

The Role of Microsimulation in Longitudinal Data Analysis

Douglas A. Wolf
Center for Policy Research
Syracuse University
New York, USA

Abstract:

Microsimulation is well known as a tool for static analysis of tax and transfer policies, for the generation of programmatic cost estimates, and dynamic analyses of socio-economic and demographic systems. However, microsimulation also has the potential to contribute to longitudinal data analysis in several ways, including extending the range of outputs generated by a model, addressing several defective-data problems, and serving as a vehicle for missing-data imputation. This paper discusses microsimulation procedures suitable for several commonly-used statistical models applied to longitudinal data. It also addresses the unique role that can be played by microsimulation in longitudinal data analysis, and the problem of accounting for the several sources of variability associated with microsimulation procedures.

Résumé

La micro-simulation est bien connu comme un outil pour l'analyse statique des politiques d'impôts et de transfert, pour l'estimation des coûts des programmes, et pour l'analyse dynamique des systèmes socio-économiques et démographiques. Cependant, le micro-simulation offre également des possibilités intéressantes pour l'analyse de données longitudinale, y compris le développement d'autres produits du modèle, la possibilité d'adresser des problèmes de données déficientes, et comme véhicule pour l'imputation de données manquantes. Cet article discute des procédures de micro-simulation appropriées à plusieurs modèles statistiques qui utilisent des données longitudinales. Il adresse également le rôle unique qui peut être joué par la micro-simulation dans l'analyse de données longitudinale, et le problème posé par les multiples sources de variabilité dans les procédures de micro-simulation.

Key Words: Longitudinal data analysis, dynamic models, event history analysis, microsimulation, Imputation

Microsimulation Defined

The term “microsimulation” encompasses a variety of methodological tools and techniques that are finding growing use in empirical social science applications. The growth in the number and variety of such applications makes the task of organizing and summarizing the field a great challenge. This paper does not attempt to provide an overview of those applications; a recent book by Gilbert and Troitzsch (1999) does an excellent job of that. Rather, the emphasis is on ways that microsimulation can serve the needs of the data *analyst* – for which we might substitute the term ‘model builder’ – rather than the model *user*. Furthermore, the focus is on longitudinal rather than cross-sectional data analysis.

Microsimulation can be described as a collection of tools that facilitate a particular approach to working with a model. The essence of that approach is (1) the use of randomization in the assignment of values to the units studied – i.e. in ‘prediction’ – and (2) the use of individual units of analysis. This description does not, admittedly, go far towards isolating a recognizable set of

analytical tools. One textbook states that simulation is a way of "... driving [a] model with certain (typically random) inputs and observing the corresponding outputs" (Bratley, Fox and Schrage 1987:2). Moreover, while rather obvious, it bears stating that microsimulation presupposes the existence of a model, as well as the availability of specific values for all its parameters, even if those values are considered 'provisional' or 'interim.' Thus the specification of a model must precede microsimulation, and parameter values must be obtained either by statistical estimation or other means (including assumption, borrowing from other sources, or pure guesswork).

In view of the preceding paragraph, a suggested definition for microsimulation relevant for social science applications is the following: *microsimulation consists of drawing a sample of realizations of a prespecified stochastic process.* Microsimulation thus entails the generation of data (a set of realizations). Again, the model (the prespecified stochastic process) must be known in advance. The generated data will look like 'real' data, and can, therefore, be analyzed and summarized just like real data, although I will argue below that additional and specialized techniques should be used to account for the uncertainty inherent in microsimulation.

The definition offered here is general enough to encompass a diverse set of empirical applications in the social sciences. Microsimulation is well known as a tool for static analysis of tax and transfer policies, and for the generation of cost estimates for proposed legislation (see, for example, Lewis and Michel 1989; Orcutt et al. 1986; or Haveman and Hollenbeck 1980). There are also several examples of efforts to develop large-scale dynamic models of socio-economic and demographic outcomes in multiple domains, such as births, deaths, marriages and divorces, education, labor force behavior, incomes, savings, retirement, health, and household arrangements. These include the DYNASIM (Orcutt et al. 1976; Zedlewski et al. 1990) and CORSIM (Caldwell 1999) projects in the U.S., Statistics Canada's DEMOGEN model (Wolfson 1989), the NEDYMAS (Nelissen 1995) in the Netherlands, and the models developed by the *Sonderforschungsbereich 3* group in Germany (e.g., Galler 1989) and at NATSEM in Australia (Harding 1993; King et al. 1999), among others. A common characteristic of these efforts is the incorporation of model elements from several non-overlapping data sources, drawn from different samples, possibly drawn at different times.

Microsimulation has received much attention from demographers, especially to study reproduction (e.g., Barrett 1971; Ridley and Sheps 1966) and the composition and evolution of kin groups (e.g., De Vos and Palloni 1989; Goldstein 1996; Ruggles 1987; Wachter 1987, 1997; Wolf 1988). The works just cited are narrower in scope than those cited in the preceding paragraph, since they simulate fewer outcomes (at most, birth, death, marriage and divorce).

However, in all the applications cited above, the emphasis is primarily on the simulation itself, and on the outputs generated by the simulation, rather than on the process of model *development, estimation, and assessment*. A goal of this paper is to argue that the microsimulation approach has much to offer in these prior steps of the modeling process, the steps that might precede the integration of disparate model elements into a large-scale, possibly policy-directed system-level application. At the same time, it is difficult to draw a line between research in which the model, rather than the simulation, is of primary emphasis.

Gilbert and Troitzsch (1999) draw a distinction between “statistical” models and ‘simulation’ models. The former consists of one or more mathematical expressions that include parameters, the numerical values of which are obtained through estimation based on empirical data. Assessment of a statistical model can depend, in part, on a comparison of the estimated model’s predictions with their real-world counterparts. In contrast, a simulation model may take the form of a computer program, and the output of the model might consist of artificial data; here, assessment of the simulation model might depend in part on a comparison of the simulated data to its real-world counterpart. Gilbert and Troitzsch (1999) include ‘microanalytic simulation models’ – more simply, ‘microsimulation’ – as a subtype of simulation. The present paper focuses on microsimulation *techniques*, which are necessarily used in, but distinct from, microsimulation *models*, and tries to point out the ways in which these microsimulation techniques can play a role in the development of what Gilbert and Troitzsch (1999) call ‘statistical models.’ Thus it attempts to link what might otherwise be viewed as quite different categories of modelling efforts.

With respect to the role of microsimulation in model development, two types of activity come immediately to mind, Monte Carlo investigation of the sampling distributions of various statistical estimators (Mooney 1997), and the recently developed simulated maximum-likelihood and method-of-moments estimators of high-dimensional latent-variable or discrete-choice models (e.g., McFadden 1989). While these techniques are of great importance, they will not be considered here. Also, by the above definition of microsimulation, the multiple-imputation approach to dealing with missing data (Rubin 1987) can be viewed as a type of microsimulation. Indeed, below I suggest that analytical results from the multiple imputation literature can be extended to deal with the several sources of uncertainty present in microsimulation output data.

There is a final distinctive way in which microsimulation differs from the ‘production’ of a model. Normally, a model construction involves the following sequence of steps: (1) specification, that is, identifying the fixed, variable, and parametric elements of the model, and the relationships among them, in some formal statement. (2) Itemizing assumptions, particularly those concerning the nature of any stochastic elements of the model. And, (3) obtaining statistical estimates of the parametric elements. In order to conduct microsimulation it is necessary to bring in a fourth element, namely the ‘baseline’ (in a static or cross-

sectional simulation) or the ‘initial conditions’ (in a dynamic, or longitudinal, simulation). The initial conditions can be chosen arbitrarily, and might represent a single ‘representative’ or otherwise interesting individual, or they may be an empirical representation of (i.e. a sample from) a large population.

In the following sections of this paper I provide specific examples of types of models that lend themselves to the microsimulation approach, discuss several ways that microsimulation can serve the data analyst, and suggest a few specific procedures to be incorporated into microsimulation exercises. These discussions are guided by a few basic principles, namely:

- microsimulation should be viewed as an exercise in *taking one’s model seriously*. That is to say, any assumptions that are imposed during the specification and estimation steps must, as well, be imposed in the microsimulation algorithm. And, if the microsimulation output produces a finding that is sharply at odds with known facts, then it is not adequate to ‘adjust’ (or ‘calibrate’) the microsimulation; rather, one must return to the model, prepared to respecify it and to reestimate its parameters (see also Klevmarken 1998, p. 22); and,
- microsimulation is fundamentally an exercise in *sampling*. Accordingly, it is important to worry about the sampling distribution of any microsimulation outputs. However, in contrast to the process of generating an empirical sample from a real and finite population, microsimulation can be viewed as the generation of a sample from a hypothetical but infinite population. Furthermore, in a microsimulation the model parameters are generally a sample from a sample space, the random numbers used in assigning simulated values are a sample from the infinite population of random numbers, and the initial conditions are often a sample from a real, finite population. Each is a distinct source of sampling variability. One can push this observation even further, noting that the random number generator, the model specification, and the microsimulation algorithm are all selected, albeit not at random, from sets of alternative choices. However, the latter sources of uncertainty or error will not be considered here.

Canonical Forms of Models and Appropriate Simulation Algorithms

There are many types of statistical models suitable for longitudinal data. Here I will list a few important classes of models that are particularly suitable for use in a microsimulation. In each case, one or more suggested simulation algorithms are also given.

Models of Duration and of Event Sequences

Duration (alternatively, ‘event-history,’ or ‘failure-time,’ or ‘survival-time’) models, have developed rapidly and achieved widespread application in recent decades. Under this heading I include only those models in which time is considered a continuous variable; discrete-time models are discussed below. In the simplest such model, interest focuses on the time elapsed from one specified event (e.g., becoming married, or being born) to the next event (e.g., becoming divorced, or widowed; dying). Closely related are models of the *number* of events occurring in a specified time interval. As pointed out by Klein and Moeschberger (1997), models of duration can be grouped into two classes: the accelerated failure-time model, and models of the hazard, or rate, of occurrence of an event.

A duration model often considers only a single random variable, that is, a single elapsed time between events, but that random variable is typically an element of a life-cycle process in which numerous events, of several types, can occur. In either case, it is common to formulate a model of duration with reference to the instantaneous rate of occurrence of an event (or, of leaving a state), i.e. the hazard. A general expression for a **multiplicative hazard model**, expressed in logarithmic form, is where i denotes an individual, j and k denote the last and the next event types, respectively, $f_{0jk}(t)$ is the (log) ‘baseline’ hazard (involving some parametric function of elapsed time, t , and possible ‘duration dependence’), and X_i is an array of predetermined (but possibly time-varying) variables. The term $g_{jk}(H_i)$ expresses the possible dependence of the current duration interval on any of several aspects of the history of the process.

$$\ln h_{ijk}(t|X_i, h_i, z_i) = f_{0jk}(t) + X_i B_{jk} + g_{jk}(H_i) + \delta_{jk} z_i \quad (1)$$

Heckman and Borjas (1980) identify several conceptually distinct types of this history-dependence, including lagged duration dependence (the dependence of the current hazard rate on the duration of a prior completed duration, in the same or some other state) and occurrence dependence (a count of the number of prior visits to the state, or an indicator that

some state has ever been visited). The term z_i represents ‘unmeasured heterogeneity’ – the combined influence of all relevant but unmeasured factors. Equation (1) represents the simplest special case, namely one in which unmeasured factors are person-specific but time-invariant. There may also exist spell-type specific, family- or other group-specific, or place-specific unmeasured heterogeneity. The existing literature includes numerous examples of these models, including some employing parametric random-effects ‘frailty’ models (e.g., Manton et al. 1986) and the multinomial mixture model developed

by Heckman and Singer (1984). In equation (1), B_{jk} and δ_{jk} are unknown parameters, while $f_{0jk}(\cdot)$ and $g_{jk}(\cdot)$ contain additional unknown parameters.

From the hazard function one can derive the conditional (on k) survival function, $S_{ijk}(t)$, which gives the probability that i will remain in state j for at least t time units prior to experiencing event k , or

$$S_{ijk}(t) = \exp\left(-\int_0^t h_{ijk}(x) dx\right) \quad (2)$$

and the density function for completed durations—the “failure density”,

$$f_{ijk}(t) = S_{ijk}(t) h_{ijk}(t) \quad (3)$$

If the random process being modeled is associated with only one type of event, then the j and k subscripts in equations (1) - (3) can be dropped, and (3) fully determines the sample paths of the process. If, however, multiple “causes of death (failure)” or destination states are explicitly represented, then one can decompose the probability that the time to the next state is T_{ij} and the next state entered is k as follows:

$$\begin{aligned} & \Pr(X_{n+1} = k, t_n = T | X_n = j) \\ &= \frac{d}{dT} \exp\left(-\sum_k \int_0^T h_{ijk}(w) dw\right) \frac{h_{ijk}(T)}{\sum_k h_{ijk}(T)} \end{aligned} \quad (4)$$

where X_n is the most recent event (the current state occupied), X_{n+1} the next state entered, and t_n the elapsed time between the most recent and the next events (see Berman, 1963; Wolf 1986).

Examples of applications of duration or event-history models in demography, economics, sociology and other social sciences are numerous. Virtually all demographic models employing life tables, whether of the simplest single-decrement type, or the more complex multiple-decrement, or even more complex ‘multistate’ types, can be viewed as special cases. In the simplest such cases hazards are treated as constant within time intervals but dependent on age, and measured and unmeasured covariates are disregarded. In more complex models, such as the semi-Markov marital-status dynamic models found in Ravanera, Rajulton and Burch (1993) or Wolf (1986), hazards depend on both

age and the time since the last event.

There are various ways that models based on continuous-time event hazards can be simulated. In a simple model with only one type of event (e.g, the time between the formation of governments in postwar Italy; the time between occurrences of crimes at a fixed location) it is sufficient to simulate times based on the survivor function, equation (2). In particular, one can

- (a) draw a random number z^* from the uniform $[0, 1]$ interval;
- (b) find t^* such that $S(t^*) = z^*$.

In practice, one often can obtain satisfactory results by finding integers $[t^*]$ and $[t^*]+1$ such that $S([t^*] + 1) \leq z^* \leq S([t^*])$ and interpolating between them. In a multistate model one can follow the preceding steps to find the time to next event, but using the “overall” survivor function [the first term on the right hand side of (4)]. One can then

- (c) draw a second random number y^* ;
- (d) divide the unit interval into K subintervals, representing the K possible destination states, $[0, h_{ij1}(t^*)], [h_{ij1}(t^*), h_{ij1}(t^*) + h_{ij2}(t^*)], \dots, [\sum_{k=1}^{K-1} h_{ijk}(t^*), \sum_{k=1}^K h_{ijk}(t^*)]$;
- (e) assign as j^* , the simulated next state entered, the index of the interval that contains y^* .

The latter algorithm is discussed more extensively in Wolf (1986).

An alternative to the multiplicative hazard model is the **accelerated failure time model**, expressed as

$$\ln T_{ij} = X_i B_j + \sigma_j \ln T_{0j} \quad (5)$$

where T_{ij} is the length of a type- j episode for individual i , X_i and B_j are as defined above, σ_j is a scaling factor, and T_{0j} is a random variable from a specified distribution (e.g., normal or gamma). Equation (5) models failure-(survival-) times directly, rather than indirectly through the hazard function as in equation (1). A person-specific random effect could presumably be added to the right hand side of equation (5). Examples of the usage of this model include Wolfson et al. (1990), who modeled survival after age 65 as a function of long-run average earnings, and Christofides and McKenna’s (1996) analysis of job tenure.

Microsimulation of the accelerated failure-time model is straightforward, requiring

- (a) drawing a random number from the specified distribution for T_{0j} ;
- (b) computing the implied value of $\ln T_{ij}$ given X_i and the estimated values of B_j and σ_j
- (c) exponentiating the result, thereby obtaining a simulated value T_{ij}^* .

This model also suggests an alternative algorithm for event-based simulation of a competing-risks model. In such a model there are J latent times, $T_{i1}^*, \dots, T_{iJ}^*$ each corresponding to the time until occurrence of event-type 1, \dots , J respectively, but what is observed is only T_{ij} , the minimum of the set of latent failure times (David and Moeschberger 1978). The other latent times are censored by the occurrence of failure due to event-type j . An algorithm for the simulation of this process simply repeats the simulation algorithm given above, for times $T_{i1}^*, \dots, T_{iJ}^*$, then chooses as the “observed” outcome the minimum of the set of simulated latent times. This approach avoids the need to evaluate several hazard functions.

Linear Models for Continuous Outcomes

When continuous outcomes for individuals are observed two or more times in panel data, analysts may model the sequence of outcomes using a generalization of the classical linear model,

$$y_{ijt} = \alpha_{ij} + X_{it} B + e_{ijt} \quad (6)$$

where i indexes individuals, j represents the j^{th} outcome, and α_{ij} is a person-specific and outcome-specific factor. The disturbance e_{ijt} may be a simple ‘pure noise’ factor, or may be generalized to exhibit serial correlation. One example of such a model is the longitudinal earnings model presented in Lillard and Willis (1978), in which the α_{ij} s are treated as normally-distributed random effects and the e_{ijt} exhibit first-order autocorrelation; for additional examples see Hsiao (1986). Microsimulation of models like (6) is straightforward depending on the distributions assumed for the person-specific factors and the disturbances. Given predetermined values for α_{ij} and X_i , one must (a) draw a random number e^* from the distribution of the disturbances, then (b) make appropriate substitutions into (6) to obtain y^* , the simulated value of y_{ijt} .

Models for Discrete-outcome Panel Data

A third class of models deals with a series of discrete-coded outcome variables, in the simplest case binary-coded (0, 1) variables. For such binary variables, a general probability expression for the observed outcome is

$$\Pr[Y_{ijt} = 1] = F(\alpha_{ij} + X_{ijt} B) \quad (7)$$

where Y_{ijt} is the observed outcome, F is a specified cumulative distribution function, and other notation is as defined above. The most common choices for F are the normal (i.e. the Probit model) and the logistic (i.e. the logit model). Alternatively (but equivalently) the outcomes can be viewed as generated by latent index functions. The logit and probit models can be derived as instances of utility-maximizing choices over a set of discrete alternatives, in which the utility to i of choice j at time t is given by

$$V_{ijt} = \alpha_{ij} + X_{ijt} B + u_{ijt} \quad (8)$$

where the u 's have independent Type I extreme-value distributions [that is, $F(u) = \exp(-\exp(-u))$] in the logistic case (see McFadden 1973) or a multivariate normal distribution in the Probit case (Hausman and Wise 1978). The model is completed with the assumption that the observed choice offers greater utility than the other available choices, i.e. $Y_{ijt} = 1$ if $V_{ijt} = \max[V_{i1t}, \dots, V_{iJt}]$. A related derivation for the binary Probit model views the linear index $X_{ijt}B$ as a 'stimulus,' an unobserved standard normal variate as a person-specific 'threshold,' e_{ijt} , and supposes that the outcome or 'response' is observed if the stimulus exceeds the threshold, or, equivalently, while $Y_{ijt} = 0$ otherwise (Finney 1971).

$$Y_{ijt} = 1 \quad \text{if } \alpha_{ij} + X_{ijt} B + e_{ijt} > 0 \quad (9)$$

In equations (7)-(9), individual-specific intercepts have been included. If a specified distribution is assumed for these intercepts, conditionally independent of X_{ijt} , a random-effects logit or Probit model results. The existing literature includes several alternative distributional assumptions for these random effects, particularly for the panel logit model, including the normal (Firth and Payne, 1999), gamma (Conaway 1990), uniform (Beggs 1988), binomial (Engberg et al. 1990; Zenger 1993), and the nonparametric discrete distribution suggested by Heckman and Singer (1984), which has been applied to discrete-response panel

data (e.g., Reader 1993). The logistic version of equation (7) has been widely used as a discrete-time duration (or event-history) model (Allison 1982), in many if not most cases without including a person-specific intercept (i.e., ignoring the possibility of unmeasured variables that persist from period to period).

There are numerous examples of empirical applications of the many varieties of models included in the above formulations. Pollard and Wu (1998) use the logistic discrete-time event-history framework (without unmeasured heterogeneity, i.e. person-specific intercepts) in their study of age at first marriage in Canada, while Ham and Rea (1987) use the logistic discrete-time event-history model with discrete unmeasured heterogeneity as suggested by Heckman and Singer (1984) in their study of the duration of unemployment in Canada. In the former study, unmarried persons contribute as many as 25 person-years of at-risk experience to the analysis, while in the latter they contribute as many as 260 person-weeks of exposure over a 5-year period.

The probability expression (7) and the latent-variable specifications (8) or (9) correspond to two different approaches to microsimulation for discrete-time discrete-outcome models. The first algorithm entails

- (a) computing the probabilities that $Y_{ijt} = 0$ or 1 (or, in a multinomial application, 0,1, ..., J), and then
- (b) drawing a random number z^* from the uniform [0, 1] distribution.

The simulated outcome Y_{ijt}^* is assigned as the value of the subinterval in which z^* falls (using subinterval definitions analogous to those discussed earlier for the competing-risks hazard model). Alternatively, one could

- (a) randomly select values $e_{i1t}^*, e_{i2t}^*, \dots$ from the appropriate random distribution;
- (b) compute the implied values V_{ijt}^* , and then
- (c) assign as j^* the maximum of the set $V_{i1t}^*, V_{i2t}^*, \dots$.

Pudney and Sutherland (1993) refer to these alternative algorithms as 'interval' and 'structural' approaches, respectively.

What Microsimulation Offers the Data Analyst

While microsimulation plays a role in some estimation techniques, and has proven to be of interest in the policy development and planning arenas, it also offers some potential advantages during the process of model development. Three areas in which microsimulation can make such a contribution are discussed below.

Extending the Range of Model Outputs

One advantage frequently noted with respect to microsimulation is its ability to produce estimates of the full distribution of an outcome, in addition to the expected value that can be produced analytically in most types of models. The full distribution for some dependent variable must itself be represented by summary statistics such as deciles or some other percentiles, graphically in the form of a density histogram, or some scalar indicator such as a Gini coefficient. However, for comparatively simple models, e.g. models depicting a single outcome (of any of the forms discussed above) microsimulation is unlikely to be able to provide any outcome measures that cannot also be obtained analytically.

For more complex models, such as a ‘multistate’ model in which several possible transitions can occur, certain summary indicators, or transformations of the underlying process, cannot be obtained analytically, or can be obtained only at great cost, or require numerical approximations to which microsimulation could be seen as a low-cost alternative. For example, in a semi-Markov (or Markov renewal) model, even in the absence of age dependence, certain outcomes that can be formally expressed with respect to the underlying hazards – such as the state probabilities (that is, the probability that an individual is in state j at time t) or the renewal function (that is, the expected number of events of a given type between time t and $t + w$) – do not have closed-form expressions except in the simplest (and least realistic) cases, such as that of no duration dependence. In general, if one wished to compute what the model predicts one of those outputs to be, the choice is between numerical inversion of Laplace transformations or microsimulation (Wolf 1986). Furthermore, if one wished to compute the state probabilities for a number of states $j = 1, \dots, J$ and a sequence of times $\tau_1, \tau_1 + 1, \tau_1 + 2, \dots$, then it would be necessary to go through the numerical-inversion process for every desired combination of state and time, whereas at least in principle a single run of a microsimulation program would provide sufficient output data to compute all the desired quantities. Moreover, the desired quantities can be obtained through the application of simple summary statistics to the microsimulation output. Note, however, that using microsimulation may introduce extra variance into those summary statistics, a topic discussed below.

Several examples of the use of microsimulation to generate a variety of indicators of model output can be found in the existing literature. For example, Dick et al. (1994) estimate a set of hazard functions describing transitions between nursing home and community-based residence, and from each residential setting to death, from age 65 onwards. They then use microsimulation to generate several indicators of life-cycle experience that depend on the full set of estimated hazards, including the number of times admitted to a nursing home, and the duration of time spent in both the

community and in nursing homes, prior to death. For each indicator, means, medians, and selected percentile figures are presented. Similarly, Moffitt and Rendall (1995) use microsimulation to develop summary indicators of women's lifetime experience as a family head, based on estimated hazard functions for entry into and exit from family headship. Wolf and Levy (1984) develop a model of job retention that includes two hazard functions, one each for jobs with and without pension coverage. They use microsimulation to generate a sample of lifetime employment histories, including outcomes such as the timing of vesting of pension benefits. The latter two examples are conditional simulations, in the sense that mortality is ignored. In all three cases cited, the use of microsimulation greatly extends the range of implications generated by the estimated model.

The ability of microsimulation to generate a data base in which numerous summary indicators of the estimated underlying model are implicit has led to several attempts to develop goodness-of-fit measures based on microsimulation output. Tuma et al. (1979) provide perhaps the first example of this use of microsimulation. They present a model of transitions among three states (partnered – whether maritally or informally – unpartnered, and attrited from the longitudinal study) based on a covariate-dependent but time and duration-independent continuous-time model. They compute, for each individual in the data file, selected state probabilities and mean event-counts, as well as finite-interval transition probabilities, and compare those predictions to their observed counterparts in the data. They note that the observed outcomes to which the predictions are compared were not used directly in estimating the model parameters, thus illustrating an important benefit of microsimulation. Heckman and Walker (1987), in a similar vein, present χ^2 goodness-of-fit statistics for simulated versus observed event-count outcomes, as well as several other ex post tests of data generated by microsimulation versus data used in parameter estimation.

Investigate Various Defective-data Problems

A second area in which microsimulation can prove helpful to the data analyst is in examining the potential seriousness of various data shortcomings, and, by extension, evaluating various procedures intended to correct for those shortcomings. Two such 'defective data problems' are errors or incompleteness in retrospective data, and attrition from a panel sample, both of which should generally lead to biased parameter estimation.

Large-scale population surveys frequently collect retrospective event-history data, and in panel surveys some such retrospective data is often collected for between-interview events. For example, Canada's 1995 General Social Survey data obtained marital-history data used by Polland and Wu (1998) to estimate a model of age at first marriage. Data of this type is, obviously, provided only by

persons who have survived to 1995 and are therefore able to be interviewed. The estimated model can be supposed to pertain to the full cohort of persons defined by a particular age, or age range, in 1995 only if prior losses from that cohort due to mortality are unrelated to the phenomenon being modeled. Yet there is ample evidence that mortality and marital status (and, by implication, marital transitions) are related, calling into question the parameters of marriage-dynamics models estimated using retrospective data.

Microsimulation could be used to investigate the degree of seriousness of such bias. For example, to the equations for marital-status transitions could be added equations for mortality, incorporating alternative assumptions regarding both the effects of unmeasured variables on the selection into a marital state, and the selection by mortality out of that state. Simulated counts of marital events based on such a model could then be compared to external information on the occurrence of marital events over time, information of the type generally readily available from vital records. An admitted problem of this approach is that it becomes difficult, even impossible, to distinguish problems due to recall error in the dating of past events from those due to selective losses from a cohort due to mortality.

One particular form of incompleteness in event-history data is that of left censoring, which gives rise to various forms of ‘initial conditions’ problems. For example, spells in progress at the beginning of an observation period are described by a different probability distribution than are fully-observed spells (Cox 1967). The problem is greatly magnified in models that explicitly incorporate unobserved heterogeneity (Heckman 1981). A number of approaches have been proposed for dealing with variant forms of initial conditions problems. Moffitt and Rendall (1995), for example, incorporate analytic probability expressions for the initial conditions directly into their estimation, which is feasible in view of the fact that their model is driven exclusively by age. In more complicated situations, however, such as those in which observed and unobserved factors interact selectively over the life cycle, microsimulation of the probabilities governing initial values may be more feasible than analytic solutions.

A second data problem for which microsimulation might prove useful is dealing with outcomes whose values are unobserved due to respondent attrition from a panel study. If a model of the joint dynamics of some outcome of interest, as well as the continued presence of a respondent in the sample (i.e. the complement of attrition) could be developed, with a common dependence of those two (or more) variables on one or more unobserved factors, then the estimated model could be used to simulate the distribution of responses among attriters, i.e. the responses otherwise unrecorded in the original data. Such an exercise is closely related to missing-value imputation in general, to which we now turn.

Imputing Missing Values

Given the claimed close association between microsimulation and missing-value imputation, it is not surprising that one of the apparent benefits of microsimulation is that of supplying values for otherwise missing variables, allowing, in turn, richer subsequent analyses. In one example of such an application, Laditka and Wolf (1998) presented a discrete-time model of functional-status transitions (e.g. transitions among states defined as 'unimpaired,' 'moderately impaired,' 'severely impaired' and 'dead'). The model was estimated using data from the Longitudinal Study of Aging (LSOA), in which subjects' functional status was observed at intervals of, on average, 27 months. Thus the estimation problem was that of identifying an embedded Markov chain (cf. Singer and Spilerman 1974). Laditka (1998) used that estimated model of functional-status transitions to impute a sequence of monthly functional-status values to respondents to the National Long-term Care Survey (NLTC), in which functional status is known only for the month of interview in waves I (1982), II (1984) and III (1989). Thus, a respondent interviewed in all three years provided, at most, observed values of three out of about 84 monthly values of functional status. Laditka (1998) simulated monthly sequences of functional statuses using microsimulation techniques, then went on to estimate a model of month-by-month probabilities of nursing home admission and discharge based on the imputed data values, pooling person-months of (observed plus imputed) data. Although multiple replications of the imputation-estimation steps would be advisable in order to correct the final-stage parameter estimates for imputation variance, Laditka (1998) performed only a single replication of the imputation step.

Caveats

The claimed advantages of microsimulation come at a price. The microsimulation approach has both substantive and procedural limitations. Among the drawbacks or limitations of the microsimulation methodology are:

- *everything is endogenous.* In order to make individual-level predictions from a dynamic model it is necessary to have updated values of explanatory variables at each temporal step in the simulation algorithm. Thus, variables taken as exogenous in the estimation stage, and whose values are therefore treated as predetermined, become problematic in a microsimulation if their value is not fixed over time. For example, the model of age at marriage found in Pollard and Wu (1998) contains among its explanatory variables several individual attributes that change over the life cycle, including educational attainment, current student status, current employment status, and current pregnancy status. All these variables are observed in the data,

and therefore present no problem for estimation, in the discrete-time hazard model approach used by Pollard and Wu. However, if someone wished to simulate the subsequent marital experience of young unmarried women found in the database, or simulate marital histories for some other population, real or hypothetical, then it would be necessary to develop auxiliary equations with which to simulate educational, employment, and pregnancy histories. Alternatively, the analyst might assume a prespecified time-path for all time-varying explanatory variables, and condition the dynamic microsimulation on that set of predetermined time paths, but this greatly limits the scope of the exercise.

- ***‘difference’ estimators generally won’t work.*** Fixed-effect specifications have been proposed for a number of the panel-data models discussed above. In the linear panel-data model [represented by equation (6)] the person-specific intercepts can be treated as fixed effects, and estimated as coefficients on person-specific dummy variables (which requires, however, that all other time-invariant variables be dropped from the model). For the panel logit model [a special case of equation (7)] Chamberlain (1980) has proposed a ‘difference’ estimator that eliminates the fixed effects from the model. For panel Probit models fixed-effects estimators are available only in special cases (Borjas and Sueyoshi 1993). The advantage of the fixed-effects estimators is that they relax the assumption, required for virtually all random-effects estimators, that the person-specific effects are uncorrelated with other components of the model, in particular the included covariates. The disadvantage of the fixed-effects estimators, for purposes of microsimulation, is that they make difficult, or even impossible, any out-of-sample simulations. In particular, if an equation or set of equations has been estimated using a fixed-effects specification, then out-of-sample simulations are possible only if (a) numerical values for the full set of empirical fixed effects can be recovered, and (b) it is possible to impute, in some fashion, the numerical values of fixed effects from the estimation sample to records in the simulation sample. Either or both of these conditions may, however, fail to be realized.
- ***software limitations.*** While there exist several choices, and at least a few widely available general-purpose statistical software systems, with which to estimate many if not all the standard types of statistical models for use with panel or longitudinal data, there are few choices facing the potential microsimulator. Thus the analyst is likely to have to develop an original program in order to realize the claimed advantages of the microsimulation technique.

- **limited inferential theory.** The summary statistics computed on microsimulation output clearly depend on data, on parameters, and on additional sources of ‘sampling’ variability of a rather specialized nature. Yet little attention has been paid so far to the problem of uncertainty, or sampling variation, or of interval estimation in the context of microsimulation. We turn to this issue below.

Uncertainty Analysis of Predictions from Microsimulations

Although much effort has gone into the development and application of microsimulation models in demography, economics, and policy analysis, relatively little attention has been paid to the issue of uncertainty surrounding the point estimates produced by microsimulation. In the words of Klevmarken “... in current practice the inference aspects [of micro simulation models] have been neglected. One has been satisfied if the model runs and approximately tracks observed data.” (Klevmarken 1998:1) Pudney and Sutherland (1994) provide analytic expressions for the variances of predictions from a static microsimulation model, recognizing three sources of variability: classical sampling error (that is, error associated with the use of a sample rather than the entire population for the initial or baseline conditions), Monte Carlo errors associated with the particular stream of random numbers used to make stochastic assignments, and parameter uncertainty. Klevmarken (1998) mentions the same three sources of uncertainty, and discusses the errors produced by microsimulation in the context of model validation. He suggests replication as a means of dealing with Monte Carlo variation, and either randomization over parameters or sample reuse methods such as the bootstrap to deal with parameter uncertainty. Wolf and Laditka (1997) provide an illustration of the former approach, while Calhoun (1997) provides an illustration of the latter (although Calhoun studies a deterministic life-table model rather than a stochastic microsimulation model). Cohen’s (1991) suggestions are similar in several respects to those found in Klevmarken’s later (1998) paper. Cohen suggests (1) the bootstrap as a means to estimate classical sampling variance, and (2) randomization over the estimated distribution of parameters to deal with parameter uncertainty. He also suggests using (3) the multiple imputation method to deal with data errors in the base or starting population caused by statistical matching, although it is not entirely clear how the three techniques are to be combined. The procedures suggested below build upon and extend the ideas first presented in Cohen (1991).

It is also worth noting that several authors advocate the usage of methods to *reduce* the variability of microsimulation output; this is particularly true in textbook treatments of operations research applications (e.g., Bratley et al. 1987). van Imhoff and Post (1998) echo this advocacy of variance-reduction techniques in the context of microsimulation models for demographic projections. The desire to minimize variation in simulation outputs appears to

be motivated by an assumption that the mean is the only summary statistic of interest once the microsimulation has been completed. Yet if, as noted before, one of the advantages of microsimulation is its ability to provide information on the entire *distribution* of outcome values as well as their expected value, then in the context of stochastic microsimulation these variance-reduction techniques seem to be misguided and limiting.

In addition to the three sources of variance identified by Pudney and Sutherland (1994) and Klevmarcken (1998), at least two additional sources of uncertainty can be identified. The first [mentioned by Cohen (1991)] consists of imputation error found in the starting-population data base. It is rare for any microdata file produced through sampling to be without missing-data fields, arising from both item and unit nonresponse. A common solution to missing-data problems is to impute values to the missing fields, a process that inevitably introduces error and, therefore, uncertainty about summary statistics based on the data. McNally and Wolf (1996) discuss another type of data-base imputation error: in their study, the starting population for a microsimulation is developed by pooling observations from two different household surveys that happen to come from partially-overlapping sampling frames. However, it is not possible to tell which observations from file B come from that part of the population that is also represented in file A. Therefore, McNally and Wolf develop a random-assignment procedure for choosing observations for pooling such that the final data file can be supposed to represent the desired population without any duplication.

Another source of uncertainty that is present in microsimulation output results from the analyst's ignorance about the true value of any 'unmeasured heterogeneity' factors imputed to individual observations in the data file. This is a special case of the more general missing-data problem.

Microsimulation shares with the multiple-imputation (MI) methodology presented in Rubin (1987) three important features. First, some sort of model is developed with which to predict an otherwise unknown value of some variable. Second, that prediction depends, in part, on the value of a randomly-selected variate. And third, the process is repeated several statistically independent times. In the case of MI, a number of repetitions of the random-assignment algorithm are performed in order to adjust any computed summary statistics for imputation variance. In other words, the analyst must be prepared to accept a penalty, in the form of larger standard errors, for making guesses at the values of otherwise missing data fields. In the case of microsimulation, replications of the microsimulation – multiple "runs" of the software – are generally performed in order to average out any Monte Carlo variation in the summary statistics.

Given the parallels between the two methods, MI would seem to provide a basis for variance estimation of summary statistics computed for microsimulation output. Rubin (1987) suggests that a small number of replications of the

imputation model be performed. If R is the number of such replications, and W_r^* and S_r^* are a sample statistic and its variance, respectively, based on the r^{th} replicate, then the overall value of the statistic in the presence of imputation error is with variance.

$$\bar{W} = \frac{1}{R} \sum_{r=1}^R W_r^* \quad (10)$$

The first term in equation (11) is the simple average of the variances produced over the R replications, while the second is the between-replication variance of the estimator adjusted by the term $1 + R^{-1}$, that is, the ‘imputation variance.’

$$\bar{S} = \frac{1}{R} \sum_{r=1}^R S_r^* + (1 + R^{-1}) \frac{\sum (W_r^* - \bar{W})^2}{R - 1} \quad (11)$$

A microsimulation exercise is, in many respects, analogous to a data-imputation exercise. First, the data elements of interest are missing; they are, in fact, 100 percent missing. Secondly, predicted values for those data elements come from a predictive model, one that includes both deterministic and stochastic elements. Accordingly, the following simple procedure is suggested for developing variances to accompany summary statistics computed using microsimulation output:

- (a) in preparing the initial-conditions data file, carry out and retain in the file K of independent random replications of each imputed element (i.e. unit imputations and/or item imputations);
- (b) select K random combinations of each random ‘factor’ present in the microsimulation. This will include each distinct imputed factor present in the starting population (above) as well as each model element that is subject to sampling error (e.g. regression coefficients) as well as random-assignment factors (e.g. error terms or random numbers used to make probabilistic assignments). The ability to sample from the ex post distribution of parameter vectors depends, in turn, on the use of an estimation technique that generates such a distribution (e.g. maximum likelihood) and a willingness to appeal to the asymptotic nature of that distribution;
- (c) run the microsimulation program (the sampling algorithm) K times, each time computing the run-specific sample statistic W_k and its variance S_k . At this stage, procedures to deal with departures from simple random sampling of the starting population, such as bootstrap or other resampling procedures (Cohen 1991) may need to be applied;

- (d) use equations (10) and (11) to derive the overall simulated point estimate and variance for each summary statistic of interest.

The preceding steps must, however, be viewed as tentative for several reasons. First, it will in general be desirable to isolate the contribution to total variance of each of the identified sources of variability. In order to do so effectively, some sort of multifactorial experimental design should be used. For example, one can easily envision the circumstance of having five separate factors contributing to overall simulation variance. If each factor were represented by, say, five randomly selected ‘levels’ there would be $5^5 = 3,125$ different possible combinations of factors, requiring 3,125 runs of the microsimulation program. Since this is clearly undesirable, and since each factor can by design be made orthogonal to all other factors, smaller ‘fractional factorial designs’ can be used. There exists a specialized literature on the application of statistical techniques, including experimental designs, to microsimulation (Kleijnan 1987), in which guidance on this approach might be found.

Second, some of the ‘factors’ over which randomization can be performed are themselves high-dimensional vectors, e.g. vectors of regression coefficients. Just as the analyst might want to investigate the contribution of an individual factor to overall variance, it might also be desirable to determine the role of sampling variances of individual parameter elements. This would, for example, allow the user to see the payoff to greater precision of parameter estimation. One problem with this objective, however, is that estimated parameters generally are not independent of other parameters (they have nonzero covariances), making it difficult to identify their unique contribution to overall variance. In particular, it is likely to require numerous replications of the microsimulation exercise to identify these effects.

Finally, an issue requiring further development is the number of replications (i.e., the value of K) necessary to adequately represent the ‘between’ replication variance due to the several sources of simulation uncertainty. In survey-data item-imputation applications of the multiple-imputation technique, a small number (say 3-6 replications) has been viewed as sufficient. However, in the microsimulation context there are both additional sources of uncertainty and 100 percent missing information, both of which might indicate a need for increasing the number of replications. Variance computations based on a small number of levels of each random factor might also be excessively subject to the influence of outliers. Thus there remains considerable developmental work to be done on the problem of quantifying the uncertainty associated with summary statistics based on microsimulated data.

Summary and Conclusion

Microsimulation is an increasingly familiar tool with which to investigate the sample paths of estimated models of socio-economic-demographic models, to obtain solutions to complex problems in which analytic solutions are infeasible, to obtain estimates of the costs and distributional implications of hypothetical policy regimes, and in many other applications. This paper presents several examples of ways in which widely-used econometric specifications can be embedded in microsimulation exercises. It also argues that microsimulation has a potentially important role to play earlier in the modeling process, namely during the process of model formulation and data analysis. Specifically, microsimulation can be used to extend the range of inferences that can be drawn from the estimated parameters of a model, can help to solve certain types of defective-data problems, and can fill gaps in available data.

A relatively underdeveloped area is that of quantifying the uncertainty inherent in summary statistics based on data produced by a microsimulation program. I have argued that due to strong parallels between the multiple imputation methodology and the structure and procedural aspects of many microsimulation exercises, the multiple imputation methodology provides a natural framework with which to develop estimates of the variances, and therefore the confidence intervals, that accompany estimates based on simulated data. There is a clear need for both additional theoretical work in this area, and for a range of experience in the application of such methods, in order to establish their feasibility and usefulness.

References:

- Allison, Paul D. 1982. "Discrete-time methods for the analysis of event histories," in S. Leinhardt (ed.), *Sociological Methodology 1982*, San Francisco: Jossey-Bass Publishers. Pp. 61-98.
- Barrett, J.C. 1971. "Use of a fertility simulation model to refine measurement techniques," *Demography* 8: 481-490.
- Beggs, John J. 1988. "A simple model for heterogeneity in binary logit models," *Economics Letters*, 27: 245-249.
- Berman, Simeon M. 1963. "Note on extreme values, competing risks and semi-Markov processes," *Annals of Mathematical Statistics* 34: 1104-1106.
- Borjas, George J. and Glenn T. Sueyoshi. 1993. "A Two-Stage Estimator for Probit Models With Structural Group Effects," National Bureau of Economic Research Technical Paper no. 146 (November).

Methodological Issues – Douglas A. Wolf

- Bratley, Paul, Bennett L. Fox, and Linus E. Schrage. 1987. *A Guide to Simulation* (Second Edition). New York: Springer.
- Caldwell, Steven B. 1999. "Dynamic Microsimulation and the Corsim 3.0 Model." Available at <http://www.strategicforecasting.com/pubs/99-misc/theory93.html>.
- Calhoun, Charles. 1997. "Bootstrapping the Multi-State Life Table: Preliminary Results." Presented at the Annual Meetings of the Population Association of America, Washington, D.C., March 27-29, 1997.
- Chamberlain, Gary. 1980. "Analysis of covariance with qualitative data," *Review of Economic Studies* XLVII: 225-238.
- Christofides, Louis N. and C.J. McKenna. 1996. "Unemployment insurance and job duration in Canada," *Journal of Labor Economics*, 14: 286-312.
- Cohen, Michael. 1991. "Variance estimation of microsimulation models through sample reuse," in C.F. Citro and E.A. Hanushek (eds.). *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling. Volume II: Technical Papers*, Washington, D.C.: National Academy Press. Pp. 237-254.
- Conaway, Mark R. 1990. "A random effects model for binary data," *Biometrics*, 46: 317-328.
- Cox, D.R. 1967. *Renewal Theory*. London: Chapman and Hall.
- David, H.A. and M.L. Moeschberger. 1978. *The Theory of Competing Risks*. New York: Macmillan Publishing Co., Inc.
- De Vos, S., Palloni A. 1989. "Formal Models and Methods for the Analysis of Kinship and Household Organization," *Population Index* 55: 174-198.
- Dick, Andrew, Alan M. Garber, and Thomas A. MaCurdy. 1994. "Forecasting nursing home utilization of elderly Americans," in D.A. Wise (ed.), *Studies in the Economics of Aging*. Chicago: The University of Chicago Press. Pp. 365-394.
- Engberg, John, Peter Gottschalk, and Douglas Wolf. 1990. "A random-effects logit model of work-welfare transitions," *Journal of Econometrics* 43: 63-75.
- Finney, D.J. 1971. *Probit Analysis*. Cambridge: Cambridge University Press.

- Firth, David and Clive Payne. 1999. "Efficacy of programmes for the unemployed: discrete time modelling of duration data from a matched comparison study," *Journal of the Royal Statistical Society, Series A*, (1621), 111-120.
- Galler, Hienz F. 1989. "Policy evaluation by microsimulation - The Frankfurt Model." Paper prepared for the 21st General Conference of the International Association for Research in Income and Wealth, Lahnstein, August 20-26, 1989.
- Gilbert, Nigel and Klaus G. Troitzsch. 1999. *Simulation for the Social Scientist*. Buckingham: Open University Press.
- Goldstein, Joshua. 1996. *The Demography of Family and Kinship in an Age of Divorce and Remarriage*. Unpublished doctoral dissertation, Department of Demography, University of California at Berkeley.
- Ham, John C. and Samuel A. Rea, Jr. 1987. "Unemployment insurance and male unemployment duration in Canada," *Journal of Labor Economics* 5: 325-353.
- Harding, A. 1993. *Lifetime Income Distribution and Redistribution: Applications of a Microsimulation Model*. Amsterdam: North-Holland.
- Hausman, Jerry A. and David A. Wise. 1978. "A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences," *Econometrica* 46: 403-426.
- Haveman, Robert H. and Kevin Hollenbeck (eds.). 1980. *Microeconomic Simulation Models for Public Policy Analysis, Volume I: Distributional Impacts*. New York: Academic Press.
- Jechkan, James J. 1981. "The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process," in C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge: MIT Press. Pp. 179-195.
- Heckman, James J. and George J. Borjas. 1980. "Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence," *Economica* 47: 247-283.
- Heckman, James J. and Burton Singer. 1984. "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica* 52: 271-320.

- Heckman, James J. and James R. Walker. 1987. "Using goodness of fit and other criteria to choose among competing duration models: A case study of Hutterite data," in C.C. Clogg (ed.), *Sociological Methodology 1987*, Washington, D.C.: American Sociological Association. Pp. 247-307.
- Hsiao, Cheng. 1986. *Analysis of Panel Data*. Cambridge: Cambridge University Press.
- King, Anthony, Hans Bækgaard, and Martin Robinson. 1999. "DYNAMOD-2: An Overview." Technical Paper no. 19 (December). Canberra: National Centre for Social and Economic Modelling.
- Kleijnan, Jack P.C. 1987. *Statistical Tools for Simulation Practitioners*. Marcel Dekker, Inc.
- Klein, John P. and Melvin L. Moeschberger. 1997. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag.
- Klevmarcken, N. Anders. 1998. "Statistical inference in micro simulation models: Incorporating external information." Department of Economics Working Paper, Uppsala University.
- Laditka, Sarah B. 1998. "Modeling lifetime nursing home use under assumptions of better health," *Journal of Gerontology: Social Sciences* 53B: S177-187.
- Laditka, Sarah B. and Douglas A. Wolf. 1998. "New methods for analyzing active life expectancy," *Journal of Aging and Health* 10: 214-241.
- Lewis, Gordon H. and Richard C. Michel (eds.). 1990. *Microsimulation Techniques for Tax and Transfer Analysis*. Washington, D.C.: The Urban Institute.
- Lillard, Lee A. and Robert J. Willis. 1978. "Dynamic aspects of earning mobility," *Econometrica* 46: 985-1012.
- Manton, Kenneth G., Eric Stallard and James W. Vaupel. 1986. "Alternative models for the heterogeneity of mortality risks among the aged," *Journal of the American Statistical Association* 81: 635-644.
- McFadden, Daniel. 1973. "Conditional logit analysis of qualitative choice behavior," in P. Zarembka (ed.), *Frontiers in Econometrics*, New York: Academic Press. Pp. 105-142.

The Role of Microsimulation in Longitudinal Data Analysis

- McFadden, Daniel. 1989. "A method of simulated moments for estimation of the multinomial probit model without numerical integration," *Econometrica*, 57: 995-1026.
- McNally, James and Douglas A. Wolf. 1996. "Family structure and institutionalization: Results from merged microdata." Papers in Microsimulation Series Paper No. 2. Syracuse: Center for Policy Research, Syracuse University (available at <http://www.cpr.maxwell.syr.edu/demogctr/micropap/micro2ab.htm>).
- Moffitt, Robert A. and Michael S. Rendall. 1995. "Cohort trends in the lifetime distribution of female family headship in the United States, 1968-1985," *Demography* 32: 407-424.
- Mooney, Christopher Z. 1997. *Monte Carlo Simulation*. Sage Publications.
- Nelissen, Jan H.M. 1995. *Demographic Projections by Means of Microsimulation: The NEDYMAS Mode*. Tilburg, Tilburg University: Work and Organization Research Centre.
- Orcutt, Guy, Steven Caldwell, and Richard Wertheimer II. 1976. *Policy Exploration Through Microanalytic Simulation*. Washington D.C.: The Urban Institute.
- Orcutt, Guy, Joachim Merz, and Hermann Quinke (eds.). 1986. *Microanalytic Simulation Models to Support Social and Financial Policy*. Amsterdam: North-Holland.
- Pollard, Michael S. and Zheng Wu. 1998. "Divergence of marriage patterns in Quebec and elsewhere in Canada," *Population and Development Review* 24: 329-356.
- Pudney, Stephen and Holly Sutherland. 1994. "Statistical reliability and microsimulation: The role of sampling, simulation and estimation error." Discussion Paper Series number 9402. Cambridge: The Microsimulation Unit.
- Ravanera, Zenaida R., Fernando Rajulton, and Thomas K. Burch. 1993. "From home-leaving to nest-emptying: A cohort analysis of life courses of Canadian men and women, 1910-1970," in *International Population Conference / Congres International de la Population, Montreal 1993, 24 August - 1st September. Volume 2*, Liege: International Union for the Scientific Study of Population. Pp. 207-218.
- Reader, S. 1993. "Unobserved heterogeneity in dynamic discrete choice models," *Environment and Planning A*, 25: 495-519.

- Ridley, Jeanne Clare, and Mindel C. Sheps. 1966. "An analytic simulation model of human reproduction with demographic and biological components," *Population Studies* 19: 297-310.
- Rowe, Geoff. 1989. "Event history analysis of marriage and divorce in Canada." Proceedings of the Statistics Canada Symposium on Analysis of Data in Time, October 1989.
- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Ruggles, Steven. 1987. *Prolonged Connections: The Rise of the Extended Family in Nineteenth-Century England and America*. Madison: The University of Wisconsin Press.
- Schmertmann, Carl P. 1999. "Estimating multistate transition hazards from last-move data," *Journal of the American Statistical Association* 94: 53-63.
- Singer, Burton and Seymour Spilerman. 1974. "Social mobility models for heterogeneous populations," in H.P. Costner (ed.), *Sociological Methodology 1973-1974*, San Francisco: Jossey-Bass Publishers. Pp. 256-401
- Tuma, Nancy B., Michael T. Hannan, and Lyle P. Groeneveld. 1979. "Dynamic analysis of event histories," *American Journal of Sociology* 84: 820-854.
- Van Imhoff, Evert, and Wendy Post. 1998. "Microsimulation Methods for Population Projection," *Population: An English Selection*, special issue *New Methodological Approaches in the Social Sciences*, 97-138
- Wachter, Kenneth W. 1987. "Microsimulation of Household Cycles," in J. Bongaarts, T. Burch, and K. Wachter (eds.), *Family Demography*. New York: Oxford University Press. Pp. 215-227.
- Wachter, Kenneth W. 1997. "Kinship Resources for the Elderly," *Philosophical Transactions of the Royal Society*, Series B 352: 1811-1817.
- Wolf, Douglas A. 1986. "Simulation methods for analyzing continuous-time event-history models," in N.B. Tuma (ed.), *Sociological Methodology 1986*. San Francisco: Jossey-Bass Publishers. Pp. 283-308.

The Role of Microsimulation in Longitudinal Data Analysis

- Wolf, Douglas A. 1988. "Kinship and Family Support in Aging Societies," in *Economic and Social Implications of Population Aging*. New York: United Nations Department of Economic and Social Affairs. Pp. 305-330.
- Wolf, Douglas A. and Frank Levy. 1984. "Pension coverage, pension vesting, and the distribution of job tenures," in H.J. Aaron and G. Burtless (eds.), *Retirement and Economic Behavior*. Washington, D.C.: The Brookings Institution. Pp. 23-61.
- Wolf, Douglas A. and Sarah B. Laditka. 1997. "Stochastic modeling of active life and its expectancy." Papers in Microsimulation Series, No. 4, Syracuse: Center for Policy Research, Syracuse University, (available at <http://www.cpr.maxwell.syr.edu/demogctr/micropap/micro4ab.htm>)
- Wolfson, Michael C. 1989. "Divorce, homemaker pensions and lifecycle analysis," *Population Research and Policy Review* 8: 25-54.
- Wolfson, Michael, Geoff Rowe, Jane F. Gentleman and Monica Tomiak. 1990. "Earnings and death: Effects over a quarter century." Analytical Studies Branch Research Paper Series, no. 30, Statistics Canada.
- Zedlewski, Sheila R., Roberta O. Barnes, Martha R. Burt, Timothy D. McBride and Jack A. Meyer. 1990. *The Needs of the Elderly in the 21st Century*. Washington, D.C.: The Urban Institute.
- Zenger, Elizabeth. 1993. "Siblings' neonatal mortality risks and birth spacing in Bangladesh," *Demography* 30: 477-488.