computing* **13**, 308–320.

Polettini, S., Franconi, L., and Stander, J. (2002). Model-based disclosure protection. In J. Domingo-Ferrer, ed., *Inference Control in Statistical Databases*, 83–96. Berlin: Springer-Verlag.

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

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

Raghunathan, T. E., Solenberger, P., and van Hoewyk, J. (2002). IVEware: Imputation and variance estimation software. Available at: `http://www.isr.umich.edu/src/smp/ive/`.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–189.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Reiter, J. P. (2005a). Estimating identification risks in microdata. *Journal of the American Statistical Association* **100**, 1103–1113.

Reiter, J. P. (2005b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

Reiter, J. P. (2005c). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365–377.

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

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika* **94**, 502–508.

Reiter, J. P. (2008a). Letter to the editor. *Journal of Official Statistics* **24**, 319–321.

Reiter, J. P. (2008b). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. *Statistics and Probability Letters* **78**, 15–20.

Reiter, J. P. and Drechsler, J. (2010). Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. *Statistica Sinica* **20**, 405–421.

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

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

Reiter, J. P., Raghunathan, T. E., and Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology* **32**, 143–150.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Royston, P. (2005). Multiple imputation of missing values: Update of ice. *The Stata Journal* **5**, 527–536.

Royston, P. (2007). Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring. *The Stata Journal* **7**, 445–464.

Royston, P. (2009). Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables. *The Stata Journal* **9**, 466–477.

Rubin, D. B. (1978). Multiple imputations in sample surveys. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 20–34. Alexandria, VA: American Statistical Association.

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 and Sons.

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

Rubin, D. B. (2004). The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician* **58**, 298–302.

Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**, 366–374.

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

Schenker, N., Raghunathan, T. E., Chiu, P. L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association* **101**, 924–933.

StataCorp (2009). *Stata 11 Multiple-Imputation Reference Manual*. College Station, TX: StataCorp.

Statistisches Bundesamt (2005). *Statistik und Wissenschaft Band 4: Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten*. Wiesbaden: Statistisches Bundesamt.

Su, Y., Gelman, A., Hill, J., and Yajima, M. (2009). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software* (forthcoming).

van Buuren, S. and Groothuis-Oudshoorn, K. (2010). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* (forthcoming).

van Buuren, S. and Oudshoorn, C. (2000). Mice v1.0 user's manual. report pg/vgz/00.038. Technical report, TNO Prevention and Health, Leiden.

Winkler, W. E. (2007a). Analytically valid discrete microdata files and reidentification. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Winkler, W. E. (2007b). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

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

Woodcock, S. D. and Benedetto, G. (2009). Distribution-preserving statistical disclosure limitation. *Computational Statistics and Data Analysis* **53**, 4228–4242.

Zwick, T. (2005). Continuing vocational training forms and establishment productivity in Germany. *German Economic Review* **6**, 155–184.

Index