ECONOMICS OF INFORMATION, STATISTICAL MATCHING METHODOLOGY AND POLICY MODELING*

by
*Mitsuo Ono***

Introduction

According to page 110 of the Philippine Statistical Development Program, 1978-82, (reference 1) one of action steps proposed to improve and expand data production within the Philippine Statistical System is to "maximize the utilization of the administrative records system of administrative/regulatory agencies". Another proposed strategy is to integrate related data files. This note provides possible directions on how these objectives can be further advanced using statistical matching techniques to merge data files for "equivalent" statistical analytical units. Although, examples of statistical matching and micro-simulation. applications cited herein are taken from research completed in the United States, basic ideas and applications of such statistical techniques are relevant here in view of ongoing studies of almost similar nature in the Philippines (references 2 and 3). This note also relates matching technology with policy modeling and the economics of information.

Economics of Information

One of the "classic" in this field is reference (4) in which Stigler reminds that "knowledge is power", i.e., information can be considered a factor of production (managerial returns) since both ignorance and misinformation can be costly. By training and application, the statistician is considered an information scientist. His job is to design statistical information systems whereby information

---

*Presented to the Second National Convention on Statistics, Philippine International Convention Center, 2-3 December, 1980.

**Consultant, Philippine Institute of Development Studies and University of Hawaii Research Corporation.

of a given quality needed for decision-making can be obtained and utilized in most cost-effective way. Key words are "quality, decision-making, and cost-effective".

The statistician, like an architect or an economist, is confronted with designing and analyzing a production function with multiple objectives (and multiple constraints). For example, the sampling statistician attempts to design a sampling survey which minimizes the mean-square error of targeted variables, given resource and time constraints. Analytical tools such as linear programming could be applied in this activity. Additional to targeting accuracy, other objectives include timeliness, flexibility, comprehensiveness and understandability. With given constraints, all of these objectives have tradeoffs. The statistician attempts to devise a plan to achieve the "best" cost-effective compromise among competing options. Thus, accuracy could be traded-off for timeliness using less rigorous methods to obtain "order of magnitude" estimates. Timely "guesstimates" although rough, would still be useful in providing policy-makers "ballpark" information upon which other pertinent information could be applied for final decision-making purposes. The basic task of a statistician then, is to design, produce and deliver the most cost-effective information that could be generated, given cost, people and time constraints. This information could range from the detailed highly accurate to the quick "order of magnitude" estimate, depending upon how the information will be used for decision-making.

A possible useful policy-analysis tool that could provide quick but adequate information could be a household micro-data computerized research file, containing both administrative record and household survey data covering various types of households involved in different government programs (reference 5). This merged file would include data obtained from a number of sources: surveys, censuses, administrative records, and so on, which would be linked using statistical matching techniques. The use of such computerized merged data files should enhance responsive policy research capabilities of statisticians as information scientists.

To do this well, the statistician will also have to be very familiar with problems faced by policymakers. He needs to understand policy

issue options, how information is needed and can be used to help policymakers make choices among issue options and the constraints involved in producing and using this information, Clearly, the statistician has a vital role in the policy decision-making process.

Policy Modeling

The preceding section keynoted the important role of the statistician (as an information scientist) in policy analysis and formation. (This section presents one of the analysis and formation.) This section presents one of the analytical tools currently being used by statisticians in integrating policy-use data. Statistical data are only means to an end, i.e., to provide policymakers a menu of available alternative choices and indication of the possible "best" choice among options, given resource and other constraints. The statistician attempts to reduce the risks of ignorance and misinformation (and bad choice) due to unavailable, defective and/or untimely data. In doing this, he formulates and manipulates policy models to integrate and test information for consistency and adequacy. As a model builder, he attempts to pull together bits and pieces of scattered information in a meaningful structure such that these data can be used to predict forthcoming events. The model-builder attempts to put some semblance of order in understanding the real world of chaotic and unpredictable state of human affairs. In doing this, he applies assumptions of rationality and behavioral patterns which are often questionable. How well his model "works" to predict future affairs will depend on how well his assumptions and methodology are operationally realistic to the particular society under study.

In this type of exercise, the conclusions derived from policy models are typically "magnitude of order" estimates. Results are based on "what if" Happenings assuming that underlying suppositions are realistic. These "what if – then" alternatives formulated under varying assumptions could then be used by the policy-maker to make decisions based upon their own insights, knowledge and observations regarding impact of probable future conditions. Any presentation of results to policy-makers should make explicit the assumptions and methods used in arriving at the conclusions. The

test of a good policy modeler is how he selects and uses the assumptions, data and the methodology built into the policy model.

Since depicting human activity in policy models is a highly complex undertaking, one of the key problems is how to identify the variables involved in measuring relationships which are strongly associated with each other. To do this well, it calls for the hierarchical disaggregation of household member activity, that is, to understand how household members interact with each other in social and economic activities, how these activities interact progressively at higher grouping levels, e.g., the community level, and so on up the locational pyramid. One example of this type of information need is that of estimating future employment requirements through the use of a macro input-output model. Typically, possible future changes in household activities of key household members, e.g., the wife, are not fully disaggregated in forecasting changes in family work and income patterns. This lack of disaggregation in analyzing labor force activity results in lower overall employment projections, as evidenced in the United States where aggregative employment estimates from its input-output model were understated because the labor force participation rates of wives (and subsequent economic impacts) were underestimated. The difficulty, of course in integrating micro-data in macro input/output tables is the need to compile individual-related data from various information sources, e.g., censuses, surveys, various administrative records, and so on. A key point here is the need for more disaggregated micro-level information for macro-level analysis. Query: How can we compile disaggregated micro-data obtained from data sources for macro policy model estimation? Matching methodology can be used as a possible vehicle for such integration, as outlined below.

Exact and Statistical Matching and Some Applications

Exact and statistical matching are not the same. Statistical matching attempts to approach exact matching of households, establishments, and individuals. In the case of exact (identical) matching, the same household covered in two or more data files (or over time in the same file) are identified, and available

data from the two files are retrieved and merged into a composite file. In this search process, there could be completely, partially matched, and unmatched households. The dual vital registration records study whereby birth events recorded in administrative records and household surveys were matched using the married name of the mother located in the sample Enumeration Districts (EDs) is an example of an exact matching study (reference 2). Another example is to assume that there is a data file from the Income and Expenditures Survey and another household data file from the Integrated Survey of Households. The identical matching process locates (using key household identifiers) and collates information on the same household which have been independently interviewed in the two surveys. In this regard, if the barangay geographic code (reference 6) was included in the data record files of the two surveys, the first computer match could be the barangay-code. This would narrow down the search to the 250 or more households typically located in barangay areas. Another computer operation could match age, sex, and other key characteristics of the household head, and which could be used to identify the same household. Another possible match could identify the number of persons in the household, and possibly some of the characteristics of household members. If resources permit, computer match could be further verified on a sample basis comparing the data obtained in the two surveys. The data obtained from the two surveys would then be reweighted (as necessary) and merged into a consolidated research file. These sample data obtained from the two surveys would be treated as those derived from a joint probability distribution. Essentially, information obtained from marginal probability distributions are being merged into a composite joint distribution from which samples are taken for analysis.

The statistical matching procedure follows the same principle but is less precise. It involves taking information about a household covered in one file and merging it with information obtained in another file for a "similar" or "equivalent" household type (instead of using the same household as in exact matching). For example, the health characteristics of members of a particular type of households can be merged with the educational characteristics of "similar"

members included in equivalent household types. The "equivalent" statistical household type is defined and implemented in the statistical match procedure as noted below depending upon the characteristics involved and the available information. For example, in matching tax reporting units with families, joint returns could be related with married couples living together. In turn, information on the size of wages/salary income could be used as a basis to link "equivalent" income groupings. The attributes and measures used in the match could also cover demographic, social and economic micro-characteristics associated with the household head, household members, and the location of the household. One type of a matching process could operate as follows:

a) Set the number of strata and the characteristics variable-mix obtained from the data files. Use chi-square analysis if needed to establish "equivalence".
b) Rank the strata according to successively greater differences in the variable-mix for each of the two data sets. For example, a mix of age, sex, educational attainment level data could be used to form a stratum and those could be ranked linearly for the two data sets.
c) Scan the two files such that for each stratum in one file, there is an "equivalent" stratum selected from the other file. The "gate" opens when a person is found from the second file starta which comes closest to the first file stratum being matched.

The complexity of the statistical matching algorithm depends upon the availability of identifier information, the nature of problem to be analyzed by the matched study the desired level of accuracy, and the extent of available resources. It may be possible that because the "gate" was not wide enough, persons from the first file stratum may not be matched. Depending on the problem to be analyzed, they could be left unmatched or further scanned with wider "gates" for the second file stratum.

In summary, statistical matching is a cost-saving stop-gap for identical matching. In statistical matching, analytical units are first identified, e.g., tax reporting units or household or families. These

are further partitioned into sub-analytical units, e.g., taxpayer, household head, family members, etc. In turn, there units are related with certain attributes, whether geographical (location), residential (housing), demographic (age, sex, and so on, of members), social (marital status, educational attainment etc.), health (disabled, health problems, etc.) and so on. Each of these attributes have categorical and/or numerical values. For example, for income, household members may or may not have received certain types of income receipts. In turn, for those receiving a certain income type, they could report certain net amounts received during the designated time period. These attributes and magnitudes could be statistically associated with particular types of behavioral relations, e.g., work activity of household members could be related with age, sex, educational attainment, and other related variables. These linkages could be formulated using both objective or subjective measures (see reference 7 and 8). In the matching process, "equivalent" units are matched.

Statistical matching has many potential uses. For estimation purposes, it can be used to impute missing information in surveys and censuses. Thus, at the U.S. Census Bureau, missing income information is imputed from "equivalent" persons fully reporting their income information. In this procedure, reported income of members of households are "given" to persons with similar characteristics but not reporting income (on an item basis). Characteristics used relate to age, race, occupation, hours worked, and sex and family member designation (see reference 9).

Another example of how statistical matching is used is to link administrative tax unit, obtained from tax reporting files against households covered in household sample surveys (see references 7 and 10). The basic idea here is to match-merge income information derived from the two sets of data files into a composite file. For example, as noted in reference 10, the U.S. Survey of Economic Opportunity did not include certain types of income data (e.g., capital gains) and these data were "lifted" from equivalent "tax unit" households. The "best" information was selected and combined using the following steps:

1) Estimate whether any members of the survey household would be expected to file a tax return.
2) For filers, imputed household survey tax units were created and actual tax returns (similar to the household survey tax unit) were randomly selected and tax data from these returns were merged with the SEO household survey information.

Approximately, 30,000 computerized matches were made in the computers. In essence, this procedure "simulated" the tax reporting procedure of households.

Statistical matching can also be used for micro-simulation policy modeling. Thus, we start-off with a data base including detailed data on individuals of families (decision units). There are attributes and magnitudes (of attributes). These could have locational (location and type of dwelling unit), demographic (age and sex of individuals) social and economic characteristics. Also, included could be problems concerning social and economic needs of decision units. Decision units interact directly or indirectly with each other through their micro-behavioral relations (forming events based on probability risk functions). These events can be time-related e.g., birth, death, marriages, etc. occuring over a year and projected annually. As in input/output analysis but disaggregated to decision unit levels, household members' market activity can enter as input variables and derived transfers and value added can be distributed as output variables. All of these interactions and events can be projected over time through Monte Carlo simulation where a randomly drawn probability is compared against a computed probability for the event to happen (and also to what degree). Derived or computed probabilities can be estimated in many ways depending upon the purpose of the results. As compared with the usual econometric model, the micro-simulation model operates at a more disaggregated level of analysis and at more frequent intervals. Hence, it is more expensive but more flexible and has a wider application. A basic reference to micro-simulation forecasting modeling is reference 11.

An example of how a micro-simulation model works is outlined below. A human development service micro-simulation model used to forecast the number of human service program participants would include the following components (reference 12):

- A population data base
- A demographic event simulation sub-model/data base
- An economic event simulation sub-model/data base
- A social functioning event sub-model/data base
- A macro-economic model component
- An input/output model component

The population data base provide the benchmark which starts off the simulation. The demographic event simulator advances the population data base for one year by simulating births, deaths, marriages, separations, and other demographic events occuring over the one-year period. For example, the probability of birth could be related to the household income, age, education, employment status of wife, size of family and so on. The probability of death could be related to income, sex, age, education, health status, and other associated characteristics. The economic event sub-model simulates economic activities of family members. For example, the DYNASIM economic event module (reference 11) simulates changes in earning rates, etc. Thus, the probability of labor force participation could be related to last year's participation status of family members, age, sex, marital status, presence of children under six years of age, and the estimated unemployment rate derived from the macro-economic model component.

The macro-economic component model projects macro activity such as the unemployment and under-employment rates, over-all GNP, personal income levels, and so on. The social functioning event simulation model identifies socio-economic problems that families and individuals encounter, what service programs are involved and the results of the application of service program. This module determines whether family members are eligible for program services, whether the services were received, what type, and outcomes. These data are usually obtained from administrative record information. Finally, the input/ output component allows users to input different sets of data tailored to policy questions being asked. The output portion is used to extract summary results that are easily computed and readily available. In all of this work, statistical matching is used to integrate micro-data files for individuals and family units.

Possible Developmental Efforts

The assumption here is that only about twenty five percent of data collected in any statistical system are utilized cost-effectively. Administrative record systems appear to be less fully used than household survey systems. What needs to be done is to raise the productivity of data utilization by integrating more statistical data files, possibly using statistical matching techniques. It is suggested that for the first stage analysis of this type data linkage, an inventory of identifiers and all micro-data available in administrative record and household survey information be completed. This effort should identify what needs to be administratively done to improve the future linking of micro-data for policy-modeling efforts. Also included in this study would be to investigate how the available barangay code number can be better utilized for geo-coding and data linkage purposes. Further efforts could also be made to conduct small experiments to test the feasibility of formulating and implementing sectorial and regional prototype micro-simulation forecasting models.

REFERENCES

1. *Philippine Statistical Development Program,* 1978-82. NEDA, Manila 1978.
2. *Development and Maintenance of a Sample Vital Registration System in the Philippines,* by Tito Mijares, NCSO, 1974.
3. "Splicing of Data Sets and the Construction of a Baseline Income Distribution" by Bruno Barros, DEPP Working Paper, April 1980.
4. "The Economic of Information" by George Stigler, *Journal of Political* Economy, June, 1961, pp. 213-225.
5. "Integration of Administration Record and Household Survey Data for the Social Service Transaction Account" by M. Ono, *Proceedings of Social Statistics Section,* American Statistical Association, 1978.
6. *Philippine Standard Geographic Code,* NEDA, Manila, 1978.
7. "Size Distribution of Family Personal Income: Methodology and Estimates for 1964" by Edward Budd, et. al., U.S. *Bureau of Economic Analysis Staff Paper No. 21, June 1973.*
8. "Application of Bayesian Estimation in Social Science Research", by Mariano Garcia, *Proceedings of First National Convention of Statistics,* December 4-5, 1978.

9. Consumer Income, *Current Population Reports,* Series P-60, N. 118, U.S. Bureau of Census, pp. 271-273.
10. "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File" by Benjamin Okner, *Annals of Economic and Social Measurement,* July 1972, pp. 325-342.
11. *Policy Exploration through Microanalytic Simulation,* by G. Orcutt, S. Caldwell, and R. Wetheimer, Urban Institute, Washington, DC., 1976.
12. "Microsimulation Forecasting Model for Human Development Services Programs" bt George Caldwell, forthcoming, *Proceedings of Social Statistics Section,* American Statistical Association, 1980.