

This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: Annals of Economic and Social Measurement, Volume 3, number 2

Volume Author/Editor: Sanford V. Berg, editor

Volume Publisher: NBER

Volume URL: <http://www.nber.org/books/aesm74-2>

Publication Date: April 1974

Chapter Title: Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey

Chapter Author: Horst E. Alter

Chapter URL: <http://www.nber.org/chapters/c10116>

Chapter pages in book: (p. 75 - 100)

CREATION OF A SYNTHETIC DATA SET BY LINKING RECORDS OF THE CANADIAN SURVEY OF CONSUMER FINANCES WITH THE FAMILY EXPENDITURE SURVEY 1970

BY HORST E. ALTER*

The synthetic data set was created by matching conceptually compatible family units from two sample surveys. This paper describes the analysis of the survey data prior to matching and for purposes of formulating matching specifications. It then explains the rationale underlying these matching specifications as well as the execution of the match. The paper concludes with a quality evaluation of the synthetic file.

INTRODUCTION

The need to match records from the 1970 Canadian Survey of Consumer Finances (SCF) with records from the 1970 Family Expenditure Survey (FEX) arose in connection with a research project which endeavored to measure and compare relative income distributions internationally. These relative income distributions were to employ income concepts based on the System of National Accounts (SNA). Neither the SCF, nor the FEX alone contained all the necessary information. However, the necessity to obtain such information had been known prior to the implementation of the surveys, and thus design, data collection, and processing had taken account of the intended use.

To issue a joint questionnaire containing all the questions asked in both surveys would not have been a feasible alternative. The response burden would have been intolerable, and the number of refusals would have been unacceptably large. Furthermore, response errors would have occurred more frequently because of the length of such a combined questionnaire.

The SCF is basically an income survey [1]. It is carried out on a national basis, and has become an annual survey as of 1971. It was a biennial survey prior to that date. The 1970 Survey of Consumer Finances (SCF), covering the 1969 reference year, contained a special section of asset and debt questions [2]. There were approximately 10,000 families and unattached individuals in the sample; whereas the 1970 Family Expenditure Survey (FEX) was the first national survey of its kind since 1948. It was based on a recall questionnaire and covered approximately 14,000 usable returns from families and unattached individuals with 1969 as the reference period [3]. Both surveys relied on the same sampling frame, the Canadian Labour Force Survey. Interviews were carried out in early 1970, with the SCF being executed about two months later than the FEX.

*The author is a research officer with the Consumer Income and Expenditure Division of Statistics Canada. The match project was carried out as part of a research program dealing with income inequality under the guidance of Mrs. G. Oja, Director of the Division. For systems analysis and programming we are indebted to H. Simon of this organization, and special thanks go to J. Lewis of the Regional Research Staff, Statistics Canada, who made his regression package available.

This paper is a condensed and revised version of a paper presented under the same title at the Conference on Price and Consumer Expenditure Data and Workshop on the Merging and Matching of Microdata sponsored by the National Bureau of Economic Research, at Williamsburg, Va., May 1973. Views expressed in this paper are entirely those of the author and do not necessarily reflect those of his associates or those of his employer.

Analysis and matching started with the assumption that both data sets would be "clean," but a number of minor flaws were detected during matching operations, and consequently some matches had to be redone. Additional shortcomings were discovered during the analysis of the matched file, but these shortcomings appear to be similar to flaws ordinarily encountered in survey data. Thus, one should bear in mind that the criticism directed towards synthetic data bases should start with the assumption that input into such bases is far from perfect.

The need for a high-quality data input into matching operations has been covered adequately elsewhere, particularly by Edward C. Budd [4]. He has also discussed the merits and limitations of matched data sets vis-à-vis other data sources. Very stimulating comments, criticisms, and rebuttals were published in conjunction with Benjamin Okner's work [5]. Most of these comments and reservations are still applicable, and the X - Y - Z problem remains unsolved [6]. Basically, the X - Y - Z problem states that a joint distribution of X , Y , and Z cannot be inferred from the known distributions of X with Y , and X with Z . The validity of the implicit assumption of independence in the context of Okner's work has been challenged by Christopher Sims, and has been defended by John E. Peck [7].

It is not intended here to reopen old arguments; but it would seem appropriate at this point to view the SCF-FEX match in the light of these unresolved problems. It cannot be avoided since approach and methodology follow that of Okner in many ways, although appreciable departures exist as well. Brief mention of these departures will now be made.

As far as the data base is concerned, Okner had to link an administrative file with a field survey. He had to deal with conceptual alignments as well as with sampling problems. The SCF-FEX match had been conceptually aligned in the planning stage of these surveys, and the sampling frame was such that both surveys were compatible in terms of coverage as well. It follows that some reservations concerning independence and the distribution of X , Y , and Z can be ignored; or at least that the impact of possibly violating the assumption of independence is even less significant in this instance than it was found to be in Okner's case.

Another important difference to be kept in mind centers on the matching criteria. Okner's matching characteristics were determined on *a-priori* grounds because of the nature of his files and the purpose to which they were to be put. The use of Internal Revenue data and the analysis of taxation questions requires the compatibility of matches within the legal framework which governs taxation. The relationship of income size, family size, marital status and other variables to taxation status, marginal tax rates and other aspects of taxation is anchored in law. One has to make use of one's knowledge of these legal provisions, and translate them into operationally feasible concepts for the purpose of matching records of this sort.

The SCF-FEX match had to rely on behavioral aspects concerning asset and debt holdings on one side and consumption patterns on the other: both related to income, demographic properties, and categorical variables, such as home-ownership.

These behavioral relationships had to be established by conducting a pre-match analysis. *A-priori* reasoning was not sufficient to determine the type of

variable to be used for matching purposes, and its relative importance vis-à-vis other matching variables. This does not imply that *a-priori* knowledge or even some form of arbitrariness could be dispensed with.

It is quite conceivable that the matching specifications and routines could have been designed without the benefit of regression analysis. It is even reasonable to assume that the outcome would have differed only marginally. But deficiencies detected in the post-match evaluation would have defied any form of explanation apart from making some sort of an intelligent guess. To analyse the faults of a method and to improve it methodically, one has to know the rationale which underlies the methodology being developed.

There is also an illustrative effect attached to the regression analysis. It constitutes a further contribution to the X - Y - Z argument. Professor Budd has drawn our attention to the fact that "the X 's are not really defined in the same way in each file [8]." But even in the event that they are so defined, one may hypothesize that they are afflicted with errors of different types, magnitudes, and frequency of occurrence. Thus, what is equal by designation, namely, X , is not really equal when we compare $X + e'$ with $X + e''$.

The problem becomes even more complex, when we realize that X is sometimes a composite variable, such as total income. It is defined as the sum of conceptually identical components in both data sets, but its actual composition varies from record to record. Furthermore, X is not a single variable, be it simple or composite. It is usually an array of variables. This may be of little consequence for the theoretical treatment, but it makes any operational solution highly complex and extremely costly.

It has also been ignored so far that Z and Y are also composite variables. Their components may reveal attributes vis-à-vis X , which are quite different from those of the composite variable.

For purposes of illustration, let Y be total consumption, and let it be explained by X_1 to X_n . When investigating components of Y , such as food consumption, clothing consumption, and others, then the explanatory variables X_1 to X_n assume different order of explanatory power vis-à-vis component parts of Y . Thus, the choice of Y or Z vis-à-vis a set of X 's further complicates what looks like a three-dimensional problem, but which in reality is multi-dimensional.

Some of the foregoing statements will become clearer, as the matching of the SCF and FEX files is presented in detail and in chronological order. The presentation starts with the regression analysis, which in turn supplies the rationale for the matching specifications. These specifications and the ensuing routine are presented with appropriate remarks for the purpose of justifying the form of these specifications in the light of the preceding analysis. A brief description of the match and some resulting "run statistics" complete that part of the paper, which in essence is a variation of known techniques tailored to a special set of circumstances.

The last section of this paper deals with attempts to assess the quality of the matched file. The result is somewhat inconclusive. With guarded optimism one may wish to say that we are on the right track, but that we have a long way to go. Or perhaps, one may wish to conclude that some aspects of matching have progressed well, while others require rethinking. The incorrigible pessimist, on the other hand, may be less charitable in assessing the results associated with this project.

Nevertheless, the facts are there. They have been presented with as much detail as is justified under the prevailing circumstances. Hence, the conclusion reached by the author and by the reader may well differ.

ANALYSIS OF DATA INPUT INTO THE MATCH

It was considered necessary to analyse SCF and FEX data prior to matching in order to establish a rationale for the selection of matching variables and the determination of matching criteria. The regression analysis to be discussed in this section follows conventional multiple-linear-least-squares assumptions: but the purpose of this exercise must be viewed differently from that of the usual interpretation associated with regression models. Neither are we interested in the significance of certain regression coefficients and the economic interpretation of this phenomenon, nor do we investigate the explanatory power of the *model* for purposes of simulating or forecasting certain events.

The purpose of this analysis is to identify matching variables, to determine the relative importance of each variable within the set of matching variables, and to find ways of translating this relative relationship into scores and other criteria. It is the formulation of these criteria which will make a matching routine operational, and it is the pre-match analysis which will form the foundation for such a matching routine.

Models to be specified for this purpose are highly dependent on the data base to be used and also on the intended use to which the match will be put. These aspects have already been covered in the introductory section.

It will be recalled that the Survey of Consumer Finances and the Family Expenditure Survey had been designed in such a way that certain variables would be collected on both questionnaires, and that these variables would be conceptually compatible. Moreover, both surveys were based on a common sampling frame. Furthermore, the field work was carried out with the smallest possible time gap between both surveys. With these precautions, and with edit and assignment procedures internal to each survey also properly co-ordinated, the input files, which were used for matching purposes, were as compatible as can be expected.

For any variable to be considered as a matching characteristic, it was required that such a variable be present in both data sets. This is a necessary but not a sufficient condition. If such variables exhibited some explanatory power vis-à-vis consumption in the Family Expenditure Survey and also vis-à-vis asset holdings and debt patterns in the Survey of Consumer Finances, only then would they be considered as matching variables. These variables were then ranked in order of importance, and their relative impact was quantified. This ranking and quantifying was done with the help of regression analysis. The mandatory matching of certain variables, on the other hand, had to be established on an *a-priori* basis in most instances.

For purposes of this paper, the regression analysis and its interpretation will have to be dealt with in a general way. It is impossible to present the full detail, and any compromise between a highly detailed presentation and a rigorous abstraction would create more questions than it would answer. Thus, rigorous abstraction will be the order of the day.

Some preliminary computer runs using a stepwise routine indicated the need for partitioning each data set because some variables seemed to be significant for some family types and not for others. This approach, of course, would later result in the specification of different matching routines for different types of families.

Partitioning followed two principal ways of classification: homeowners versus non-homeowners, and families of two or more versus unattached individuals. The resultant models were labelled "SCF and FEX analytical models I to IV." Analytical model number one (AM I) always refers to families living in their own home, "AM II" refers to unattached individuals living in their own home, while "AM III" and "AM IV" correspond to families and unattached individuals who are *not* living in their own home.

Depending on the nature of the partitioned set, some explanatory variables were common to several sets, while others were set-specific: e.g., income was common to all sets, mortgage status and home equity were common to "homeowner sets" (AM I and AM II), whereas number of children and wife's earning status applied to family sets (AM I and AM III) only.

The dependent variable for the survey of Consumer Finances was basically "net worth," although its disaggregated form in terms of assets only and debt only was also used. Moreover, assets and debts were defined in different ways depending on the liquidity of the assets and the type of debt: i.e., consumer or personal debt. Net worth, of course, is defined as the difference between assets and debts.

The dependent variable for the Family Expenditure Survey was originally "total expenditure," but later "total current consumption" was chosen because "total expenditure" was explained almost entirely by income. Current consumption was further disaggregated into its major components such as food, shelter, clothing expenditures, and others.

These modified models should not be viewed as systems of equations. The reason for keeping the set of explanatory variables fixed and using it in conjunction with parts of a larger and already explained variable was to see, whether a variable or variables would explain secondary and primary phenomena equally well. For example, total income would explain all forms of consumption to a very high degree. It was the top-ranking variable in all but two instances. In case of the two homeowner models, mortgage status contributed more towards the explanation of shelter expenses than did income. Area-of-residence code, on the other hand, helped to explain shelter expenses in all models, but it failed to explain any other form of consumption outlay.

These two examples must suffice, but the principle to be kept in mind is this: the greater the explanatory power of one independent variable versus another independent variable, the greater its value in the matching process. The more frequent the occurrence of a variable in equations explaining parts of a larger unit, and the more frequent the appearance of a variable in a number of models, the greater its value as a matching variable. Unfortunately, a conflict arises where low frequency is associated with great explanatory power. The question now arises, what is meant by explanatory power in this context?

Ordinarily the explanatory power of an *equation* is judged by its *R*-square. In this instance, *R*-square values differed appreciably between equations because

model specifications in the conventional sense were inadequate. Adding variables to explain certain forms of consumption, or asset portfolios, or indebtedness would have been of little interest since such additional variables are not contained in both data files to be matched. Thus the relative contribution of a *variable* to the *R*-square of an *equation*, however small or large it may be, was used to measure the *explanatory power of a variable*.

If a variable contributed 0.16 to an *R*-square of 0.32, it explained 50 percent of an equation with inferior specifications: if another variable, or the same variable, contributed 0.45 to an *R*-square of 0.90, it had the same explanatory power as the variable in the preceding example, but the specification of the model is obviously superior. A variable contributing 0.20 to an *R*-square of 0.60 had less explanatory power than any of the two earlier cited, namely 33 percent, but the equation as such was more powerful than the first-cited example, and less powerful than the second. To repeat, explanatory power of a variable was determined by its percentage contribution to the *R*-square of the equation in question.

Let us now register a few caveats against this notion. In the event that the explanatory power of two variables in any equation was equal, the one with the greater "partial *F*" was assigned the higher rank. Similarly, if the explanatory power of variables was less than 10 percent, they were ranked according to their "partial *F*'s" only. Finally, the explanatory power of variables with insignificant regression coefficients was ignored.

The explanatory power of a variable and its frequency of appearing as a significant contributor in various models and equations of both data sets would determine its ultimate usefulness as a matching tool. But, as was mentioned earlier, conflicts would arise between relative explanatory power and frequency of occurrence within models or equations related to the same data base. Moreover, some variables would be considered very useful in connection with one data base, and only marginal in connection with the other.

The resolution of these conflicts together with other aspects will be discussed in the following section, where the matching specifications are outlined, and where the matching routines are described.

The reader should also be prepared to face elements of arbitrary solutions, some based on *a-priori* knowledge, others on intuition. It is well to admit that the creation of synthetic data files by way of matching is just as much an art as it is a science. This admission, however, does not diminish the value of the product. Whether it enhances its value is beyond proof. However, value and quality will not be discussed until specifications and matching routines have been reviewed, and it is this topic which will be presented in the following section. Only upon completion of the section dealing primarily with operational aspects will our attention be shifted to questions of quality.

MATCHING SPECIFICATIONS AND COMPUTERIZED ROUTINES

Specifications assigning matching variables and quantifying their relative importance varied between data sets. Four such data sets had been created from each survey data base. These data sets correspond to the four analytical models described in the preceding section; they were classified according to home-

ownership and family type. Specifically, there was a Survey of Consumer Finance (SCF) data set of families living in their own home to be matched with an identically classified data set from the Family Expenditure Survey (FEX). For future reference these two sets will be called "match base I." Similarly "match base II" resulted from the SCF and FEX sets containing unattached individuals living in their own home. "match base III" contained family units not living in their own home, and "match base IV" contained unattached individuals who failed to live in their own home.

Matching specifications distinguished between mandatory matches and desirable matches. As the name implies, mandatory matches imposed a necessary condition upon a match. Two records could not be considered for matching, unless all mandatory conditions were met. However, a number of desirable matches also had to take place. It is here that trade-offs were possible and that a point system had to be constructed. More will have to be said about this aspect a little later.

Most mandatory matches were universally applicable, and the reason for their selection cannot always be found in the regression analysis. For example, the five regions of Canada were designated a mandatory match. The main reason for this decision can be found in the weighting scheme, which differs between regions. On the other hand, it should be pointed out now that weights of the SCF file were retained at all times. In other words, the SCF file was considered the main or primary file, and supplementary information was assigned thereto from the secondary FEX file.

Major-source-of-income categories were also made mandatory. These had revealed elements of multicollinearity in the regression analysis and thus were omitted therefrom, although some form of explanatory power could not be ruled out for these categories. Making them mandatory seemed to overcome possible ambiguities.

Broad age groups of heads of families were made mandatory matches because life-cycle variables employed in the regression analysis had also posed problems of interpretation. These life-cycle variables are of the categorical type and embrace age of head, number of children, age of children, and family type.

Other classificatory variables also became mandatory matches because they had been part of the composite life-cycle variable. These other variables, however, did not apply to all match sets; e.g., sex codes pertaining to heads of families could have been applied to all match sets, but were considered important for unattached individuals only. For families of two or more, male heads were in the overwhelming majority. Thus it was highly probable that male heads with spouses would be matched with their sex-specific counterparts from the opposite file. On the other hand, male or female heads with children but without a spouse would be matched virtually at random. Considering future uses of the file, this element of randomness was considered of little importance.

Child status code was also mandatory and applied to match bases I and III only. This condition was built into the specifications in order to avoid the matching of childless families with those having children. While the *number* of children was used as a desirable match, it was hypothesized that the impact resulting from matching a record containing "*n*" children with a record containing "*n* + 1" or "*n* - 1"

children is less harmful than that resulting from matching a childless family with a family having one child. In other words, the impact of a marginal child is less pronounced, given that at least one child exists, as compared with the impact from an incremental child from zero to one. The behavioral pattern of childless families undoubtedly differs from those having children, whereas families of similar sizes with children will exhibit similar net-worth configurations and consumption patterns. *ceteris paribus*.

Some of these mandatory matches had to be converted into desirable matches later on, while others had to be discarded completely. This change was instituted as matching became more difficult near the extremities of the distributions of certain variables, and as fewer records remained to be matched. Under these circumstances one had to ask, for example, whether a match of a family with a very high income with another family of similar income should be done within regions but at the cost of a greater income deviation, or whether one should match across regions, thereby minimizing the income difference of matched records. It was argued that socio-economic compatibility should outweigh regional concordance, and thus regional constraints were removed under these circumstances. More will be said about the removal or conversion of mandatory matches at a later time. First let us turn our attention to desirable matches.

Variables being used as desirable matches fall into two categories: they either represent quantities, such as income in dollars, number of children per family unit and others, or they are categorical variables, such as mortgage status. In the first instance, agreement between conceptually identical fields to be matched can be sought within specified limits. The better the agreement, that is, the smaller the numerical difference between the quantities, the greater the score to be assigned. In the second case, namely that of categorical variables, a score will be assigned if, and only if, the categorical statement agrees on both records to be matched.

Considering the nature of the match base, the number of variables used as desirable matches, and the point score assigned to each variable, the maximum number of attainable points differed between match bases. It was as high as 130 points for match base I, and as low as 97 points for match base IV.

The actual number of points attained was expressed as a percentage of the number attainable. This percentage was called the "union score," and a high union score in conjunction with the compliance with all mandatory conditions determined the acceptability of a match.

Union score limits were initially set at a level of 95 percent, but were gradually reduced as matches could no longer be obtained at the specified level. More will be said about this aspect when the discussion of matching routines takes place. For now, a few paragraphs will be devoted towards the detailed presentation of the four match bases, their desirable match variables and corresponding point scores, as well as reasons for placing certain variables in their relative positions.

Since the nature of programmed routines is repetition, components common to certain match sets were awarded identical point scores. Obviously, some match sets had fewer components than others: "component" being a group of related matching characteristics or variables. Such a group will also be called a *decision module*.

The description of match base I will be fairly detailed. This set contains a large number of decision modules, whereas sets II to IV contain only some of these modules, therefore requiring only a brief description of their overall make-up.

Desirable matches pertaining to set I can be described in terms of five decision modules. The first one handles major-source-of-income amounts, the second one handles total family income, the third one specifies age of family head and place of residence, the fourth one deals with homeownership aspects, and the fifth module handles variables pertaining to attributes associated with families of two or more.

Total family income had been used for analytical purposes as a continuous variable, while major source of income had entered the regression model as a categorical variable. In this form it became a mandatory match, while the major-source amount had not been taken into consideration so far. With the decision to use the major-source category as a mandatory match, it was reasoned that the major-source amount should be given some weight as well. Since the major-source amount and total income are the same in many instances, the combined effect of these two items should not be greater than that of total income when used by itself. Because of this interrelationship between total income and major source, the first two decision modules will be discussed simultaneously.

Total family income had contributed anywhere from 26 to 58 percent to the R-square in SCF analytical models. It had occupied either first or second rank among other explanatory variables. The impact of total income was even greater on the expenditure side, where it ranked first in fourteen out of sixteen cases, and where its explanatory power ranged between 44 and 95 percent. Thus, the combined income effect was allotted 70 points in the first two decision modules. In view of the fact that the total number of points attainable varies between match sets, income accounts for anywhere from 54 to 72 percent of the union score. To be more specific, in the case of match base I income accounts for a possible 70 out of 130 points, or for approximately 54 percent.

The first decision module would assign 40 points, if the major-source amounts considered for matching were in agreement within two percent of the compared amount shown on the primary record. A discrepancy of ten percent constituted the lower limit, and contributed 30 points to the total score. Anywhere from 36 to 31 points were assigned for matching differences ranging between two and eight percent.

Total income was handled similarly, but since it is subject to greater variation, a maximum discrepancy of 25 percent was permitted between amounts compared between primary and secondary records. As before, agreement within two percent was awarded the maximum number of points, namely 30, declining gradually to 20 points at the lower acceptance limit.

It should be mentioned at this point that income differences of \$100 in each of the first two decision modules were given the maximum number of points, regardless of the percentage difference represented by this amount. This provision was added in order to overcome problems at the lower end of the income distribution. It also accounts for the fact that reporting errors are apt to exceed \$100.

The weakest possible combination of income would contribute 50 points to the total score, or 38 percent to the union score for match base I. The same number of points would correspond to 52 percent of the union score for match base IV.

Under these circumstances, income could never bring about a match by itself, given that mandatory conditions had been satisfied, although it contributed the lion's share to the union score. Income could not bring about a match, even though it made a large contribution to the union score, because matches were initially consummated at a level of 95 percent, and the acceptance level was never lowered to less than 65 percent of the total score.

We now turn to the third decision module. It assigned five points if the ages of heads of families agreed within five years of each other. The decision to use age differences is based on the composite life-cycle variable in the regression analysis. The age effect, child effect, and family size effect could not be ascertained precisely, and the decision to assign five points, in the event that ages agreed within these specified limits, is somewhat arbitrary.

Generally speaking, life-cycle variables contributed less than ten percent to any R-square in SCF models. Usually, these variables ranked last or second last, and they were insignificant in many subsets. In other words, their frequency of occurrence was low. FEX models performed slightly better, not in terms of frequency, but in terms of the explanatory power of life cycles. In five instances, they exceeded a ten-percent contribution: e.g., explanatory power was as high as 30 percent for food consumption of families not living in their own home. Nevertheless, the overall impact of life-cycle variables was marginal.

The place-of-residence decision departed from the usual treatment applied to categorical variables. Place of residence is identified as either metropolitan or non-metropolitan, and will be referred to henceforth as "met" and "non-met" respectively. It was recognized that agreement in the "met" category almost precludes agreement in the next item, namely "farm affiliation [9]," whereas agreement in the "non-met" category increases the likelihood of finding concurrence in terms of farm affiliation. In order to make the combined effect of "non-met" with farm affiliation equal to that of "met" residence, the combined effect of the first and the single effect of the second were given 15 points each. The breakdown for the combined effect is eight points for "non-met," and seven points for farm affiliation.

Residence categories had been insignificant in SCF models, yet FEX models showed "met" residence ranking either second or third in four instances. The relative contribution to R-square was 12 percent in one of these cases, and less than ten percent in all others. Whenever residence categories had sufficient explanatory power, they ranked higher than life-cycle categories.

Farm residence as an explanatory variable was based exclusively on FEX analytical models. But apart from supporting regression analysis, farm compatibility had to be stressed in view of the intended use to which the matched file were to be put: i.e., the imputation of income in kind for home-grown produce consumed by farmers. Consequently, special attention towards farm compatibility was desirable for this reason alone.

The fourth decision module deals with aspects of home ownership. Agreement of mortgage status, namely the fact that both records considered for matching either had or failed to have a mortgage, was good for ten points. Home equity was also good for ten points, provided equities agreed within \$1,000 when compared. Fewer points were assigned on a sliding scale with a minimum of two for those

equities which differed anywhere from \$10,000 to \$15,000. Such a wide margin of error was tolerated because home equities fluctuate widely between locations. Moreover, the reporting of equities left much to be desired, and assigned equities contained in the original files were also subject to wide margins of error.

Home equity ranked highly in some SCF models: for example, when explaining the holding of liquid assets. For SCF families living in their own home, equity contributed about 40 percent to total *R*-square, and it outranked income in this one isolated case. On the FEX side equity ranked reasonably high in two instances, namely third out of seven, and fourth out of ten contributing variables.

The fifth and final decision module dealt with properties inherent in families of two or more. The wife's earning status is one of these properties, and if it happened to be identical for records to be matched, three points were awarded towards the cumulative point score. If the number of children agreed perfectly, five points were assigned, and if the number of children differed by one, provided some children existed in each record, then two points were allotted. Similarly, five points were awarded for perfect agreement in the number of adults, whereas zero points augmented the total score if any difference existed in the number of adults on records to be matched.

It was argued on *a-priori* grounds that the marginal difference of one child, *ceteris paribus*, kept family units sufficiently similar. The impact of the marginal adult on the degree of similarity between prospective matches, on the other hand, was considered substantial. Consumption patterns, asset portfolios, and debt patterns of such families would depart sufficiently, thereby precluding the assignment of any points under these circumstances.

Considerations for the inclusion of the number of adults and the number of children go back to life-cycle variables and their relative impact revealed in the various regression models. The disaggregation of life-cycle categories into number of adults and number of children is based on intuition and thus similar to circumstances outlined in connection with those surrounding the ages of family heads as discussed in the third decision module.

The foregoing decision modules apply to match base I, and permit the accumulation of a maximum number of 130 points. Match base II excludes the fifth decision module, since we are dealing with unattached individuals. All other aspects remain as discussed above, and a total of 117 points can be scored under ideal conditions. In contradistinction to match base I, match base III excludes the fourth decision module because we are dealing with non-homeowners. The maximum cumulative score is 105 for this set. Finally, match base IV, which consists of unattached individuals who fail to live in their own home, contains the first three decision modules only. This set could accumulate up to 97 points.

The specifications outlined above had now to be translated into a computerized routine, and this routine and its various stages will be the subject of the following paragraphs. However, before initiating computer runs and interrogating records for the purpose of computing union scores, a number of preparatory tasks had to be carried out, such as creating work files of the four match bases. Match sets were sorted by mandatory categories and size of total income. Thus, defaults on mandatory conditions could be bypassed quickly [10].

The matching routine was then initiated at the union score level of 95 percent. If matches were attained at that level, then the file was searched again until no further matches could be found. The union score was then lowered by five percentage points, and the process was repeated. The lowest acceptable union score limit was 65 percent.

The routine just outlined caused the first acceptable match to become the final match: It was not necessarily the best match within the specified limits. Commencing the routine at a high union score level and reducing levels of acceptability gradually, on the other hand, should have kept departures from the ideal at a minimum. Whether these departures in fact were minimized remains an open question.

This brief note must suffice for now. More will have to be said about the matching routine and the approach used whenever the limit of 65 percent in the union score had been reached, but unmatched records remained in the primary file. First, however, let us clarify notions of *primary* and *secondary* file or record.

The SCF file was the *primary* file, and the FEX file was the *secondary* one, and these attributes applied to family records taken from these files for matching purposes. Any SCF record which had been incorporated into a matched pair was removed from the primary file. The matched pair was then stored in order to be merged later with the full-length record of the original file. Any FEX record having entered a matched pair remained in the secondary file, thereby being available for repeated use.

As the union score was gradually lowered whenever matches failed to materialize, a level of 65 percent was eventually reached. The union score was not lowered beyond this limit, but matching specifications were revised instead.

The matching routine, which incorporates these revised specifications, will be referred to as "stage two." Mandatory matches pertaining to broad age codes of heads of economic families were omitted during this stage. The desirable match for age of head was modified insofar as now a maximum of 20 points could be awarded in case of perfect agreement. The points were reduced at a sliding scale until no further awards were made when the age difference exceeded ten years. The net gain resulting from this revision was 15 points for all match sets.

Matches were again initiated at a union score level of 95 percent. This limit, however, was now reduced at ten-percent intervals each time matches had been exhausted. The minimum acceptance level was retained at 65 percent. At the conclusion of stage two, 774 SCF records out of a total of 9,962 remained unmatched. To accommodate these records, a further revision in the matching specifications was carried out. These revised specifications will be designated as "stage three."

Major source of income was omitted as a mandatory match from stage three, and the major-source amount was no longer used in computing points based on desirable matching characteristics. Instead, total income accounted for all 70 points, which earlier had to be accumulated out of a combination of total income and that part which was classified as the major source of income. A maximum permissible discrepancy of 25 percent was retained for total income, and 50 points were assigned in this event. Points falling between the limits of 70 and 50 were assigned on a sliding scale depending on the percentage difference in total income.

The absolute deviation which used to be \$100 for total income and the amount of major-source income each, or a possible combined departure of \$200 under unfavorable circumstances, this absolute deviation was now increased to \$500. Consequently, 70 points were also awarded if total income on two records to be matched agreed within \$500, regardless of the percentage difference represented by this amount.

Stage three was processed just like the previous stage: i.e., computer runs were started at an acceptance level of 95 percent and this level was lowered at ten-percent intervals until matches had been exhausted at a union score of 65 percent. Thirty-eight SCF records remained unmatched at the termination of stage three.

These remaining records were examined individually, and the best possible match was constructed in accordance with the merits of each case. In assessing these merits, reasoning followed that which had been employed in each of the decision modules, but one or another mandatory condition usually had to be violated, or the margin of permissible errors had to be slightly extended: e.g., unattached individuals of the opposite sex were matched, or matches were obtained under established criteria but at a union score level of less than 65 percent.

Before concluding this section of the paper, a few run statistics will be of interest. A total of 9,962 matches had to be carried out. Slightly less than one-half, namely 4,297, were matched uniquely; i.e., the FEX record matched to the SCF record was used only once. FEX records used more than once in a match are summarized as follows: 1,535 records were used twice, 476 records were used three times, 143 records entered a match four times, and 96 records were employed five times or more, but always less than sixteen times.

This concludes the description of the matching routines, but it is not quite the end of the paper. Hopefully, the most interesting part is yet to come. Undoubtedly, it is important to explain the rationale which underlies the matching specifications; undoubtedly it is necessary to justify the methodology employed, but the question of quality cannot be ignored under any circumstances.

How good, or how bad, is the match? How useful is the final product? What reservations must be voiced, and what sort of caveats should be attached? These and other questions will be asked in the next section. Unfortunately, not all of them will be answered. It is always easier to ask a question than to answer it precisely, truthfully, and intelligently.

EVALUATION OF THE QUALITY OF THE MATCHED FILE

Upon completion of the matching operation, each record contained three types of information: first, data common to both files and used in the matching process; secondly, information common to both files but *not* used in the matching process; and thirdly, information originating with either the SCF or the FEX file only. It is this latter type of data which should enhance the usefulness of the matched file; but it is the information common to both files, which will help to assess the quality of the match.

There is no need to spend much time on assessing those items which were used as mandatory matches. Variables used as desirable matches, on the other

hand, may show varying forms of agreement. It is here that a number of tests were performed.

Categorical variables or codes were cross-tabulated. Those classifications which had been matched properly would occupy the main diagonal of the tabulation, while incompatibly matched codes would occupy cells off the main diagonal depending on their particular combination. Tabulations of this sort were not only run for categorical variables which had been specified as desirable matches, but also for those which had never been used in the matching process, but which happened to be common to both files [11]. With these prefatory remarks in mind, the presentation of specific tests will now begin.

Region codes, age codes, and those designating major source of income suffer from the effects of mixed application: i.e., they were mandatory for matching at one time, and they were then converted into desirable categories or disregarded altogether. As a result, out of 9,962 matched records, 1,736 region codes were incompatible. This amounts to 17.4 percent of all records. It should be noted in this context that there are five major regions in Canada, and that the number of incompatible matches is fairly evenly distributed: namely 290 in the Atlantic region, 350 in Quebec, 437 in Ontario, 347 in the Prairies, and 312 in British Columbia. Since original codes and weights from the SCF file will be retained, the mismatched regional codes are of no practical consequence.

Three age codes had been used specifically for matching purposes. They correspond to head's age under 45, from 45 to 65 years old, and 66 years and over. Records had been coded "one" to "three" respectively. Only 658 out of 9,962 records, or 6.6 percent of the file, showed incompatible age codes. The real impact of this discrepancy is further reduced if one realizes that a match of adjacent age codes may signify only a minor difference in actual ages, as long as these ages are close to the common class limit. Only matches between codes "one" and "three" are in severe disagreement, but there are only 82 such records which constitute less than one percent of the file.

It seems to be of interest to compare these results with age differences ascertained through re-interviews. Too often the fact is ignored that survey data are afflicted with non-sampling errors, and that synthetic matches may introduce irregularities which are hardly greater than these non-sampling errors. Unfortunately, although much has been written on the subject of response errors, little has been expressed in quantitative terms. Gladys L. Palmer is one of the few who have supplied some supporting evidence [12]. She shows among other findings that ages reported in re-interviews differed by one year or more for 10.3 percent of the cases where the original respondent supplied the information: but the error increased to 17.1 percent whenever different respondents answered the questions during the re-interview [13]. Since the comparison of matched data is not for age differences of one year or more, a valid inference cannot be drawn for our case from Gladys Palmer's study. Nevertheless, the presence and incidence of reporting errors should be kept in mind when challenging differences arising out of matching processes. With this brief aside, let us now turn to major sources of incomes.

Six major-source-of-income codes are contained in the matched file, and final tabulations show that 554 incompatible codes of that type existed after matching had been completed. This is equivalent to 5.6 percent of the file. The

number of incompatible codes is as high as 131 for code "one" (MSI from wages and salaries) and as low as 48 for code "three" (MSI from farm self-employment). Although the code for wages and salaries shows the greatest number of mismatches, it also shows the smallest percentage error, namely 1.8 percent, whereas the 48 code-three records out of a total of 316 SCF codes constitute 15.2 percent of this category.

It will be recalled that major source of income had been a mandatory match until the end of stage two. Thus it would seem to be more appropriate to assess these 554 incompatible matches against the 774 records which remained to be matched at the commencement of this stage. But it should also be acknowledged that the likelihood of finding compatible major-source matches, given all other constraints, was extremely low with only 774 records remaining. It was precisely for this reason that the mandatory requirement for major-source matches had been removed.

The discussion will now shift to codes designating variables which had been used as desirable matches. Since these codes are of a binomial nature, the outcome of the match is best explained with the help of a table, and the following paragraphs should be read with reference to table 1. The nature of these codes provides for two types of incompatibilities: either type "A," where an SCF code "one" has been matched with a FEX code "two," or type "B," where an SCF code "two" has been matched with a FEX code "one."

Before going into detail, a few general remarks concerning table 1 should be offered. Type A and type B incompatibilities have already been explained, and the sum of these two is the total number of incompatibly matched records for a certain categorical variable. The number of SCF records originally coded "one" or "two" constitutes the base against which the percentage deviation of type A or type B is measured. The number of all incompatible records as a percentage of all records in the relevant universe yields the percentage deviation for all types and is shown in the last column.

In terms of met/non-met matches, incompatibilities of both types would seem to be of equal importance. The fact that SCF non-met codes show a greater incidence of mismatches can be attributed to the matching routine which assigned only eight points for non-met agreement, whereas concordance in terms of metropolitan residence had produced 15 points. The anticipated re-inforcement of non-met codes with farm affiliation matches apparently failed to materialize.

The absolute size of type A and B departures for farm affiliates is about the same, namely 240 and 225 respectively; but the percentage deviation of incompatible farm affiliates based on all farm affiliates by far exceeds the mismatched families not affiliated with farming relative to all non-farm affiliates. The fact that 41 percent of SCF farm affiliates have not been matched with FEX farm affiliates is reason for concern, especially when the intended use of the matched file is kept in mind. Additional internal assignments, as is commonly done for missing data within a survey file, may well be in order.

As far as the matching routine is concerned, specifications should be changed in this respect. Whenever specific applications are foreseen, mandatory matches or relatively inflated scores should be employed in order to increase the number of compatible matches. But it should also be kept in mind that this approach is apt

TABLE 1
SUMMARY OF INCOMPATIBLE CODES DESIGNATING DESIRABLE MATCHES FOR SELECTED BINOMIAL SCF CATEGORIES

	Incompatibility		All incompatible Records	Total SCF			Percentage Deviation		
	Type A	Type B		Code 1	Code 2	All Codes (Universe)	Type A	Type B	All types
Metropolitan (1)	502	890	1,392	5,555	4,407	9,962	9.0	20.1	14.0
Non-Met (2)	240	225	465	585	9,377	9,962	41.0	2.4	4.7
Farm Affiliate (1)	420	422	842	2,464	3,464	5,928	17.0	12.2	14.0
Mortgaged Prop. (1)	1,359	869	2,228	2,878	4,741	7,619	47.2	18.3	29.2
Mortgage-Free Prop. (2)									
Wife With Earnings (1)									
Wife Without Earnings (2)									

Note: Incompatibility "Type A" exists if an SCF code 1 was matched with a FEX code 2 for the relevant category.

Incompatibility "Type B" exists if an SCF code 2 was matched with a FEX code 1 for the relevant category.

The universe consists of all records for metropolitan and farm categories. For mortgage codes, the universe is that of "homeowners," and for wives income status the universe consists of families of two or more with a male head.

to increase the number of repetitions with which secondary records enter the matched file.

With further reference to table 1 it should be remarked that mortgage codes show little difference in terms of incompatibility types. Their percentage deviations fall within three percentage points of the combined deviation of 14 percent. Type A and B incompatibilities are of equal importance in this respect.

Codes pertaining to wife's earning status show the greatest incidence of incompatible matches as compared to other variables shown in table 1. The low analytical value of this variable, and consequently the low point score assigned, probably resulted in a large number of random matches for this category.

This concludes the discussion of codes and their compatibility as far as matching characteristics are concerned. Two intermediate cases will now be presented, namely family-sized codes and life-cycle codes [15]. These did not enter the matching routine, but their subcategories, such as age of head, presence of children, number of children and number of adults, were used as matching criteria. Results, therefore, should compare favourably with those presented for variables used as desirable matches and related to life-cycle categories.

Life cycle and family size both employ seven codes, but the clustering effect around the main diagonal is more pronounced for life-cycle codes than for family-size categories. The universe contains 9,962 records in each case. There are 8,331 compatible matches for life-cycle codes, and family-size codes show 5,702 entries along the main diagonal.

It will now be argued that a deviation of one code within a category can be considered a "near-match." These code numbers are functionally related to quantities and thus lines on our cross-tabulation which are parallel and immediately adjacent to the main diagonal define "near-matches." In other words, SCF_n with FEX_n is a true match and appears on the main diagonal, whereas SCF_n with FEX_{n-1} or with FEX_{n+1} is a code combination differing numerically by "one" and called a "near-match".

With 264 near-matches for life-cycle codes, the total of matches and near-matches amounts to 8,595, while family-size codes show 2,403 near-matches bringing the combined total for this category to 8,105.

One may conclude that the disaggregation of life cycles and of family size had no adverse effect on the code compatibility for these categories. Using 1,365 and 1,857 respectively as the number of matches outside the band of central lines, the percentage deviation would come to 13.7 and 18.6 for life-cycle and family-size codes respectively.

Perhaps, family-size codes should better be assessed for that subset of the universe which eliminates 1,723 unattached individuals, since these had been subjected to mandatory conditions. This alternate approach would yield 6,382 matches and near-matches, or 1,857 severely incompatible codes out of 8,239 families of two or more, thereby constituting a deviation from the norm of 22.6 percent. The magnitude of this departure seems to remain within reasonable expectations, and we shall now turn to variables which are common to both data sets, but which were not employed in the matching process.

Codes never employed in any of the matching routines appear to be randomly distributed. This contention is not very obvious when examining "sex-of-head"

codes. Restricting the comparison to "families of two or more," since sex codes for unattached individuals were made mandatory, we observe 812 incompatible codes, or 9.9 percent of the relevant universe. However, there are 7,619 male heads of economic families and only 620 female heads in the SCF. Assuming that the FEX distribution is similar, the likelihood of a male head from the SCF matching a male head from the FEX remains relatively high. The fact that 620 female heads had found a compatible match in only 150 cases is important to note. Conversely, 470 female heads of families out of 620 possible matches or 75.8 percent were incompatible.

The situation for mother-tongue codes is similar to that in the preceding paragraph because 61 percent of the original SCF records stated English as their mother tongue, while 23 percent stated French, and the remaining 16 percent belonged to other linguistic groups. But six different codes had been assigned to the smallest group in addition to one code each for English and French. The deviation of the code for English against the norm was 24.6 percent, while French departed by 46.3 percent from the norm. The remaining six codes never really found their counterpart, and percentage deviations ranged from 91.0 to 99.0 percent.

Occupations had also been disregarded in the matching process, although randomness had been reduced for very broad categories through the use of major-source concepts. Nevertheless, the universe of occupational codes using 13 classes shows 6,627 incompatible matches, which means that 66.5 percent of all occupational codes on the SCF file have not been matched with their counterpart from the FEX. Farmers, and the heterogeneous group of "miscellaneous occupations" compare favorably with the global statement attached to occupational codes. They show a percentage deviation from the norm of 48.5 and 25.1 respectively. All other occupation codes show deviations from the norm ranging from 72.2 to 98.6 percent.

Finally, the comparative evaluation of education codes will be discussed. It proves to be particularly interesting in the light of prior knowledge which postulates a high degree of association between education and income. Since income size was a predominant influence in the matching process one would expect education codes to reveal a high degree of compatibility. There were seven codes to be compared between both files, and results will show that prior knowledge could have failed us. While a certain level of education may correspond to a certain level of income at one point in the life cycle, the obverse is not true: the level of income does not permit any inference as to the educational attainment.

There were 7,483 records with incompatible education codes, or 75.1 percent of the file. It does not come as a surprise that the highest percentage deviation is associated with the smallest group, namely those reporting no schooling. Out of 126 SCF records coded in this fashion originally, 116 records remained incompatible after the match had been completed, and this amounts to 91.1 percent. The second largest error for a single code belongs to high school graduates, where 1,527 incompatible codes confront a population of 1,886 for a percentage deviation of 81.0. Percentage deviations for other education classes range from 70.4 to 77.5.

Since education progresses along a continuum, and since educational codes are functionally related to educational attainment, the notion of "near-matches" introduced earlier will be applied in this instance as well. The result is that 3,564

near-matches and 2,479 true matches combine for a total of 6,043 acceptable matches, which translates into a percentage deviation of 39.3.

The validity of this notion of near-matches, as interpreted above, is further strengthened by reporting errors which may lead to appreciable distortions in the coding of this category. Supporting evidence will be taken again from Gladys Palmer's study [14]. She had discovered that response errors for educational attainment, when measured as the difference between initial reporting and results from a re-interview, were as high as 26.3 percent whenever different respondents were involved. But even when the respondent in the first and second interview did not change, educational attainment was reported differently in 21.8 percent of all cases. These results are not all that surprising, and the survey literature is full of explanations and hypotheses concerning questionnaire design, recall capabilities and other aspects, all related to the impact and possible ways of ameliorating reporting errors. But if these deficiencies are a fact of life, what sort of tolerance is one inclined to attach to a synthetic match? With this rhetorical question in mind, the comparison of actual quantities will now be discussed.

Actual quantities when used as desirable matches had to agree within the limits laid down in the matching specifications. However, some of these variables could still have been matched at random, provided a sufficiently high union score had been attained with the help of other variables. Moreover, it can be assumed that total income had been employed in the matching process in almost all instances because of attainment of a sufficiently high score without the use of income would have been virtually impossible. Other quantities may show greater elements of randomness because they have entered the matching decision less frequently. However, the null hypothesis is the same in all cases. It is hypothesized that matched FEX quantities will be greater than conceptually identical SCF quantities in just as many cases as they will be smaller. Consequently, the average of these differences should not depart significantly from zero. If this hypothesis cannot be supported, then the alternate hypothesis is phrased in terms of the average difference not being zero: i.e., there is no reason to believe that this average should be either positive or negative. All hypotheses will be tested at the five-percent level of significance.

Table 2 summarizes some of the computational results. Tests have been carried out for the complete file, and also for major subsets, such as families of two or more, unattached individuals, and for those portions of the file associated with the so-called match bases I to IV.

Except for match base II, the null hypothesis for total income differences had to be rejected. While it indicates that matching is not carried out with the precision hypothesized, it does not necessarily render the results useless.

The average difference (\bar{D}) for total income obtained from matched SCF and FEX fields was \$196 for all records, it amounted to \$223 for families of two or more, and to \$69 for unattached individuals. The null hypothesis had to be rejected in all cases.

Let us compare these results with differences between average incomes based on a direct match of a sample of the 1960 U.S. Census of Population and records from the Internal Revenue Service (IRS). The relevant report shows in table A that total money income for married persons filing a joint return differs by \$197

TABLE 2
AVERAGE DIFFERENCES OF MATCHED VALUES FOR SELECTED INCOME COMPONENTS BY FAMILY TYPE
OF FOUR MATCHING SUBSETS

	Total Income			Wages and Salaries		
	\bar{D}	t -Statistic	$H_0: \bar{D} = 0$	\bar{D}	t -Statistic	$H_0: \bar{D} = 0$
All records	196	18.68	No	-23	-1.18	Yes
Families 2+	223	18.08	No	-9	-0.38	Yes
Unattached Individuals	69	4.83	No	-91	-2.80	No
Match Base I	259	16.52	No	-6	-0.20	Yes
Match Base II	58	1.90	Yes	-146	-2.08	No
Match Base III	152	7.76	No	-13	-0.53	Yes
Match Base IV	74	4.62	No	-70	-1.93	Yes

Note: Match bases I to IV contain (I) Families living in own home, (II) Unattached Individuals living in own home, (III) Families *not* living in own home, (IV) Unattached individuals *not* living in own home.

(6,611-6,414) between Census and IRS [16]. This figure compares favorably with our difference of \$196 for all records. A more appropriate comparison might be made by referring to families of two or more in table 2. This comparison is less satisfactory since our difference is \$223 versus \$197 in the Census-IRS comparison study. Wages and salaries are also compared in table A for married persons filing joint returns [17]. The absolute difference between Census and IRS is \$144, whereas our absolute difference for wages and salaries of families of two or more is nine dollars.

Wages and salaries, of course, present a more gratifying picture than total income, and the null hypothesis could not be rejected in the majority of cases. In any event, the comparison of the synthetic match with the direct match shows that matching discrepancies fall within the range of reporting errors. Greater precision, therefore, may not imply greater realism. Nevertheless, efforts to improve the quality of matched data sets, to the point where the null hypothesis cannot be rejected, should be made. But in the meantime, the data will be used because they constitute the best possible match, given the present state of the arts.

Comparisons for other income components were carried out, and led to a rejection of the null hypothesis in almost all instances. In addition to investment income and family allowance receipts, old age pensions, and government transfer payments were tested. Moreover, non-income items, such as the equity of homes and the value of passenger cars, were subjected to the same sort of testing. At present, these results do not add anything to our knowledge. Comparisons with conventional reporting errors through re-interviews or direct matching are not available for these items.

The number of adults per matched family, number of children, and age of head in years were also tested against the familiar null hypothesis. The null hypothesis could not be rejected for age differences, where the average turned out to be 0.02 years with a t -statistic of 0.22. All records formed the universe for this test.

The number of children and the number of adults was tested for records representing families of two or more. For children, the average difference was 0.04

with a *t*-statistic of 2.82: for adults the corresponding values were 0.08 with 8.38. When testing the null hypothesis for family size, computations seem to indicate that the child and adult effect is cumulative, since the average difference for family size is 0.13 with a *t*-statistic of 7.04. Obviously, the null hypothesis was rejected at the five-percent level in all three cases.

CONCLUSION

The quality evaluation undoubtedly indicated a number of imperfections. However, one should remember that the tests were applied to individual records and that the cumulative effect reflected in the *t*-statistic will somewhat overstate the case. The intended use of this data base will be in form of grouped data and for purposes of measuring distributional effects. The group membership of any particular record will not be affected to the same degree as the difference between a matched pair of variables would indicate. Moreover, the SCF variable will always govern in those cases where both files contain the same variable. It is actually the effect of the difference between matching variables on variables which are specific to one set only that counts. This effect is always smaller than the difference, given that no matching variable has 100 percent explanatory power.

One should also recall that survey data, or other data bases, are not free from imperfections in their original state. The matching process cannot overcome these imperfections, and matching to a level of precision which exceeds that of the original data may well be utter folly. While the need to improve matching routines cannot be denied, such improvement must go hand in hand with the amelioration of those data which serve as input into the matching process.

Future matching attempts may have to emphasize quality assessments of input files prior to matching. When combined with operational improvements, such as seeking out the *best match* rather than the *first acceptable match*, synthetic data files undoubtedly will become useful tools for the social scientist. But even the most useful tool has to be used with discretion. Knowing the origin of such a research tool, the process which forged it, and the constraints which had to be imposed upon it, will help to develop this sort of discretion. Whatever use will be made of the combined SCF-FEX file, hopefully will also be made with this word of caution in mind.

Statistics Canada

REFERENCES

- [1] Statistics Canada, *Income Distributions by Size in Canada 1969*, Catalogue No. 13-544.
- [2] Statistics Canada, *Income, Assets and Indebtedness of Families in Canada 1969*, Catalogue No. 13-547.
- [3] Statistics Canada, *Family Expenditure in Canada 1969, Vol. 1*, Catalogue No. 62-535.
- [4] Edward C. Budd, "The Creation of a Microdata File for Estimating the Size Distribution of Income," *The Review of Income and Wealth*, December 1971, pp. 317-333.
- [5] Benjamin Okner, "Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File," *Annals of Economic and Social Measurement* July 1972, pp. 325-342.
- [6] Christopher A. Sims, "Comments" to Okner's 1966 Merge File op. cit., *Annals of Economic and Social Measurement*, July 1972, pp. 343-345.

- [7] John K. Peck, "Comments" to Sims' comments in conjunction with Okner's 1966 Merge File op. cit., *Annals of Economic and Social Measurement*, July 1972, pp. 347-348.
- [8] Edward C. Budd, "Comments" to Okner's 1966 Merge File op. cit., *Annals of Economic and Social Measurement*, July 1972, p. 350.
- [9] Any family was coded a "farm affiliate" if at least one family member had some income from farm self-employment. The correlation between "farm affiliate" and "farm residence" was quite high, and farm relationships were further re-inforced for families with "farm self-employment income equivalent to major-source".
- [10] All programming, including the development of a general program for the matching of microdata, was carried out by Henri Simon of Statistics Canada. Without his help the project would not have come to fruition and I shall remain indebted to him for his support.
- [11] Attempts to measure the number of incompatibly matched codes per record and to classify records in terms of their "error frequency" did not lead to conclusive results. This aspect is only supplementary to the present approach of using cross-tabulations. Moreover, its usefulness is small compared with the tests applied to differences of actual quantities after matching. Therefore, a detailed discussion of "error frequencies" has been omitted from this paper.
- [12] Gladys L. Palmer, "Factors in Variability of Response in Enumeration studies," *Journal of American Statistical Association*, June 1943, pp. 143-152.
- [13] Palmer, "Factors in Variability of Response in Enumeration Studies" op. cit., Table I, p. 146.
- [14] Palmer, op. cit.
- [15] Life-cycle categories are compound variables embracing age of family head in terms of broad age groups, number and age of children, and family type.
- [16] U.S. Bureau of the Census, *Evaluation of Research Program of the U.S. Censuses of Population and Housing, 1960: Record Check Study of Accuracy of Income Reporting*, Series ER 60, No. 8, U.S. Government Printing Office, Washington, D.C., 1970, p. 2.
- [17] U.S. Bureau of the Census, op. cit.

COMMENT

BY CHRISTOPHER A. SIMS. JANUARY 17, 1974

The "Comment" and "Rejoinder" I contributed to the January 1972 issue of this journal, criticizing "matching" work by Benjamin Okner, apply almost without alteration to these two more recent papers. Alter and Ruggles are in places more explicit about what assumptions are necessary to justify matching than was Okner. Alter's checks on the accuracy of the match and the Ruggleses' declared intention to study match accuracy with split samples are also welcome methodological advances over Okner's work. Nonetheless, it remains true that matching incorporates a very poor method of estimating regression functions in sparse regions of the joint distribution being estimated, that it therefore creates certain systematic biases, and that simple modifications of the procedures used by the Ruggleses and by Alter could improve the estimates and help users of the artificial sample to avoid misuse of it.

The Ruggleses make an argument against "the technique of imputation by regression" as a way of generating artificial samples without matching. As they give no references, it is not clear who may have used or proposed this technique, which has certain obvious flaws. It seems to be a technique in which the mean of, say, Y conditional on X is assumed to be some simple function of X over the whole range of the sample, this regression function is estimated, and the estimated function f is applied to each observation (X_i, Z_i) to yield an artificial sample of observations $(X_i, Z_i, f(X_i))$ to be treated as if it followed the true joint distribution of X, Z , and Y . The Ruggleses' objections to this technique are certainly justifiable. In fact they may concede too much in suggesting that this technique could be better than matching in sparse samples when the functional form is well known.

However, they make a misleading assertion in their discussion of this point, in claiming that,

For matching purposes no specific functional relationship need be determined in advance. Non-linear relationships will automatically be handled as efficiently as linear relationships, without explicit recognition that the relationships are non-linear.

As I pointed out in my earlier "Comment" and "Rejoinder," to justify a matching procedure one requires an assumption that the regression relation giving the conditional distribution of (say) Y as a function of X is *constant* and *a fortiori* that the mean of Y is a constant conditional on X . This is a much stronger requirement than the assumption that the conditional mean of Y be linear in X . Of course if one does not believe that the conditional distribution of Y is independent of X , but one is able to specify certain multidimensional intervals in X -space within which the conditional distribution of Y given X is approximately constant, then one can approximately justify separate matching procedures within these intervals. The Ruggleses partially recognize this point when they concede that, "The success of the matching procedure depends on the sample being quite dense." But the

definition of "denseness" which applies here must be that any observation always has a neighboring observation within an interval small enough that within it Y is independent of X . Thus knowing where in X -space a sample is dense requires knowing something about the form of the regression function.

In the kind of application for which matching has thus far been used, X includes income or income components. Income distributions have long, fat tails. This means that there will be intervals which are large in dollar terms which will contain few individuals. Thus it is almost inevitable that there will be regions of the sample—economically very important regions at that—in which the sample is *not* dense in the sense required to justify matching.¹

I have suggested that these sparse regions of the sample must be identified and that in them either the match should not be carried out,² thus preventing misuses of the artificial sample in applications for which those regions are important, or the direct matching technique be replaced with one which requires less stringent assumptions.

An example of the latter sort of technique would be one which began by estimating a linear regression function for the mean of Y as a function of X , $\bar{Y} = f(x)$, within a region of the sample X small enough to justify the approximate assumption of linearity for f . (This region would obviously be larger than the subregions of it within which $f(x)$ could be treated as constant.) Then to form the artificial sample, (X_i, Z_i) observations for X_i in the region would be extended artificially to (X_i, Z_i, Y_i^*) observations by setting $Y_i^* = f(X_i) + U_j$, where $U_j = Y_j - f(X_j)$, and (X_j, Y_j) is an observation from the other sample, within the region, which has been "matched" with (X_i, Z_i) . This proposed method is only a more detailed version of what I suggested in the earlier "Comment" and "Rejoinder." It probably would be somewhat, but not much, more expensive than the alternative of not matching data points which cannot be matched reasonably well under the independence assumption. Also, this method can only improve accuracy along the "fringe" of the sample. There are likely still to be sections of the sample so sparse that linearity of the mean is not a justifiable assumption, and here one should again refrain from matching. It is possible that in some applications more elaborate treatments of the functional dependence of the Y -distribution on X than the simple linear mean-shift would allow a reasonably accurate artificial sample over a wider range of X -space than the local linearity assumption would allow. The benefits of extending the sample this way as compared to the increased computational cost can be determined only by experimentation.

¹ Stratified sampling can partially offset this problem, but in a match using a considerable number of variables it is unlikely to be feasible to stratify in all the relevant dimensions. Simply increasing the size of the sample will increase density everywhere, but is also likely to result in people attempting to use the sample for more finely disaggregated studies. Thus though the Ruggleses are right that for any given study a bigger matched sample should be better, the need for accuracy in the fringes of the sample and for warning flags to users who attempt to use unmatched or badly matched observations is no smaller in large-sample matches.

² Alter goes some way toward implementing this suggestion by attaching a point score to each match. A use of the sample could determine whether a region of the sample which interested him contained badly biased data by computing a mean score over the region. This score is what the Ruggleses call a metric. Alter shows by example that a metric can be computed for diagnostic purposes without creating the computational problems the Ruggleses rightly attribute to a method which tries to choose the optimal match for each observation.

In summary, my fundamental objection to exhaustively matching a pair of samples to generate a new sample is that this is a head-in-the-sand approach. The Ruggleses claim that their procedure provides "systematic processing of information based on objective rules and criteria." But there is no way to avoid "subjective" use of simple economic theory in deciding when a match is bad. In their eagerness to avoid "subjective" assumptions about the nature of the distribution they are estimating, matchers have been letting the computer make foolish assumptions for them.

Lest it be lost sight of, let me point once again to the importance of the independence assumption for the validity of matching. The only reason for using a matched sample is to study relations between the Y and Z variables. If the R^2 in the regression of a particular Y_i on the X vector used in matching is R_{xi}^2 , then the assumption of independence implies that no set of X and Z variables can produce an R^2 higher than R_{xi}^2 in explaining Y_i . Alter's procedure of exploring regressions of Y and Z variables on X -variables is therefore likely to be useful, both as a diagnostic check on the validity of the matching procedure and as a benchmark for users of the artificial sample in regression analysis.

Finally a technical remark about the choice of matching criteria by Alter and the Ruggleses. The Ruggleses " $I(x)$ " method is in principle adaptable to a serious treatment of regions of the sample where matches are bad, and their stated criterion for $I(x)$ size choice is correct: Choose $I(x)$ such that the conditional distribution of Y given X over the interval does not change. However, by insisting that differences in the conditional distribution between intervals be statistically significant under a chi-squared test, they guarantee that no interval will be sparsely populated. The serious problem with matching occurs in regions of the sample over which we *know* the distribution of Y must shift, but within which the sample is sparse. If we pretend we don't know that economic behavior of people with incomes of \$12,000 and \$24,000 is different, we will not, with the Ruggleses sample, be able to *prove* that it does differ. But this does not justify pretending that random matching within the "\$11,800 and over" interval is justified. Prior judgments *must* enter the enforcement of a maximal size for $I(x)$, even when this results in sparsely populated intervals.

More technically, the Ruggleses procedure works with univariate X -intervals in potentially collinear data. Their criterion of " R^2 " on the distribution function may lead to acceptance of $I(x)$ intervals for collinear data which are too broad to allow reproduction of partial correlations in the sample. Conversely, the procedure might lead to a proliferation of redundant fine intervals for a family of highly collinear data. In this respect Alter's method for choosing matching criteria could be better than the Ruggleses': Alter uses *multivariate* regression in deciding which X -variables are important.

University of Minnesota

