Preface

Readers will find this book a mixture of practical advice, mathematical rigor, management insight, and philosophy. Our intended audience is the working analyst. Our approach is to work by real life examples. Most illustrations come out of our successful practice. A few are contrived to make a point. Sometimes they come out of failed experience, ours and others.

We have written this book to help the reader gain a deeper understanding, at an applied level, of the issues involved in improving data quality through editing, imputation, and record linkage. We hope that the bulk of the material is easily accessible to most readers although some of it does require a background in statistics equivalent to a 1-year course in mathematical statistics. Readers who are less comfortable with statistical methods might want to omit Section 8.5, Chapter 9, and Section 18.6 on first reading. In addition, Chapter 7 may be primarily of interest to those whose professional focus is on sample surveys. We provide a long list of references at the end of the book so that those wishing to delve more deeply into the subjects discussed here can do so.

Basic editing techniques are discussed in Chapter 5, with more advanced editing and imputation techniques being the topic of Chapter 7. Chapter 14 illustrates some of the basic techniques. Chapter 8 is the essence of our material on record linkage. In Chapter 9, we describe computational techniques for implementing the models of Chapter 8. Chapters 9–13 contain techniques that may enhance the record linkage process. In Chapters 15–17, we describe a wide variety of applications of record linkage. Chapter 18 is our chapter on data confidentiality, while Chapter 19 is concerned with record linkage software. Chapter 20 is our summary chapter.

Three recent books on data quality – Redman [1996], English [1999], and Loshin [2001] – are particularly useful in effectively dealing with many management issues associated with the use of data and provide an instructive overview of the costs of some of the errors that occur in representative databases. Using as their starting point the work of quality pioneers such as Deming, Ishakawa, and Juran whose original focus was on manufacturing processes, the recent books cover two important topics not discussed by those seminal authors: (1) errors that affect data quality even when the underlying processes are operating properly and (2) processes that are controlled by others (e.g., other organizational units within one's company or other companies).

Dasu and Johnson [2003] provide an overview of some statistical summaries and other conditions that must exist for a database to be useable for

specific statistical purposes. They also summarize some methods from the database literature that can be used to preserve the integrity and quality of a database. Two other interesting books on data quality – Huang, Wang and Lee [1999] and Wang, Ziad, and Lee [2001] – supplement our discussion. Readers will find further useful references in The International Monetary Fund's (IMF) Data Quality Reference Site on the Internet at http://dsbb.imf.org/Applications/web/dqrs/dqrshome/.

We realize that organizations attempting to improve the quality of the data within their key databases do best when the top management of the organization is leading the way and is totally committed to such efforts. This is discussed in many books on management. See, for example, Deming [1986], Juran and Godfrey [1999], or Redman [1996]. Nevertheless, even in organizations not committed to making major advances, analysts can still use the tools described here to make substantial quality improvement.

A working title of this book – *Playing with Matches* – was meant to warn readers of the danger of data handling techniques such as editing, imputation, and record linkage unless they are tightly controlled, measurable, and as transparent as possible. Over-editing typically occurs unless there is a way to measure the costs and benefits of additional editing; imputation always adds uncertainty; and errors resulting from the record linkage process, however small, need to be taken into account during future uses of the data.

We would like to thank the following people for their support and encouragement in writing this text: Martha Aliaga, Patrick Ball, Max Brandstetter, Linda Del Bene, William Dollarhide, Mary Goulet, Barry I. Graubard, Nancy J. Kirkendall, Susan Lehmann, Sam Phillips, Stephanie A. Smith, Steven Sullivan, and Gerald I. Webber.

We would especially like to thank the following people for their support and encouragement as well as for writing various parts of the text: Patrick Baier, Charles D. Day, William J. Eilerman, Bertram M. Kestenbaum, Michael D. Larsen, Kevin J. Pledge, Scott Schumacher, and Felicity Skidmore.

http://www.springer.com/978-0-387-69502-0

Data Quality and Record Linkage Techniques

Herzog, Th.N.; Scheuren, F.J.; Winkler, W.E.

2007, XIV, 234 p., Softcover
ISBN: 978-0-387-69502-0