Preface

Introduction

An increasing amount of data on persons and establishments are collected by statistical organizations. Also, the demand for microdata for researchers is increasing since economic or empirical analysis, and to make statements about our society on empirical basis is often only possible when investigating in data with detailed information. Moreover, a considerable increase in the production of socioeconomic data and their accessibility by researchers have been observed in recent years. Statistical agencies are making more of their survey and census microdata available, government agencies are publishing more of their administrative data, and the private sector has become a major provider of big data.

This, however, comes with a variety of legal, ethical, and technical challenges. Privacy protection principles and regulations impose restrictions on access and use of individual data. Statistical producers are faced with the challenge of ensuring respondents' confidentiality when making microdata files accessible. Not only data producers are obligated to protect confidentiality, but also confidentiality is crucial for maintaining the trust of respondents and ensuring the honesty and validity of their responses. Statistical disclosure control (SDC) is thus an emerging field of research. Proper and secure microdata dissemination requires statistical agencies to establish policies and procedures that formally define the conditions for accessing microdata (Dupriez and Boyko 2010) and to apply statistical disclosure control (SDC) methods to data before release.

Also, the demand for complex microdata for training purposes for students seems to increase. Research projects under the 5th, 6th, or 7th framework program for research of the European Union generated confidential microdata for public or scientific use. Public-use files are also generated for researchers to be able to run complex simulation studies to compare methods. The *BLUE-ETS*, *AMELI*, *DACSEIS*, and *EUREDIT* research projects of this framework program are examples using anonymized data for simulation tasks. Outside Europe, there is a long tradition to anonymize data sets. For example, the US CENSUS Bureau provides

public-use data since the 1970s, and current releases of public-use data set include the current population survey, the survey of income and program participation, the American housing survey, the survey of program dynamics, the American community survey, and the consumer expenditure survey [see, e.g., Task Force on the Conference of European Statisticians 2007]. In addition, almost every statistical agency in the world and institutions holding confidential data, e.g., on health statistics, are at least interested in providing public- or scientific-use files of high quality and low disclosure risk. In any case, due to national laws on privacy, microdata cannot be distributed to the public or to researchers whenever re-identification of persons or establishments is possible.

This book gives a detailed view on well-known SDC methods for data, complex sample surveys, censuses, and administrative sources. It discusses the traditional approach of data anonymization by perturbation of the data, the disclosure risk, and data utility of anonymized data sets. It also includes methods to simulate synthetic data.

This is not a book about the software environment `R` and related packages for statistical disclosure control. The aim of the book was to explain the theory of statistical disclosure control methods. However, in order to better understand the theory, a lot of practical examples are given. Almost every example can be reproduced by the readers using `R` and the `R` packages **sdcMicro** and **simPop**. The code for each example is included in the book and provided as supplementary files at

http://www.statistik.tuwien.ac.at/public/templ/sdcbook

The free and open-source software provided with this book allows the reader to also apply the presented methods on their own data sets provided by the mentioned software.

Overview of the Book

This book is intended for statistical producers at National Statistical Offices (NSOs), data holders, and international and national agencies, as well as data users who are interested in the subject. It does not require no prior knowledge of statistical disclosure control (SDC). This book is focused on SDC methods for microdata. It does not cover SDC methods for protecting tabular outputs [see Castro 2010, for more details].

Chapter 1 includes a very brief introduction to R. It is intended to make further `R` code in the book understandable, but it does not replace a conventional introduction to R. If readers want to learn `R` in a broader context, we refer to the official manuals on `R` on the CRAN Webpage or to further contributed documentation on CRAN or books about R. After this short introduction, SDC tools and the differences between each of them are discussed. This section closes with a general discussion about the `R` package **sdcMicro**. The **sdcMicro** and the **simPop** packages are applied in various other chapters, especially in the case studies of Chap. 8.

The methods' parts start with an introduction to the basic concepts regarding statistical disclosure in Chap. 2. These concepts include the definition of terms used in the area of SDC and briefly present different kinds of disclosure scenarios. Chapter 2 finishes with risk-utility maps that show the trade-off between disclosure risk and data utility.

Chapter 3 includes methods for measuring disclosure risk. The aim is not only to discuss one concept of measuring/estimating disclosure risk, but also to describe several concepts. The chapter starts with basic methods such as *k*-anonymity and *l*-diversity. To measure risk on subsets of key variables, the SUDA method is explained. A large part of this chapter is dedicated to measuring the individual risk and the global risk with log-linear models.

In Chap. 4, the most common SDC methods are presented. The chapter is basically divided into two parts: one for categorical data and the other for continuous scaled data. Not only popular, but also new methods, such as shuffling, are discussed.

An introduction to common approaches for assessing information loss and data utility is given in Chap. 5. This chapter includes more than just typical comparisons, such as comparing means and covariances, since these methods—even regularly applied—are often too general and not suitable for specific data sets. Therefore, further concepts based on estimating certain indicators are presented as well.

Even if this book is mainly focused on traditional methods for SDC, Chap. 6 is dedicated to the simulation of synthetic data sets. A lot of different concepts exist in literature, but we focus on only one particular approach. We point out some advantages of the approach of model-based simulation of synthetic population data and describe the framework of this concept.

Chapter 7 provides practical guidelines on how to implement SDC on data sets in practice. Basically, the work flow for anonymizing data is shown. All the learned lessons are included in this workflow.

This workflow is also applied in Chap. 8 on several very popular data sets such as the Structure of Earnings Statistics (SES), the European Statistics on Income and Living Conditions (EU-SILC), the Family Income and Expenditure Survey (FIES), the International Income Distribution Database (I2D2), the Purchase Power Parity Data set (P4), and the Survey-based Harmonized Indicators Data (SHIP) that are also harmonized data files from household surveys.

Note that almost all methods introduced in this book can be implemented using **sdcMicroGUI**, an `R` -based, user-friendly, point-and-click graphical user interface (Kowarik et al. 2013). However, all methods are shown in applications using the `R`-package **sdcMicro** (Templ et al. 2015) that differs in that sense that working with the package is command line only and more methods are included than in **sdcMicroGUI** . Note that a successor, a new version of **sdcMicroGUI** that works in a browser and is included in **sdcMicro**, version > 4.7. The app can be opened via the function call `sdcApp()` in R. For simulating synthetic data, the package

simPop is used. Readers are encouraged to explore the implementation of methods using this book along with the detailed user manuals of **sdcMicroGUI** (Kowarik et al. 2013), **sdcMicro** (Templ et al. 2015), and **simPop** (Templ et al. 2017).

Winterthur, Switzerland Matthias Templ
March 2017

References

Castro, J. (2010). Statistical disclosure control in tabular data. In J. Nin & J. Herranz (Eds.), *Privacy and anonymity in information management systems, Advanced information and knowledge processing* (pp. 113–131). London: Springer.

Dupriez, O., & Boyko, E. (2010). Dissemination of microdata files. *Formulating policies and procedures*. IHSN Working Paper No 005, International Household Survey Network, Paris.

Kowarik, A., Templ, M., Meindl, B., & Fonteneau, F. (2013). *sdcMicroGUI: Graphical user interface for package sdcMicro*. URL:http://CRAN.R-project.org/package=sdcMicroGUI. *R* package version 1.1.1.

Task Force on the Conference of European Statisticians. (2007). Managing statistical confidentiality & microdata access. *Principles and guidelines of good practice*. Technical report, United Nations Economic Commission for Europe, New York and Geneva.

Templ, M., Meindl, B., & Kowarik, A. (2015). Statistical disclosure control for micro-data using the *R* package sdcMicro. *Journal of Statistical Software*, *67*(1), 1–37.

Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). Simulation of synthetic complex data: The R-package simPop. *Journal of Statistical Software,* 1–38. Accepted for publication in December 2015.

http://www.springer.com/978-3-319-50270-0

Statistical Disclosure Control for Microdata
Methods and Applications in R
Templ, M.
2017, XIX, 287 p. 37 illus., 27 illus. in color., Hardcover
ISBN: 978-3-319-50270-0