MICROANALYTIC SIMULATION MODELS

Robert Michielutte, Wake Forest University - J. Timothy Sprehe, Agency for International Development

Introduction

In order to place our discussion of microanalytic simulation models in proper perspective, it is helpful to make explicit a definition of simulation which best fits the models to be discussed.

In its broadest sense, simulation is "essentially a technique that involves setting up a model of a real situation and then performing experiments on the model" (Naylor, et al., 1966). The types of simulation we are concerned with, however, are more restricted in scope. They are all computer simulations, based on models that have been developed through some combination of logical and statistical information, which attempt to describe the functioning of specific systems over time. Naylor and others (1966) have suggested a working definition of simulation, which explicitly incorporates these characteristics. It is as follows:

> Simulation is a numerical technique for conducting experiments on a digital computer, which involves certain types of mathematical and logical models that describe the behavior of a business or economic system (or some component thereof) over extended periods of real time.

We would extend this definition to include any system or subsystem of interest to the social scientist that is amenable to the development of logical and mathematical models.

Within the framework of this definition, it seems essential, both in designing and evaluating computer simulation models, to ask the question: why do we simulate in the first place? A number of justifications for the use of simulation techniques have been suggested:

> 1) It may be either impossible or too costly to observe or experiment with certain processes in the real world. It is, for example, impossible to control the birth rate of even a small sample of women for the purposes of observing its impact on other processes and of formulating hypotheses.
>
> 2) The observed system may be too complex to describe mathematically in such a way that analytic solutions could be obtained.
>
> 3) In some situations, it is possible to develop a mathematical model, but it cannot be solved by straightforward analytic methods.
>
> 4) It may be possible to derive a set of mathematical equations which can be solved to describe a system, but validating experiments to test the model cannot be performed. This differs from the first problem only in the manner in which the simulated data would be utilized. "In the first case simulated data are necessary to formulate hypotheses whereas in the latter case simulated data are required to test hypotheses" (Naylor et al., 1966: 7).

It would appear that the basic rationale for the simulation models to be discussed here lies in points one and four. In attempting to answer questions concerned with the future distribution of income, hospital utilization, or occupational prestige, one approach is to develop simulation models which will allow the researcher to project current trends and to conduct hypothetical experiments on the simulated populations.

In this brief discussion, we propose to examine the methods and objectives of four simulation projects--those of the Urban Institute, The Research Triangle Institute, USC's Human Resources Center, and our own--in the light of some general facing simulations of this type.

The Simulation Models

I. The Urban Institute has undertaken the construction of a microanalytic model for the simulation of income activity at the family level (Orcutt et al., 1970). The model provides a dynamic description of the processes of birth, death, and changes in marital status. In addition, it has been designed to simulate education as a first step in the simulation of income variables.

The present stage of development operates on what the authors term a "persons" model, with a "family" model projected for the future. This distinction appears to be analogous to the open and closed populations discussed in the other models. Basically, the distinction involves the grouping and identification of individuals in the simulated population. In the "family" or "closed" model, marriage mates are selected from within the population and the households/families are treated as a unit. In the "person" or "open" model, marriage mates are in a sense hypothetical and exist only as information on an individual's record (one in the smulation population). In addition, no attempt is made to keep household the family members together. The relative merits of the two models will be discussed in a later section.

One particularly interesting feature of the Urban Institute simulation is the fact that an explicitly defined macro-model is incorporated as a means of obtaining closure in a simulation run. The authors recognize the fact that a limited model based on person and families cannot present a picture of an entire economic system. Thus the macro-model is utilized to supply parameters such as total employment, price level changes, and GNP information to the micro-model.

The use of a macro model such as this one is not unique. In fact, all the models under consideration supply limiting parameters to the simulations in varying degrees. The Urban Institute model, however, makes the supplementary model most explicit and, by doing so, focuses attention on an important principle. No microanalytic simulation can be expected to represent an entire system or subsystem. Certain types of external information must be provided to the model before either projections or experiments can be run.

II. The POPSIM model developed at the Research Triangle Institute is a microanalytic model which, in its basic form, will simulate the basic demographic processes of birth, death, and changes in marital status (Horvitz et al., 1970). The simulated population generated from this model can then be used as input for more specific simulations.

One valuable feature of the basic POPSIM model is its ability to generate an input population for the purposes of simulation. Therefore, the POPSIM user is afforded the capability of generating a population with characteristics suitable to the user's research interests. For example, one could generate a population which is representative, on a number of measures, of the entire United States population at a given time, or one can work with a more specialized population, say, males aged 20 to 45. The major limitation, of course, is the availability of suitable aggregated tables which contain the information desired.

The version of POPSIM that is currently being utilized appears to be an open (person) model where no attempt is made to keep families intact. The record for each individual in the simulation population contains information on the individual, secondary individuals (children), and family characteristics such as family income.

POPSIM has been applied to the problem of hospital utilization in the United States. The basic model was modified to include the variables of race, residence, and family income for the purpose of simulating hospital usage. In this application, inputs (a simulated population)from POPSIM were introduced into a hospital-episodes (HOSPEP) model which simulates admission, duration of stay, bedsize of hospital, insurance status and residence for each individual in the population. The purpose has been to estimate the effects of changes in population composition on utilization of hospitals, and correspondingly, the magnitude and kinds of hospital manpower which would be required under various projected conditions.

III. The simulation model in the process of development at the University of Southern California's Human Resources Research Center (Yett, et al., 1970) has been aimed at creating an analytical model which could cope with the national health-care crisis. The investigators intend that their ideal model, the Mark I, could be used for studying the structure of the health care system, for forecasting the effects of demographic changes on the system, for policy analysis and evaluation, and as a guide for research on the health care system.

The Mark I model would consist of three interacting modules: health services, health manpower, and health education. Technical problems and data deficiencies, however, have led the authors to conclude that the Mark I could not be developed in the near future. Thus, a simplified model called the Mark II has been proposed. This model is based on an open (persons) population rather than a closed model. The notion of an open population can be conceptualized in exactly the same manner as for the previous models. The closed population, however, means something slightly different in this case. The other models define a closed population as one in which marriage mates are selected from within the population and individuals in households are treated as a unit. A closed population in this context would be one which matches specific individuals to specific medical institutions.

Since the model is fundamentally concerned with simulating health care, less emphasis has been placed on the demographic processes than in the case of the other models. Births, deaths, and changes in marital status are treated as exogenous events in the model and are generally based on few variables. Death, for example, is treated as a function of age only. Since an accurate simulation of changes in health services, health manpower, and health education is dependent upon accurately projecting the population, it may be that this model does not devote enough attention to these basic processes.

One additional feature of the Mark II is of particular interest. All of the other models under consideration are based on what is termed a "files" approach by Yett and his associates. Data based on this approach are stored internally by individual records. The Mark II, on the other hand, is based on a "cell" approach. Data based on this approach are stored internally as contingency tables which tabulate the number of individuals with particular combinations of characteristics (e.g. age, sex, race). As long as the cross-tabulations are by all variables included in the simulation, the amount of information retained is exactly the same for both approaches.

In general, the superiority of one approach over the other appears to be based on the computer storage required and the number of decision making operations required in the course of the simulation. Briefly stated, it is suggested that a situation with many variables (age, sex, race, education, etc.) and few cases is best examined

by a files approach, while a situation with many cases and few levels is most efficiently handled by the cell approach.

It would appear, however, that another factor must be entered into the choice of best method. If the researcher has developed the simulation model in such a way as to require entry simultaneously of all input information on all individuals in the simulation population, storage space will become a problem regardless of which approach is selected. Twenty variables for 10,000 individuals would require 200,000 words for the files approach and much more space for the cell approach. Either approach will far exceed the capacity of most computers.

However, if the input data are stored on tape or disks and one individual at a time entered into memory and operated upon, the type of approach becomes almost irrelevant. Memory storage becomes more a matter of the length of the program, the operations to be performed, and the output desired. In general, it would appear that the files approach is most appropriate for microanalytic simulations that operate in this manner.

In fact, most simulations that would use a cell approach are generally macroanalytic as opposed to microanalytic models. The exact manner in which the simulation under development by the Human Resources Institute will utilize the cell approach is not completely clear. It may be, however, that the Mark II could be more appropriately termed a macroanalytic simulation. The Mark I, on the other hand, proposes to use a files approach and is clearly microanalytic in nature.

IV. The fourth computer simulation model to be considered is one we have developed within the framework of social accounting (Michielutte and Sprehe, 1970). The model represents an initial step in the development of a set of "social indicators" which could be used to monitor changes in American society.

The computer simulation program that has been developed is based on a microanalytic model which will simulate the basic demographic processes of birth, death, marriage and divorce. In a limited way, the model also incorporates international migration into the United States by adding in a sample of migrants each simulation year. In addition, the program in its present form will simulate changes in the educational and occupational structure (males only) of the United States population.

The simulation model deals with individuals as they pass through a life-cycle. Events occur within households principally on the basis of where individuals are located in relation to their time in life and accompanying circumstances. A set of data records representing characteristics of individuals in households in passed through the computer simulation program. Each pass represents a time period of one year and the simulation is repeated over many passes to effect the lapse of a number of years. At the end of each simulated year, a census of the population is taken, and the basic demographic changes in the population monitored. The simulated population is also stored on magnetic tape and becomes the input for the next year of simulation.

The data base for the simulation is a 1/10,000 census sample of the United States population. The population is defined as being closed, which means that marriage mates are selected and matched from within the population, and households are kept intact as units of analysis.

The simulation model can be used for both projection and experimentation. Basic twenty-year projections have been run by estimating probabilities for birth, death, and marital status, and by projecting these probabilities according to part trends. Hypothetical experiments have also been conducted by altering probabilities (e.g. probability of birth for women with two or more children) and observing the consequences for population growth.

Basic output for the program in its current form includes population counts by age, race, and sex; current marital status by age and previous marital status; vital rates (birth and death) by age, race, and sex; educational distributions by age, race, and sex; and occupational prestige distributions for employed males by age and race. This output is presented for each simulation year. In addition, individual records can be an output if desired.

Problems in Microanalytic Simulation

Although this brief summary of each model does not present a full picture of the different approaches, it does provide the background for a discussion of some of the more important problems in microanalytic simulation.

Monte Carlo vs. Analytic Techniques. All of the simulation models under consideration use or propose to use Monte Carlo techniques in varying degrees. Briefly stated, this means that in a dynamic simulation model, the occurence of events (e.g. death, hospital admission) is controlled by (1) finding a probability value which represents the likelihood of the event's occurring, (2) comparing this probability with a random number that has been generated, and (3) depending on the comparison, causing the event either to occur or not to occur.

Monte Carlo methods, while extremely valuable, have at least two major limitations. First they represent a stochastic process and thus are subject to sampling error. This means that one must either conduct repeated simulations to determine the degree of sampling error, or have

a sample large enough to be assured of very small errors. Even in the latter case, repeated simulations are desirable.

Secondly, Monte Carlo methods are extremely time-consuming and thus expensive. In our own simulation, for example, it requires approximately five minutes of central processor time to simulate one year for an initial sample of approximately 17,000 persons. The computer is a very fast CDC 6400.

Essentially, Monte Carlo simulation is a brute force technique and should be recognized as such. Hammersley and Handscomb (1964) suggest as a general principle what wherever analytic techniques can be employed in place of Monte Carlo techniques, the analytic techniques are preferable. They suggest, in effect, that the researcher interested in simulations of the type under consideration be constantly alert for ways of doing thing analytically.

We suspect that all of the models previously discussed, our own included, could profit from the advice. From the information available to us, it appears that the model of income distribution under development at the Urban Institute, the POPSIM model, and our own all make relatively heavy use of Monte Carlo methods, while the Mark II model of hospital care will not rely as heavily on these techniques.

Each of the models makes use of analytic techniques at various stages in the simulation. The decision as to whether an analytic procedure can profitably be substituted for a Monte Carlo technique will depend on the particular variable being simulated and the output desired. However, we feel that all simulation models should present clear justification for the use of Monte Carlo methods, which means an explicit statement of why analytic procedures are not applicable.

In the case of our own model, the inefficiencies of the Monte Carlo methods led us to employ more analytic methods in the simulation of occupation. For this routine, occupation was defined as a function of race, education, marital status, age and the interaction between these variables. Occupational prestige is assigned to every eligible male by means of a regression equation which predicts a mean score for males with various combinations of characteristics. This procedure at present still has many flaws, but its application resulted in a considerable savings in time, and no loss of information for our particular problem.

Open vs. Closed Populations. Another problem related to the long running times associated with Monte Carlo simulations is the question of whether to simulate on a basis of an open or closed population. Three of the simulations discussed, the Urban Institute model of income distribution, the Human Resources Research Center model of Health Care, and the Research Triangle Institute's POPSIM are currently based on open populations. Our simulation, on the other hand, at present operates on a closed population.

Based on our experience with the closed population, we would suggest that the open population approach is superior for most purposes. The closed population has a number of disadvantages which hamper the efficiency of the simulation. In our closed population, it is necessary to match marriage partners from within the population and treat households as units for a number of decision-making processes. In the event of death, for example, a series of decisions must be made with respect to re-organizing the family and choosing a new head of household if the previous head is the one who dies. In the case of divorce, decisions must be made with respect to the children and to the creation of new households. Essentially, the problem becomes one of increasing effort for diminishing returns. A great deal of additional decision-making and programming is required which contributes nothing to the desired output, but which must be done in order to allow more important parts of the simulation to operate.

The increased number of decision steps and simple clean-up operations necessary to maintain intact households adds to both the running time of a simulation and the amount of memory required for the program.

A further limitation is the increased data requirements for the closed model. In the model of health care, for example, the authors point out that the use of a closed model would require knowledge of the attributes of individuals employed in each health occupation, by each type of institution. In our model, use of the closed population requires knowledge of relationships between variables for which data could not be obtained (e.g. are children institutionalized if both parents die?)

Estimation Procedures. The problem of unavailable data is one which affects every simulation model. There is nothing inherent in simulations, however, that make the data problems a particular failing of the technique itself. It simply reflects the fact that simulations tend to be complex and include a relatively large number of interacting variables. The simple addition of even one variable into the simulation problem geometrically increases the data required.

Ideally, a simulation model would be able to specify the exact relationship between every variable included in the simulation. This is nearly impossible in even a moderately complex simulation and the limiting assumptions must be explicitly recognized when interpreting the output.

It is important to emphasize that the quality of

the data, with respect to both accuracy and complexity, determines what output is feasible. If, for example, one simulates the occurrence of death on the basis of age and sex, death rates computed by age, race, and sex would be accurate only if there were no relationship between death rates and race. In general, any time a researcher excludes a variable such as race from one part of the simulation and then produces output which includes that variable, he is implicitly assuming that no relationship exists or that it is so small as to have little effect.

The four simulation models vary considerably with respect to the complexity of estimation procedures. The Urban Institute model has defined death as a function of year (trends over time), age, race, sex marital status, education, and parity. The POPSIM model bases mortality on year, age, race, sex, and marital status. Our simulation model defines death as a function of year, age, race, and sex. Finally, the model of health care appears to base death only on age.

The implications of omitting marital status and education from the simulation of mortality in our model extend beyond the results pertaining to death. We could (and do) have the output of death rates by age, race, and sex with a relatively high degree of accuracy; no attempt is made to examine death rates by marital status. However, if marital status and death are related, then the results obtained for the simulation of marital status are likely to be somewhat inaccurate. Since the relationship tends to be in the direction of lower death rates for married people, omission of marital status from the mortality routine will result in too many married people dying in the course of a simulation year. This will introduce some error into the results for percent married, etc.

Accuracy of the Models. In general, surprisingly little mention is made of the manner in which the simulation models under consideration will be tested. The process of model verification requires data collection for at least two points in time. The first set of data will be used either as the input population or as a base to generate the input population. The simulation results should then be compared to data obtained at later points in time in order to obtain some estimate of the accuracy with which the different processes have been simulated.

All of the models discussed propose to simulate events based on real data and which have some relevance for policy decisions in future years. Since this is the case, confidence in the results of projections and hypothetical experiments must be based on the results of earlier verification procedures.

A Strategy for Simulation

An important direction for further development of microanalytic simulation models would seem to be with respect to causal models. Implicit in the models discussed here are causal relationships between many of the variables in a given system. Although in many cases specification of the causal links would be different or uncertain, such a development would aid in the evaluation and improvement of existing models.

Once simulation models are viewed within the framework of a set of causally related variables, the possibilities for development of expanded models should become apparent. The models viewed in this paper tend to focus on the simulation of variables which are basically demographic in nature. There is no reason, however, why simulation models could not also include measures of attitudes and values. This has already been done to some extent in the Urban Institute model. Birth is defined partially as a function of contraceptive usage in the simulation.

The first step in the development of a simulation model would be to hypothesize a set of causal links between the variables to be included in the simulation. If the causal scheme could be expressed in terms of a set of regression equations, one could then start with an initial population and simulate the consequences of the particular causal structure. Processes such a mortality would continue to operate and each individual in the population would be simulated through the set of equations. Ideally, the basic demographic events such as mortality would be incorporated into the causal model. For example, one would be able to specify the effect of a father's death on the son's educational chances.

The difficulties in implementing this approach have not been underestimated. It is likely that the ideal situation outlined here simply will not be possible for some time. However, computer simulation models are frequently accused of being atheoretical and having little relationship to the real world. The approach suggested here would provide both a theoretical orientation to the simulation and aid in interpretation of the results.

References:

1964 - Hammersley, J.M. and D.C. Handscomb, Monte Carlo Methods, New York: John Wiley & Sons.

1970 - Horvitz, D.C., et al., "Simulation of Hospital Utilization." paper presented to the Conference on a Health Manpower Simulation Model.

1970 - Michielutte, Robert and J. Timothy Sprehe, Simulation of Large-Scale Social Mobility: Toward the Development of a System of Social Accounts. Final Report: National Science Foundation Grant #GS-2311.

1966 Naylor, Thomas H., et al., Computer Simulation Techniques. New York: John Wiley & Sons.

1970 - Orcutt, Guy H. et al., "Some Demographic Results of a Dynamic Simulation." mimeographed paper.

1970 - Yett, Donald, et al., The Development of a Microsimulation Model of Health Manpower Demand and Supply. Final Report: U.S. Public Health Service Grant #PH-108-69-69.

Note: Work on this paper and on the Michielutte - Sprehe simulation model was carried out under NSF Grant No. GS-2311.