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