STATISTICAL POLICY
WORKING PAPER 22 (Second version, 2005)

Report on Statistical Disclosure Limitation Methodology

Federal Committee on Statistical Methodology

Originally Prepared by Subcommittee on Disclosure Limitation Methodology 1994

Revised by Confidentiality and Data Access Committee 2005

Statistical and Science Policy
Office of Information and Regulatory Affairs
Office of Management and Budget

December 2005

The Federal Committee on Statistical Methodology
(December 2005)

Members

Brian A. Harris-Kojetin, Chair, Office of Management and Budget

Wendy L. Alvey, Secretary, U.S. Census Bureau

Lynda Carlson, National Science Foundation

Steven B. Cohen, Agency for Healthcare Research and Quality

Steve H. Cohen, Bureau of Labor Statistics

Lawrence H. Cox, National Center for Health Statistics

Robert E. Fay, U.S. Census Bureau

Ronald Fecso, National Science Foundation

Dennis Fixler, Bureau of Economic Analysis

Gerald Gates, U.S. Census Bureau

Barry Graubard, National Cancer Institute

William Iwig, National Agricultural Statistics Service

Arthur Kennickell, Federal Reserve Board

Nancy J. Kirkendall, Energy Information Administration

Susan Schechter, Office of Management and Budget

Rolf R. Schmitt, Federal Highway Administration

Marilyn Seastrom, National Center for Education Statistics

Monroe G. Sirken, National Center for Health Statistics

Nancy L. Spruill, Department of Defense

Clyde Tucker, Bureau of Labor Statistics

Alan R. Tupek, U.S. Census Bureau

G. David Williamson, Centers for Disease Control and Prevention

Expert Consultant

Robert Groves, University of Michigan and Joint Program in Survey Methodology

Preface

The Federal Committee on Statistical Methodology (FCSM) was organized by the Office of Management and Budget (OMB) in 1975 to investigate issues of data quality affecting Federal statistics. Members of the committee, selected by OMB on the basis of their individual expertise and interest in statistical methods, serve in a personal capacity rather than as agency representatives. The committee conducts its work through subcommittees that are organized to study particular issues. Statistical Policy Working Papers are prepared by the subcommittee members and are reviewed and approved by FCSM members.

The Confidentiality and Data Access Committee (CDAC) is a special interest subcommittee of the FCSM that was formed in 1995 as a result of recommendations contained in the original Statistical Policy Working Paper 22. The committee consists primarily of statisticians working in federal agencies who are involved with issues relating to protecting data confidentiality, and providing selective and controlled access to confidential data. CDAC provides a unique forum for discussing these issues and sharing information and research ideas among the federal agencies. CDAC's website may be accessed at http://www.fcsm.gov/committees/cdac. The 2005 revision to Statistical Policy Working Paper 22 is the second version of the 1994 work by the Subcommittee on Disclosure Limitation and Methodology. The Subcommittee on Disclosure Limitation Methodology was formed in 1992 to describe and evaluate existing disclosure limitation methods for tabular and microdata files and to update previous work presented in Statistical Policy Working Paper 2, "Report on Statistical Disclosure and Disclosure-Avoidance Techniques" published in 1978. See Cover and Introductory Material in the 1994 version of Statistical Policy Working Paper 22 for a discussion of the Subcommittee on Disclosure Limitation Methodology.

The Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper 22, discusses both tables and microdata and describes current practices of the principal Federal statistical agencies. The original report includes a tutorial, guidelines, and recommendations for good practice; recommendations for further research; and an annotated bibliography. In 2004, the Confidentiality and Data Access Committee (CDAC) revised Statistical Policy Working Paper 22 to include research and new methodologies that were developed over the past ten years, and to reflect current agency practices. The annotated bibliography was partially updated. The CDAC members who worked on the revision:

Jacob Bournazian, Energy Information Administration
Nancy Kirkendall, Energy Information Administration
Steve Cohen, Bureau of Labor Statistics
Philip Steel, Bureau of Census
Alvan O. Zarate, National Center for Health Statistics
Arnold Reznek, Bureau of Census
Paul Massell, Bureau of Census

Acknowledgements

We thank the agency representatives of CDAC for their contributions to this working paper and updating the descriptions of agency practices in Chapter 3.

Table of Contents

CHAPTER I - Introduction 1
A. Subject and Purposes of This Report 1
B. Some Definitions 2
B.1. Confidentiality and Disclosure 2
B.2. Tables, Microdata, and On-Line Query Systems 4
B.3. Restricted Data and Restricted Access 4
C. Organization of the Report 5
D. Underlying Themes of the Report 6
CHAPTER II - Statistical Disclosure Limitation Methods: A Primer 8
A. Background 8
B. Definitions 9
B.1. Tables of Magnitude Data Versus Tables of Frequency Data 9
B.2. Table Dimensionality 9
B.3. Hierarchical Structure of Variables 10
B.4. What is Disclosure? 10
C. On-Line Query Systems 11
D. Tables of Counts or Frequencies 12
D.1. Sampling as a Statistical Disclosure Limitation Method 12
D.2. Defining Sensitive Cells 14
D.2.a Special Rules 14
D.2.b. The Threshold Rule 15
D.3. Protecting Sensitive Cells After Tabulation 16
D.3.a. Suppression 16
D.3.b. Random Rounding 18
D.3.c. Controlled Rounding 19
D.3.d. Controlled Tabular Adjustment 19
D.4. Protecting Sensitive Cells Before Tabulation 21
E. Tables of Magnitude Data 22
E.1. Defining Sensitive Cells – Linear Sensitivity Rules 22
E.2 Protecting Sensitive Cells After Tabulation 22
E.3. Protecting Sensitive Cells Before Tabulation 23
F. Microdata 24
F.1. Sampling, Removing Identifiers and Limiting Geographic Detail 25
F.2. High Risk Variables 25
F.2.a. Top-coding, Bottom-coding, Recoding into Intervals 26
F.2.b. Adding Random Noise 27
F.2.c. Data Swapping and Rank Swapping 28
F.2.d. Blank and Impute for Randomly Selected Records 32
F.2.e. Blurring 33
F.2.f. Targeted Suppression 33
G. Summary 33

CHAPTER III – Current Federal Statistical Agency Practices 34
A. Agency Summaries 34
A.1. Department of Agriculture 34
A.1.a. Economic Research Service (ERS) 34
A.1.b. National Agricultural Statistics Service (NASS) 35
A.2. Department of Commerce 37
A.2.a. Bureau of Economic Analysis (BEA) 37
A.2.b. Bureau of the Census (BOC) 38
A.3. Department of Education: National Center for Education Statistics (NCES) 40
A.4. Department of Energy: Energy Information Administration (EIA) 42
A.5. Department of Health and Human Services 44
A.5.a. Agency for Healthcare Research & Quality (AHRQ) 44
A.5.b. National Center for Health Statistics (NCHS) 45
A.6. Department of Justice: Bureau of Justice Statistics (BJS) 46
A.7. Department of Labor: Bureau of Labor Statistics (BLS) 46
A.8. Department of the Transportation: Bureau of Transportation Statistics (BTS) 48
A.9. Department of the Treasury: Internal Revenue Service, Statistics of Income Division (IRS, SOI) 49
A.10. National Science Foundation (NSF) 50
A.11. Social Security Administration (SSA) 51
B. Summary 52
B.1. Magnitude and Frequency Data 52
B.2. Microdata 52
CHAPTER IV – Methods for Tabular Data 57
A. Tables of Frequency Data 57
A.1. Controlled Rounding 58
B. Tables of Magnitude Data 59
B.1. Definition of Sensitive Cells – Linear Sensitivity Rules 60
B.1.a. The p-Percent Rule 61
B.1.b. The pq Rule 62
B.1.c. The (n, k) Rule 63
B.1.d. The Relationship Between (n, k) and p-Percent or pq Rules 64
B.1.e. Information in Parameter Values 65
B.2. Complementary Suppression 66
B.2.a. Audits of Proposed Complementary Suppressions 66
B.2.a.i. Implicitly Published Unions of Suppressed Cells Are Sensitive 66
B.2.a.ii. Row, Column and/or Layer Equations Can Be Solved for Suppressed Cells 67
B.2.a.iii. Software For Auditing A Suppression Pattern 67
B.2.b. Automatic Selection of Cells for Complementary Suppression 68
B.3. Controlled Tabular Adjustment 70
B.4. Adding Noise to Microdata Prior to Tabulating Data 71
C. Online Data Query Systems 73
D. Technical Notes: Relationships Between Common Linear Sensitivity Measures 74

CHAPTER V – Methods for Public-Use Microdata Files 81
A. Disclosure Risk of Microdata 81
A.1. Disclosure Risk and Intruders 82
A.2. Factors Contributing to Risk 82
A.3. Factors that Naturally Decrease Risk 83
A.4 Disclosure Risks Associated with Regression Models 84
B. Mathematical Methods of Addressing the Problem 85
B.1. Proposed Measures of Risk 86
B.1.a. MASSC. 87
B.1.b. R-U Confidentiality Map. 87
B.2. Methods of Reducing Risk by Reducing the Amount of Information Released 88
B.3. Methods of Reducing Risk by Disturbing Microdata 88
B.3.a. Data Swapping 89
B.3.b. Data Shuffling 89
B.3.c. Data Blurring and Microaggregation 91
B.3.d. Micro Agglomeration, Substitution, Subsampling, and Calibration (MASSC) 92
B.4. Methods of Reducing Risk by Using Simulated Microdata 92
B.4.a. Latin Hypercube Sampling. 92
B.4.b. Inference-Valid Synthetic Data. 92
B.4.c. The FRITZ Algorithm for Disclosure Limitation. 93
B.5. Methods of Analyzing Disturbed Microdata to Determine Usefulness 93
C. Necessary Procedures for Releasing Microdata Files 94
C.1. Removal of Identifiers 94
C.2. Limiting Geographic Detail 94
C.3. Top-Coding High Risk Variables That Are Continuous. 95
C.4. Precautions for Certain Types of Microdata 95
C.4.a. Establishment Microdata 96
C.4.b. Longitudinal Microdata 96
C.4.c. Microdata Containing Administrative Data 96
C.4.d. Consideration of Potentially Matchable Files and Population Uniques 96
D. Stringent Methods of Limiting Disclosure Risk 97
D.1. Do Not Release the Microdata 97
D.2. Recode Data to Eliminate Uniques 97
D.3. Disturb Data to Prevent Matching to External Files 98
E. Conclusion 98
CHAPTER VI – Recommended Practices For Federal Agencies 99
A. Introduction 99
B. Recommendations 99
B.1. General Recommendations for Tables and Microdata 99
B.2. Tables of Frequency Count Data 101
B.3. Tables of Magnitude Data 101
B.4. Microdata 103
GLOSSARY 104

APPENDIX A – Technical Notes: Extending Primary Suppression Rules To Other Common Situations 106
APPENDIX B – Government References and Websites 110
APPENDIX C – References 112
Books 112
Reports Of Conferences and Workshops 113
Special Issues of Journals 114
Online References 114
Manual 115
Articles 115
APPENDIX D – Confidentiality and Data Access Committee 128