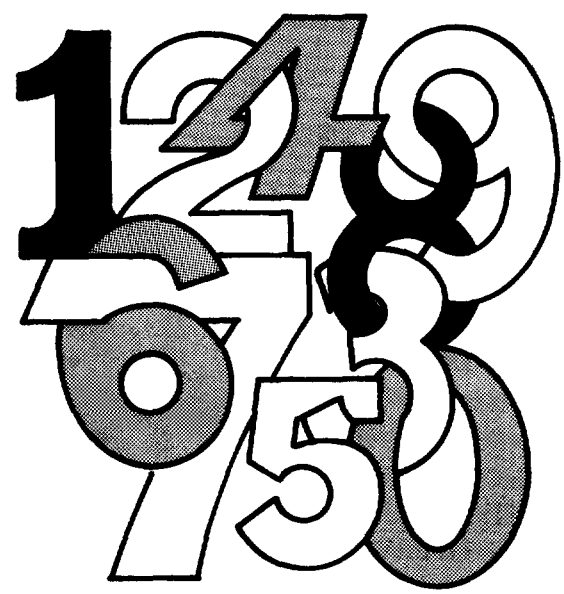
Record Linkage Techniques—1985

Proceedings of the Workshop on Exact Matching Methodologies
Arlington, Virginia
May 9-10, 1985

Co-Sponsored with the
Washington Statistical Society
and the
Federal Committee on Statistical Methodology



Compiled and Edited by
Beth Kilss and Wendy Alvey

Department of the Treasury
Internal Revenue Service
Statistics of Income Division
December 1985

PREFACE

The Workshop on Exact Matching Methodologies was held on May 9-10, 1985, at the Rosslyn Westpark Hotel in Arlington, Virginia. The conference grew out of the efforts of the Matching Group, Administrative Records Subcommittee, of the Federal Committee on Statistical Methodology. It was co-sponsored with the Washington Statistical Society. This volume contains the papers from that event.

The current volume, Record Linkage Techniques -- 1985, is more than just a proceedings of the May conference. It is intended to serve as a handbook on modern matching theory, as well as to report on the current state of the art. For this reason, not only were the papers from the Workshop included here, but extensive background material and bibliographic citations have also been added.

Contents. -- The format for this volume essentially follows that of the Workshop agenda, with several sections added to help round out the actual presentations. The collection begins with an Introduction, which summarizes the objectives of the Matching Group in conducting a Workshop of this sort. It also proposes some recommendations for the statistical community to consider with regard to the future of exact matching. (This latter portion is based on comments made by the participants during and after the Workshop.)

The rest of the volume is set up as follows:

- Section I provides selected background material, which lays the historical groundwork for current methodological thought. Some of these papers were distributed at the conference, but they were not presented as part of the agenda.

- Section II begins the program for the Workshop. This contains three papers presented at the Opening Session, to introduce the theory and provide an overview of matching applications. A fourth (contributed) paper, describing the present state of general methodological issues, is also included.

- Section III focuses on current theory and practice. It is comprised of three invited papers and their resulting discussions, as well as two related contributed papers.

- Sections IV and V follow with papers which describe recent application case studies. Once again, two relevant papers have been added to the six invited papers and their discussions presented at the conference.

- Finally, Section VI deals with computer software for exact matching. It provides the papers presented during the last portion of the Workshop.

The volume also contains two appendices:

- Appendix A consists of selected bibliographies of exact matching methodologies and applications. Five separate collections of references are provided, each with a slightly different orientation.

- Appendix B concludes the volume with information specific to the workshop, itself -- the agenda, the list of attendees, and the list of sponsors.

Copy Preparation. -- The contents of the papers included here are the responsibility of the authors. With the exception of previously published background papers, which were simply reproduced as is, all of the papers in this volume underwent only a limited peer review process. Each paper was read by at least one person familiar with the subject matter. It should be noted, however, that reviewers were instructed to focus on editorial concerns and gross factual problems. Since this did not constitute a formal referee process, authors were also encouraged to obtain their own technical review. Corrections and changes were either made by the authors themselves or cleared through them by the editors. Final layout of the papers was done by the editorial staff, with minor changes of a cosmetic nature considered the prerogative of the editors.

Acknowledgments. -- First of all, I would like to thank the Steering Committee -- Maria Gonzalez, Thomas B. Jabine, Matthew Jaro, Nancy Kirkendall, and Carol Utter -- and the Workshop Coordinators -- Beth Kilss and Wendy Alvey -- for their outstanding job in organizing such a successful conference.

Next, I join the editors of this volume in thanking the Workshop participants for their insightful remarks and helpful suggestions regarding both bibliographic references and recommendations for future research. Along those lines, a special thanks goes to Thomas B. Jabine and Nancy Kirkendall for their work in drafting the Matching Group Pecommendations, which appear following the Introduction. Appreciation is also due to all of the reviewers, for providing their thoughtful comments in such a timely manner. Finally, many thanks go out to the typists -- Michelle Cobb, Denise Herbert, Dawn Nester, Denise Reeder, and Susan Rhodes -- and especially, the editorial staff -- Clementine Brittain, Nancy Dutton, and Bettye Jamerson -- for their extensive contributions to the copy preparation process.

Fritz Scheuren
Director
Statistics of Income Division

December 1985

TABLE OF CONTENTS

Page

Application Case Studies I

Application Case Studies II

Page

Computer Software

Appendix A: Selected Bibliographies of Exact Matching Methodologies and Applications

Page

Appendix B: Workshop Particulars

In June 1984, the Administrative Records Subcommittee recommended to the Federal Committee on Statistical Methodology that a subcommittee be set up to explore integration of surveys and administrative records. The result was the creation of a Matching Group, whose initial (ambitious) goals were to examine policy strategies in conducting data linkages; look at such methodological issues as measurability of matching and analysis of statistical techniques in view of matching errors; and study previous linkage studies for suggestion of possible alternative approaches to matching problems. Both population-based linkages and establishment matches were of interest.

The Matching Group began by reviewing the available literature on exact matching. It soon became apparent that some gaps in knowledge existed that perhaps could be addressed by a workshop conducted by experts currently working in that field. Thus was born the Workshop on Exact Matching Methodologies.

The Workshop was designed to balance the disparate interests of the many different people involved in exact matching: statisticians, research analysts, and computer programmers. Subject matter interests ranged from the epidemiologists' concerns about person-matches to occupation and mortality data, to economists' desires to create estimates based on establishment linkages. As such, the Workshop was viewed as a means of summarizing the work done on matching over the past ten to fifteen years, filling in some of the holes we had discovered, and drawing this more current information together in one place -- this volume -- for use as a ready resource aid by the statistical community and its users. The conference was also seen as a means of building a network of people interested in matching, with a view towards establishing a more coordinated approach to future policy and research efforts.

The Workshop

The Workshop drew 140 registrants from both the U.S. and Canada, representing 47 different agencies, universities, and businesses -- a very sizable segment of the major contributors to the field of exact matching today. Not surprisingly, well over half of those who attended represented Federal agencies; most notably, the Bureau of the Census, Internal Revenue Service, Social Security Administration, Energy Information Administration, and the Bureau of Labor Statistics. Furthermore, about half of those who came expressed primary interest in application issues. The remainder were about equally divided between statistical theory and computational developments.

Workshop Results

In addition to the interactions which took place at the Workshop, there were also several important tangible products which resulted from that effort. First, based on discussions at the Workshop and subsequent correspondence, some recommendations for next steps in exact matching were developed. These were summarized by the Matching Group and appear following this Introduction. Next, selected papers representing the historical development of modern matching methodological thought were assembled. These have been represented here in Section I. Then, in Sections II through VI, presentations from the Workshop and a few additional related papers are provided. Their inclusion is intended to document the current state of exact matching methodology, application and computer software development.

Finally, extensive efforts were made to develop a comprehensive bibliography of exact matching literature. What resulted was a collection of five separate reference lists, each slightly different in orientation. These are provided in Appendix A. Also, along similar lines, the Matching Group developed a special software package containing a menu-prompt library of information on recent exact matching studies. This effort was dubbed Project LINK-LINK and is described in Section VI of this volume.

One of the most important outgrowths of the Workshop, however, was that it provided a long overdue forum for persons working in the area of exact data linkage. It not only sparked new interest in matching, but provided an atmosphere where participants could interact on a more personal basis -- a very important factor which, in the past, has been lacking, resulting in unnecessary duplication of effort in some cases. If nothing else, the Workshop has served its purpose if it acts as a catalyst for initiating more concerted efforts with regard to matching policy, methodological development and application. It was with this aim in mind that the Matching Group assembled the Recommendations which follow.

rive recommendations are provided here from the Matching Group, Administrative Records Subcommittee, Federal Committee on Statistical Methodology. The first recommendation calls for the establishment of a continuing interagency working group on record linkage systems and techniques: such a working group would be expected to play a significant role in implementing recommendations 2 through 5. The second recommendation calls for careful monitoring of external developments that might affect the prospects for undertaking record linkages for statistical purposes. Recommendations 3, 4 and 5 identify specific aspects of record linkage systems and techniques that deserve special emphasis in future research, development and evaluation activities. The five recommendations are:

1. Documentation should be improved and information on record linkage systems and techniques should be shared.
It is recommended that the Matching Group of the Administrative Records Subcommittee be reconstituted as a Technical Working Group on Record Linkage Systems and Techniques, continuing to function under the auspices of the Federal Committee on Statistical Methodology. The main goal of the Working Group would be to promote the effective use of record linkage techniques for statistical purposes by encouraging the documentation of individual record linkage systems and techniques and the sharing of relevant technical information. A primary activity would be sponsorship and organization of workshops and meetings of professional societies to discuss relevant new developments and research, and to disseminate information on existing systems and techniques. In addition, the reconstituted working group would contribute, in appropriate ways, to the implementation of recommendations 2 through 5 below.

2. Changes in the external environment for record linkages should be monitored.
Statistical users of record linkage techniques should track external developments that may influence their ability to perform record linkages. Such developments include changes in laws, regulations and policies affecting access to records and changes in the content of data files used in record linkages. Examples of the latter would include increased use of four-digit ZIP code add-ons ("ZIP + 4") and steps taken to promote the use of unique addresses in rural areas. In so far as possible, statistical users of record linkage techniques, working through the reconstituted Working Group (see recommendation 1), should attempt to influence the course of these developments in ways that will facilitate statistical applications. For example, the Working Group might try to promote the development of standards for reporting names and addresses of both businesses and individuals.

3. Comparative evaluation studies of record linkage systems should be undertaken.
Several agencies of the United States and Canadian governments have invested substantial resources in the development of automated record linkage systems for use in a variety of statistical programs. For many new applications, use of an existing system is likely to be more cost-effective than development of a new one. To aid potential users of record linkage systems, it is recommended that resources be sought for comparative evaluations of existing systems and some of their components, such as name and address standardizers and blocking rules. The evaluation design should recognize that record linkage systems vary in their objectives, especially with respect to the kinds of units for which records are to be matched: persons or businesses. A much-needed first step is the development of a detailed evaluation plan that specifies the measures of quality and cost to be used in the evaluation and the nature of the files to be matched. Such evaluations may require data sets for which true match status is known. One possibility would be to create such data sets by simulation.

4. Research and development aimed at the improvement of record linkage systems and techniques should give priority to selected aspects.
Recognizing that resources for the development of improved record linkage systems are limited, it is recommended that priority be given to the following aspects: (1) systems for linking business records, (2) name and address standardizers, (3) string comparators, (4) the choice of blocking strategies, (5) the development of "learning" systems, and (6) the role of manual intervention.

5. Errors associated with record linkages and their effects on analyses should be measured.
It is recommended that more research be carried out on the error characteristics of record linkage systems and on the effects of errors on analyses performed with the linked data sets. To enhance the value of such research, consensus is desirable on standard measures of record linkage errors and on methods of measuring them. Promising error measurement methods include

multiple matching techniques and direct contacts with samples of linked pairs to determine their true match status.

By design, the principal focus of the Workshop discussions and followup comments by participants was on methodological aspects of record linkages for statistical purposes. Legal and ethical considerations in such linkages were not part of the main agenda.

Nevertheless, the Matching Group of the Administrative Records Subcommittee recognizes that legal and ethical considerations must be weighed carefully by any organization that links records from different sources and that public perceptions of the appropriateness of various kinds of record linkages are also of critical importance. More research in these areas would also be desirable, addressing, in particular: (1) public understanding of and attitudes toward linkages performed for statistical and other purposes; (2) survey respondents' comprehension of informed consent statements currently being used, especially when such statements cover linkages of survey data and administrative records; and (3) the effects on survey response of varying the amount and kinds of information included in informed consent statements to respondents.

Members of the

MATCHING GROUP: ADMINISTRATIVE RECORDS SUBCOMMITTEE

(as of May 1985)

Mary Bentz
Internal Revenue Service

David Cordray
General Accounting Office

Jane Crane
National Center for Education Statistics

Maria Elena Gonzalez
Office of Management and Budget

Nick Greenia
Internal Revenue Service

David Hirschberg
Small Business Administration

Howard Hogan
Bureau of the Census

Thomas B. Jabine
Consultant, Committee on National Statistics

Matt Jaro
Bureau of the Census

Patrick Kelley
Bureau of the Census

Beth Kilss
Internal Revenue Service

Nancy Kirkendall
Energy Information Administration

Douglas Kleweno
Department of Agriculture

Thomas Reilly
Bureau of the Census

Douglas Sater
Bureau of the Census

Fritz Scheuren (Chair)
Internal Revenue Service

Wray Smith
Mathematica Policy Research

Wendel Thompson
Department of Energy

Carol Utter
Bureau of Labor Statistics

William Winkler
Energy Information Administration