The Synthetic Micro Data File: A New Tool for Economists

By Charles A. Sisson*

Detailed data files required to fill many economic models are not available Direct construction of a needed file often proves to be prohibitively expensive The author of this article poses one alternative to synthesize a file by merging two or more existing ones that, between them, contain the needed information For example, consider a researcher who wishes to know how economic variables affect sociological behavior and who has found one file with economic and demographic information and another with sociological and demographic information By matching demographic characteristics, the researcher can construct a synthetic file to use in analyzing the relationship of economic and sociological characteristics

Keywords

Data
Synthetic data files
Research methods

Economists require detailed information about the characteristics of the economy to formulate a rational economic policy The more complete the micro files they employ in their research, the more confident they can be in their policy recommendations National income accounts and other summary figures cannot provide precise enough detail about the interactions at the micro level that are the foundation of economics (*9, 17,* and *13*) [1] As Wassily Leontief noted

> The time is past when the best that could be done with large sets of variables was to reduce their number by averaging them out or what is essentially the same, combining them into broad aggregates, now we can manipulate complicated analytical systems without suppressing the identity of their individual elements (*9*, p 6)

Unfortunately, our ability to process large masses of data has exceeded our means to generate empirical bases for hypothesis testing and policy formulation Ideally, all micro-based studies would stem from a carefully chosen sample of the population that included relevant income, expenditure, tax, and demographic factors The direct approach to constructing such a file—collecting a sample—is prohibitively expensive The impacts of many important policy changes are thus estimated by gross approximation Yet it is obviouslv self defeating to use macro subtotals to examine policy changes that have impacts only through their influence on individuals

Seeking to improve their methodology within the formidable constraints barring construction of a true micro file, researchers have turned to synthetic micro files as a practical and improved base for their policy prescriptions Synthetic micro files are not a true sample of the population They are formed by a matching or merging of two different micro files that, between them, contain information about the desired variables This technique might be useful, for example, if one had a microeconomic source of demographic characteristics for a specific socioeconomic group and another microeconomic source of information on their economic status, but wished to know how economic variables affected sociological structure Suppose a researcher had a microeconomic data file, such as the agricultural census, and wished to extend its usefulness without creating a new microeconomic data file The researcher might consider enlarging its applicability by "merging" it into a second, appropriate data file to create a file that would supplv information on the missing relationships in the first file

In this article, I examine merits and shortcomings of synthetic micro data files, review examples of such files, and explain procedures for constructing them Results should be useful to each researcher in economics

Synthetic data files are no panacea Although they may be a useful research tool, they have shortcomings They may produce the data base for useful and varied micro studies, but results are contingent on the appropriateness of the matching process Although the creation of synthetic data files can significantly save money over the expense of designing and collecting a microeconomic survey, their construction requires huge investments of human and computer time, and patience Studies employing such files can be costly Also, it takes 2 or 3 years to collect the data and another year or more before the file can be constructed If results from studies using such a file are to be more than an historical exer

*At the time this article was written, the author was an economist with the National Economic Analysis Division, ESCS He is now with the International Monetary Fund, Washington, D C

[1] Italicized numbers in parentheses refer to items in Bibliography at the end of this article

Detailed data files required to fill many economic models are not available Direct construction of a needed file often proves to be prohibitively expensive The author of this article poses one alternative to synthesize a file by merging two or more existing ones that, between them, contain the needed information

cise, one must presume that the relations depicted have not changed over the intervening span of years [2]

Several questions arise How is a synthetic micro file constructed? How relevant are these files? How big can a synthetic data file be—that is, if it makes sense to merge data files A and B to form C, does it make sense to merge C and some other data file D to form E? What types of synthetic data files have been constructed? What are their relative merits?

THE BROOKINGS MERGE FILE [3]

The following indepth view of the Brookings MERGE file should make the abstract concepts of synthetic micro files more understandable

The Brookings MERGE file synthetically links individual records from two sources The U S Internal Revenue Service (IRS) tax file for 1966, which contains individual Federal income tax returns, and the 1967 Survey of Economic Opportunity (SEO) data file, which samples the total U S population through field interviews. Both files reference individual family income for calendar year 1966, but each contains information not found in the other The IRS tax file contains more complete tax information, the SEO file, more complete demographic information

The Brookings MERGE file links information from the 87,000 individual records in the IRS tax file with the 30,000 household records in the SEO file Each family record in the SEO file was examined to determine if any member of the family would be expected to have filed a tax return in 1966 If so, tax information from the return judged most likely to have been filed by that individual was added to the demographic information in the sample record The process optimizes a "distance function" after certain basic criteria for a match are satisfied How this is done is explained below

The Brookings MERGE file is not perfectly synthetic Low-income records from the SEO file have no tax data associated with them because no tax was paid High income tax file records in the final version of the MERGE file provide no demographic records because tax information alone is available for families with incomes above $30,000 [4] It is the vast middle range of records that have been artificially linked by the matching process

Once an individual from a SEO family record is judged likely to have filed a tax return, the matching process consists of finding a return in the IRS tax file that closely represents the actual tax return that he or she would have filed First, a set of "cells" or "equivalence classes" is constructed to serve as first-stage, or rough sort All IRS tax returns occupying the same cell as the supposed (constructed) tax return from the SEO record are compared in a second-stage sort, based on income Finally (ideally) a match is determined based on a distance function Each match is randomly determined from all returns that fall within a standard income range of the SEO record

The "equivalence classes" are formed on the basis of four criteria (1) type of return filed—single, joint, surviving spouse, head of household, (2) age of the household head or spouse—65 years old or over, (3) number of exemptions—1, 2, 3, 4, 5 or more, and (4) reported pattern of income—major and minor sources of income (in absolute terms) Four classes of income were considered wages, business, farm, and property

If the cells as defined had been interpreted strictly, there would have been 1,420 different categories So many cells would have been left empty that many SEO records would have been impossible to match, and still more would have yielded improbable income matches Accordingly, the 1,420 original cells were collapsed to 74 somewhat densely populated cells

The original criterion for the second stage, or major income source match, was the major income for the

[2] In the short run, this is usually a reasonable assumption Joseph Pechman stated at the National Tax Association-Tax Institute of America Symposium held in Washington, D C , July 10, 1975 "The 1966 MERGE file shows the gist of income and tax distributions even today "

[3] The material in this section is drawn from (*11* and *14*, pp 84-92, unless otherwise noted

[4] Groups with high and those with negative incomes are not really separate as returns with substantial losses are generally filed by wealthy persons See (*11*, p 335)

There may be a means of eliminating some bias in the SEO survey Personal surveys are notorious for underreporting income for those at high income levels See (*14*, p 85)

SEO unit, plus or minus 2 percent[5] All returns in the acceptable income range were subjected to a final sieve, a "consistency score," to help narrow the choice Hitherto unused information was used to effect the most suitable match Six criteria were used, each with different weights If the tax return matched the related characteristic on the SEO record, it was awarded points

Tax return	SEO record	Points
Home mortgage interest deduction on property tax return	Home ownership or debt (or house value in farm value)	12
Interest or dividend income	Interest or dividend income or ownership of stocks, bonds, and others	8
Farm income	Farm income or farm assets or debt	10
Business income	Business income or business assets or debt	10
Rental income or real estate property tax deduction	Rental income or real estate assets or debt	9
Nonzero capital gains income[1]	Dividends or interest on stocks, bonds, and others	8

[1] Capital gains equal to zero on the tax return and earnings from property in the SEO file are consistent

The return with the highest consistency score was not necessarily matched to the SEO record Any return in the highest 25 percent of those for which consistency scores were calculated was equally likely to be selected, if it scored 25 points or more This procedure was sufficient to make most of the matches (97 percent)

SEO records that could not be matched by this technique were reprocessed, and an iterative process was begun The income acceptability range was widened by 1 percent,[6] consistency scores for eligible tax returns were recalculated, and a match was determined based on the same consistency criteria Records that failed this test were reprocessed six times, or until a match was determined, each time the acceptable income range was widened 1 percent Records still lacking a match—0 5 percent of the 28,643 returns—were hand matched

THEORETICAL BASIS FOR SYNTHETIC MICROECONOMIC DATA FILES

Constructing a synthetic, merged file generally involves merging two samples whose overlap is insignificant Certain variables, denoted by the vector X, appear in both samples Other variables, represented by the vector Y, appear in one sample, others, Z, appear in the second sample[7] The ideal is a single sample with information on the joint distribution F(X,Y,Z) As this does not exist, an artificial one must be generated This construction is a special case of the following general problem Given samples from two marginal distributions of a joint distribution, estimate the joint distribution and generate a sample from it[8] The difficulty is estimating the joint distribution

The problem involves so many variables that it is difficult to conceptualize Graphical presentation is also difficult Two partial views that may assist the reader are in figures 1 and 2 Figure 1 shows the crux of the matter The joint distribution of X and Y and the joint distribution of X and Z are known, but the joint distribution of Y and Z is unknown If X and Y are single variables, the joint distribution of X and Y

[5] Limitations on the total amount of error excluded impossible (at lower limits) or overgenerous (at higher limits) margins of error

[6] With corresponding increases in the maximum and minimum amounts that were permissible

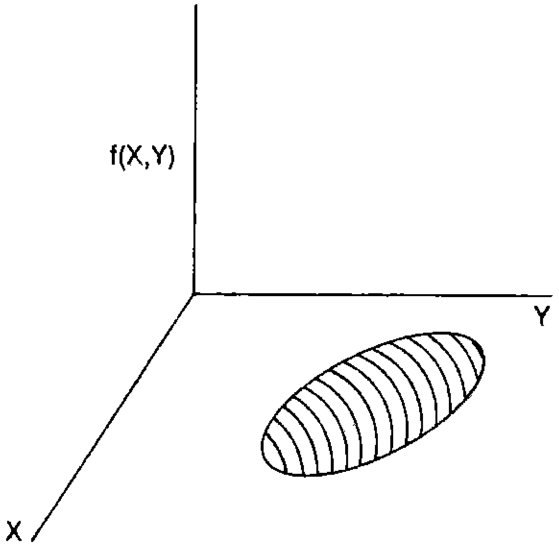
[7] This notation, which relies on (*21*), is standard in the literature

[8] This approach relies on (*21*)

FIGURE 1
Distribution of Variables in Joint Sample of (X,Y,Z)

	X	Y
X	Known	Known
Z	Known	Not Known

FIGURE 2
Joint Distribution of X and Y



in 3-space might look as depicted in figure 2 The joint distribution F(X,Y,Z) occupies 4-space and cannot be represented here It can, however, be defined in terms of conditional probabilities

$$F(X,Y,Z) = F(Y/Z,X) \quad F(Z/X) \cdot F(X)$$

The latter two terms in this expression are known However, as the joint distribution of Y and Z is not known, F(Y/Z,X) is unknown If we could estimate F(Y/Z,X), the joint distribution F(X,Y,Z) could be computed from the above equation

As a first step, let us suppose that X is a *k*-dimensional variable, furthermore, let us suppose that it is divided into *k*-dimensional cells *k*(X) small enough for the distribution of Y to be essentially independent of X within each cell [9] Then within each cell

$$F(Y/X) \simeq F(Y)$$

and

$$F(Y/Z,X) \simeq F(Y/X)$$

so

$$F(X,Y,Z) \simeq F(X) \cdot F(Z/X) \cdot F(Y)$$

As the information needed to estimate F(Z/X) is in the original data file, the joint distribution F(X,Y,Z) can be estimated

PROBLEMS OF APPLICATION

In theory, it is possible to calculate F(X,Y,Z) from knowledge about the joint distributions F(X,Y) and F(X,Z), but clearly there may be problems applying this technique First, a means must be derived for determining when the cells *k*(X) are small enough to consider Y independent of X, within cells Such a sieve need not be of uniform dimension with X In fact, it would be better if the dimensions varied to reflect the density of the data The more dense the data within cells, the better the estimate of the distribution F(Y/X) will be, however, the

[9] If the cells are densely filled, F(Y/X) = F(Y) can be estimated directly If the cells are sparsely populated, regression techniques can be used in conjunction with some smoothness assumptions

The more dense the data within cells, the better the estimate of the distribution F(Y/X) will be, however, the cells must be small enough to justify the assumption of independence between X and Y

cells must be small enough to justify the assumption of independence between X and Y

As the data will vary in density, this trade off can be handled by varying the cell size as conditions warrant Nonetheless, the sieve must be composed of cells in which X and Y are independent A means of testing this assumption is required

A first step in this direction has been proposed by Nancy and Richard Ruggles (*16*, pp 360-362) They suggest a chi-squared test of the hypothesis that the samples in a cell came from different universes, and a second (correlation) test to evaluate the relative importance of these differences before making necessary adjustments These tests do provide some basis for hypothesis testing, but I am unconvinced that they are reliable (*20*, p 397)

An alternative to testing the validity of the synthetic micro file would be to introduce restrictions into the matching process deliberately One could ignore some common variables in the two files (let us call these X′) during the creation of the synthetic file, and estimate the mean square error of the artificial matches for X_1 to their actual value This approach, however, would lead to inefficiencies in the actual matching process as some information instrumental in making the union would be deliberately sacrificed At this point, the Heisenberg Uncertainty Principle becomes a factor

For hypothesis testing to be meaningful, the matching cells must be densely occupied—otherwise the testing procedure cannot have statistical validity This condition might be expected in the core (or central portion) of a large matched file, but it would require a large file Even then, the fringe (or outer portions) of the file will be too scattered For example, suppose X_1 is individual incomes and Y_1 is tax liability At high values of X_1, the individuals will be too diverse to allow sufficient expansion of cell intervals along Y_1 to encompass a large number of wealthy individuals while maintaining the premise that Y_1 and X_1 are independent

Obviously, the fewer the outliers and the more dense the data, the more technically correct the finished synthetic file will be Thus, the more valid it is to "stack" this file with another—that is, to use this file as a basis for another synthetic file However, this should not imply that outliers are "bad" They represent valuable information, and if their absence implies a "better" file, it is only "better" in a statistical sense related to ease of matching when pyramiding artificial files A file that lacks outliers may be unrepresentative of the population and it may be quite misleading [10] One reason statisticians square the distance from the regression line to sample points is to give greater weight to "extraordinary" points, and it behooves the synthetic file builder to be aware of the informative value of outlying points

Yet outliers do pose a special problem for the type of matching technique proposed here Their existence implies that some regions in the sample are so sparsely populated that some cells will lack match records In practice, the samples—even for the largest files—will have cell-vacancy problems The practical solution is to combine some of the X variables, which thereby collapses some of the cells (*16*, p 357) The Brookings MERGE file, for example, has 74 nonincome classifications instead of the 1,000-plus first envisioned If these cells had not been telescoped, many original cells in the grid would have produced obvious mismatches or no match at all

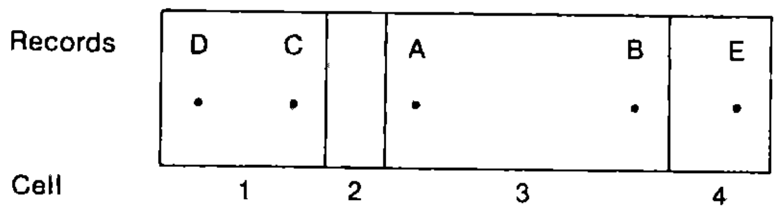
This difficulty is usually treated by using a metric technique, generally a distance function Distance functions rank possible matches by their "closeness" to the record to be matched, not by whether they occupy the same cell If a cell technique were strictly applied, only sample records sorted into the same cell would be linked This situation might lead to some matched cells that were unnecessarily diverse

Figure 3 depicts such a case for a single variable X_1 Sample A is a record for file (X,Z), and it is to be matched with a record from file (X,Y) The records B, C, D, and E are the leading candidates for matching What is the best match for A? The cell technique would signify B should be chosen, because it occupies cell 3 in common with A However, the distance function would rank B as the next to poorest choice of the four possibilities The "closest" record to A is C, and the distance criterion would indicate it should be chosen, regardless of the difference in cells Which is the best?

Following the premise that the cells were constructed such that the conditional probability of Y is not independent of X between cells, B is the best choice But if

[10] The problem occurs particularly with income and expenditure distributions, which are generally skewed

FIGURE 3
Individual Records in a Cell Structure



B did not exist and point E were the alternative to C and D, which should then be chosen? These are usual circumstances with a limited sample, and here the cell-matching technique breaks down completely Of course, if the cells have been chosen to reflect Y's independence of X, the question is one of minimizing damage Given that an ideal match is impossible, which is the best match? If one is willing to assume that the distribution will not be markedly asymmetrical, choosing the record that is closest or among those closest is a reasonable standard [11]

SPECIFIC EXAMPLES OF SYNTHETIC MICRO DATA FILES

Several existing synthetic files have used variants of the distance function concept Researchers creating the Canadian Survey of Consumer Finances-Family Expenditure Survey synthetic data file (SCF-FES) used multivariate analysis to determine their distance function ranking The variables Y_i and Z_i were regressed individually on all the variables X The explanatory power of the various X_i in the regressions on Y_i and Z_i was used to determine the variable's weight in the distance function If a variable X_j had high partial R-squares for a wide range of the Y_i and Z_i variables, it was considered a crucial element in the matching process Records from the two files that had similar values for that variable were awarded a relatively large number of points toward qualifying for matching A variable X_k having low partial R-squares for most Y_i and Z_i variables was considered relatively inconsequential to a good match It was assigned either a low or zero point contribution for the matching criteria Matches between the two files were little influenced by the correspondence of these variables Pairs of records with high match scores were linked (*1*)

The Brookings MERGE file uses a more *ad hoc* approach to the distance function The relative importance of the X variables in distinguishing a good match was predetermined on what were considered reasonable grounds rather than by quantitative analysis (*11*, *12*, and *14*)

The Bureau of Economic Analysis (BEA), U S Department of Commerce, developed its synthetic file, a matching of the Current Population Survey and the Tax Model for the year 1964, in a similar manner to the Brookings effort However, instead of using the Brookings technique of sampling for matches, the BEA file involved a one-to-one match between the two files Each tax record was assigned to a unique population record

[11] It does, however, entail an implicit assumption of independence between the Y and Z, given X See (*21*, p 343)

The final match between two records (from different files) is determined by the closeness of the match between the corresponding X_i and by the preselected weights

by matching records having the same rank order within broad *a priori*-determined equivalence classes Each cell was defined so that it has the same weighted number of records from each file, which avoids the issue of improper population aggregates extrapolated from individual records [12]

The synthetic file built by Nancy and Richard Ruggles under the auspices of the National Bureau of Economic Research (NBER) uses a distance function that is less arbitrary in cutting across previously defined cell structures The Ruggles' success in adhering closely to the prescribed sampling structure is due chiefly to one advantage more data The files described earlier involved matching files on the order of 50,000 records each The NBER file matches the 1970 Public Use Sample with the Social Security Longitudinal Employer-Employee Data file, each has 2 million records These files are so large, in fact, that they not only permitted the use of a cell-structure technique but also eliminated explicit use of a distance function of the type employed by other researchers Metric calculations take up computer time, and one must consider efficiency when processing 2 million records A cell technique is not only theoretically more desirable, in this case it is practical [13]

THE WEIGHTING PROBLEM

The NBER synthetic file technique does not necessarily approach optimal efficiency The cell dimensions may well be larger than they should be As noted, this will be especially true in the fringe, or outer portions of the file Collapsing the cells in these portions of the file implies a distance function of a rudimentary sort It is important to recognize the significance of the various weights or, as is the usual case, points assigned to each variable X_i in the distance function The final match between two records (from different files) is determined by the closeness of the match between the corresponding X_i and by the preselected weights

Obviously, as concepts about which variables are most important to a "good" match change and as those decisions are reflected in a different weighting scheme, the synthetic file changes Certain records that would have been considered matches will no longer be considered satisfactory and will be dropped, and others that would have been considered unfit will now be linked

The importance of the weighting scheme finally adopted lies in its determination of the accuracy of the file Whether or not these weights are determined empirically, as Alter and the Ruggles did, or theoretically, as Okner did, there is considerable subjectivity in the final determination It is also true that the file will have relative strengths and weaknesses according to the use that is made of it The file is usually designed for general purposes, and the weights are chosen to provide a mean between conflicting goals Ideally, all common variables X are in correspondence before a match is determined, but generally, only the most basic variables will approximate each other For example, total income is generally such a basic variable, and records from two different files need to have very similar incomes to be eligible for matching However, the source of that income is less important, and greater margins of error for individual sources of income are consistent with a "good" match

The points awarded toward a match reflect an income source's importance and determine the trade off *vis-a-vis* other aspects of the X sector In a specific study of farm taxation, for example, better results would be obtained if farm income were emphasized as a basic variable in the matching process There would then be a greater likelihood that the (X,Y) and (X,Z) records would both have farm income, and the file quality for such a specific purpose would be improved Given the expense of the matching process, however, it is more reasonable to construct a multipurpose file and use it for specific tasks rather than to construct a special file for each research task

OTHER APPROACHES

Only one means of effecting file links—matching—has been considered although it is not the only process available Perhaps the most important alternative to

[12] This is known as the alignment problem For a discussion of this problem, see (*14*, p 88) For a discussion of the BEA synthetic file, see (*3* and *2*)

[13] (*16*, pp 370-371) For a general review of the NBER data file, also see (*17*)

Synthetic data files, when properly constructed, can provide more conclusive answers to policy questions than other more traditional approaches

matching is a regression technique.[14] One file can be used to define the functional relationship between the common variables (X) and the disjunct variables it contains (Z). This relationship can then be used to append estimated Z values to the information in the second file, using the X values in file (X,Y) as a basis for the imputation (*16*, p. 354). That is, Z = f(X) would be estimated from the first file, and the X values in file (X,Y) would be used as the basis for calculating (X,Y,f(X)).

There are, however, several deficiencies in this approach. Perhaps the most grievous of the econometric problems (regressions imply their existence) is equation specification. The relationship between X and Z is unlikely to be well known—if it were, there would be little need for the first data file—and this relationship is even less likely to be linear throughout the domain. Thus it is extremely unlikely that the true joint distribution (X,Y,Z) can be well approximated by (X,Y,f(X)) (*20*, p. 395). A second major problem might be multicollinearity. A complex set of economic information, such as budget outlays, usually has highly inter-related components. Separate estimates of each outlay would lead to inconsistent estimates of the aggregate.[15] Another likely problem is heteroskedasticity. Many econometric studies using cross-sectional data find a changing variance in the disturbance term (7, p. 214).

For these and other reasons, regression analysis seems an inappropriate alternative to matching. From a methodological standpoint, it is inferior because it fails to produce the original variance of the data set when the imputations are generated. The regressions always assign mean values, whereas a matching process reproduces the distribution of variables in the original set over repeated imputations (7, p. 214). However, the basis for choice is hardly as one-sided as has sometimes been claimed. The Ruggleses have espoused a matching process over a regression technique because

> for matching purposes no specific functional relationship need be determined in advance. Nonlinear relationships will automatically be handled as efficiently as linear relationships, without explicit recognition that the relationships are nonlinear (*16*, p. 354).

If this observation were pertinent, it would be sufficient cause to rely exclusively on matching techniques, however it misses the point. Under the simplest conditions, when Y and Z are independent, the two techniques give identical results (although the regression equation does not reproduce the original variance in the data set), but when there are interdependencies and nonlinearities, the two techniques differ. Sims has neatly summarized the problem.

> To justify a matching procedure one requires an assumption that the regression relation giving the conditional distribution of (say) X 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 (*20*, p. 395).

A matching technique does have greater flexibility than a regression, but the assumptions necessary for its success are more stringent.

CONCLUSIONS

Synthetic data files, when properly constructed, can provide more conclusive answers to policy questions than other more traditional approaches. However, in practice, several expedients are used which entail a loss in validity. Further, a micro data file cannot provide quick, rough approximations on exceedingly broad topics, summary tables still have their place for quick estimates and general guides. A microfile is more unwieldy and complicated, but—if well constructed—capable of precise calculations and it can be used for a range of topics. Although it would be possible to

[14] Other choices might be averaging or interpolation techniques.

[15] The Ruggleses consider this property to exemplify the superiority of a matching process, arguing that it is a simpler and more satisfactory way of transferring complete sets of budget information from observations in one sample to observations in another (*16*, p. 354). Admittedly, it does retain the integrity of each set of information, but it should be possible to improve on such a naive estimating approach. The regression analysis could be modified by including a constraint to produce consistent answers. See (7, pp. 155-159).

pyramid this process and to create synthetic files, the conditions necessary for this to be practical would prevent the file from being of more than academic interest

Some rough guidelines for construction of a synthetic file can be these If the functional relationships linking two of the three variable sets X, Y, and Z are well-known and the data are scattered, regression analysis would probably be superior to matching Otherwise, matching is the best strategy In matching, the more dense the data, the more closely one can approximate the conditions of the ideal scheme, therefore, larger data sets are preferred Of course, a researcher may find that the relevant variables are only available in small samples—a situation which prevents him from constructing a more accurate synthetic file

This may seem inconclusive The ambiguity stems from the nature of the matching problem As Sims has noted

> there is no way to avoid "subjective" use of economic theory in deciding when a match is bad In their (the researchers') 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 (*20*, p 397)

The problem may not be easily answered, but we need to solve it because synthetic data do offer the promise of a better understanding of economic relationships

In Earlier Issues

Studies of relationships between agriculture and the rest of the economy must continually weigh the conveniences of aggregation against losses of relevant detail At one extreme are simple models which treat all agriculture as one enterprise selling a single composite product But the diversity of conditions within agriculture generally forces us to frame price and production programs in terms of individual commodities Modern techniques of analysis, such as the input-output or "interindustry relations" approach of Leontief and the "linear programming" methods of Dantzig, Koopmans and others, are creating a demand for more accurate data   Electronic computers can handle the formidable calculations required for such studies, but the accuracy of the final results must depend on that of the basic data For this reason, agricultural economists should take an active interest in the interpretation, application, and further development, of the interindustry relations approach most recently exemplified by the Bureau of Labor Statistics study of the U S economy in 1947 As time goes on, we need to supplement the input-output approach with one that permits us to use, among other things, our knowledge of demand and supply curves for agricultural commodities Conceptually, this leads us into a very large system of simultaneous equations—a sort of "econometric map' of the agricultural economy in the framework of total economic activity Our single-equation demand analyses, and sub models of moderate complexity, would be as useful as ever But the over-all model would force upon us a keener awareness of the nature of the approximations we were making, and of the variables or sets of economic relationships that we were assuming constant

Karl A Fox
and Harry C Norcross
AER, Vol IV, No 1,
Jan 1952, pp 13 and 21

BIBLIOGRAPHY

(1) Alter, Horst E "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970 " *Annals of Economic and Social Measurement* 3 (Spring 1974) 373-394

(2) Budd, Edward C "Comments " *Annals of Economic and Social Measurement* 1 (Summer 1972) 349 354

(3) ___________ "The Creation of a Microdata File for Estimating the Size Distribution of Income " *Review of Income and Wealth 17* (December 1971) 317 334

(4) Diesing, Paul *Patterns of Discovery in the Social Sciences* Chicago Aldine-Atherton, 1971

(5) Georgescu Roegen, N "Economic Theorv and Agrarian Economics " *Oxford Economic Papers* 12 (Feb 1960) 1 40

(6) Goldsmith, Selma F "Appraisal of Basic Data Available for Constructing Income Size Distributions " In *Conference on Research in Income and Wealth,* Vol 13, pp 267-377 New York National Bureau of Economic Research, 1951

(7) Johnston, J *Econometric Methods* New York McGraw Hill, 1972

(8) Klein, Lawrence R *A Textbook of Econometrics* Evanston, Illinois Row, Peterson and Company, 1953

(9) Leontief, Wassily "Theoretical Assumptions and Nonobserved Facts " *The American Economic Review* 61 (Mar 1971) 1-7

(10) Morgenstern, Oskar *On the Accuracy of Economic Observations* Princeton Princeton University Press, 1973

(11) Okner, Benjamin A "Constructing a New Data Base from Existing Microdata Sets The 1966 MERGE File " *Annals of Economic and Social Measurement* 1 (Summer 1972) 325-342

(12) ___________ "Data Matching and Merging An Overview " *Annals of Economic and Social Measurement* 3 (Sept 1974) 347-352

(13) ___________ "Reply and Comments " *Annals of Economic and Social Measurement* 1 (Summer 1972) 359-362

(14) Pechman, Joseph A , and Benjamin A Okner *Who Bears the Tax Burden?* Washington, D C The Brookings Institution, 1974

(15) Peck, Jon K "Comments " *Annals of Economic and Social Measurement* 1 (Summer 1972) 347-348

(16) Ruggles, Nancy, and Richard Ruggles "A Strategy for Merging and Matching Microdata Sets " *Annals of Economic and Social Measurement* 3 (Spring 1974) 353-371

(17) Ruggles, Richard, and Nancy Ruggles "The Role of Microdata in the National Economic and Social Accounts " *Review of Income and Wealth* 21 (June 1975) 203-216

(18) Shoup, Carl S *Quantitative Research in Taxation and Government Expenditure* New York National Bureau of Economic Research, 1973

(19) Siegel, Sidney, *Nonparametric Statistics for the Behavioral Sciences* New York McGraw Hill, 1956

(20) Sims, Christopher A "Comment " *Annals of Economic and Social Measurement* 3 (Spring 1974) 395 397

(21) ___________ "Comments " *Annals of Economic and Social Measurement* 1 (Summer 1972) 343-345

(22) ___________ "Rejoinder " *Annals of Economic and Social Measurement 1 (Summer 1972) 355-357*

(23) Theil, Henry *Linear Aggregation of Economic Relations* Amsterdam North-Holland Publishing Co , 1954