C-3

SURVEY METHODOLOGY

STATISTICS CANADA STATISTIQUE CANADA
JUL 27 1993
LIBRARY BIBLIOTHÈQUE
C.3

Catalogue 12-001

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 1993

•

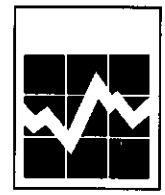
VOLUME 19

•

NUMBER 1

Statistics Canada Statistique Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 1993 • VOLUME 19 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry,
Science and Technology, 1993

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

July 1993

Price: Canada: $35.00
United States: US$42.00
Other Countries: US$49.00

Catalogue No. 12-001

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members
B.N. Chinnappa
G.J.C. Hole
F. Mayda (Production Manager)
R. Platek (Past Chairman)
C. Patrick
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D. Binder, *Statistics Canada*
E.B. Dagum, *Statistics Canada*
J.-C. Deville, *INSEE*
D. Drew, *Statistics Canada*
R.E. Fay, *U.S. Bureau of the Census*
W.A. Fuller, *Iowa State University*
J.F. Gentleman, *Statistics Canada*
M. Gonzalez, *U.S. Office of Management and Budget*
R.M. Groves, *U.S. Bureau of the Census*
D. Holt, *University of Southampton*
G. Kalton, *University of Michigan*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Laval University*
D.B. Rubin, *Harvard University*
I. Sande, *Bell Communications Research, U.S.A.*
C.-E. Särndal, *University of Montreal*
W.L. Schaible, *U.S. Bureau of Labor Statistics*
F.J. Scheuren, *U.S. Internal Revenue Service*
J. Sedransk, *State University of New York*
C.M. Suchindran, *University of North Carolina*
J. Waksberg, *Westat Inc.*
K.M. Wolter, *A.C. Nielsen, U.S.A.*

Assistant Editors

P. Lavallée, L. Mach and H. Mantel, *Statistics Canada*

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $35 per year in Canada, US $42 in the United States, and US $49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal of Statistics Canada
Volume 19, Number 1, June 1993

CONTENTS

Survey Methodology, June 1993
Vol. 19, No. 1, p. 1
Statistics Canada

In This Issue

This issue of *Survey Methodology* features a special section on **Record Linkage and Statistical Matching**. Special thanks are due to Fritz Scheuren for coordinating the editorial work of this special section. One or two papers which also deal with this topic, and which were too late to be included in this issue, may appear in a later issue.

In record linkage two datafiles are combined by linking records which refer to the same unit. The objective may be to create an enriched datafile containing variables from both of the source files, or it may be to identify records referring to common units. In situations where record linkage is not possible, statistical matching could be used to create an enriched datafile. A datafile created by statistical matching may contain synthetic records in the sense that variables obtained from the different data sources need not refer to the same unit; however, it is hoped that the matched file still accurately reflects statistical relationships among the variables.

Bartlett, Krewski, Wang and Zielinski discuss the advantages and disadvantages of record linkage in epidemiological studies. Record linkage methodology and methodological issues are reviewed and illustrated with examples of two large scale record linkage studies in epidemiology. Issues in the analysis of data from linked files are also reviewed.

Belin describes an experimental approach to the evaluation of alternative record linkage procedures. The approach is illustrated through a factorial experiment investigating the effect of such factors as the choice of matching variables, assignment of weights, and other factors. The experiment uses data from the 1988 U.S. census/post-enumeration survey dress rehearsal.

Thibaudeau considers an alternative to the commonly used conditional independence model for the probabilities of matches in different comparison fields. Data from the 1988 St. Louis census/post-enumeration survey dress rehearsal is used for illustration. It is found that the conditional independence model is reasonable for the true links; however, a hierarchical log-linear model with some interaction terms is used for the true nonlinks.

Scheuren and Winkler consider the analysis of data from linked files. In particular they consider the problem of regression of a dependent variable from one source file onto an independent variable from another source file. The approach taken is to estimate and correct for biases due to possibly incorrectly linked records. The approach works well if the probability of a match being a true link (and hence the biases in the regression estimation) can be well estimated. Some empirical results are presented.

The last paper in this special section, by Singh, Mantel, Kinack and Rowe, deals with statistical matching rather than record linkage. The authors develop methods of matching which use auxiliary data to avoid the conditional independence assumption. They also consider imposing categorical constraints so that the matched file agrees with appropriate marginal or conditional categorical distributions obtained from the source files or from auxiliary information. The main conclusion of an empirical evaluation is that the use of appropriate auxiliary information can considerably improve the quality of the matched file.

Hidiroglou, Drew and Gray present standards for the definitions of nonresponse to surveys that are being adopted at Statistics Canada. This will facilitate the analysis of global trends in nonresponse and better understanding of differences in nonresponse to different surveys. Factors affecting nonresponse and measures taken to reduce it are also discussed and nonresponse for two major Statistics Canada surveys is examined.

Treder and Sedransk compare simple random sampling and three allocation methods for double sampling. The three allocation methods are proportional, Rao's and optimal.

Casady and Lepkowski propose stratified telephone survey designs, based on commercial lists of telephone numbers, as alternatives to the widely used two stage random digit dialing procedure known as the Mitofsky-Waksberg technique.The efficiencies of various sampling schemes for this stratified design, simple random digit dialing and the Mitofsky-Waksberg procedure are compared.

Ouyang, Schreuder, Max and Williams consider the problem of estimation in Poisson-Poisson and binomial-Poisson sampling. A number of estimators of totals and standard errors are developed and empirically evaluated in the context of estimation of total volume of usable wood in a stand of trees.

Starting with this issue, *Survey Methodology* is changing to a larger page size.This larger size is less expensive to print and will allow *Survey Methodology* to reduce its continuing production deficit. We also took this opportunity to redesign the cover. I hope you like the result of our efforts.

The Editor

Survey Methodology, June 1993
Vol. 19, No. 1, pp. 3–12
Statistics Canada

Evaluation of Error Rates in Large Scale Computerized Record Linkage Studies

S. BARTLETT, D. KREWSKI, Y. WANG and J.M. ZIELINSKI[1]

ABSTRACT

Matching records in different administrative data bases is a useful tool for conducting epidemiological studies to study relationships between environmental hazards and health status. With large data bases, sophisticated computerized record linkage algorithms can be used to evaluate the likelihood of a match between two records based on a comparison of one or more identifying variables for those records. Since matching errors are inevitable, consideration needs to be given to the effects of such errors on statistical inferences based on the linked files. This article provides an overview of record linkage methodology, and a discussion of the statistical issues associated with linkage errors.

KEY WORDS: Computerized record linkage; Canadian Farm Operators Study; National Dose Registry Mortality Study; Threshold selection.

1. INTRODUCTION

In recent years, there has been a trend in environmental epidemiology towards the use of existing administrative databases as sources of information for health studies (Howe and Spasoff 1986; Carpenter and Fair 1990). In general terms, this involves linking records of human exposure to environmental hazards with records on health status, often using computerized methods for matching individual records from different databases (Newcombe 1988).

Computerized record linkage (CRL) methods have recently been used to examine the mortality experience of over 326,000 farm operators in Canada in relation to farm practices (Jordan-Simpson *et al.* 1990). This study involved linking the Canadian Mortality Data Base (CMDB) with the 1971 Census of Population and the 1971 Census of Agriculture. Preliminary results based on 70,000 male farm operators in Saskatchewan have indicated that, although the cohort as a whole demonstrated no excess mortality for specific causes of death, there was some evidence of a dose-response relationship between mortality due to non-Hodkins lymphoma and acres sprayed with herbicides among farms less than 1,000 acres in size (Wigle *et al.* 1990).

Another ongoing large-scale study which involves record linkage is based on the National Dose Registry (NDR) of Canada. The NDR contains information on occupational exposures to ionizing radiation experienced by approximately 255,000 Canadians dating back to 1950. The NDR has recently been linked to the CMDB to investigate associations between exposure to ionizing radiation and cancer mortality (Ashmore *et al.* 1993).

A number of other health studies have been conducted by linking exposure data to the CMDB. Howe *et al.* (1987) determined significantly elevated lung cancer in uranium miners in the Northwest Territories. Significant associations were determined between lung cancer and diesel fumes and coal dust in a cohort study of male pensioners of the Canadian National Railway Company (Howe *et al.* 1983). Shannon *et al.* (1984) linked employment records of nickel workers in Ontario to the CMDB and found an excess in laryngeal and lung cancer mortality. Morrison *et al.* (1988) found significantly elevated risk of cancer of the lung, salivary gland, buceal cavity and pharynx among Newfoundland underground fluorspar miners. Mao *et al.* (1988) used CRL to link the CMDB to the Alberta Cancer Registry to determine survival rates after diagnosis for a wide range of cancers. The Canadian Labor Force Survey data base has been linked to the CMDB to examine the mortality experience of different occupations (Howe and Lindsay 1983). A comprehensive list of other health studies based on linking exposure data with the CMDB was compiled by Fair (1989).

Record linkage is the process of bringing together two or more separately recorded pieces of information pertaining to the same individual. The procedures for CRL have become highly refined, using sophisticated algorithms to evaluate the likelihood of a correct match between two records (Hill 1988; Newcombe 1988). Statistics Canada has developed a CRL system called CANLINK which is capable of handling both one-file or internal linkages as well as linkages between two separate files (Howe and Lindsay 1981; Smith and Silins 1981).

[1] S. Bartlett, D. Krewski, Y. Wang and J.M. Zielinski, Environmental Health Directorate, Health Protection Branch, Health and Welfare Canada, Ottawa, Ontario, Canada K1A 0L2.

The confidentiality of records protected under the Statistics Act is strictly maintained if they are to be used in a study requiring record linkage. All studies requiring linkage with protected data bases must satisfy a rigorous review and approval process prior to implementation. All linked files with identifying information remain in the custody of Statistics Canada (Labossière 1986).

Record linkage studies have several advantages over traditional epidemiological studies. By using existing administrative databases, the need to collect new data for health studies is circumvented. By accessing existing data, large sample sizes can often be achieved with relatively little effort. Depending on the nature of the databases utilized, record linkage provides an inexpensive way of exploring many possible associations in epidemiological studies.

Record linkage also has a number of disadvantages. Matching errors may occur due to coding differences or nonuniqueness of the identifiers. There is generally little control over the information collected and there can be appreciable loss to follow-up. Record linkage studies also suffer from the same deficiencies as conventional epidemiological studies, including possible biases, confounding, and insensitivity to weak associations between the environment and health.

The purpose of this article is to explore the use of computerized record linkage in epidemiological studies based on administrative health and environmental records. Of particular interest is the impact of false links on statistical inferences about environmental health hazards. Algorithms for computerized record linkage are discussed in section 2. Applications of record linkage in studies of occupational exposure to ionizing radiation and agricultural chemicals are described in section 3. A discussion of statistical issues in the analysis of data bases formed by record linkage is given in section 4. Our conclusions concerning the use of record linkage as a tool for use in environmental epidemiology are presented in section 5.

2. ISSUES IN RECORD LINKAGE

2.1 Problem Definition

Consider two computer files, **A** and **B**, consisting of health data and environmental exposure data, respectively, for two groups of individuals. Each file consists of a number of records or "observations", each containing a number of fields or "components". Typically, each observation corresponds to an individual member of the population. Fields are attributes such as name, address, age, and sex which characterize the observations. Record linkage is used to identify and link observations on each file that correspond to the same individual (Figure 1). In this example, record 1 of file A matches record 1 in file B, and record 2 in data base A matches record 3 in file B. Record 3 in file A does not match any records in file B, nor does record 2 in file B match any records in file A.

If the records contain unique identifiers which were accurately assigned, then the matching operation is trivial. The social insurance number is an example of an identifier that is unique to an individual. However, unique identifiers may not be available, in which case a "hard" linkage cannot be performed and thus some form of probabilistic linkage must be considered (see section 2.3). With this latter form of linkage, the likelihood of a correct match is computed, and a system of linkage weights is used to determine links and nonlinks.

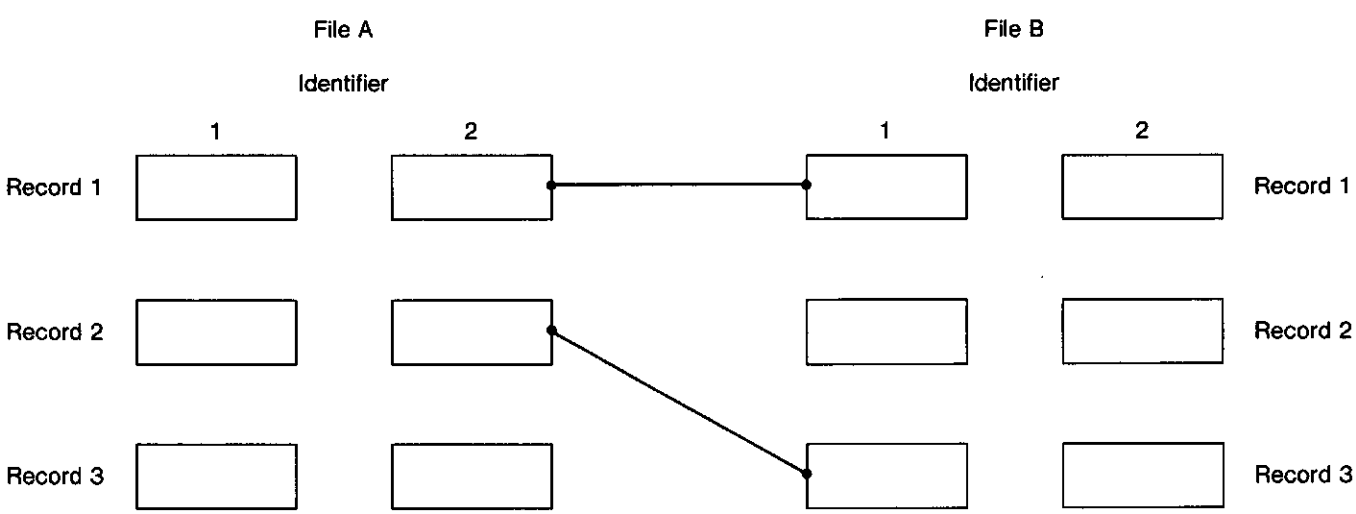


Figure 1. Schematic Diagram of Linking Two Files

2.2 Computerized Record Linkage System (CRL)

In probabilistic record linkage systems, the preliminary matching decision is based on a weight obtained from the comparisons of components of a pair of records (Newcombe 1988). The weight reflects the degree to which the pair is likely to be a true link: the higher the weight, the more likely the pair is a true link. The weight is commonly based on the odds in favour of a match when comparing two records,

$$\frac{P(M|AB\ldots Z)}{P(\bar{M}|AB\ldots Z)} = \frac{P(A|M)P(B|M)\ldots P(Z|M)P(M)}{P(A|\bar{M})P(B|\bar{M})\ldots P(Z|\bar{M})P(\bar{M})}.$$

Here, M is the event that two records match and $\{A, B, \ldots, Z\}$ are outcomes of the comparisons of individual identitiers. The weight w is defined by the log-odds

$$w = log_2\left\{\frac{P(M|AB\ldots Z)}{P(\bar{M}|AB\ldots Z)}\right\}$$

$$= W_a + W_b + \ldots + W_z + W,$$

where

$$W_j = log_2\left\{\frac{P(J|M)}{P(J|\bar{M})}\right\}$$

for all $J \in \{A, B, \ldots, Z\}$, and

$$W = log_2\left\{\frac{P(M)}{P(\bar{M})}\right\}.$$

It should be noted that in order to obtain an absolute odds, it is necessary to know the number of true matches and the number of non-matches. Otherwise, only the relative odds ratio can be determined. The weight determined by the CRL system used by Statistics Canada is the relative log-odds ratio.

Algorithms have been developed for assigning weights for the likelihood of a link between two records, based on the assumption that the likelihoods of the match for the individual identifiers are statistically independent (Howe and Lindsay 1981). Some identifiers, however, may be correlated leading to bias in the assignment of the overall weight.

Fellegi and Sunter (1969) proposed a mathematical model to provide a theoretical framework for record linkage. In the Fellegi-Sunter model, the weight takes into account the error probabilities for each field by using a likelihood ratio, with the weight w defined by

$$w = \sum_{i \in \{\text{fields}\}} w_i,$$

where

$$w_i = \begin{cases} log_2\{m_i/u_i\} & \text{if field } i \text{ of a record pair agrees} \\ log_2\{(1-m_i)/(1-u_i)\} & \text{if field } i \text{ of a record pair disagrees,} \end{cases}$$

with

$$m_i = \Pr\{\text{field } i \text{ agrees} \mid \text{record pair} \in M\} \quad (1)$$

and

$$u_i = \Pr\{\text{field } i \text{ agrees} \mid \text{record pair} \in U\}. \quad (2)$$

Here, M is a set of true matched record pairs and U is a set of un-matched pairs of records. The outcomes of each field comparison are also assumed to be statistically independent (Jaro 1989).

Newcombe (1988), Fellegi and Sunter (1969), Tepping (1968), Copas and Hilton (1990) developed various probabilistic and model-based approaches for assigning weights to components (fields) of records. A probabilistic system like the one used at Statistics Canada determines linkage weights by computing the logarithm of observed odds in favour of a match; other model-based systems use the EM algorithm (Dempster *et al.* 1977) to estimate linkage weights (Jaro 1989; Belin 1989; Winkler 1988).

2.3 Sources of Error

There are a number of sources of potential errors in record linkage that may lead to mismatching of records. Coding errors, such as the wrong birthdate, may occur when records are entered into data bases. There could be variations in the codes, such as different versions of the given name or surname.

In addition to coding errors and coding variations, missing data, especially for important identifiers, will significantly increase the error rate for record linkage (Fair and Lalonde 1988). Duplicate records, which occur when the same record in one file is matched with more than one record in the second file, could also lead to linkage errors (Jabine and Scheuren 1986). Because of this, CRL systems need to include rules that permit multiple matches.

One technique used for increasing the reliability of the surname identifier is to use a phonetic coding system. For example, two observations of an identifier, **ANDERSON** and **ANDERSEN**, will both be recoded as **ANDAR** by using the New York State Intelligence and Identification System (NYSIIS) (Newcombe 1988). Thus, the impact of variations in the name on the linkage would be minimized. However, in compressing the name, the power to discriminate between records may be diminished since two different names may have the same NYSIIS code. The likelihood of making incorrect links also increases (Newcombe 1988).

The given name may have variations with different versions entered on different data bases. Examples are William and Bill, Cynthia and Cindy, and David and Dave. Newcombe *et al.* (1992) discuss methods of using knowledge about variations in given names to increase the likelihood of a correct link.

Sometimes, the available identifiers may not adequately discriminate between individual records. The linkage algorithm may also under use the information contained in the the identifying fields used in the linkage process. Both situations can lead to matching errors.

For large files, it becomes impractical to compare all possible pairs of records. To reduce the number of comparisons, the records for the two files to be linked can be partitioned into mutually exclusive and exhaustive blocks and comparisons be made within blocks. Blocking is generally implemented by sorting the two files using one or more identifying variables. A disadvantage of doing this is that pairs of records, assigned to different blocks would not be compared, and hence would be classified as non-matching. The pairs to be compared would only be drawn from those records where the sorting variables agree. Thus, the number of false negative links would increase (Newcombe 1987; Jaro 1989). Good blocking variables are those based upon blocks that contain nearly the same number of records (Jaro 1989).

In most applications of the Fellegi-Sunter method, results of comparisons for different matching fields are assumed to be independent. Kelley (1986) performed simulation studies to investigate the robustness of the U.S. Census Bureau's linkage system against violations of the independence assumption. For certain populations and linkage variables, it was found that violation of the independence assumption can have an appreciable effect on the linkage error rates.

Newcombe *et al.* (1983) compared the accuracy of computerized matching with that of corresponding manual searches in an epidemiological follow-up study. They found that the computerized matching was more successful than the manual searches, and less likely to yield false links with records not related to the study population. In both approaches, accuracy was strongly dependent on the degree of personal identifying information available on the records being linked. Fair and Lalonde (1987) reached the same conclusion after examining the influence of the availability or non-availability of various identifiers on linkage error rates.

Schnatter *et al.* (1990) tested the adequacy of the CRL system used at Statistics Canada for correctly identifying deaths. Deaths known to have occurred in a cohort of 17,446 refinery and petroleum workers were compared to deaths determined through record linkage to the CMDB. Of the deaths occurring in Canada, 98% were detected by the CRL system.

2.4 Threshold Selection and Error Rate Estimation

After weights have been assigned to all potential matched pairs, a decision is made about the likelihood of the match being a true link. With the Fellegi-Sunter method, each weight is compared to upper and lower thresholds and a decision made as follows.

$$\text{Potential link} = \begin{cases} \text{a link} & \text{if} \quad w \geq w_u \\ \text{a possible link} & \text{if} \quad w_l < w < w_u \\ \text{a non-link} & \text{if} \quad w \leq w_l. \end{cases}$$

Here, w_l and w_u are the lower and upper linkage thresholds, respectively, which ideally are selected to minimize the number of possible links, holding the two types of classification errors (true links classified as non-links and true nonlinks classified as links) at or below given levels.

Where feasible, any matches classified as possible links are resolved manually. Additional information may be used to aid in making decisions about possible links. In many applications, however, manual resolution is not practical, especially for linkages with a large number of possible links. In these situations, a single threshold, $w_t = w_l = w_u$, may be determined so that only two outcomes are possible. Those links with weights greater than w_t are declared links; those with weights less than w_t are declared non-links.

The choice of the threshold w_t is not straightforward. Existing methods are based on knowledge of the linkage error rates which are estimated either by manually resolving a sample of (if not all) possible links, or analytically. The former is a sample based approach since it involves the collection of data to estimate the linkage error rate.

The error rates for record linkage depend on how the thresholds are set. The larger the difference between the upper and lower thresholds, the more possible links there are. With a single threshold, the number of false negatives increases and the number of false positives decreases as the threshold increases.

A simple sample based approach for selecting the threshold entails a pilot study. First, a sample of the smaller of the two files to be linked is selected. Second, links are determined both manually and using a computerized probabilistic record linkage system. Third, assuming that the manually matched links are true links, the threshold is chosen as the weight at which the number of false positives plus the number of false negatives is minimized. Even so, linkage errors could still occur in the manually resolved links due to coding errors, insufficient discriminatory power in the identifiers used, or other linkage problems.

To estimate the error rates of a CRL system, a 2 × 2 contingency table can be constructed as follows.

CRL	Manual	
	Linked	Unlinked
Linked	n_{11}	n_{12}
Unlinked	n_{21}	n_{22}

The false positive (FP) and false negative (FN) rates are then estimated by

$$\mathbf{FP} = \frac{n_{12}}{n_{11} + n_{12}}$$

and

$$\mathbf{FN} = \frac{n_{21}}{n_{21} + n_{22}}.$$

Fellegi and Sunter (1969) point out that the error rates associated with given thresholds are functions of the agreement probabilities for true matches and true non-matches. Consequently, estimates of the agreement probabilities can be used to determine thresholds. This approach is also discussed by Jaro (1989).

For model-based record linkage systems, the principles for linking pairs of records and the strategy for setting thresholds are, with some modifications, similar to the sample based described above. The emphasis of this approach is to fit models for estimating conditional probabilities given by (1) and (2) and for estimating error rate using log odds of two estimated conditional probabilities. One such system uses the EM algorithm to estimate the conditional probabilities m_i and u_i given in (1) and (2) for the ith field of the record by assuming independence of the comparisons among fields,

$$Pr(\gamma^j|M) = \Pi_{i=1}^{n} m_i^{\gamma_i^j}(1 - m_i)^{1-\gamma_i^j},$$

$$Pr(\gamma^j|U) = \Pi_{i=1}^{n} u_i^{\gamma_i^j}(1 - u_i)^{1-\gamma_i^j},$$

($i = 1, \ldots, n$ and $j = 1, \ldots, N$), where n is the number of fields, N is the number of all comparison pairs, and

$$\gamma_i^j = \begin{cases} 1 & \text{if field } i \text{ agrees for record pair } j \\ 0 & \text{if field } i \text{ disagrees for record pair } j. \end{cases}$$

Iterating between the expectation step (E-step) and the maximization step (M-step) in the EM algorithm yields estimates of conditional probabilities $\hat{m}_i$ and $\hat{u}_i$. The overall probability of correct matches may then be estimated based upon $\hat{m}_i$ and $\hat{u}_i$ (Jaro 1989).

Belin and Rubin (1991) provide a procedure which uses previous computer matching experience to fit a mixture model for estimating the linkage error rate. A Box-Cox transformation (Box and Cox 1964) is applied to the weights for matches and for non-matches so that the transformed weights w_i^* form Gaussian distributions, φ_T and φ_F, with means μ_T and μ_F, and variances σ_T^2 and σ_F^2, respectively. All transformed weights are then assumed to come from a mixture distribution

$$\lambda\varphi_T\left(\frac{w_i^* - \mu_T}{\sigma_T}\right) + (1 - \lambda)\varphi_F\left(\frac{w_i^* - \mu_F}{\sigma_F}\right).$$

After estimating the mixture coefficient λ using information obtained from previous matching experience, the above model can be fit using weights obtained from linkage procedure. The error rate for the record linkage algorithm given a particular threshold can then be estimated using the fitted model. The associated standard error of the estimated error rate is also estimated by using the SEM algorithm in which the covariance of the estimated parameters provided by the EM algorithm is estimated (Meng and Rubin 1991).

3. EXAMPLES OF LARGE RECORD STUDIES

3.1 Canadian Farm Operators Study

The Canadian Farm Operators Study was initiated to investigate possible relationships between causes of death in farm operators and various socio-demographic and farming variables. In particular, relationships between pesticide use and mortality are of interest. Mortality data was obtained from the CMDB, while the socio-demographic and farming variables were obtained from the Census of Population and the Census of Agriculture. Since exposure to pesticides was not directly available in the census data bases, variables such as the number of acres sprayed for the control of insects or weeds and the cost of agricultural chemicals was used as surrogate information. The analysis file containing the pertinent information was constructed using probabilistic record linkage.

3.1.1 Cohort Definition

The cohort consists of all male farmers who met the definition for farm operator in the 1971 Census. A farm operator is defined as the person responsible for the daily decisions to be made about the operation of the farm. Farm operators are not necessarily owners, but could be tenants or hired managers. Only one operator was designated for each farm. A farm as determined in the 1971 census was an agricultural holding with one or more acres and with sales of agricultural products of $50 or more. There were 326,000 male individuals who were classified as farm operators (Jordan-Simpson *et al.* 1990). The mortality experience of the cohort was followed up to 1987.

3.1.2 Record Linkage Methodology

The analysis file for the Canadian Farm Operator Study was formed as a result of three separate linkages, the last of which was the linkage of the farm operator cohort file to the CMDB. Before this linkage was done, the farm operator cohort file needed to be constructed.

Socio-demographic data was available from the 1971 Census of Population and information on farming practices was available from the 1971 Census of Agriculture. The Census of Population contains records for every individual in Canada and was collected in two versions, a short form and a long form. The long form asked for more information than the short form and was randomly administered to one third of the households. The Census of Agriculture was administered at all agricultural holdings.

Farm operators are not specifically identified by name in the Census of Agriculture file nor in the Census of Population file. The name and addresses of farm operators are contained in the Central Farm Registry which was created as a mailing list for agriculture questionnaires.

CMDB contains records for all registered deaths reported by the provinces and territories since 1950 and is stored in a standardized, computerized format under the custody of Statistics Canada (Smith and Newcombe 1982). The total number of death registrations on the CMDB from 1950 to 1987 is 5.9 million. The file contains identifying information, plus the date, place and underlying cause of death coded using the International Classification of Disease (ICD) code.

The Statistics Act protects the confidentiality of all records in the CMDB and the Census of Population and Census of Agriculture. As stated previously, all studies requiring linking with these data bases must satisfy a rigorous review and approval process prior to implementation and the resulting linked files with identifying information remain in the custody of Statistics Canada.

To form the analysis file, all the files described above were linked together in three phases.

(a) **Follow-up.** The 1971 and the 1981 Central Farm Registers were linked using CRL to determine if farmers listed in 1971 were still alive in 1981. This information was added to the 1971 Central Farm Registry to increase the probability of linkage to the correct individual in the CMDB.

(b) **Farm operator cohort data base.** The 1971 Central Farm Registry with the follow-up information was merged with the Census of Agriculture in order to add names to the cohort data base. This was necessary for linkage to the mortality data base. The resulting file was then linked to the Census of Populaton file using CRL to form the farm operator cohort data base.

(c) **Analysis file.** The farm operator cohort data base was linked to the CMDB using CRL. The resulting file contained sociodemographic, exposure and death data and was then suitable for analysis.

3.1.3 Threshold Selection

Thresholds were required for each of the three linkages completed to form the analysis file and for linkages based on the short form and the long form. For the mortality linkage, Statistics Canada used a sample based procedure for setting thresholds. This procedure is illustrated for the final linkage of the farm operater cohort data base for those who filled out the short census form with the mortality data base.

A sample of approximately 10% of the short form records filled out by the cohort of farm operators was selected (Statistics Canada 1991a). The links were then determined in two ways, by using the Statistics Canada CRL and by manual resolution using information from death records. The results of the linkages were then compared assuming that the linkages determined by manual resolution were true linkages.

Numbers of false positives and false negatives at a series of link weight thresholds are shown in Figure 2 for the short form linkage. The threshold was selected to minimize the total number of false positives and false negatives, and occurs at a threshold value of 8. The false positive error rate is estimated to be $(36/453) \times 100 = 7.9\%$, while the false negative error rate is estimated to be $(38/20{,}847) \times 100 = 0.2\%$ leading to an overall error rate of 8.1% (Table 1).

Table 1

Comparisons of Linked and Unlinked Records Using CLR and Manual Resolution Based on a Sample of Census Records in the Farm Operator's Study

Computerized Record Linkage (CLR)	Manual Resolution		Total
	Linked	Unlinked	
	Short Form		
Linked	417	36	453
Unlinked	38	20,809	20,847
Total	455	20,845	21,300
	Long Form		
Linked	286	13	299
Unlinked	15	18,498	18,513
Total	301	18,511	18,812

To illustrate the impact of additional identifying information, a similar table can be constructed for the long form records (Table 1). The false positive and false negative error rates are $(13/299) \times 100 = 4.3\%$ and $(15/18{,}511) \times 100 = 0.1\%$, respectively, for an overall rate of 4.4%. Thus, with more identifying information, the error rates can be reduced.

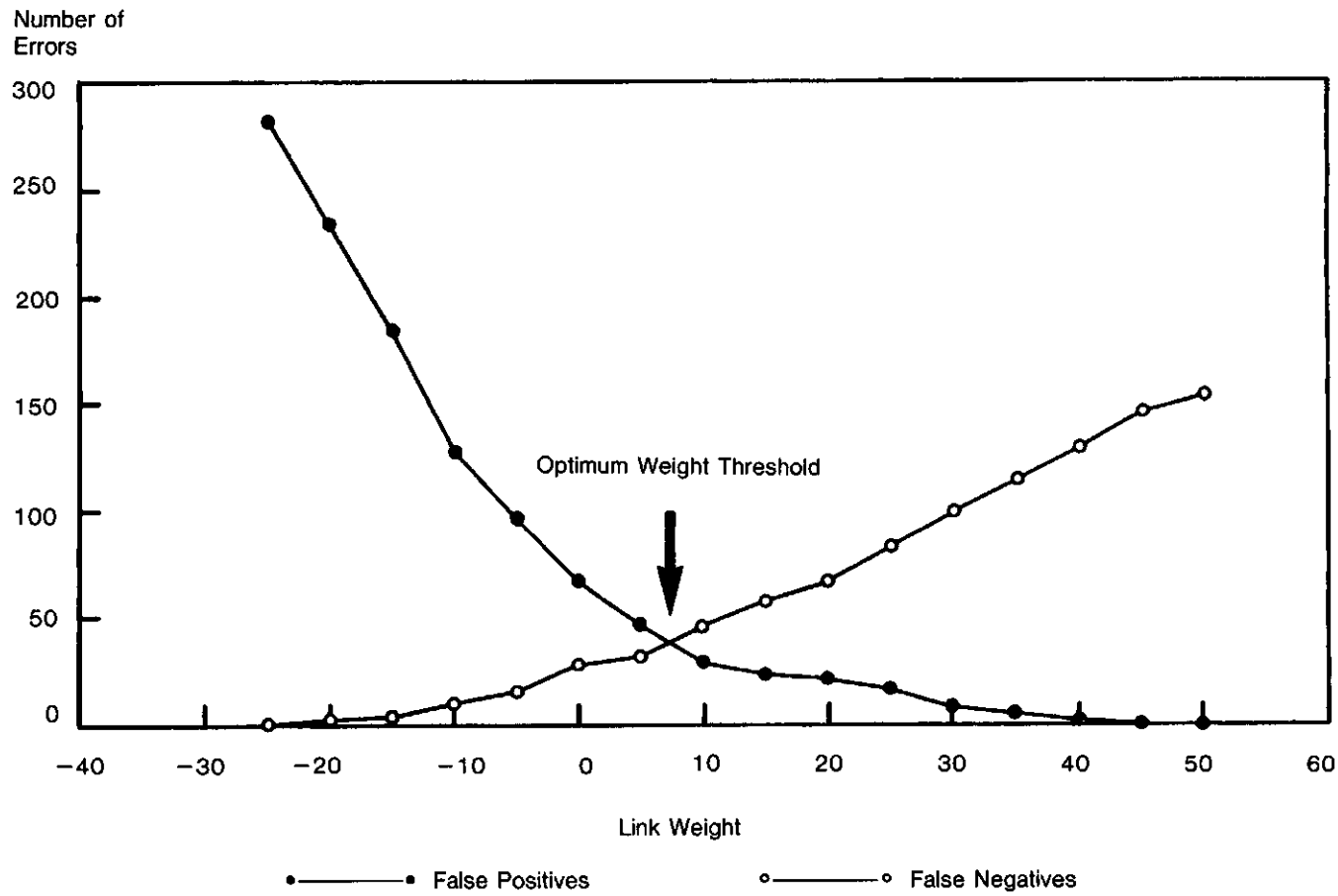


Figure 2. False Positive and False Negative Links Canadian Farmer Operators Mortality Linkage: Short Form

3.2 National Dose Registry Mortality Study

The National Dose Registry of Canada contains records of occupational radiation exposures for approximately 255,000 Canadians dating back to 1951. The NDR has recently been linked to the CMDB. The purpose of the National Dose Registry mortality study is to determine associations between excess mortality due to cancer and other causes and occupational exposure to low levels of ionizing radiation (Ashmore *et al.* 1993).

3.2.1 Cohort Definition

The cohort consists of all workers monitored for ionizing radiation, including tritium and radon daughters, whose records were contained in the National Dose Registry as of December 31, 1983. It contains radiation exposure records of virtually all monitored radiation workers in Canada, with some records providing 37 years of exposure data. In addition, the Registry includes 80 different job categories ranging from nuclear power generating station workers to hospital radiologists to dentists. A total 248,940 people were included in the study cohort.

Depending on the type of radiation and the levels of exposure anticipated within specific job categories, radiation exposure records have been collected annually, quarterly, monthly, or biweekly. Each year, a summary measure of the annual dose experienced by each individual is recorded in the Lifetime Dose History System (LDHS). The annual exposure records maintained in the LDHS will be used as the basis for examining potential relationships between occupational radiation exposure and health status.

The individual data in the LDHS also permit the calculation of a cumulative lifetime dose for each individual. Although individuals will not experience the same level of exposure each year, an average annual dose for an individual can be obtained by dividing the cumulative lifetime dose by the number of years which have elapsed since the time of first exposure. Statistical analysis can be based on the cumulative lifetime dose, the average annual dose, or the annual doses as recorded on the LDHS. Up to 1986, personal identifying information such as surname, given name, sex, year of birth, and assigned identification numbers used to identify the individuals' dose records were stored separately in the Master Identification File (MIF) (Ashmore and Grogan 1985).

3.2.2 Record Linkage Methodology

Identifying variables changed form a number of times during the history of the NDR making tracing an individual's dose history difficult, at times. Because of these and related problems, the Social Insurance Number has been used as the key to the individuals' records from 1977 onward.

There were several linkages required to bring together the appropriate personal identifiers, dose histories, and death information.

(a) **Dose history linkage.** Since 1984 Statistics Canada has been conducting dynamic merges to their LDHS database in order to regroup dose records; reducing the number of fragmented records and consolidating the records into comprehensive dose histories for each study member. The file resulting from the internal linkages indicated which records on the NDR appear to belong to the same individual.

(b) **CMDB linkage.** The internally linked MIF cohort was linked to the mortality records (two-file linkage). By linking the two, it is possible to measure the cohort members' subsequent risk of death. In this study, the CMDB was used to obtain the underlying cause, year of death, the place of death, place of birth, and birth year information.

(c) **Analysis file.** A match of data from MIF, the CMDB and the LDHS was performed to create a comprehensive record for each member of the study cohort. Where the information is available, each record includes birth month and year, sex, the death data listed above, the death linkage weight, and a dose history. Any unmatched records from the MIF or dose history file have undergone special scrutiny.

3.2.3 Threshold Selection

Threshold selection for the link of the cohort file to the CMDB was done in a manner similar to that used in the Canadian Farm Operator Study. First, the weights of potential links were determined. All potential links that had weights less than -30 were considered to be nonlinks. There were 4,429 female and 8,686 cohort members with linkage weights above this value. A sample of these remaining individuals was selected and manually resolved by reviewing death certificates to determine if the links were true links or nonlinks. The threshold was selected at the link weight for which the number of false positive links was equal to the number of false negative links for females and males separately. For females the selected threshold was 53 and for males, 27 (figure not shown).

4. ISSUES IN ANALYSIS OF LINKED DATA SETS

Relatively little work has been done to determine the impact of record linkage on the results of regression analysis. Neter *et al.* (1965) recognized that errors introduced during the matching process could adversely affect analysis based on the resultant linked files. Suppose that true values of a random variable of interest are recorded on a data file comprised of N records. Let Y_i denote the true value for record $i = 1, \ldots, N$. This file is linked to a second file containing identifying information, following which a value of Z_i is assigned to record $i = 1, \ldots, N$. Assuming that all matching errors are equally probable, we have,

$$Z_i = \begin{cases} Y_i & \text{with probability } p \\ Y_j & \text{with probability } q (j \neq i), \end{cases}$$

where $p + (N - 1)q = 1$.

Neter *et al.* (1965) used this model to study the impact of matching errors on the sample mean and variance of the variable Z. The effect of matching errors on the correlation between Z and a second random variable X contained on the same file as well on parameter estimates of the regression between Z_i and X_i, was also investigated. It was shown that (1) the estimate of the mean of Z is unbiased for the mean of the Y; (2) if "X" is positively correlated with Y, the residual variance from a regression of Z on X will be larger than the variance from a regression of Y on "X"; and (3) the slope of the regression line will be underestimated when Z is used rather than Y.

Belin and Rubin (1991) and Winkler and Thibaudeau (1991) discuss theoretical framework, computational algorithms, and software for estimating matching probabilities. These advances motivated Scheuren and Winkler (1991) to update the work of Neter *et al.* (1965). They used the model

$$Z_i = \begin{cases} Y_i & \text{with probability } p_i \\ Y_j & \text{with probability } q_{ij} (j \neq i), \end{cases}$$

where $p_i + \sum_{j \neq i} q_{ij} = 1$, to study the impact of matching errors on the estimates of the coefficients β in the linear regression model

$$Y = X\beta + \epsilon.$$

The effect of matching errors on the above regression model may be expressed as

$$E(Z_i) = Y_i + B_i,$$

where the bias term is given by

$$B_i = (p_i - 1)Y_i + \sum_{j \neq i} q_{ij} Y_i.$$

Instead of using the pair of independent and dependent variables (X_i, Y_i), the pair of independent and linked dependent variables (X_i, Z_i) is used to fit the model. Noticing that the linked dependent variable may be written as $Z = Y + B$, the coefficients are estimated as

$$\hat{C} = (X^T X)^{-1} X^T Z = \hat{\beta} + (X^T X)^{-1} XB,$$

where $\hat{\beta} = (X^T X)^{-1} X^T Y$, so that the bias adjustment is $(X^T X)^{-1} XB$.

Scheuren and Winkler (1991) used these models conduct simulation studies based on real data. Their approach was to take a file of linked and nonlinked cases and re-link them using different matching variables. This simulation demonstrates that the ability to accurately estimate matching probabilities critically effects the accuracy of the coefficient estimates. If the matching probabilities can be accurately estimated, the adjustment procedure works reasonably well.

5. DISCUSSION

Record linkage provides an attractive methodology for exploring relationships between exposures and health outcomes by making use of existing data bases. However, linkage errors are possible, resulting from coding errors, variations in identifiers, missing data, and insufficient discrimination power in the identifiers.

Error rates depend on the amount of identifying information, as seen in the farm operators study. Here, the error rate decreased for the linkage of the long census form where more identifying information was available than in the short census form. Thus, it is important that good identifying information be available for record linkage.

Relatively little attention has been paid to the impact of linkage errors on statistical inferences based on record linkage studies. Such errors can lead to biases in estimates of measures of association between health and environmental variables, such as regression coefficients. Work is in progress to investigate the impact of these errors on the results of the epidemiological studies presented in this paper.

ACKNOWLEDGMENTS

We would like to thank Martha Fair, Statistics Canada, and Dr. Howard Morrison, Health and Welfare Canada for their helpful comments on this article. We also thank two anonymous reviewers for many constructive suggestions.

REFERENCES

ASHMORE, J.P., and GROGAN, D. (1985). The National Dose Registry of Canada. *Radiation Protection Dosimetry*, 11, 95-100.

ASHMORE, J.P., KREWSKI, D., and ZIELINSKI, J.M. (1993). National Dose Registry Study. *European Journal of Cancer*, submitted.

BELIN, T.R. (1989). Results from evaluation of computer matching. Memorandum, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BELIN, T.R., and RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.

BOX, G., and COX, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, 26, 211-246.

CARPENTER, M., and FAIR, M.E. (Eds.) (1990). *Canadian Epidemiology Research Conference - 1989: Proceedings of Record Linkage Sessions and Workshop*. Ottawa, Ontario: Ottawa Select Printing.

COPAS, J.B., and HILTON, F.J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society*, Series A, 153, 287-320.

DEMPSTER, A.D., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via EM algorithm, (with discussion). *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

FAIR, M.E. (1989). Studies and references relating to uses of the Canadian Mortality Data Base. Report from the Occupational and Environmental Health Research Unit, Health Division, Statistics Canada, Ottawa.

FAIR, M.E., and LALONDE, P. (1988). Missing identifiers and the accuracy of individual follow-up. *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada, Ottawa, 95-107.

FELLEGI, I.P., and SUNTER, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

HILL, T. (1988). Generalized Iterative Record Linkage System: GIRLS Strategy. Release 2.7. Report from Research and General Systems, Informatics Services and Development Division, Statistics Canada, Ottawa.

HOWE, G.R., and LINDSAY, J. (1981). A Generalized Iterative Record Linkage computer system for use in medical follow-up studies. *Computers and Biomedical Research*, 14, 327-340.

HOWE, G.R., and LINDSAY, J. (1983). A follow-up study of a ten-percent sample of the Canadian Labor Force. I. Cancer mortality in males, 1965-73. *Journal of the National Cancer Institute*, 70, 37-44.

HOWE, G.R., FRASER, D., LINDSAY, J., PRESNAL, B., and YU, S.Z. (1983). Cancer mortality (1965-77) in relation to diesel fume and coal exposure in a cohort of retired railway workers. *Journal of the National Cancer Institute*, 70, 1015-1019.

HOWE, G.R., NAIR, R.C., NEWCOMBE, H.B., MILLER, A.B., BURCH, J.D., and ABBATT, J.D. (1987). Lung cancer mortality (1950-80) in relation to radon daughter exposure in a cohort of workers at the Eldorado Port radium uranium mine: Possible modification of risk by exposure rate. *Journal of the National Cancer Institute*, 79, 1255-1260.

HOWE, G.R., and SPASOFF, R.A. (Eds.) (1986). *Proceeding of the Workshop on Computerized Linkage in Health Research.* Toronto: University of Toronto Press.

JARO, M.A., (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: methodological issues. *Journal of Official Statistical*, 2, 255-277.

JORDAN-SIMPSON, D.A., FAIR, M.E., and POLIQUIN, C. (1990). Canadian Farm Operator Study: Methodology. *Health Reports*. Catalogue 82-003, Statistics Canada, 2, 141-155.

KELLEY, R.P. (1986). Robustness of the Census Bureau's record linkage system. Paper Presented at the August 1986 meeting of the American Statistical Association.

LABOSSIÈRE, G. (1986). Confidentiality and access to data: the practice at Statistics Canada. *Proceedings of the Workshop on Computerized Record Linkage in Health Research.* Toronto: University of Toronto Press.

MAO, Y., SEMENCIW, R., MORRISON, H., KOCH, M., HILL, G., FAIR, M., and WIGLE, D. (1988). Survival rates among patients with cancer in Alberta in 1974-78. *Canadian Medical Association Journal*, 138, 1107-1113.

MENG, L., and RUBIN, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86, 899-911.

MORRISON, H.I., SEMENCIW, R.W., MAO, Y., and WIGLE, D.T. (1988). Cancer mortality among a group of fluorspar miners exposed to radon progeny. *American Journal of Epidemiology*, 128, 1266-1275.

NETER, J., MAYNES, E.S., and RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.

NEWCOMBE, H.B., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., and ABBATT, J.D. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Computers in Biology and Medicine*, 13, 157-169.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: methods for health and statistical studies, administration and business.* Oxford: Oxford Medical Publications.

NEWCOMBE, H.B., FAIR, M.E., and LALONDE, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association*, 87, 1193-1204.

SCHEUREN, F., and WINKLER, W.E. (1991). An error model for regression analysis of data files that are computer matched. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C.

SCHNATTER, A.R., ACQUIVELLA, J.F., THOMPSON, F.S., DONALESKI, D., and THERIAULT, G. (1990). An analysis of death ascertainment and follow-up through Statistics Canada's Mortality Data Base system. *Canadian Journal of Public Health*, 81, 60-65.

SHANNON, H.S., JULIAN, J.A., and ROBERTS, R.S. (1984). A mortality study of 11,500 nickel workers. *Journal of the National Cancer Institute*, 73, 1251-1258.

SMITH, M.E., and NEWCOMBE, H.B. (1982). Use of the Canadian Mortality Data Base for epidemiological follow-up. *Canadian Journal of Public Health*, 73, 39-46.

SMITH, M.E., and SILINS, J. (1981). Generalized Iterative Record Linkage System. *Proceedings of the Social Statistics Section, American Statistical Association*, 128-137.

STATISTICS CANADA, (1991a). Canadian Farm Operators' mortality study general work plan. Internal report of the Occupational and Environmental Health Research Section, Statistics Canada, Ottawa.

STATISTICS CANADA, (1991b). Unpublished data.

TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.

WIGLE, D.T., SEMENCIW, R.M., WILKINS, K., RIEDEL, D., RITTER, L., MORRISON, H.I., and MAO, Y. (1990). Mortality study of Canadian male farm operators: Non-Hodgkin's lymphoma mortality and agriculture practices in Saskatchewan. *Journal of the National Cancer Institute*, 82, 575-581.

WINKLER, W.E. (1988). Using the E.M. algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.

WINKLER, W.E., and THIBAUDEAU, Y. (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division, Technical Report.

Survey Methodology, June 1993
Vol. 19, No. 1, pp. 13–29
Statistics Canada

Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment

THOMAS R. BELIN[1]

ABSTRACT

Record linkage refers to the use of an algorithmic technique for identifying pairs of records in separate data files that correspond to the same individual. This paper discusses a framework for evaluating sources of variation in record linkage based on viewing the procedure as a "black box" that takes input data and produces output (a set of declared matched pairs) that has certain properties. We illustrate the idea with a factorial experiment using census/post-enumeration survey data to assess the influence of a variety of factors thought to affect the accuracy of the procedure. The evaluation of record linkage becomes a standard statistical problem using this experimental framework. The investigation provides answers to several research questions, and it is argued that taking an experimental approach similar to that offered here is essential if progress is to be made in understanding the factors that contribute to the error properties of record-linkage procedures.

KEY WORDS: Cutoff weight; False-match rate; Fellegi-Sunter algorithm; Matching variables; Post-enumeration survey; String comparison; Weighting scheme.

1. EVALUATING RECORD-LINKAGE PROCEDURES

Record linkage refers to the use of an algorithmic technique to identify pairs of records, one from each of two data files, that correspond to the same individual. The goal is to identify, using a computerized approach, the records from the respective data files that should be declared "matched" as well as the records that should be declared "not matched" without an excessive rate of error, thereby avoiding the cost of manual processing.

Specifying a record-linkage procedure requires both a method for measuring closeness of agreement between records and a rule for deciding when to classify records as matches or non-matches. Much attention has been paid in the record-linkage literature to the problem of assigning so-called "weights" to individual fields of information in a multivariate record to obtain a "composite weight" that summarizes the closeness of agreement between two individuals (*e.g.*, Newcombe *et al.* 1959; Fellegi and Sunter 1969; Newcombe 1988; Copas and Hilton 1990). Less attention has been paid to other aspects of record-linkage procedures, such as the handling of close but inexact agreement between fields of information, and to the effects of using various approaches (treatments) in combination with one another.

In some settings, a personal identifier, such as a social security number, can serve as a basis for linkage. However, such an identifier is not always available, and even when one is present, it still may be necessary to rely on other identifying information for a substantial subset of cases (*e.g.*, Rogot, Sorlie and Johnson 1986).

This paper describes a large factorial experiment contrasting various procedures for matching census and post-enumeration survey (PES) records. Social security number is not collected in the census, so we are in a setting where closeness of agreement is based on several variables. Interest focuses on two questions:

(1) What are the most important factors affecting the accuracy of record linkage?

(2) What combination of factors works best in practice?

Beyond addressing these questions in the census/PES setting, perhaps the most important contribution of this investigation is the idea that record-linkage procedures should be studied by conducting careful experiments. With many factors at the discretion of the operator of the program, there is little hope of understanding the full complexities of a matching algorithm by varying factors one at a time (or worse, not even conducting any systematic evaluation at all). The idea of conducting an experiment would seem quite natural to an agricultural scientist or an industrial quality-control engineer, although it seems that such an approach has not been taken in the context of record linkage aside from this investigation and earlier work by the author (Belin 1989a, 1989b).

2. APPLIED CONTEXT FOR RECORD LINKAGE

2.1 Applications of Record Linkage

Record-linkage methods have been used in a variety of settings. Applications can be characterized as falling into two broad groups: problems where it is desired to draw

[1] Thomas R. Belin, Department of Biomathematics, UCLA School of Medicine, Los Angeles, CA, 90024-1766, U.S.A.

inferences about relationships between variables collected in separate large data files, and problems where interest focuses directly on the number of individuals represented in one or both data files (or a function of those quantities).

Examples of the first type of application are numerous. Studies have been conducted linking data from health and nutrition surveys to registries of mortality data to study relationships between dietary risk factors and death from various causes (Johansen 1986), linking labor force survey data to mortality data to assess health effects of uranium mining (Newcombe, Smith, Howe, Mingay, Strugnell and Abbatt 1983; Abbatt 1986), linking information on educational background to records of earnings of individuals some years later to assess the benefit of a college education (Fagerlind 1975), comparing reported income on welfare records to reported income on tax records (Kershaw and Fair 1979), and linking records of individuals exposed to radiation during atomic-bomb tests and records of a cohort of control individuals to national death records to assess differences in mortality patterns between exposed and control individuals (Dulberg, Spasoff and Raman 1986). Using record-linkage methodologies in such studies is attractive primarily for reasons of cost and timeliness, since for any of the research endeavors just described, it would take much longer and would have been much more expensive to conduct studies with one or more stages of followup than it was to make use of existing data.

The primary motivating example in this article is representative of the other type of application, where the goal is to determine the number of overlapping cases in two data files. In this example, a record-linkage procedure is used as the first step of an extensive matching operation in which records from a census are compared to records from a large-scale post-enumeration survey (PES) conducted after the census to evaluate census coverage. Other examples where the goal is to determine the number of overlapping cases between data files are the investigation by Nicholl (1986) of classification errors regarding the types of injuries sustained by road accident victims (based on linking hospital records to police reports of accidents), the investigation by Johnson (1991) into caseloads for U.S. Attorneys in different districts around the country (based on linking a list of cases assembled by the Department of Justice to a list of cases assembled by federal district courts), and a variety of investigations into the accuracy and coverage of mortality data files (Wentworth *et al.* 1983; Curb *et al.* 1985; Boyle and Decouflé 1990; Williams *et al.* 1992).

Census undercount estimation has been a prominent and at times controversial topic in statistical research, especially during the past decade. Much of the controversy revolves around a proposed adjustment of the census based on undercount estimates from a PES. For general background on issues involved in census undercount estimation, see Ericksen and Kadane (1985), Citro and Cohen (1985), Freedman and Navidi (1986), Wolter (1986), Schirm and Preston (1987), Ericksen, Kadane, and Tukey (1989), Cohen (1990), and the special sections on census coverage error in the June and December, 1988, issues of this journal. A record-linkage procedure is the first step of matching census records to PES records; it is followed by matching of records by clerks, subsequent followup interviewing of households when there appear to be discrepancies between the census and PES findings, and an additional round of clerical matching after followup interviewing. Based on assessments from the matching operation and certain assumptions about the probability that individuals would be included only in the census, only in the PES, in both the census and PES, or in neither the census nor PES, it is possible to estimate undercount (or overcount) rates in the census.

2.2 Background on Record-Linkage Theory

The development probabilistic reasoning in record-linkage theory can be traced to Newcombe, Kennedy, Axford, and James (1959), who develop a weighting scheme in an effort to reflect the odds that a pair of records is correctly matched. Fellegi and Sunter (1969) enhance the theoretical underpinnings of commonly-used weighting rules, noting that the procedure proposed by Newcombe *et al.*, corresponds to calculating a likelihood ratio under a simple model for the record-linkage problem that supposes independence of agreement among all fields of information within records. They show that a weighting scheme similar to that of Newcombe *et al.*, combined with cutoff weights that depend on a specified false-match rate and a specified false non-match rate, define a linkage procedure that is optimal in the sense of minimizing the proportion of records that will be assigned neither as definitely matched nor as definitely not matched, assuming the underlying model is valid.

Much of the ensuing development of record-linkage technology has taken place in the context of applications, as investigators put the theoretical ideas outlined in the earlier literature to practical use. Prominent applications include the Oxford Record Linkage Study (Acheson 1967; Goldacre 1986); the three-way match among records from the Current Population Survey, the Social Security Administration, and the Internal Revenue Service (Kilss and Scheuren 1978); and the National Longitudinal Mortality Study (Rogot, Sorlie, Johnson, Glover and Treasure 1988). The proceedings volumes from conferences on record linkage (Kilss and Alvey 1985; Howe and Spasoff 1986; Carpenter and Fair 1990), compilations of papers from annual conferences (Kilss and Alvey 1984a; Kilss and Alvey 1984b; Kilss and Alvey 1984c; Kilss and Alvey 1987; Kilss and Jamerson 1990), and proceedings volumes from conferences more broadly focused on uses of administrative

data (Coombs and Singh 1988) document numerous other applications that make use of record-linkage methodology.

Software development has enhanced the ability to pursue research into record linkage. Software incorporating refinements of weighting methods and blocking strategies has been developed for use in a variety of applications at Statistics Canada and the U.S. Bureau of the Census. Background on the Statistics Canada "Generalized Iterative Record Linkage System" (GIRLS) is discussed in Howe and Lindsay (1981); documentation is contained in Hill (1981) and Hill and Pring-Mill (1986). Background on the matching system developed by the Record Linkage Staff at the U.S. Bureau of the Census can be found in Jaro (1989), Winkler (1989), and Winkler and Thibaudeau (1992), with documentation found in Laplant (1988), Laplant (1989), and Winkler (1991).

New models that reflect subtleties within data files that could be used in developing a probabilistic weighting scheme are offered by Copas and Hilton (1990). Other extensions to record-linkage methodology designed to take advantage of information in person names are described in Newcombe, Fair and Lalonde (1992). A review paper by Jabine and Scheuren (1986), a textbook by Newcombe (1988), and a compilation by Baldwin, Acheson and Graham (1987) serve as broad references on record-linkage methodology.

2.3 Flow of a Standard Record-Linkage Procedure

Typical steps in a record linkage procedure can be described as follows: (1) data collection, (2) preprocessing of data, (3) determination of rules for assessing closeness of agreement between candidate matched pairs, (4) assignment of candidate matched pairs, and (5) declaration of matched pairs. We use the term "candidate matched pairs" to describe pairs of records that are brought together as being the best potential match for each other from the respective data files (*cf.* "hits" in Rogot, Sorlie, and Johnson (1986); "pairs" in Winkler (1989); "assigned pairs" in Jaro (1989)). Candidate matched pairs might be declared matched after the application of a decision rule in step (5), but they will not necessarily be declared matched by the decision rule.

As indicated earlier, closeness of agreement between candidate matched pairs is assessed in many record-linkage procedures by a univariate summary statistic, often referred to as a "composite weight". In such procedures, step (3) above would refer to the determination of weighting rules, and step (5) above would involve the setting of a cutoff weight above which record pairs will be declared matched.

Record linkage may be viewed as a decision problem with two or more actions to be taken by the computer. Typically, three actions are considered (*e.g.*, declare records matched, declare records as not matched, or send record to be reviewed more closely by a human observer, as in Fellegi and Sunter 1969), although sometimes only two actions (declare matched, declare not matched) are contemplated, and as many as five actions have been considered in some instances (Tepping 1968).

Postulating that distance between multivariate records can be summarized by a univariate composite weight narrows the scope of possible procedures that could be used to perform record linkage. The author is aware of very little research exploring alternatives to such univariate-composite-weight approaches, other than merely specifying a deterministic set of rules for when to declare records matched; one exception is Smith and Newcombe (1975). Such alternatives are beyond the scope of this paper.

2.4 Detailed Description of the Procedure Used to Match Census/PES Records

A variety of separate techniques may be involved in each of the five steps outlined above. Figure 1 provides a flowchart illustration of the main steps used in the linkage of census/PES records.

The frame of the census is a compilation of housing-unit address listings. Addresses are assembled by a variety of techniques, generally depending on whether the area is urban or rural. In urban and suburban areas, census forms are mailed to households with the hope that residents will respond by mailing back a completed form; in other areas census enumerators visit households. When there is no response from a household that was sent a census form by mail, an enumerator will visit the household in person. Data are entered into Census Bureau computer files by a combination of computerized scanning techniques and clerical keying operations. An overview of census methodology can be found in Citro and Cohen (1985); detailed descriptions of various census operations can be found in the Census Bureau's 1990 Decennial Census Information Memorandum Series (Bureau of the Census 1988-1991).

Data collection in the type of post-enumeration survey conducted in 1990 (and in test censuses leading up to the 1990 PES) begins with a process of listing addresses that is conducted by enumerators canvassing neighborhoods. Information is obtained entirely through interviewing operations as opposed to the mailout-mailback approach. Data are entered into computer files entirely by clerical keypunching. Hogan (1992) provides an overview of the PES; details of PES operations can be found in the Census Bureau's STSD Decennial Census Memorandum Series (Bureau of the Census 1987-1991).

Preprocessing of data is rarely discussed in the literature on record linkage, even though this stage provides opportunities both for squeezing available information from the data at hand and for unwisely discarding information available from the data. Winkler (1985a, 1985b) presents

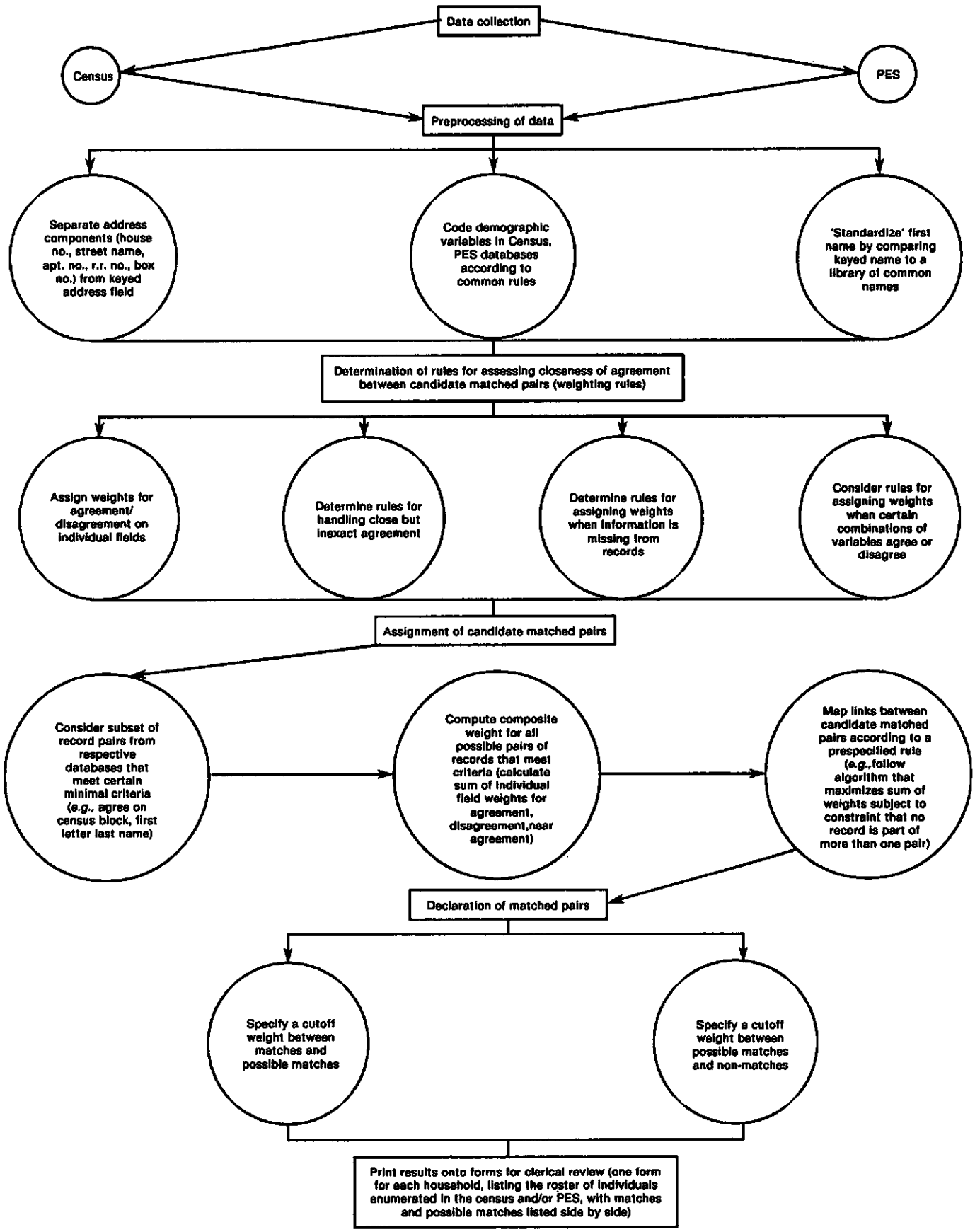


Figure 1. Flowchart of Census/PES Record Linkage Procedures

some specific strategies that are shown to make it easier to distinguish true matches from false matches, and Jabine and Scheuren (1986) and Newcombe (1988) offer some broad guidelines in this area. In the census/PES matching operation, preprocessing of data includes coding demographic variables according to common rules, identifying and separating address components (such as house number, street name, apartment number, rural route number, and post office box number) from the keyed address field (Laplant 1989), and "standardizing" an individual's first name by comparing the keyed first name to a library of nicknames and converting nicknames observed in the data to their common antecedent names (Paletz 1989).

The census/PES record-linkage procedure is a weight-based procedure. The determination of a weighting method includes consideration of both model-based and *ad hoc* rules for assigning weights for agreement and disagreement on individual fields of information, rules for assigning weights for close but inexact agreement on particular fields, rules for assigning weights when information is missing from records, and rules for assigning weights when certain combinations of variables are found to be in agreement or disagreement.

The designation of candidate matched pairs in census/PES matching reflects certain constraints that are placed on the matching process. First, time and resource constraints make it impractical to compare each record in one data file to every record in the other data file. Accordingly, comparisons are made only between pairs of records that meet certain minimal criteria, such as that they fall in the same census block and share the same first letter of last name. The subset of records formed by this restriction is referred to as a "block", and the variables required to be in agreement for a match to be declared are referred to as "blocking variables" (Jaro 1989).

Another constraint placed on the census/PES matching operation is that a given record in one data file is not allowed to be declared matched to more than one record in the other data file. The approach that is used to perform the assignment of candidate matched pairs draws on operations-research techniques for solving the so-called transportation problem (Jaro 1989). The algorithm assigns candidate matches so as to maximize the sum of composite weights among all possible pairs of records within a block defined by the blocking variables, subject to the aforementioned restriction that no record is allowed to match more than one record in the other data file. For example, suppose that within a particular block record A from file 1 has a higher agreement weight with record B from file 2 than with any other record in file 2. The assignment algorithm still might link record A to another record, say C, and link B to another record, say D, if the sum of the agreement weights for (A,C) and (D,B) are higher than for other permutations of candidate match assignment.

The current approach to census/PES matching contemplates three possible actions to be taken by the computer: declare a record pair to be a match, declare a record pair to be a "possible match", or declare a record to be not matched. All non-matches and possible matches are sent to clerks to be reviewed, and an attempt to obtain a followup interview is made for households where there is a discrepancy between the census and the PES. The distinction between possible matches and non-matches only has to do with the procedures applied by clerks when they review these cases (Childers 1989; Donoghue 1990). In the processing of 1990 census/PES data, the operator of the matching program set cutoff weights manually to distinguish matches, possible matches, and non-matches after scanning sets of candidate matched pairs with weights in a certain range. A new technique by Belin and Rubin (1991) offers an alternative for automating the setting of cutoffs.

3. AN EXPERIMENT

3.1 Factors Influencing the Output of Record-Linkage Procedures

The performance of a record-linkage procedure can depend on a number of factors, including:

(1) The choice of matching variables;

(2) The choice of blocking variables;

(3) The assignment of weights to agreement or disagreement on various matching variables;

(4) The handling of close but not exact agreement between matching variables;

(5) The handling of missing data in one or both of a pair of records;

(6) The algorithm for assigning candidate matches;

(7) The choice of a cutoff weight above which record pairs will be declared matched;

(8) The site or setting from which the data are obtained.

Among these factors, only (8) represents a source of variation over which the operator of the matching program does not have control. As mentioned earlier, two lines of inquiry are of primary interest in the experiment. Identifying major sources of variability in record linkage could help to focus future record-linkage research and to offer a deeper understanding of the process that generates errors in linkage procedures. Further, it is of interest to identify the combination of factors that works best in achieving a maximum number of matches while maintaining low error rates, since in practice the user generally must make a single choice among a myriad of possibilities for each factor just described.

3.2 Factorial Experiment Using Census/ Post-Enumeration Survey Data

A study was conducted using data from each of the three sites (St. Louis, Missouri; East Central Missouri including the Columbia, Missouri area; and a rural area in eastern Washington state) of the 1988 dress rehearsal census and PES. These data sets had been matched by computer and then reviewed by clerks. For the purposes of subsequent analysis, the final clerical determinations of true and false match status are taken as the truth. Thus, although subsequent analyses will only be as accurate as the determinations by clerks, these data files offer an excellent opportunity to study record linkage.

Descriptions of the specific methods used in linking records between the census and PES can be found in Jaro (1989), Winkler (1991), and Winkler and Thibaudeau (1992). The current implementation of the record-linkage procedure allows the user a variety of options over all of the factors listed in Section 3.1 except for the choice of an algorithm for assigning candidate matches (a "linear-sum assignment" algorithm is used; see Jaro 1989).

The variables available for matching census/PES records include name, address, age, race, sex, telephone number, marital status, and relationship to head of household. In practice, name is usually broken down into first name, last name, and middle initial, with these three used as separate matching variables. A preprocessing program is typically used to parse address information into house number, street name, apartment number, rural route number, and box number (Laplant 1989). Sometimes "irregularities" in address information, perhaps caused by clerical typing errors or by recording errors on the part of a census or post-enumeration survey interviewer, result in an inability to parse an address into various components; in these cases, the entire address field (referred to as the "conglomerated address") is used as a matching variable. An available preprocessing program also can be used to convert nicknames to a "standardized" name using a library of names and their common variants (Paletz 1989). A variety of schemes are available for assigning weights based on close agreement between variables, and a procedure is also available for adding or subtracting weight to the composite weight for a record pair when certain combinations of fields are in agreement or disagreement (Winkler 1991).

The experiment consisted of eight "treatment" factors and one "blocking" factor (where "blocking" here refers to the experimental-design notion of a grouping of units expected to yield results as similar as possible in the absence of treatment effects) with replication across three sites in a $2^5 \times 3^3 \times 5 \times 13$ factorial design. The outcome variable in the experiment, described further in Section 3.5, was a transformation of the false-match rate, where the transformation was used to stabilize the variance of the outcome. The factors in the experiment can be described as follows:

Label	Description of factor	Number of levels of factors	Description of levels of factor
A	Assignment of weight for name fields.	5	1. Assign weights of ± 2 for agreement/ disagreement on first, last name.
			2. Assign weights of ± 4 for agreement/ disagreement on first, last name.
			3. Assign weights of ± 6 for agreement/ disagreement on first, last name.
			4. Assign weights based on estimates of probabilities of agreement on first, last name from Fellegi-Sunter algorithm (see Winkler and Thibaudeau 1992).
			5. Use frequency-based weighting for first, last name (see Winkler and Thibaudeau 1992).

Label	Description of factor	Number of levels of factors	Description of levels of factor
B	Assignment of weight for close but inexact agreement on name fields.	3	1. Assign disagreement weight for any discrepancy in first, last name. 2. Assign fraction of agreement weight for close agreement on first, last name using Jaro string comparison metric (Jaro 1989; Winkler 1991). 3. Assign fraction of agreement weight for close agreement on first, last name using piecewise linear metric described in Winkler (1991).
C	Assignment of weight for non-name fields.	2	1. Assign weights of ± 2 for agreement/ disagreement on age, phone number, and address fields, and assign weights of ± 1 for agreement/ disagreement on sex, race, marital status, relationship to head of household, middle initial. 2. Assign weights based on estimates of probabilities of agreement from Fellegi-Sunter algorithm.
D	Assignment of weight for close but inexact agreement on non-name fields.	3	1. Assign disagreement weight for any discrepency in non-name fields. 2. Assign fraction of agreement weight for close agreement on house number, street name, phone number, age, using Jaro string comparator. 3. Assign fraction of agreement weight for close agreement on street name using Jaro string comparator, for age using Jaro pro-rated-to-absolute-difference metric, for house number and phone number using Winkler piecewise-linear string comparator.
E	Use of keyed first name or standardized version of first name.	2	1. Use the version of the individual's first name that was keyed into each data file for comparison of first name. 2. Use the version of the individual's first name that is obtained as output from name standardization software (Paletz 1989).

Label	Description of factor	Number of levels of factors	Description of levels of factor
F	Adjustment of weights for correlated agreement.	2	1. Do not adjust the composite weight for possible correlated agreement. 2. Adjust composite weights for possible correlated agreement between first name, middle initial and among first name, sex, age.
G	Inclusion of marital status, relationship to head of household as matching variables.	2	1. Do not include marital status, relationship as matching variables. 2. Include marital status, relationship as matching variables.
H	Use of four or seven digits of phone number.	2	1. Use only last four digits of phone number as a matching variable. 2. Use all seven digits of phone number.
I	Site of census/post-enumeration survey.	3	1. Eastern Washington state. 2. Columbia, Missouri. 3. St. Louis, Missouri.
J	Proportion of PES file declared matched.	13	1.-13. Let the number of records accepted as declared matches equal 60%, 62.5%, 65%, 67.5%, 70%, 72.5%, 75%, 77.5%, 80%, 82.5%, 85%, 87.5%, 90% of the number of PES records in the given site.

With reference to the sources of variation described in Section 3.1, factors E, G, and H relate to the choice of matching variables; factors A, C, and F relate to the choice of a weighting scheme; factors B and D relate to the handling of close but inexact agreement; factor J reflects the choice of a cutoff; and factor I reflects the influence of the particular site on the performance of the matching procedure.

Consideration of resource limitations led to a decision not to address the effect of varying missing data treatments or the effect of different choices of blocking variables in this experiment, and the lack of available software precluded any investigation of alternative alogrithms for assigning candidate matches. Belin (1989a, 1989b) studied the influence of missing data treatments and of different choices of blocking variables in an experiment similar to the factorial experiment described here. The results of that investigation suggested that alternative treatments of missing data had no substantial effect on false-match rates associated with different cutoffs in matching of census/PES data, but the choice of blocking variables did have a substantial effect.

In this investigation, as in Belin (1989a, 1989b), only "one-pass" matching procedures are considered. That is, the entire computer-matching operation consists of a single cycle of choosing blocking variables, establishing weights, and setting a cutoff, as opposed to "multiple-pass" procedures that first use very restrictive blocking variables to skim off the nearly perfect matches, then relax the blocking criteria in successive passes through the data. The author is aware of very little research on multiple-pass matching procedures. Belin (1989b) reports that when single-pass procedures are used, procedures that use relatively less restrictive blocking criteria enjoy advantages over procedures that use relatively more restrictive blocking criteria, confirming the intuitive notion that the blocking process can exclude true matches from consideration as an unfortunate side effect.

3.3 Subtleties in Experimental Treatments

3.3.1 Treatments for Assigning Weights for Agreement/ Disagreement on Fields of Information

To clarify the experiment, we describe each of the experimental factors in greater detail. Factors A and C are concerned with the assignment of weights for agreement and disagreement on the various matching variables. The different weighting approaches used in factors A and C include completely *ad hoc* methods and methods that are based on estimates of parameters in explicit probability models. The study of *ad hoc* weights provides an opportunity to gauge the importance of incorporating more complicated approaches to weighting.

The *ad hoc* weighting schemes call for a weight of U, say, to be added to the composite weight if the fields being compared agree, and for an identical weight U to be subtracted from the composite weight if the fields being compared disagree. Three different values of U are studied in factor A, with the same value of U being assigned for agreement on first name as for agreement on last name. In factor C, an *ad hoc* scheme that weights some variables more than others is studied, with the decision about which variables to weight more being based on *a priori* judgments. Belin (1989b) suggests that such a "modified-equal-weighting" scheme has advantages over an "equal-weighting" scheme in which all matching variables are assigned the same weights for agreement or disagreement.

The "Fellegi-Sunter algorithm" refers to the method outlined in Fellegi and Sunter (1969), which is based on a probabilistic model that incorporates information about patterns of agreement and disagreement between pairs of records. The model postulates that probabilities of agreement on individual fields of information given that a pair is a true match are independent across all fields of information, and that independence across fields also holds given that a pair is a false match. The paper by Fellegi and Sunter shows that such a model implies certain optimality properties for the type of weighting scheme used by Newcombe *et al.* (1959), in which weights for individual fields of information are calculated by taking the logarithm of the ratio of probability of agreement given true match to the probability of agreement given false match, and in which composite weights are obtained by summing individual field weights.

In applications, the probabilities of agreement given true match and agreement given false match need to be estimated. For the treatments in the experiment characterized as relying on the Fellegi-Sunter weighting approach, the probabilities of agreement given true match are estimated using a version of an EM algorithm (Dempster, Laird and Rubin 1977) to obtain maximum likelihood estimates of these probabilities based on counts of all possible patterns of agreement observed in the data files at hand (Winkler 1989; Jaro 1989). The probabilities of agreement given false match are estimated based on counts of agreement on individual fields between all record pairs that agree on blocking variables, making use of the fact that most of the pairs that could possibly be brought together as matches are not true matches (Winkler and Thibaudeau 1992).

Another weighting approach that has been implemented in the Census Bureau's record linkage software considers the relative frequency of names in the data files at hand, assigning more weight for agreement on names such as Abramowicz, which may be relatively rare, than for agreement on names such as Smith, which may be common. Of course, it could happen that in a particular area Abramowicz is a more common name than Smith, in which case the frequency-based weighting approach would assign greater weight to agreement on the name Smith. The idea of incorporating information on marginal frequencies from the current data files was mentioned by Newcombe *et al.* (1959), and has been noted by many authors since then, including Fellegi and Sunter (1969). (Thus, the distinction drawn here between the "Fellegi-Sunter algorithm" and "frequency-based weighting" is actually a distinction between two methods of calculating weights that are both discussed by Fellegi and Sunter.) Details on the implementation of frequency-based weighting in the Census Bureau's software can be found in Winkler and Thibaudeau (1992).

3.3.2 Treatments for Handling Close but Inexact Agreement

Factors B and D deal with the handling of fields that may agree closely but do not agree exactly with one another. Several techniques have been proposed for handling close but inexact agreement between fields of information, often reflecting different perspectives on probable departures from exact agreement.

The Jaro string comparator is designed to measure the closeness of agreement of two multi-character fields; the metric that defines closeness is a function of the lengths of the character fields in the two files, the number of characters in common between the character fields, and the number of transpositions of characters between the character fields. The weight that gets assigned for partial agreement is between the weight for agreement on the field and the weight for disagreement on the field, and is a linear function of the string comparator metric between the agreement weight and the disagreement weight.

The Winkler piecewise-linear approach uses the same metric as the Jaro string comparator to define closeness of agreement, but the rate at which partial agreement weights decrease from the agreement weight to the disagreement weight is a piecewise linear function of the string comparator metric, requiring two user-supplied rate parameters and two user-supplied thresholds where the slope changes.

The Jaro pro-rated method assigns a weight between the agreement weight and the disagreement weight based on the absolute value of the difference between two numeric fields. As with the aforementioned techniques, the partial agreement weight falls off as a linear function of the absolute value of the difference.

Even for some numeric fields (*e.g.*, telephone number), a comparison method designed to accommodate slight typographical variation would seem more sensible than a method based on absolute numerical difference. However, for variables such as year of birth or age, it may not be clear whether to target efforts toward accommodating typographical errors (for which a string comparison method would be best suited), reporting errors (for which the absolute-difference method may be most appropriate), or other types of errors such as "heaping" or rounding of reported ages on multiples of five years (for which neither of the previously mentioned comparison methods would be ideally suited). Accordingly, we pursue our empirical evaluations in an attempt to shed light on these issues.

3.3.3 Treatments Involving the Choice of Matching Variables

As mentioned previously, an approach has been developed at the Census Bureau for converting nicknames to a standardized root. Software developed by Paletz (1989) implements the name-standardization routine.

The treatment that omits marital status and relationship to head of household as matching variables allows for an assessment of the importance of two background demographic variables on the quality of matching. Chernoff (1980) develops theory for the information carried by a matching variable and shows that a variable recorded in error even a small percentage of the time can lose a substantial amount of information for matching purposes (*e.g.*, the Kullback-Leibler information associated with a binary variable recorded in error three percent of the time is only about half that of a binary variable recorded without error). Considering that relationship to head of household could differ between the census and PES if the person listed as the head of household is different, and that marital status will change for some individuals in the intervening time, it is not clear in advance how much information for matching is provided by these variables. On the other hand, it is hard to imagine that using additional matching variables would be deleterious, so that this treatment provides a standard for assessing the practical significance of some of the other treatments.

The treatment of using either four or seven digits of phone number as a matching variable is self-explanatory. A motivation for considering this treatment is that one of the specific piecewise-linear string comparator methods proposed by Winkler was developed based on analysis of the last four digits of phone number as a matching variable.

3.3.4 Treatment for Adjusting Composite Weights for Correlated Agreement

The method described as adjusting the composite weight to reflect the possibility of correlated agreement is also due to Winkler and is described in Winkler and Thibaudeau (1992). Research by Kelley (1986) and Thibaudeau (1989) reveals that agreement on the various fields available for matching between the census and PES data files is far from being independent across fields. In particular, analyses suggested that agreement on first name was correlated with agreement on middle initial and that agreement on first name, age, and sex were mutually correlated. These findings led to the implementation of modifications to the composite weight when certain patterns appear (*e.g.*, if first name, age, and sex all disagree, then a large value is subtracted from the composite weight). The current scheme for adjusting the composite weight is entirely *ad hoc*; research into methods that reflect correlated agreement still appears to be in its infancy.

3.4 Data Files Used in Experiment

As mentioned before, the three sites of the 1988 dress rehearsal census and post-enumeration survey provided separate data files on which these analyses of record linkage could be performed. There were 12,072 records in the PES file from St. Louis, 6,581 records in the PES file from East Central Missouri, and 2,782 records in the PES file from eastern Washington state. As was also noted earlier, the final determinations by clerks who reviewed these files were taken as the truth for purposes of evaluation. Other test censuses were conducted during the 1980's; the primary reason for not including the data from other test censuses in this experiment is that a considerable amount of "overhead" time is required to prepare a data set for the analyses performed here.

3.5 Outcome Variable

The primary outcome variable considered in this experiment was a transformation of the false-match rate. The false-match rate is defined as the number of false matches divided by number of declared matches, and is a common measure of performance in the literature on record linkage (*e.g.*, Fellegi and Sunter (1969) attempt to provide output that satisfies a fixed false-match rate criterion supplied by the operator of the program). In order to stabilize the variance of the outcome, the analyses here use the arcsine of the square root of the false-match rate as an outcome variable.

3.6 Choice of Cutoff Weight as a Blocking Factor

It is clear that the false-match rate in record linkage is apt to depend heavily on the choice of a cutoff between declared matches and declared non-matches. Accordingly,

a blocking factor (Factor J) is introduced to fix the determination of cutoffs so as to facilitate comparison of other record-linkage treatments. To provide a standard for comparisons across sites having different numbers of records, the cutoff level is defined in terms of the proportion of the PES data file declared matched.

Because of the discreteness of record-linkage weights, it is possible to have ties among the weights of record pairs on the boundary where the cutoff should be assigned. For example, in a file of 10,000 records, there may be 40 records with weight W (of which 10 may be false matches), 7,980 records with weight greater than W (of which 3 may be false matches), and 1,980 records with weight less than W. If the treatment in factor J calls for 80% of the PES file to be matched, then it may not be obvious how to calculate the false-match rate, since there are 40 records with the same weight straddling the point where the cutoff should be set. Calculations of the false-match rate in such a case are based on the following relationship:

$$\text{fmr} = \frac{f_{\text{abv}} + \frac{f_{\text{bdy}}}{n_{\text{bdy}}} \times (n_{\text{cut}} - n_{\text{abv}})}{n_{\text{cut}}},$$

where fmr denotes false-match rate, f_{abv} is the number of false matches and n_{abv} the number of declared matches with weights above the cutoff weight, f_{bdy} is the number of false matches and n_{bdy} the number of declared matches with weights equal to the boundary cutoff weight, and n_{cut} is the number of declared matches needed to satisfy the condition that a certain percentage of the PES data file be declared matched. If we were to calculate the false-match rate by randomly selecting the appropriate number of boundary records to satisfy the cutoff criterion, then the expression above would give the expected false-match rate over repetitions of such a procedure; thus, the logic behind this definition is clear.

In the example above, one fourth of the boundary cases are false matches, and twenty additional records are needed to satisfy the stipulation that 80% of the file be declared matched. Effectively five false matches are added to the three among the records among the pairs with weights above the cutoff weight, giving a false-match rate of $(3 + 0.25(40 - 20))/8{,}000 = 8/8{,}000 = 0.001$.

3.7 Further Considerations Relevant to the Analysis of Experimental Results

Analysis of the experimental results proceeded from the standpoint that general indications of significance are more important than precise p-values, especially because the experiment itself is exploratory. Belin (1991) points out that appropriate methods for assessing significance from these data are somewhat complicated; this is because site should be thought of as a random factor (since we would like to generalize about treatment effects from the sample of three sites to a population of many possible sites), but standard procedures that use the site by treatment interaction as the error term for a particular treatment suffer from low power given the small number of available sites. Belin (1991) uses the Johnson-Tukey display-ratio plot (Johnson and Tukey 1987), which is a close relative of the half-normal plot of Daniel (1959), to estimate underlying noise levels in assessing the significance of effects. In this paper, we do not attempt to present formal significance findings.

4. RESULTS

4.1 ANOVA Breakdown of Experimental Results

We begin by breaking down the results of the factorial experiment into an analysis of variance, distinguishing treatment effects, site effects, cutoff effects, and their interactions from one another, grouping effects of the same order. Table 4.1 is an excerpt from the complete ANOVA breakdown of the experiment, showing treatment interactions up to four-way along with corresponding error terms.

F-statistics are calculated dividing the mean square for the given effect by the mean square for the effect-by-site interaction term. Thus, for example, the F-statistic for three-way interactions among treatments is calculated as $0.0120/0.00470 = 2.551$, with the denominator coming from the line for the four-way treatment-by-site interaction.

If the F-statistics are interpreted in the usual way, then statistical significance at the 0.0001-level is achieved for all of the F-statistics reported in Table 4.1 except the treatment-by cutoff four-way interactions; however, caution should be used in interpreting these results. First, the magnitudes of the various mean-square terms suggest that the higher-order effects are not of substantial practical importance. Further, the comparison of the F-statistics calculated above to a reference F-distribution relies on certain exchangeability assumptions (*e.g.*, that site-to-site variability in main effects is the same for all main effects) that are not necessarily well-founded. For example, it may not make sense to pool site-to-site variability in the effect of four versus seven digits of phone number with site-to-site variability in the effect of the different weighting schemes in estimating an error term for main effects.

4.2 Importance of Choice of Cutoff as Compared to Other Controllable Factors

It is evident (*e.g.*, from the mean squares for main effects) that site-to-site variability and variability due to the choice of a cutoff are considerably larger than the variability explained by differences in treatments. Although

Table 4.1

Excerpt from ANOVA Breakdown of Factorial Experiment, Grouping Effects of the Same Order

Source	*df*	Sums of squares	Mean square	*F*
Site main effects	2	35.195	17.598	
Treatment main effects	13	30.917	2.378	10.570
Cutoff main effects	12	147.515	12.293	7.548
Treatment/site 2-way interactions	26	5.850	0.225	
Cutoff/site 2-way interactions	24	39.089	1.629	
Treatment/treatment 2-way ints	70	6.992	0.100	4.041
Treatment/cutoff 2-way ints	156	1.410	0.009	3.553
Treatment/site 3-way interactions	140	3.461	0.0247	
Cutoff/treatment/site 3-way ints	312	0.794	0.0025	
Treatment 3-way interactions	206	2.472	0.0120	2.551
Treatment/cutoff 3-way ints	840	0.530	0.0006	1.866
Treatment/site 4-way interactions	412	1.938	0.00470	
Cutoff/treatment/site 4-way ints	1,680	0.568	0.00034	
Treatment 4-way interactions	365	0.747	0.00205	2.365
Treatment/cutoff 4-way ints	2,472	0.267	0.00011	0.236
Treatment/site 5-way interactions	730	0.632	0.00087	
Cutoff/treatment/site 5-way ints	4,944	0.226	0.00046	
...				
Total	**56,159**	**279.169**		

this result may be explained in part by the fact that some treatments are very close to one another (*e.g.*, using four digits versus seven digits of phone number), it is nevertheless the case that some of the qualitative differences between treatments are quite substantial (*e.g.*, leaving out two matching variables versus keeping them in). The ANOVA breakdown also highlights the fact that we can expect substantial site-to-site variability in false-match rates. In their approach to calibrating record-linkage procedures, Belin (1991) and Belin and Rubin (1991) explicitly accommodate site-to-site variability in providing estimates of false-match rates corresponding to different cutoffs.

4.3 The Main Effects of Treatments

In Table 4.2, we give the mean of the outcome variable observed for each level of the treatment factors. Since arcsine (x) is a monotone increasing function of x, lower values of the outcome signify lower false-match rates and thus better performance.

Table 4.2

Marginal Values of arcsine($\sqrt{\text{fmr}}$) for each Level of Experimental Treatments Averaged over all other Experimental Conditions

Factor	A	(name wts)	Factor	B	(inexact agree, name wts)	Factor	C	(non-name wts)
Level	1	0.106	Level	1	0.113	Level	1	0.101
	2	0.096		2	0.094		2	0.101
	3	0.093		3	0.095			
	4	0.130						
	5	0.079						
Factor	D	(inexact agree, non-name wts)	Factor	E	(Standardize name)	Factor	F	(Adjust for correlated agree)
Level	1	0.111	Level	1	0.102	Level	1	0.106
	2	0.108		2	0.100		2	0.095
	3	0.084						
Factor	G	(Include marit/rel)	Factor	H	(Four or seven digits phone #			
Level	1	0.103	Level	1	0.102			
	2	0.098		2	0.100			

Belin (1991) breaks down the experimental findings into a set of complementary orthogonal contrasts. The largest main-effect contrasts among those prespecified by Belin (1991) were those between frequency name weights (A = 5) and Fellegi-Sunter name weights (A = 4), between Winkler's string comparators on non-name fields (D = 3) and Jaro's corresponding string comparators (D = 2), between some string comparator for names (B = 2 or 3) and no string comparator for names (B = 1), between some string comparator for non-name fields (D = 2 or 3) and no string comparator for these fields (D = 1), and between performing an adjustment for correlated agreement (F = 2) and not performing such an adjustment (F = 1).

4.4 Two-Way Treatment Interactions

The largest two-way treatment interaction contrast among those reviewed by Belin (1991) was the F × G effect, which is the interaction of performing an adjustment for correlated agreement (among first name and middle initial and among first name, age, and sex) with including or not including marital status and relationship to head of household as matching variables. This contrast was statistically significant according to any of the procedures used in Belin (1991) for estimating a background noise level. We show the average levels of the outcome across the four treatment combinations above in Table 4.3.

Table 4.3
Average Performance for Combinations of F and G Treatments

F	G	False-match rate	Arcsine($\sqrt{\text{fmr}}$)
1	1	0.0182	0.116
1	2	0.0143	0.097
2	1	0.0128	0.091
2	2	0.0151	0.100

This result suggests that the adjustment for correlated agreement (level 2 of factor F) helps a great deal when marital status and relationship are not included as matching variables (level 1 of factor G), but the adjustment for correlated agreement does not help on average when marital status and relationship are included as matching variables. That we are able to identify this type of effect emphasizes the importance of pursuing empirical evaluations in an experimental framework.

The next two largest two-way treatment interaction contrasts cited by Belin (1991) after the F × G interaction comprise part of the A × B interaction (involving the choice of name weights and the choice of string comparisons to use for name fields). We show the average results for all of the combinations of treatments for factors A and B below as Table 4.4.

Table 4.4
Average Performance for Combinations of A and B Treatments

A	B	False-match rate	Arcsine ($\sqrt{\text{fmr}}$)
1	1	0.0192	0.120
1	2	0.0140	0.099
1	3	0.0143	0.100
2	1	0.0170	0.110
2	2	0.0120	0.087
2	3	0.0123	0.089
3	1	0.0177	0.113
3	2	0.0118	0.084
3	3	0.0119	0.083
4	1	0.0254	0.145
4	2	0.0193	0.123
4	3	0.0189	0.122
5	1	0.0109	0.079
5	2	0.0109	0.079
5	3	0.0109	0.078

Thus, we find that when we use frequency-based name weights (A = 5), it hardly matters whether we use any string comparison method, but when we use *ad hoc* name weights or Fellegi-Sunter name weights, the use of string comparison methods substantially improves the average performance of the computer-matching procedure.

We highlight some of the other interesting findings noted in Belin (1991) based on exploring the largest two-way treatment interaction effects:

(1) The Winkler approach to inexact agreement on non-name variables (*i.e.*, D = 3), which is the best treatment on average for factor D, has more of a helpful effect on average when marital status and relationship to head of household are included as matching variables (*i.e.*, G = 2), even though the latter variables are not included in any of the treatments for handling inexact agreement.

(2) Unlike the other treatments for name weights, which appear to be helped by the inclusion of marital status and relationship, frequency-based name weighting appears to be adversely affected by the inclusion of these variables.

(3) *Ad hoc* weights of ±6 for agreement on name perform better on average when combined with the *ad hoc* weighting approach to non-name variables; *ad hoc* name weights of ±4 and ±2 work better with the weights assigned by the Fellegi-Sunter algorithm to non-name variables.

(4) Without the adjustment for correlated agreement, Fellegi-Sunter weights for non-name variables worked better for these data than *ad hoc* weights, but the *ad hoc* weights worked better when the adjustment for correlated agreement was included. (However, based on the method of estimating the background noise level described in Belin (1991), this phenomenon should not necessarily be expected to carry over to other sites.)

4.5 Which Treatment Combination Works Best?

To wrap up the analysis of the experimental results, we consider now the question of which treatment combination works best. To measure the performance for a given treatment combination, we take the average outcome from using that procedure across the three available sites. The outcomes we examine are the false-match rates corresponding to 60%, 62.5%, 65%, 67.5%, 70%, 72.5%, 75%, 77.5%, 80%, 82.5%, 85%, 87.5%, and 90% of the PES file declared matched. The results from the experiment are provided in Table 4.5.

Table 4.5

Best Treatment Combination for each of Thirteen Cutoffs from Factorial Experiment

Cutoff level	Levels of factors in best treatment combination (A B C D E F G H)								False-match rate for best treatment combination averaged over three sites
60% matched	3	3	2	3	2	1	1	1	0.00042
62.5% matched	3	3	1	3	1	1	2	1	0.00047
65% matched	3	3	1	3	2	2	2	2	0.00052
67.5% matched	3	3	2	3	2	2	1	1	0.00071
70% matched	2	3	2	3	1	2	1	2	0.00079
72.5% matched	5	2	2	3	1	1	1	2	0.00081
75% matched	5	1	1	3	2	1	2	1	0.00112
77.5% matched	3	3	1	3	2	1	2	1	0.00133
80% matched	2	3	2	3	1	2	1	2	0.00188
82.5% matched	3	3	1	3	2	2	1	1	0.00571
85% matched	5	1	2	3	2	2	1	2	0.01556
87.5% matched	2	3	2	3	1	2	1	2	0.03023
90% matched	2	3	2	3	1	2	1	2	0.05174

These results contrast with the earlier result suggesting that frequency-based weighting for names (level 5 for factor A) is better on average than using *ad hoc* name weights of ±6 (level 3 for factor A). Apparently, the reason that the latter is worse on average is due to certain interaction effects. When the *ad hoc* weighting approach is combined with the appropriate levels of other factors, it appears to perform at least as well as the frequency-weighting approach. We also note that the best combination of factors F and G is not always treatments 2 and 1, respectively, despite our earlier finding that this treatment combination for these two factors performs best on average. Only treatment 3 of factor D (using Winkler modifications in handling inexact agreement on non-name variables) is an unequivocal choice for the best treatment no matter how we measure the outcome of the experiment. The choice for the best treatment for name weights is between deterministic weights of ±6 or ±4 and the frequency name-weighting approach. If one of the deterministic weighting schemes is used, the Winkler approach to string comparisons for names is to be recommended; with frequency name weights, it is not clear that any string comparison approach should be used on names.

Between Fellegi-Sunter weights for non-name variables and *ad hoc* weights, the choice is not obvious, but earlier analysis suggested that the effect either way is small. Similar remarks apply to the choice of whether to use standardized or unstandardized first names and to the choice of whether to use four or seven digits of the phone number.

Considering the fact that there is not a single treatment combination that is uniformly superior to all other treatment combinations, one might look to the performance of different treatment combinations in a particular region of interest (*e.g.*, where the false-match rate is around 0.001). However, if we look at the best treatment combinations in the region where 70%-80% of the PES file is declared matched (*i.e.*, restricting attention to five cutoffs), we still find no obvious choice for a preferred treatment combination. Averaged across those five cutoffs, the best treatment combination is (2,3,2,3,1,2,1,2); that is, using name weights of ±4, incorporating Winkler's modifications to inexact agreement on name, estimating weights using the Fellegi-Sunter algorithm for non-name variables, using Winkler's approach to inexact agreement for non-name variables, using the original unstandardized version of first name, adjusting the composite weight for correlated agreement, not including marital status and relationship to head of household as matching variables, and using all seven digits of phone number.

For comparison, we display in Table 4.6 the average performance of some of the other candidates for best treatment combination. Thus it appears that the best alternatives to (2,3,2,3,1,2,1,2) are treatment combinations (3,3,1,3,2,2,2,2) and (3,3,1,3,2,1,2,1). Both of these procedures feature name weights of ±6, predetermined

Table 4.6

Average False-match Rates for Different Treatment Combinations Across Three Sites and across Five Cutoff Levels (70%, 72.5%, 75%, 77.5%, and 80% of PES File Declared Matched)

Levels of factors in treatment combination (A B C D E F G H)	Average false-match rate across sites and across cutoffs with 70%, 72.5%, 75%, 77.5%, and 80% of PES file declared matched
3 3 2 3 2 1 1 1	0.00493
3 3 1 3 1 1 2 1	0.00154
3 3 1 3 2 2 2 2	0.00137
3 3 2 3 2 2 1 1	0.00161
2 3 2 3 1 2 1 2	0.00124
5 2 2 3 1 1 1 2	0.00191
5 1 1 3 2 1 2 1	0.00153
3 3 1 3 2 1 2 1	0.00138
3 3 1 3 2 2 1 1	0.00156
5 1 2 3 2 2 1 2	0.00155

ad hoc weights for non-name variables, Winkler's approaches to inexact agreement for both name and non-name variables, standardized first names, and inclusion of marital status and relationship as matching variables. These treatment combinations differ from each other in that one includes an adjustment of the composite weight for correlated agreement and calls for using seven digits of phone number, whereas the other features no adjustment of weights for correlated agreement and only four digits of phone number. The treatment combinations involving the use of frequency-based name weighting do not perform as well as the best treatment combinations using *ad hoc* name weights according to this standard.

In the 1990 PES, the treatment combination that was used in computer-matching operations was very close to treatment combination (5,3,2,3,2,2,2,1). In the test-census data sets studied here, this treatment combination produced an average false-match rate across the five cutoffs of 0.00179.

4.6 Concluding Remarks

While the results in this paper address the tradeoff between the number of records declared matched and false-match rates, an anonymous referee noted that "every gain which is achieved by a superior record linkage procedure must be justified by the cost of implementing that procedure." This is another tradeoff that any practitioner can appreciate. Hopefully, the findings presented here about the relative importance of various factors in record linkage will provide some guidance to those who develop and implement linkage software. Because some of the results may depend on specific features of the census/PES data being matched, there may be some question as to how these results relate to other record-linkage settings. But as was emphasized at the outset, one practical recommendation that does generalize across data settings is the call for taking an experimental approach to the study of record linkage. Empirical study through designed experiments is a tried and true source of guidance, offering a clear framework for adding to the accumulated insights of record-linkage specialists.

ACKNOWLEDGMENTS

Much of this work was done while the author was working for the Record Linkage Staff of the U.S. Bureau of the Census in Washington, D.C. The author gratefully acknowledges helpful discussions and comments from Don Rubin, Bill Winkler, Alan Zaslavsky, and an anonymous referee, as well as earlier support from JSA 88-02 and JSA 89-07 while the author was a doctoral candidate at Harvard University.

REFERENCES

ABBATT, J.D. (1986). A cohort study of eldorado uranium workers. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 51-57.

ACHESON, E.D. (1967). *Medical Record Linkage*. Oxford: Oxford University Press.

ACHESON, E.D. (Ed.) (1968). *Record Linkage in Medicine*, Edinburgh: E. & S. Livingstone.

BALDWIN, J.A., ACHESON, E.D., and GRAHAM, W.J. (Eds.) (1987). *A Textbook of Medical Record Linkage*. Oxford: Oxford University Press.

BELIN, T.R. (1989a). Outline of procedure for evaluating computer matching in a factorial experiment. Unpublished memorandum, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BELIN, T.R. (1989b). Results from evaluation of computer matching. Unpublished memorandum, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BELIN, T.R. (1991). Using mixture models to calibrate error rates in record-linkage procedures, with application to computer-matching for census undercount estimation. Ph.D. thesis, Department of Statistics, Harvard University. (Published by University Microfilms, Inc.)

BELIN, T.R., and RUBIN, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C.

BOYLE, C.A., and DECOUFLÉ, P. (1990). National sources of vital status information: Extent of coverage and possible selectivity in reporting. *American Journal of Epidemiology*, 131, 160-168.

BROWN, P., LAPLANT, W., LYNCH, M., ODELL, S., THIBAUDEAU, Y., and WINKLER, W. (1988). Collective Documentation for the 1988 PES Computer Match Processing and Printing. Vols. I-III, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

BUREAU OF THE CENSUS (1988-1991). 1990 Decennial Census Information Memorandum Series, Decennial Planning Division, Bureau of the Census, Washington, D.C.

[Note: To all of the reports in the aforementioned memorandum series, the following statement is attached:

"These overviews are prepared for use by planning and operating divisions within the Census Bureau who are conversant with the background, previous experiences, terminology, and processes, as well as with the overall framework of the decennial census design, goals, and inter-relationships of operations and systems. They are NOT [emphasis in original] intended or appropriate for external distribution and should not be sent outside the Census Bureau without prior approval from Jim Dinwiddie ([301]-763-5270) of the Decennial Planning Division."]

BUREAU OF THE CENSUS (1987-1991). STSD Decennial Census Memorandum Series, Statistical Support Division, U.S. Bureau of the Census, Washington, D.C.

CARPENTER, M., and FAIR, M.E. (Eds.) (1990). Canadian Epidemiology Research Conference - 1989: *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Centre for Health Information, Statistics Canada, Ottawa, Ontario.

CHERNOFF, H. (1980). The identification of an element of a Large population in the presence of noise. *Annals of Statistics*, 8, 1179-1197.

CHILDERS, D. (1989). 1990 PES Within Block Matching - Clerical Matching Group. STSD Decennial Census Memorandum Series #V-69, U.S. Bureau of the Census, Washington, D.C.

CITRO, C.F., and COHEN, M.L. (Eds.) (1985). *The Bicentennial Census: New Directions for Methodology in 1990*, Washington, D.C.: National Academy Press.

COHEN, M.L. (1990). Adjustment and reapportionment - Analyzing the 1980 decision. *Journal of Official Statistics*, 6, 241-250.

COOMBS, J.W., and SINGH, M.P. (Eds.) (1988). *Proceedings of the Symposium on Statistical Uses of Administrative Data*. Statistics Canada, Ottawa, Ontario.

COPAS, J., and HILTON, F. (1990). Record Linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society*, Series A, 153, 287-320.

CURB, J.D., FORD, C.E., PRESSEL, S., PALMER, M., BABCOCK, C., and HAWKINS, C.M. (1985). Ascertainment of vital status through the National Death Index and the Social Security Administration. *American Journal of Epidemiology*, 121, 754-766.

DANIEL, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1, 311-341.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

DONOGHUE, G. (1990). Clerical Specifications for the 1990 Post Enumeration Survey Before Followup Matching - Special Matching Group. STSD Decennial Census Memorandum Series #V-92, U.S. Bureau of the Census, Washington, D.C.

DULBERG, C.S., SPASOFF, R.A., and RAMAN, S. (1986). Reactor clean-up and bomb test exposure study. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 59-62.

ERICKSEN, E.P., and KADANE, J.B. (1985). Estimating the population in a census year: 1980 and Beyond (with discussion). *Journal of the American Statistical Association*, 80, 98-131.

ERICKSEN, E.P., KADANE, J.B., and TUKEY, J.W. (1989). Adjusting the 1980 Census of Population and Housing. *Journal of the American Statistical Association*, 84, 927-944.

FAGERLIND, I. (1975). *Formal Education and Adult Earnings: A Longitudinal Study on the Economic Benefits of Education*, Stockholm: Almquist and Wiksell.

FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

FREEDMAN, D.A., and NAVIDI, W.C. (1986). Regression models for adjusting the 1980 census (with discussion). *Statistical Science*, 1, 1-39.

GOLDACRE, M.J. (1986). The Oxford record linkage study: Current position and future prospects. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press, 106-129.

HILL, T. (1981). Generalized Iterative Record Linkage System: GIRLS. (Glossary, Concepts, Strategy Guide, User Guide), Systems Development Division, Statistics Canada, Ottawa, Ontario.

HILL, T., and PRING-MILL, F. (1986). Generalized iterative record linkage system: GIRLS, (revised edition). Systems Development Division, Statistics Canada, Ottawa, Ontario.

HOGAN, H. (1992). The 1990 Post-Enumeration Survey: An Overview. *The American Statistician*, 46, 261-269.

HOWE, G.R., and LINDSAY, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers in Biomedical Research*, 14, 327-340.

HOWE, G.R., and SPASOFF, R.A. (Eds.) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto: University of Toronto Press.

JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: Methodological issues. *Journal of Official Statistics*, 2, 255-277.

JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

JOHANSEN, H.L. (1986). Record linkage of national surveys: The Nutrition Canada example. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 153-163.

JOHNSON, E.G., and TUKEY, J.W. (1987). Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao Data. In *Design, Data, and Analysis*, (Ed. C.L. Mallows) New York: John Wiley and Sons.

JOHNSON, R.A. (1991). Methodology for Evaluating Errors in U.S. Department of Justice Attorney Workload Data. Unpublished technical report, General Accounting Office, Washington, D.C.

KELLEY, R.P. (1986). Robustness of the Census Bureau's record linkage system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 620-624.

KERSHAW, D., and FAIR, J. (1979). *The New Jersey Income and Maintenance Experiment: Operations, Surveys, and Administration*, Volume I. New York: Academic Press.

KILSS, B., and ALVEY, W. (Eds.) (1984a). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. I, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and ALVEY, W. (Eds.) (1984b). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Vol. II, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and ALVEY, W. (Eds.) (1984c). *Statistics of Income and Related Administrative Record Research:* 1984, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and ALVEY, W. (Eds.) (1985). *Record Linkage Techniques - 1985*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and ALVEY, W. (Eds.) (1987). *Statistics of Income and Related Administrative Record Research: 1986-1987*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and JAMERSON, B. (Eds.) (1990). *Statistics of Income and Related Administrative Record Research 1988-1989*, Statistics of Income Division, Internal Revenue Service, Washington, D.C.

KILSS, B., and SCHEUREN, F. (1978). The 1973 CPS-IRS-SSA Exact Match Study. *Social Security Bulletin*, Vol. 41, 10, 14-22.

LAPLANT, W. (1988). User's Guide for the Generalized Record Linkage Program Generator (GENLINK). Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

LAPLANT, W. (1989). User's Guide for the Generalized Address Standardizer (GENSTAN). Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.

NEWCOMBE, H.B., FAIR, M.E., and LALONDE, P. (1992). The use of names for linking personal records (with discussion). *Journal of the American Statistical Association*, 87, 1193-1208.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

NEWCOMBE, H.B., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., and ABBATT, J.D. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado Uranium Workers. *Computers in Biology and Medicine*, 13, 157-169.

NICHOLL, J.P. (1986). The use of hospital in-patient data in the analysis of the injuries sustained by road accident casualties. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds. G.R. Howe and R.A. Spasoff). Toronto: University of Toronto Press, 243-244.

PALETZ, D. (1989). Name standardization software. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

ROGOT, E., SORLIE, P.D., and JOHNSON, N.J. (1986). Probabilistic methods in matching census samples to the National Death Index. *Journal of Chronic Disease*, 39, 719-734.

ROGOT, E., SORLIE, P.D., JOHNSON, N.J., GLOVER, C.S., and TREASURE, D.W. (1988). A Mortality Study of One Million Persons. Public Health Service, National Institutes of Health, Washington, D.C.

SCHIRM, A.L., and PRESTON, S.H. (1987). Census undercount adjustment and the quality of geographic population distributions (with discussion). *Journal of the American Statistical Association*, 82, 965-990.

SMITH, M.E., and NEWCOMBE, H.B. (1975). Methods for computer linkage of hospital admission-separation records for cumulative health histories. *Methods of Information in Medicine*, 14, 118-125.

TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.

THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Section on Statistical Computing, American Statistical Association* 283-288.

WENTWORTH, D.N., NEATON, J.D., and RASMUSSEN, W.L. (1983). An evaluation of the Social Security Administration Master Beneficiary Record File and the National Death Index in the ascertainment of vital status. *American Journal of Public Health*, 73, 1270-1274.

WILLIAMS, B.C., DEMITRACK, L.B., and FRIES, B.E. (1992). The accuracy of the National Death Index when personal identifiers other than Social Security Number are used. *American Journal of Public Health*, 82, 1145-1147.

WINKLER, W.E. (1985a). Preprocessing of lists and string comparison. In *Record Linkage Techniques - 1985*, (Eds. W. Alvey and B. Kilss). Internal Revenue Service, Publication 1299 (2-86), 181-187.

WINKLER, W.E. (1985b). Exact matching lists of businesses: blocking, subfield identification, and Information Theory. In *Record Linkage Techniques - 1985*, (Eds. W. Alvey and B. Kilss). Internal Revenue Service, Publication 1299 (2-86), 227-241.

WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter Model of record linkage. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 145-155.

WINKLER, W.E. (1991). Documentation of record-linkage software. Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

WINKLER, W.E., and THIBAUDEAU, Y. (1992). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

WOLTER, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.

Survey Methodology, June 1993
Vol. 19, No. 1, pp. 31–38
Statistics Canada

The Discrimination Power of Dependency Structures in Record Linkage

YVES THIBAUDEAU[1]

ABSTRACT

A record-linkage process brings together records from two files into pairs of two records, one from each file, for the purpose of comparison. Each record represents an individual. The status of the pair is a "matched pair" status if the two records in the pair represent the same individual. The status is an "unmatched pair" status if the two records do not represent the same individual. The record-linkage process is governed by an underlying probabilistic process. A record-linkage rule infers the status of each pair of records based on the value of the comparison. The pair is declared a "link" if the inferred status is that of a matched pair, and it is declared a "non-link" if the inferred status is that of an unmatched pair. The discrimination power of a record-linkage rule is the capacity of the rule to designate a maximum number of matched pairs as links, while keeping the rate of unmatched pairs designated as links to a minimum. In general, to construct a discriminatory record-linkage rule, some assumptions must be made on the structure of the underlying probabilistic process. In most of the existing literature, it is assumed that the underlying probabilistic process is an instance of the conditional independence latent class model. However, in many situations, this assumption is false. In fact, many underlying probabilistic processes do not exhibit key properties associated with conditional independence latent class models. The paper introduces more general models. In particular, latent class models with dependencies are studied and it is shown how they can improve the discrimination power of particular record-linkage rules.

KEY WORDS: Record-linkage rule; Latent class model; Expectation-Maximization procedures.

1. INTRODUCTION

The goal of the paper is to show how record-linkage rules can gain in discriminatory power when probabilistic models more descriptive of the underlying probabilistic processes, are elicited. For this purpose, a particular record-linkage situation is chosen and the conditional independence model, traditionally used in record linkage, is compared to a more descriptive model, in the sense that the new model allows for the expression of more complex relations of dependency between some of the variables involved.

First some terminology must be reviewed. In section 2, the definition of record-linkage process is stated and a general formulation of the probabilistic process underlying a record-linkage process is given. This formulation leads to the expression of two central concepts: the concepts of record-linkage rule and that of most discriminatory record-linkage rule.

In section 3, probabilistic models for record linkage are considered. In the first part of section 3, the family of latent class models is introduced and it is shown how this family provides natural models for the probabilistic process underlying a record-linkage process. In the second part of the section, the focus is on a particular model in the family of latent class models: the latent class model with conditional independence. This model is of interest because it is easy to handle computationally. In the third part, inference techniques adapted to the conditional independence model are reviewed.

In section 4, an application is presented. For this application, truth and falsehood are available, that is, it is known which pairs are matched and which aren't. The first part describes how the information on truth and falsehood was obtained. The second part shows how dependencies between the comparison fields are generated. In the third part of section 4, the knowledge on truth and falsehood is used to evaluate the dependencies between the comparison fields. This leads in the fourth part to the formulation of a model more descriptive of the underlying probabilistic structure of the record-linkage process. The final part is a brief discussion regarding the techniques of parameter estimation for generalized latent class models.

In section 5, an alternative methodology to construct approximate probabilistic models is presented. The model produced by this methodology is compared to those introduced in sections 3 and 4, in terms of discrimination power of the record-linkage rules derived from the models. The results of the comparisons are reported in section 6. In section 7, the suggestions of an anonymous referee to improve the methodology of the paper are presented. In section 8, conclusions are drawn and guidelines are provided.

[1] Yves Thibaudeau, U.S. Bureau of the Census, Federal Bldg. 4, Room 3000, Washington, D.C. 20233.

2. THE FELLEGI-SUNTER MODEL FOR RECORD-LINKAGE

2.1 Record-Linkage Processes

The paper is geared toward building new record-linkage techniques. Before expanding on new record-linkage techniques, some background is necessary. The concept of record-linkage process first needs to be reviewed. Consider two files; file A and file B, both containing records, each record representing an individual. A record-linkage process brings together one record from file A with one record from file B. The records are compared, producing the comparison pattern γ. For the purpose of this paper, this comparison pattern is a vector $\gamma = [\gamma^1, \ldots, \gamma^N]$, where N is the dimensionality of the vector. Each dimension corresponds to a comparison field recorded for each individual, such as last name, age, address, *etc.* With no loss of generality, γ^i is assigned the value 0 if the records disagree over comparison field i and it is assigned 1 if they agree. The comparison space Γ is assumed to be the set of all binary vectors (*i.e.* whose components are 0 or 1) of dimension N.

2.2 Underlying Probabilistic Processes

A record-linkage process is governed by an underlying probabilistic process. A good knowledge of the probabilistic process is needed to extract information from the record-linkage process. The formulation of the underlying probabilistic process is presented here in general terms. It is made more specific in the next section.

Consider a particular comparison pattern γ, define $m(\gamma)$ as the probability of observing γ, given that the two records producing γ, when brought together, represent the same individual. Similarly, define $u(\gamma)$ as the probability of observing γ, given that the two records producing γ, when brought together, do not represent the same individual. These two conditional probabilities, along with the probability of a match, define the underlying probabilistic process. The probabilistic process drives the record-linkage process. $m(\gamma)$ and $u(\gamma)$ are fundamental in the construction of record linkage rules; in particular most discriminatory rules. Record-linkage rules are devices to retrieve matches. They are defined next.

2.3 Record-Linkage Rules

In practice, a record-linkage rule classifies the pairs generated by a record-linkage process in one of three possible categories: a link, a non-link or a possible link. A link is an inferred matched pair and a non-link is an inferred unmatched pair. The pairs classified as possible links are set aside for further examination and eventually they are reclassified as links or non-links. The rule is based only on the value of the comparison vectors corresponding to each pair. The errors induced by a record-linkage rule are of two types: the type I error measuring the proportion of unmatched pairs among the pairs classified as links under the linkage rule, and the type II error measuring the proportion of matches among the pairs classified as non-links.

The objective of record-linkage, from the standpoint of the paper, is to construct a most discriminatory record-linkage rule; that is one that will retrieve a maximum number of links while keeping the type I error under control. To accomplish this, let the comparison patterns be indexed according to decreasing value of $m(\gamma)/u(\gamma)$ to obtain the sequence $\{\gamma_1, \gamma_2, \ldots \gamma_m\}$, where M is the total number of pairs. Fellegi and Sunter (1969) show that the rule declaring the pairs whose index is smaller than some upper bound K "links" is the most discriminatory record linkage rule. The upper bound K is a function of the maximum type I error tolerated. The rule is most discriminatory in the sense that for the same tolerance on the type I error, it is impossible to find another rule which, in the long run, will retrieve more matched pairs. This fact is a direct application of the Neyman-Pearson Lemma (DeGroot 1986, pp. 444-445). Two uses of the Fellegi-Sunter rule are illustrated in section 6.

The Fellegi-Sunter record-linkage rule is articulated around the ratio $m(\gamma)/u(\gamma)$. Usually this ratio is estimated from the data through a model of the underlying probabilistic process. It is assumed that the model is a genuine representation of the probabilistic process. If the representation is not genuine, then substituting $m(\gamma)/u(\gamma)$ in the Fellegi-Sunter rule may not yield a most discriminatory record-linkage rule. Therefore, particular care must be taken in the choice of the model. The next section introduces models designed to describe the underlying probabilistic process in given situations.

3. MODELS FOR RECORD-LINKAGE

Two models formulating underlying probabilistic processes are presented in this section. The first model is a general formulation of any underlying process. The second model is an application of the first. In some situations, the second model is a good representation of the underlying probabilistic process and the Fellegi-Sunter rule based on this model is most discriminatory. Parameter estimation is discussed so that the expressions involved in the Fellegi-Sunter rule can be evaluated.

3.1 Latent Class Models

Because of the particular nature of a record-linkage process, the underlying probabilistic process can always be represented by a latent class model. A latent class model is built around latent variables. Generally speaking, a latent variable is a variable not observable, characterizing any observation generated by the probabilistic process. Latent variables classify the observations into latent

classes. In this problem, the observations are the comparison vectors (*i.e.* comparison patterns). An obvious latent variable categorizing the observations into two latent classes is the status of the pair associated with each comparison vector. This status is that of a matched pair status or of an unmatched pair status. The corresponding latent classes are the class of matched pairs and the class of unmatched pairs. A mathematical representation is given next to enable development of specific latent class models.

Let $\nu_{k,i_1,\ldots,i_N}$ represent the count of pairs with the following attributes: if $k = 0$ the corresponding pairs have an unmatched pair status and if $k = 1$ they have a matched pair status. Furthermore, whenever $i_s = 0$, the corresponding pairs do not exhibit record agreement over the comparison field s and whenever $i_s = 1$, the pairs do exhibit record agreement over the comparison field s. Note that $s = 1, \ldots, N$, where N is the number of comparison fields. It is important to keep in mind that the counts $\nu_{k,i_1,\ldots,i_N}$ cannot be observed. Rather, what is observed are the counts aggregated over the latent classes. The aggregated counts are denoted by $\nu_{i_1,\ldots,i_N}$, where

$$\nu_{i_1,\ldots,i_N} = \nu_{0,i_1,\ldots,i_N} + \nu_{1,i_1,\ldots,i_N}. \quad (1)$$

While only the aggregated counts are observable in record-linkage situations, models are usually expressed in terms of the basic counts. This is done only for convenience. The following subsection is more specific and a simple latent class model for record linkage is introduced.

3.2 Conditional Independence

The conditional independence models are the simplest latent class models. Despite their simplicity, these models are an accurate representation of the underlying probabilistic process in some situations. Goodman (1974) gives a thorough analysis of several conditional independence models. Haberman (1979) gives a presentation of several conditional independence models, along with appropriate techniques of parameter estimation.

In this section, the conditional independence model for record linkage is introduced and its implications in terms of the underlying probabilistic process are exposed. The model is best described in its log-linear representation:

$$\log(\nu_{k,i_1,\ldots,i_N}) = \mu + \lambda_k + \sum_{j=1}^{N} \alpha_{i_j}^{j} + \sum_{j=1}^{N} \zeta_{k,i_j}^{j}. \quad (2)$$

Naturally, there are constraints attached to the parameters of the model given in (2):

$$\lambda_1 = -\lambda_0;\ \alpha_1^j = -\alpha_0^j;\ \zeta_{k,1}^j = -\zeta_{k,0}^j;\ \zeta_{1,ij}^j = -\zeta_{0,ij}^j,$$
$$k = 0,1;\ j = 1, \ldots, N;\ i_j = 0,1. \quad (3)$$

The expression on the right-hand side of (2) includes one term for the latent variable (λ_k) and one term for each comparison field ($\alpha_{i_j}^j$). It also includes interaction terms ($\zeta_{k,ij}^j$). Each interaction is between a field and the latent variable. There are no direct interaction between the comparison fields. In other words, conditional on each latent class, agreements and disagreements over the comparison fields occur independently.

The assumption that the comparison variables are independent given the value of the latent variable is implicit when deriving inference through a conditional independence model. In practice, however, the underlying probabilistic process often conflicts with this assumption. Then the Fellegi-Sunter record-linkage rule constructed assuming model (2) may not be most discriminatory. In that situation, the discriminatory power can be raised through a better elicitation of the model. In fact, more elaborate latent class models integrate a higher degree of complexity in the relationships between the comparison fields themselves and between the comparison fields and the latent variable. These models can take a large number of forms according to the nature of a particular record-linkage situation. An instance of such models is presented in Section 4.

3.3 Parameter Estimation for the Conditional Independence Model

Once a model has been formulated, the values of its parameters must be evaluated. Then the Fellegi-Sunter rule is constructed from the model using the corresponding estimated values for $m(\gamma)$ and $u(\gamma)$. The parameter estimation process shall be reliable enough to prevent a significant loss of discriminatory power by way of the estimation error.

One feature of the latent class models makes them prone to estimation error: unidentifiability. Latent class models typically are unidentifiable in the sense that the equations maximizing the likelihood admit more than one solution. Parameter estimation with unidentifiable models remains difficult and confusing. However, from experience, the author found that for the conditional independence models, unidentifiability is usually not a determinant factor in the estimation error. A larger part of the error typically comes from the inadequacy of the model as a genuine representation of the underlying probabilistic process.

A suitable parameter-estimation technique for conditional independence models stems from approaching the problem as one of finding a maximum likelihood estimator in the presence of "missing observations". The missing observation in this case is the latent variable, the status of each pair. In the general context of parameter estimation with missing observations, Expectation-Maximization (E.M.) algorithms are quite popular. In fact, the E.M. algorithm is implemented without difficulty in the estimation

of the parameters of the conditional independence model given in (2) (Winkler 1988). But if there is considerable departure from the independence assumption, the value of the estimates becomes difficult to interpret (An example of this is given in section 4).

4. THE ST. LOUIS DATA: AN EXAMPLE OF A COMPLEX RECORD-LINKAGE PROCESS

This section introduces a particular example of a record-linkage process. A model is developed specifically to represent the underlying probabilistic process supporting this record-linkage process. It is expected that this model will induce more discrimination power in the application of the Fellegi-Sunter rule than the conditional independence model would.

4.1 Observable Latent Variable

The example is based on data collected in 1988 during a dress rehearsal in preparation for the Decennial Census Operations. Basically, there are two separate and presumably exhaustive surveys of all the individuals living in a defined geographical area within the city of St. Louis, Missouri. For each survey and for each individual available at the time of the survey, a record is created and various characteristics of the individual are recorded. These characteristics are: house number, phone number, street name, first name, last name, middle initial, marital status, age, race, sex, relationship with the respondent. The records of the two surveys are linked together.

For this particular application, the latent variable is made observable through an extensive follow-up study for the purpose of this and other researches. In the present situation, the information extracted from the latent variable leads to the construction of a model representative of the probabilistic process underlying the record-linkage process. Ultimately the discrimination power of this model is compared with that of the conditional independence model. The motivations leading to the construction of the model are presented in the following subsections.

4.2 Blocking and Dependencies

The goal of record-linkage is to retrieve as many matched pairs as possible given an upper bound on the type I error. The first obstacle is often the size of the files. The files may be quite large, making it impossible to examine all the pairs consisting of one record of file A and one record of file B. Blocking is considered whenever an exhaustive review of all the pairs is too costly and/or too time consuming.

The principle of blocking is as follows: To bring down the number of comparisons and other associated operations, the records of each file are assigned to blocks according to the value of a few key characteristics. These characteristics are called the blocking variables. Only the records whose blocking variables take the same values may be brought into pairs. Since the records forming a matched pair tend to agree on the blocking characteristics, it is natural to expect the vast majority of the pairs discarded to be unmatched, as a result of the blocking scheme.

In the St. Louis example, the census file has 15,048 records, while the PES file contains 12,072 records. Potentially, there are over 180,000,000 pairs available for review. This number is excessive and blocking must be used to keep the size of the problem manageable. Therefore, the records are blocked on the first character of the surname and on a geographical unit called geocode. The geographical area encompassed by a given value of the geocode may consist of several street blocks, or two or more nearby perpendicular or parallel streets. This scheme yields blocks of reasonable sizes. Under this design, 116,305 pairs provide the information to construct inference.

Unfortunately, while it brings down the size of the problem, blocking on geocode also has undesirable side effects: it induces strong dependencies between the household variables among the unmatched pairs. The household variables are the last name, house number, street name and telephone number. For instance, consider two individuals forming an unmatched pair but who are part of the same block. Now, suppose these two individuals agree on the last name. Intuitively, given this information, chances are higher that the two individuals are from the same household. Therefore, the probabilities of agreement over the other household fields, given the information of agreement on the last name, are higher than the marginal probabilities. The nature of the dependencies between the household variables is studied next.

4.3 Measuring the Dependencies

To construct a model representative of the St. Louis record-linkage process, the dependencies between the household variables must be assessed. The information on the latent variable allows this. Table 1 gives the correlations of the responses of record comparisons over the comparison fields for the matched pairs. Table 2 gives the correlations of the responses of the record comparisons over the comparison fields for the unmatched pairs. For both matrices, all the correlations greater or equal to .01 are given. A correlation is not shown only if it is smaller than .01.

The correlations in Table 1, are rather small and overall do not suggest a significant pattern of dependency among the comparison variables restricted to the matched pairs. Note in particular that the correlations between the household variables are small among the matched pairs, suggesting little or no dependency. This can be explained by the fact that among the matched pairs, the agreement rate over any household field is very high and has a behavior close to that of a constant.

Table 1
Correlations Between Selected Comparison Fields over the Set of Links

	Middle In.	Street	Phone	Marital
First Name	.123	0.	.045	.032
Middle In.	1	.010	.161	.079
House No.	.017	.194	.037	0.
Street	.01	1	.035	0.
Phone	.161	.035	1	.107
Age	.051	.004	.075	.118
Marital	.079	0.	.107	1

Table 2
Correlations Between Selected Comparison Fields over the Set of Non-Links

	House No.	Street	Phone	Marital	Race
Last N.	.748	.326	.642	.099	.101
House No.	1	.400	.699	.111	.105
Street	.400	1	.292	.043	.086
Age	.104	.054	.086	.165	.024
Rel	.121	.068	.084	.394	.049

But in Table 2, the effects of blocking are evident in the high values of the correlations associated with the household variables restricted to the unmatched pairs. A sensible design for the model of the underlying probabilistic process should account for these high correlations by incorporating dependency components.

4.4 A Model Tailored for the St. Louis Data

In order to make valid inference on the status of the pairs, a model descriptive of the underlying probabilistic process must be elicited. The conditional independence model presented in (2) is attractive because of its simplicity. However, it is clear at this point that this model does not correctly represent the probabilistic process underlying the St. Louis record-linkage process. An educated model is introduced, motivated by the information made available on the dependencies between the household variables.

To appreciate the more general structure of the educated model, some conventions must be set regarding the indexing of the comparison fields: comparison field 1 is the last name, comparison field 2 is the house number, comparison field 3 is the street name, and comparison field 4 is the phone number. The seven remaining comparison fields are indexed arbitrarily by the values 5-11. The educated model accounts for all possible interaction effects between fields 1 through 4 among the unmatched pairs. The log-linear representation of the educated model is as follows:

$$\log(\nu_{k,i_1,\ldots,i_{11}}) = \mu + \lambda_k + \sum_{j=1}^{11} \alpha_{i_j}^{j} + \sum_{j=1}^{11} \zeta_{k,i_j}^{j} + (1-k)\left(\sum_{\{1\le j<l\le 4\}} \eta_{i_j,i_l}^{j,l} + \sum_{\{1\le j<l<m\le 4\}} \Phi_{i_j,i_l,i_m}^{j,l,m} + \Psi_{i_1,i_2,i_3,i_4}^{1,2,3,4} \right). \quad (4)$$

Note the coefficient $(1 - k)$ multiplying the household interaction terms, indicating that the dependency relation between the household variables is only among the unmatched pairs. This contrasts with the symmetry of the conditional independence model in (2).

The restrictions in (3) apply here as well. In addition, more constraints must be satisfied. The following constraints are imposed on the interaction terms of the second order:

$$\eta_{i_j,1}^{j,l} = -\eta_{i_j,0}^{j,l}; \quad \eta_{1,i_l}^{j,l} = -\eta_{0,i_l}^{j,l}. \quad (5)$$

The range of the indices is $1 \le j < l \le 4$. The constraints on the interaction terms of the third order are:

$$\Phi_{i_j,i_l,1}^{j,l,m} = -\Phi_{i_j,i_l,0}^{j,l,m}; \quad \Phi_{i_j,1,i_m}^{j,l,m} = -\Phi_{i_j,0,i_m}^{j,l,m}; \quad \Phi_{1,i_l,i_m}^{j,l,m} = -\Phi_{0,i_l,i_m}^{j,l,m}. \quad (6)$$

The range of the indices in this case is: $1 \le j < l < m \le 4$. Finally, the constraints on the fourth order interaction terms are:

$$\Psi_{i_1,i_2,i_3,1}^{1,2,3,4} = -\Psi_{i_1,i_2,i_3,0}^{1,2,3,4}; \quad \Psi_{i_1,i_2,1,i_4}^{1,2,3,4} = -\Psi_{i_1,i_2,0,i_4}^{1,2,3,4}; \quad \Psi_{i_1,1,i_3,i_4}^{1,2,3,4} = -\Psi_{i_1,0,i_3,i_4}^{1,2,3,4}; \quad \Psi_{1,i_2,i_3,i_4}^{1,2,3,4} = -\Psi_{0,i_2,i_3,i_4}^{1,2,3,4}. \quad (7)$$

It is natural to expect the educated model (4) to be more discriminatory since it accounts for interactions between the household variables. In section 6, the performances of the two models are presented.

4.5 Parameter Estimation for Models with Dependencies

Parameter estimation for models with dependencies is far more difficult than for conditional independence models. For the St. Louis example, the scoring algorithm given by Haberman (1979, p. 547) was used to estimate the parameters of the educated model (4). This technique can be regarded as an E.M. algorithm where the maximization part (M. step) is an application of the Newton-Ralphson algorithm.

The most important difficulty when using this technique is the choice of a starting point. The following strategy is adopted to choose a starting point. First, the parameters of the conditional independence model (2) are estimated via the E.M. algorithm presented in subsection 3.3. Then an intermediate model is constructed. The intermediate model, in this case, embeds all the second and lower order interaction terms of the educated model (4). The estimated parameters of the conditional independence model can serve to construct the starting point to estimate the parameters of the intermediate model through the scoring algorithm. Finally, the estimates of the parameters of the intermediate model are used as a starting point to estimate the parameters of the educated model (4), via the scoring algorithm.

5. THE AD-HOC APPROACH

In the last section, a complex model representing an underlying probabilistic process was elicited for the St. Louis data. In this situation, the elicitation is easy since follow-up information is available. Of course in practice, follow-up information is not available. It is often too difficult and/or too expensive to go through the elicitation and estimation procedures to determine the structure of the underlying process and the values of the parameters. In those cases, an ad-hoc approach might be appropriate. In the St. Louis example, the ad-hoc approach consists of adjusting the parameters of the process derived from the conditional independence model (2) to obtain a more discriminatory model.

Note that under both model (2) and model (4), for the matched pair, the agreement or disagreements over the comparison fields are independent. This means that the following formula applies in both situations.

$$m(\gamma) = \prod_{i=1}^{N} m_i^{x_i}(1 - m_i)^{1-x_i},$$

m_i is the probability of agreement over field i of two records forming a matched pair. Furthermore, $x_i = 0$ if the pattern γ calls for a disagreement over field i and $x_i = 1$ if it calls for an agreement. The idea behind the ad-hoc method is to keep the conditional independence structure in (2), but to adjust the values of the m_i's.

The probabilities of agreement, conditional on a matched pair, evaluated under the conditional independence model and the educated model are given in Table 3. The difference between the probability corresponding to the educated model with the probability corresponding to the conditional independence model can be quite substantial for some fields. In particular, the difference is important in the case of the first name field.

Table 3
Probabilities of Agreement Conditional on a Matched Pair

Comparison Field	Cond'l Indep.	Educated
Last Name	.9430	.9561
First Name	.3319	.9140
Mid. Init.	.2125	.5222
House No.	.9692	.9724
Street Name	.9179	.9194
Phone	.6619	.6887
Age	.3903	.8602
Relation	.3353	.4986
Marital Status	.6072	.8547
Sex	.6134	.4842
Race	.9672	.9018

In general, experience shows that the conditional probability of agreement over first name, conditional on a matched pair, is around .99, closer to the .91 value obtained under the educated model. Therefore, after estimating the parameters of the conditional independence model through the E.M. algorithm, the probability of agreement over the first name given a match status is replaced by the value .99. The probability of agreement over the last name given a matched pair is also replaced by the value .99. This procedure increases the discriminatory power associated with the conditional independence model in the application of the Fellegi-Sunter rule.

6. APPLYING THE FELLEGI-SUNTER RULE

6.1 St.Louis

This subsection evaluates the discrimination power of the Fellegi-Sunter rule when applied to the St-Louis record-linkage data and assuming, in turn, three different underlying probabilistic processes. The three underlying probabilistic processes assumed are derived directly from the conditional model (2), directly from the educated model (4), and finally, from the conditional model (2), through the ad-hoc procedure. The following table gives a comparative measure of the performance of the Fellegi-Sunter rule under each of the 3 assumptions regarding the underlying process. The performance is evaluated making use of the privileged information available on the latent variable.

Each cell of Table 4 contains three entries. The first of these entries is the number of matched pairs that were designated links through the Fellegi-Sunter record-linkage rule, assuming each of the three underlying processes, and under four different controlled Type I errors. The total number of matched pairs that could theoretically be recovered is 9,823. The second entry of each cell is the total number of pairs designated link through the Fellegi-Sunter rule. The third entry of the cell is the upper bound on the

Type I error. Recall that the Fellegi-Sunter rule maximizes the number of links under a fixed type I error provided it is based on the correct underlying process. The first column of Table 4 gives the counts assuming an underlying process derived from the conditional independence model (2). The second column gives the same quantities assuming an underlying process derived from the educated model (4). Finally, the third column gives the same numbers assuming an underlying process derived from the conditional independence model and adjusted through the ad-hoc procedure.

Table 4

St. Louis: Links Recovered via Three Approaches under Four Error Levels

	Independence Assumption	Household Interactions	Ad-hoc Procedure
Links	6,404	9,012	6,476
Pairs	6,436	9,056	6,508
Error Bound	.005	.005	.005
Links	7,273	9,712	9,562
Pairs	7,346	9,808	9,659
Error Bound	.01	.01	.01
Links	9,636	9,758	9,765
Pairs	9,824	9,952	9,960
Error Bound	.02	.02	.02
Links	9,740	9,776	9,783
Pairs	10,038	10,062	10,097
Error Bound	.03	.03	.03

There are two important facts that can be deduced from this table. First, the rule based on an underlying process derived from the educated model (4) does consistently better than the rule based on an underlying process derived from the conditional independence model in terms of matches retrieved. Secondly, the performances of the rules differ most when the bound on the type I error is small and at that level (.005), the rule based on an underlying probability process derived from the educated model is clearly superior. When the bound is larger (.03), the underlying probabilistic models are more or less equivalent in terms of induced discrimination power.

6.2 Columbia

The same type of data were collected throughout the area of Columbia, Missouri. The data are slightly different because some of the records have a rural format, that is the street name is replaced by the rural route number and the house number by the box number. Nevertheless, the same relations of dependencies emerge and the same model is appropriate. Table 5 gives a summary of the discrimination achieved at 2 levels of tolerance on the type I error. Taking into account the blocking scheme, there are 6,780 retrievable pairs.

Table 5

Columbia: Links Recovered via Three Approaches under Two Error Levels

	Independence Assumption	Household Interactions	Ad-hoc Procedure
Links	700	1,268	2,035
Pairs	704	1,276	2,046
Type I Error	.005	.005	.005
Links	5,954	6,607	6,545
Pairs	6,016	6,675	6,612
Type I Error	.01	.01	.01

In the case of Columbia, it is clear again than the educated model does better than the conditional independence model. It should be noted that in practice, the ad-hoc approach built on the conditional independence model performs as well as the educated model. The educated model however, is preferred because of its sound theoretical basis.

7. A SUGGESTION FROM AN ANONYMOUS REFEREEE

Another ad-hoc technique is suggested by an anonymous referee. The referee points out that a large majority of the pairs examined in situations like these are unmatched. In the case of St. Louis, 91.5% of the pairs examined turn out to be unmatched. Given this proportion, the trends animating the comparison variables over the set of all pairs, mostly reflect the activity of the unmatched pairs. This reasoning can be extended further to conclude that the estimation of the parameters of the dependency structure underlying the unmatched pairs can be carried through successfully by treating the set of all pairs as if it were the set of unmatched pairs. The parameter estimation becomes trivial. The parameters that must be estimated characterize a simple log-linear model, without any latent variable (Fienberg, Bishop and Holland, p. 24). The parameters descriptive of the matches can be estimated separately through a simple iterative technique such as the E.M. algorithm, combined with *a priori* information.

The approach of the referee does proceed from a realistic model of the process, and in that way, it is in agreement with the thrust of this paper. But the effort of the paper is also to devise discriminatory rules, while sticking to the latent structure constraint. In situations where the proportion of matched pairs is high, or dependencies are manifest among the matches, the approach of the referee fails. A parameter estimation derived directly from the natural model, if feasible, is recommended.

8. CONCLUSIONS

The goal of the research was to show how a better elicitation of the probabilistic models supporting record-linkage processes can induce accrued discriminatory power in the Fellegi-Sunter record-linkage rule. In the cases of the St. Louis and Columbia examples, this goal was certainly achieved. The educated model given in (4) is indeed more descriptive of the underlying probabilistic process and it induces a good deal more discrimination power in the Fellegi-Sunter rule than the conditional independence model (2).

The techniques used for the St. Louis and Columbia data can also be used for the analysis of other data set generated by record-linkage processes supported by a probabilistic process with a similar dependency structure. This dependency structure is certain to surface in any record-linkage application involving the matching of records of individuals on a set of household variables (last name, street name, house number, phone, rural address *etc.*). It is also likely to occur when matching records of businesses on household variables.

There are two major difficulties in the way, when seeking improved discriminatory power by model elicitation. First, since the probability structure underlying the process is usually unknown, to elicit the structure or the corresponding statistical model involves a considerable investigative effort and the cost involved may be prohibitive. Second, even assuming that the correct model is available, the estimation procedures available for the parameter estimation are difficult to handle and poorly understood. More research and work are needed to understand and, to a degree, overcome these two difficulties.

It must also be pointed out that methods based on ad-hoc adjustments of the type described in section 5, and on approximations, as suggested by an anonymous referee, also increase the discriminatory power of the Fellegi-Sunter rule substantially in situations of the type of St. Louis or Columbia. Techniques of this type are serious competitors. The parameter estimation is easy and the associated Fellegi-Sunter rule can be just as crisp in some cases. However, the assumptions supporting these techniques are flawed and the resulting Fellegi-Sunter rule is pathological, providing an unsteady basis on which to make decisions. A model with parameters estimated "naturally" is preferable. The ad-hoc techniques and approximations are recommended when the elicitation of an educated model seems not possible, or the estimation of the parameters of the educated model appears excessively difficult.

A word must be said about the St. Louis and Columbia data. These data are of very high quality. This explains in part the very successful rate of matching exhibited in both the St. Louis and Columbia examples. It is also reasonable to expect a less clear-cut difference between the various linkage techniques had the data been lower quality.

ACKNOWLEDGEMENTS

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau. The author is grateful to the anonymous referee for his/her patience and constructive suggestions. The author is also indebted to William E. Winkler for the guidance he provided throughout the learning process that led to this paper.

REFERENCES

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

DeGROOT, M.H. (1986). *Probability and Statistics*, 2nd. Edition. Reading, MA: Addison-Wesley.

FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 40, 1183-1210.

GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 2, 215-231.

HABERMAN, S.J. (1979). *Analysis of Qualitative Data*, Vol. 2. New York: Academic Press.

HABERMAN, S.J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proceedings of the Statistical Computing Section, American Statistical Association*, 45-50.

THIBAUDEAU, Y. (1989). Fitting log-linear models in computer matching. *Proceedings of the Statistical Computing Section, American Statistical Association*, 283-288.

WINKLER, W.E. (1989). Near automatic weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Bureau of The Census Fifth Annual Research Conference*, 145-155.

WINKLER, W.E. (1988). Using The E.M. algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.

Survey Methodology, June 1993
Vol. 19, No. 1, pp. 39–58
Statistics Canada

Regression Analysis of Data Files that are Computer Matched

FRITZ SCHEUREN and WILLIAM E. WINKLER[1]

ABSTRACT

This paper focuses on how to deal with record linkage errors when engaged in regression analysis. Recent work by Rubin and Belin (1991) and by Winkler and Thibaudeau (1991) provides the theory, computational algorithms, and software necessary for estimating matching probabilities. These advances allow us to update the work of Neter, Maynes, and Ramanathan (1965). Adjustment procedures are outlined and some successful simulations are described. Our results are preliminary and intended largely to stimulate further work.

KEY WORDS: Record linkage; Matching error; Regression analysis.

1. INTRODUCTION

Information that resides in two separate computer data bases can be combined for analysis and policy decisions. For instance, an epidemiologist might wish to evaluate the effect of a new cancer treatment by matching information from a collection of medical case studies against a death registry in order to obtain information about the cause and date of death (*e.g.*, Beebe 1985). An economist might wish to evaluate energy policy decisions by matching a data base containing fuel and commodity information for a set of companies against a data base containing the values and types of goods produced by the companies (*e.g.*, Winkler 1985). If unique identifiers, such as verified social security numbers or employer identification numbers, are available, then matching data sources can be straightforward and standard methods of statistical analysis may be applicable directly.

When unique identifiers are not available (*e.g.*, Jabine and Scheuren 1986), then the linkage must be performed using information such as company or individual name, address, age, and other descriptive items. Even when typographical variations and errors are absent, name information such as "Smith" and "Robert" may not be sufficient, by itself, to identify an individual. Furthermore, the use of addresses is often subject to formatting errors because existing parsing or standardization software does not effectively allow comparison of, say, a house number with a house number and a street name with a street name. The addresses of an individual we wish to match may also differ because one is erroneous or because the individual has moved.

Over the last few years, there has been an outpouring of new work on record linkage techniques in North America (*e.g.*, Jaro 1989; and Newcombe, Fair and Lalonde 1992). Some of these results were spurred on by a series of conferences beginning in the mid-1980s (*e.g.*, Kilss and Alvey 1985; Howe and Spasoff 1986; Coombs and Singh 1987; Carpenter and Fair 1989); a further major stimulus in the U.S. has been the effort to study undercoverage in the 1990 Decennial Census (*e.g.*, Winkler and Thibaudeau 1991). The new book by Newcombe (1988) has also had an important role in this ferment. Finally, efforts elsewhere have also been considerable (*e.g.*, Copas and Hilton 1990).

What is surprising about all of this recent work is that the main theoretical underpinnings for computer-oriented matching methods are quite mature. Sound practice dates back at least to the 1950s and the work of Newcombe and his collaborators (*e.g.*, Newcombe *et al.* 1959). About a decade later, the underlying theory for these basic ideas was firmly established with the papers of Tepping (1968) and, especially, Fellegi and Sunter (1969).

Part of the reason for the continuing interest in record linkage is that the computer revolution has made possible better and better techniques. The proliferation of machine readable files has also widened the range of application. Still another factor has been the need to build bridges between the relatively narrow (even obscure) field of computer matching and the rest of statistics (*e.g.*, Scheuren 1985). Our present paper falls under this last category and is intended to look at what is special about regression analyses with matched data sets.

By and large we will not discuss linkage techniques here. Instead, we will discuss what happens *after* the link status has been determined. The setting, we will assume, is the typical one where the linker does his or her work separately from the analyst. We will also suppose that the analyst (or user) may want to apply conventional statistical techniques - regression, contingency tables, life tables, *etc.* - to the linked file. A key question we want to explore then is "What should the linker do to help the analyst?" A

[1] Fritz Scheuren, U.S. Internal Revenue Service, Washington DC 20224; William E. Winkler, U.S. Bureau of the Census, Washington DC 20233.

related question is "What should the analyst know about the linkage and how should that information be used?"

In our opinion it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly. Obviously the quality of the linkage effort may directly impact on any analyses done. Despite this, rarely are we given direct measures of that impact (*e.g.*, Scheuren and Oh 1975). Rubin (1990) has noted the need to make inferential statements that are designed to summarize evidence in the data being analyzed. Rubin's ideas were presented in the connotation of data housekeeping techniques like editing and imputation, where nonresponse can often invalidate standard statistical procedures that are available in existing software packages. We believe Rubin's perspective applies at least with equal force in record linkage work.

Organizationally, our discussion is divided into four sections. First, we provide some background on the linkage setting, because any answers - even partial ones - will depend on the files to be linked and the uses of the matched data. In the next section we discuss our methodological approach, focusing, as already noted, just on regression analysis. A few results are presented in section 4 from some exploratory simulations. These simulations are intended to help the reader weigh our ideas and get a feel for some of the difficulties. A final section consists of preliminary conclusions and ideas for future research. A short appendix containing more on theoretical considerations is also provided.

2. RECORD LINKAGE BACKGROUND

When linking two or more files, an individual record on one file may not be linked with the correct corresponding record on the other file. If a unique identifier for corresponding records on two files is not available - or is subject to inaccuracy - then the matching process is subject to error. If the resultant linked data base contains a substantial proportion of information from pairs of records that have been brought together erroneously or a significant proportion of records that need to be brought together are erroneously left apart, then statistical analyses may be sufficiently compromised that results of standard statistical techniques could be misleading. For the bulk of this paper we will only be treating the situation of how erroneous links affect analyses. The impact of problems caused by erroneous nonlinks (an implicit type of sampling that can yield selection biases) is discussed briefly in the final section.

2.1 Fellegi-Sunter Record Linkage Model

The record linkage process attempts to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (*e.g.*, Newcombe *et al.* 1959), Fellegi and Sunter (1969) considered ratios of probabilities of the form:

$$R = Pr(\gamma \in \Gamma \mid M)/Pr(\gamma \in \Gamma \mid U), \qquad (2.1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Smith or Zabrinsky, occur. The fields that are compared (surname, first name, age) are referred to as *matching variables*.

The decision rule is given by:

If $R > Upper$, then designate pair as a link.

If $Lower \leq R \leq Upper$, then designate pair as a possible link and hold for clerical review. (2.2)

If $R < Lower$, then designate pair as a nonlink.

Fellegi and Sunter (1969) showed that the decision rule is optimal in the sense that for any pair of fixed bounds on R, the middle region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds *Upper* and *Lower* are determined by the error bounds. We call the ratio R or any monotonely increasing transformation of it (such as given by a logarithm) a *matching weight* or *total agreement weight*.

In actual applications, the optimality of the decision rule (2.2) is heavily dependent on the accuracy of the estimates of the probabilities given in (2.1). The probabilities in (2.1) are called *matching parameters*. Estimated parameters are (nearly) *optimal* if they yield decision rules that perform (nearly) as well as rule (2.2) does when the true parameters are used.

The Fellegi-Sunter approach is basically a direct extension of the classical theory of hypothesis testing to record linkage. To describe the model further, suppose there are two files of size n and m where - without loss of generality - we will assume that $n \leq m$. As part of the linkage process, a comparison might be carried out between all possible $n \times m$ pairs of records (one component of the pair coming from each file). A decision is, then, made as to whether or not the members of each comparison-pair represent the same unit or whether there is insufficient evidence to determine link status.

Schematically, it is conventional to look at the $n \times m$ pairs arrayed by some measure of the probability that the pair represent records for the same unit. In Figure 1, for example, we have plotted two curves. The curve on the right is a hypothetical distribution of the n true links by the "matching weight" (computed from (2.1) but in natural logarithms). The curve on the left is the remaining of the $n \times (m - 1)$ pairs - the true nonlinks - plotted by their matching weights again in logarithms.

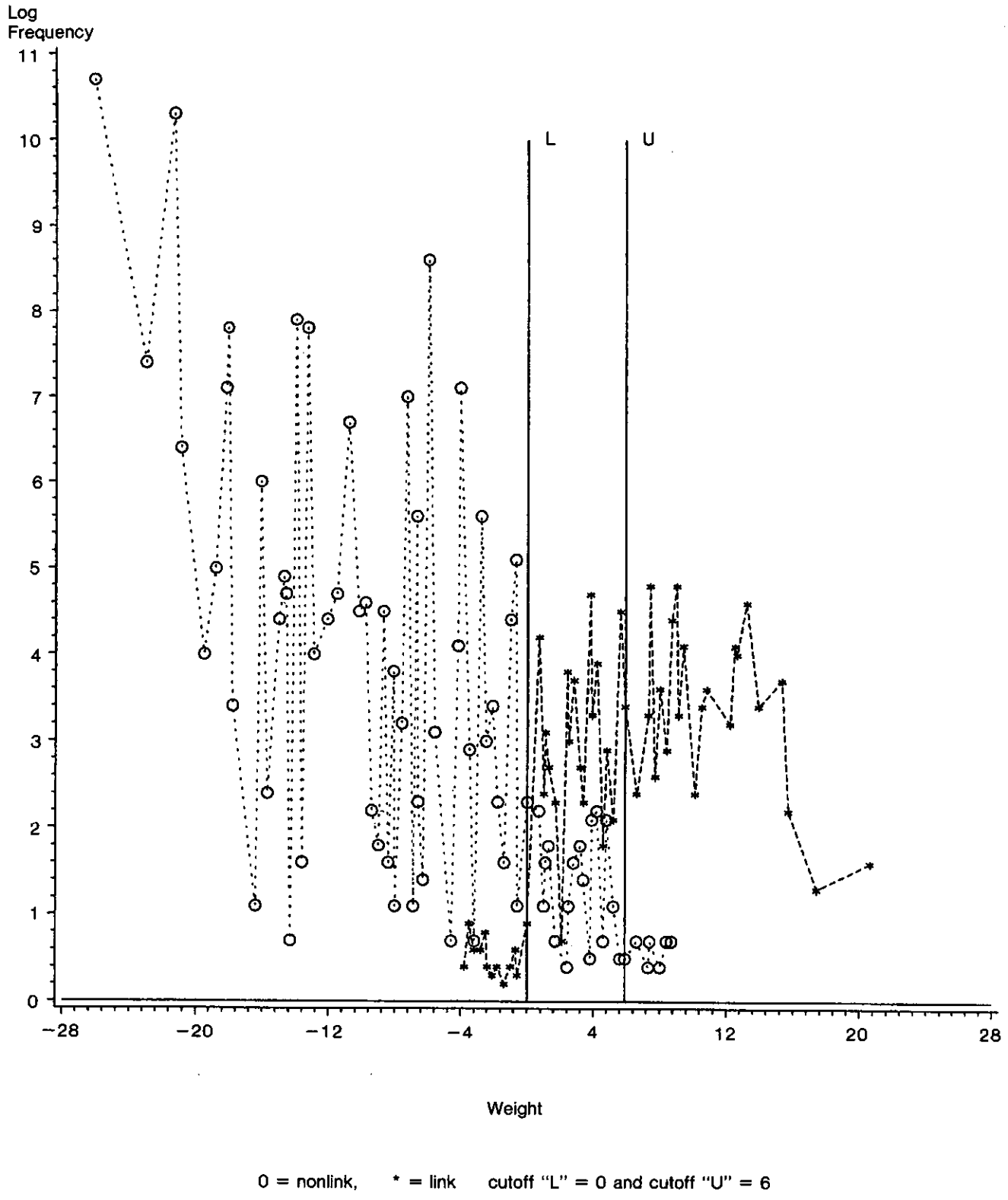


Figure 1. Log Frequency *vs* Weight, Links and Nonlinks

Typically, as Figure 1 indicates, the link and nonlink distributions overlap. At the extremes the overlap is of no consequence in arriving at linkage decisions; however, there is a middle region of potential links, say between "L" and "U", where it would be hard, based on Figure 1 alone, to distinguish with any degree of accuracy between links and nonlinks.

The Fellegi-Sunter model is valid on any set of pairs we consider. However, for computational convenience, rather than consider all possible pairs in $\mathbf{A} \times \mathbf{B}$, we might consider only a subset of pairs where the records from both files agree on key or "blocking" information that is thought to be highly accurate. Examples of the *logical blocking criteria* include items such as a geographical identifier like Postal (*e.g.*, ZIP) code or a surname identifier such as a Soundex or NYSIIS code (see *e.g.*, Newcombe 1988, pp. 182-184). Incidentally, the Fellegi-Sunter Model does not presuppose (as Figure 1 did) that among the $n \times m$ pairs there will be n links but rather, if there are no duplicates on A or B, that there will be at most n links.

2.2 Handling Potential Links

Even when a computer matching system uses the Fellegi-Sunter decision rule to designate some pairs as almost certain *true links* or *true nonlinks*, it could leave a large subset of pairs that are only potential links. One way to address potentially linked pairs is to clerically review them in an attempt to delineate true links correctly. A way to deal with erroneously nonlinked pairs is to perform additional (again possibly clerical) searches. Both of these approaches are costly, time-consuming, and subject to error.

Not surprisingly, the main focus of record linkage research since the beginning work of Newcombe has been how to reduce the clerical review steps caused by the potential links. Great progress has been made in improving linkage rules through better utilization of information in pairs of records and at estimating error rates via probabilistic models.

Record linkage decision rules have been improved through a variety of methods. To deal with minor typographical errors such as "Smith" versus "Smoth", Winkler and Thibaudeau (1991) extended the string comparator metrics introduced by Jaro (1989). Alternatively, Newcombe *et al.* (1989) developed methods for creating and using partial agreement tables. For certain classes of files, Winkler and Thibaudeau (1991) (see also Winkler 1992; Jaro 1989) developed Expectation-Maximization procedures and *ad hoc* modelling procedures based on *a priori* information that automatically yielded the optimal parameters in (2.1) for use in the decision rules (2.2).

Rubin and Belin (1991) introduced a method for estimating error rates, when error rates could not be reliably estimated via conventional methods (Belin 1991, pp. 19-20). Using a model that specified that the curves of weights versus log frequency produced by the matching process could be expressed as a mixture of two curves (links and nonlinks), Rubin and Belin estimated the curves which, in turn, gave estimates of error rates. To apply their method, Rubin and Belin needed a training sample to yield an *a priori* estimate of the shape of the two curves.

While many linkage problems arise in retrospective, often epidemiological settings, occasionally linkers have been able to designate what information is needed in both data sets to be linked based on known analytic needs. Requiring better matching information, such as was done with the 1990 Census Post-Enumeration Survey (see *e.g.*, Winkler and Thibaudeau 1991), assured that sets of potential links were minimized.

Despite these strides, eventually, the linker and analyst still may have to face a possible clerical review step. Even today, the remaining costs in time, money and hidden residual errors can still be considerable. Are there safe alternatives short of a full review? We believe so and this belief motivates our perspective in section 3, where we examine linkage errors in a regression analysis context. Other approaches, however, might be needed for different analytical frameworks.

3. REGRESSION WITH LINKED DATA

Our discussion of regression will presuppose that the linker has helped the analyst by providing a combined data file consisting of pairs of records - one from each input file - along with the match probability and the link status of each pair. Link, nonlink, and potential links would all be included and identified as such. Keeping likely links and potential links seems an obvious step; keeping likely nonlinks, less so. However, as Newcombe has pointed out, information from likely nonlinks is needed for computing biases. We conjecture that it will suffice to keep no more than two or three pairs of matches from the B file for each record on the A file. The two or three pairs with the highest matching weights would be retained.

In particular, we will assume that the file of linked cases has been augmented so that every record on the smaller of the two files has been paired with, say, the *two* records on the larger file having the highest matching weights. As $n \leq m$, we are keeping $2n$ of the $n \times m$ possible pairs. For each record we keep the linkage indicators and the probabilities associated with the records to which it is paired. Some of these cases will consist of (link, nonlink) combinations or (nonlink, nonlink) combinations. For simplicity's sake, we are not going to deal with settings where more than one true link could occur; hence, (link,link) combinations are by definition ruled out.

As may be quite apparent, such a data structure allows different methods of analysis. For example, we can partition

the file back into three parts - identified links, nonlinks, and potential links. Whatever analysis we are doing could be repeated separately for each group or for subsets of these groups. In the application here, we will use nonlinks to adjust the potential links, and, thereby, gain an additional perspective that could lead to reductions in the Mean Square Error (MSE) over statistics calculated only from the linked data.

For statistical analyses, if we were to use only data arising from pairs of records that we were highly confident were links, then we might be throwing away much additional information from the set of potentially linked pairs, which, as a subset, could contain as many true links as the set of pairs which we designate as links. Additionally, we could seriously bias results because certain subsets of the true links that we might be interested in might reside primarily in the set of potential links. For instance, if we were considering affirmative action and income questions, certain records (such as those associated with lower income individuals) might be more difficult to match using name and address information and, thus, might be heavily concentrated among the set of potential links.

3.1 Motivating Theory

Neter, Maynes, and Ramanathan (1965) recognized that errors introduced during the matching process could adversely affect analyses based on the resultant linked files. To show how the ideas of Neter *et al.* motivate the ideas in this paper, we provide additional details of their model. Neter *et al.* assumed that the set of records from one file (1) always could be matched, (2) always had the same probability p of being correctly matched, and (3) had the same probability q of being mismatched to any remaining records in the second file (*i.e.* $p + (N - 1)q = 1$ where N is file size). They generalized their basic results by assuming that the sets of pairs from the two files could be partitioned into classes in which (1), (2) and (3) held.

Our approach follows that of Neter *et al.* because we believe their approach is sensible. We concur with their results showing that if matching errors are moderate then regression coefficients could be severely biased. We do not believe, however, that condition (3) - which was their main means of simplifying computational formulas - will ever hold in practice. If matching is based on unique identifiers such as social security numbers subject to typographical error, it is unlikely that a typographical error will mean that a given record has the same probability of being incorrectly matched to all remaining records in the second file. If matching variables consist of name and address information (which is often subject to substantially greater typographical error), then condition (3) is even more unlikely to hold.

To fix ideas on how our work builds on and generalizes results of Neter *et al.* we consider a special case. Suppose we are conducting ordinary least squares using a simple regression of the form,

$$y = a_0 + a_1 x + \epsilon. \tag{3.1}$$

Next, assume mismatches have occurred, so that the y variables (from one file) and the x variables (from another file) are *not* always for the *same unit*.

Now in this setting, the unadjusted estimator of a_1 would be biased; however, under assumptions such as that x and y are independent when a mismatch occurs, it can be shown that, if we know the mismatch rate, h, that an unbiased adjusted estimator can be obtained by simply correcting the ordinary estimator by multiplying it by $(1/(1 - h))$. Intuitively, the erroneously linked pairs lead to an understatement of the true correlation (positive or negative) between x and y. The adjusted coefficient removes this understatement. With the adjusted slope coefficent $\hat{a}_1$, the proper intercept can be obtained from the usual expression $\hat{a}_0 = \bar{y} - \hat{a}_1\bar{x}$, where $\hat{a}_1$ has been adjusted.

Methods for estimating regression standard errors can also be devised in the presence of matching errors. Rather than just continuing to discuss this special case, though, we will look at how the idea of making a multiplicative adjustment can be generalized. Consider

$$Y = X\beta + \epsilon, \tag{3.2}$$

the ordinary univariate regression model, for which error terms all have mean zero and are independent with constant variance σ^2. If we were working with a data base of size n, Y would be regressed on X in the usual manner. Now, given that each case has two matches, we have $2n$ pairs altogether. We wish to use (X_i, Y_i), but instead use (X_i, Z_i). Z_i could be Y_i, but may take some other value, Y_j, due to matching error.

For $i = 1, \ldots, n$,

$$Z_i = \begin{cases} Y_i \text{ with probability } p_i \\ Y_j \text{ with probability } q_{ij} \text{ for } j \neq i, \end{cases} \tag{3.3}$$

$p_i + \sum_j q_{ij} = 1$.

The probability p_i may be zero or one. We define $h_i = 1 - p_i$ and divide the set of pairs into n mutually exclusive classes. The classes are determined by records from one of the files. Each class consists of the independent x-variable X_i, the true value of the dependent y-variable, the values of the y-variables from records in the second file to which the record in the first file containing X_i have been paired, and computer matching probabilities (or weights). Included are links, nonlinks, and potential links. Under an assumption of one-to-one matching, for each $i = 1, \ldots, n$, there exists at most one j such that $q_{ij} > 0$. We let ϕ be defined by $\phi(i) = j$.

The intuitive idea of our approach (and that of Neter *et al.*) is that we can, under the model assumptions, express each observed data point pair (X,Z) in terms of the true values (X,Y) and a bias term (X,b). All equations needed for the usual regression techniques can then be obtained. Our computational formulas are much more complicated than those of Neter *et al.* because their strong assumption (3) made considerable simplification possible in the computational formulas. In particular, under their model assumptions, Neter *et al.* proved that both the mean and variance of the observed Z-values were necessarily equal the mean and variance of the true Y-values.

Under the model of this paper, we observe (see Appendix) that

$$E(Z) = (1/n)\sum_i E(Z|i) = (1/n)\sum_i (Y_i p_i + \sum_j Y_j q_{ij})$$

$$= (1/n)\sum_i Y_i + (1/n)\sum_i [Y_i(-h_i) + Y_{\phi(i)}h_i]$$

$$= \bar{Y} + B. \qquad (3.4)$$

As each X_i, $i = 1, \ldots, n$, can be paired with either Y_i or $Y_{\phi(i)}$, the second equality in (3.4) represents $2n$ points. Similarly, we can represent σ_{zy} in terms of σ_{xy} and a bias term B_{xy}, and σ_z^2 in terms of σ_y^2 and a bias term B_{yy}. We neither assume that the bias terms have expectation zero nor that they are uncorrelated with the observed data.

With the different representations, we can adjust the regression coefficients β_{zx} and their associated standard errors back to the true values β_{yx} and their associated standard errors. Our assumption of one-to-one matching (which is not needed for the general theory) is done for computational tractability and to reduce the number of records and amount of information that must be tracked during the matching process.

In implementing the adjustments, we make two crucial assumptions. The first is that, for $i = 1, \ldots, n$, we can accurately estimate the true probabilities of a match p_i. See Appendix for the method of Rubin and Belin (1991). The second is that, for each $i = 1, \ldots, n$, the true value Y_i associated with independent variable X_i is the pair with the highest matching weight and the false value $Y_{\phi(i)}$ is associated with the second highest matching weight. (From the simulations conducted it appears that at least the first of these two assumptions matters greatly when a significant portion of the pairs are potential links.)

3.2 Simulated Application

Using the methods just described, we attempted a simulation with real data. Our basic approach was to take two files for which true linkage statuses were known and re-link them using different matching variables – or really versions of the same variables with different degrees of distortion introduced, making it harder and harder to distinguish a link from a nonlink. This created a setting where there was enough discrimination power for the Rubin-Belin algorithm for estimating probabilities to work, but not so much discriminating power that the overlap area of potential links becomes insignificant.

The basic simulation results were obtained by starting with a pair of files of size 10,000 that had good information for matching and for which true match status was known. To conduct the simulations a range of error was introduced into the matching variables, different amounts of data were used for matching, and greater deviations from optimal matching probabilities were allowed.

Three matching scenarios were considered: (1) *good*, (2) *mediocre*, and (3) *poor*. The good matching scenario consisted of using most of the available procedures that had been developed for matching during the 1990 U.S. Census (*e.g.*, Winkler and Thibaudeau 1991). Matching variables consisted of last name, first name, middle initial, house number, street name, apartment or unit identifier, telephone, age, marital status, relationship to head of household, sex, and race. Matching probabilities used in crucial likelihood ratios needed for the decision rules were chosen close to optimal.

The mediocre matching scenario consisted of using last name, first name, middle initial, two address variations, apartment or unit identifier, and age. Minor typographical errors were introduced independently into one seventh of the last names and one fifth of the first names. Matching probabilities were chosen to deviate from optimal but were still considered to be consistent with those that might be selected by an experienced computer matching expert.

The poor matching scenario consisted of using last name, first name, one address variation, and age. Minor typographical errors were introduced independently into one fifth of the last names and one third of the first names. Moderately severe typographical errors were made in one fourth of the addresses. Matching probabilities were chosen that deviated substantially from optimal. The intent was for them to be selected in a manner that a practitioner might choose after gaining only a little experience.

With the various scenarios, our ability to distinguish between true links and true nonlinks differs significantly. For the good scenario, we see that the scatter for true links and nonlinks is almost completely separated (Figure 2). With the mediocre scheme, the corresponding sets of points overlap moderately (Figure 3); and, with the poor, the overlap is substantial (Figure 4).

We primarily caused the good matching scenario to degenerate to the poor matching error (Figures 2-4) by using less matching information and inducing typographical error in the matching variables. Even if we had kept the same matching variables as in the good matching scenario (Figure 2), we could have caused curve overlap (as in Figure 4) merely by varying the matching

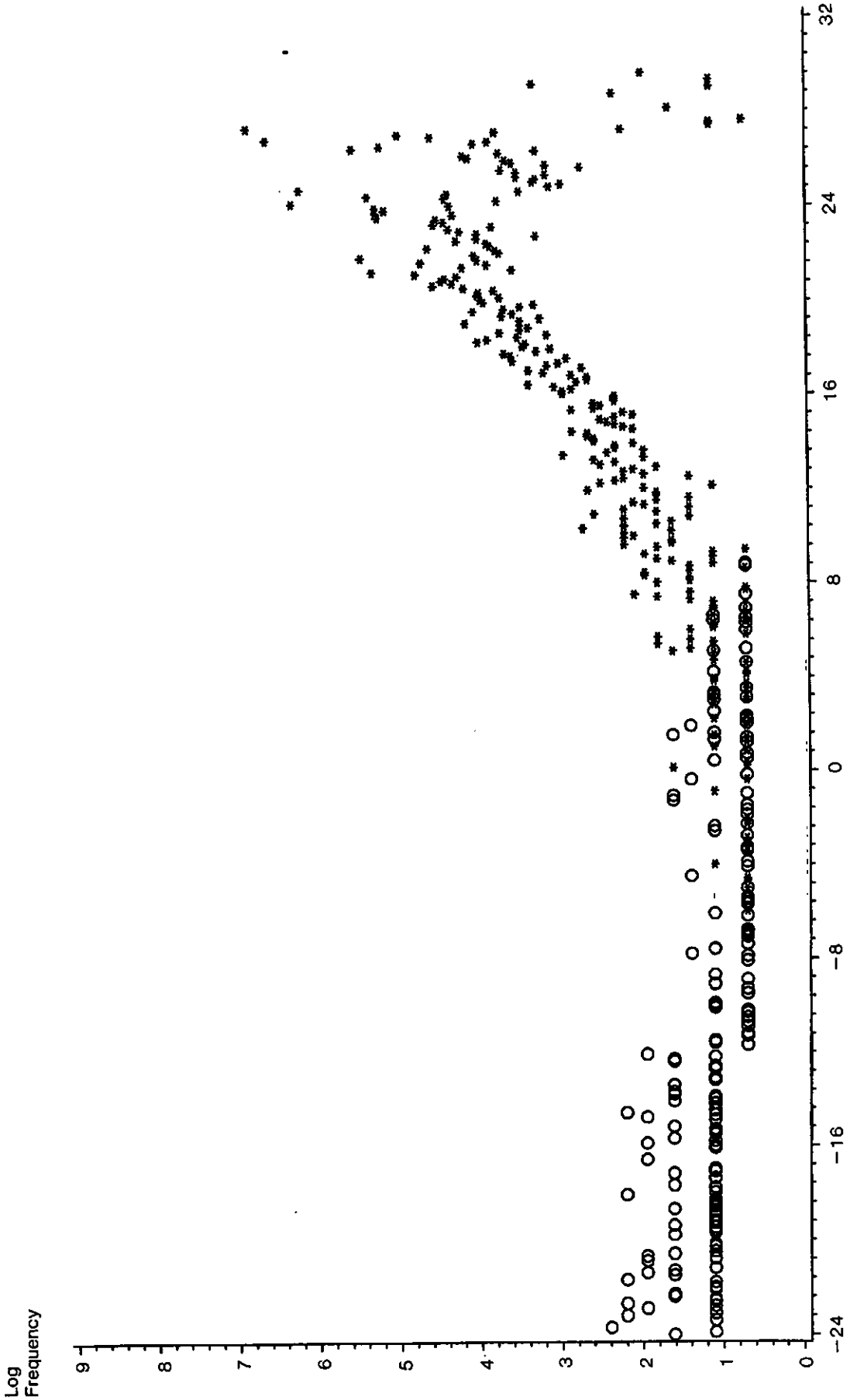


Figure 2. Log of Frequency *vs* Weight Good Matching Scenario, Links and Nonlinks

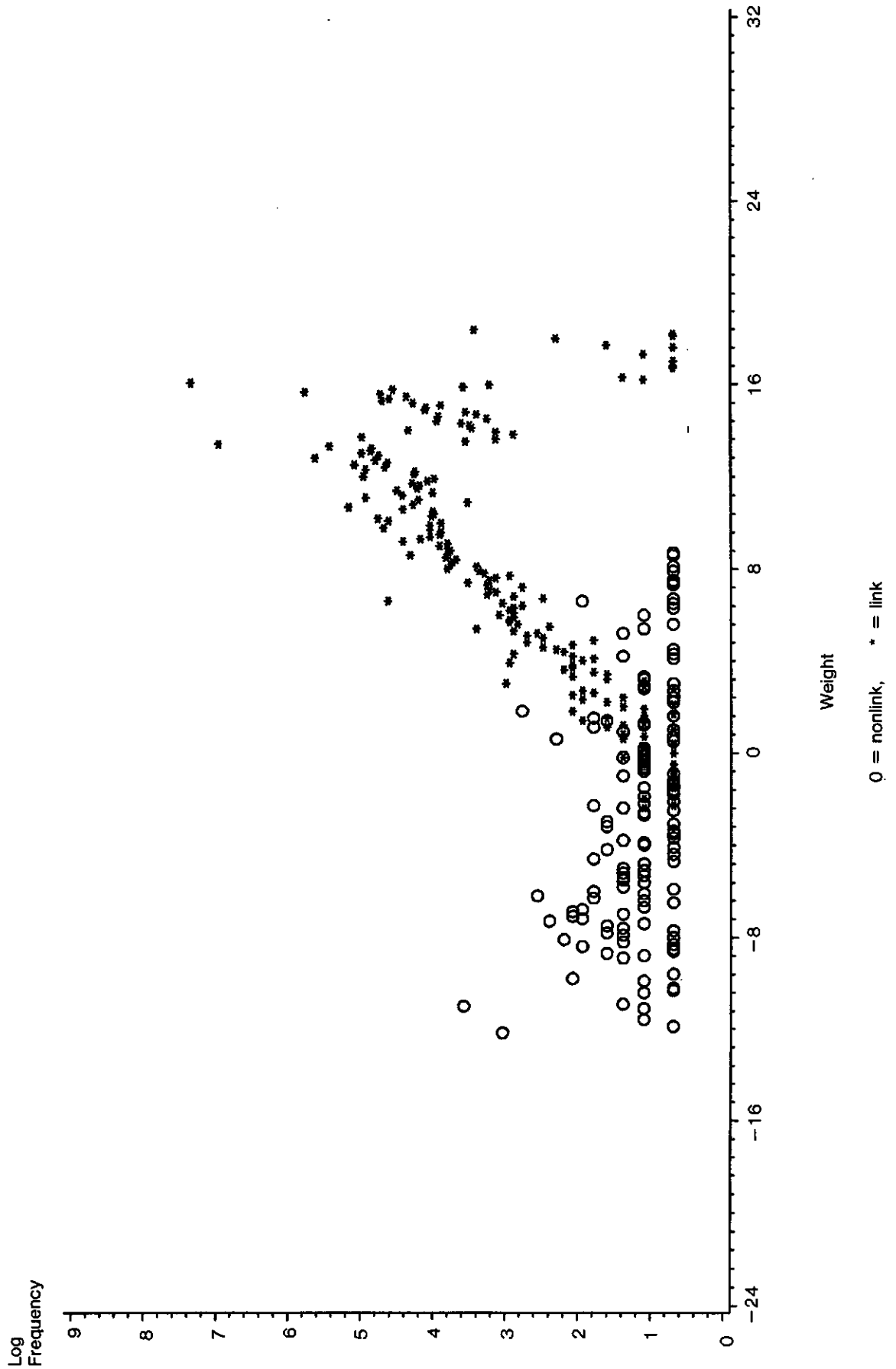


Figure 3. Log of Frequency *vs* Weight Mediocre Matching Scenario, Links and Nonlinks

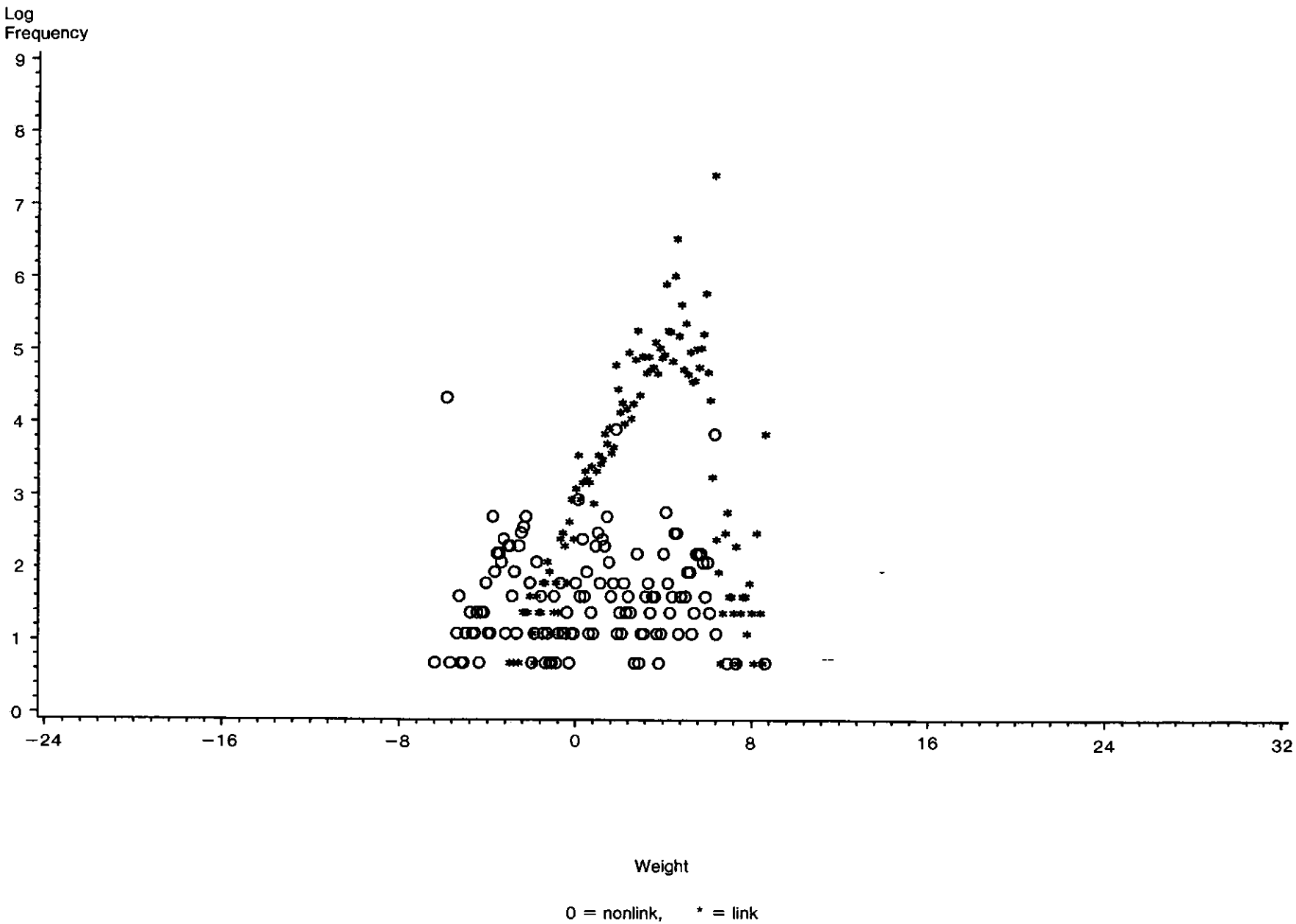


Figure 4. Log of Frequency *vs* Weight Poor Matching Scenario, Links and Nonlinks

Table 1

Counts of True Links and True Nonlinks and Probabilities of an Erroneous Link in Weight Ranges for Various Matching Cases; Estimated Probabilities via Rubin-Belin Methodology

Weight	False match rates											
	Good				Mediocre				Poor			
	True		Prob		True		Prob		True		Prob	
	Link	NL	True	Est	Link	NL	True	Est	Link	NL	True	Est
15+	9,176	0	.00	.00	2,621	0	.00	.00	0	1	.00	.00
14	111	0	.00	.00	418	0	.00	.00	0	1	.00	.00
13	91	0	.00	.01	1,877	0	.00	.00	0	1	.00	.00
12	69	0	.00	.02	1,202	0	.00	.00	0	1	.00	.00
11	59	0	.00	.03	832	0	.00	.00	0	1	.00	.00
10	69	0	.00	.05	785	0	.00	.00	0	1	.00	.00
9	42	0	.00	.08	610	0	.00	.00	0	1	.00	.00
8	36	2	.05	.13	439	3	.00	.00	65	1	.02	.00
7	30	1	.03	.20	250	4	.00	.01	39	1	.03	.00
6	14	7	.33	.29	265	9	.03	.03	1,859	57	.03	.03
5	28	4	.12	.40	167	8	.05	.06	1,638	56	.03	.03
4	6	3	.33	.51	89	6	.06	.11	2,664	62	.02	.05
3	12	7	.37	.61	84	5	.06	.20	1,334	31	.02	.11
2	8	6	.43	.70	38	7	.16	.31	947	30	.03	.19
1	7	13	.65	.78	33	34	.51	.46	516	114	.18	.25
0	7	4	.36	.83	13	19	.59	.61	258	65	.20	.28
−1	3	5	.62	.89	7	20	.74	.74	93	23	.20	.31
−2	0	11	.99	.91	3	11	.79	.84	38	23	.38	.41
−3	4	6	.60	.94	4	19	.83	.89	15	69	.82	.60
−4	4	3	.43	.95	0	15	.99	.94	1	70	.99	.70
−5	4	4	.50	.97	0	15	.99	.96	0	25	.99	.68
−6	0	5	.99	.98	0	27	.99	.98	0	85	.99	.67
−7	1	6	.86	.98	0	40	.99	.99			.99	.99
−8	0	8	.99	.99	0	41		.99			.99	.99
−9	0	4	.99	.99	0	4		.99			.99	.99
−10−	0	22			0	22		.99			.99	.99

Notes: In the first column, weight 10 means weight range from 10 to 11. Weight ranges 15 and above and weight ranges −9 and below are added together. Weights are log ratios that are based on estimated agreement probabilities. **NL** is nonlinks and **Prob** is probability.

parameters given by equation (2.1). The poor matching scenario can arise when we do not have suitable name parsing software that allows comparison of corresponding surnames and first names or suitable address parsing software that allows comparison of corresponding house numbers and street names. Lack of proper parsing means that corresponding matching variables associated with many true links will not be properly utilized.

Our ability to estimate the probability of a match varies significantly. In Table 1 we have displayed these probabilities, both true and estimated, by weight classes. For the good and mediocre matching scenarios, estimated probabilities were fairly close to the true values. For the poor scenario, in which most pairs are potential links, deviations are quite substantial.

For each matching scenario, empirical data were created. Each data base contained a computer matching weight, true and estimated matching probabilities, the independent x-variable for the regression, the true dependent y-variable, the observed y-variables in the record having the highest match weight, and the observed y-variable from the record having the second highest matching weight.

The independent x-variables for the regression were constructed using the SAS RANUNI procedure, so as to be uniformly distributed between 1 and 101. For this paper, they were chosen independently of any matching variables. (While we have considered the situation for which regression variables are dependent on one or more matching variables (Winkler and Scheuren 1991), we do not present any such results in this paper.)

Three regression scenarios were then considered. They correspond to progressively lower R^2 values: (1) R^2 between 0.75 and 0.80; (2) between 0.40 and 0.45; and (3) between 0.20 and 0.22. The dependent variables were generated with independent seeds using the SAS RANNOR procedure. Within each matching scenario (good, mediocre, or poor), all pairing of records obtained by the matching process and, thus, matching error was fixed.

It should be noted that there are two reasons why we generated the (x,y)-data used in the analyses. First, we wanted to be able to control the regression data sufficiently well to determine what the effect of matching error was. This was an important consideration in the very large Monte Carlo simulations reported in Winkler and Scheuren (1991). Second, there existed no available pairs of data files in which highly precise matching information is available and which contain suitable quantitative data.

In performing the simulations for our investigation, some of which are reported here, we created more than 900 data bases, corresponding to a large number of variants of the three basic matching scenarios. Each data base contained three pairs of (x,y)-variables corresponding to the three basic regression scenarios. An examination of these data bases was undertaken to look at some of the matching sensitivity of the regressions and associated adjustments to the sampling procedure. The different data bases determined by different seed numbers are called *different samples*.

The regression adjustments were made separately for each weight class shown in Table 1, using both the estimated and true probabilities of linkage. In Table 1, weight class 10 refers to pairs having weights between 10 and 11 and weight class -1 refers to pairs having weights between -0 and -1. All pairs having weights 15 and above are combined into class 15+ and all pairs having weights -9 and below are combined into class $-10-$. While it was possible with the Rubin-Belin results to make individual adjustments for linkage probabilities, we chose to make average adjustments, by each weight class in Table 1. (See Czajka *et al.* 1992, for discussion of a related decision. Our approach has some of the flavor of the work on propensity scores (*e.g.*, Rosenbaum and Rubin 1983, 1985). Propensity scoring techniques, while proposed for other classes of problems, may have application here as well.

4. SOME HIGHLIGHTS AND LIMITATIONS OF THE SIMULATION RESULTS

Because of space limitations, we will present only a few representative results from the simulations conducted. For more information, including an extensive set of tables, see Winkler and Scheuren (1991).

The two outcome measures from our simulation that we consider are the relative bias and relative standard error. We will only discuss the mediocre matching scenario in detail and only for the case R^2 between 0.40 and 0.45. Figures 5-7 shows the relative bias results from a single representative sample. An overall summary, though, for the other scenarios is presented in Table 2. Some limitations on the simulation are also noted at the end of this section.

4.1 Illustrative Results for Mediocre Matching

Rather than use all pairs, we only consider pairs having weights 10 or less. Use of the smaller subset of pairs allows us to examine regression adjustment procedures for weight classes having low to high proportions of true nonlinks. We note that the eliminated pairs (having weight 10 and above) are associated only with true links. Figures 5 and 6 present our results for adjusted and unadjusted regression data, respectively. Results obtained with unadjusted data are based on conventional regression formulas (*e.g.*, Draper and Smith 1981). The weight classes displayed are cumulative beginning with pairs having the highest weight. Weight class w refers to all pairs having weights between w and 10.

We observe the following:

- The *accumulation* is by decreasing matching weight (*i.e.* from classes most likely to consist almost solely of true links to the classes containing increasing higher proportions of true nonlinks). In particular, for weight class $w = 8$, the first data point shown in Figures 5-7, there were 3 nonlinks and 439 links. By the time, say, we had cumulated the data through weight class $w = 5$, there were 24 nonlinks; the links, however, had grown to 1,121 - affording us a much larger overall sample size with a corresponding reduction in the regression standard error.
- Relative *biases* are provided for the original and adjusted slope coefficient $\hat{a}_1$ by taking the ratio of the true coefficient (about 2) and the calculated one for each cumulative weight class.
- Adjusted regression results are shown employing both estimated and true match probabilities. In particular, Figure 5 corresponds to the results obtained using estimated probabilities (all that would ordinarily be available in practice). Figure 7 corresponds to the unrealistic situation for which we knew the true probabilities.
- Relative *root mean square errors* (not shown) are obtained by calculating MSEs for each cumulative weight class. For each class, the bias is squared, added to the square of the standard errors, and square roots taken.

Observations on the results we obtained are fairly straightforward and about what we expected. For example, as sample size increased, we found the relative root mean square errors decreasd substantially for the adjusted coefficients. If the regression coefficients were not adjusted,

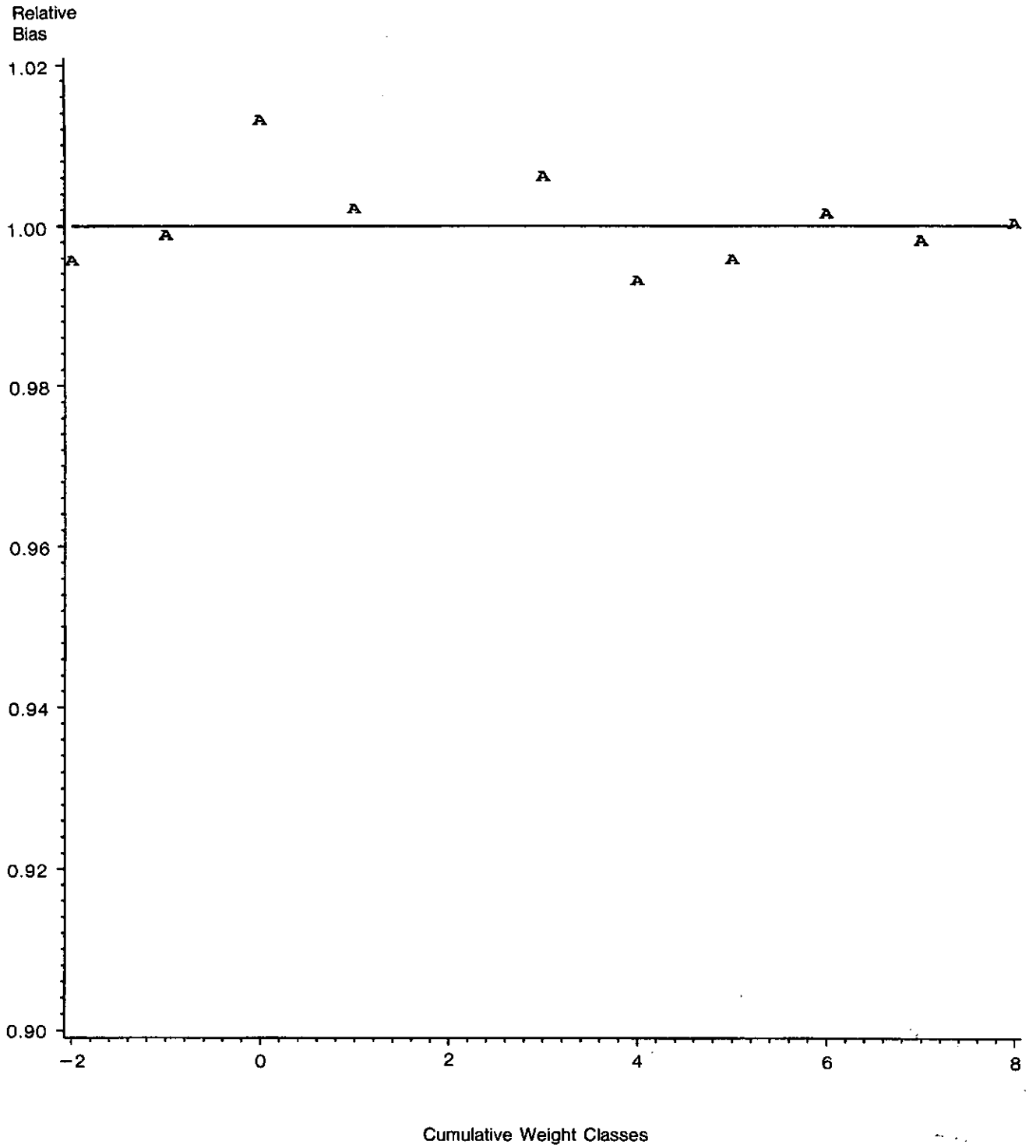


Figure 5. Relative Bias For Adjusted Estimators, Estimated Probabilities

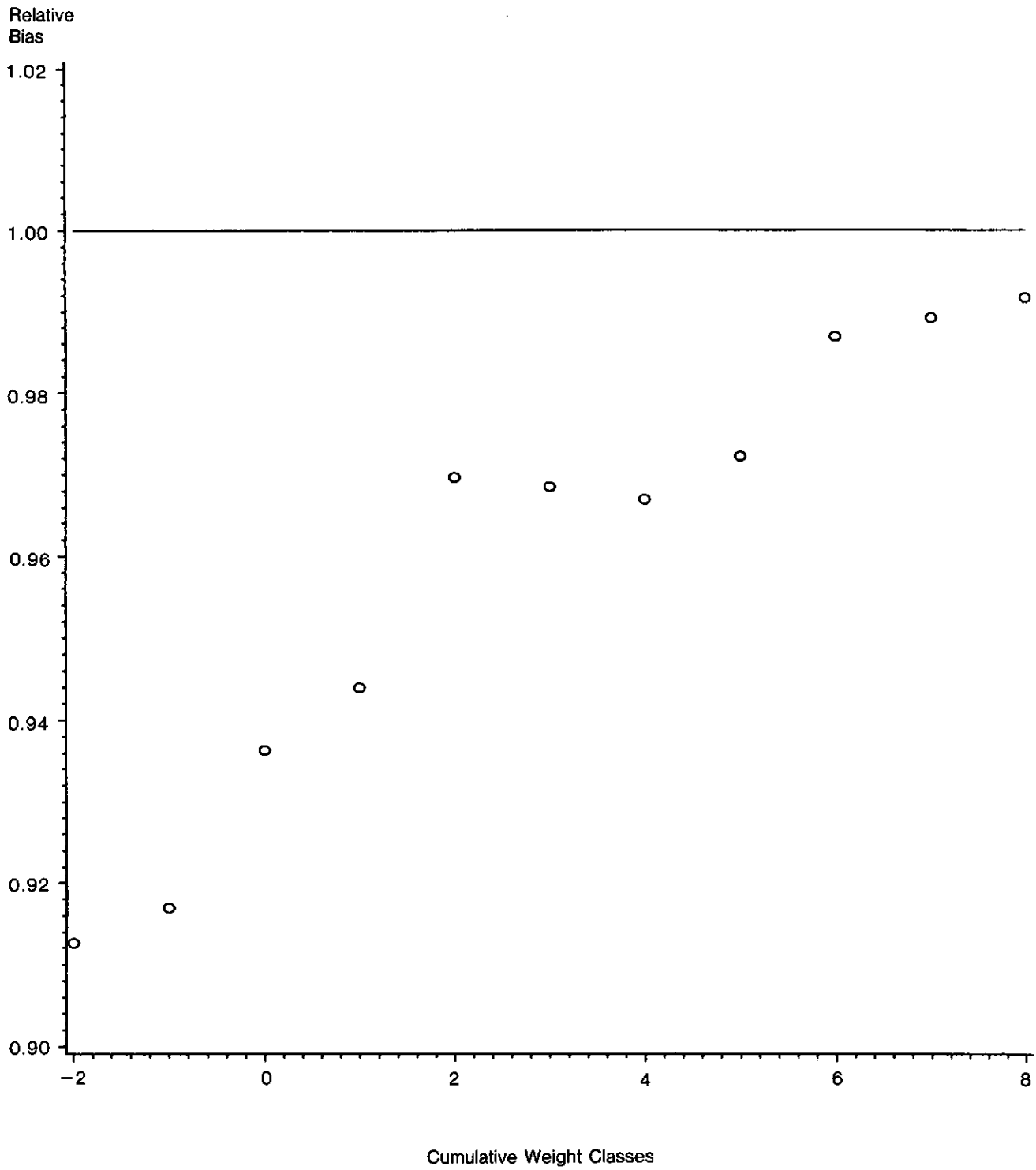


Figure 6. Relative Bias For Unadjusted Estimators

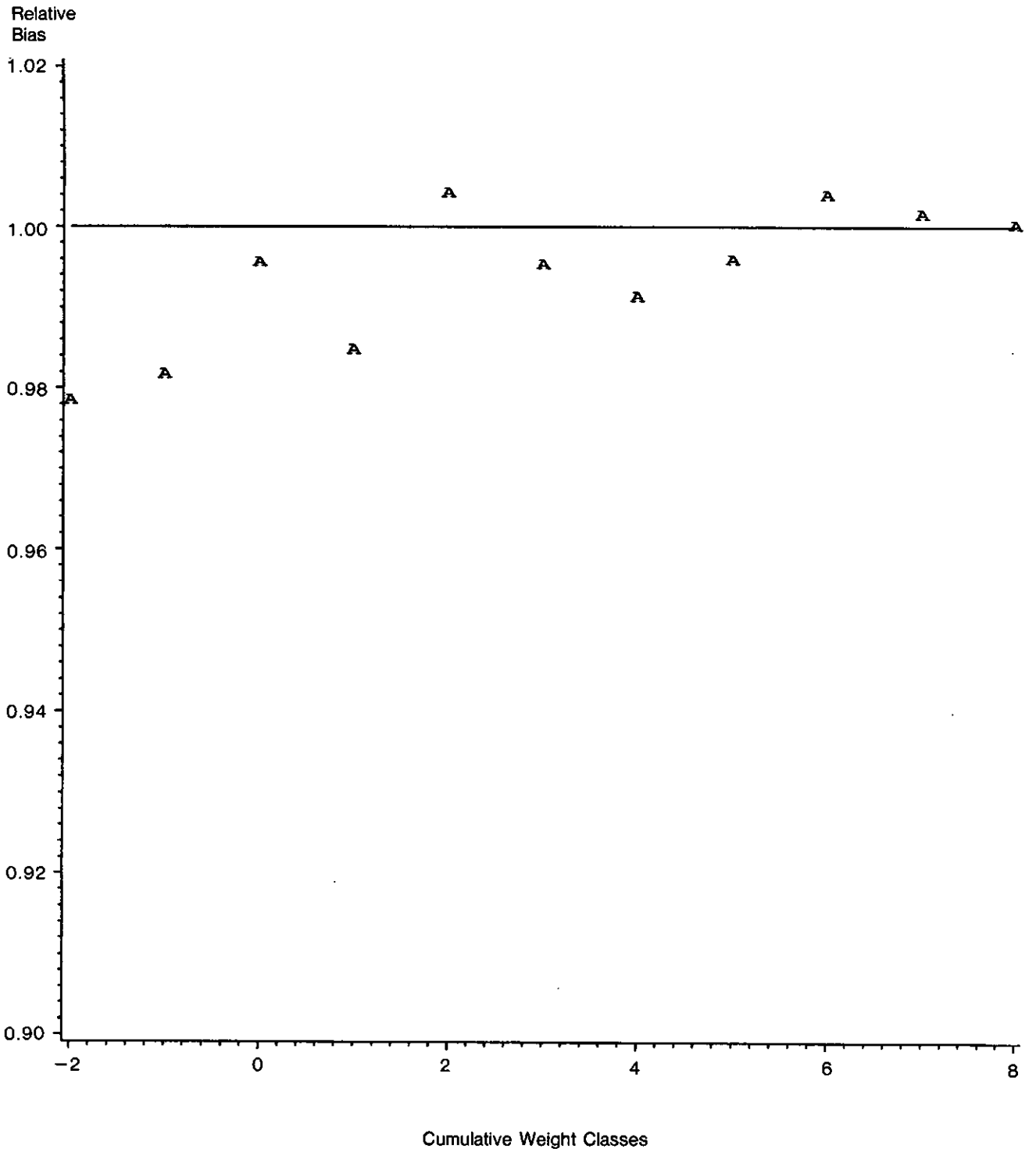


Figure 7. Relative Bias For Adjusted Estimators, True Probabilities

standard errors still decreased as the sample size grew, but at an unacceptably high price in increased bias.

One point of concern is that our ability to accurately estimate matching probabilities critically affects the accuracy of the coefficient estimates. If we can accurately estimate the probabilities (as in this case), then the adjustment procedure works reasonably well; if we cannot (see below), then the adjustment could perform badly.

4.2 Overall Results Summary

Our results varied somewhat for the three different values of R^2 - being better for larger R^2 values. These R^2 differences, however, do not change our main conclusions; hence, Table 2 does not address them. Notice that, for the good matching scenario, attempting to adjust does little good and may even cause some minor harm. Certainly it is pointless, in any case, and we only included it in our simulations for the sake of completeness. At the other extreme, even for poor matches, we obtained satisfactory results, but only when using the true probabilities - something not possible in practice.

Table 2

Summary of Adjustment Results for Illustrative Simulations

Basis of adjustments	Matching scenarios		
	Good	Mediocre	Poor
True probabilities	Adjustment was not helpful because it was not needed	Good results like those in Section 4.1	Good results like those in Section 4.1
Estimated probabilities	Same as above	Same as above	Poor results because Rubin-Belin could not estimate the probabilities

Any statistical estimation procedure will have difficulty with the poor matching scenario because of the extreme overlap of the curves. See Figure 4. We believe the mediocre scenario covers a wide range of typical settings. Nonetheless, the poor matching scenario might arise fairly often too, especially with less experienced linkers. Either new estimation procedures will have to be developed for the poor case or the Rubin-Belin probability estimation procedure - which was not designed for this situation - will have to be enhanced.

4.3 Some Simulation Limitations

The simulation results are subject to a number of limitations. Some of these are of possible major practical significance; others less so. A partial list follows:

- In conducting simulations for this paper, we assumed that the highest weight pair was a true link and the second highest a true nonlink. This assumption fails because, sometimes, the second highest is the true link and the highest a true nonlink. (We do not have a clear sense of how important this issue might be in practice. It would certainly have to be a factor in poor matching scenarios.)
- A second limitation of the data sets employed for the simulations is that the truly linked record may not be present at all in the file to which the first file is being matched. (This could be important. In many practical settings, we would expect the "logical blocking criteria" also to cause both pairs used in the adjustment to be false links.)
- A third limitation of our approach is that no use has been made of conventional regression diagnostic tools. (Depending on the environment, outliers created because of nonlinks could wreak havoc with underlying relationships. In our simulations this did not show up as much of a problem, largely, perhaps, because the X and Y values generated were bounded in a moderately narrow range.)

5. CONCLUSIONS AND FUTURE WORK

The theoretical and related simulation results presented here are obviously somewhat contrived and artificial. A lot more needs to be done, therefore, to validate and generalize our beginning efforts. Nonetheless, some recommendations for current practice stand out, as well as areas for future research. We will cover first a few of the topics that intrigued us as worthy of more study to improve the adjustment of potential links. Second, some remarks are made about the related problem of what to do with the (remaining) nonlinks. Finally, the section ends with some summary ideas and a revisitation of our perspective concerning the unity of the tasks that linkers and analysts do.

5.1 Improvements in Linkage Adjustment

An obvious question is whether our adjustment procedures could borrow ideas from general methods for errors-in-variables (*e.g.*, Johnston 1972). We have not explored this, but there may be some payoffs.

Of more interest to us are techniques that grow out of conventional regression diagnostics. A blend of these with our approach has a lot of appeal. Remember we are making adjustments, weight class by weight class. Suppose we looked ahead of time at the residual scatter in a particular weight class, where the residuals were calculated around the regression obtained from the cumulative weight classes above the class in question. Outliers, say, could then be identified and might be treated as nonlinks rather than potential links.

We intend to explore this possibility with simulated data that is heavier-tailed than what was used here. Also we will explore consciously varying the length of the weight classes and the minimum number of cases in each class. We have an uneasy feeling that the number of cases in each class may have been too small in places. (See Table 1.) On the other hand, we did not use the fact that the weight classes were of equal length nor did we study what would have happened had they been of differing lengths.

One final point, as noted already: we believe our approach has much in common with propensity scoring, but we did not explicitly appeal to that more general theory for aid and this could be something worth doing. For example, propensity scoring ideas may be especially helpful in the case where the regression variables and the linkage variables are dependent. (See Winkler and Scheuren (1991) for a report on the limited simulations undertaken and the additional difficulties encountered.)

5.2 Handling Erroneous Nonlinks

In the use of record linkage methods the general problem of selection bias arises because of erroneous nonlinks. There are a number of ways to handle this. For example, the links could be adjusted by the analyst for lack of representativeness, using the approaches familiar to those who adjust for unit or, conceivably, item nonresponse (*e.g.*, Scheuren *et al.* 1981).

The present approach for handling potential links could help reduce the size of the erroneous nonlink problem but, generally, would not eliminate it. To be specific, suppose we had a linkage setting where, for resource reasons, it was infeasible to follow up on the potential links. Many practitioners might simply drop the potential links, thereby, increasing the number of erroneous nonlinks. (For instance, in ascertaining which of a cohort's members is alive or dead, a third possibility - unascertained - is often used.)

Our approach to the potential links would have *implicitly* adjusted for that portion of the erroneous nonlinks which were potentially linkable (with a followup step, say). Other erroneous nonlinks would generally remain and another adjustment for them might still be an issue to consider.

Often we can be faced with linkage settings where the files being linked have subgroups with matching information of varying quality, resulting in differing rates of erroneous links and nonlinks. In principle, we could employ the techniques in this paper to each subgroup separately. How to handle very small subgroups is an open problem and the effect on estimated differences between subgroups, even when both are of modest size, while seemingly straightforward, deserves study.

5.3 Concluding Comments

At the start of this paper we asked two "key" questions. Now that we are concluding, it might make sense to reconsider these questions and try, in summary fashion, to give some answers.

- "*What should the linker do to help the analyst?*" If possible, the linker should play a role in designing the datasets to be matched, so that the identifying information on both is of high quality. Powerful algorithms exist now in several places to do an excellent job of linkage (*e.g.*, at Statistics Canada or the U.S. Bureau of the Census, to name two). Linkers should resist the temptation to design and develop their own software. In most cases, modifying or simply using existing software is highly recommended (Scheuren 1985). Obviously, for the analyst's sake, the linker needs to provide as much linkage information as possible on the files matched so that the analyst can make informed choices in his or her work. In the present paper we have proposed that the links, nonlinks, and potential links be provided to the analyst - not just links. We strongly recommend this, even if a clerical review step has been undertaken. We do *not* necessarily recommend the particular choices we made about the file structure, at least not without further study. We would argue, though, that our choices are serviceable.
- "*What should the analyst know about the linkage and how should this be used?*" The analyst needs to have information like link, nonlink, and potential link status, along with linkage probabilities, if available. Many settings could arise where simply doing the data analysis steps separately by link status will reveal a great deal about the sensitivity of one's results. The present paper provides some initial ideas about how this use might be approached in a regression context. There also appears to be some improvements possible using the adjustments carried out here, particularly for the mediocre matching scenario. How general these improvements are remains to be seen. Even so, we are relatively pleased with our results and look forward to doing more. Indeed, there are direct connections to be made between our approach to the regression problem and other standard techniques, like contingency table loglinear models.

Clearly, we have not developed complete, general answers to the questions we raised. We hope, though, that this paper will at least stimulate interest on the part of others that could lead us all to better practice.

ACKNOWLEDGMENTS AND DISCLAIMERS

The authors would like to thank Yahia Ahmed and Mary Batcher for their help in preparing this paper and two referees for detailed and discerning comments. Fruitful discussions were held with Tom Belin. Wendy Alvey also provided considerable editorial assistance.

The usual disclaimers are appropriate here: in particular, this paper reflects the views of the authors and not necessarily those of their respective agencies. Problems, like a lack of clarity in our thinking or in our exposition, are entirely the authors' responsibility.

APPENDIX

The appendix is divided into four sections. The first provides details on how matching error affects regression models for the simple univariate case. The approach most closely resembles the approach introduced by Neter *et al.* (1965) and provides motivation for the generalizations presented in appendix sections two and three. Computational formulas are considerably more complicated than those presented by Neter *et al.* because we use a more realistic model of the matching process. In the second section, we extend the univariate model to the case for which all independent variables arise from one file, while the dependent variable comes from the other, and, in the third, we extend the second case to that in which some independent variables come from one file and some come from another. The fourth section summarizes methods of Rubin and Belin (1991) (see also Belin 1991) for estimating the probability of a link.

A.1. Univariate Regression Model

In this section we address the simplest regression situation in which we match two files and consider a set of numeric pairs in which the independent variable is taken from a record in one file and the dependent variable is taken from the corresponding matched record from the other file.

Let $Y = X\beta + \epsilon$ be the ordinary univariate regression model for which error terms are independent with expectation zero and constant variance σ^2. If we were working with a single data base, Y would be regressed on X in the usual manner. For $i = 1, \ldots, n$, we wish to use (X_i, Y_i) but we will use (X_i, Z_i), where Z_i is usually Y_i but it may take some other value Y_j due to matching error.

That is, for $i = 1, \ldots, n$,

$$z_i = \begin{cases} Y_i & \text{with probability } p_i \\ Y_j & \text{with probability } q_{ij} \text{ for } j \neq i, \end{cases}$$

where $p_i + \sum_{j \neq i} q_{ij} = 1$.

The probability p_i may be zero or one. We define $h_i = 1 - p_i$. As in Neter *et al.* (1965), we divide the set of pairs into n mutually exclusive classes. Each class consists of exactly one (X_i, Z_i) and, thus, there are n classes. The intuitive idea of our procedure is that we basically adjust Z_i in each (X_i, Z_i) for the bias induced by the matching process. The accuracy of the adjustment is heavily dependent on the accuracy of the estimates of the matching probabilities in our model.

To simplify the computational formulas in the explanation, we assume one-to-one matching; that is, for each $i = 1, \ldots, n$, there exists at most one j such that $q_{ij} > 0$. We let ϕ be defined by $\phi(i) = j$. Our model still applies if we do not assume one-to-one matching.

As intermediate steps in estimating regression coefficients and their standard errors, we need to find $\mu_z \equiv E(Z)$, σ_z^2, and σ_{zx}. As in Neter *et al.* (1965),

$$\begin{aligned} E(Z) &\equiv (1/n)\sum_i E(Z|i) \equiv (1/n)\sum_i (Y_i p_i + \sum_{j \neq i} Y_j q_{ij}) \\ &= (1/n)\sum_i Y_i \\ &\quad + (1/n)\sum_i [Y_i(-h_i) + Y_{\phi(i)} h_i] \\ &\equiv \bar{Y} + B. \end{aligned} \tag{A.1.1}$$

The first and second equalities are by definition and the third is by addition and subtraction. The third inequality is the first time we apply the one-to-one matching assumption. The last term on the right hand side of the equality is the bias which we denote by B. Note that the overall bias B is the statistical average (expectation) of the individual biases $[Y_i(-h_i) + Y_{\phi(i)} h_i]$ for $i = 1, \ldots, n$. Similarly, we have

$$\begin{aligned} \sigma_z^2 &\equiv E(Z - EZ)^2 = E(Z - (\bar{Y} + B))^2 \\ &= (1/n)\sum_i (Y_i - \bar{Y})^2 p_i + (1/n)\sum_{j \neq i} \\ &\quad (Y_j - \bar{Y})^2 q_{ij} - 2B\,E(Z - \bar{Y}) + B^2 \\ &= (1/n)S_{yy} + B_{yy} - B^2 = \sigma_y^2 + B_{yy} - B^2, \end{aligned} \tag{A.1.2}$$

where $B_{yy} = (1/n)\sum_i [(Y_i - \bar{Y})^2(-h_i) + (Y_{\phi(i)} - \bar{Y})^2 h_i]$, $S_{yy} = \sum_i (Y_i - \bar{Y})^2$ and $\sigma_y^2 = (1/n)S_{yy}$.

$$\begin{aligned} \sigma_{zx} &\equiv E[(Z - EZ)(X - EX)] \\ &= (1/n)\sum_i (Y_i - \bar{Y})(X_i - \bar{X})p_i \\ &\quad + (1/n)\sum_{j \neq i} (Y_j - \bar{Y})(X_i - \bar{X})\, q_{ij} \\ &= (1/n)S_{yx} + B_{yx} = \sigma_{yx} + B_{yx}, \end{aligned} \tag{A.1.3}$$

where $B_{yx} = (1/n)\sum_i[(Y_i - \bar{Y})(X_i - \bar{X})(-h_i) + (Y_{\phi(i)} - \bar{Y})(X_i - \bar{X})h_i]$, $S_{yx} = \sum_i (Y_i - \bar{Y})(X_i - \bar{X})$, and $\sigma_{yx} = (1/n)S_{yx}$. The term B_{yy} is the bias for the second moments and the term B_{yx} is the bias for the cross-product of Y and X. Formulas (A.1.1), (A.1.2), and (A.1.3), respectively, correspond to formulas (A.1), (A.2), and (A.3) in Neter *et al.* The formulas necessarily differ in detail because we use a more general model of the matching process.

The regression coefficients are related by

$$\beta_{zx} \equiv \sigma_{zx}/\sigma_x^2 = \sigma_{yx}/\sigma_x^2 + B_{yx}/\sigma_x^2 = \beta_{yx} + B_{yx}/\sigma_x^2. \quad \text{(A.1.4)}$$

To get an estimate of the variance of β_{yx}, we first derive an estimate s^2 for the variance σ^2 in the usual manner.

$$(n-2)\,s^2 = \sum_i (y_i - \hat{y}_i)^2 = S_{yy} + \beta_{yx}\,S_{xy}$$
$$= n\,\sigma_y^2 - n\,\beta_{yx}\,\sigma_x^2. \quad \text{(A.1.5)}$$

Using (A.1.2) and (A.1.3) allows us to express s^2 in terms of the observable quantities σ_z^2 and σ_{zx} and the bias terms B_{yy}, B_{yx}, and B that are computable under our assumptions. The estimated variance of β_{yx} is then computed by the usual formula (*e.g.*, Draper and Smith 1981, 18-20)

$$\text{Var}(\beta_{yx}) = s^2/(n\,\sigma_x^2).$$

We observe that the first equality in (A.1.5) involves the usual regression assumption that the error terms are independent with identical variance.

In the numeric examples of this paper we assumed that the true independent value X_i associated with each Y_i was from the record with the highest matching weight and the false independent value was taken from the record with the second highest matching weight. This assumption is plausible because we have only addressed simple regression in this paper and because the second highest matching weight was typically much lower than the highest. Thus, it is much more natural to assume that the record with the second highest matching weight is false. In our empirical examples we use straightforward adjustments and make simplistic assumptions that work well because they are consistent with the data and the matching process. In more complicated regression situations or with other models such as loglinear we will likely have to make additional modelling assumptions. The additional assumptions can be likened to the manner in which simple models for nonresponse require additional assumptions as the models progress from ignorable to nonignorable (see Rubin 1987).

In this section, we chose to adjust independent x-values and leave dependent y-values as fixed in order to achieve consistency with the reasoning of Neter *et al.* We could have just as easily adjusted dependent y-values leaving x-values as fixed.

A.2. Multiple Regression with Independent Variables from One File and Dependent Variables from the Other File

At this point we pass to the usual matrix notation (*e.g.*, Graybill 1976). Our basic model is

$$Y = X\beta + \epsilon,$$

where Y is a $n \times 1$ array, X is a $n \times p$ array, β is a $p \times 1$ array, and ϵ is a $n \times 1$ array.

Analogous to the reasoning we used in (A.1.1), we can represent

$$Z = Y + B, \quad \text{(A.2.1)}$$

where Z, Y, and B are $n \times 1$ arrays having terms that correspond, for $i = 1, \ldots, n$, via

$$z_i = y_i + p_i y_i + h_i y_{\phi(i)}.$$

Because we observe Z and X only, we consider the equation

$$Z = XC + \epsilon. \quad \text{(A.2.2)}$$

We obtain an estimate $\hat{C}$ by regressing on the observed data in the usual manner. We wish to adjust the estimate $\hat{C}$ to an estimate $\hat{\beta}$ of β in a manner analogous to (A.1.1).

Using (A.2.1) and (A.2.2) we obtain

$$(X^TX)^{-1}X^TY + (X^TX)^{-1}XB = \hat{C}. \quad \text{(A.2.3)}$$

The first term on the left hand side of (A.2.3) is the usual estimate $\hat{\beta}$. The second term on the left hand side of (A.2.3) is our bias adjustment. X^T is the transpose of X.

The usual formula (Graybill 1976, p. 176) allows estimation of the variance σ^2 associated with the i.i.d. error components of ϵ,

$$(n-p)\,\hat{\sigma}^2 = (Y - X\beta)^T(Y - X\beta)$$
$$= Y^TY - \hat{\beta}\,X^TY, \quad \text{(A.2.4)}$$

where $\hat{\beta} = (X^TX)^{-1}X^TY$.

Via (A.2.1) $\hat{\beta}\,X^TY$ can be represented in terms of the observable Z and X in a manner similar to (A.1.2) and (A.1.3). As

$$Y^TY = Z^TZ - B^TZ - Z^TB + B^TB, \quad \text{(A.2.5)}$$

we can obtain the remaining portion of the right hand side of (A.2.4) that allows estimation of σ^2.

Via the usual formula (*e.g.*, Graybill 1976, p. 276), the covariance of $\hat{\beta}$ is

$$\text{cov}[\hat{\beta}] = \sigma^2\,(X^TX)^{-1}, \quad \text{(A.2.6)}$$

which we can estimate.

A.3. Multiple Regression with Independent Variables from Both Files

When some of the independent variables come from the same file as Y we must adjust them in a manner similar to the way in which we adjust Y in equations (A.1.1) and (A.2.1). Then data array X can be written in the form

$$X_d = X + D, \qquad \text{(A.3.1)}$$

where D is the array of bias adjustments taking those terms of X arising from the same file as Y back to their true values that are represented in X_d. Using (A.2.1) and (A.2.2), we obtain

$$Y + B = (X_d - D)C. \qquad \text{(A.3.2)}$$

With algebra (A.3.2) becomes

$$\begin{aligned}(X_d^T X_d)^{-1} X_d^T Y &= (X_d^T X_d)^{-1} X_d^T(-B) \\ &\quad + (X_d^T X_d)^{-1} X_d^T (X_d + D)C \\ &= (X_d^T X_d)^{-1} X_d^T(-B) \\ &\quad + (X_d^T X_d)^{-1} X_d^T DC + C. \qquad \text{(A.3.3)}\end{aligned}$$

If D is zero (*i.e.*, all independent x-values arise from a single file), then (A.3.3) agrees with (A.2.3). The first term on the left hand side of (A.2.3) is the estimate of $\hat{\beta}$. The estimate $\hat{\sigma}^2$ is obtained analogously to the way (A.2.3), (A.2.4) and (A.2.5) were used. The covariance of $\hat{\beta}$ follows from (A.2.6).

A.4. Rubin-Belin Model

To estimate the probabilty of a true link within any weight range, Rubin and Belin (1991) consider the set of pairs that are produced by the computer matching program and that are ranked by decreasing weight. They assume that the probability of a true link is a montone function of the weight; that is, the higher the weight, the higher the probability of a true link. They assume that the distribution of the observed weights is a mixture of the distributions for true links and true nonlinks.

Their estimation procedure is:

1. Model each of the two components of the mixture as normal with unknown mean and variance after separate power transformations.
2. Estimate the power of the two transformations from a training sample.
3. Taking the two transformations as known, fit a normal mixture model to the current weight data to obtain maximum likelihood estimates (and standard errors).
4. Use the parameters from the fitted model to obtain point estimates of the false-link rate as a function of cutoff level and obtain standard errors for the false-link rate using the delta-method approximation.

While the Rubin-Belin method requires a training sample, the training sample is primarily used to get the shape of the curves. That is, if the power transformation is given by

$$\psi(w_i; \delta, \omega) = \begin{cases} (w_i^\delta - 1)/(\delta\, \omega^{\delta - 1}) & \text{if } \delta \neq 0 \\ \omega \log(w_i) & \text{if } \delta = 0, \end{cases}$$

where ω is the geometric mean of the weights w_i, $i = 1, \ldots, n$, then ω and δ can be estimated for the two curves. For the examples of this paper and a large class of other matching situations (Winkler and Thibaudeau 1991), the Rubin-Belin estimation procedure works well. In some other situations a different method (Winkler 1992) that uses more information than the Rubin-Belin method and does not require a training sample yields accurate estimates, while software (see *e.g.*, Belin 1991) based on the Rubin-Belin method fails to converge even if new calibration data are obtained. Because the calibration data for the good and mediocre scenarios of this paper are appropriate, the Rubin-Belin method provides better estimates than the method of Winkler.

REFERENCES

BEEBE, G. W. (1985). Why are epidemiologists interested in matching algorithms? In *Record Linkage Techniques* – 1985. U.S. Internal Revenue Service.

BELIN, T. (1991). Using Mixture Models to Calibrate Error Rates in Record Linkage Procedures, with Application to Computer Matching for Census Undercount Estimation. Harvard Ph.D. Thesis.

CARPENTER, M., and FAIR, M.E. (Editors) (1989). *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Epidemiological Research Conference, Statistics Canada.

COOMBS, J.W., and SINGH, M.P. (Editors) (1987). *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada.

COPAS, J.B., and HILTON, F.J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society A*, 153, 287-320.

CZJAKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., and RUBIN, D.B. (1992). Evaluation of a new procedure for estimating income and tax aggregrates from advance data. *Journal of Business and Economic Statistics*, 10, 117-131.

DRAPER, N.R., and SMITH, H. (1981). *Applied Regression Analysis*, 2nd Edition. New York: J. Wiley.

FELLEGI, I.P., and SUNTER, A. (1969). A theory of record linkage. *Journal of the of the American Statistical Association*, 64, 1183-1210.

GRAYBILL, F.A. (1976). *Theory and Application of the Linear Model*. Belmont, CA: Wadsworth.

HOWE, G., and SPASOFF, R.A. (Editors) (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*. Toronto, Ontario, Canada: University of Toronto Press.

JABINE, T.B., and SCHEUREN, F.J. (1986). Record linkages for statistical purposes: methodological issues. *Journal of Official Statistics*, 2, 255-277.

JARO, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

JOHNSTON, J. (1972). *Econometric Methods*, 2nd Edition. New York: McGraw-Hill.

KILSS, B., and ALVEY, W. (Editors) (1985). *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service, Publication 1299, 2-86.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business*. Oxford: Oxford University Press.

NEWCOMBE, H.B., FAIR, M.E., and LALONDE, P. (1992). The use of names for linking personal records. *Journal of the Americal Statistical Association*, 87, 1193-1208.

NEWCOMBE, H.B., KENNEDY, J.M., AXFORD, S.J., and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

NETER, J., MAYNES, E.S., and RAMANATHAN, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.

ROSENBAUM, P., and RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

ROSENBAUM, P., and RUBIN, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39, 33-38.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: J. Wiley.

RUBIN, D.B. (1990). Discussion (of Imputation Session). *Proceedings of the 1990 Annual Research Conference*, U.S. Bureau of the Census, 676-678.

RUBIN, D., and BELIN, T. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.

SCHEUREN, F. (1985). Methodologic issues in linkage of multiple data bases. *Record Linkage Techniques - 1985*. U.S. Internal Revenue Service.

SCHEUREN, F., and OH, H.L. (1975). Fiddling Around with Nonmatches and Mismatches. *Proceedings of the Social Statistics Section, American Statistical Association*, 627-633.

SCHEUREN, F., OH, H.L., VOGEL, L., and YUSKAVAGE, R. (1981). Methods of Estimation for the 1973 Exact Match Study. Studies from Interagency Data Linkages, U.S. Department of Health and Human Services, Social Security Administration, Publication 13-11750.

TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.

WINKLER, W.E. (1985). Exact matching list of businesses: blocking, subfield identification, and information theory. In *Record Linkage Techniques - 1985*, (Eds. B. Kilss and W. Alvey). U.S. Internal Revenue Service, Publication 1299, 2-86.

WINKLER, W.E. (1992). Comparative analysis of record linkage decision rules. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.

WINKLER, W.E., and SCHEUREN, F. (1991). How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis. U.S. Bureau of the Census, Statistical Research Division Technical Report.

WINKLER, W.E., and THIBAUDEAU, Y. (1991). An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division Technical Report.

Survey Methodology, June 1993
Vol. 19, No. 1, pp. 59–79
Statistics Canada

Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption

A.C. SINGH, H.J. MANTEL, M.D. KINACK and G. ROWE[1]

ABSTRACT

In the creation of micro-simulation databases which are frequently used by policy analysts and planners, several datafiles are combined by statistical matching techniques for enriching the host datafile. This process requires the conditional independence assumption (CIA) which could lead to serious bias in the resulting joint relationships among variables. Appropriate auxiliary information could be used to avoid the CIA. In this report, methods of statistical matching corresponding to three methods of imputation, namely, regression, hot deck, and log linear, with and without auxiliary information are considered. The log linear methods consist of adding categorical constraints to either the regression or hot deck methods. Based on an extensive simulation study with synthetic data, sensitivity analyses for departures from the CIA are performed and gains from using auxiliary information are discussed. Different scenarios for the underlying distribution and relationships, such as symmetric versus skewed data and proxy versus nonproxy auxiliary data, are created using synthetic data. Some recommendations on the use of statistical matching methods are also made. Specifically, it was confirmed that the CIA could be a serious limitation which could be overcome by the use of appropriate auxiliary information. Hot deck methods were found to be generally preferable to regression methods. Also, when auxiliary information is available, log linear categorical constraints can improve performance of hot deck methods. This study was motivated by concerns about the use of the CIA in the construction of the Social Policy Simulation Database at Statistics Canada.

KEY WORDS: Categorical constraints; Conditional correlation; Log normal contaminations; Shrinkage to the mean.

1. INTRODUCTION

Statistical matching can be viewed as a special case of imputation in which we have two distinct micro-data sources containing different information on different units. One data source serves as a host or recipient file to which new information is imputed for each record using data from the other source which is the donor file. Statistical matching, however, differs from the usual problem of imputation whenever the host file contains information about additional variables which are not present in the donor file. A typical use for the matched file is as input to micro-simulation models for which a complete file with all variables is required. Available micro-datafiles may correspond to samples from administrative files or survey data. Since the records from the different source files correspond to different units, the process of merging the information from the various files is unlike exact matching in which one would search through these other data sources for specific units. In fact, even if exact matching were possible, confidentiality concerns could prevent an exact matching of the files.

A general formulation is as follows. A host file A will contain information on variables (X, Y) and a donor file B will contain information on variables (X,Z). The common variable X can be used to identify similar units in the two files. The problem is to complete the records in file A by imputing live values for Z, using the information on the (X,Z) relationship in file B. In practice, the variables X, Y, and Z would generally be multivariate. An important advantage of imputing live values of Z is that relationships among components of multivariate Z are preserved. Throughout this paper, it will be assumed, for convenience, that X, Y and Z are univariate.

The Social Policy Simulation Database (SPSD; see Wolfson *et al.* 1987), a micro-simulation database created at Statistics Canada, provides an important application of statistical matching for use in economic policy analysis, *e.g.*, calculations of taxes and transfers for families on the database. The multistage construction process of the SPSD uses the technique of statistical matching at a number of points in order to enrich the host datafile, the Survey of Consumer Finance (SCF), with additional information from other data sources. Specifically, information from unemployment insurance claim histories, personal income tax returns, and the Family Expenditure Survey is added to the SCF records. If file A corresponds to the SCF and file B to the tax file, then X variables may represent

[1] A.C. Singh, H.J. Mantel and M.D. Kinack, Social Survey Methods Division; G. Rowe, Social and Economic Studies Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

demographic and economic variables, Y may denote transfer income, and Z may correspond to tax liability, investment income and deductions.

Statistical matching, as described above, suffers from a serious limitation in that information on the variable Y is completely ignored. This limitation amounts to the assumption of conditional independence of Y and Z given X ($Y \perp Z \mid X$), denoted CIA (conditional independence assumption). The importance of the CIA is obvious, since the purpose of the match is to analyze the joint relationships of X, Y and Z. If the true relationships of the variables are such that conditional independence does not hold, then the CIA would mask an important component of these relationships, and would bias some analyses involving the full set of variables. The potential seriousness of the CIA was noted by Sims (1978) and Rubin (1986), and, although statistical matches based on the CIA are not necessarily seriously flawed, Paass (1986) and Armstrong (1989) offer some empirical evidence that the problem is often real. The present study, in fact, is motivated from considerations of improving the content of the SPSD which assumes the CIA for the process of statistical matching; see also comments of Scheuren (1989) on the methodology used in the SPSD.

The literature on statistical matching extends over more than two decades. Early references are Budd and Radner (1969), Budd (1971) and Okner (1972). Sims (1972), in his comments on Okner's paper, was the first to point out the potential risk of statistical matching because of the implicit conditional independence assumption. Concerns were also expressed by Fellegi (1977) about the validity of joint distributions in the matched file and he suggested that thorough empirical testing of matching methods should be done. U.S. Department of Commerce (1980) provides a good review of statistical matching as well as exact matching methods; see also Kadane (1978) and Rodgers (1984). Barr and Turner (1990) describe a detailed empirical investigation of quality issues for file merging, and also present a good list of references. For a more recent review see Cohen (1991).

In this paper we consider the use of auxiliary information as an alternative to the CIA in statistical matching. Thus, it is assumed that there exists a third file C representing auxiliary information about the full set (X,Y,Z) or the reduced set (Y,Z). This information could be outdated, proxy (*i.e.* different but similar variables), or in the form of frequency tables and could come from small scale specially conducted surveys or from confidential datafiles. We wish to complete records in file A by adding Z from file B using information from files A, B, and C on the joint relationships of X, Y, and Z. A measure of success would be the extent to which the Z values on the completed file A could reasonably have come from the true underlying distribution conditional on X and Y. In the context of a simulation study we can compare the matched Z values to the suppressed true Z values by evaluation measures at the unit level or at the aggregate level. Some examples of unit level measures are mean absolute distance from the true Z values and the deviation of conditional covariance, $\mathrm{Cov}(Y,Z \mid X)$, from the true value. Some examples of aggregate level evaluation measures are chi-square distance and P-values based on likelihood ratio tests for categorical distributions. It is often the case in practice that the completed file A is used to produce cross-classified tables of counts and, therefore, the aggregate level measures based on categorical distributions would generally be of main interest. Moreover, for any arbitrary distribution for (X,Y,Z), which could be quite complex in practice, the categorical transformation provides a simple unified approach for summarizing the joint distribution.

The statistical matching problem as mentioned above is clearly important from practical considerations. In practice, for a given problem the matching method should be appropriately chosen for the type of auxiliary information available. The methods proposed earlier in the literature are mainly due to Rubin (1986) and Paass (1986). Rubin proposed versions of parametric regression while Paass proposed versions of nonparametric regression. These are related respectively to the familiar regression (REG) and hot deck (HOD) methods of imputation.

Rubin's method (a version of which is denoted in this paper by REG*) basically consists of first finding an intermediate value, Z_{int}, from the regression predictor of Z on X and Y (obtained by using information about the unconditional correlation $\rho_{Y,Z}$ or the conditional correlation $\rho_{Y,Z|X}$ from file C) and then a live Z-value is determined from file B using hot deck with (X,Z) Euclidean distance; see Section 3 for details. If the form of the regression predictor function is known, then the REG* procedure for statistical matching could be easily implemented in practice. However, finding a suitable predictor for Z is in general not easy, especially when Z is multivariate. Moreover, if information in file C is in the form of a categorical distribution, which may be quite common in practice, the REG* method would not be applicable.

Paass's method (a version of which is denoted in this paper by HOD*), on the other hand, basically consists of first finding an intermediate value, Z_{int}, from file C by hot deck imputation (with Y- or (X,Y)-distance as the case may be) and then a live Z-value from file B is obtained using again hot deck with (X,Z)-distance. This is a simplified version of the original Paass's method which is iterative such that values of Z for file A, Y for file B, and X for file C (assuming C has only (Y,Z) information) are updated successively using files C, A and B respectively until some convergence criterion is satisfied; see Section 3 for details. To start the iteration, initial values of Z for A, Y for B and X for C are imputed suitably. In the evaluation study considered in this paper, we have considered only the simplified version of Paass's method due to the

considerable computational effort required for the original method. As in the case of the REG* method, the HOD* method is not applicable if file C is in the form of a frequency table. Moreover, even if file C contains micro-data but its size is small (as in the case of a small scale specially conducted survey) or it is proxy or outdated, it may be better to extract some macro-level information such as the categorical distribution based on a fairly coarse partition.

It may be remarked that in the absence of auxiliary information, *i.e.* file C, both REG* and HOD* methods reduce simply to the usual methods of imputation, namely regression (REG) and hot deck (HOD). As part of the evaluation study, these methods are also included.

We propose modifications of Rubin's and Paass's methods, denoted by REG.LOGLIN* and HOD.LOGLIN* respectively, which are based on the log linear method of imputation as introduced by Singh (1988). The proposed modifications use auxiliary information to impose categorical constraints on the matched files obtained from REG* and HOD* methods. In this way, categorical association parameters (estimated via log linear modelling) which measure departure from conditional independence (in the categorical sense) are preserved in the matched file. These categorical constraints are expected to render joint distributions for the completed file A data robust to inferior quality or imperfect nature of the auxiliary data from file C. If auxiliary information is in the form of a categorical distribution and not at the micro-level, then CIA based matching methods can be modified by imposing categorical constraints; in this case the CIA is being used only within X, Y categories. For example, with the usual methods of imputation REG and HOD, which could be used to match by ignoring Y, we can get the corresponding modified versions as REG.LOGLIN and HOD.LOGLIN. These two methods are also considered in this paper.

Note that the categorically constrained matching methods are different from the usual constrained statistical matching methods where the constraints are in the form of a few characteristic measures from file B (such as mean and variance) that variables in the matched file must satisfy. Another key distinction is that the usual constrained matching methods focus on the marginal distribution of Z, whereas the focus here is on the conditional distribution, albeit categorical, which is more relevant for file A; thus there is a basic difference between the two approaches to constrained matching.

Following Rubin (1986) and Paass (1986), we investigate the performance of matching methods empirically. A Monte Carlo study was carried out to investigate the effect of the proposed modifications to the existing methods for the two cases, with and without auxiliary information. This would allow analysis of sensitivity to failure of the CIA and gains from using auxiliary data. The synthetic data for the simulation study was generated from multivariate normal distributions with some log normal contamination to induce asymmetry. An important advantage of using synthetic data is that relevant control parameters could be modified to yield different distributional scenarios for the matching problem. Eight methods (four existing ones, REG, REG*, HOD, HOD*, and four proposed ones, REG.LOGLIN, REG.LOGLIN*, HOD.LOGLIN, HOD.LOGLIN*) were compared by four evaluation measures (two at the unit level and two at the aggregate level) as mentioned earlier; see Section 6 for details. The main findings of the empirical study can be summarized as follows.

(i) Use of auxiliary information to avoid the CIA could considerably improve the quality of the matched file. However, if there is no auxiliary information, then among CIA based methods (*i.e.* REG and HOD), the HOD method has better overall performance. Furthermore, an interesting finding was that for small departures from conditional independence, use of auxiliary information may not improve performance of the HOD method with respect to aggregate level evaluation measures. This should have important practical implications in the absence of readily available auxiliary information.

(ii) The REG* method has very favourable performance with respect to unit level measures. By contrast, it has extremely unfavourable performance with respect to aggregate level measures. This is probably due to the shrinkage towards the mean phenomenon for regressions procedures.

(iii) The HOD* method does considerably better than REG* at the aggregate level but performs, in general, marginally worse than REG* at the unit level.

(iv) Categorical constraints, in general, improve performance of REG* and HOD* methods. Specifically, the REG.LOGLIN* method shows slight improvement at the aggregate level, but HOD.LOGLIN* shows considerable improvement at the aggregate level. Their performances at the unit level remain essentially unaffected.

(v) At the aggregate level, the HOD.LOGLIN method based only on categorical auxiliary information performs generally better than HOD.LOGLIN* based on micro-level auxiliary information. At the unit level, however, HOD.LOGLIN shows marginal deterioration in comparison to HOD.LOGLIN*. This finding may be important from practical considerations because HOD.LOGLIN is computationally much less demanding than HOD.LOGLIN* and does not require micro-level auxiliary information. The REG.LOGLIN method does not have such favourable performance, probably again due to the shrinkage to the mean effect.

(vi) If the auxiliary data is outdated or proxy, there may still be gain in using it. In this context, the HOD.LOGLIN method performs quite favourably and in fact, has fairly robust behaviour with respect to imperfect auxiliary information. Note that since this method uses only information about categorical associations from auxiliary data, it would seem reasonable for this to be affected only slightly by a limited degree of outdatedness or proxyness in file C. The REG.LOGLIN method, however, does not share this property.

It should be noted that there have been several empirical investigations in the past to evaluate statistical matching methods. Among those that do not consider the use of auxiliary information, some main references are Ruggles, Ruggles and Wolff (1977), Paass and Wauschkuhn (1980), Barr, Stewart and Turner (1981) and Rodgers and DeVol (1982). Paass (1986) provides an excellent review of these empirical tests on the quality of matching methods.

All of the studies cited above confirmed the seriousness of the CIA. This stresses the need for additional information to be incorporated in the matching process. There have been few empirical studies considering the use of auxiliary information and the impact of the CIA; Paass (1986) considered an evaluation with synthetic data only, whereas Armstrong (1989) considered simulations with both synthetic and real data. The present study could be considered as complementary to these studies in the sense that some new methods are included and the choice of underlying population distributions is reasonably broad.

The organization of this paper is as follows. Section 2 describes different types of auxiliary information. A brief review of alternative matching methods using auxiliary information is given in Section 3 and the proposed modifications using categorical constraints are described in Section 4. Different types of matching methods are illustrated in Section 5 by means of a simple numerical example. The description of the design of the empirical study on the proposed matching methods is given in Section 6 and the discussion of results in Section 7. Finally, Section 8 contains concluding remarks and some directions for further research.

2. TYPES OF AUXILIARY INFORMATION

Although a current and sufficiently large micro-datafile with information on the full set of variables is not available, it may be the case that an additional auxiliary source exists containing information on some of the joint relationships of either the full set of variables (X,Y,Z) or perhaps the reduced set (Y,Z). When this is the case it can be incorporated into the matching process to avoid the CIA and improve the quality of the completed file by reducing distortions in the joint relationships in the matched file.

Such auxiliary information may emanate from various possible sources and may reside in several different forms. Since the purpose of the auxiliary information is only to aid in avoiding the CIA, we limit its use in that information from the host or donor files is never overridden or modified by the auxiliary information. In other words, the objective is to borrow additional information from the auxiliary source not available in the source files. This is accomplished in such a way that confidentiality concerns associated with the auxiliary source would not be violated and implies that the auxiliary source could be a specially conducted small scale survey or a confidential datafile.

Another implication is that the auxiliary information need not be perfect. That is, it may be deficient in some sense. For instance, it may come from an outdated data source (perhaps a previous census or survey), but from which the required auxiliary information may still be valid, or at least represent an improvement over the otherwise default CIA. On the other hand, the auxiliary information may refer to a set of proxy variables expected to behave similarly to the variables of interest.

Auxiliary information could be at the macro-level or micro-level. At the macro-level, it could take the form of either correlations or categorical cell proportions or possibly some other parameters. If the auxiliary information in file C is on the conditional correlation of Y and Z given X, *i.e.* $\rho_{Y,Z|X}$, it can be used with the (X,Y) and (X,Z) correlations from files A and B to estimate the unconditional correlation of (Y,Z) using

$$\rho_{Y,Z} = \rho_{X,Y}\,\rho_{X,Z} + \rho_{Y,Z|X}\,(1-\rho_{X,Y}^2)^{1/2}\,(1-\rho_{X,Z}^2)^{1/2}. \qquad (2.1)$$

Now data from files A and B can be used to obtain a linear regression of Z on X and Y for the REG* method (see Section 3.1). If auxiliary information on only the unconditional correlation of Y and Z is available, then it can also be used in a similar manner.

The second type of macro-level auxiliary information from file C would be in the form of a categorical distribution for (X^*,Y^*,Z^*) where '*' denotes the categorical transformation of the original variable. If some variables were categorical to begin with, then it may not be necessary to change them. The frequency table required for categorically constrained matching methods can be obtained by raking the (X^*,Y^*,Z^*) table corresponding to file C such that its marginal tables (X^*,Y^*) and (X^*,Z^*) match respectively with the (X^*,Y^*) table from file A and (X^*,Z^*) table from file B. Note that the (X^*,Z^*) table from file B would have to be raked first to match its X^* marginal with that from file A. The method of raking preserves the (Y^*,Z^*) and (X^*,Y^*,Z^*) associations of the (X^*,Y^*,Z^*) table from file C in deriving the categorical constraints. The above adjustment of the (X^*,Y^*,Z^*) table

from file C is reasonable on the grounds that information about the (X^*,Y^*) distribution from file A and about the (X^*,Z^*) distribution from B are believed to be more precise or appropriate than those from file C. If only the (Y^*,Z^*) distribution is available (or used) from file C, then the above raking procedure could be modified to obtain suitable categorical constraints. In this case, the (Y^*,Z^*) association from file C would be preserved and the three factor (X^*,Y^*,Z^*) association term would be assumed to be zero. To achieve this, first the (X^*,Z^*) table from B is raked as before to match the X^* margin from A and then the (Y^*,Z^*) table from C is raked to match the Y^* margin from A and the Z^* margin from B. Then, a three dimensional table of ones is raked to match the (X^*,Y^*) table from A, the adjusted (X^*,Z^*) table from B and the adjusted (Y^*,Z^*) table from C. The categorical counts obtained by these procedures need not be integer values. They are rounded randomly by redistributing fractional counts by sampling cells randomly without replacement with probabilities proportional to the fractions for each cell. This is done independently for each (X^*,Y^*) category.

The next section elaborates on the use of auxiliary information in statistical matching. It also describes the use of auxiliary micro-data. In most cases when micro-level auxiliary information is available, it is possible to roll it up to the macro-level and obtain reliable information on correlations and categorical cell proportions. The validity and reasonableness of this would depend in part on the size of the micro-level datafile.

3. REVIEW OF ALTERNATIVE STATISTICAL MATCHING METHODS

3.1 The Regression Method

We first describe a regression method which uses auxiliary information. This is a version of the method due to Rubin (1986). A parametric form of the regression of Z on X and Y is assumed and the corresponding parameters are then estimated from data in files A, B, and C. For example, in the case of a linear regression, we have the model

$$E(Z \mid X,Y) = \beta_0 + \beta_1 X + \beta_2 Y,$$

$$V(Z \mid X,Y) = \sigma^2, \qquad (3.1)$$

where β_0, β_1, and β_2 are estimated from equations similar to the usual least squares equations by combining information from files A, B, C suitably. Below we describe a procedure for doing this which is somewhat different from the one described in Rubin (1986). If file C has (X,Y,Z) information, then estimates can be obtained of the conditional correlation $\rho_{Y,Z|X}$ from C, the correlation $\rho_{X,Z}$, mean μ_Z, and standard deviation σ_Z from B and the correlation $\rho_{X,Y}$, means μ_X, μ_Y, and standard deviations σ_X,σ_Y from A. Thus file B will be used only if file A is deficient in information about the quantity of interest and file C will be used for some information only when A and B are deficient. Thus we assume a hierarchy of reliability or relevance of the files A, B, and C. Such a hierarchy was not assumed by Rubin. We can then get the required estimates from

$$\beta_2 = \rho_{Y,Z|X}\frac{\sigma_{Z|X}}{\sigma_{Y|X}}, \quad \beta_1 = \rho_{X,Z|Y}\frac{\sigma_{Z|Y}}{\sigma_{X|Y}},$$

$$\beta_0 = \mu_Z - \beta_1\mu_X - \beta_2\mu_Y, \qquad (3.2)$$

where

$$\sigma_{Z|X} = (1 - \rho^2_{X,Z})^{1/2}\sigma_Z, \quad \sigma_{Y|X} = (1 - \rho^2_{X,Y})^{1/2}\sigma_Y,$$

$$\sigma_{Z|Y} = (1 - \rho^2_{Y,Z})^{1/2}\sigma_Z, \quad \sigma_{X|Y} = (1 - \rho^2_{X,Y})^{1/2}\sigma_X, \qquad (3.3)$$

and $\rho_{X,Z|Y}$ is obtained from the standard formula after first calculating $\rho_{Y,Z}$ from (2.1), *i.e.*

$$\rho_{X,Z|Y} = (\rho_{X,Z} - \rho_{X,Y}\rho_{Y,Z})(1 - \rho^2_{X,Y})^{-1/2}. \qquad (3.4)$$

It may be noted that under the normality assumption, departures from conditional independence are parametrized by $\rho_{Y,Z|X}$. Under conditional independence, $\rho_{Y,Z|X} = 0$ and the model (3.1) reduces to the simple linear regression of Z on X, *i.e.*

$$E(Z \mid X) = \beta_0 + \beta_1 X, \quad V(Z \mid X) = \sigma^2, \qquad (3.5)$$

which can be specified by combining information from files A and B or from file B alone. The formulas (3.2) reduce to

$$\beta_2 = 0, \quad \beta_1 = \rho_{X,Z}\frac{\sigma_Z}{\sigma_X}, \quad \beta_0 = \mu_Z - \beta_1\mu_X. \qquad (3.6)$$

For the case when file C contains information about $\rho_{Y,Z}$ only, the parameters of (3.1) can be easily estimated in a similar manner by combining information from A, B and C.

After the regression model is determined, the REG* method can be applied in the following two steps. Step II is important because we want to have live values of Z so that relationships among components of multivariate Z are preserved.

REG* (Step I) For each (X,Y) in A, find an intermediate value Z_{int} from the regression model (3.1).

REG* (Step II) Replace each (X,Y,Z_{int}) obtained in Step I with (X,Y,Z_{match}) where Z_{match} denotes a live Z-value from B which is nearest under the Euclidean distance in (X,Z) where the components X and Z would be scaled by their respective standard deviations. In other words, the hot deck distance method is used to find the live value. This was termed "regression with predictive mean matching" by Rubin; see Little and Rubin (1987).

Another point of departure from the method described by Rubin (1986) is that in his method a predicted Y is found for records on file B using an equation analogous to (3.1) and then corresponding predicted Z values are found; then records on file A are matched to records on file B based on the difference in predicted Z-values.

If auxiliary information is not available then the matching method REG under CIA can be used. The two steps are

REG (Step I) For each (X,Y) in A, find Z_{int} from the simple regression model (3.5).

REG (Step II) Same as in REG*.

The method described by Rubin (1986) differs in that a predicted Z is also obtained for records on file B using (3.5), and then records on file A are matched to records on file B based on the difference in predicted Z-values. In the present example, where X is univariate, this is equivalent to matching on X.

3.2 The Hot Deck Method

We first describe a hot deck method using auxiliary data. This is a version of the method due to Paass (1986). Here, ideas of nonparametric regression are used. In parametric regression, the conditional distribution of Z given X and Y is specified in a wide sense by mean and variance functions in terms of a few parameters. In nonparametric regression the techniques of nonparametric density estimation are used to estimate the conditional distribution itself. For instance, in the case of the nearest neighbour method of density estimation, for each (X,Y), K nearest neighbours (with respect to a distance function such as the Euclidean distance in (X,Y) are determined and then the conditional distribution is represented by this sample (possibly weighted) of the K neighbours where K is an integer specified suitably. Thus, $P(Z \in U \mid X,Y)$ can be specified as a conditional expectation,

$$E(I_U(Z) \mid X,Y) = \sum_{i=1}^{K} w_i(X,Y)\, I_U(Z_i), \qquad (3.7)$$

where w_i's denote weights which decrease with growing distance of (X_i,Y_i) from (X,Y) and I_U is the indicator function for the set U.

In Paass's method, first the conditional distribution of Z for each (X,Y) in A is determined by representing it with a set of K Z-values using nonparametric regression. In other words, K Z-values are added to each (X,Y). Then for each (X,Y) in A, a single live Z-value, Z_{match}, from B is obtained which is nearest under (X,Z)-distance. This gives the matched file with (X,Y,Z_{match}). The conditional distributions for file A are obtained by an iterative process in the case of file C with (Y,Z) information, as follows. Choose K initial values for nearest neighbours for Z in file A, for Y in file B, and for X in file C. This can be done by the usual hot deck method of imputation. Now each cycle consists of determining conditional distributions for elements (X,Y) in A from information in C, *i.e.* suitably updating K Z-values in A from Z-values in C using (X,Y) distance, and then updating K Y-values in file B from those of file A using (X,Z) distance, and finally updating K X-values in C from those of file B using (Y,Z) distance. This cycle is repeated until the maximal difference between some statistics for the three-dimensional distribution of (X,Y,Z) of successive iterations (*e.g.* covariance matrix) falls below a given threshold. At convergence, each file has K added values representing respective conditional distributions. In the other case in which file C has information about (X,Y,Z) the process becomes noniterative. We simply use file C to get K Z-values for A using (X,Y) distance and then get Z_{match} from B for each (X,Y) in A using (X,Z)-distance. This case was, however, not considered by Paass.

In the empirical study considered in this paper we did not use the above iterative version of Paass's method when file C had (Y,Z) data, because of its computationally intensive nature. Instead, we used a simplified noniterative version with $K = 1$. This method, denoted by HOD*, consists of the following two steps.

HOD* (Step I) For each (X,Y) in A, find an intermediate value Z_{int} from C using hot deck with Y-distance in the case of (Y,Z) auxiliary information and with (X,Y) Euclidean distance in the case of (X,Y,Z) auxiliary information.

HOD* (Step II) Replace each (X,Y,Z_{int}) obtained in Step I by (X,Y,Z_{match}) where Z_{match} is obtained from B using hot deck with (X,Z) Euclidean distance.

If file C were not available, then the matching method HOD under CIA can be used. The two steps for HOD are

HOD (Step I) Determine suitable X-categories as in usual hot deck imputation.

HOD (Step II) For each (X,Y) in A, impute a live Z-value from the corresponding X-category from B using hot deck with X-distance.

4. THE PROPOSED MODIFICATIONS BY CATEGORICALLY CONSTRAINED MATCHING

We propose modifications to REG, REG*, HOD and HOD* matching methods by imposing categorical constraints on the Z-values selected from B for completing A. The purpose of these constraints is to preserve categorical associations (as defined by log linear modelling) under a suitable partition of (X,Y,Z) for the matched file. These

associations are obtained by combining information from A, B and C. The idea of categorically constrained matching is based on the method of log linear imputation (*cf.* Singh 1988, Singh *et al.* 1988). Here the constraints could be based on auxiliary information which could be used to estimate the categorical conditional distribution, or some aspects of it, but which would not be of sufficient quality to estimate the full conditional distribution.

We start with a suitable partition of X,Y and Z variables. Let X^*, Y^*, Z^* denote the corresponding categorically transformed variables. Now the distribution of cell proportions for the (X^*,Y^*,Z^*) table can be parametrized by a log linear model

$$\log p_{ijk} = u + u_{1i} + u_{2j} + u_{3k} + u_{12ij} + u_{13ik} + u_{23jk} + u_{123ijk}, \qquad (4.1)$$

where p_{ijk} denotes the proportion for (i,j,k)th cell and 1, 2, 3 denote respectively X^*, Y^*, and Z^*. It should be noted that the parametrization (4.1) holds for arbitrary underlying distributions of the original variables (X,Y,Z). The files A and B, of course, do not contain any information about the two-factor effects u_{23} and three-factor effects u_{123}. If these are set to zero, this amounts to assuming CIA in the categorical sense, *i.e.* $Y^* \perp Z^* \mid X^*$. However, with auxiliary information in file C, this assumption can be avoided because the parameters u_{23} and u_{123} could be estimable from C. Thus, regardless of the form of the joint distribution of (X,Y,Z), the above log linear modelling provides a unified approach for gauging departures from CIA at least in the categorical sense. In the linear regression approach, on the other hand, departures from CIA are parametrized by $\rho_{Y,Z|X}$ only in the case of normality.

As was explained in Section 2, the auxiliary information from file C (either on (Y,Z) or on (X,Y,Z)) is first used to construct categorical constraints in the form of a (X^*,Y^*,Z^*) distribution. This is done by means of raking such that u_{23} and u_{123} effects from file C are preserved. The categorically constrained version of REG*, denoted by REG.LOGLIN*, can now be defined by the following two steps.

REG.LOGLIN* (Step I) Same as in Step I of REG*.

REG.LOGLIN* (Step II) Same as in Step II of REG* except that categorical constraints are imposed, implying that match order is required when obtaining live Z-values from B. We first find the match with minimum distance in (X,Z). The (X^*,Y^*,Z^*) category of the completed record would be noted and if the resulting number of matched records in that (X^*,Y^*,Z^*) category does not exceed the count imposed by the categorical constraints that match is allowed. Otherwise, that match is rejected and the match with the second smallest distance is examined. The process continues until file A is completed, and then the distribution of (X^*,Y^*,Z^*) in the completed file must satisfy the categorical constraints.

Similarly, the categorically constrained version of HOD*, denoted by HOD.LOGLIN*, consists of the following two steps.

HOD.LOGLIN* (Step I) For each (X,Y) in A, find an intermediate value, Z_{int}, from C using hot deck with Y- or (X,Y)-distance as the case may be such that the categorical constraints are satisfied. This step is similar to Step II of REG.LOGLIN*.

HOD.LOGLIN* (Step II) For each (X,Y,Z_{int}), a live value, Z_{match}, from B is determined using hot deck with (X,Z)-distance while respecting the category of Z_{int}.

An alternative approach for HOD.LOGLIN* would have been to impute an intermediate Z_{int} without constraints and then to use categorically constrained distance matching to get a live value from file B, as in Step II of REG.LOGLIN*. This was also tried but did not work well so it was dropped from the study becuase of computational burden. One possible explanation for its poor performance is shrinkage to the mean for the Z_{int} values from file C due to file C being too small. That is, the Z_{int} values would tend to be near the centre of the distribution and when the categorical constraints are then imposed the final Z values would tend to be clumped at the inside boundaries of the outer Z categories.

Suppose file C has information only at the macro-level in the form of a categorical distribution, or the micro-level information in C is considered unreliable but the information in the categorical distribution under a somewhat coarse partition is considered reliable. We can then define categorically constrained versions of the REG and HOD methods, to be denoted by REG.LOGLIN and HOD.LOGLIN respectively. The two steps for REG.LOGLIN are

REG.LOGLIN (Step I) Same as in Step I of REG.

REG.LOGLIN (Step II) Same as in Step II of REG.LOGLIN*.

Similarly, HOD.LOGLIN consists of the following two steps.

HOD.LOGLIN (Step I) Same as in Step I of HOD.

HOD.LOGLIN (Step II) Same as in Step II of REG.LOGLIN* except that no intermediate values Z_{int} exist, so that matching is based on X-distance instead of (X,Z)-distance.

For both REG.LOGLIN and HOD.LOGLIN, which do not require micro-level information on file C, the CIA is being used only within X,Y categories. Thus a reduced form of conditional independence is being assumed and the consequences of this assumption should not be as severe as those of the full CIA.

5. AN ILLUSTRATIVE EXAMPLE

Before we investigate the empirical properties of the proposed modifications in relation to the previously proposed methods, it may be instructive to consider a simple numerical example to illustrate the types of computation involved with the eight methods. Suppose files A, B and C are as shown in Table 1 which are based on random samples drawn from a multivariate normal with mean 0 and covariance matrix specified by $\sigma_X = \sigma_Y = \sigma_Z = 1$, $\rho_{X,Y} = \rho_{X,Z} = .5$ and $\rho_{Y,Z} = .7$ (which implies that $\rho_{Y,Z|X} = .6$). Here, file C is assumed to have only (Y,Z) information. For file A, Z-values are suppressed in Table 1 but are shown in Table 3 for computing evaluation measures. Suppose we employ, for simplicity and in view of small file sizes, a rather coarse categorical transformation for X,Y,Z by considering only two categories, $(-\infty,0)$ and $[0,\infty)$. Then, the three two dimensional count tables corresponding to files A, B and C can be constructed as in Table 2(a). Table 2(b) shows the adjusted tables for B and C so that they match the appropriate marginals as described in Section 2. Table 2(c) gives the three-dimensional table obtained after raking and Table 2(d) gives the desired categorical constraints after random rounding of entries of Table 2(c) as explained earlier in Section 2.

The eight methods were applied to the data of Table 1 and the matching results are shown in Table 3 along with the true values of Z which were suppressed in Table 1.

The evaluation measures shown in Table 3 were briefly introduced earlier in the introduction and are fully explained in the next section. The categorical partition for the χ^2 measure was the same as the one used for deriving categorical constraints. Note that since the partitioning is not changed for evaluation, the χ^2 values for M3, M4, M7 and M8 would be identical. It should be pointed out that the evaluation measures are given only for the sake of illustrating the calculation and should not be construed as indicators for the relative performance of various methods because they are based on just one small sample realization.

The method M8 (HOD.LOGLIN*) happens to be the most computationally intensive, the details of which are shown in Table 4. From this, it would be relatively easy to visualize the computational steps required for other methods.

Table 1

Data for Files A, B, C

Record Identifier	File A		Record Identifier	File B		Record Identifier	File C	
	X	Y		X	Y		Y	Z
A1	−0.86	−0.32	B1	−0.95	−0.69	C1	−0.40	−0.60
A2	−0.77	−0.33	B2	−0.64	−0.83	C2	−2.33	−2.81
A3	−0.09	−0.26	B3	−1.58	−0.11	C3	−0.79	−0.47
A4	−0.42	0.62	B4	−0.42	0.36	C4	0.67	−0.29
A5	−0.81	0.56	B5	0.97	−0.42	C5	−0.65	1.19
A6	−0.56	0.00	B6	1.09	−1.16	C6	−1.32	0.05
A7	0.37	−0.04	B7	0.44	−0.49	C7	−0.55	0.70
A8	0.06	−1.29	B8	0.14	−0.38	C8	0.55	0.66
A9	0.95	−2.15	B9	1.33	1.24	C9	1.31	1.12
A10	1.90	−1.07	B10	0.80	0.85	C10	1.46	2.58
A11	1.32	0.61	B11	1.60	0.31			
A12	1.38	0.79	B12	1.42	0.99			
A13	1.63	1.03						
A14	0.50	1.24						
A15	0.90	1.19						

Table 2

Categorial distributions for files A, B, C under the given 2 × 2 × 2 partition

(a)

File A			File B			File C		
	$Y < 0$	$Y \geq 0$		$Z < 0$	$Z \geq 0$		$Z < 0$	$Z \geq 0$
$X < 0$	3	3	$X < 0$	3	1	$Y < 0$	3	3
$X \geq 0$	4	5	$X \geq 0$	4	4	$Y \geq 0$	1	3

(b)

Unadjusted File A Table			Adjusted File B Table			Adjusted File C Table		
	$Y < 0$	$Y \geq 0$		$Z < 0$	$Z \geq 0$		$Z < 0$	$Z \geq 0$
$X < 0$	3	3	$X < 0$	4.5	1.5	$Y < 0$	5.15	1.85
$X \geq 0$	4	5	$X \geq 0$	4.5	4.5	$Y \geq 0$	3.85	4.15

(c) Raked 2 × 2 × 2 table of ones to match the marginals in Table 2(b)

	$Z < 0$		$Z \geq 0$	
	$Y < 0$	$Y \geq 0$	$Y < 0$	$Y \geq 0$
$X < 0$	2.55	1.95	0.45	1.05
$X \geq 0$	2.60	1.90	1.40	3.10

(d) Categorical constraints by randomly rounding entries of Table 2(c)

	$Z < 0$		$Z \geq 0$	
	$Y < 0$	$Y \geq 0$	$Y < 0$	$Y \geq 0$
$X < 0$	2	2	1	1
$X \geq 0$	2	2	2	3

Table 3

Comparison of Eight Matching Methods for Completing File A

File A			Matched Z-Values							
			Versions of REG Method				Versions of HOD method			
X	Y	Z	M1	M2	M3	M4	M5	M6	M7	M8
−0.86	−0.32	−0.97	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69
−0.77	−0.33	0.16	−0.69	−0.69	−0.69	−0.69	−0.83	−0.69	−0.83	−0.69
−0.09	−0.26	0.19	−0.38	−0.38	0.36	0.36	0.36	−0.38	0.36	0.36
−0.42	0.62	−0.44	−0.38	0.36	0.36	−0.38	0.36	−0.38	0.36	−0.38
−0.81	0.56	−0.76	−0.69	0.36	−0.69	0.36	−0.69	0.36	−0.69	0.36
−0.56	0.00	1.06	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83
0.37	−0.04	−1.18	−0.38	−0.38	−0.38	0.36	−0.49	−0.49	−0.49	−0.49
0.06	−1.29	0.33	−0.38	−0.38	−0.36	−0.38	−0.38	−0.38	0.85	0.36
0.95	−2.15	−1.26	−0.42	−1.16	−0.42	−1.16	−0.42	−1.16	−0.42	−1.16
1.90	−1.07	0.01	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31
1.32	0.61	2.08	0.31	0.99	0.31	0.99	1.24	−0.42	1.24	−0.42
1.38	0.79	0.32	0.31	0.99	0.31	0.99	0.99	−0.42	0.99	−0.42
1.63	1.03	1.53	0.31	0.99	0.31	0.99	0.31	0.99	0.31	0.99
0.50	1.24	1.34	−0.49	0.85	−0.49	−0.38	−0.49	0.85	−0.49	0.85
0.90	1.19	−1.01	−0.42	0.85	−0.42	−0.42	−0.42	0.85	−0.42	0.85
Evaluation Measures		MAD-Z	0.79	0.81	0.76	0.78	0.79	0.85	0.78	0.78
		x^2	13.07	13.34	1.75	1.75	2.70	10.78	1.75	1.75

Note: M1: REG M2: REG* M3: REG.LOGLIN M4: REG.LOGLIN* M5: HOD M6: HOD* M7: HOD.LOGLIN M8: HOD.LOGLIN*.

Table 4

Computational Steps Required for M8 (HOD.LOGLIN*)

(X,Y) cell	X	Y		Match Order	Y-dist	Z_{int}		(X,Z)-dist	Z_{match}	
$X < 0$	−0.86	−0.32	(A1)	2	.08	−0.60	(C1)	.12	−0.69	(B1)
$Y < 0$	−0.77	−0.33	(A2)	1	.07	−0.60	(C1)	.19	−0.69	(B1)
	−0.09	−0.26	(A3)	3	.29	0.70	(C7)	.47	0.36	(B4)
$X < 0$	−0.42	0.62	(A4)	2	.05	−0.29	(C4)	.56	−0.38	(B8)
$Y \geq 0$	−0.81	0.56	(A5)	1	.01	0.66	(C8)	.50	0.36	(B4)
	−0.56	0.00	(A6)	3	.40	−0.60	(C1)	.25	−0.83	(B2)
$X \geq 0$	0.37	−0.04	(A7)	4	.36	−0.60	(C1)	.14	−0.49	(B7)
$Y < 0$	0.06	−1.29	(A8)	1	.03	0.05	(C6)	.57	0.36	(B4)
	0.95	−2.15	(A9)	2	.18	−2.81	(C2)	1.66	−1.16	(B6)
	1.90	−1.07	(A10)	3	.25	0.05	(C6)	.40	0.31	(B11)
$X \geq 0$	1.32	0.61	(A11)	1	.06	−0.29	(C4)	.37	−0.42	(B5)
$Y \geq 0$	1.38	0.79	(A12)	3	.12	−0.29	(C4)	.43	−0.42	(B5)
	1.63	1.03	(A13)	5	.28	1.12	(C9)	.25	0.99	(B12)
	0.50	1.24	(A14)	2	.07	1.12	(C9)	.41	0.85	(B10)
	0.90	1.19	(A15)	4	.12	1.12	(C9)	.29	0.85	(B10)

Note: Match order between records in files A and C is within (X, Y) cell under the categorical constraints given by Table 2(d).

6. EMPIRICAL INVESTIGATION OF PROPOSED MATCHING METHODS

This section presents the details of an empirical evaluation through an extensive simulation study with synthetic data generated from symmetric as well as skewed multivariate distributions. Symmetry was introduced via normal distributions, while skewness was introduced via contaminations by multivariate log normal distributions. The reason for using synthetic data is to have control over all of the relevant parameters, including those specifying the joint relationships of the different variables. This permits evaluation of the various approaches to matching as the joint relationships are allowed to depart in a systematic manner from conditional independence. It also permits comparisons of the methods as the underlying distribution generating the data moves away from symmetry. Proxy auxiliary information was generated by changing parameters of the normal distribution generating file C or by inducing log normal contaminations. We thus have four types of matching problems; the two corresponding to symmetric and skewed distributions with nonproxy data for C and the two corresponding to symmetric distributions with two types of proxy data for C. Programming was done on micro-computers using the software GAUSS.

6.1 Design of the Monte Carlo Study

In order to simulate statistical matching three datafiles are needed: a host file A, a donor file B, and an auxiliary file C. These are generated synthetically from specified distributions, with each file containing the three variables X, Y and Z. In file A the variable Z is suppressed and in file B the variable Y is dropped. The suppressed Z-values in file A are used to evaluate the performance of the various methods of statistical matching. File C could have only (Y,Z) information (by suppressing X) or the full (X,Y,Z) information. The empirical results presented in this paper correspond to file C with only (Y,Z) variables although file C with (X,Y,Z) variables was also included in the study (see Singh *et al.* 1990).

Runs of 100 simulations apiece were performed for each combination of design parameters considered. Four evaluation measures were calculated for each simulation and then were combined over all 100 simulations.

Files A and B were always generated from the same underlying distribution, with each containing 500 independent and identically distributed observations. File C contained 250 observations, not necessarily from the same distribution as that for files A and B; that is, file C could contain either proxy or nonproxy auxiliary information.

The distribution of observations (X,Y,Z) was multivariate normal with some log normal contamination introduced by taking the exponentials of X, Y and Z for some of the observations. Individual observations were contaminations or not according to a Bernoulli process with probability fixed for any particular run of 100 simulations. Prior to contamination X, Y and Z were standard normal. The covariances of (X,Y) and (X,Z) prior to contamination were always .5, with the covariance of (Y,Z) varying from run to run. Consequently, the conditional correlation of Y and Z given X, $\rho_{Y,Z|X}$, was also varied from run to run.

For most runs the distribution of observations in the auxiliary file C was the same as that in files A and B. However, if in an application the source of auxiliary information is historical or via proxy variables this assumption may be unreasonable. Two series of runs were carried out with proxy auxiliary information. In the first series the auxiliary data had a different $\rho_{Y,Z|X}$. In the second series the auxiliary data had some log normal contamination.

For the proposed methods which use categorical constraints and for defining matching categories for the HOD method, it was necessary to choose a categorical partition. Two partitions were used. The first, called standard interval, divided the ranges of the X, Y and Z variables into the categories < -1, $[-1,0)$, $[0,1)$, ≥ 1; that is, the partition was centred on the mean of the marginal distribution before contamination, with break points at the centre and at plus or minus one standard deviation. The second partition, called equal probability, was similar but had break points at the quartiles of the pre-contamination marginal distributions; that is, the partition had the categories $< -.6745$, $[-.6745,0)$, $[0,.6745)$, $\geq .6745$. The partitions were defined in terms of the pre-contamination distributions; for simplicity the same partitions were used when there were log normal contaminations. It would, however, have been more realistic to let the partitions be data dependent.

6.2 The Matching Methods

The eight methods as defined earlier were considered. Except for REG and HOD, all others use auxiliary information. Thus, we have two variants for each depending on whether (Y,Z) or (X,Y,Z) information is available in file C. For the methods HOD and HOD.LOGLIN, three versions of hot deck (namely, rank, random, and X-distance) were considered for finding live Z-values from B although only results based on X-distance are reported here. For the other six methods, although we considered three types of hot deck (namely, Z-distance, (X,Z)-distance, and (X,Y,Z)-distance) for finding live Z-values from B, we show only results for (X,Z) distance here for simplicity. Section 7.3 does contain a brief description of results obtained with different distance measures. The report by Singh *et al.* (1990) contains other details not included here. It may be noted that for using hot deck with (X,Y,Z)-distance to get a live Z-value from B, intermediate Y-values

would have to be first obtained for B from file C, analogous to Z_{int} for file A. Note also that the Euclidean distance was always employed whenever hot deck with distance metric was used. However, variables were not preadjusted by their standard deviations for convenience and because all the variables in the synthetic population had common variances.

6.3 The Evaluation Measures

Four evaluation measures were used to measure how well the different matching methods performed. All of the evaluations are based on comparisons of the matched file to the file with the suppressed true Z-values. Two of the measures are based on categorical comparisons, but the categories used for evaluations need not be the same as those used for categorical constraints by the LOGLIN procedures. The results reported here correspond to using the equal probability partition (see Section 6.1) for matching and the standard interval partition for evaluations. The first of the four evaluation measures is based on unit by unit comparison of the matched and suppressed Z-values. However, the objective of a statistical matching procedures cannot be to reproduce the suppressed Z-values exactly, but to produce Z-values that come from the same distribution given what is known, in this case given X and Y. The last three evaluation measures are based more on comparisons of the conditional distributional properties of Z.

(i) Average of Mean Absolute Differences of Z ($\overline{\text{MAD-}Z}$)

The simplest measure of performance is the mean absolute difference between the matched and suppressed Z-values for records in file A. Monte Carlo averages of these means as well as standard errors were obtained.

The formula for the MAD-Z statistic for any given simulation, is

$$\text{MAD-}Z = \sum_i | Z_{s,i} - Z_{m,i} | /500, \qquad (6.1)$$

where $Z_{s,i}$ is the suppressed Z-value for the ith record in file A, $Z_{m,i}$ is the matched Z-value, and the sum is over all 500 records of file A. $\overline{\text{MAD-}Z}$ denotes the average of the MAD-Z statistics over simulations.

(ii) Average of Absolute Difference of Covariances ($\overline{\text{AD-Cov}}$)

The second measure of performance is the absolute difference of the conditional covariances of Y and Z given X in the matched and suppressed files. Monte Carlo averages of these absolute differences as well as standard errors were obtained.

For a file with variables X, Y and Z we may define

$$\text{Cov}(Y,Z \mid X) = \text{Cov}(Y,Z) - \text{Cov}(X,Y)\text{Cov}(X,Z)/\text{Var}(X), \qquad (6.2)$$

where Cov and Var are the sample covariance and variance operators respectively. In the multivariate normal case this corresponds to the covariance of Y and Z given X. Otherwise it may be interpreted as the covariance of the residuals of a linear regression of Y on X with the residuals of a linear regression of Z on X. The AD-Cov statistic for any given simulation, would be the absolute difference between these quantities for the matched and suppressed files. $\overline{\text{AD-Cov}}$ denotes as usual the average over simulations.

(iii) Average of Chi-square Statistics ($\overline{\chi^2}$)

The third measure of performance, based on categorical comparisons, is a distance measure based on the Pearson chi-square statistic. What is reported is the average chi-square statistic over the 100 simulations, transformed to lie in the interval (0,1).

The formula for the chi-square statistic, is

$$\chi^2 = \sum_{i,j,k} (m_{ijk} - n_{ijk})^2/(m_{ijk} + .5), \qquad (6.3)$$

where m_{ijk} is the number of records in X^* category i, Y^* category j, and Z^* category k in the matched file, n_{ijk} is the same for the suppressed file, and the sum is over all (X^*, Y^*, Z^*) categories. A constant .5 is added to all of the denominators in this sum to avoid the problem of zeros.

Once the mean of the chi-square statistics from 100 simulations, say $\overline{\chi^2}$, is obtained, it is transformed to lie in the interval (0,1) using the transformation (see Bishop, Fienberg and Holland 1975, p 383; here 500 is the size of file A)

$$\text{Transformed } \overline{\chi^2} = [\overline{\chi^2}/(\overline{\chi^2} + 500)]^{1/2}. \qquad (6.4)$$

(iv) Likelihood Ratio Test (LRT)

The final measure of performance is also based on categorical comparisons. Within each (X^*, Y^*) category that has a minimum number of observations (in the present study, we set it at 20) a likelihood ratio test that the categorical Z-values from the matched and suppressed files come from the same multinominal distribution is performed. The tests for different (X^*, Y^*) categories are then combined to obtain an overall P-value. What is reported is the proportion of times, out of 100 simulations, that the overall P-value was less than .05. The larger this proportion, the greater the difference between the true and matched categorical distributions of Z^* given the (X^*, Y^*) categories.

The minimum sample size of 20 for (X^*, Y^*) categories in file A was required so that the chi-square approximation to the distribution of the test statistic might be reasonable. If the number of Z^* categories was increased, this minimum sample size might also need to be increased.

Using the same notation as in the previous measure, the formula for the likelihood ratio test statistic from the (i,j) (X^*, Y^*) category is

$$\begin{aligned} \text{LRT} = 2 \sum_k \{ & (n_{ijk} + .5) ln((n_{ijk} + .5)/ \\ & (n_{ijk} + m_{ijk} + 1)) + (m_{ijk} + .5) \\ & ln((m_{ijk} + .5)/(n_{ijk} + m_{ijk} + 1))\} \\ & + (4n_{ij} + 2K) ln2, \end{aligned} \tag{6.5}$$

where

$$n_{ij} = \sum_k n_{ijk} = \sum_k m_{ijk}, \quad i = 1, \ldots, I,$$

$$j = 1, \ldots, J, \quad k = 1, \ldots, K. \tag{6.6}$$

The asymptotic distribution of this statistic, when the m_{ijk}'s and n_{ijk}'s come from the same multinominal distribution, is chi-square with $(K - 1)$ degrees of freedom. An overall P-value is obtained by adding these statistics and their degrees of freedom for each (X^*, Y^*) category meeting the minimum sample size criterion, and finding the probability of a chi-square variable with the appropriate degrees of freedom being larger than the observed value.

7. RESULTS OF THE MONTE CARLO STUDY

In this section we describe the results of the simulation study. A more complete description is given in Singh *et al.* (1990). Tables of actual numbers underlying Figures 1 through 5 are available upon request.

We have not paid much attention to Monte Carlo standard errors of the evaluation measures in the presentation. This is because they were generally quite small, for example, coefficients of variation were generally less than two percent for the $\overline{\text{AD-Cov}}$ evaluation measure. Furthermore, the evaluations of different methods would be expected to be positively correlated so that the relative differences between matching methods would be even more precisely estimated than suggested by the standard errors. A further indication of the quality of the Monte Carlo evaluations of the various methods is the general smoothness of observed trends, for example, see Figures 2 to 5. In short, any discernible difference in the figures is likely to indicate a real difference.

7.1 Methods with no Auxiliary Information (REG and HOD)

Figures 2 through 5 show how departures from conditional independence affect performance of matching methods which use CIA. Apparently the use of such methods may result in serious bias in the joint relationship of (X,Y,Z) in the matched file. For example, Figure 2 shows a progressive deterioration as the true conditional correlation, $\rho_{Y,Z|X}$, moves away from zero with respect to all measures except $\overline{\text{MAD-Z}}$ which actually shows no deterioration at all. It may be due to the fact that $\overline{\text{MAD-Z}}$ is an unconditional measure which is based on unit by unit comparison of the matched and suppressed Z-values, while the other measures are based on comparisons of the conditional distributions of Z. It is interesting to note from Figure 2 that when the true value of $\rho_{Y,Z|X}$ is small, the performance of the HOD* method, which uses auxiliary information, can be worse with respect to the categorical or aggregate level evaluation measures than the performance of the HOD method which does not make use of auxiliary information. The point at which the use of auxiliary information would become advantageous would depend on the precision of the auxiliary information.

7.2 Methods with Auxiliary Information

Our empirical results do confirm, as expected, that the use of auxiliary information does protect against the failure of the CIA. The degree of protection would depend on the method and the type of auxiliary information used. A brief summary of performances of various methods was presented earlier in the introduction. Here, we will provide some details based on Figures 2 to 5.

In the regression family, the methods using auxiliary information on conditional correlations, namely REG* and REG.LOGLIN*, show very favourable performance with respect to the unit level measures (*i.e.* $\overline{\text{MAD-Z}}$ and $\overline{\text{AD-Cov}}$) for symmetric populations (see Figure 2). They continue to outperform hot deck methods for skewed populations (Figure 3) although the bias tends to increase as the degree of skewness grows. However, for proxy auxiliary information having different conditional correlation (Figure 4), the regression methods perform in a mixed fashion, *i.e.* they could be better or worse than hot deck methods at the unit level. In fact, for the second type of proxy auxiliary information (namely, with log normal contamination; see Figure 5), they tend to be slightly inferior to the HOD.LOGLIN method with respect to the $\overline{\text{AD-Cov}}$ measure. If we restrict ourselves to the regression family, then the REG* method can be recommended with regard to the unit level evaluation measures. However, with respect to the aggregate level, all regression methods show very unfavourable performance. This can probably be explained by the shrinkage to the mean effect as discussed in subsection 7.3.

Matching Methods for Figures 1 to 5

REG	Z_{int} obtained from regression of Z on X, Z_{match} based on (X,Z) distance
REG*	Z_{int} obtained from regression of Z on X and Y, Z_{match} based on (X,Z) distance
REG.LOGLIN	Z_{int} obtained from regression of Z on X, Z_{match} based on (X,Z) distance using categorical constraints
REG.LOGLIN*	Z_{int} obtained from regression of Z on X and Y, Z_{match} based on (X,Z) distance using categorical constraints
HOD	Hot deck using X distance within X categories
HOD*	Z_{int} obtained from file C using hot deck with Y distance, Z_{match} obtained from file B using hot deck with (X,Z) distance
HOD.LOGLIN	Hot deck using X distance within X categories and using categorical constraints
HOD.LOGLIN*	Z_{int} obtained using hot deck with Y distance and using categorical constraints, Z_{match} obtained using hot deck with (X,Z) distance within (X,Y,Z) categories

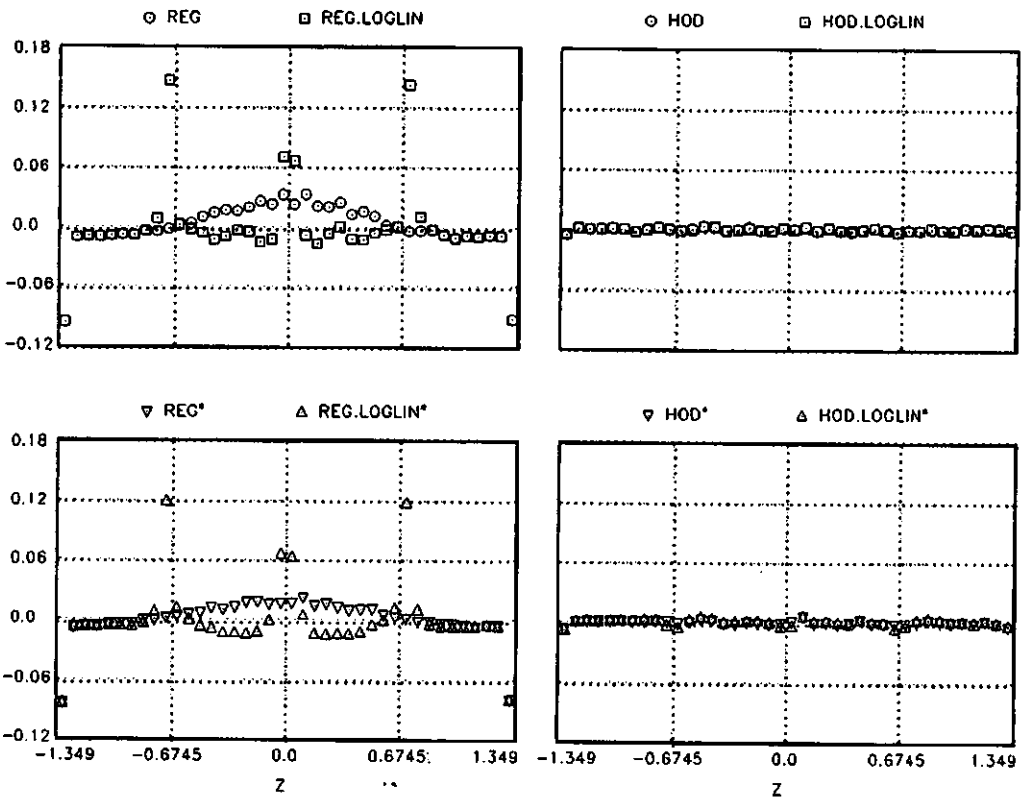


Figure 1. Difference of matched and suppressed marginal Z-histograms (symmetric data, $\rho_{Y,Z|X} = .4$)

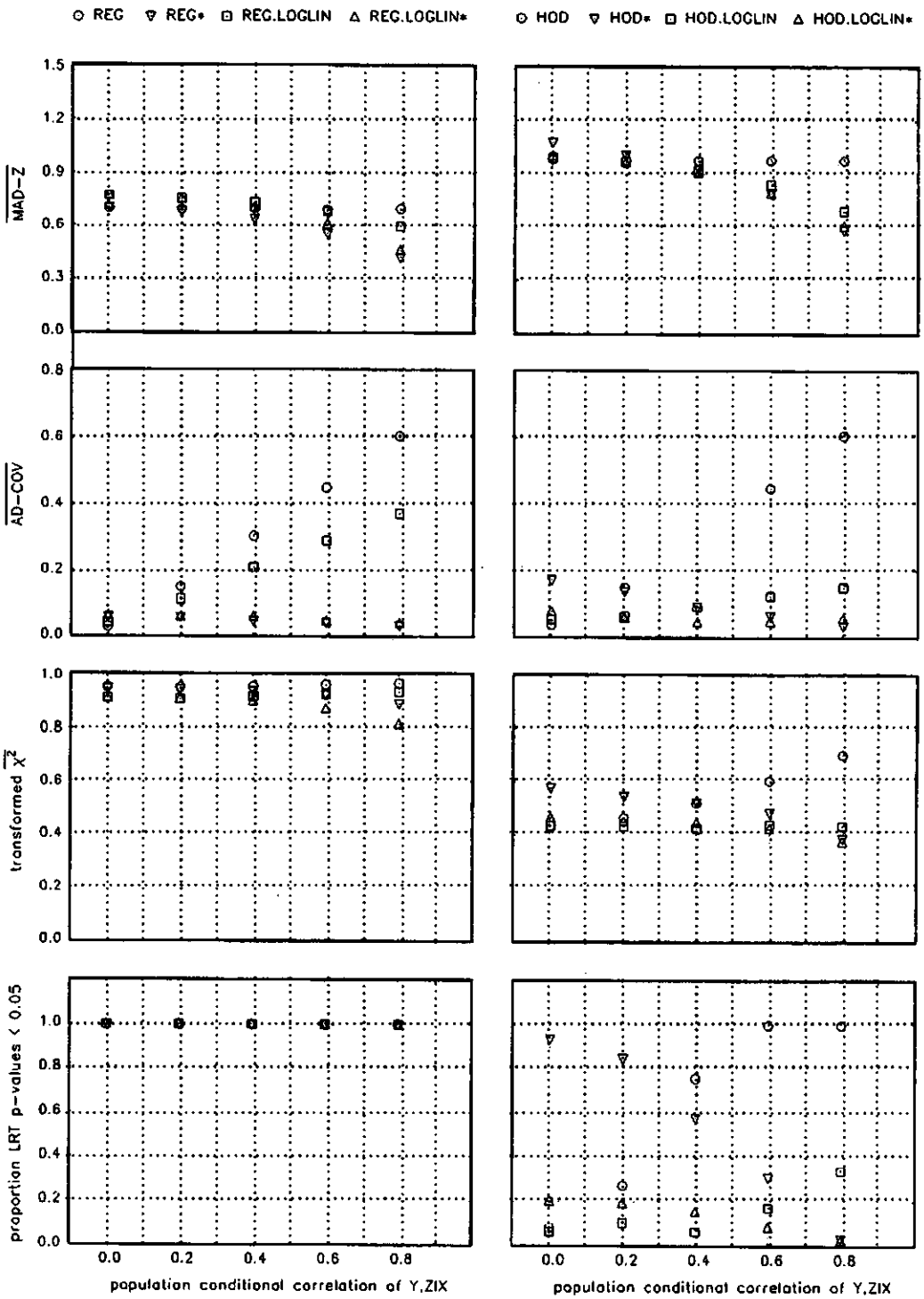


Figure 2. Comparison of statistical matching methods as $\rho_{Y,Z|X}$ varies for the symmetric population, non-proxy auxiliary information

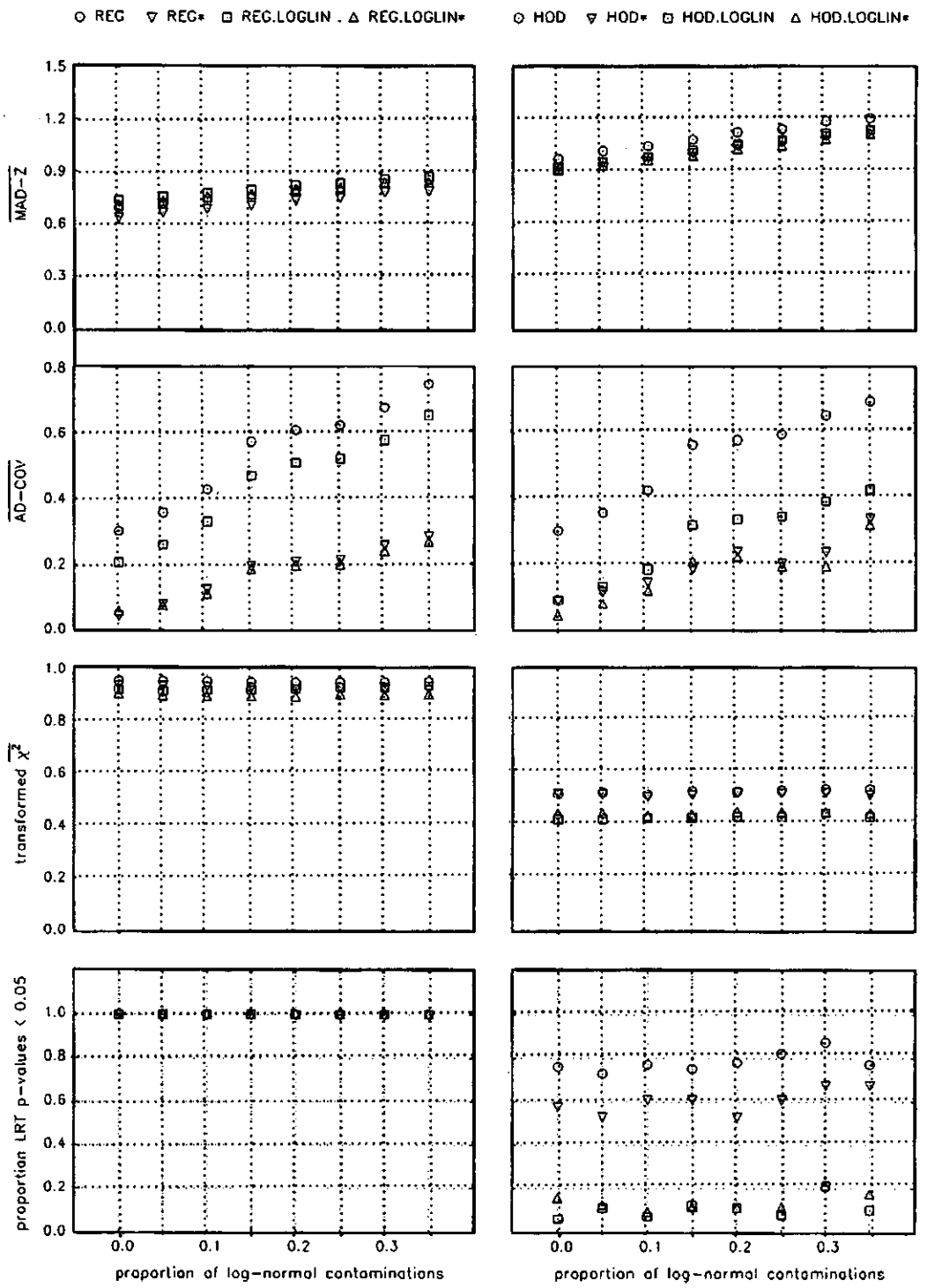


Figure 3. Comparison of statistical matching methods as the proportion of log-normal contamination varies ($\rho_{Y,Z|X}$ before contamination), non-proxy auxiliary information

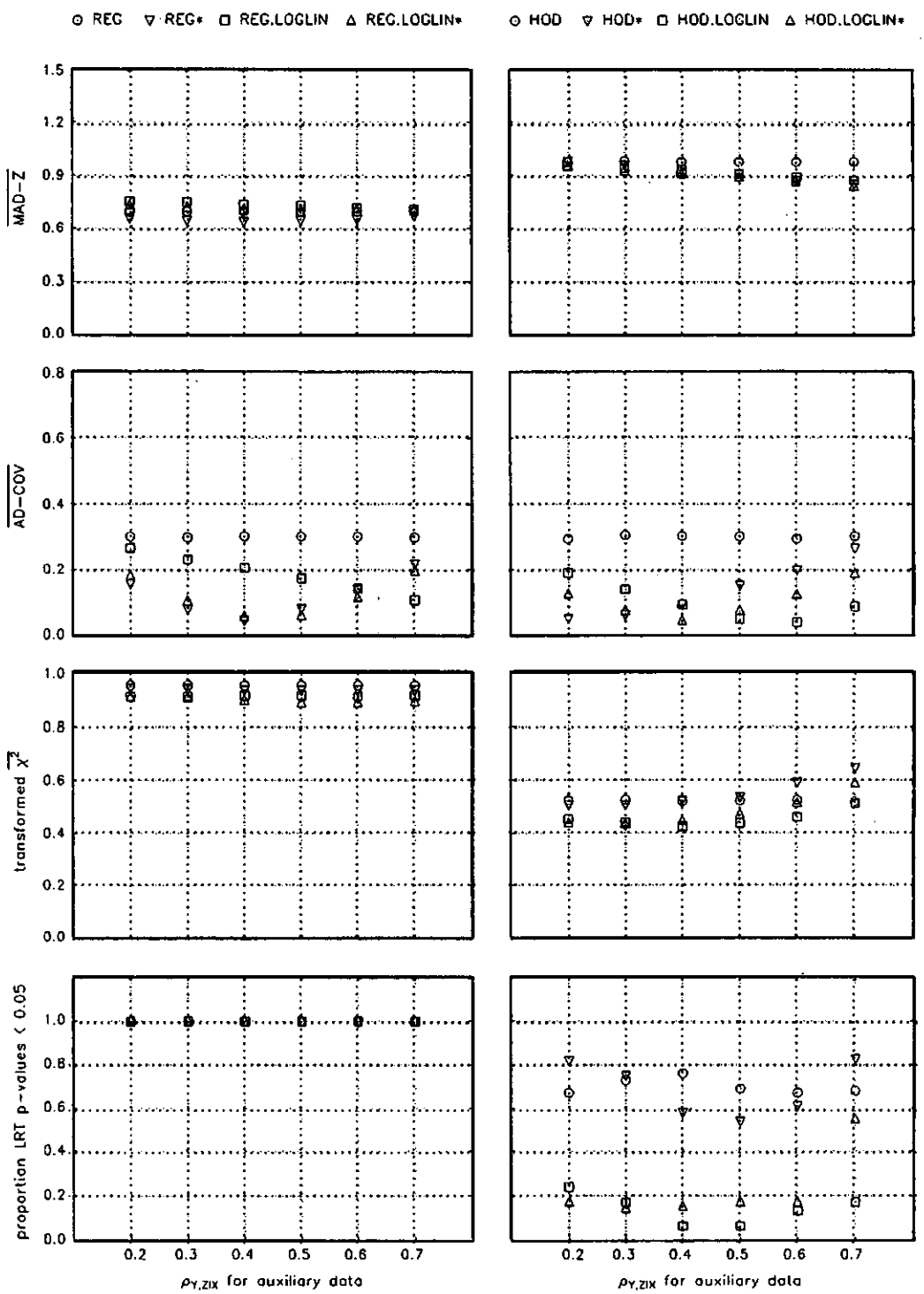


Figure 4. Comparison of statistical matching methods as $\rho_{Y,Z|X}$ varies for the auxiliary data file C ($\rho_{Y,Z|X} = .4$ for files A and B)

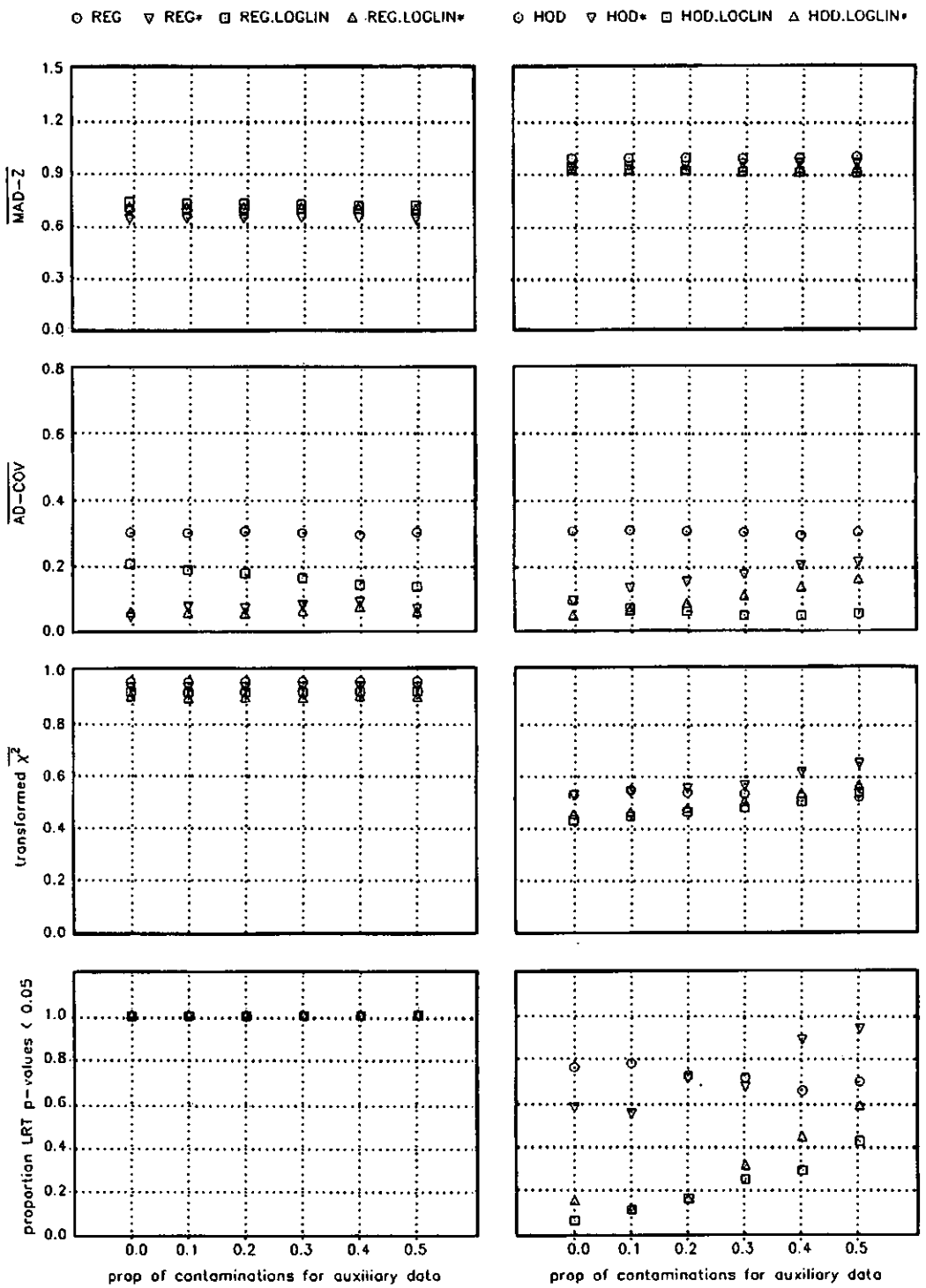


Figure 5. Comparison of statistical matching methods as the proportion of log normal contaminations varies for the auxiliary data file C ($\rho_{Y,Z|X} = .4$ before contamination and files A and B have no log-normal contamination)

In the hot deck family of methods with auxiliary information, the two methods with categorical constraints, namely, HOD.LOGLIN and HOD.LOGLIN* show vary favourable performance at the aggregate level (*i.e.* with respect to transformed χ^2 and LRT) for all types of underlying populations; see Figures 2 to 5. In this case, the method HOD.LOGLIN generally outperforms HOD.LOGLIN*. Next we consider unit level measures. For symmetric and skewed populations (Figures 2 and 3), generally speaking, HOD* and HOD.LOGLIN* perform very similarly to each other and somewhat better than HOD.LOGLIN but slightly worse than REG*. However, for proxy auxiliary data (Figures 4 and 5) with respect to the $\overline{\text{AD-Cov}}$ measure, the method HOD.LOGLIN could be better or worse than the HOD.LOGLIN* and the REG* methods, and is more often better in the case of proxy data with log normal contaminations. Also HOD.LOGLIN in the case of proxy data tends to have fairly robust behaviour with respect to all four evaluation measures. Thus, in the hot deck family, based on overall performance, HOD.LOGLIN* can be recommended. However, in practice, HOD.LOGLIN may be preferable as a compromise because it performs moderately well at the unit level, extremely well at the aggregate level, is computationally much less demanding and shows robustness with respect to the proxy auxiliary data. Furthermore, HOD.LOGLIN does not require micro-level auxiliary information.

7.3 Miscellaneous Observations

In this subsection we describe separately some interesting findings, the corresponding empirical results for some of which are not included here, but are presented in Singh *et al.* (1990).

(i) Shrinkage to the Mean

An important and consistent finding was that matching methods in the regression family do not perform well with respect to the categorical measures. This can be explained by shrinkage towards the mean; that is, the matched Z-values are more tightly distributed about their mean than are the suppressed true Z-values. This is displayed in Figure 1 which shows the difference between the marginal histograms of matched and suppressed Z-values for various matching methods.

The positive differences for REG and REG* near the centre indicate that there are more Z-values in that region in the matched file than in the suppressed file. The very large negative observations at the extreme points of this plot are associated with open ended intervals, and it seems quite likely that had these intervals been broken down into several smaller intervals the plot would have shown several smaller negative numbers in the extreme tails, so that the interpretation of the plot should be that these methods are putting too many Z-values at the centre of the distribution at the expense of the extreme tails.

Figure 1 also shows shrinkage towards the mean for the REG.LOGLIN and REG.LOGLIN* methods. However, in this case the shrinkage is limited by the categorical constraints so that, while we still see that the tails of the Z-distribution of the matched file are too short, the displaced values are now not going to the centre of the distribution, but only to the partition boundary points which act like walls. The large positive values to either side of the central boundary point can be explained similarly if one bears in mind that what this plot is showing is actually an average of differences of histograms over 100 independent simulations. It seems reasonable that if we were to examine each of the 100 differences of histograms individually we would sometimes see a large positive value just to the left of the central boundary point, and sometimes just to the right, but never both at the same time.

Figure 1 also shows that shrinkage to the mean and boundary effects are not serious for methods in the hot deck family.

(ii) *(Y,Z) vs (X,Y,Z)* Auxiliary Information

Although only results based on (Y,Z) auxiliary information were presented in this paper, (X,Y,Z) auxiliary information was also considered as part of the simulation study as mentioned in Section 6. An interesting finding was that for the HOD.LOGLIN and HOD.LOGLIN* methods, the use of (Y,Z) auxiliary information leads, in general, to somewhat better performance at the aggregate level than the use of (X,Y,Z) information. This does not seem to be the case with the HOD* method. This phenomenon is probably due to instability in the estimation of (X^*,Y^*,Z^*) factor effects used in the categorical constraints on account of insufficient size of auxiliary data. An implication is that the true values were probably close to zero and so taking them as zero leads to better results. This suggests that the impact of different sample sizes on performance of matching methods should be considered, if possible, in future investigations. The above consideration also suggests an interesting new class of methods which would combine (X,Y,Z) micro-level auxiliary information for finding Z_{int} values along with the derived (Y^*,Z^*) categorical distribution only from file C for imposing constraints. These methods were, however, not included in the present study.

(iii) Comparison of Different Versions of Hot Deck Methods

In all the matching methods considered, except HOD and HOD.LOGLIN, the second step for finding Z_{match} consists of using hot deck imputation in which (X,Z)-distance is employed. For the remaining two, X-distance was considered. Some other options (for methods other than

HOD and HOD.LOGLIN), consist of using Z-distance or (X,Y,Z)-distance. For the latter, Y_{int} would have to be added first to file B. This was included in the original simulation study, although empirical results are not reported here. It was found that there is generally no difference though, for REG and REG* methods, (X,Z)-distance sometimes showed superior performance with respect to the $\overline{\text{AD-Cov}}$ measure. This is the reason for our choice of (X,Z)-distance in the methods considered here. However, in practice, it may be preferable to use Z-distance with hot deck matching methods because of computational convenience.

Further, it should be noted that for HOD and HOD.LOGLIN methods, there is the option of using random or rank in Step II instead of X-distance. In hot deck rank, records from files A and B are ranked separately according to the value of X, and then are matched based on ranks. This was proposed by G. Rowe for the SPSD application mentioned in the introduction. Clearly, this method is suitable for univariate X only. An advantage of ranking is that there will not be one record from file B acting as donor for many records from file A. The above three versions of hot deck were included in the Monte Carlo study although results for X-distance only are reported here. It was found that it generally does not make much difference which version is used. The choice of X-distance was made for HOD and HOD.LOGLIN because it was consistent with the hot deck distance version used for other methods. In practice the hot deck random version would be least demanding computationally; however, in a real application we would not know how much might be lost by using random matching instead of ranking or distance, and we would probably want to use as much information as would be feasible.

8. CONCLUDING REMARKS

In this paper, the problem of using auxiliary information in statistical matching was considered. The two main methods previously proposed are due to Rubin (1986) and Paass (1986), versions of which were denoted by REG* and HOD*. Some modifications of these methods, denoted by REG.LOGLIN* and HOD.LOGLIN*, were proposed by imposing categorical constraints derived from auxiliary information. These would reduce to REG.LOGLIN and HOD.LOGLIN if only categorical auxiliary information is available or useable. In the absence of auxiliary information, the usual methods of imputation, REG and HOD would be used. An empirical study was conducted to evaluate performance of the above eight methods with respect to four evaluation measures (two at the unit level, and two at the aggregate level). It was found that for the case of no auxiliary information, the HOD method is preferable. The case of auxiliary information is, however, more complex. If only unit level evaluation measures are deemed important, then the REG* method is recommended. If aggregate level measures are also considered important then if there is nonproxy auxiliary data HOD.LOGLIN* is recommended. As an alternative, a good compromise would be HOD.LOGLIN if computational burden is an important consideration or if proxy auxiliary data is believed to be present. If unit level measures are less important or are not of interest (this may often be the case because the matched data would generally be presented in tabular forms in practice), then HOD.LOGLIN would be recommended. With both HOD and HOD.LOGLIN methods, the similar performances of distance, random and rank versions might suggest the use of random versions in practice in view of its computational simplicity.

It may be remarked that we did not consider the fully iterative version of Paass's method. It would be interesting to find out in future investigations how this might perform. Another point that requires investigation is the implementation of categorical constraints with many variables. The application of the raking algorithm may be computationally prohibitive. In this connection, the results of Paass (1989) are expected to be useful.

In the present study we did not, due to limitations of computing, systematically vary the accuracy of the auxiliary data source; that is, we did not vary the size of the file C. We also did not vary the size of the files A or B. An interesting question that might have been addressed is how the performance of various methods might be affected by the size of these files.

Finally, it should be pointed out that although the results of this study are based on synthetic data (which was necessary to produce various scenarios mimicking real data), it is believed that the results would be relevant for real applications. Clearly, it would be interesting and useful to carry out a simulation study with real data to check whether the findings continue to hold and to see what sorts of substantive impact the biases in the joint distribution of the matched file have. A related question is how to account for such biases in inferences based on the matched file; that is, how to produce measures of uncertainty for parameter estimates from the matched file that reflect not only the variability within the matched file, but also the uncertainty inherent in the matching procedure itself. Although we are unable to answer this question, it is clear that matching procedures using auxiliary information would enhance the overall utility of the matched file. These and some other related questions will be investigated in the future.

ACKNOWLEDGEMENT

The authors would like to thank J. Armstrong, G. Gray, G. Hole, D. Royce and M. Wolfson for helpful comments. The first author's research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University.

REFERENCES

ARMSTRONG, J. (1989). An evaluation of statistical matching methods. Methodology Branch Working Paper, BSMD, 90-003E. Statistics Canada.

BARR, R.S., and TURNER, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Surveys. Technical report, U.S. Department of the Treasury, Office of Tax Analysis.

BARR, R.S., and TURNER, J.S. (1990). Quality issues and evidence in statistical file merging. In *Data Quality Control: Theory and Pragmatics* (Eds. G.E. Liepins and V.R.R. Uppuluri). New York: Marcel Dekker, 245-313.

BARR, R.S., STEWART, W.H., and TURNER, J.S. (1981). An empirical evaluation of statistical matching methodologies. Technical report, Edwin L. Cox School of Business, Southern Methodist University, Dallas, Texas.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass: MIT Press.

BUDD, E.C. (1971). The creation of a microdata file for estimating the size distribution of income. *The Review of Income and Wealth*, 17, 317-333.

BUDD, E.C., and RADNER, D.B. (1969). The OBE size distribution series: methods and tentative results for 1964. *American Economic Review, Papers and Proceedings*, LIX, 435-449.

COHEN, M.L. (1991). Statistical matching and microsimulation models. In *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, Volume II, Technical Papers, (Eds. C.F. Citro and E.A. Hanushek). Washington, D.C.: National Academy Press, 62-85.

FELLEGI, I.P. (1977). Discussion paper. *Proceedings of the Section on Social Statistics, American Statistical Association*, 762-764.

FORD, B.L. (1983). An overview of hot-deck procedures. In *Incomplete Data in Sample Surveys*, (Vol. 2), (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 185-207.

KADANE, J.B. (1978). Some statistical problems in merging data files. In *1978 Compendium of Tax Research*. Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 159-171.

KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

OKNER, B.A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement*, 1, 325-342.

PAASS, G. (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. In *Microanalytic Simulation Models to Support Social and Financial Policy* (Eds. G.H. Orcutt, J. Merz and H. Quinke). Amsterdam: Elsevier Science.

PAASS, G. (1989). Stochastic generation of a synthetic sample from marginal information. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 431-445.

PAASS, G., and WAUSCHKUHN, U. (1980). Experimentelle erprobung und vergleichende Bewertung statistischer Matchverfahren. Internal report, IPES.80.201, St. Augustin, *Gesellschaft für Mathematik und Datenverarbeitung.*

PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (or Domains). *International Statistical Review*, 48,3-18.

RODGERS, W.L. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2, 91-102.

RODGERS, W.L., and DeVOL, E. (1982). An evaluation of statistical matching. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 128-132.

RUBIN, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4, 87-94.

RUGGLES, N., RUGGLES, R., and WOLFF, E. (1977). Merging microdata: Rationale, practice and testing. *Annals of Economic and Social Measurement*, 6, 407-428.

SCHEUREN, F.J. (1989). Comment on Wolfson *et al.* (1989). *Survey of Current Business*, 69, 40-41.

SIMS, C.A. (1972). Comment on Okner (1972). *Annals of Economic and Social Measurement*, 1, 343-345.

SIMS, C.A. (1978). Comment on Kadane (1978). In *1978 Compendium of Tax Research*, Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 172-177.

SINGH, A.C. (1988). Log-linear imputation. Methodology Branch Working Paper, SSMD, 88-029E, Statistics Canada; also published in *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 118-132.

SINGH, A.C., ARMSTRONG, J.B., and LEMAÎTRE, G.E. (1988). Statistical matching using log linear imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 672-677.

SINGH, A.C., MANTEL, H., KINACK, M., and ROWE, G. (1990). On methods of statistical matching with and without auxiliary information: Some modifications and an empirical evaluation. Methodology Branch Working Paper, SSMD, 90-016E. Statistics Canada.

U.S. DEPARTMENT OF COMMERCE (1980). Report on exact and statistical matching techniques. Statistical Policy Working Paper 5, Washington, D.C.: Federal Committee on Statistical Methodology.

WOLFSON, M., GRIBBLE, S., BORDT, M., MURPHY, B., and ROWE, G. (1987). The social policy simulation database: an example of survey and administrative data integration. *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada, Ottawa, (Eds. J.W. Coombs and M.P. Singh), 201-229; another version published in *Survey of Current Business* (1989), 69, 36-40.

Survey Methodology, June 1993
Vol. 19, No. 1, pp. 81–94
Statistics Canada

A Framework for Measuring and Reducing Nonresponse in Surveys

MICHAEL A. HIDIROGLOU, J. DOUGLAS DREW, and GERALD B. GRAY[1]

ABSTRACT

The need for standards introduced for the gathering and reporting of information on nonresponse across surveys within a statistical agency is discussed. Standards being adopted at Statistics Canada are then described. Measures to reduce nonresponse undertaken at different stages in the design of surveys at Statistics Canada that have a bearing on nonresponse are described. These points are illustrated by examining nonresponse experiences for two major surveys at Statistics Canada.

KEY WORDS: Nonresponse rates; Incentives; Follow-ups; Data collection.

1. INTRODUCTION

National agencies such as Statistics Canada conduct a large number of different surveys every year. These vary in their subject matter, units of response, periodicity, sample design and collection methodologies. They also have varied experiences with respect to the nonresponse incurred. There is a need for agency-wide standards for the gathering and reporting of information on response and nonresponse. If they are sufficiently flexible to accommodate the requirements of the variety of surveys that are conducted, it is logical to have standard definitions. A distinction needs to be made though between standard definitions and standards of acceptable levels of different components of nonresponse to surveys. It is the former and not the latter that is under discussion.

There are major differences between surveys that result in different levels of nonresponse achieved; for example, longitudinal and cross-sectional surveys face somewhat different missing data problems. Standard definitions can provide a common lexicon that will help in isolating and understanding better the differences. A common lexicon helps in the ongoing analysis of trends in nonresponse. Information on survey response and nonresponse can serve multiple purposes, such as the potential for nonresponse biases, pointing to weak areas that need to be strengthened in future rounds of the survey. They provide measures of frame coverage, for developing methods to compensate for and to reduce nonresponse. They also give an important input to survey design, collection methodologies, evaluation of data quality and operations for different surveys.

Nonresponse rates can be defined differently, depending on whether they are used to diagnose sampling activities, data collection activities or to analyze published data. For example, in the case of sampling requirements, the unit for which nonresponse is measured ought to be the sampled unit. Correspondingly, for data collection activities, the unit of measure for computing nonresponse would be based on the unit of response. It should be noted that for business surveys there is often not a one-to-one correspondence between sampled units and units of response (*e.g.*, the sampled unit may be the head office and the unit of response is its branches). For published data, the measure of nonresponse could be weighted size measures or weighted key variables to estimate the contribution of nonrespondents to the key aggregates. In business surveys, such measures can be important because of the skewed populations where a few units contribute to a disproportionately large share of the estimate.

Breakdowns of the nonresponse rates should be available at pre-determined geographical levels, industrial and size levels and combinations of it. If possible, the reasons for nonresponse also should be available *e.g.*, unable to contact, refusal *etc.* These can be used to produce diagnostics to establish causes of nonresponse. If the data are collected by using interviewers located throughout nationwide regional offices, then nonresponse rates by interviewers within each regional office and nonresponse rates aggregated by regional office can be used as measures of operational performance. Questionnaire item nonresponse rates can be used to point to questions that need to be rethought in terms of wording or data availability.

This paper deals with total nonresponse, where nonresponse occurs at the level of the unit for which data are being collected. It does not deal with partial nonresponse, where the respondent provides usable information for some items but not for others. We start with a conceptual

[1] Michael A. Hidiroglou, Business Survey Methods Division; J. Douglas Drew, Household Surveys Division; Gerald B. Gray, Social Survey Methods Division, Statistics Canada.

framework for the definition of response and nonresponse that is suitable for both business and social surveys. The next section is devoted to general causes of nonresponse and to means for reducing nonresponse. Finally, we look at the experiences with nonresponse for two major surveys conducted at Statistics Canada.

2. DEFINITIONS OF NONRESPONSE RATES

Nonresponse rates and their complements, response rates, are defined as ratios of variables that represent a given category of response/nonresponse in some domain of interest. The important variable may be a simple count or it may be weighted by some factor. It may be the sample weights of the unit or the unit's expected contribution to the estimate of some major statistic of the survey. Figure 1 represents a conceptual framework developed by Drew and Gray (1991) for classifying sampled units in a survey into responding, nonresponding and out-of-scope units. The hierarchical representation is similar to one initially proposed by Platek and Gray (1986). The framework has been evaluated for several business and social surveys at Statistics Canada. The agency has adopted these standards for the gathering and reporting of information on nonresponse. Starting with the 1993 reference year, several major surveys will be required to report detailed nonresponse using the standard definitions. A data base of nonresponse rates will be maintained for use in agency-wide monitoring and analysis of trends in nonresponse.

We begin with the **Total number of Units** (weighted or unweighted). The total number of units consists of those that are thought to belong to the survey of interest before the survey process begins. The total (Box 1 in Figure 1) is broken down into two main categories: resolved (Box 2) and unresolved (Box 3) units. **Resolved Units** are those whose status as belonging or not belonging to the target universe is known by the cutoff date of the survey data collection. For some surveys all units can be resolved. For other surveys, it is either impossible or impractical to resolve all units. For example, in a telephone survey there are telephone numbers that ring but do not correspond to working numbers. Without checking the status of each so-called ring-no-answer case with the telephone company, there is no way to determine whether such a number represents a working number. Similarly for a survey with mail collection, without a follow-up of units not returning a questionnaire, it may not be known which units are

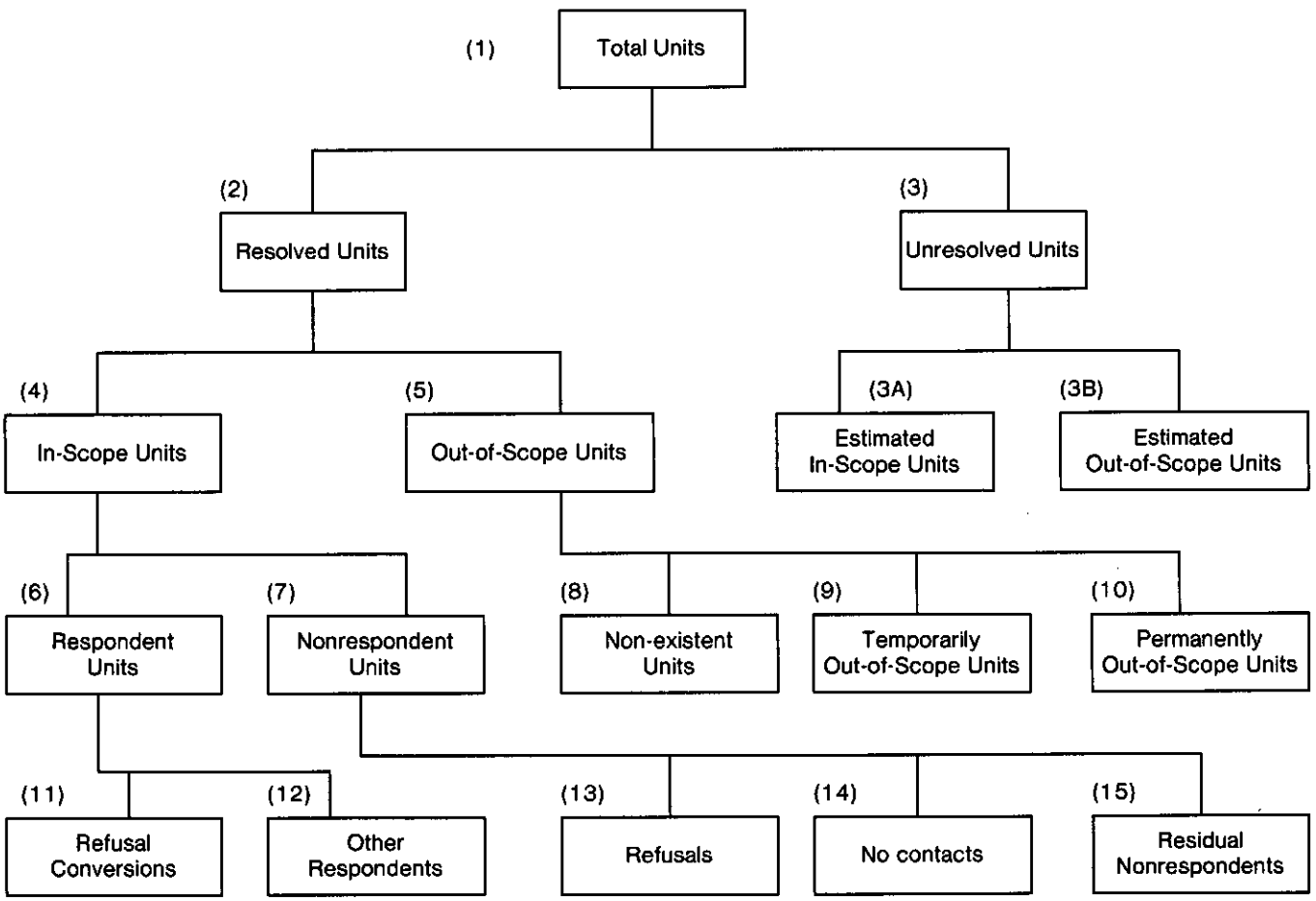


Figure 1. Respondent/Nonrespondent Components at the Data Collection Phase

out-of-scope (*e.g.*, the unit no longer exists, or it exists but it is out-of-scope), versus those that are in-scope and should have responded. **Unresolved Units** are units whose status cannot be determined by the end of data collection for the survey. The number of Unresolved Units may be broken down into **Estimated In-Scope Units** and **Estimated Out-of-Scope Units** by apportioning the nuber in the same ratio as for the Resolved Units for example. **A Resolved Rate** may then be defined as the *ratio of the number of Resolved Units to the Total number of Units*. The two components of the Resolved Units, *i.e.* In-Scope (Box 4) and Out-of-Scope (Box 5), lead to two complementary rates: the **In-Scope Rate**, defined as the *ratio of the number of In-Scope Units to the number of Resolved Units* and its complement, the **Out-of-Scope Rate**.

The Out-of-Scope Units (Box 5) may be split up into as many as three categories, some of which may not be applicable to a particular survey. These include Non-existent (Box 8), Temporarily Out-of-Scope (Box 9) and Permanently Out-of-Scope (Box 10) Units. The **Non-existent Units** include business deaths, that is, companies that have gone out of business, and dwellings that have been demolished. For recurring surveys, once it is determined that a unit is non-existent, it is excluded from data collection on future survey occasions. The **Temporarily Out-of-Scope Units** are units that were Out-of-Scope at the time of the survey, but which might be in-scope later. Hence, units can be temporarily Out-of-Scope even for single occasion surveys. For recurring surveys, it is necessary to recontact temporarily out of scope cases periodically in case their status has changed. Examples include businesses that are inactive due to seasonal factors, seasonal dwellings whose occupants have a usual place of residence elsewhere, and vacant dwellings. The **Permanently Out-of-Scope Units** result from improper classification on the frame because of changes in the classification since the frame was last updated. These cases may be screened out during the first stage of response. The Out-of-Scope Rate may be split up into three component rates: the **Non-existent Rate**, defined as the ratio of the number of Non-existent Units to the number of Resolved Units. The **Temporarily Out-of-Scope Rate** and **Permanently Out-of-Scope Rates** rates are similarly defined.

The In-scope Units (Box 4) may be broken down into Respondent (Box 6) and Nonrespondent Units (Box 7). The **Respondent Units** include in-scope units that have responded by the cutoff date for the data collection and have provided "usable information". The notion of "usable information" applies to respondents who provide only partial information. A threshold is needed in terms of level of completion of the questionnaire, below which units are considered nonrespondents. The **Response Rate** may be defined in different ways, depending upon the intended analysis. We prefer to define it as ratio of the *number of Respondent Units to the number of In-Scope and Unresolved Units*. This ratio is a conservative measure of the quality of the frame and the data collection procedures, since some Unresolved Units may be Out-of-Scope. An alternative definition would include only the number of In-Scope Units in the denominator. That rate, a conditional response rate given the known status of the units in the sample, measures the efficiency of the data collection procedure alone. The **Nonrespondent Units** (Box 7) are the remainder of the In-Scope Units. The **Nonresponse Rate** is defined as the complement of the Response Rate. It is the *ratio of the number of Nonrespondent and Unresolved Units to the number of In-Scope and Unresolved Units*. Alternative definitions omit the Unresolved Units in the numerator and denominator or apportion the Unresolved Units between estimated numbers of In-scope and out-of-scope units.

To determine the effort needed to convert Refusals to Respondents at the data collection stage, the Respondent Units are divided between Refusal Conversions (Box 11) and Other Respondents (Box 12). The **Refusal Conversions** are those who refuse initially in the current or previous collection period, and are successfully converted to be respondents because of follow-up interviews. The **Refusal Conversion Rate** is a measure of the success in converting refusals to respondents. Instead of being merely a component of the Response Rate with the same denominator, the **Refusal Conversion Rate** is defined as the *ratio of the number of Refusal Conversions to the number of Refusals and Refusal Conversions*. For completeness, we label respondents who were not refusal conversions as **Other Respondents**.

Finally, the Nonrespondent Units (Box 7) may be broken down into three components; *viz.* Refusals (Box 13), No Contacts (Box 14) and all remaining categories, that is, the Residual Nonrespondents (Box 15). The **Refusals** are nonresponding units that have been contacted but refuse to participate in the survey. The **Refusal Rate** is defined as *the ratio of the number of Refusals to the number of In-Scope Units*. The **No Contacts** are in-scope units that cannot be contacted. For social surveys, these include dwellings whose occupants were temporarily absent and households where no one was at home when interviews were attempted. The occupancy status of such dwellings is determined through observation, or where applicable by speaking to building superintendents. For business surveys, these include telephone respondents who cannot be reached, and mail nonrespondents known to be in-scope, but who were not contacted as part of any nonresponse follow-up. The **No-Contact Rate** is defined as the *ratio of number of No-Contacts and Unresolved Units to the number of In-Scope and Unresolved Units*. The **Residual Nonrespondents** include units that did not respond due to special conditions (for example, language problems, or inaccessibility) as well as respondents who provided no usable information. Special conditions also include in-scope units for which interviews

were not attempted. This is to avoid unwanted overlap between samples for different surveys, as a measure to prevent undue respondent burden. While these latter units differ from other nonresponse in that interviews are not attempted, it is important that they be considered as nonrespondents in deriving nonresponse adjustment factors at the estimation stage. The **Residual Nonrespondent Rate** is the *ratio of the number of Residual Nonrespondents to the number of In-Scope units.*

The rates defined above are at the unit level. Clearly, rates can also be defined at the item level, that is, for individual items on the questionnaire. Typically an item tends to be completed, missing, or in error as detected during editing. Hence, at an item level one can define a response rate, a missing rate, and an edit failure rate. If the missing or edit failures are imputed, one can define an imputation rate. These rates can be defined for unit respondents only, which would generally be preferable if unit nonresponse is treated by reweighting. Alternatively, unit nonrespondents can be included in both the numerators and denominators for the rates, which would be preferable in cases where unit nonresponse is treated by imputation.

We apply the definitions of nonresponse as provided above to several business and social surveys. Table 1 presents annual average nonresponse rates for several Statistics Canada surveys.

Nonresponse rates are highest for three of the social surveys and stem from: (i) the sensitivity of income as a subject matter in the case of the Survey of Consumer Finances, (ii) the respondent burden due to the length of the interview in the case of the Family Expenditure Survey, and (iii) the combination of inexperienced interviewers, telephone survey methodology and nonproxy reporting for the General Social Survey. Nonresponse is very low for the LFS because this is a long-standing flagship survey where many steps are taken to keep nonresponse low. Nonresponse rates are low for the business surveys in Table 1. Some initiatives were undertaken during the recent business survey redesign program, at reducing nonresponse.

3. FACTORS AFFECTING NONRESPONSE

Several survey design factors impact on response and nonresponse. In this section, we begin by briefly examining the influence of the frame and sample design. We follow with a more in depth examination of data collection, in terms of its organization, interviewer training, technology, mode of primary collection, and questionnaire design. Also, methods used for follow-up of nonresponse and edit failures, and use of administrative data to replace direct collection are considered.

Table 1
Response Rate Components for Selected Surveys Data Collection Stage (Rates in %)

	COMPUTATION	FAMEX	ASM	GSS	SCF	LFS	RTS
Resolved rate	(2)/(1)	100.0	100.0	98.1	100.0	100.0	95.8
In-Scope rate	(4)/(2)	92.0	95.3	51.2	86.3	85.1	97.0
Response rate	(6)/[(3)+(4)]	72.9	92.8	75.9	73.9	94.4	94.0
Refusal Conversion rate	(11)/[(11)+(13)]	N.A.	N.A.	26.7	N.A.	N.A.	N.A.
Nonresponse rate	[(7)+(3)]/[(3)+(4)]	27.1	7.2	24.1	26.1	5.6	6.0
Refusal rate	(13)/(4)	16.2	7.2	13.2	23.7	1.5	1.7
No-Contact rate	[(14)+(3)]/[(3)+(4)]	5.1	0.0	5.9	2.3	3.6	4.3
Residual nonresponse rate	(15)/(4)	5.8	0.0	5.8	0.0	0.4	0.0
Out-of-Scope rate	(5)/(2)	8.0	4.7	48.8	13.7	14.9	3.0
Non-Existent rate	(8)/(2)	0.8	2.5	0.0	0.3	0.3	2.3
Temporarily Out-of-Scope rate	(9)/(2)	7.1	1.2	0.0	13.4	14.6	0.5
Permanently Out-of-Scope rate	(10)/(2)	0.0	1.0	48.8	0.0	0.0	0.3

FAMEX: Family Expenditure Survey (1990).
ASM: Annual Survey of Manufactures (1989).
GSS: General Social Survey Cycle 5 (January-March 1990).
SCF: Survey of Consumer Finances (1991).
LFS: Labour Force Survey (1990).
RTS: Retail Trade Survey (December 1990).

3.1 Frame

Duplication or overcoverage on a frame can be irritating and lead to nonresponse if there are no procedures for unduplication, or if the procedures are not always successful. For business surveys, accurate classification information is essential if the survey is industry specific or uses industry specific questionnaires. For example, if a sampled business receives a questionnaire not pertaining to its industrial activity, it is unlikely to respond. Accurate information on the coverage of complexly structured businesses is necessary to provide respondents with a good description of the required geographical and/or industrial information. Similarly, information on contact persons within the business is needed to establish good reporting arrangements with the respondent. Inaccuracies in the contact information will cause delays in getting the required data. Inaccurate coverage description will result in improper or incomplete data being provided by the respondent.

The samples for Business surveys at Statistics Canada are drawn from a file known as the Business Register. It is a list frame that contains relevant information for selecting and contacting samples of business respondents. It has recently been redesigned using a comprehensive model reflecting the real-world complexity of business respondents. The processes incorporated in the Business Register minimize the impact of the above causes for nonresponse. Duplication is kept to a minimum by continually linking the changes that are occurring to existing units on the Business Register. These changes include births, amalgamations, splits and mergers of business respondents. Several events can signal changes to the structure of large businesses, including different administrative sources and direct survey feedback. These signals trigger a "profiling" action, *i.e.* contact with the business to redefine its structure. In the absence of signals, structures are profiled on a periodic basis, at a frequency depending upon their significance and their propensity to change. The profiling exercise gathers the necessary information to update the model. More details of the required actions are provided by Colledge (1989). The source of updates is a combination of administrative updates, profile updates and direct survey feedback. Contact, coverage and questionnaire type is kept up-to-date for each sampled unit by setting up and maintaining a computerized collection system for sampled businesses for each survey of interest. The resulting collection units are automatically built and kept up-to-date using well defined rules that vary from survey to survey. The questionnaire type takes account of factors such as: the periodicity of data collection, industrial classification, any seasonal considerations for sub-annual surveys, and fiscal year ends for annual surveys. Automatic maintenance of these collection units is carried out using a wide range of updates to the Business Register. These updates encompass activity status (live, dead, seasonal), name, address and telephone changes as well as structural changes to the surveyed unit.

The adequacy of the frame plays a similar role for social surveys in reducing nonresponse. The frame in combination with the sample design and collection procedures is important: in ensuring manageable interviewer workloads, in providing information to facilitate contact of respondents by interviewers, and in preventing unwanted overlaps in the sample across surveys. The Labour Force Survey (LFS) serves as the main vehicle for the conduct of social surveys based on area sampling. Presently, most other social surveys are supplements to the LFS, that are administered through add on questions to LFS respondents. Some surveys, due to the length of the interview or sensitivity of the subject matter are not suitable as supplements. Instead, they are based on separate samples of households drawn from the LFS frame and design.

The LFS is based for the most part on an area frame, and initial contact with sampled households is generally by face-to-face interview. The efficiency of the area frame deteriorates over time; dwelling counts for the sampling units used to determine the selection probabilities of the sampling units and interval of sampling become out-of-date. This makes it harder to plan and maintain manageable interviewer workloads. The principal mechanism for keeping the area frame up-to-date is a sample redesign following each decennial census of population. Other measures have included *ad hoc* frame updating restricted to high growth areas identified by the mid-decade census. Another measure taken in the 1981 redesign was the creation of so-called buffer strata on the outskirts of large urban centres. This involved a simple design that could be readily updated without affecting the remainder of the frame in the event that growth of the urban centre reached out into the buffer zone. To prevent interviewer workloads from becoming unwieldy when units experiencing large growth enter the sample, sub-sampling is done. For cases of extreme growth, area sub-sampling is resorted to, in which the areal unit is sub-divided into new units, a sub-sample of which is selected. If the growth is not too high, the original sample unit is retained. The rate of sampling is modified to reduce the number of dwellings selected to the point where it no longer poses a problem in terms of the interviewer's workload.

Besides the area frame, a list frame of apartment buildings is used by the LFS in larger cities. This list is kept current using information on building permits. To facilitate contact with sampled dwellings in the apartment sample, telephone numbers are obtained, where possible, by matching address information to telephone company files. Supplying interviewers with telephone numbers in this fashion has proven useful since it gives them an additional means of contacting selected dwellings that are difficult to access due to security systems, or where it is difficult to find people at home. Since the introduction of this procedure, while the nonresponse rate for the apartment frame remains higher than that for the area frame, the gap has narrowed from 8.6% to 6.2%. An alternative

to the area frame used by a few social surveys is a telephone frame. Sampling is based on Random Digit Dialing of numbers within "banks" of numbers containing working residential numbers. The banks are updated using files purchased from telephone companies. To prevent undue respondent burden, telephone numbers of households currently or recently in the LFS or other surveys using the area frame are excluded from the telephone surveys.

The LFS is currently being redesigned. Consideration is being given to adopting an address register as a list frame in urban areas. An address register of residential dwellings was created as a coverage improvement tool in the 1991 Census, and is being updated to reflect the Census enumeration of dwellings (Swain *et al.* 1992). Ways of updating the address register on an ongoing basis using administrative records or information from the postal service, and using it as a frame for social surveys are currently under study. An address register based frame should impact positively on field operations and nonresponse. Telephone numbers will be available for up to 70% of dwellings as a tool for interviewers to facilitate contacting households. Due to its regular updating, the sample can be designed to have good control on interviewer workloads, without having to resort to measures such as sub-sampling as are required under the area frame. Additionally, for the redesign, it is planned to build in mechanisms for both area and list frames to track all dwellings that are selected for Statistics Canada surveys.

3.2 Sample Design

The sample size for a survey is arrived at by taking into account budgets, survey objectives and desired level of reliability for key variables for the primary domains of interest. The overall sample size and survey design strategy should also allow for follow-up of non-responding units. In Section 4, we illustrate this point for the recently redesigned Monthly Wholesale and Retail Trade Surveys at Statistics Canada.

Business and Agricultural Surveys are stratified by a number of key variables including the size of the units. Because of the highly skewed nature of the distribution of key variables in the population, the size stratification results in a take-all and a number of take-some strata. Units in the take-all stratum cannot be rotated out of the sample, unless they become smaller in size over time. Optimum sampling plans that minimize the overall sample size for given levels of reliability may require too many units in the take-all stratum. To minimize response burden, some surveys restrict the number of take-all units; for example, the National Farm Survey (Julien and Maranda, 1990). Another means under consideration to reduce the response burden among the large units is the integration of questionnaires and/or data collection for several surveys. This implies that only distinct statistical data need to be collected for the different surveys.

Response burden among the smaller units can be reduced by periodic rotation of sampled units. However, rotation of units increases the cost of the survey because of additional sample maintenance, additional training of interviewers and difficulties in grooming new units to provide data. Partial rotation of sampled units at some fixed rate is undertaken as a compromise between 100% rotation which is very expensive and gives poor estimates of change, versus no rotation at all which would result in an unacceptable distribution of response burden. The rotation schemes keep a unit in the sample for a given period of time, after which the unit would be ineligible for reselection by the same survey for a minimum period. Surveys using such a scheme include: the Survey of Employment, Payrolls and Hours (with rotation of approximately 1/12th of the take-some units of the sample every month), the Monthly Wholesale and Retail Trade Survey (with rotation of approximately 1/24th of the smaller sized units every month), and the Labour Force Survey (with rotation of 1/6th of the sample every month). Another way to reduce response burden for individual units of Business and Agricultural Surveys is to minimize the overlap between surveys. This can be accomplished using a technique known as synchronized sampling. This technique attaches a permanent random number between 0 and 1 to each unit in the population. Different surveys are then allotted subsets of the interval (0,1) and all units whose random number falls within a survey's allotted subset are selected for that survey.

One of the objectives in the redesign of the Labour Force Survey to be introduced in 1995-1996 is to achieve a general household survey vehicle. Several new recurring social surveys are scheduled to start up in the mid-1990's, including a longitudinal survey of labour and income dynamics, and a health survey. The LFS redesign will consider not only LFS requirements, but requirements of these other surveys. Elements of the general survey orientation will include a common frame and similar sample designs with general purpose stratification. It will also feature co-ordinated sampling with overlap of selected primary sampling units (PSU's) to permit common interviewers across surveys. Unduplication of samples of dwellings between surveys to avoid respondent burden will also be carried out.

3.3 Data Collection Procedures

While all facets of the survey design can influence the survey response rates, data collection procedures and operations have the most direct and important bearing. In this section we examine the data collection procedures for business and social surveys, and the impact that factors such as the organization, the interviewer, mode of collection, technology, follow-up strategies, and response incentives have on nonresponse.

3.3.1 Organization of Data Collection

Data for business surveys are collected primarily through mail surveys with telephone follow-up. Before the mid-1980's, the collection and editing of business survey data was carried out principally in the subject matter divisions of Statistics Canada at its Head Office. This resulted in over seventy percent of the staff in these divisions being assigned to the processing of survey data. For many business surveys, regional offices had the responsibility of collecting data for nonrespondents to the surveys. During the mid-1980's, it was recognized that better use of Head Office and regional office resources could be made by a shift in the organization of data collection. The shift resulted in the concentration of collection and data capture activities within one division at Head Office specializing in the collection of annual data, and the regionalization of data collection for sub-annual surveys to the regional offices. The benefits of this reorganization were as follows: (i) operational resources could be used more effectively, (ii) the division of resources between the Head Office and regional offices could be better allocated, (iii) the increasing complexity of data collection could be handled by groups specialized in this activity, and could more readily exploit technical innovations and movement towards more integrated collection procedures, (iv) regional offices could establish "warm" contacts with the potential respondents on account of their geographical proximity to them, and (v) regional offices could offer services to users that would enhance Statistics Canada's presence among the potential responding units. All this helped in reducing the nonresponse rates.

Data for the social surveys are collected through a combination of face-to-face and telephone interviews. The monthly Labour Force Survey and most other social surveys conducted by Statistics Canada use a dispersed field force of approximately 1,000 interviewers across the country. The interviewers do a mixture of telephone interviewing from their homes and face-to-face interviewing. They are supervised by 100 senior interviewers. Project managers located in each of Statistics Canada's regional offices are responsible for the work of 3-4 senior interviewers. For the LFS, project managers and seniors are provided with performance reports each month for the interviewers they supervise. The reports include measures such as edit failure rates, nonresponse and costs. This continual feedback improves data collection procedures, thereby having a positive impact on response rates. For social surveys, there was no alternative to the dispersed organization before the advent of telephone survey methods. From 1985-1989 a program of research and testing of telephone survey methods was carried out (Drew 1991), in which a mixed organization was considered. Under this organization the role of local interviewers would be restricted largely to one of conducting face-to-face interviews, and telephone interviewing would be carried out from the regional offices. The mixed organization would provide less opportunity for face-to-face follow-up of households that could not be contacted by telephone, leading to somewhat higher nonresponse. Also, the mixed organization would have higher overhead costs for extra office space and equipment in the regional offices. It would result in a much smaller field force, reducing the flexibility to carry out large scale *ad hoc* surveys requiring face-to-face interviews. Also, the pool of experienced field staff would be reduced to tap into each 5 years for the census of population. Based on these considerations, it was decided to retain the dispersed organization.

3.3.2 Interviewers

When new interviewers are hired for the Labour Force Survey, they are paid for 5 hours of home exercises and reading material, followed by three days of classroom training. During their first two days of interviewing in each first two months, new interviewers are observed by the senior interviewer. In addition, interviewers are routinely provided with material to read at home, and with exercises to complete dealing with different aspects of the survey taking procedures. Also, home studies are available to deal with specific problems identified in head office editing of the data. All interviewers receive an additional three days of classroom training per year. For supplements, training generally takes the form of reading material and self-study exercises to complete at home. For business surveys, the number of interviewers is much smaller, 260 in total. Training and monitoring are similar to those in the Labour Force Survey.

In a comprehensive study of nonresponse, Gower (1979) found that nonresponse rates vary greatly among interviewers. Particularly of interest, Gower found that about 15% of interviewers regularly encounter little or no nonresponse to the LFS. A focus group study is planned involving groups of superior and average interviewers. It will determine how they differ both in terms of locating respondents and in convincing them to participate in the survey. The latter will be looked at from the point of view of compliance theory, drawing on the work of Cialdinni (1991). The objective will be to identify techniques being used by superior interviewers so as to teach them to other interviewers.

3.3.3 Mode of Collection

Statistics Canada places high priority on allowing respondents to choose the mode of reporting that best fits their circumstances, including the official language of their choice. Such flexibility helps in improving response rates.

Business surveys conducted at Statistics Canada can be classified in two main groups: annual and sub-annual surveys. For the annual surveys, most of the data collection

is via questionnaire mailout and mailback administered from Ottawa, with some respondents providing data via magnetic tapes or floppy disks. The timing for mailout of annual business surveys should be linked to the respondent's fiscal year end for tax reporting purposes. This is because the required data are readily available at this time, and ambiguity about the reference year is minimized. Bilocq and Fontaine (1988), in a study on the Annual Census of Manufactures, found that the best response rates were obtained by contacting respondents three months after their fiscal year end. This implies a staggered mailout that takes fiscal year end into account. For sub-annual business surveys, data collection is mostly by mailout from Head Office and mailback to the regional offices. Most of the non-mail units respond by telephone to the regional offices, while a few respondents provide computer readable responses directly to Ottawa. It is important to respect bookkeeping practices of respondents. Most respondents use the calendar month for bookkeeping, whereas others use four and five week cycles. In both cases, data are usually available to the survey agency one or two weeks after the end of the monthly period. Telephone interviewing is used to collect data in business surveys for a variety of reasons that range from clarification of instructions to follow-up action. The quality of response may suffer if this mode of collection is improperly used. For instance, a respondent may be forced to estimate the data due to lack of availability of records near the telephone. If telephone interviewing is used on a periodic basis, such as in monthly surveys, then a best day and time arrangement with the respondent will improve response rates as well as the quality of response.

For social surveys, such as the Labour Force Survey, the mode of collection is "warm" telephone interviewing, that is, households receive an initial face-to-face interview during their first month in the sample, with predominantly telephone interviews in later months. When the initial contact with the household is made, the interviewer presents his/her identification badge. The respondent is then provided with a description of the purposes of the survey, and given assurances of the confidentiality of the responses before proceeding with the interview. The face-to-face visit is preceded by an advance letter from the Regional Director, notifying the household of its selection in the survey and describing the purpose of the survey. Respondents are invited to call on a toll free number if they have any questions before or during the survey. In a program of research and testing of telephone survey methods from 1985-1989, the feasibility of replacing the initial face-to-face interview with a telephone interview was examined. The alternative of conducting the LFS as a central telephone survey led to a 68-75% increase in nonresponse rates. There was evidence of increased nonresponse bias stemming from differences in the labour force characteristics of respondents and the additional nonrespondents (Drew 1991). The only recurring household survey at Statistics Canada to use telephone survey methods for all its data collection is the annual General Social Survey (GSS). It uses Random Digit Dialing (RDD) in a survey of 10,000 households. On occasion the GSS sample has been augmented with households rotated out of the LFS. For example, a sample of elderly persons who had been in the LFS was selected during one round of the survey when this age group was of special interest.

3.3.4 Questionnaire Design and Introductory Material

Good questionnaire design practices contribute not only to the accuracy of the data collected, but also to the response rates. The questionnaire and introductory material are particularly important in mail collection since they are the only contact with the respondent. Material sent to respondents should include a description of the purposes of the survey, the authority under which it is conducted, assurances of confidentiality of responses, and a phone number in the agency for answering any queries on the survey questionnaire.

Questionnaires should go through a review process that is independent of the questionnaire design. This process takes the form of peer reviews by experts within the agency or focus groups of survey participants. The use of focus groups or cognitive research has resulted in several improvements aimed at respondent motivation. It has also resulted in simplification of the task of completing the questionnaires for several surveys at Statistics Canada. These include the Census of Population, the Labour Force Survey, the Census of Construction Industry and the Survey of Employment, Payrolls and Hours (Gower 1990).

3.3.5 Follow-up Strategies

For both business and social surveys, follow-up is an integral part of the overall survey design. It is only through intensive follow-up that low levels of no-contact nonresponse can be achieved. Since follow-up usually costs more per unit than primary collection (assuming a fixed survey cost), the amount of follow-up has a direct bearing on the sample size and therefore the variance, on the response rate and therefore the nonresponse bias. Design strategies range from a large sample with little follow-up to a smaller sample with intensive follow-up. In the redesign of the Monthly Wholesale and Retail Trade Survey during the 1980's, improving response rates was a priority, and this led to adoption of the strategy of a smaller sample with more intensive follow-up.

For business surveys, follow-up is undertaken both to obtain data from nonrespondents and to recontact respondents with edit failures. Most business surveys use mail as a primary mode of collection as it is inexpensive, and it gives businesses the opportunity to consult their records in responding. Nonresponse follow-up is often

restricted to a subsample of nonrespondents to reduce costs. The allocation and selection of the nonresponding units is usually based on the following factors: (i) a take-all stratum of units that must be followed-up to concentrate effort on the larger nonresponding units; (ii) an equalization of response rates across design strata; and (iii) rotation of the smaller sized nonresponding units targeted for follow-up. Nonresponse follow-up is generally by telephone for sub-annual surveys, as time constraints do not permit mail follow-up. For annual surveys, where timeliness of the collection is not as critical, mail has tended to be used for both primary collection and for initial attempts at nonresponse follow-up, with a telephone follow-up as the last resort. Increasingly, though, in recent years more of the follow-up has been by telephone for the annual surveys as well.

For social surveys, there is not as clear a distinction between primary collection and nonresponse follow-up. Follow-up consists for the most part of second and subsequent attempts to contact and interview households during the survey period. Some distinctions exist depending on the status of the dwelling. Newly sampled dwellings are initially visited to identify those that are out-of-scope and to attempt a face-to-face interview with occupants of in-scope dwellings. In cases where an interview cannot be obtained, the interviewer attempts to obtain information such as name, telephone number, and best time to call from a neighbour. Interviewers are instructed to make two to three additional attempts to interview. These follow-ups can be either by telephone or face-to-face. Occupants of previously sampled dwellings are generally interviewed by telephone. However, if repeated attempts at telephone contact are unsuccessful, a face-to-face visit is made, to insure the dwelling is still in-scope and to attempt an interview.

While follow-up is needed to bring nonresponse to acceptable levels, there is a point after which further follow-up yields diminishing returns for the money expended. There has been little work aimed at addressing the question of appropriate strategies for the scheduling and the number of follow-ups based on cost and total error considerations. Studying this issue would require cost studies to estimate parameters in a cost and mean squared error model. The factors would include contact attempts, outcomes, costs, and characteristics of respondents at different stages of follow-up. The increased automation of data collection in the years ahead should make it more feasible to collect and use such information to optimize data collection strategies.

3.3.6 Technology

Data collection for business surveys is mostly by paper and pencil. Notable exceptions are the Monthly Survey of Manufacturing where CATI is currently being used (Coutts *et al.* 1992), and the Annual Survey of Manufacturing where CATI has been used experimentally to collect data from the smaller manufacturers. With the successful implementation of CATI for the Monthly Survey of Manufacturing, plans are under way to employ CATI for other business surveys. Experiments are also currently being carried out to test other data collection technologies for business surveys. These include: a hand-held computer for the Consumer Price Index, the Grid Pad for the Quarterly For-Hire Trucking survey, and touch tone data entry for the Survey of Employment Payrolls and Hours.

Data collection for social surveys is also based on paper and pencil technology. A decision has been taken to move to Computer Assisted Interviewing (CAI) over the next few years. The dispersed interviewing staff will be equipped with portable computers for face-to-face interviewing and for telephone interviewing from their homes. The decision was made based on positive findings from two tests of CAI on the LFS. The first test (Catlin and Ingram 1988) showed: data quality improvements such as better enumeration of persons within sampled dwellings, and fewer edit failures, with no detectable impact on survey estimates or response rates. The second test in 1991 (Coutts *et al.* 1992) demonstrated the operational viability of portable computers for CAI by interviewers in the field. Plans are to begin converting social surveys to CAI as early as 1993. These will depend on the results obtained from more extensive testing during 1992. Factors to be considered will include its impact on survey estimates, and on data quality, including response rates.

3.3.7 Response Incentives

Under the Statistics Act that sets out the legal framework governing Statistics Canada, participation in Statistics Canada surveys is mandatory for those businesses and individuals selected for survey unless the Chief Statistician designates the survey as voluntary. An example of a mandatory program is the Census of Population, where an outright refusal can lead to prosecution. For other programs, the agency relies on obtaining the co-operation of potential respondents via advance written material or publicity explaining the purpose of the survey and the confidentiality of the data, and "door step diplomacy" measures such as display of badges by face-to-face interviewers, and informing respondents about purpose and confidentiality.

Several studies of the use of response incentives have been carried out for social surveys. The first was on the Labour Force Survey (Gower 1979). In a split sample test, the Canada Handbook was given to half the households when first contacted. The result was a marginally lower refusal rate in later months for the sample receiving the incentive. Interviewers believed that the incentive was of marginal benefit, and that existing door-step procedures

were more important in reducing nonresponse. More recently, in an incentive study in the 1990 Family Expenditure Survey three treatments were administered at the interviewer level: one in which each selected household received a clipboard with the Statistics Canada logo, a second receiving the Statistics Canada publication "A Portrait of Canada," and a control sample receiving no incentive. At the national level, there was no significant change in the response rates (Kumar and Durning 1992). A study of response incentives is also planned for an upcoming longitudinal survey of income and labour.

3.4 Selective Editing

Another potential cause for nonresponse is faulty editing procedures that result in several recontacts with the respondent for the same questionnaire, lessening their willingness to cooperate on future occasions. To streamline and optimize the editing process to minimize recontacts, the following three measures should be followed. First, editing at the data capture, follow-up and imputation stages should be consistent. Second, selective editing ought to be applied to numeric data especially in business and agricultural surveys. Records that have a significant impact on the estimates are identified, and follow-up is restricted to those records. The records with a small impact should be subjected to an automated edit and imputation process to ensure consistency. Third, to keep response burden to a minimum, all errors should be identified for the units to be followed-up so that most errors can be cleared up in a single contact. The use of an inter-field edit analyzer and error localizer, such as the one in the Generalized Edit and Imputation System developed at Statistics Canada, is recommended for this requirement (Kovar, MacMillan and Whitridge 1988). If too many items fail edit but prove to be correct on follow-up, the edits should be adjusted to alleviate unnecessary response burden.

Selective editing procedures for numeric data developed at Statistics Canada can be grouped in three sets: (i) statistical editing, (ii) grouping of variables and (iii) a score function. For statistical editing, Hidiroglou and Berthelot (1986) have developed a transformation that allows more emphasis on detecting units that show unusual changes from occasion to occasion. It recognizes that period to period changes for small units are inherently more variable than changes for large units. The cut-off bounds for edit failures are thus funnel shaped, allowing large relative changes in small units. These bounds are calculated using medians and quartiles, and are thus robust to outlier observations in the data. This method can also be used to detect outlier ratios between two variables. However, the number of pair wise comparisons can become prohibitively large. Bilocq and Berthelot (1990) recommended a method of grouping the variables into subsets of related variables and then only cross editing variables within the subsets. The procedure used for this partitioning is based on principal component correlation methods. The significance of the errors as measured by their influence on the estimates must be considered as well. In the case of edit failure for completed questionnaires, Latouche and Berthelot (1992) have developed a score function that assigns a relative score of error importance to each respondent based on the size of the unit, the size and number of suspicious data items on the questionnaire and the relative importance of the variables. It has been demonstrated in a simulation study using this idea, that recontacting a few units is sufficient to ensure acceptable data quality for the final estimates.

3.5 Administrative Data Considerations

Response burden for Business and Agricultural Surveys at Statistics Canada is being alleviated by obtaining some data for the smaller sized units from administrative sources. Such data are also used to replace illegible, inconsistent or missing survey data. For example, the data for the smaller sized nonresponding units is imputed using tax data.

3.6 Management System for Data Collection

A good tracking system is required to determine the status of the collection process at any time. For Business Surveys, collection status codes, whose history is kept for each surveyed unit, are used to control the collection process. These collection status codes, stored in the time sequence that the survey is being carried out, are used with other codes that reflect the activity status of the unit (active, seasonal with operating dates provided, out of business, temporarily closed, *etc.*). Examples of collection status codes are: i) mode of data collection at different time points of the data collection process, ii) contact initiation codes for units (known to be active during the reference period), and for exclusions (which include closed units, out of business, temporarily closed), and iii) expected dates for return of the information to prompt additional follow-up. The management system receives information from sources external to the survey indicating a change in the status of units, and tracks the collection status from initial data collection to follow-up until all the units are ultimately classified into one of the categories under the framework described in Section 2.

4. ANALYSIS OF NONRESPONSE FOR SELECTED SURVEYS

We will briefly examine nonresponse for two surveys at Statistics Canada, to illustrate some general factors impacting on nonresponse described in Section 3.

4.1 The Monthly Retail Trade Survey

The Monthly Retail Trade Survey (MRTS) is a survey that collects sales from a sample of retail locations and inventories for a sub-sample of them. Estimates of the level and change are generated for these two variables. The sample design is a rotating simple random sample of companies stratified by province, industry and gross business income. The population size is approximately 165,000 companies, and the sample size is about 13,000. Data are collected by telephone for approximately 40% of the units and by mail for the remaining 60%. Preliminary estimates are published 7 weeks after the survey reference period, and final estimates, which include more respondents because of nonresponse follow-up are released a month later.

A redesign of the survey was implemented in January 1990. The new design differed in several aspects from the old one that had been in place since the early seventies. First, to increase the design efficiency, the number of industry groups was reduced from 34 to 18 and three size strata were used in place of two. Second, the levels of reliability were relaxed with the new design. These changes permitted a sample reduction of 35%, allowing intensive follow-up of nonrespondents. Third, data collection was decentralized to the regional offices. Under this strategy, data collection costs were higher on a per unit basis on account of the extra follow-up. There was, however, an overall gain in quality of survey results due to the reduction in nonresponse.

Both preliminary and revised weighted response rates, defined as the ratio of the estimate of sales contributed by the respondents to the estimate of sales for all in-scope units are provided in Figure 2 for the period 1986-1992. From this graph, both preliminary and revised response rates are substantially higher for the new survey than for the old survey. Preliminary rates have risen from 75% to 93%, while final rates have risen from 85% to 95%. It is also clear that the gap between the preliminary and revised response rates is much smaller for the new survey. It should be noted that in September, 1991 the preliminary rates were lower than expected because of a strike by the clerical staff handling the documents.

Several factors have contributed to the improvement in the response rates, the most important ones being mode of data collection and follow-up procedures. In the old survey, questionnaires were mailed out from and returned to Head Office (Industry Division). The mailout was carried out using manually controlled reporting arrangements. Immediate follow-up of nonrespondents was restricted to large units, and was done by telephone from Ottawa. Smaller sized nonresponding units were followed up by mail one month later, and the mail follow-up was continued for up to two additional months. Nonrespondents which had not responded for three consecutive months were referred to the regional offices for a telephone follow-up.

For the new survey, prior to their first occasion in the survey, newly sampled units (new entrants) are mailed an advance letter explaining the survey and the importance

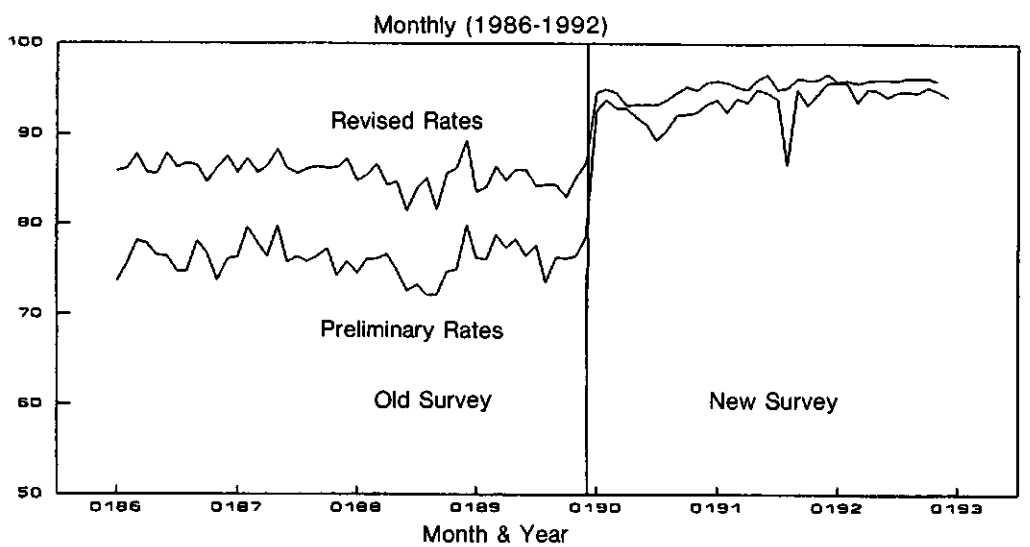


Figure 2. MRTS Résponse Rates

of their participation. A blank questionnaire is included. Also, each new entrant is telephoned about a week after the expected receipt of the advance letter to explain difficult ideas, to answer questions, and to offer a choice of mail or telephone data collection. For the mail respondents, questionnaires are mailed out by Industry Division using automated collection arrangements that are derived from the information on the Business Register. These collection arrangements are updated on the Business Register via profiles carried out by the Business Register Division, as well as new information found out by the regional offices during their contact with the respondent. Regional offices request data from the telephone respondents at pre-arranged dates and times, and the collected data are transmitted to Head Office after each monthly collection cycle.

4.2 Labour Force Survey

The Labour Force Survey is the largest continuous social survey conducted by Statistics Canada with a sample size of approximately 62,000 households per month. The impact of different aspects of the survey design on LFS nonresponse were discussed in Section 3. In this section, we examine historical trends in nonresponse and consider in more detail the role of nonresponse follow-up.

Table 2 below shows that the overall nonresponse rates have been steady in the 4% - 5% range throughout most of the period 1977-1991, as have refusal rates, in the 1.0% - 1.5% range. However, a few patterns are evident. One is the positive effect of the Census of Population on the nonresponse rates for the LFS, pointing to the benefit of the publicity surrounding the Census spilling over to household surveys. Nonresponse rates dropped by 1.0% between 1980 and 1981, and by 0.6% between 1985 and 1986, and by 0.4% between 1990 and 1991, the only years in which substantial drops in nonresponse rates have occurred. In 1986 virtually all the decrease was in refusals, while these accounted for over half the reduction in 1981. While the changes in nonresponse over the period are not dramatic, a gradual lessening of the positive effects of the Census is apparent. There is a slight increase in the last four years in both nonresponse and refusal rates as compared to the period from 1981 to 1987.

The graph below (Figure 3) giving the nonresponse and temporarily absent rates by month shows: the seasonal trends in the rates, with a peak in the summer months for the overall nonresponse rates, accompanied by a parallel increase in the Temporarily Absent rate. The strong relationship between the overall nonresponse rate and the Temporary Absent rate is apparent in the graph. The data collection period for the survey is normally a six day period from the Monday to the Saturday following the reference week. By Saturday of interview week the interviewers have returned all their cases to the regional offices. To reduce the seasonal peak, a Monday follow-up procedure was started in the late 1970's for the July and August surveys. Occasionally, the Monday follow-up is extended to June depending on the school year. The Monday follow-up of nonrespondents who could not be reached during the survey week is carried out from the regional offices. It has been observed that it reduces the number of cases of Temporarily Absent nonresponse.

From 1984 onwards, there has been a change in the pattern of seasonal peaks in Temporarily Absent Nonresponse. The summer peaks are less severe, but a second peak in February and March is becoming more pronounced. This seems to reflect a shift in vacation patterns of households toward more winter breaks. Consequently, in recent years, the Monday follow-up has been carried out in March if the survey week coincides with the school break.

Another noticeable feature in the LFS nonresponse pattern is higher nonresponse for households that are in the sample for the first time than for the other households. In 1980, the nonresponse rate for the first month interview households was 6.9% versus 3.5% for later months. Most of this difference occurs in the No Contact component of nonresponse. Since interviewers employ mostly face-to-face interviewing in the first month, they are limited in the number of contact attempts they can make. In later months, telephone interviewing and information obtained during the initial interview on the best time to call lead to a substantially improved contact rate.

During the 1981 post-censal redesign of the LFS, a detailed time and cost study was undertaken. The primary purpose of the study was to obtain cost information needed to carry out a cost/variance optimization of the

Table 2

LFS Nonresponse and Refusal Rates by Year

	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
NR	5.42	5.39	5.35	5.37	4.41	4.67	4.65	4.57	4.69	4.08	4.23	5.07	5.18	5.57	5.20
REF	1.34	1.45	1.41	1.47	1.16	1.19	1.14	1.18	1.18	0.99	1.06	1.30	1.31	1.51	1.38

NR = Overall Nonresponse rate.
REF = Refusal rate.

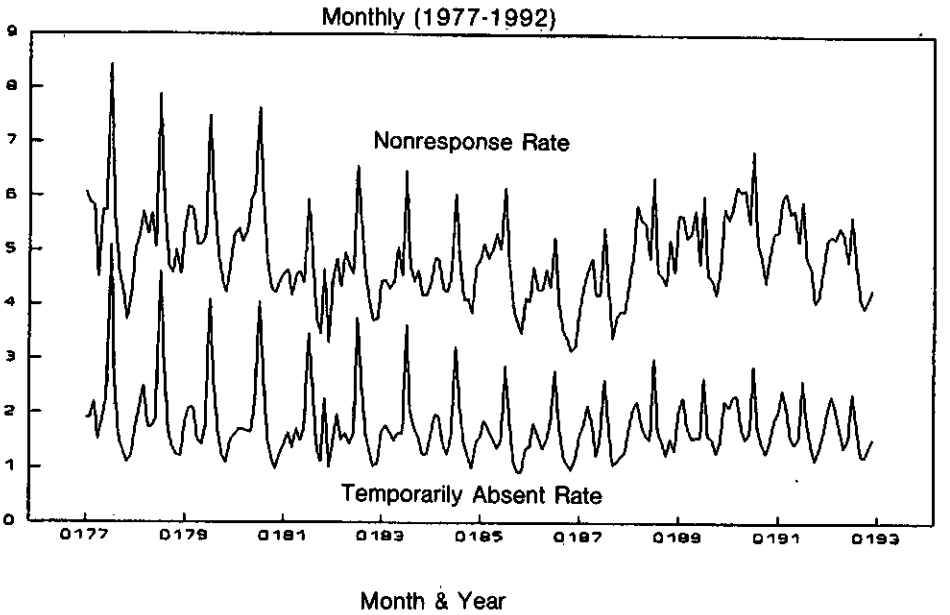


Figure 3. LFS Nonresponse Rates

survey design. The study reported by Lemaître (1983) also yielded interesting information on interviewer movement and household visit patterns, and the effect of nonresponse follow-up on response rates under face-to-face interviewing. He found a response rate of 92.4% was achieved after 3 visits, with the ratio of responses per visit consistently high at 56-61% for each round of visits. More extensive follow-up was carried out for only 3.5% of dwellings. These dwellings were visited on average another 2.5 times, with only 29% of such visits resulting in a response. The extra visits for these households accounted for 5.8% of all dwelling visits, and increased the response rate by 3.1% to 95.1%.

The 1983 Time and Cost Study was undertaken before the introduction of telephone interviewing in smaller urban and rural areas for non-first month in the sample cases, and before the introduction of telephone follow-up of first month nonresponse cases. Consideration is being given to repeating the study under the current survey conditions. One of the questions such a study could address is the cost benefit of extra visits to reduce nonresponse rates. While fourth and subsequent visits may not represent a high proportion of visits, their contribution to collection costs may be considerably higher due to the dispersion of such dwellings. Costs of such visits, coupled with information on their characteristics relative to those of other respondents, would permit an assessment of how much follow-up is warranted based on cost and mean squared error considerations.

5. SUMMARY

In this paper we have presented standards for the definition of nonresponse. In a pilot study of 7 major business and social surveys at Statistics Canada, no difficulties were found in applying the standard definitions. Beginning with the 1993 reference year, information on nonresponse for major surveys according to these standards will be reported and maintained in a central repository within the agency. This will facilitate analysis of global trends affecting response and nonresponse to surveys.

We have discussed what measures can be taken in various aspects of the survey design to help minimize nonresponse, and have illustrated their application for two major recurring surveys. Although we have restricted our focus to the role such measures play in nonresponse, they constitute good survey taking practice whose benefits encompass more than improved response rates.

In speculating about what the future holds for survey response rates in Canada, there is nothing in current trends to be alarmed about, despite a slight increase in nonresponse rates for social surveys over the last decade. However, Statistics Canada is pursuing cognitive research efforts in nonresponse aimed at better understanding respondents' attitudes and concerns about issues such as privacy, confidentiality, response burden, and record linkage. Selective editing studies are also being undertaken to focus on editing and follow-up efforts on large units. There is much scope for reducing response burden and

costs, with little impact on estimates. Findings from these studies will be helpful in designing our surveys and statistical programs in ways that respect respondents' concerns. This will permit us to continue the high levels of cooperation from the Canadian public and businesses.

ACKNOWLEDGMENTS

The authors thank B.N. Chinnappa, Statistics Canada, and the referees for their helpful comments, and to Statistics Canada: Methods and Standards Committee for its guidance and support in the development of the nonresponse framework.

REFERENCES

BILOCQ, F., and BERTHELOT, J.-M. (1990). Analysis on Grouping of Variables and on the Detection of Questionable Units. Methodology Branch Working Paper, BSMD, 90-005E. Statistics Canada,

BILOCQ, F., and FONTAINE, C. (1988). Étude sur la mise à la poste échelonnée pour le recencement des manufacturiers. Statistics Canada report.

CIALDINNI R.B. (1991). Deriving Psychological Concepts relevant to survey participation from the literatures on compliance, helping and persuasion. International Workshop on Household Survey Non-response, Sweden, October 1990.

COLLEDGE, M.J. (1989). Coverage and classification maintenance issues in economic surveys. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons, 80-107.

CATLIN, G., and INGRAM, S. (1988). The effects of CATI on cost and quality. *Telephone Survey Methodology*, (Eds. R. Groves *et al.*). New York: Wiley, 437-450.

COUTTS, M., JAMIESON, R., WILLIAMS, B., and BRASLINS, A. (1992). The building of an integrated collection operation in Statistics Canada's regional offices. *Proceedings of the 1992 Annual Research Conference*. US Bureau of the Census, 395-411.

DREW, J.D. (1991). Research and testing of telephone surveys methods at Statistics Canada. *Survey Methodology*, 17, 57-68.

DREW, J.D., and GRAY, G.B. (1991). Standards and guidelines for definition and reporting of nonresponse to surveys. Prepared for the Second International Workshop on Household Survey Non-response, Washington, DC.

GOWER, A.R. (1979). Nonresponse in the Canadian Labour Force Survey. *Survey Methodology*, 5, 29-58.

GOWER, A., and ZYLSTRA, P.D. (1990). The use of qualitative methods in the design of a business survey questionnaire. Presented at the *International Conference on Measurement Errors in Surveys*, Tucson, Arizona.

HIDIROGLOU, M.A., and BERTHELOT, J.-M. (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology*, 12, 73-83.

JULIEN, C., and MARANDA F. (1990). Sample design of the 1988 National Farm Survey. *Survey Methodology*, 16, 117-129.

KOVAR, J.G., MACMILLAN, J.H., and WHITRIDGE P. (1988). Overview and Strategy for the Generalized Edit and Imputation System. Methodology Branch Working Paper, BSMD, 88-007E. Statistics Canada.

KUMAR, S., and DURNING, A. (1992). The Impact of Incentives on the Response Rates for FAMEX 1990: an Evaluation. Methodology Branch Working Paper, SSMD 92-001E. Statistics Canada.

LATOUCHE, M., and BERTHELOT, J.-M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics*, 8, 389-400.

LEMAÎTRE, G. (1983). Results from the Labour Force Survey Time and Cost Study. Internal report, Household Survey Methods Division, Statistics Canada.

PLATEK, R., and GRAY, G.B. (1986). On the definitions of response rates. *Survey Methodology*, 12, 17-27.

SWAIN, L., DREW, J.D., LAFRANCE, B., and LANCE, K. (1992). The creation of a residential address register for coverage improvement in the 1991 Canadian Census. *Survey Methodology*, 18, 127-141.

Survey Methodology, June 1993
Vol. 19, No. 1, pp. 95–101
Statistics Canada

Double Sampling for Stratification

R.P. TREDER and J. SEDRANSK[1]

ABSTRACT

Double sampling is a common alternative to simple random sampling when there are expected to be gains from using stratified sampling, but the units cannot be assigned to strata prior to sampling. It is assumed throughout that the survey objective is estimation of the finite population mean. We compare simple random sampling and three allocation methods for double sampling: (a) proportional, (b) Rao's (Rao 1973a,b) and (c) optimal. There is also an investigation of the effect on sample size selection of misspecification of an important design parameter.

KEY WORDS: Optimal sample sizes; Two phase sampling.

1. INTRODUCTION

Suppose we wish to estimate the finite population mean in a stratified population, but the units cannot be assigned to strata prior to sampling. Typically, the number of units in each stratum is unknown. Then, double sampling is commonly considered as an alternative to simple random sampling. With double sampling, a simple random sample of size n' is selected from a finite population of N units with n_i' units identified as members of stratum i, $i = 1, \ldots, L$. The second phase sample is a set of L independent simple random subsamples where, in stratum i, n_i units are selected from the n_i' identified in the first phase. Letting y_{ij} denote the value of Y for the j-th unit in the second phase sample in stratum i, the finite population mean, $\bar{Y}$, is estimated by

$$\hat{\bar{Y}} = \sum_{i=1}^{L} w_i \bar{y}_i,$$

where $w_i = n_i'/n'$ and $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$.

Let $\sigma(n_i')$ and $\sigma(n_i)$ denote, respectively, the set of values for first phase and second phase sample units in stratum i, $n' = (n_1', \ldots, n_L')$ and $\sigma(n')$ the set of values for all first phase sample units. Also, let $\bar{y}_{n'}$ be the mean of the values in $\sigma(n')$, $\bar{y}_i'$ the sample mean of $\sigma(n_i')$, $s_i'^2 = \sum_{j=1}^{n_i'} (y_{ij} - \bar{y}_i')^2/(n_i' - 1)$ the sample variance of $\sigma(n_i')$, $S_i^2 = \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2/(N_i - 1)$ the population variance in stratum i and S^2 the analogous finite population variance. It is assumed throughout that n' is sufficiently large that $Pr(n_i' = 0)$ is negligible. Noting that $1 \leq n_i \leq n_i'$,

$$E(\hat{\bar{Y}}) = E_{\sigma(n')}\{E(\hat{\bar{Y}} \mid \sigma(n'))\} = \bar{Y}$$

and

$$\begin{aligned} V(\hat{\bar{Y}}) &= V_{\sigma(n')}E\{\hat{\bar{Y}} \mid \sigma(n')\} \\ &\quad + E_{\sigma(n')}\{V(\hat{\bar{Y}} \mid \sigma(n'))\} \\ &= V_{\sigma(n')}(\bar{y}_{n'}) \\ &\quad + E_{\sigma(n')}\left\{\sum_{i=1}^{L} w_i^2 s_i'^2 \left(\frac{1}{n_i} - \frac{1}{n_i'}\right)\right\} \qquad (1.1) \\ &= S^2\left(\frac{1}{n'} - \frac{1}{N}\right) \\ &\quad + E_{n'}\left\{\sum_{i=1}^{L} w_i^2 S_i^2 \left(\frac{1}{n_i} - \frac{1}{n_i'}\right)\right\}. \qquad (1.2) \end{aligned}$$

We assume the linear cost function

$$C = c'n' + \sum_{i=1}^{L} c_i n_i, \qquad (1.3)$$

where c' is the per unit cost associated with sampling a first phase unit, and c_i is the per unit cost of measuring Y in stratum i. The sample sizes, n' and the n_i, are selected subject to fixed total cost or to fixed total expected cost.

In this paper we compare three double sampling designs, differentiated by the way that the sample sizes, n' and the n_i, are chosen. We also compare these methods with a simple random sample having the same fixed total cost.

The alternative designs are presented in Section 2 and compared in Section 3. Section 4 presents the results of an investigation of the effect on sample size selection of misspecification of an important design parameter.

[1] R.P. Treder, Statistical Sciences, Inc. Seattle, Washington; J. Sedransk, State University of New York at Albany, Albany, New York.

2. ALTERNATIVE METHODS

2.1 Proportional Allocation

For proportional allocation, $n_i = nw_i$ where $n = \sum_{i=1}^{L} n_i$. Then, using (1.2), the variance of $\hat{Y}$ under proportional allocation, V_P, can be shown to be

$$V_P = \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) \sum_{i=1}^{L} W_i S_i^2, \quad (2.1)$$

where $W_i = N_i/N$ is the population proportion of units in stratum i. Substituting $n_i = nw_i$ in (1.3), the expected total cost is

$$\bar{C}_P = c'n' + cn, \quad (2.2)$$

where $c = \sum_{i=1}^{L} W_i c_i$. Choosing n' and n to minimize (2.1) subject to fixed total expected cost, $\bar{C}_P = C^*$, yields

$$n' = \frac{C^*}{c' + \sqrt{c'cG}}, \quad (2.3a)$$

$$n = \frac{C^*}{c + \sqrt{c'c/G}}, \quad (2.3b)$$

where $G = S_W^2/S_B^2$, $S_W^2 = \sum_{i=1}^{L} W_i S_i^2$ and $S_B^2 = S^2 - S_W^2$.

Using (2.3),

$$V_P = \frac{1}{C^*}\left\{\left(c' + \sqrt{c'cG}\right) S_B^2 + \left(c + \sqrt{c'c/G}\right) S_W^2\right\} - \frac{S^2}{N}. \quad (2.4)$$

2.2 Rao's Allocation

Rao (1973a,b) proposes selecting $n_i = \nu_i n_i'$ where the ν_i $(0 < \nu_i \leq 1)$ are constants fixed in advance of sampling. Using this allocation in (1.2), the variance of $\hat{Y}$ under Rao's allocation, V_R, can be shown to be

$$V_R = \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \frac{1}{n'} \sum_{i=1}^{L} W_i S_i^2 \left(\frac{1}{\nu_i} - 1\right). \quad (2.5)$$

The corresponding expected cost, $\bar{C}_R$, is

$$\bar{C}_R = c'n' + n' \sum_{i=1}^{L} c_i \nu_i W_i. \quad (2.6)$$

The ν_i which minimize (2.5) subject to $\bar{C}_R = C^*$ satisfy

$$\nu_i^0 = \frac{S_i\sqrt{c'}}{S_B\sqrt{c_i}}, \quad (2.7)$$

provided that the right side of (2.7) does not exceed 1 for any i. Otherwise, an algorithm is required to determine the optimal ν_i (see Rao 1973a,b). Since Rao minimizes the *unconditional* variance, the optimal ν_i do not depend on the observed n_i'. After determining the ν_i, n' is obtained from (2.6). Assuming that $\nu_i^0 \leq 1$ for each i,

$$V_R = \frac{1}{C^*}\left(\sum_{i=1}^{L} W_i S_i \sqrt{c_i} + S_B\sqrt{c'}\right)^2 - \frac{S^2}{N}. \quad (2.8)$$

2.3 Optimal Allocation

The optimal allocation of the sample sizes can be obtained by minimizing (1.2) directly. For *fixed* n' and n', select the n_i to minimize

$$\sum_{i=1}^{L} w_i^2 S_i^2 \left(\frac{1}{n_i} - \frac{1}{n_i'}\right), \quad (2.9)$$

subject to fixed remaining cost, $C^* - c'n' = \sum_{i=1}^{L} c_i n_i$ and $n_i \leq n_i'$. An algorithm is required to determine the optimal n_i given the n_i'; see Hughes and Rao (1979) and Treder (1989). One may find the optimal value of n' by evaluating (1.2) for a sequence of "trial" values of n'. For each such n', one estimates the expected value of (2.9) using Monte Carlo sampling of n' (see Booth and Sedransk (1969) and Treder (1989)). Note that the algorithm needed to find the optimal n_i is straightforward, and the Monte Carlo sampling of n' given n' is simple. There are several differences between the optimal allocation and Rao's allocation. In the former, total costs will not exceed C^* while in the latter the allocation only guarantees that the budget will be satisfied on the average. In the latter, the ν_i are fixed in repeated sampling while in the former, allocation of the n_i depends on the observed n'. Of course, additional effort (*i.e.*, the Monte Carlo sampling) is needed to find the optimal allocation. In contrast to the optimal allocation, Rao's method permits selection of the second phase sampling fractions *prior* to observing the n_i' (see (2.7)). See Sections 3 and 4 for additional discussion.

3. COMPARISONS

3.1 Proportional *vs* Rao's Allocation

Assuming that $\nu_i^0 \leq 1$, $i = 1, \ldots, L$, and using (2.4) and (2.8), it can be shown that

$$V_P - V_R = \frac{1}{C^*}\left(S_W - \frac{\bar{S}_c}{\sqrt{c}}\right) \times \left\{2S_B\sqrt{c'c} + c\left(S_W + \frac{\bar{S}_c}{\sqrt{c}}\right)\right\}, \quad (3.1)$$

Table 1

Percent decrease in variance, R, for Rao's allocation compared to proportional allocation for a selection of textbook examples

Reference	L	S^2	S_W^2	G	C^*	R for $c' = 1$ and $c =$			
						1	2	5	25
Cochran (1977), p. 93	2	52,448	17,646	0.51	30	15.1	16.6	18.6	21.6
Hansen *et al.* (1953), p. 205	3	2,835,856	1,467,632	1.07	1,000	48.7	55.1	62.3	70.9
Sukhatme *et al.* (1984), p. 118	4	72,238	23,509	0.48	100	11.8	13.5	15.7	18.9
Cochran (1977), p. 111	7	619	343	1.25	1,000	11.2	11.7	12.4	13.7
Hansen *et al.* (1953), p. 202	8	47,393	45,595	25.36	1,000	10.5	11.0	11.5	12.0
Hansen *et al.* (1953), p. 202	11	47,393	44,974	18.59	1,000	22.9	24.1	25.4	26.7
Hansen *et al.* (1953), p. 235	11	2,039,184	820,722	0.67	1,000	21.3	24.8	29.1	35.1
Hansen *et al.* (1953), p. 202	12	47,393	40,252	5.64	1,000	16.7	18.3	19.8	21.6

Note: $R = 100(V_P - V_R)/V_P$ with V_P and V_R defined in (2.1) and (2.5) and C^* is the total budget. The cost function is defined in (1.3), and the variances (S^2, S_W^2, G) in (2.3).

where $\bar{S}_c = \sum_{i=1}^{L} W_i S_i \sqrt{c_i}$. Recalling that $c = \sum_{i=1}^{L} W_i c_i$ and using the Cauchy-Schwarz inequality, $S_w - \bar{S}_c/\sqrt{c} \geq 0$. Thus, as expected, $V_P - V_R \geq 0$. Defining $\bar{S} = \sum_{i=1}^{L} W_i S_i$ and $\bar{S}_\gamma = \sum_{i=1}^{L} W_i S_i \sqrt{\gamma_i}$ with $\gamma_i = c_i / \sum_{j=1}^{L} W_j c_j$, and using (3.1), it can be shown that

$$V_P - V_R = \frac{1}{C^*}\left\{2\sqrt{c'c}\left(\frac{S_B}{S_W + \bar{S}}\right) + c\right\} \times \left(S_W^2 - \bar{S}^2\right)$$

$$+ \frac{1}{C^*}\left\{2\sqrt{c'c}S_B + c(\bar{S} + \bar{S}_\gamma)\right\} \times \left(\bar{S} - \bar{S}_\gamma\right). \tag{3.2}$$

The first term in (3.2) is the reduction in variance if all sampling costs are equal while the second term in (3.2) is the reduction if all strata variances are equal. As expected, if $c_i = c$ and $S_i = S$, $V_P = V_R$.

We present in Table 1, the values of $R = 100(V_P - V_R)/V_P$ corresponding to a set of textbook examples with $c_i = c$. In parallel columns we give characteristics of the associated populations (L, S^2, S_W^2, $G = S_W^2/S_B^2$) and C^* together with the values of R corresponding to $c/c' = 1$, 2, 5 and 25. This set of examples represents a broad range of conditions where stratified sampling may be used. For a given value of c, the range of values of R indicates the wide range of gains that may be attained. It is clear from Table 1 that there may be substantial reductions in variance if one uses Rao's allocation, even when second phase strata sampling costs are equal and in situations when the stratification is not especially effective (note the large values of G for three examples). As c increases, R increases at a rate that is approximately constant (see Table 1).

3.2 Comparisons with Simple Random Sampling

For comparability with Rao and proportional allocations, assume a simple random sample of size n^* with expected cost $n^* \sum_{i=1}^{L} W_i c_i = n^* c$ (see (1.3)). Thus, for a fixed expected cost, C^*, $n^* = C^*/c$ and

$$\mathrm{Var}(\bar{y}_{n^*}) = S^2\left(\frac{c}{C^*} - \frac{1}{N}\right) \equiv V_S, \tag{3.3}$$

where $\bar{y}_{n^*}$ is the sample mean. Using (2.4) and (3.3),

$$V_S - V_P = \frac{1}{C^*}\left\{(c - c')S_B^2 - 2S_B S_W\sqrt{c'c}\right\}. \tag{3.4}$$

It can be shown that $V_S - V_P \geq 0$ if, and only if,

$$\frac{c}{c'} \geq \left(\sqrt{G} + \sqrt{1 + G}\right)^2 = LB_P, \tag{3.5}$$

where $G = S_W^2/S_B^2$. Using (2.8) and (3.3),

Table 2

Percent decrease in variance for proportional (R_P) and Rao's (R_R) allocation compared to simple random sampling for a selection of textbook examples

Reference	L	LB_P	LB_R	R_P			R_R		
				$c = 1$	5	25	$c = 1$	5	25
Cochran (1977), p. 93	2	3.8	2.6	−177.9	11.9	45.7	−136.0	28.3	57.4
Hansen *et al.* (1953), p. 205	3	6.1	1.1	−102.8	−6.1	26.4	−4.1	59.9	78.6
Sukhatme *et al.* (1984), p. 118	4	3.7	2.7	−132.8	12.8	46.6	−105.3	26.5	56.7
Hansen *et al.* (1953), p. 210	4	17.4	0.7	−127.7	−21.3	3.6	23.0	58.9	69.4
Cochran (1977), p. 111	7	6.8	4.5	−197.8	−9.8	23.3	−164.5	3.9	33.8
Hansen *et al.* (1953), p. 202	8	103.4	5.6	−38.2	−14.1	−4.0	−23.7	−0.9	8.5
Hansen *et al.* (1953), p. 202	11	76.4	1.7	−44.0	−15.6	−3.9	−11.0	13.7	23.8
Hansen *et al.* (1953), p. 235	11	4.5	2.2	−105.8	4.0	37.9	−62.0	32.0	59.7
Hansen *et al.* (1953), p. 202	12	24.5	4.0	−71.6	−19.9	0.2	−42.8	3.9	21.8

Note: Using (2.4), (2.8) and (3.3), $R_P = 100(V_S - V_P)/V_S$, $R_R = 100(V_S - V_R)/V_S$, and (LB_P, LB_R) are defined in (3.5) and (3.7). For these examples, $c' = 1$ and C^*, the total budget for each of the methods, is as in Table 1.

$$V_S - V_R = \frac{c}{C^*}\left\{S^2 - \left(\bar{S}_\gamma + S_B\sqrt{c'/c}\right)^2\right\}, \qquad (3.6)$$

where it is again assumed that $\nu_i^0 \leq 1$ for all i (see (2.7)). It is easily seen that $V_S - V_R \geq 0$ if, and only if,

$$\frac{c}{c'} \geq \frac{S_B^2}{(S - \bar{S}_\gamma)^2} = LB_R. \qquad (3.7)$$

In practice, one will estimate LB_P and LB_R in (3.5) and (3.7) and compare them with the cost ratio, c/c′, to decide if it will be beneficial to use double sampling with proportional or Rao's allocation rather than simple random sampling. In Table 2 we present the values of LB_P and LB_R for each of the examples in Table 1. We also include for $c = 1$, 5, and 25 the values of R, the per cent reduction in variance accruing from using a double sampling method rather than simple random sampling. As noted above, this set of examples represents a broad range of conditions where stratified sampling may be used. For a given value of c, the range of values of R_P and R_R indicates the wide range of gains (over simple random sampling) that may be obtained.

While $LB_P \geq LB_R$ is true in general, $LB_P \gg LB_R$ for many of the examples. The results point to potentially large gains for double sampling, especially using Rao's allocation, when c/c' is large. Conversely, if c/c' is relatively small, gains are modest and, in some cases, simple random sampling is preferred. This argues for careful estimation of LB_P, LB_R and c/c'.

3.3 Optimal *vs* Rao's Allocation

To compare the optimal allocation with that proposed by Rao, we have considered a wide range of values of the design parameters c', S^2 and $\{(c_i, S_i^2, W_i) : i = 1, \ldots, L\}$. We took $C^* = 1{,}000$ and considered $L = 2$ and 3. The values of the design parameters for $L = 2$ are listed in Table 3. Note that for these examples $G = S_W^2/S_B^2$ ranges from 0.01 to 10.00. We assume throughout that N is sufficiently large that S^2/N in (1.2) is negligible.

Table 3

Values of design parameters for the case of $L = 2$ strata

Parameter	Values
c'	0.125, 0.250, 0.500, 1.000
c_1	1, 4, 16
c_2	16
W_1	0.5, 0.6, 0.7, 0.8, 0.9
S^2	70.4, 128, 704
S_1^2	1, 4, 16, 64
S_2^2	64

Note: All 720 combinations of the above parameters were used. In addition, we also studied all arrangements of c', S^2, and S_1^2 as above together with

(a) $c_1 = 16$; $c_2 = 1, 4, 16$ and $W_1 = 0.5, 0.6, 0.7, 0.8, 0.9$,

(b) $W_1 = 0.1, 0.2, 0.3, 0.4$; $c_1 = 1, 4, 16$; $c_2 = 16$, and

(c) $W_1 = 0.1, 0.2, 0.3, 0.4$; $c_1 = 16$; $c_2 = 1, 4, 16$.

To ensure comparability of the two allocations we proceeded as indicated below for each specification of the design parameters.

1. Fix a single value of n'. We used both the value of n' identified as best using (a) Rao's method and (b) the optimal allocation.
2. From each of K Monte Carlo replications ($K = 200$ or 500) we obtain $n' = (n_1', \ldots, n_L')$ and then $n = (n_1, \ldots, n_L)$ using the optimal allocation and $\nu = (\nu_1, \ldots, \nu_L)$ from Rao's method. For the latter we use the algorithm which makes appropriate adjustments when the right side of (2.7) exceeds 1 for one or more strata.

   Since neither n from the optimal method nor n from Rao's method ($n_i = \nu_i n_i'$) are necessarily integers we round the n_i and adjust them so that for each sample the budget is satisfied (up to the approximation necessitated by having integer values of n' and n). We found that if these adjustments were not made there were anomalous results where the variance of $\hat{\bar{Y}}$ using Rao's allocation was less than the corresponding variance using the optimal allocation. This occurred when the total cost associated with Rao's procedure was larger than that for the optimal procedure.
3. To obtain estimates, $\bar{V}_{(c)O}$ and $\bar{V}_{(c)R}$, of the conditional variances, $E_{n'}\{\sum_{i=1}^{L} w_i^2 S_i^2 (1/n_i - 1/n_i')\}$, corresponding to the optimal and Rao's allocation, we used the average of $\sum_{i=1}^{L} w_i^2 S_i^2 (1/n_i - 1/n_i')$ over the K replications. The estimates of the unconditional variance, $\mathrm{Var}(\hat{\bar{Y}})$, in (1.2) are denoted by $\bar{V}_{(u)O}$ and $\bar{V}_{(u)R}$ where $\bar{V}_{(u)R} = \bar{V}_{(c)R} + (S^2/n')$.

The precision of these estimates was assessed by estimating the standard errors and coefficients of variation of $\bar{V}_{(u)R}$ and $\bar{V}_{(c)R}$. All of the standard errors were less than 0.0022. The coefficients of variation for $\bar{V}_{(u)R}$ and $\bar{V}_{(c)R}$ were below 0.0074 and 0.023, respectively. Thus, $\bar{V}_u$ and $\bar{V}_c$ provide precise estimates of the unconditional and conditional variances.

We present in Table 4 estimates of the per cent increase in the average unconditional variance for Rao's allocation, $I_u = 100\,(\bar{V}_{(u)R} - \bar{V}_{(u)O})/\bar{V}_{(u)O}$, for some of the design parameters listed in Table 3. We include results only for the value of n' identified as optimal by the optimal procedure. These results are typical of those seen for the other specifications in Table 3, those that we considered for the case $L = 3$, and those which use the value of n' identified as optimal by Rao's method. It is clear from Table 4 that improvements in precision are small, ranging from none to about 4%.

We obtained somewhat similar results for the per cent increase in the *conditional* variance for Rao's allocation, $I_c = 100(\bar{V}_{(c)R} - \bar{V}_{(c)O})/\bar{V}_{(c)O}$, where $\bar{V}_{(c)R}$ and $\bar{V}_{(c)O}$ are obtained by estimating $E\{\sum_{i=1}^{L} w_i^2 S_i^2 (1/n_i - 1/n_i')\}$ using, respectively, Rao's allocation and the optimal allocation. The results, based on 200 Monte Carlo replications and presented using boxplots in Treder (1989, Figures 2.8.2 and C.1 - C.3), can be summarized as follows. For *all* parameter specifications, the medians of the distributions of I_c are near 0. Most of the values of I_c are small: about 95% of the parameter specifications have distributions of I_c with third quartiles less than 10%. However, occasionally, there are large values of I_c: about 15% of the parameter specifications have the maximal value of I_c larger than 20%.

Table 4

Percent increase, I_u, in the average unconditional variance $\bar{V}_{(u)}$ for Rao's allocation compared to optimal allocation for a selection of design parameters with $S^2 = 70.4$, $S_2^2 = 64$, $c_2 = 16$ and $c' = 1$

S_1^2	G	c_1		
		16	4	1
	a. $(W_1, W_2) = (.9, .1)$			
64	10.000	0.0	0.4	1.4
16	0.419	0.1	0.1	0.1
4	0.166	0.1	0.1	0.4
1	0.116	0.1	0.3	0.8
	b. $(W_1, W_2) = (.7, .3)$			
64	10.000	0.0	0.7	3.6
16	0.760	0.0	0.2	0.7
4	0.455	0.1	0.3	1.4
1	0.394	0.0	0.7	0.9
	c. $(W_1, W_2) = (.5, .5)$			
64	10.000	0.0	1.0	4.1
16	1.316	0.0	0.4	0.9
4	0.934	0.0	0.6	1.8
1	0.858	0.0	0.2	0.0

Note: $I_u = 100(\bar{V}_{(u)R} - \bar{V}_{(u)O})/\bar{V}_{(u)O}$. See the note to Table 1 for definitions of the costs and variances.

These results can be explained, in part, by defining the *optimal* second phase sample size in stratum i by $n_i = \xi_i(n') \cdot n_i'$ where the dependence of n_i on the observed n' is emphasized by writing $\xi_i(n')$ and $0 < \xi_i(n') \leq 1$. Then, one may find the optimal allocation by choosing the $\xi_i(n')$ to minimize (for fixed n')

$$\frac{1}{n'} \sum_{i=1}^{L} \frac{w_i S_i^2}{\xi_i(n')}, \tag{3.8}$$

subject to $\sum_{i=1}^{L} c_i n_i' \cdot \xi_i(n') = C^* - c'n'$ (see 2.9).

By contrast, for the Rao allocation, for fixed n', one selects the ν_i to minimize

$$\frac{1}{n'}\sum_{i=1}^{L}\frac{W_iS_i^2}{\nu_i}, \tag{3.9}$$

subject to $n'\sum_{i=1}^{L} c_iW_i\nu_i = C^* - c'n'$, *i.e.* fixed expected cost.

Minimizing (3.8) rather than (3.9) will yield a smaller conditional and, thus, unconditional variance. However, when n' is large, the difference between (3.8) and (3.9) will be small.

3.4 Recommendations

Given reasonable estimates of the design parameters, one should first compare the cost ratio, c/c', with lower bounds, LB_P and LB_R, in (3.5) and (3.7) to see whether it is preferable to use double sampling rather than simple random sampling. These assessments must be done carefully because inappropriate use of double sampling may result in a *reduction* in precision. If there are good estimates of the design parameters, using Rao's allocation is preferable to proportional allocation.

Given the importance of adhering to a fixed budget we recommend the use of a modification of Rao's procedure:

> Use Rao's procedure to find the "optimal" value of n'. Then, given the n_i', use the optimal allocation procedure (*i.e.* minimize (2.9)) to find the n_i. This method guarantees that the budget will be satisfied for each sample, preserves most of the (small) gain in precision from using the optimal allocation and is easy to implement.

An alternative is to use Rao's procedure to find the "optimal" values of n' and the ν_i. Then implement an algorithm to round and modify the n_i ($n_i = \nu_i n_i'$) to ensure that the budget is satisfied for each sample. Unfortunately, it is difficult to develop the part of the algorithm needed to insure against cost overruns.

However, to avoid the large values of the proportional error in the *conditional* variance (*i.e.* I_c) that occur occasionally, one must use the *optimal* values of n' and the n_i.

Each of these methods requires knowledge of some design parameters. For Rao's allocation, the optimal ν_i require that the W_i and S_i^2 be specified. One can see from (2.9) that for the optimal allocation, the optimal n_i depend on the S_i^2 but not on the W_i. However, the optimal choice of n' requires that the W_i be specified. Alternatively, Srinath (1971) and Rao (1973a) have suggested a procedure which requires knowledge of the S_i^2 but not the W_i. Clearly, Rao's allocation requires the greatest knowledge of the design parameters and Srinath's procedure the least. Since the choice of n' is, typically, robust to misspecification of design parameters (see, *e.g.*, Sedransk 1965, Section 4.2.3), the optimal method may work well in the circumstances for which Srinath's method was designed.

4. SENSITIVITY OF ALLOCATIONS TO ESTIMATION OF DESIGN PARAMETERS

The preceding analysis assumes that the sample allocations are minimally affected by errors in the specification of the design parameters. In this section we investigate, in a simple case, the effect on $\mathrm{Var}(\hat{\bar{Y}})$ of the misspecification of an important design parameter. With proportional allocation, the choice of n' and n depends only on $G = S_W^2/S_B^2$, c' and c (see (2.3)). Estimating G by $\hat{G}$ and substituting the resulting values of n' and n from (2.3) in (2.1),

$$\frac{V_P(\hat{\bar{Y}})_{\hat{G}}}{S_W^2} = \frac{1}{C^*}\left(\frac{c' + \sqrt{c'c\hat{G}}}{G} + c + \sqrt{c'c/\hat{G}}\right) - \frac{1}{N}\left(1 + \frac{1}{G}\right), \tag{4.1}$$

where G is the correct value of S_W^2/S_B^2 and $\hat{G}$ is used only to determine n' and n.

Table 5

Per cent increase in unconditional variance, I, for proportional allocation when G is estimated by $\hat{G}$. $C^* = 1{,}000$, $c' = 1$ and $c_1 = c_2 = 16$

G	$\hat{G}$								
	1/100	1/36	1/16	1/4	1	4	16	36	100
1/100	0.0	6.0	19.8	69.3	174.9	389.8	817.3	817.3	817.3
1/36	6.2	0.0	4.4	33.1	103.8	251.4	547.7	547.7	547.7
1/16	21.9	3.9	0.0	12.1	57.1	156.2	357.9	357.9	357.9
1/4	71.7	30.7	12.6	0.0	11.8	51.2	138.5	138.5	138.5
1	128.9	67.2	37.3	7.3	0.0	7.5	35.9	35.9	35.9
4	179.1	101.6	63.3	22.3	5.7	0.0	5.4	5.4	5.4
16	210.2	123.4	80.3	33.5	12.9	2.3	0.0	0.0	0.0
36	220.4	130.7	86.0	37.4	15.7	4.0	0.0	0.0	0.0
100	225.9	134.6	89.1	39.5	17.2	4.9	0.0	0.0	0.0

Note: I is defined in (4.3), $G = S_W^2/S_B^2$ and the cost function is given by (1.3).

The optimal value of Var $(\hat{\bar{Y}})$ (*i.e.* when using G) in (2.4) can be expressed as

$$\frac{V_P(\hat{\bar{Y}})_G}{S_W^2} = \frac{1}{C^*}\left(\frac{c'}{G} + c + 2\sqrt{c'c/G}\right) - \frac{1}{N}\left(1 + \frac{1}{G}\right). \qquad (4.2)$$

If $(1/N)(1 + 1/G)$ is negligible, the per cent increase in variance due to estimating G, $I = 100\{V_P(\hat{\bar{Y}})_G - V_P(\hat{\bar{Y}})_G\} / V_P(\hat{\bar{Y}})_G$, is, from (4.1) and (4.2),

$$I = \frac{(1 - G) + \sqrt{c/c'}\{\sqrt{\hat{G}} - 2\sqrt{G} + (G/\hat{G})\}}{(1 + \sqrt{cG/c'})^2} \times 100. \qquad (4.3)$$

Note that (4.3) depends only on G, $\hat{G}$ and c/c'.

We present in Table 5 the values of I for $C^* = 1{,}000$, $c' = 1$, $c_1 = c_2 = 16$ and nine values of G and $\hat{G}$. The following conclusions are based on the results in Table 2.10.1 of Treder (1989) which includes additional values of G and $\hat{G}$. As long as $\hat{G}$ is within the interval $[G/4, 4G]$, using $\hat{G}$ to find (n', n) increases the variance by no more than 15%, typically less. If $\hat{G}$ is in the interval $[G/2, 2G]$, the increase in variance due to misspecification is about 4% or less. As G increases, the increase in variance associated with such intervals (*e.g.*, $[G/4, 4G]$) decreases. This happens because for large G, one has $n' = n$ and both $\hat{G}$ and G yield the same allocation. One manifestation of this result is the array of zeros in the lower right corner of Table 5. When G is small, that is when stratification is good, the sample allocation is more sensitive to incorrect specification of G than when G is large. These findings are little influenced by the values assigned to $c_1 = c_2$. In summary, for proportional allocation, fairly large misspecifications of the design parameter (G) lead to relatively small increases in variance.

REFERENCES

BOOTH, G., and SEDRANSK, J. (1969). Planning some two-factor comparative surveys. *Journal of the American Statistical Association*, 64, 560-573.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd Ed.). New York: John Wiley.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, (Vol. 1). New York: John Wiley.

HUGHES, E., and RAO, J.N.K. (1979). Some problems of optimal allocation in sample surveys involving inequality constraints. *Communications in Statistics - Theory and Methods A*, 8(15), 1551-1574.

RAO, J.N.K. (1973a). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

RAO, J.N.K. (1973b). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 669.

SEDRANSK, J. (1965). A double sampling scheme for analytical surveys. *Journal of the American Statistical Association*, 60, 985-1004.

SRINATH, K.P. (1971). Multiphase sampling in nonresponse problems. *Journal of the American Statistical Association*, 66, 583-586.

SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S., and ASOK, C. (1984). *Sampling Theory of Surveys with Applications*, (3rd Ed.). Ames, IA: Iowa State University Press.

TREDER, R.P. (1989). Some problems in double sampling for stratification. Unpublished Ph.D. dissertation, University of Washington.

Survey Methodology, June 1993
Vol. 19, No. 1, pp. 103–113
Statistics Canada

Stratified Telephone Survey Designs

ROBERT J. CASADY and JAMES M. LEPKOWSKI[1]

ABSTRACT

Two stage random digit dialing procedures as developed by Mitofsky and elaborated by Waksberg are widely used in telephone sampling of the U.S. household population. Current alternative approaches have, relative to this procedure, coverage and cost deficiencies. These deficiencies are addressed through telephone sample designs which use listed number information to improve the cost-efficiency of random digit dialing. The telephone number frame is divided into a stratum in which listed number information is available at the 100-bank level and one for which no such information is available. The efficiencies of various sampling schemes for this stratified design are compared to simple random digit dialing and the Mitofsky-Waksberg technique. Gains in efficiency are demonstrated for nearly all such designs. Simplifying assumptions about the values of population parameters in each stratum are shown to have little overall impact on the estimated efficiency.

KEY WORDS: Random digit dialing; Optimal allocation; Coverage; Relative efficiency.

1. THE CURRENT STATUS OF TELEPHONE SURVEY DESIGNS

The two stage random digit dialing design for sampling telephone households, first proposed by Mitofsky (1970) and more fully developed by Waksberg (1978), has been widely employed in telephone surveys. The Mitofsky-Waksberg technique capitalizes on a feature of the distribution of working residential numbers (hereafter referred to as WRNs) in the U.S.: specifically, the WRNs tend to be highly clustered within banks of consecutive telephone numbers. Currently, only about twenty percent of the possible telephone numbers within the known area code, three digit prefix combinations are WRNs for the United States as a whole. However, if a bank of 100 consecutive telephone numbers can be identified that has at least one known WRN then, on average, over 50 percent of the numbers in the bank will be WRNs. A technique which can identify 100-banks containing WRNs will greatly reduce the amount of screening necessary to identify telephone numbers assigned to households.

The two-stage Mitofsky-Waksberg technique starts by obtaining a list of area code, prefix combinations for the study area (available nationally from BellCore Research; see Lepkowski 1988). A frame of telephone numbers, hereafter referred to as the BellCore Research or BCR frame, is generated by appending all 10,000 four digit suffixes (*i.e.*, 0000 to 9999) to the area code-prefix combinations. The telephone numbers in the frame are grouped into banks of 100 numbers using the area code, three digit prefix, and the first two digits of the suffix to specify each bank. For example, the area code, prefix combination 313/764 will have 100 different 100-banks: 313/764-00, 313/764-01, . . . , 313/764-99. Next, a sample of 100-banks is selected and a single complete telephone number is generated for each selected bank by appending a two digit, randomly selected, number to the bank identifier. Each of these generated telephone numbers is dialed in the first sampling stage and the residential status of each number is determined and recorded. All 100-banks for which the randomly generated number is not a WRN are discarded. A second stage sample of WRNs is selected from all 100-banks for which the randomly generated number is a WRN. Typically an equal number of numbers, say k, are generated in each bank to start the second stage sampling process. When one of these second stage numbers is found to be non-residential, it is replaced by another randomly generated number from the same bank. This process is continued until k WRNs are identified in each bank. The result is a two stage sample based on selection of 100-banks with probabilities proportional to the number of residential numbers in the bank. This methodology has proven to be an excellent technique for identifying 100-banks with WRNs.

This technique has obvious advantages. The proportion of residential numbers within the 100-banks retained for second stage sampling is much higher than for the BCR frame in general, which results in a substantial improvement in efficiency over simple random digit dialing (RDD). It only requires that the complete set of area code, prefix combinations for the study area be known, and that the study staff have access to a random number generator for sampling telephone numbers. Finally, it also affords, in principle, complete coverage of all telephone households in the study area.

The Mitofsky-Waksberg technique also has several disadvantages. For example, not every 100-bank has the required k residential numbers so the second stage random number generation can use all 99 remaining numbers and

[1] Robert J. Casady, Bureau of Labor Statistics, U.S. Department of Labor and James M. Lepkowski, Survey Research Center, University of Michigan.

still not achieve the required k WRNs. In addition, determining the residential status of each generated number, especially at the first stage, can be difficult. For instance, in many rural areas recording equipment which notifies the caller that a number is not in service is not used. Calls to unassigned numbers are switched to a "ringing" machine. In these areas it is difficult to distinguish unassigned numbers from residential numbers where no one is at home during the study period. This difficulty is more noticeable at the end of a study period due to the need to replace non-residential numbers. Numbers generated at the end of the study period as replacements for non-residential numbers at the second stage of sampling have less time to be called. A small residual of unresolved numbers accumulates at the end of the study period, and final determination of residential status is impossible within study time constraints. Procedures for handling these unresolved numbers have been proposed (Burkheimer and Levinsohn 1988), but they often detract from the simplicity of the overall method.

Many of the difficulties with the Mitofsky-Waksberg technique can be reduced in importance through prescreening of telephone numbers and the use of computer assisted interviewing systems. However, these difficulties are not eliminated unless departures are made from the basic simplicity and/or underlying probability sampling principles of the method (see for example Potthoff 1987 and Brick and Waksberg 1991).

Alternatively, lists of published telephone numbers have been employed as a frame. These lists of published numbers are available for the entire country from commercial firms such as Donnelley Marketing Information Systems. A straightforward selection of telephone numbers from such lists provides a very high rate of WRNs (typically at least 85%) but unfortunately does not cover households with unpublished numbers. Comparisons of telephone households with and without published numbers (see, for example, Brunner and Brunner 1971) indicates that substantial bias may result.

Lists of published numbers can be employed in a manner to provide coverage of households with unlisted numbers as well. Groves and Lepkowski (1986) describe a dual frame approach in which a sample of listed numbers is combined with a random digit dialed sample through post-stratification estimation. If coverage of the population is less important, lists of published numbers can be used to identify 100-banks with at least one listed residential number, and sampling can be restricted to such banks. Survey Sampling Inc. (1986), and previously Stock (1962) and Sudman (1973) using reverse directories, selected samples of telephone numbers from this type of 100-bank. Clearly this approach does not provide complete coverage of unlisted telephone households, but it can greatly improve sampling efficiency. In fact these "truncated frame" methods have rates of residential numbers comparable or higher than the Mitofsky-Waksberg technique, and the troublesome replacement of non-residential numbers is not needed. Unfortunately, for many survey organizations, the coverage deficiency caused by truncating the frame is considered to be unacceptable.

The purpose of this paper is to examine stratified designs for the BCR frame as an alternative to frame truncation and Mitofsky-Waksberg designs. As an example of frame stratification, the BCR frame could be partitioned into two strata: a "high density" stratum consisting of residential numbers in 100-banks with one or more listed numbers and a "low density" stratum consisting of all the remaining numbers in the BCR frame. The "cut-off" point between high and low density strata is somewhat arbitrary; a cut-off of two or more listed numbers could reduce the chance that 100-banks are inadvertently included due to a keying error in a telephone number. Direct access to all listed numbers is not required for this stratification scheme. Counts of listed numbers, or any other indicator of the presence of listed telephone numbers in a 100-bank obtained from a reverse-directory (in metropolitan areas with such commercial services) or a commercial list for the entire country, would be sufficient. Preliminary work indicates that approximately 50% of the numbers in the high density stratum are WRNs while only about 2% of the numbers in the low density stratum are WRNs. The obvious cost difference of sampling from the two strata can be exploited through differential sample allocation. The telephone numbers in the low density stratum could be further stratified by careful examination of the characteristics of the 100-banks as determined by other data available from the BCR frame and/or the Donnelley list which may result in even further sampling efficiency.

The next section examines the question of the appropriate allocation of sample between the strata when simple random sampling is utilized within each stratum. A key feature of the stratified telephone sample approach is that it permits alternative approaches to sample selection within in the different strata. Several alternatives are presented and discussed in Section 3. Section 4 presents a study of the impact of "non-optimal" sample allocation on design efficiency. The paper concludes with a general discussion contrasting the Mitofsky-Waksberg procedure and stratified designs.

2. THE ALLOCATION PROBLEM FOR STRATIFIED TELEPHONE DESIGNS

2.1 Background

We assume that the basic sampling frame is the collection of all telephone numbers generated by appending four digit suffixes to the BCR list of area-prefix codes. Further, we assume that each household in the target population

is "linked" to one and only one telephone number in the basic sampling frame (to avoid complications of unequal probability of selection).

We also assume that we have access (possibly only indirect) to a directory based, machine readable list of telephone numbers. It should be noted that because many households choose not to list their telephone numbers in a directory, any such directory based frame will not contain all of the WRNs. Directory based lists are also by nature out of date so they will omit some numbers that are currently published WRNs while including others that are no longer WRNs.

From a survey design point of view these two frames tend to be radically different. The BCR frame includes all WRNs so it provides complete "coverage" of the households in the target population, but only about 20 percent of the telephone numbers included in the BCR frame are actually WRNs. Thus, the "hit rate" (and hence sampling efficiency) will be quite low for a simple RDD sample design utilizing the BCR frame. In contrast, a typical directory/list frame covers only about 70 percent of the target households, but the hit rate is 85 to 90 percent. In general the sampling efficiency for a simple RDD design using a directory/list frame is far better than can be attained for the BCR frame using the Mitofsky-Waksberg technique. Unfortunately, the low coverage rates associated with directory based frames preclude their use in many cases.

The basic idea of the proposed stratification approach is to utilize information from the directory based frame to partition the BCR frame into two or more strata with disparate hit rates and then allocate the sample to the strata so as to minimize cost (variance) for a specified variance (cost). Hereafter the stratum with the lowest hit rate will be referred to as the residual stratum. The truncated designs discussed earlier can be included in this general type of design if we allow the allocation of no sample to the residual stratum, and use mean squared error in place of variance.

2.2 Basic Notation

Assume that the BCR frame of telephone numbers has been partitioned into H strata based on a 100-bank attribute which can be determined from either the BCR or the directory based frame of telephone numbers. The choice of 100-banks is somewhat arbitrary; banks of from 10 to 500 consecutive numbers could be considered. For the ith stratum let

P_i = proportion of the frame included in the stratum,

h_i = proportion of the telephone numbers in the stratum that are WRNs (*i.e.* the hit rate),

w_i = average proportion of WRNs in the non-empty 100-banks (*i.e.* the average hit rate for non-empty banks),

z_i = proportion of the target population included in the stratum, and

t_i = proportion of 100-banks in the stratum that contain no WRNs.

The average hit rate for the frame is given by $\bar{h} = \sum_{i=1}^{H} h_i P_i$ and the proportion of empty 100-banks in the frame is given by $\bar{t} = \sum_{i=1}^{H} t_i P_i$.

In general only the P_i's will be known with certainty. Data from a joint research project involving the Bureau of Labor Statistics and the University of Michigan were used to provide approximate values for the parameters h_i and w_i for the two strata in the example. Values for the remaining parameters were calculated using the algebraic relationships $t_i = 1 - (h_i/w_i)$ and $z_i = h_i P_i/\bar{h}$. The approximations for all of the frame parameters for the two stratum design are given in Table 1 below; note that for the BCR frame and $\bar{h} \cong .211$ and $\bar{t} \cong .605$. The value of $\bar{h}$ is in close agreement with that given in Waksberg (1978) but the value of $\bar{t}$ is somewhat smaller than the .65 provided by Groves (1977). At this time it is impossible to determine which value of $\bar{t}$ is more accurate; in fact, the value may have changed since 1977. More recently, Tucker, Casady and Lepkowski (1992) estimated the value of $\bar{t}$ to be .616 for 10-banks which supports the lower estimate $\bar{t}$ of for 100-banks.

Table 1

Approximate values of the frame parameters for a two stratum design based on the BCR frame and Donnelley directory list. Stratum 1 consists of all telephone numbers in 100-banks with at least one telephone number on the Donnelley list frame; stratum 2 contains all remaining numbers

Stratum	Proportion of Frame (P_i)	Proportion of Population (z_i)	Hit Rate (h_i)	Proportion of Empty 100-Banks (t_i)	Hit Rate Within Non-empty Banks (w_i)
1	.3804	.9402	.5210	.0300	.5371
2	.6196	.0598	.0204	.9584	.4900

2.3 The Basic Estimation Problem, Sample Designs and Estimators

We assume the telephone numbers in the ith stratum are labeled 1 through M_i. Let

$$d_{ij} = \begin{cases} 1 & \text{if the } j\text{th telephone number in the } i\text{th stratum is a WRN,} \\ 0 & \text{otherwise.} \end{cases}$$

The variable of interest is the household characteristic Y, and y represents the value of Y for a particular household. The population parameter to be estimated is the population mean $\mu = Y./N.$, where $N. = \sum_{i=1}^{H} \sum_{j=1}^{M_i} d_{ij} = \sum_{i=1}^{H} N_i$ and $Y. = \sum_{i=1}^{H} \sum_{j=1}^{M_i} d_{ij}\, y_{ij}$. The term N_i denotes the number of WRNs in the ith stratum and $N.$ denotes the number of WRNs in the population.

Consider two sample designs: (1) simple random sampling without replacement (*i.e.* simple RDD) from the telephone numbers in the BCR frame, denoted as design D_0 and (2) stratified simple random sampling from the BCR frame (*i.e.* independent simple RDD samples are selected from each stratum), denoted as design D_1. Under design D_0 the standard ratio estimator for μ is given $\bar{Y}_0 = \hat{Y}_0/\hat{N}_0$ where $\hat{Y}_0$ and $\hat{N}_0$ are the usual inflation estimators for $Y.$ and $N.$ respectively. The estimator $\bar{Y}_0$ is asymptotically unbiased for μ and its variance is given by $\text{var}(\bar{Y}_0) \cong \sigma^2/m\bar{h}$ where m is the sample size of telephone numbers and σ^2 is the population variance of the y's. For the design D_1 the standard ratio estimator of μ is given by $\bar{Y}_1 = \hat{Y}_1/\hat{N}_1$ where $\hat{Y}_1$ and $\hat{N}_1$ are the standard inflation estimators for $Y.$ and $N.$ under stratified sampling. The estimator $\bar{Y}_1$ is also asymptotically unbiased for μ and

$$\text{var}(\bar{Y}_1) \cong \sum_{i=1}^{H} \frac{z_i^2 \sigma_i^2 (1 + (1 - h_i)\lambda_i)}{m_i h_i}, \tag{2.1}$$

where $\lambda_i = (\mu_i - \mu)^2/\sigma_i^2$ and m_i, μ_i, and σ_i^2 are the stratum sample sizes, means, and variances, respectively.

2.4 The Cost Model

There are costs associated both with determining the value of the indicator variable d and the value of the characteristic of interest Y. The cost function for determining the indicator variable is denoted by $C_1(\cdot)$, with

$$C_1(d) = \begin{cases} c_1 & \text{if } d = 1 \\ c_0 & \text{if } d = 0. \end{cases}$$

This model allows for the possibility that the cost of determining that a telephone number is not a WRN may be different than determining that a telephone number is a WRN. In fact the cost of determining the status of telephone numbers that are WRNs is usually less. The cost of determining the value of the characteristic Y includes only the *additional cost* of determining the value of y after the value of d has been determined. Letting $C_2(\cdot,\cdot)$ represent this additional cost, with

$$C_2(d,y) = \begin{cases} 0 & \text{if } d = 0 \\ c_2 & \text{if } d = 1. \end{cases}$$

The sum $c_1 + c_2$ represents the cost of a "productive" sample selection and c_0 represents the cost of an "unproductive" selection, then, following Waksberg (1978), $\gamma = (c_1 + c_2)/c_0$ represents the ratio of the cost of a productive selection to an unproductive selection.

The total cost for sample selection and the determination of the values of Y is a random variable for both design D_0 and D_1. Letting $C(D_0)$ and $C(D_1)$ represent the total cost of conducting a survey under the two respective designs it is straightforward to show that

$$E[C(D_0)] = mc_0(1 + (\gamma - 1)\bar{h}) \tag{2.2}$$

and

$$E[C(D_1)] = c_0 \sum_{i=1}^{H} m_i(1 + (\gamma - 1)h_i). \tag{2.3}$$

2.5 Optimal Allocation for $\bar{Y}_1$

The stratum sample allocation that minimizes $\text{var}(\bar{Y}_1)$ for a fixed expected total cost C^* (or that minimizes $E[C(D_1)]$ for a fixed variance V^*) is specified up to a proportionality constant by

$$m_i \propto \frac{z_i \sigma_i}{\sqrt{h_i}} \left(\frac{1 + (1 - h_i)\lambda_i}{1 + (\gamma - 1)h_i} \right)^{1/2}, \tag{2.4}$$

where the proportionality constant is determined by substitution into the expected cost equation (or the variance equation, as appropriate). The proportional reduction in variance, relative to RDD sampling, under optimal allocation for fixed cost C^* (or the proportional reduction in cost under optimal allocation for fixed variance V^*) is given by

$$R(\bar{Y}_1, \bar{Y}_0) \cong 1 - \frac{\left[\sum_{i=1}^{H} \frac{z_i \sigma_i}{\sqrt{h_i}} \left[(1 + (1 - h_i)\lambda_i)(1 + (\gamma - 1)h_i) \right]^{1/2} \right]^2}{\bar{h}^{-1}\sigma^2(1 + (\gamma - 1)\bar{h})}. \tag{2.5}$$

2.6 Practical Problems Associated With Optimal Allocation

The problem of specifying the values for the parameters in the allocation equations is generic to optimal allocation schemes. For our particular case there are three basic types of parameters: frame related (z_i and h_i), cost related (γ and c_0) and those specific to the variable of interest (λ_i and σ_i^2). Currently, we have a fairly good working knowledge of the frame related parameters for the two stratum example and certain other specific stratification schemes. In Section 5, we will discuss several active research projects which should further expand our knowledge in this area.

It is clear that $\gamma \geq 1$, but the actual value can vary widely. For example, in the case of a multipurpose survey information is collected for several variables, so the costs of determining the status of a telephone number, c_0 and c_1, are in effect amortized over the variables of interest, and γ will probably be considerably larger than unity. On the other hand, if the survey is intended to collect information on only a single variable then the value of γ is probably not much larger than two or three. Waksberg (1978) considers values of γ between 2 and 20.

Potentially the variable specific parameters pose the most serious problem. Usually our knowledge regarding the values of these parameters is limited and, in the case of multipurpose surveys, we must decide which variable(s) to use for the purposes of allocation. Fortunately, in many practical applications, two factors combine to somewhat lessen this problem. First, the allocation tends to be relatively "flat" in a neighborhood of the optimum allocation so that the reduction in variance is relatively robust with respect to allocation. Secondly, in most cases the variables of interest will not be highly related to variables of the type we are using for stratification. Therefore, with caution, we assume that $\lambda_i = 0$ and $\sigma_i^2 = \sigma^2$ for $i = 1, 2, \ldots, H$. Optimal allocation is achieved by

$$m_i \propto \frac{z_i}{\sqrt{h_i}}(1 + (\gamma - 1)h_i)^{-1/2} \tag{2.6}$$

and the proportional reduction in variance is

$$R(\bar{Y}_1, \bar{Y}_0) \cong 1 - \bar{h}\frac{\left[\sum_{i=1}^{H} z_i\left(\frac{1 + (\gamma - 1)h_i}{h_i}\right)^{1/2}\right]^2}{(1 + (\gamma - 1)\bar{h})}. \tag{2.7}$$

In the case of the two stratum example, the allocation specified by (2.6) implies that allocation relative to the residual stratum (*i.e.* m_1/m_2) is 2.54 when $\gamma = 2$ and 1.42 when $\gamma = 10$. In the first case the projected proportional reduction in variance is $R = .283$ and in the second $R = .077$. In fact, it follows from (2.7) that as the relative cost of determining the value of the variable of interest increases, the relative benefit of optimal allocation decreases.

The Mitofsky-Waksberg sample design, denoted by D_3, employs two stages of sample selection (*i.e.* non-empty 100-banks are selected in the first stage and WRNs are selected in the second stage). Following Waksberg (1978), we let $(k + 1)$ be the total number of WRNs selected from each sample 100-bank. The Mitofsky-Waksberg estimator, denoted by $\bar{Y}_3$, is unbiased for μ, and its variance is minimized when

$$k + 1 = \max\left\{1, \left(\frac{(1 - \rho)\bar{t}}{(1 + (\gamma - 1)\bar{h} - \bar{t})\rho}\right)^{1/2}\right\}, \tag{2.8}$$

where ρ is intra-bank correlation. Under this "optimal" within 100-bank sample allocation the reduction in variance, relative to simple RDD, for the estimator $\bar{Y}_3$ is given by

$$R(\bar{Y}_3, \bar{Y}_0) \cong 1 - \frac{\left[(1 + (\gamma - 1)\bar{h} - \bar{t})^{1/2}(1 - \rho)^{1/2} + (\rho\bar{t})^{1/2}\right]^2}{1 + (\gamma - 1)\bar{h}}. \tag{2.9}$$

At the national level Groves (1977) reports that $\rho \cong .05$ for economic or social statistics. Using this value of ρ, together with the values of $\bar{h}$ and $\bar{t}$ from the two stratum example, the projected proportional reduction in variance for the Mitofsky-Waksberg procedure is $R = .281$ when $\gamma = 2$ and $R = .060$ when $\gamma = 10$.

The two methodologies appear to produce essentially identical variance reduction for both values of the cost ratio. However, too much should not be read into this simple comparison as the projected reduction for each of the procedures is based on simplifying assumptions that will not be strictly true for any application. The only inference intended is that the two procedures appear to highly competitive under a general set of circumstances typically encountered in application.

3. ALTERNATIVE SAMPLE DESIGNS

3.1 Truncated Designs

The designs presented in the previous section produce unbiased estimates of the population mean. Incorrect assumptions regarding the various frame, cost, and population parameters only affect the efficiency of the estimators, not their expectations. Unfortunately an extremely high price is paid for the assurance of unbiasedness because sampling from the residual stratum provides information on only a small proportion of the population and at a relatively high cost. For example, suppose we are willing to settle for an estimate of the population mean exclusive of those households linked to telephone numbers in the residual stratum (*i.e.* we "truncate" the original frame by eliminating the residual stratum and select a stratified RDD sample from the remaining telephone numbers). For the two stratum example the "truncated frame" would consist only of those telephone numbers in the first stratum. The hit rate for the sample from the truncated frame would be .521, in contrast to a hit rate of .211 for the entire frame. However, only about 94% of the target population would remain in scope.

In what follows we assume that the truncated frame is simply the original BCR frame less the residual stratum which (without loss of generality) we assume to be stratum H. Accordingly, for the truncated frame $\bar{h}^* = (\bar{h} - P_H h_H)/(1 - P_H)$ is the hit rate, $\bar{t}^* = (\bar{t} - P_H t_H)/(1 - P_H)$ is

the proportion of empty 100-banks and $\mu^* = (\mu - z_H\mu_H)/(1 - z_H)$ is the population mean. Let design D_4 be stratified simple random sampling from the truncated frame, and $\bar{Y}_4$ the standard ratio estimator of the population mean. The estimator $\bar{Y}_4$ is asymptotically unbiased for μ^*, and, in general, it is biased for μ. The (asymptotic) bias is given by

$$B(\bar{Y}_4) = \mu^* - \mu = \frac{z_H(\mu - \mu_H)}{(1 - z_H)}. \qquad (3.1)$$

In most practical circumstances the bias tends to zero monotonically as the proportion of the target population in the residual stratum becomes small, although, as indicated by (3.1), this is not necessarily the case. In any event, since the value of $\mu - \mu_H$ is never known, an upper limit on the proportion of the population in the residual stratum is usually the key specification to be determined when considering the use of a truncated frame. For the two stratum example approximately 6% of the target population is excluded from the sampling frame and, in almost all cases, this would not be tolerable for Federal agencies.

The equations for cost, variance, allocation, and proportional reduction in variance (or cost) are essentially the same as those presented in Section 2. In fact the only modifications required for equation (2.1) and equations (2.3) through (2.7) are to replace μ by μ^* and, for $i = 1, 2, \ldots, H - 1$, replace z_i with $z_i^* = z_i/(1 - z_H)$, and replace λ_i with $\lambda_i^* = (\mu_i - \mu^*)^2/\sigma_i^2$. Obviously all sums are only over the remaining $H - 1$ strata. For the special case where only one stratum remains after truncation the proportional reduction in variance (cost) reduces to

$$R(\bar{Y}_4, \bar{Y}_0) = 1 - \frac{\bar{h}(1 + \bar{h}^*(\gamma - 1))}{\bar{h}^*(1 + \bar{h}(\gamma - 1))}. \qquad (3.2)$$

Thus for the two stratum design, the proportional reduction in variance (cost) is approximately $R = .492$ when $\gamma = 2$ and $R = .206$ when $\gamma = 10$. In both cases the reduction is substantially greater than achieved by the two methods in the previous section. However, nearly 6% of the population is not covered by the frame.

In an attempt to retain the relative efficiency of truncation while reducing the magnitude of the coverage problem, BLS and the University of Michigan are investigating several alternative stratification plans in an effort to reduce the proportion of the population in the residual stratum. One promising approach calls for the partition of the residual stratum into two or more residual strata. For example, the partitioning could create a residual stratum 3 consisting of telephone numbers in 100-banks thought to be primarily assigned to commercial establishments or not yet activated for either residential or commercial use. Residual stratum 2 will now contain all other telephone numbers in the residual stratum from the two stratum design D_2. Estimated frame parameters for the resulting three stratum design are given in Table 2.

Table 2

Estimated frame parameters for a proposed three stratum design based on the BCR frame and the Donnelley list frame

Stratum	Proportion of Frame (P_i)	Proportion of Population (z_i)	Hit Rate (h_i)	Proportion of Empty 100-Banks (t_i)	Hit Rate Within Non-empty Banks (w_i)
1	.3804	.9402	.5210	.0300	.5371
2	.2000	.0399	.0420	.9143	.4900
3	.4196	.0199	.0100	.9796	.4900

These data were used to compute the projected proportional reduction in variance for both a three stratum design and a truncated three stratum design in which Stratum 3 is excluded. These results, together with a summary of the results for the two stratum designs and the Mitofsky-Waksberg design, are presented in Table 3 below. (Although not discussed in the text, Table 3 also includes the projected reduction in variance for a cost ratio of 20.)

Table 3

Projected proportional reduction in variance (or cost) relative to simple RDD sampling for five alternative telephone sample designs

Sample Design	Proportional Reduction in Variance or Cost			Proportion of Frame not in Scope
	$\gamma = 2$	$\gamma = 10$	$\gamma = 20$	
Two Stratum	.2829	.0766	.0320	.0000
Two Stratum (Truncated)	.4917	.2055	.1189	.0598
Mitofsky-Waksberg	.2811	.0597	.0135	.0000
Three Stratum	.3001	.0866	.0389	.0000
Three Stratum (Truncated)	.4095	.1574	.0879	.0199

The proposed partitioning strategy successfully reduces the percent of the population out of scope from nearly 6% to approximately 2%. The projected proportional reduction in variance for the truncated three stratum design is approximately $R = .410$ when $\gamma = 2$ and $R = .157$ when $\gamma = 10$. From an efficiency point of view, it occupies the middle ground between the highly efficient truncated two stratum design and unbiased designs.

Of course the issue to be faced when considering such a design is the coverage problem. The design is already subject to non-coverage of the non-telephone household population. Truncating the frame may add to any non-coverage bias already due to this source. For any particular application the risk inherent in sampling from a frame that does not include all of the target population must be weighed against the potential gain in efficiency. As expected, the standard three stratum design is slightly more efficient than the two stratum design. However, the increase in efficiency is so small that it is doubtful that the added cost of partitioning the BCR frame into an additional stratum is justified except for the purpose of truncation.

3.2 Designs Using Optimal Allocation and the Mitofsky-Waksberg Procedure

The final design to be considered is based on the stratified BCR frame. Depending on the proportion of empty 100-banks in the stratum, we use simple RDD sampling in some strata and Mitofsky-Waksberg sampling in others. The motivation for this type of design is based on the following two considerations:

(a) Mitofsky-Waksberg sampling tends to be "administratively complex", and if the gain in efficiency is small, simple RDD is preferred.

(b) It follows from (2.9), applied at the stratum level, that if the proportion of empty banks in a stratum is "small" then Mitofsky-Waksberg sampling offers little, if any, increase in efficiency.

Thus, we propose to utilize simple RDD sampling in strata with a "small" proportion of empty hundred banks and Mitofsky-Waksberg sampling in the remaining strata. The criterion for determining the type of sampling to be utilized is based on equation (2.8) applied at the stratum level. Specifically, if the "optimal" total number of WRNs, as determined by equation (2.8), to be selected from sample 100-banks in a particular stratum is equal to one, then the stratum is designated a simple RDD stratum; otherwise it is designated a Mitofsky-Waksberg stratum. In terms of the proportion of empty hundred banks, the ith stratum will be an RDD stratum if

$$t_i \leq \frac{2.25\rho(1 + h_i(\gamma - 1))}{(1 + 1.25\rho)} \tag{3.3}$$

and a Mitofsky-Waksberg stratum otherwise. For the two stratum example, the first stratum is a RDD stratum, and the second is a Mitofsky-Waksberg stratum for γ equal either 2 or 10.

Formally the proposed sample design is as follows. The BCR frame has been partitioned into H strata and, according to the criteria given in (3.3), simple RDD sampling is specified for the first H_1 strata and Mitofsky-Waksberg sampling is specified for the remaining strata.

Let:

m_i = the number of telephone numbers selected from the ith RDD stratum,

m_i' = the number of WRNs in the sample from the ith RDD stratum,

$\tilde{m}_i$ = the number of 100-banks selected from the ith Mitofsky-Waksberg stratum,

$\tilde{m}_i'$ = the number of retained 100-banks in the ith Mitofsky-Waksberg stratum,

k_i = number of additional WRNs selected from each retained 100-bank, and

$y_{i\cdot}$ = aggregate of y values for the sample WRNs from the ith stratum.

The combined ratio estimator $\bar{Y}_5 = \hat{Y}_5/\hat{N}_5$, where $\hat{Y}_5 = \sum_{i=1}^{H_1} M_i/m_i\, y_{i\cdot} + \sum_{i=H_1+1}^{H} M_i/\tilde{m}_i\, (y_{i\cdot}/k_i + 1)$ and $\hat{N}_5 = \sum_{i=1}^{H_1} M_i/m_i\, m_i' + \sum_{i=H_1+1}^{H} M_i/\tilde{m}_i\, \tilde{m}_i'$, is utilized to estimate the population mean μ and the values of m_i, $\tilde{m}_i$ and k_i are to be chosen to minimize $\text{var}(\bar{Y}_5)$ or the expected cost as specified.

The estimator $\bar{Y}_5$ is asymptotically unbiased for μ and it is straightforward to show that

$$\text{var}(\bar{Y}_5) \cong \sum_{i=1}^{H_1} \frac{z_i^2\sigma_i^2}{m_i h_i}(1 + (1 - h_i)\lambda_i) + \sum_{i=H_1+1}^{H} \frac{z_i^2\sigma_i^2}{\tilde{m}_i h_i}\left[1 + (1 - h_i)\lambda_i - k_i(1 - \rho)(k_i + 1)^{-1}\right] \tag{3.4}$$

and

$$E[C(D_5)] = c_0\left\{\sum_{i=1}^{H_1} m_i\left[1 + h_i(\gamma - 1)\right] + \sum_{i=H_1+1}^{H} \tilde{m}_i\left[1 + k_i(1 - t_i) + h_i(k_i + 1)(\gamma - 1)\right]\right\}. \tag{3.5}$$

The optimal values of m_i and $\tilde{m}_i$, specified up to a proportionality constant, are given by

$$m_i \propto z_i\sigma_i\left(\frac{1 + (1 - h_i)\lambda_i}{h_i(1 + h_i(\gamma - 1))}\right)^{1/2}, \tag{3.6}$$

for $i = 1, \ldots, H_1$ and

$$\tilde{m}_i \propto z_i\sigma_i\left(\frac{\lambda_i(1 - h_i) + \rho}{h_i t_i}\right)^{1/2}, \tag{3.7}$$

for $i = H_1 + 1, \ldots, H$. The optimal value of $(k_i + 1)$, for $i = H_1 + 1, \ldots, H$, is given by

$$k_i + 1 = \max \left\{1, \left(\frac{t_i(1-\rho)}{(1 + h_i(\gamma - 1) - t_i)(\lambda_i(1 - h_i) + \rho)}\right)^{1/2}\right\}. \tag{3.8}$$

The proportionality constant for (3.6) and (3.7) is found by substitution into the expected cost equation or the variance equation as appropriate.

Under optimal allocation the reduction in variance (or cost) relative to simple RDD, is given by

$$R(\bar{Y}_S, \bar{Y}_0) = 1 - \frac{\bar{h}\Phi^2}{\sigma^2(1 + (\gamma - 1)\bar{h})}, \tag{3.9}$$

where

$$\Phi = \sum_{i=1}^{H_1} \frac{z_i\sigma_i}{h_i^{1/2}}(1 + (1 - h_i)\lambda_i)^{1/2}(1 + (\gamma - 1)h_i)^{1/2} + \sum_{i=H_1+1}^{H} \frac{z_i\sigma_i}{h_i^{1/2}}\Big[(\rho + (1 - h_i)\lambda_i)^{1/2}t_i^{1/2} + (1 - t_i + (\gamma - 1)h_i)^{1/2}(1 - \rho)^{1/2}\Big]. \tag{3.10}$$

Under the simplifying assumptions $\lambda_i = 0$ and $\sigma_i^2 = \sigma^2$ for $i = 1, 2, \ldots, H$,

$$\Phi = \sigma\left[\sum_{i=1}^{H_1} \frac{z_i}{h_i^{1/2}}(1 + (\gamma - 1)h_i)^{1/2}\right] + \sigma\left[\sum_{i=H_1+1}^{H} \frac{z_i}{h_i^{1/2}}\big((\rho t_i)^{1/2} + ((1 - t_i + (\gamma - 1)h_i)(1 - \rho))^{1/2}\big)\right]. \tag{3.11}$$

When applied to the two stratum frame, this combined sampling strategy yields a proportional reduction in variance of approximately $R = .440$ for $\gamma = 2$ and $R = .157$ for $\gamma = 10$. For both of the cost ratios, the reduction in variance is considerably larger than achieved by any of the unbiased procedures considered previously. In fact, the variance reduction is essentially equivalent to that attained by the three stratum truncated design (which is subject to a bias of unknown magnitude). Thus, on first consideration, this combined sampling strategy appears to be superior to all of the other methods.

Unfortunately there are practical problems which may preclude the use of this sampling design in certain situations. For example, the hit rate in the Mitofsky-Waksberg stratum is very low (only .02) so the number of first stage sample 100-banks must be fairly large in order that the expected number of retained 100-banks is not too small. On the other hand, the *relative* number of first stage sample units allocated to the RDD stratum is considerably larger than allocated to the Mitofsky-Waksberg stratum, therefore a large overall sample size is required (see Table 4). Also, from Table 4, the number of WRNs required from each of the retained 100-banks is relatively large and may actually exceed the number of WRNs in some banks. Clearly both of these problems are more acute for $\gamma = 2$ than for $\gamma = 10$. Therefore, the use of this design is restricted to situations where resources can support a "large" sample, and the cost ratio is moderate to large.

Table 4

First stage allocation ratios and second stage sample sizes for the combined RDD/Mitofsky-Waksberg sample design applied to the two stratum BCR frame

Stratum	$\gamma = 2$		$\gamma = 10$	
	$m_1/\bar{m}_2$	Sample Size Second Stage	$m_1/\bar{m}_2$	Sample Size Second Stage
1	28.17	N.A.	14.56	N.A.
2	N.A.	17.00	N.A.	9.00

4. SAMPLE ALLOCATION AND DESIGN EFFICIENCY

In Section 2.6 the problem of specifying the parameters required to optimally allocate the sample to the various strata was considered. It was noted that the variable specific parameters (*i.e.* the λ_i and σ_i^2) tend to pose the most serious problem since we usually have little information regarding their values. For most cases the variables of analytic interest will not be very highly related to the variables used for stratification. Thus it is reasonable to assume that $\lambda_i = 0$ and $\sigma_i^2 = \sigma^2$ for $i = 1, 2, \ldots, H$. Under these assumptions the optimal allocation is given by (2.6) and the proportional reduction in variance is given by (2.7).

It is obvious that for any particular application these assumptions will never be strictly true, so when we allocate according to (2.6) the actual proportional reduction in variance will not be that given exactly by (2.7). Furthermore, allocating according to (2.6) will not provide the maximum reduction in variance which is achieved under the optimal allocation specified by (2.4). Assuming that we plan to allocate according to (2.6) two questions need to be addressed: (1) does (2.7) give a reasonable approximation to the actual reduction in variance, and (2) is the actual

reduction in variance reasonably close to the maximum possible reduction in variance? A single simple answer is not possible for either question because the outcome depends on exactly how and to what extent the assumptions failed. In the following we address these question for the two stratum design under three specific cases of model failure which are typical of situations encountered in the "real world". In all three cases the results indicate strongly affirmative answers for both questions.

In the first case we assume that $\sigma_1^2 = \sigma_2^2 \equiv W^2$ but $\lambda_1 \neq \lambda_2$. The projected, the actual, and the maximum reduction in variance were computed for selected values of $\beta = |\sqrt{\lambda_1} - \sqrt{\lambda_2}| = |\mu_1 - \mu_2|/W$ between 0.00 to 0.50 and the results are presented in Table 5 below. Based on our previous discussion regarding the weak relationship between the analytic and stratification variables it would seem highly unlikely that β will ever be larger than 0.50. The results in Table 5 indicate that for both cost ratios and for all selected values of β the actual reduction in variance achieved by allocation under the simplifying assumptions is essentially equivalent to that which would be attained under "optimal" allocation. For both cost ratios the projected reduction in variance is always larger than the reduction actually attained and the difference increases as β becomes larger. However, it should be noted that for $\beta \leq .35$ the percentage difference between the projected reduction and the actual reduction is less than 10% when $\gamma = 10$, and less than 4% when $\gamma = 2$.

Table 5

The projected, the actual, and the maximum proportional reduction in variance for cost ratios of 2 and 10 and values of β between 0.00 and 0.50

β	$\gamma = 2$			$\gamma = 10$		
	Projected Reduction	Actual Reduction	Maximum Reduction	Projected Reduction	Actual Reduction	Maximum Reduction
0.00	.2829	.2829	.2829	.0766	.0766	.0766
0.10	.2829	.2820	.2820	.0766	.0761	.0761
0.20	.2829	.2793	.2794	.0766	.0745	.0746
0.30	.2829	.2748	.2750	.0766	.0720	.0721
0.40	.2829	.2686	.2692	.0766	.0684	.0689
0.50	.2829	.2607	.2619	.0766	.0639	.0649

The second general case considered assumes that the analytic variable is Bernoulli, where p_1 and p_2 represent the proportion of the population with the attribute of interest in stratum 1 and stratum 2, respectively. The projected, the actual, and the maximum proportional reduction in variance were computed for two specific cases of assumption failure, namely $p_2 = .90p_1$ and $p_2 = 1.10p_1$; p_1 was allowed to vary from .05 to .50 and cost ratios of 2 and 10 were considered.

As discussed before it is probably reasonable to assume that p_2 will be within 10% of p_1 in most "real world" situations so these results can be considered general for Bernoulli type analytic variables. The actual reduction in variance was virtually identical to that attained under optimal allocation in all cases; thus, allocation under (2.6) can be considered (near) optimal. The projected reduction in variance was also very close to the actual reduction. When p_2 was smaller than p_1 the actual reduction was always larger than the predicted reduction, and the converse was true when p_2 was larger than p_1. In both cases the maximum difference (which was only about 3.5% of the actual reduction when $\gamma = 2$ and 8.3% of the actual reduction when $\gamma = 10$) occurred when $p_1 = 0.05$ and monotonically decreased as p_1 increased.

In summary; the two cases considered seem to indicate that so long as the assumptions which yield the allocation specified by (2.6) are not radically violated, the variance will be very near that attained under optimal allocation. Furthermore, the proportional reduction in variance given by (2.7) provides an approximation for the actual reduction in variance which is at least accurate enough for the purposes of survey design.

5. CONCLUDING REMARKS

The strengths of the Mitofsky-Waksberg technique for generating telephone samples are clear: high hit rates in the second stage of selection, an efficient method for screening empty banks of telephone numbers, and a conceptually ingenious approach to sample generation. It is a remarkable testimony to the strength of the technique that it is widely considered to be the standard method of random digit dialing with few serious competitors after many years. The weakness of the technique (first stage screening and replacement of non-residential numbers during the data collection) does not, on the surface, seem to be important relative to its general strength. However, these features can cause substantial difficulty, especially in short time-period telephone survey operations.

In this paper stratified designs, based on commercial lists of telephone numbers, are proposed as alternatives to the Mitofsky-Waksberg technique. Both two and three stratum designs are studied in detail. In addition to simple random sampling within each stratum, two general alternatives are considered:

(1) Simple random sampling from all strata except the low density stratum frame where the Mitofsky-Waksberg method is used.

(2) Simple random sampling from all strata except the low density stratum which is not sampled at all.

The basic thesis of this paper is that stratified sampling methods, using strata based on counts of listed telephone

numbers, are at least as efficient as the Mitofsky-Waksberg technique. Furthermore, these designs can eliminate the need for the troublesome replacement of non-residential numbers at the second stage, since the only telephone numbers that must be dialed in the high density stratum are those that are generated at the beginning of the study. Specific conclusions include the following:

- For low cost ratios, the two and three stratum designs are as efficient as the Mitofsky-Waksberg approach.
- When numbers can be dropped from the low density stratum, these alternatives designs are much more efficient, but at the price of unknown bias due to excluding part of the target population.
- When cost ratios are high, the two and three stratum approaches are clearly superior.

A critical issue is the magnitude of the bias introduced by dropping the low density stratum. As noted previously, approximately 7% of U.S. households do not have a telephone and truncating the frame may add to the non-coverage bias. As less than 5% of the U.S. household population is expected to be contained in the low density stratum it is likely that the additional coverage bias will not be substantial for many characteristics of the total population. On the other hand, for some characteristics, and for some subgroups of the population, the magnitude of the additional bias may be large enough to be of concern. Further empirical investigations of this population must be conducted.

There are two costs associated with the use of stratified designs that may detract from their use: the cost of the commercial list used to stratify the BCR frame and the overall lower hit rate. The cost of stratifying the frame into high and low density strata is not addressed in this investigation because the requisite information was derived from a specialized research file. The cost of stratification is a fixed cost and therefore will reduce the resources available for data collection. It is not known what the fixed cost will be in the future as arrangements are made with commercial vendors to routinely provide such data. Furthermore, this fixed stratification cost can be amortized over multiple studies to greatly reduce its impact on any single sample. It is unlikely that data collection for one time surveys will find either the Mitofsky-Waksberg or the stratification method described here to be as cost-effective as indicated. Further investigation is needed into the frame costs before a complete answer can be found.

The second cost issue concerns the lower hit rates presented in this paper. Given the relative competitive efficiencies of the alternatives considered here, it appears that the lower hit rates do not seriously detract from the efficiency of the alternatives. It may be possible to improve the hit rates in the high density stratum if smaller banks of numbers are used. For example, in another investigation we have found that 10-banks will have hit rates in the neighborhood of .57 compared to the .52 reported here for 100-banks. Of course, working with 10-banks substantially increases the size of files and processing operations that must be used to generate samples and the cost of a 10-bank frame is likely to be much higher than the 100-bank frame.

The cost models as shown in (2.2) and (2.3) are relatively simple, ignoring many cost differences in the telephone survey process that may be important for comparisons of relative efficiencies of the designs. These cost models allow the allocations to be expressed in a straightforward way, but they do not specifically address the cost components associated with two features of the Mitofsky-Waksberg technique that the alternative designs address; replacement of nonworking numbers and weighting to compensate for exhausted clusters. Thus, the cost models ignore structural cost differences between the Mitofsky-Waksberg approach and the proposed alternatives that, if properly taken into account, could effect the relative efficiency of the two methods.

Clearly the results presented here are insufficient to draw final conclusions about the overall value of these alternative designs. Further cost data and empirical evidence on the size of the bias caused by eliminating the numbers from the low density stratum is required before a final conclusion can be reached.

ACKNOWLEDGMENTS

The support and assistance of Clyde Tucker and Bob Groves is gratefully acknowledged. The findings and opinions expressed in this article are those of the authors and do not necessarily reflect those of the U. S. Bureau of Labor Statistics or the University of Michigan.

REFERENCES

BRICK, J.M., and WAKSBERG, J. (1991). Avoiding sequential sampling with random digit dialing. *Survey Methodology*, 17, 27-41.

BRUNNER, J.A., and BRUNNER, G.A. (1971). Are voluntarily unlisted telephone subscribers really different? *Journal of Marketing Research*, 8, 121-124.

BURKHEIMER, G.J., and LEVINSOHN, J.R. (1988). Implementing the Mitofsky-Waksberg sampling design with accelerated sequential replacement. In *Telephone Survey Methodology*, (Eds. R. Groves, *et al.*) 99-112. New York: John Wiley and Sons.

GROVES, R.M. (1977). An Empirical Comparison of Two Telephone Designs. Unpublished report of the Survey Research Center of the University of Michigan, Ann Arbor, MI.

GROVES, R.M., and LEPKOWSKI, J.M. (1986). An experimental implementation of a dual frame telephone sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 340-345.

LEPKOWSKI, J.M. (1988). Telephone sampling methods in the United States. In *Telephone Survey Methodology*, (Eds. R. Groves, *et al.*) 73-98. New York: John Wiley and Sons.

MITOFSKY, W. (1970). Sampling of telephone households. Unpublished CBS News memorandum, 1970.

POTTHOFF, R.F. (1987). Generalizations of the Mitofsky-Waksberg technique for random digit dialing. *Journal of the American Statistical Association*, 82, 409-418.

STOCK, J.S. (1962). How to improve samples based on telephone listings. *Journal of Advertising Research*, 2, 55-51.

SUDMAN, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10, 204-207.

SURVEY SAMPLING, INC. (1986). Statistical characteristics of random digit telephone samples produced by Survey Sampling, Inc. Westport, CT: Survey Sampling, Inc.

TUCKER, C., CASADY, R.J., and LEPKOWSKI, J.M. (1992). Sample allocation for stratified telephone sample designs. To appear, *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Survey Methodology, June 1993
Vol. 19, No. 1, pp. 115–121
Statistics Canada

Poisson-Poisson and Binomial-Poisson Sampling in Forestry

Z. OUYANG, H.T. SCHREUDER, T. MAX and M. WILLIAMS[1]

ABSTRACT

Binomial-Poisson and Poisson-Poisson sampling are introduced for use in forest sampling. Several estimators of the population total are discussed for these designs. Simulation comparisons of the properties of the estimators were made for three small forestry populations. A modification of the standard estimator used for Poisson sampling and a new estimator, called a modified Srivastava estimator, appear to be most efficient. The latter is unfortunately badly biased for all 3 populations.

KEY WORDS: High value timber; Volume estimation; Estimators for Poisson-Poisson sampling; Simulation comparisons; Forest sampling; Srivastava estimation.

1. INTRODUCTION

Volume estimation in forestry has been highly developed in the sense that very efficient sampling strategies are available to estimate total volume (Schreuder and Ouyang 1992). Estimating and measuring defect is often not built into these strategies since measuring defect is difficult and not economically justified in most stands. But in high value stands two-phase strategies such as Poisson-Poisson sampling may be suitable where defect is measured on trees at the second phase. To sample truck loads of logs, binomial-Poisson sampling may be a suitable sampling design.

The purpose of this article is to present the theory of binomial-Poisson and Poisson-Poisson sampling and discuss some of the properties of estimators for these designs based on simulation.

2. REVIEW OF LITERATURE

Singh and Singh (1965) developed the theory for two-phase sampling with probability proportional to size (*pps*) sampling at the second phase. Furthermore, Särndal and Swensson (1987) gave a general theory of two-phase sampling. A list of sampling units is assumed to be available at the first phase prior to sampling.

Hajek (1957) developed Poisson sampling and Grosenbaugh (1964) suggested its use for one-phase unequal probability sampling when no list is available. Poisson sampling is a scheme such that each unit in a population, say unit i, is drawn into the sample independently with probability p_i. Thus the inclusion probability of unit i is equal to p_i, and joint inclusion probability of units i and j is equal to p_ip_j. Binomial sampling, also often called Bernoulli sampling, is a special case of Poisson sampling when all p_i are equal.

In forest survey, Poisson sampling is often implemented as follows (Schreuder *et al.* 1968).

1. Visit the N units (say trees) in the population in any order and measure or ocularly estimate the value of a covariate x_i $(i = 1, \ldots, N)$ highly correlated with the value of interest y_i $(i = 1, \ldots, N)$.
2. As each x_i is observed, compare it with a random integer, δ_i, randomly selected from the range $1 \leq \delta_i \leq L$, where L is an integer selected prior to sampling. L is picked such that $L = X/n_e$ where X = total for the covariate in the population and n_e is the desired sample size. X is usually not known before sampling and needs to be estimated.
3. If $\delta_i \leq x_i$, select the unit for the sample and measure y_i.

Implementation of this method results in a sample of size n, where $E(n) = n_e$ (if a good estimate of X was made prior to sampling). In binomial sampling all the $x_i (i = 1, \ldots, N)$ are the same (Goodman 1949).

3. SAMPLING METHODS

The United States Forest Service Region 6 (Wendall L. Jones - personal communication) uses a truck load sampling method as follows: as trucks pull up to the mill a binomial sampling technique is used to randomly select

[1] Z. Ouyang, Formerly post-doctoral fellow, Statistics Dept., Colorado State University Fort Collins, Colorado, now Research Statistician, ICI Seeds, Inc., Slater, Iowa; H.T. Schreuder, Project Leader, Multiresource Inventory Techniques Project, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado; T. Max, Station Biometrician, USDA Forest Service, Pacific Northwest Experiment Station, Portland, Oregon; M. Williams, Statistician, Multiresource Inventory Techniques Project, USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado.

trucks to be sampled, with $p = 0.10$ say. These truck loads are measured for volume. A problem with this approach is that there are long runs of no trucks being sampled. As communicated to one of the authors, this was considered highly undesirable from a practical point of view. An alternative approach, which should decrease the frequency of long runs of no samples, and could be more efficient is to use binomial - Poisson sampling instead as follows:

Apply binomial sampling with a larger p (say $p = 0.30$). The scaler visually estimates volume on the selected loads. A Poisson subsample of these loads is then selected with probability proportional to the estimated volumes and the loads selected at this phase are scaled for volume. This is binomial-Poisson sampling.

For high-value timber stands in the Pacific Northwest of the United States highly accurate estimates of net volume, that is, usable volume is often desired. Actually cutting down and destructively measuring sample trees is the most reliable method of determining net volume, *i.e.* total volume minus defective volume (Johnson and Hartman 1972). Poisson-Poisson sampling may be a good sampling design in this situation. The procedure is:

1. Select n_1 out of the N trees in the population by Poisson sampling, selecting the trees proportional to some estimate of gross volume, say x_1 = diameter at breast height squared (d^2). With Poisson sampling actual sample size is random, say n_1 where $E(n_1) = n_{e1}$. Ocularly estimate say x_2 = ocular net volume.
2. Select n_2 out of the n_1 sample trees proportional to x_2, by Poisson sampling. Here $E(n_2) = n_{e2}$ is the expected sample size at the second phase.

The n_2 sample trees are then cut and destructively measured for gross, net, and defective volume. To maintain maximum efficiency in both inventory and operations it is probably best to implement both sampling phases at once and mark the n_2 sample trees at inventory time. Ascertaining usable volume for these n_2 trees is done later either by a different crew or by carrying the sample trees into a sawmill to process them for actual wood products. Binomial-Poisson sampling is a special case of this. (If the second phase is implemented separately from the first phase then a list of sampling units is available to implement the second phase and some *pps* procedure with fixed sampling size should be used instead of Poisson sampling. This approach is usually inefficient because it requires two trips to the field location).

4. NOTATION

N = Population size (not known until sampling is completed).

n_e = Expected sample size in one-phase Poisson sampling.

n = Achieved sample size in one-phase Poisson sampling.

n_{e1} = Expected sample size of first phase in two-phase Poisson sampling.

n_1 = Achieved sample size of first phase in two-phase Poisson sampling.

n_{e2} = Expected sample size of second phase in two-phase Poisson sampling.

n_2 = Achieved sample size of second phase in two-phase Poisson sampling.

Y = Total usable volume in the population (to be estimated by two-phase sampling), $Y = \sum_{i=1}^{N} y_i$.

x_{1i} = Covariate value for tree i at phase 1, say tree diameter at breast height squared (D^2).

X_1 = $\sum_{i=1}^{N} x_{1i}$ (known after implementing the first phase in the entire population).

$\pi_i(P)$ = Probability of selecting tree i in one-phase Poisson sampling ($= n_e x_{1i}/X_1$). If all the $\pi_i(P)$ are equal, this is one-phase binomial sampling.

π_{1i} = Probability of selecting tree i at phase 1 ($= n_{e1} x_{1i}/X_1$).

x_{2i} = Covariate value for tree i at phase 2, say ocular estimate of net volume.

X_2 = Total amount of ocularly-estimated volume in the population (only obtained for the n_1 sample trees at the first phase so X_2 can only be estimated).

π_{2i} = Probability of selecting tree i at the second phase ($= n_{e2} x_{2i}/\sum_{i=1}^{n_1} x_{2i}$).

y_i = Value of interest for tree i (say net volume).

π_i = Probability of selecting tree i through both sampling phases ($= \pi_{1i}\pi_{2i}$).

π_i^* = Approximate probability of selecting tree i through both sampling phases ($= \pi_{1i}^*\pi_{2i}^*$ where $\pi_{1i}^* = n_1 x_{1i}/X_1$ and $\pi_{2i}^* = n_2 x_{2i}/\sum_{i=1}^{n_1} x_{2i}$).

5. THEORY

For Poisson sampling, the estimator

$$\hat{Y}_u = \sum_{i=1}^{n} y_i/\pi_i(P), \tag{1}$$

is unbiased but very inefficient and should be replaced by the following approximately unbiased estimator (Grosenbaugh 1964):

$$\hat{Y}_a = \left\{\frac{n_e}{n}\hat{Y}_u \quad \text{if} \quad n > 0\right\}$$
$$= 0 \quad \text{if} \quad n = 0. \tag{2}$$

The variance of $\hat{Y}_a$, as given in Brewer and Hanif (1983), is

$$V(\hat{Y}_a) = \sum_{i=1}^{W} \pi_i(P)[1 - \pi_i(P)]\left[\frac{y_i}{\pi_i(P)} - \frac{Y}{n_e}\right]^2 + p_0 Y^2,$$

where $p_0 = P(n = 0)$.

For Poisson-Poisson (PP) sampling, an estimator for Y analogous to $\hat{Y}_u$ above is the unbiased estimator

$$\hat{Y}_1 = \sum_{i=1}^{n_2} y_i/\pi_i. \tag{3}$$

This estimator can be horribly inefficient as pointed out for $\hat{Y}_u$ in Poisson sampling (Schreuder *et al.* 1968).

The variance of $\hat{Y}_1$ can be written down by using the general formulas developed by Särndal and Swensson (1987) for unbiased estimation in double sampling:

$$V(\hat{Y}_1) = \sum_{i=1}^{N}\left(\frac{1-\pi_{1i}}{\pi_{1i}}\right)y_i^2 + E_1\left\{\sum_{i=1}^{n_1}\left(\frac{1-\pi_{2i}}{\pi_{2i}}\right)\left(\frac{y_i}{\pi_{1i}}\right)^2\right\},$$

where E_1 denotes expectation over the first-phase sample. Since $\hat{Y}_1$ is not efficient we do not give its variance estimator. Analogous to the more efficient adjusted estimator in Poisson sampling we have the approximately unbiased estimator

$$\hat{Y}_2 = \sum_{i=1}^{n_2} y_i/\pi_i^* = \hat{Y}_1(n_{e1}/n_1)(n_{e2}/n_2). \tag{4}$$

The variance of $\hat{Y}_2$ is:

$$V(\hat{Y}_2) = p(\phi)\,Y^2 + \sum_{i=1}^{N}\pi_{1i}(1-\pi_{1i})\left(\frac{y_i}{\pi_{1i}} - \frac{Y}{n_{e1}}\right)^2$$

$$+ \sum_{s_1 \neq \phi} p_1(s_1)\left\{\sum_{i\in s_1}\pi_{2i}(1-\pi_{2i})\left(\frac{y_i}{\pi_{1i}\pi_{2i}} - \frac{n_1 Y}{n_{e1}n_{e2}}\right)^2\right\},$$

where s_1 denotes the first-phase sample, $p(\phi)$ is the probability of drawing an empty sample, which is equal to

$$p(\phi) = p_1(\phi) + \sum_{s_1 \neq \phi} p_1(s_1)p_2(\phi),$$

and p_1 and p_2 denote respectively the sampling design for the first-phase and the second-phase sampling design conditional on the sample drawn in the first-phase.

Usually, population size is large and the first phase sample size is also large (compared to the second phase sample size). Thus we can safely assume $p_1(\phi) \doteq 0$ (compared to $p_2(\phi)$). For example, if we draw a first phase sample with expected sample size 50 out of a population of size 500, and then we draw a second phase sample with expected sample size 20 out of the first phase sample, all by using bionomial sampling, the inclusion probability in the first phase is 0.1 and the probability to draw an empty first phase sample is $(0.9)^{500}$; but the inclusion probability in the second phase is roughly .04 and the probability to draw an empty second phase sample is $(0.6)^{50}$. Notice that $(0.9)^{500} \doteq (0.3487)^{50} << (0.6)^{50}$ Thus, in most practical applications,

$$p_1(\phi) \doteq 0.$$

A variance estimator of $\hat{Y}_2$ can hence be easily given:

$$v_1(\hat{Y}_2) = p_2(\phi)\,\hat{Y}_2^2$$

$$+ \frac{n_{e1}n_{e2}}{n_1 n_2}\sum_{i=1}^{n_2}(1-\pi_{1i})(y_i/\pi_{1i} - \hat{Y}_2/n_{e1})^2/\pi_{2i}$$

$$+ \frac{n_{e2}}{n_2}\left[\sum_{i=1}^{n_2}(1-\pi_{2i})\left(\frac{y_i}{\pi_{1i}\pi_{2i}} - \frac{n_1\hat{Y}_2}{n_{e1}n_{e2}}\right)^2\right]. \tag{5}$$

Estimator (5) should work well in usual applications. Sometimes when ocularly estimating net volume, however, the field worker may estimate that a tree has no value but turns out to be incorrect. Thus, some x_{2i}, hence π_{2i}, will be zero (in the simulations a small value is added to those so that $\pi_{2i} > 0$). In this case, a more stable term is needed to replace the last term in (5). Notice that

$$\frac{n_{e2}}{n_2}\sum_{i=1}^{n_2}\pi_{1i}(1-\pi_{1i})(y_i/\pi_{1i} - \hat{Y}_2/n_{e1})^2/\pi_{2i}$$

is an improved estimator of

$$\sum_{i=1}^{n_1}\pi_{1i}(1-\pi_{1i})(y_i/\pi_{1i} - \hat{Y}_2/n_{e1})^2. \tag{6}$$

To ensure that the estimator does not become too large when one or more π_{2i} are close to zero, we use the following estimator

$$\left\{\left[\sum_{i=1}^{n_2}\pi_{1i}(1-\pi_{1i})\left(\frac{y_i}{\pi_{1i}}-\frac{\hat{Y}_2}{n_{e1}}\right)^2\right]\Big/\left[\sum_{i=1}^{n_2}\pi_{2i}\right]\right\}n_{e2}. \quad (7)$$

If we consider x_{2i} as the auxiliary characteristic of $\pi_{1i}(1-\pi_{1i})\ (y_i/\pi_{1i}-\hat{Y}_2/n_{e1})^2$, then (7) is a ratio estimator of (6), since $\pi_{2i} \propto x_{2i}$ for $i = 1, \ldots, n_1$. But since x_{2i} is not necessarily approximately proportional to $\pi_{1i}(1-\pi_{1i})\ (y_i/\pi_{1i}-\hat{Y}_2/n_{e1})^2$, (7) may not be a very efficient estimator of (6). The advantage of using (7) is that $\sum_{i=1}^{n_2}\pi_{2i}$ will not be close to zero, so that (7) will be stable.

This leads to the following variance estimator:

$$v_2(\hat{Y}_2) = p_2(\phi)\ \hat{Y}_2^2 + \frac{n_{e1}n_{e2}}{n_1}\left[\sum_{i=1}^{n_2}(1-\pi_{1i})\left(\frac{y_i}{\pi_{1i}}-\frac{\hat{Y}_2}{n_{e1}}\right)^2\right]\Big/\sum_{i=1}^{n_2}\pi_{2i} + \frac{n_{e2}}{n_2}\sum_{i=1}^{n_2}(1-\pi_{2i})\left(\frac{y_i}{\pi_{1i}\pi_{2i}}-\frac{n_1\hat{Y}_2}{n_{e1}n_{e2}}\right)^2, \quad (8)$$

which is less affected by small probabilities than (5) and hence is more stable. We will use (8) instead of (5) as a variance estimator of $\hat{Y}_2$.

Let E_1 denote the expectation with respect to the first phase and E_2 denote the expectation with respect to the second phase. Since n_2 is the actual sample size and $E\,n_2 = E_1E_2n_2 = E_1n_{e2}$, the adjusted estimator in *PP* sampling should be $E_1n_{e2}/n_2\ \hat{Y}_1$. But the quantity E_1n_{e2} is not available and is replaced by n_{e2} to obtain the following estimator:

$$\hat{Y}_3 = \frac{n_{e2}}{n_2}\hat{Y}_1. \quad (9)$$

$\hat{Y}_3$ should also have very small bias and the variance of $\hat{Y}_3$ is

$$V(\hat{Y}_3) = p(\phi)Y^2 + \sum_{i=1}^{N}\frac{1-\pi_{1i}}{\pi_{1i}}y_i^2 + \sum_{s\neq\phi}p_1(s_1)\left\{\sum_{i\in S_1}\pi_{2i}(1-\pi_{2i})\left(\frac{y_i}{\pi_{i1}\pi_{2i}}-\frac{Y}{n_{e2}}\right)^2\right\}.$$

A variance estimator of $\hat{Y}_3$ is

$$v(\hat{Y}_3) = p_2(\phi)\hat{Y}_3^2 + \frac{n_{e2}}{n_2}\left[\sum_{i=1}^{n_2}\pi_{2i}(1-\pi_{1i})\left(\frac{y_i}{\pi_{1i}\pi_{2i}}\right)^2\right] + \sum_{i=1}^{n_2}(1-\pi_{2i})\left(\frac{y_i}{\pi_{1i}\pi_{2i}}-\frac{\hat{Y}_3}{n_{e2}}\right)^2. \quad (10)$$

Another possible estimator is based on the idea that we first want an efficient estimator of the first-phase information. This is accomplished by an analogous estimator to $\hat{Y}_a$ in eq. (2):

$$\hat{Y}_a(2) = \sum_{i=1}^{n_2}(y_i/\pi_{2i})n_{e2}/n_2 \quad \text{if} \quad n_2 > 0.$$

This estimator can be expanded to estimate Y by dividing the first-phase sample by its probability of selection and we obtain

$$\hat{Y}_4 = \left[\hat{Y}_a(2)\Big/\left\{\prod_{i\in s}^{n_1}p_{1i}\prod_{j\notin s}^{N}(1-p_{1j})\right\}\right]\Big/2^{N-1}, \quad (11)$$

where $i\in s$ indicates that unit i is in the sample, $j\notin s$ indicates that j is not in the sample, $p_{1i} = n_{e1}x_{1i}/X_1$, and 2^{N-1} is the number of all samples.

The variance of $\hat{Y}_4$ is

$$V(\hat{Y}_4) \doteq (2^{-2(N-1)})\left[\sum_{s_1\neq\phi}T(s_1)^2/p_1(s_1)\right] - Y^2 + (2^{-2(N-1)})\sum_{s_1\neq\phi}\left\{\sum_{i\in s_1}\pi_{2i}(1-\pi_{2i})\left[\frac{y_i}{\pi_{2i}}-\frac{1}{n_{e2}}T(s_1)\right]^2 + p_2(\phi)T(s_1)^2\right\}/p_1(s_1),$$

where $T(s_1)$ is the total of y over s_1. It can be easily derived by using the formula

$$V(\hat{Y}_4) = V_1E_2(\hat{Y}_4) + E_1V_2(\hat{Y}_4),$$

and the variance given for $\hat{Y}_a$.

This estimator is expected to be highly unstable. A possible improvement is to condition the estimator on the actual sample size obtained, *i.e.*,

$$\hat{Y}_5 = \left[\hat{Y}_a(2) \Big/ \left\{ \frac{\left[\prod_{i \in s}^{n_1} \pi_{1i} \prod_{j \notin s} (1 - \pi_{1j}) \right]}{P_1(n_1)} \cdot \binom{N-1}{n_1 - 1} \right\} \right], \quad (12)$$

where $P_1(n_1)$ is the probability of drawing a first phase sample of size n_1.

To compute this probability, let I_i be the random variable which is 1 if unit i is in the sample and 0 otherwise. Hence $n_1 = \sum_{i=1}^{N} I_i$, and

$$E(n_1) = n_{e1}, \text{Var}(n_1) = \sum_{i=1}^{N} \pi_{1i}(1 - \pi_{1i}) = d.$$

If

$$r = \frac{n_0 - n_e}{\sqrt{d}},$$

$$\phi(r) = (2\pi)^{-1/2} \exp[-\frac{1}{2}r^2],$$

$$f_m(r) = \left[\frac{1}{\sqrt{d}}\right] \phi(r) \left[1 + \sum_{j=1}^{m} p_j(r)\right],$$

where $P_j(r)$ are Edgeworth polynomials. Then

$P_1(n_1) \doteq f_{m1}(r)$ and specifically, for $m = 2$

$$P_1(n_1) \doteq f_2(r) = \left[\frac{1}{\sqrt{d}}\right] \phi(r) \left[1 + \frac{1 - 2\bar{\bar{\pi}}}{6\sqrt{d}} (r^3 - 3r) + \frac{1}{4!} \frac{1 - \overline{6\pi(1-\pi)}}{d} (r^4 - 6r^2 + 3) + \frac{10}{6!} \frac{(1 - 2\bar{\bar{\pi}})^2}{d} (r^6 - 15r^4 + 45r^2 - 15)\right],$$

where

$$\bar{\bar{\pi}} = \frac{\sum_{i=1}^{N} \pi_i^2(1 - \pi_i)}{\sum_{i=1}^{N} \pi_i(1 - \pi_i)}, \quad \overline{\pi(1-\pi)} = \frac{\sum_{i=1}^{N} \pi_{1i}^2(1 - \pi_{1i})^2}{\sum_{i=1}^{N} \pi_{1i}(1 - \pi_{1i})}$$

(Hájek 1981).

$\hat{Y}_4$ and $\hat{Y}_5$ are only given for completeness. They are not considered further since both are unstable.

An alternative to $\hat{Y}_4$ and $\hat{Y}_5$ is to correct $\hat{Y}_a(2)$ using an expansion factor based on the information for covariate x_1. These estimators are sensible if $\hat{Y}_a(2) / \sum_{i=1}^{n_1} x_{1i}$ is an approximately unbiased estimator of $R = Y/X_1$ which is true for binomial-Poisson (BP) but not for PP sampling. This fact is verified by simulation, but the reason why approximate unbiasedness holds for binomial-Poisson is that $\hat{Y}_a(2) / \sum_{i=1}^{n_1} x_{1i}$ under binomial sampling is similar to the ratio estimator under simple random sampling. Hence the following estimator is only appropriate for *BP* sampling.

$$\hat{Y}_6 = X_1 \left[\hat{Y}_a(2) \Big/ \sum_{i=1}^{n_1} x_{1i} \right]. \quad (13)$$

The variance of $\hat{Y}_6$ is

$$V(\hat{Y}_6) = \frac{N^2}{n_{e1}^2} \sum_{i=1}^{N} (y_i - Rx_i)^2 \pi_{1i} + E_1 \left\{ \frac{x_i}{n_1 \bar{x}_1 s_1} \left[\sum_{i=1}^{n_1} \pi_{2i}(1 - \pi_{2i}) \left(\frac{y_i}{\pi_{2i}} - \frac{n\bar{y}_{s1}}{n_{e2}} \right)^2 + p_2(\phi)\, n_1^2 \bar{y}_{s1}^2 \right] \right\}.$$

Another promising estimator is based on Srivastava's (1985) proposed unbiased estimator $\hat{Y}_{sr1}$ based on the sample weight function concept. Srivastava and Ouyang (1992) developed a structure for the sample weight in order that $\hat{Y}_{sr1}$ has zero variance at some points of the parameter space $\{y_1, \ldots, y_N\}$. The sample weight function can use any information other than that given in a sample. Examples of this kind of information have been given in Srivastava and Ouyang (1992) and Ouyang and Schreuder (1992). If the information can be formulated as a model

$$y_i = \alpha + \beta x_i + e_i, \; i = 1, \ldots, N, \quad (14)$$

then the so called "generalized ratio estimator approximation" (Ouyang *et al.* (1992)) can be used which gives the following estimator of the population total:

$$\hat{Y}_7 = \left[\frac{\hat{Y}_1}{\sum_{i=1}^{n_2} y_i^*/(\pi_{1i}\pi_{2i})} \right] Y^*, \qquad (15)$$

with $\hat{a}$ and $\hat{\beta}$ weighted regression coefficients, and y_i^* calculated by $y_i^* = \hat{a} + \hat{\beta}x_{1i}$ and $Y^* = \sum_{i=1}^{N} y_i^*$.

Note that $\hat{Y}_7$ is dependent on the model assumption.

6. SIMULATIONS

Simulation samples with first-and second-phase samples of expected sizes 50 and 20 in Poisson-Poisson and binomial-Poisson sampling were each drawn from three populations. Two populations were high-value fir, cedar and pine trees. Population 1, called BLM1 (Data from unpublished report "Comparison of volume estimates made by several timber measurement methods in western Oregon" by G. B. Hartman. Feb., 1971. Bureau of Land Management, Portland, Oregon), contained 331 trees and population 2, called BLM2, included 510 trees (Johnson and Hartman 1972). Measured variables on each tree were: net volume scaled (nvs), net volume dendrometered (nvd), and diameter at breast height (d). Here nvs ($= y$) is the variable of interest, $x_1 = d^2$ is used in the first phase of PP sampling and $x_2 =$ nvd is the more expensively but presumably additionally useful covariate obtained at the second level of PP sampling; 200,000 simulations were performed. Ideally, one would like the first- and second-level covariates to be relatively uncorrelated yet both highly correlated with y. These would be d^2 or nvd at the first phase and some measure of defect at the second-phase. Unfortunately, to do this in a satisfactory manner requires separating trees into a class where the field worker is comfortable estimating defect and another class for which he does not. This was not done for the available data. In BP sampling trees were selected with equal probabilities at the first phase and proportional to x_2 at the second phase. Population 3, a mapped data set, called Surinam, was also used since it was cleaner than the other populations in terms of having available more sensible variables for Poisson-Poisson sampling. The population consists of a 60-ha mapped Surinam forest for which only species and diameters were recorded (Schreuder *et al.* 1987). Tree heights and standing tree volumes for other species were superimposed on these trees as described in Schreuder *et al.* (1992). The resulting population consists of 5,525 trees for which tree diameter (d), height (h) and volume (v) were available. This yielded covariates $x_1 = h_i^2$ and $x_2 =$ standing gross tree volume for PP sampling. For BP sampling x_2 was used at the second phase. Board foot volume (y) was also added to the data set. Included are 10 trees for which d^2h is large ($\geq$ 60,000) but bd. ft. volume is essentially zero; 10,000 simulations were performed for the Surinam data. Results for BLM1, BLM2, and Surinam are given in Tables 1, 2 and 3 respectively.

Table 1

Simulation results for BLM1 ($N = 331$) population. 200,000 simulations were performed using $x_1 = D^2$ and $x_2 =$ nvd as covariates*

Estimator	Bias		SE		EASE	
	BP	PP	BP	PP	BP	PP
$\hat{Y}_1$	0.021	0.011	42.495	53.228		
$\hat{Y}_2$	−0.045	−0.770	37.272	48.219	97.787	97.806
$\hat{Y}_3$	−0.050	−0.777	39.819	49.349	97.492	96.763
$\hat{Y}_6$	0.012		39.992			
$\hat{Y}_7$	−0.036	3.650	18.881	21.885		

Table 2

Simulation results for BLM2 ($N = 510$) population. 200,000 simulations were performed using $x_1 = D^2$ and $x_2 =$ nvd as covariates*

Estimator	Bias		SE		EASE	
	BP	PP	BP	PP	BP	PP
$\hat{Y}_1$	0.146	0.059	95.708	62.500		
$\hat{Y}_2$	0.055	−0.424	90.247	55.876	100.325	98.583
$\hat{Y}_3$	0.050	−0.411	91.259	58.701	99.779	98.679
$\hat{Y}_6$	0.146		94.100			
$\hat{Y}_7$	0.486	4.391	26.788	19.855		

Table 3

Simulation results for Surinam ($N = 5{,}525$) population. 10,000 simulations were performed using $x_1 = D^2$ and $x_2 =$ ocular estimate of net volume*

Estimator	Bias		SE		EASE	
	BP	PP	BP	PP	BP	PP
$\hat{Y}_1$	0.764	0.364	25.709	25.924		
$\hat{Y}_2$	0.290	−0.402	15.636	10.845	97.492	97.37
$\hat{Y}_3$	0.019	−0.463	20.989	17.886	100.364	98.945
$\hat{Y}_6$	1.013		20.822			
$\hat{Y}_7$	2.277	2.426	22.428	17.397		

* All tables give bias and standard error (SE) expressed as a percentage of the population net volume. The estimated average standard error (EASE) is expressed as a percentage of the simulation standard error. Expected sample sizes are $n_{e1} = 50$ and $n_{e2} = 20$ for both binomial-Poisson (BP) and Poisson-Poisson (PP) sampling.

7. RESULTS AND DISCUSSION

For PP sampling $\hat{Y}_2$ is the most efficient estimator of the three ($\hat{Y}_1$, $\hat{Y}_2$, and $\hat{Y}_3$) relatively assumption-free estimators for BLM1 and BLM2; $\hat{Y}_3$ is slightly less efficient than $\hat{Y}_2$. Note that $\hat{Y}_7$ is even more efficient than $\hat{Y}_2$ but $\hat{Y}_7$ has a serious bias in some cases. The variance estimators for $\hat{Y}_2$ and $\hat{Y}_3$, $v(\hat{Y}_2)$ and $v(\hat{Y}_3)$, in eq. (8) and (10) are approximately unbiased.

For BP sampling, $\hat{Y}_7$ has negligible bias and the smallest standard error of all the estimators. $\hat{Y}_2$ is considerably less efficient than $\hat{Y}_7$ for BLM1 and BLM2 but more efficient than the other estimators. The variance estimators for both $\hat{Y}_2$ and $\hat{Y}_3$ are approximately unbiased.

Note for BLM1, BP sampling is always more efficient than PP sampling whereas for BLM2 PP sampling is more efficient with $\hat{Y}_1$, $\hat{Y}_2$ and $\hat{Y}_3$. This is because x_2 is not the logical variable to measure after the effect of x_1 is removed. Unfortunately a better variable to assess defect was not available for these data. For BLM1 x_2 did not but for BLM2 it did improve estimation.

For both PP and BP sampling, using population Surinam, $\hat{Y}_2$ is again the most efficient estimator of the three ($\hat{Y}_1$, $\hat{Y}_2$ and $\hat{Y}_3$) relatively assumption-free estimators. $\hat{Y}_3$ is considerably less efficient than $\hat{Y}_2$. $\hat{Y}_7$ is less efficient than $\hat{Y}_2$ and is substantially more biased for this population. $v(\hat{Y}_2)$ and $v(\hat{Y}_3)$ seems to be a approximately unbiased variance estimators for $\hat{Y}_2$ and $\hat{Y}_3$. For this population PP sampling is more efficient than BP sampling with $\hat{Y}_2$ showing that in this case both $x_1 = d^2h$ and $x_2 =$ standing gross total volume are useful in sampling.

Actually, it is not surprising to see $\hat{Y}_2$ is the most efficient estimator, since it uses the most amount of information at both the design and estimation stages. Estimator $\hat{Y}_7$ tends to be even more efficient in terms of mean squared error, but with larger bias. This is because $\hat{Y}_7$ is based on the model given in equation (14). If the model is correct, $\hat{Y}_7$ should be preferred over $\hat{Y}_2$, since $\hat{Y}_7$ incorporates even more information from the population. But otherwise, $\hat{Y}_2$ should be preferred. $\hat{Y}_7$ is not recommended if model (14) is not justified.

8. RECOMMENDATIONS

1. Both Poisson-Poisson and binomial-Poisson sampling are useful in practical forest sampling. With either procedure, estimator $\hat{Y}_2$ should be used. This estimator, with negligible bias and high efficiency, is analogous to the adjusted estimator $\hat{Y}_a$ used in Poisson sampling and has a reliable variance estimator.
2. Estimator $\hat{Y}_7$ is considerably more efficient than $\hat{Y}_2$ for 2 populations but should not be used in preference to $\hat{Y}_2$ until it has been more fully investigated in additional studies. $\hat{Y}_7$ tends to be seriously biased in these simulations.

ACKNOWLEDGEMENT

We appreciate valuable comments by a referee.

REFERENCES

BREWER, K.R.W., and HANIF, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.

GOODMAN, L.A. (1949). On the estimation of the number of classes in a population. *Annals Mathematical Statistics*, 20, 572-579.

GROSENBAUGH, L.R. (1964). Some suggestions for better sample-tree measurement. *Society of American Foresters. Proceedings*, 36-42.

HAJAK, J. (1957). Some contributions to the theory of probability sampling. *Bulletin of the International Statistical Institute*, 36,127-133.

JOHNSON, F.A., and HARTMAN, G.B. (1972). Fall, buck and scale cruising. *Journal of Forestry*, 566-568.

OUYANG, Z. (1990). Investigation of some estimators and strategies in sampling proposed by Srivastava. PhD thesis. Colorado State University Fort Collins, CO, 83.

OUYANG, Z., and SCHREUDER, H.T. (1992). Srivastava estimation in forestry. Submitted to *Forest Science*.

OUYANG, Z., SRIVASTAVA, J.N., and SCHREUDER, H.T. (1992). A general ratio estimator and its application in model based inference. *Annals. Institute of Statistical Mathematics*, (in press).

SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

SCHREUDER, H.T., SEDRANSK, J., and WARE, K.D. (1968). Sampling and some alternatives, I. *Forest Science*, 14, 429–454.

SCHREUDER, H.T., BANYARD, S.C., and BRINK, G.E. (1987). Comparison of three sampling methods in estimating stand parameters for a tropical forest. *Forest Ecology and Management*, 21, 119-128.

SCHREUDER, H.T., and OUYANG, Z. (1992). Optimal sampling strategies for weighted linear regression estimation. *Canadian Journal of Forest Research*, 22, 239-247.

SCHREUDER, H.T., OUYANG, Z., and WILLIAMS, M. (1992). Point-Poisson, point-pps, and modified point-pps sampling: Efficiency and variance estimation. *Canadian Journal of Forest Research*, (in press).

SINGH, D., and SINGH, B.D. (1965). Some contributions to two-phase sampling. *Australian Journal of Statistics*, 7, 45-47.

SRIVASTAVA, J.N. (1985). On a general theory of sampling, using experimental design. Concepts I: Estimation. *Bulletin of the International Statistical Institute*, 51, 1-16.

SRIVASTAVA, J.N., and OUYANG, Z. (1992). Studies on a general estimator in sampling, utilizing extraveous information through a sampling weight function. *Journal of Statistical Planning and Inference*, 31, 177-196.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 8, Number 4, 1992

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Division, Statistics Sweden, S-115 81 Stockholm, Sweden

Applied Statistics

JOURNAL OF THE ROYAL STATISTICAL SOCIETY (SERIES C)

CONTENTS

Volume 42, No. 2, 1993

Printed in Great Britain at the Alden Press, Oxford

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.