Current Development in Microsimulation and Assessment of Uncertainty in JUTTA Model

Meng Zhou[1,2] and Maria Valaste[2]

[1]University of Helsinki, email: meng.zhou@helsinki.fi
[2]Finnish Social Insurance Institution

Abstract

Nowadays, microsimulation method has been introduced to different fields, such as Social Science, Medicine research and Economic study. This method evaluates the effects of the proposed interventions or policies before they are implemented in the real world. In this article, we will concentrate on microsimulation method used in Social Science by firstly explaining two main streams in microsimulation world, Static approach and Dynamic approach. In the following section, the uncertainty of a Finnish static microsimulation model JUTTA is assessed and Toimtuki model one of the sub-model in JUTTA is detected to have space to be more accurate. In order to do so, two statistical models- Linear Regression model and Two-Stage Least Squares (2SLS) model are applied to it. From the results, we could conclude that both the Linear Regression and 2SLS successfully improves the accuracy of TOIMTUKI to some extent.
Keywords: Static microsimulation, Dynamic microsimulation, JUTTA, assessment, 2SLS

1 Introduction

What and Why

A microsimulation model differs from other types of models in that it operates on individual units rather than on aggregate information (TRIM3. 2012a). Typically, in social sciences, those units are individual substantial or economic units. The database used as input to a microsimulation model contains records describing persons, households or business. And the simulation model applies a set of rules to each individual record. The result of the computations might be the amount of taxes owed by the unit to which the unit is entitled under certain government legislation. Also, if we are interested in the total tax, each individual result should be multiplied by whatever weight is associated with the unit in the microdata file, then the weighted individual results are added together to obtain the aggregate result. Thus, different policies could apply to the same microdata file, and the report of the comparisons among the different results could be a good helper to the wise government.

The purpose of the microsimulation is mainly to evaluate the effects of the proposed interventions or policies before they are implemented in the real world. By using microsimulation, people can easily estimate the impacts of a new scheme by producing outputs on a wide range of measures of effectiveness.

Currently, there are two main streams in the microsimulation field, which are Static microsimulation and Dynamic microsimulation.

2 State of the art

Static microsimulation

The Static microsimulation has one important character that it does not take the individual behavior into account, which means once the rules are made, they will be obeyed 100% without any variation. It suits for performing detailed simulations of past, the present, and the near future. It typically use static aging techniques, changing certain variables on the original microdata file to produce a file with the demographic and economic characteristics expected in the future year. Person weights are modified to change the total population and the weighted characteristics of the population; labor force status may be changed to alter the unemployment rate; and incomes are adjusted for price changes. Simulations can then be run on the aged microdata file to estimate the impact of a change to be implemented in the future year (TRIM3. 2012b).

Dynamic microsimulation

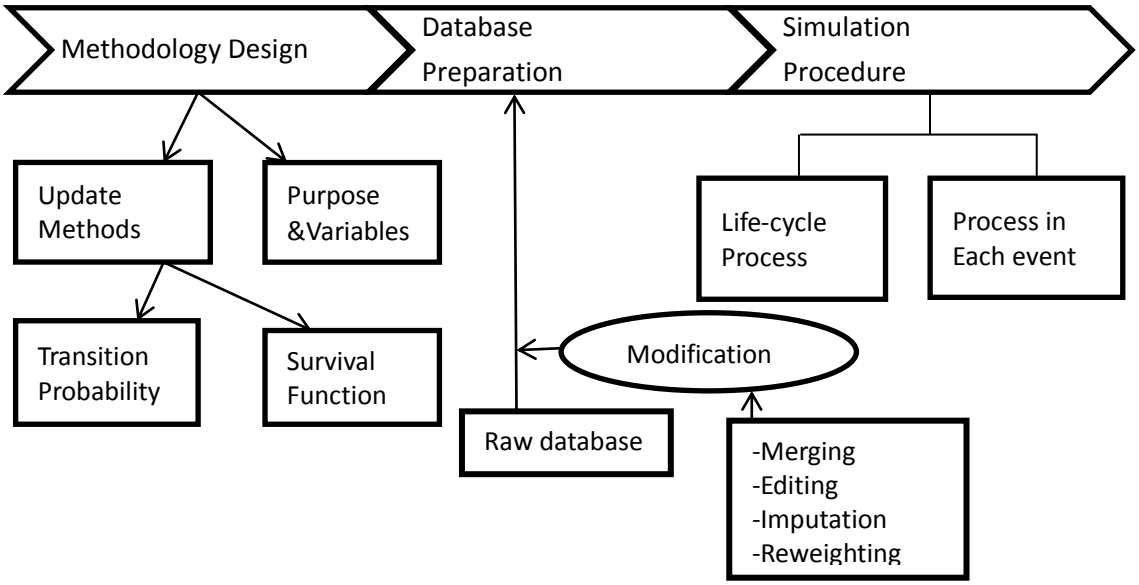
Dynamic microsimulation models age each person in the microdata file from one year to the next by probabilistically deciding whether or not that person will get married, get divorced, have a child, drop out of school, get a job, change jobs, become unemployed, retire, or die, then the same procedure is repeated as many times as the user wants to achieve the final simulation year. Simulations of government legislations can be run in the current year, the final year of the aging process, or any interim year. The simulation of the government program in one year may affect a person's characteristics in the subsequent year (TRIM3. 2012b). For example, whether or not someone will drop out of school could be programmed to depend partly on family income, which could, in turn, be affected by government transfer payments. This kind of models could create the synthetic database for a future year, which is capable of performing simulations into the distant future, but it couldn't capture as much details as static models do.

In Dynamic microsimulation models the transition probabilities play the important role, because they are used to create the synthetic database about the individuals' life paths on the demographic events, personal events and so on.

3 The detailed Dynamic microsimulation

There are three components in the Dynamic Microsimulation: methodology design, database preparation and simulation procedure. The figure 1 presents the basic structure of it.

Figure 1: Simulation Procedure



Updating methods and statistical model application

In the methodology step, deciding the update methods is the main point. Usually, there are two options, transition probability and survival function. Nowadays, the most popular updating method is based on the transition probabilities, and the space in time between the updating processes is one year. Thus, the estimation of the transition probability becomes the hot point, where statistic models are mainly applied. The often used models are linear model, generalized linear model, mixed model and so on. The survival function method is sometimes used, for instance, the cox model is used when estimating the event fertility, please see the example in Anthony *et al.* (1999).

Here, we will talk about the application of Generalized Linear Model (GLM). The generalized linear model is a flexible generalization of ordinary linear regression. The linear model can be transformed to a generalized linear model by linked function g(). The model could be represented as: $E(Y) = \mu = g^{-1}(X\beta)$ Where $E(Y)$ is the expected value of *Y*, $X\beta$ is the linear predictor, a linear combination of unknown parameter vector β, g is the link function.

There are many commonly used link functions. In Dynamic microsimulation, the logit and probit are the two most useful models when we are estimating the transition probabilities. Here one example will be given, the event is the "Employment Status", for more information, please see Lennart Flood *et al.* (2005).

It would be a good case to illustrate the Monte Carlo Simulation. Monte Carlo technic gives the model stochastic property. For the binary variable employment status, we have a Bernoulli distribution, i.e. $Y_i \sim bernoulli(\pi_i)$, where $\Pr[Y_i = 1] = \pi_i$ and $\Pr[Y_i = 0] = 1 - \pi_i$. As an illustration, let Y_i denote unemployment for individual i during the period of interest. Let Y_i=1 denote unemployment and Y_i=0 denote employment, π_i denote the probability that the individual is unemployed during the year. This event is simulated by comparing π_i with a uniform random number u_i . If $u_i < \pi_i$ the event is realized and individual i become unemployed.

The propensity of becoming unemployed is determined by π_i , by allowing π_i to be determined by individual or household attributes these attributes also determine the probability of unemployment. This is typically accomplished by a logit regression model. The logit model is given as $\pi_i = [1 + \exp(-X_i\beta)]^{-1}$,

where X_i is a vector of individual or household characteristics like gender, age, working history or any other characteristic relevant for explaining unemployment, i.e. rate of regional unemployment and β is a vector of parameters.

In order to calculate the estimator $\hat{\pi}_i$, firstly, we need to know the expected parameter vector $\hat{\beta}$, where it could be estimated from the outsource databases, such as registered database and official survey database. Then, the dependent variable $\hat{\pi}_i$ is calculated this way:

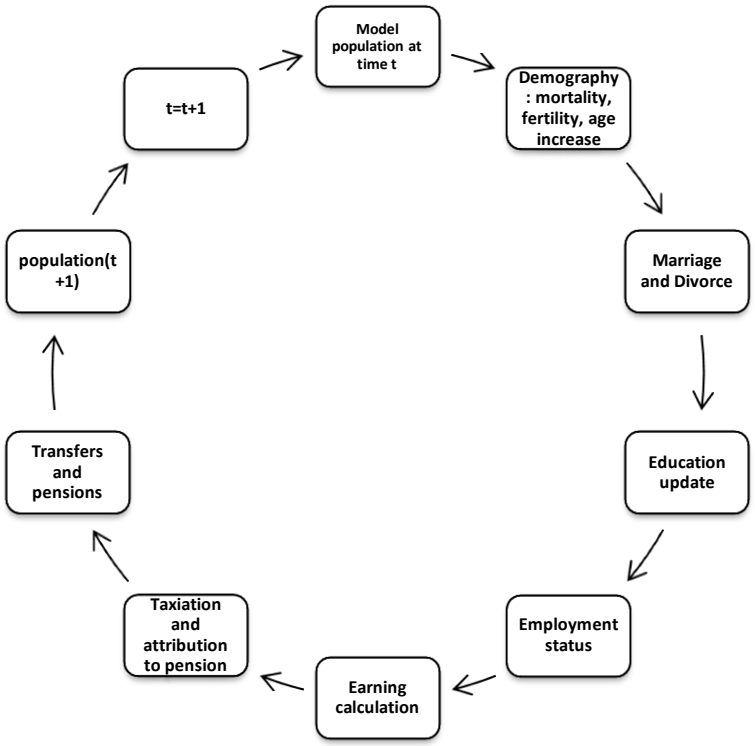
$$\hat{\pi}_i = [1 + \exp(-X_i \hat{\beta})]^{-1}$$

After obtaining $\hat{\pi}_i$, u_i is chosen randomly from the uniform distribution: $u_i \sim U(0,1)$. Finally simulated binary variable employment status is assigned to 1 or 0 by comparing $\hat{\pi}_i$ and u_i.

Simulation procedure: Model structure

The dynamic microsimulation model ages the underlying data base by one year, and that is run repeatedly to generate the multi-year demographic evolution needed for the whole simulation. Figure 2 describes us the life-cycle structure in the normal dynamic microsimulation model.

Figure 2: Life-cycle events process



Its "kernel" ages an input database by one year in any given pass. During each such pass, it simulates all of the births, deaths, marriages, labor force entry and exit and earnings, etc., that occur during that simulation year, and ages each of the individuals in the database by one year. It then outputs another database that is itself a new, representative population, but one that reflects the situation one year later than did the previous input database. The cycle is repeated over and over again for the length of the simulation run; in each cycle, the output data base from one pass through the kernel is used as the input for the next pass. (Anthony *et al.* 1999)

We could consider the aging process as a sequence of modules (events) that this step consists of a number of modules executed in(operated in) sequence, each of them modifying the in-memory population for that module's event for the current year. Each module processes all of the population for which that module/event is relevant, updating that aspect of the individuals' lives. However, not all individuals are eligible for all modules; e.g. individuals who have previously died will not give birth, and individuals who are presently married are not, in the same year, eligible to enter the marriage market (Rick Morrison, 1998).

Once the full set of modules has been executed, they have collectively aged the in memory population by one year. That is, they have implemented all of the events that effectively transform the base from one year's representative population to the next year's representative population.

4 Assessment of uncertainty of the JUTTA model and innovation method in TOIMTUKI

Background of JUTTA model

The JUTTA model is a static microsimulation model developed by Social Insurance Institution of Finland, it is also called tax-benefit model. JUTTA has 10989 households and around 30000 individuals sample size, and the data resources came from Statistics Finland. It has ten sub-models and one main model. The sub-models are designed for each branch of legislations and the main model is designed for running all the sub-models and producing the final results of the key data based on household level. The sub-models include: SAIRVAK, TTURVA, KOTIHUKI, OPINTUKI, KANSEL, VERO, LLISA, ELASUMTUKI, ASUMTUKI, TOIMTUKI. They represent sickness insurance benefits, unemployment benefits, child care benefits and day-care fees, study grant, the national pension system, personal taxes, benefits for families with children, pensioner's housing allowances, general housing allowances, means-tested income support, respectively. For each of these sub-models, parameter system and function system were built. (Honkanen Pertti, 2010)

Assessment of JUTTA model

In all the models, the accuracy is calculated in two different forms, one is the absolute difference percentage and the other one is the relative difference percentage. The absolute difference percentage is calculated based on the classifying the absolute errors between the real value and estimated value to the intervals [0, 1), [1, 10), [10, 100), [100, 1000) and [1000, ∞), and then dividing the number of the observations in each intervals by the total number of the observations. The relative difference percentage is similar to the absolute one, but classifying the errors in the intervals [0%, 0.1%), [0.1%, 1%), [1%, 10%), [10%, ∞).

After obtaining the percentage results, we could see that most of the models perform quite well, with their variables' accuracy high enough in the interval (60%, 100%] for both absolute different and relative different in first level called [0, 1) and [0, 0.1%) respectively. However, there is one extremely inaccurate model called TOIMTUKI meaning income-related supplementary benefit, which with both zero percentage in the first level intervals and more than 60% in the last intervals([1000, ∞) and [10%, ∞)).

The Toimtuki calculates the last benefit the people could apply after house benefit, heath benefit, student benefit and so on. In the other word, the Toimtuki could be regarded as the "residual" benefit in the JUTTA model, where the people apply when no other benefits could be applied.

In order to improve Toimtuki's accuracy, the statistical models are used. The first method is the Linear Regression method. After checking the relevant variables, seven of them are significant, which are tyot, ttyotpr, svatva, maksvuok, jasenia, desmod, meaning number of month of person's unemployment or forced leaving, unemployment allowance, household yearly income, monthly house rent, number of people in the household, decile group the household belonging to (according to OECD).

Method one: The linear regression model

$$\hat{y} = X\hat{\beta} \quad (1)$$

where X is the characteristics mentioned vector, $\hat{\beta}$ is the vector of estimated coefficients and $\hat{y}$ is the estimated benefits the household should receive. Now we plug the numeric coefficients in to equation (1):

$\hat{y}$ =195.67209+158.40442*tyot-0.17366*ttyotpr-0.10823*svatva+2.02230*maksvuok

+1308.86787*jasenia-786.69217*lapsia+520.13951*desmod; (R-Square=0.3744)

Table 1 shows us how efficient this method is when compared with the original Toimtuki model. It tells that the TOIMTUKI has been improved to some extent, especially when we consider the absolute difference.

Table 1: Comparison between JUTTA and Linear Regression model

Model		JUTTA	Method 1
Variable		TUKI	TUKI
Number of Observation		1012	621
Absolute Error Percentage	[0,1)	0	0.00161
	[1, 10)	0.00296	0.01288
	[10, 100)	0.02569	0.05314
	[100, 1000)	0.33103	0.37037
	[1000, ∞)	0.64032	0.562
Relative Error Percentage	[0, 0.1%)	0	0.00322
	[0.1%, 1%)	0.00494	0.01771
	[1%, 10%)	0.03458	0.08535
	[10%, ∞)	0.96047	0.89372

Method two: 2-Stage Least Squares

Algorithm:

- Estimate the binary variable status, which describes the weather the person gets this benefit or not, meaning if he/she gets, then status=1, if he/she doesn't get, then status=0. This step is using Monte Carlo method. Firstly, by logistic regression, the estimated parameters are calculated, then by using $\pi_i = \exp(X\beta)/(1+\exp(X\beta))$, where π_i is the probability t of being status=1. Finally, generating random value from the uniform distribution, and compare this value with π_i the probability, if the random value is larger than the probability, giving status value 0, if not, giving value 1.
- Estimating the TUKI value by regression model in case the status=1, otherwise, give value 0. However, the estimated value could be negative, but in reality, it should be nonnegative value, so change the negative value to 0.

Table 2: Logistic estimated coefficients

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr>ChiSq
Intercept	-2.1593	0.1185	332.0326	<.0001
tyot	0.2023	0.0107	354.7804	<.0001
svatva	-0.00005	6.957E-6	59.0412	<.0001
maksvuok	0.00216	0.000184	137.2076	<.0001
vvvmk1	-0.00011	0.000019	35.2189	<.0001
desmod	-0.1751	0.0390	20.1498	<.0001
tyotseu	0.0429	0.0120	12.8695	0.0003
lpaktyva	0.00952	0.00287	10.9831	0.0009

So, by calculating $\pi_i = \exp(X\beta)/(1+\exp(X\beta))$, where β is the vector and its estimation has been shown in table 2. Next, a random number u_i is drawn from the standard uniform distribution, that is $u_i \sim U(0,1)$. Finally, by comparing u_i and π_i, we give the estimated status 1 and 0. In cases that the individual estimated status is 1, regression model is set to calculate the TUKI, while in other cases, TUKI will be given value 0 directly. Next step is the linear regression as showed in method one, by using equation (1), we could get:

$\hat{y}$=-152.15070+202.66061*tyot-0.07511*svatva+1.78485*maksvuok-0.13827*ttyotpr

+394.85153*jasenia+341.67733*desmod;(R-Square=0.3930)

Table 3 shows us how efficient this method is when compared with the original Toimtuki model.

Table 3: Comparison between JUTTA and 2SLS model

Model		TOIMTUKI	Method 2
Variable		TUKI	TUKI
Number of Observation		1012	894
Absolute Error Percentage	[0, 1)	0	0
	[1, 10)	0.00296	0.00224
	[10, 100)	0.02569	0.02685
	[100, 1000)	0.33103	0.42953
	[1000, ∞)	0.64032	0.54139
Relative Error Percentage	[0%, 0.1%)	0	0
	[0.1%, 1%)	0.00494	0.00224
	[1%, 10%)	0.03458	0.01454
	[10%, ∞)	0.96047	0.98322

From the table 3, we see that the second method is better than the original method in the absolute difference view, however, it is almost the same as the original method in the relative difference point of view.

Conclusion:

JUTTA model as a static microsimulation model performs quite well in all sub-models, only except for the "residual" model-Toimtuki. The Linear Regression model and 2SLS model both improved the accuracy of

the Toimtuki to some extent, especially in the absolute difference percentage view, and there might be more potential significant variables which could help TOIMTUKI to be more accurate.

References:

TRIM3. The urban institute of US, 2012a. Available at: <*http://trim3.urban.org/T3IntroMicrosimulation.php*>. Accessed 2011.

TRIM3. The urban institute of US, 2012b. Available at: <*http://trim3.urban.org/documentation/Static%20versus%20Dynamic%20Microsimulation.html*>. Accessed 2011.

Anthony King, Hans Bækgaard, Martin Robinson(December 1999). *DYNAMOD-2: AN OVERVIEW*. Technical Paper no. 19, ISSN 1443-5098, ISBN 0858898004, in NATSEM, University of Canberra, Australia.

Lennart Flood, Fredrik Jansson, Thomas Pettersson, Tomas Pettersson, Olle Sundberg, Anna Westerberg(2005). *SESIM III- a Swedish dynamic micro simulation model*. In Handbook of SESIM051222, Swedish Ministry of Finance.

Rick Morrison(1998), Bernard Dussault(Edited 2000). *Overview of DYNACAN : a full-fledged Canadian actuarial stochastic model designed for the fiscal and policy analysis of social security schemes*. Statistics Cananda. Slightly adapted by Bernard Dussault(March 2000) for inclusion on the IAA website.

Honkanen Pertti(2010). *JUTTA-käsikirja*. Tulonsiirtojen ja verotuksen mikrosimulointijärjestelmä. Kela, Helsinki.

Appendix

htyotper:Basic unemployment allowance paid by KELA in Euros.

tyot: Number of month of person's unemployment or forced leaving.

tyotseu: Number of month of person's unemployment or forced leaving in year 2010.

ttyotpr: Unemployment allowance.

lpaktyva: Employee compulsory unemployment insurance.

vvvmk1: Paid earnings-related unemployment allowance.

vvvpvt1: Paid earning-related unemployment allowance days in total.