Current Development in Microsimulation and Experimental Innovation method in JUTTA Model

Meng Zhou 1,*, Yuting Men 2, Xuwen Qing 2
1 School of Sciences, Southwest Petroleum University, Chengdu, China
2 School of Mechanical and Electrical Engineering, Southwest Petroleum University, Chengdu, China

*Corresponding author e-mail: 201599010133@swpu.edu.cn

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-models in JUTTA is detected to have space to be more accurate. In order to do so, two experimental 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; Experimental Innovation method; 2SLS.

1. Introduction
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 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. 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 uses 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).

3. 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). 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.

There are three components in the Dynamic Microsimulation: methodology design, database
preparation and simulation procedure. The figure1 presents the basic structure.

![Figure 1. Presents the basic structure.](image)

3.1. Methodology design
The design for the model should suit for obtaining the desired results. Generally, the demographic items
such as age, gender, fertility, mortality are default variables in the simulation. The other necessary
variables are inputted in the original database according to which direction the model wants to estimate.
The directions are often classified into three main fields: Taxation, Pension and labor market.

A mainstream dynamic microsimulation in the sense is that the variables (events) are updated in a
sequence based on transition probabilities, and the space in time between the updating processes is a
year. Also, there exist the outliers that the events happen do not depend on the transition probabilities,
but survival functions, such as DYNAMOD-2. The transition probabilities are simulated by the Monte
Carlo method.
3.2. Database preparation

Normally, the initial survey data (register data) sets could not be applied to the microsimulation process directly, they need some modification. But, firstly, we will start choosing the suitable raw database from different resources.

Merging, editing, imputation and reweighting are the main tools helping with merging initial databases into one piece dataset, deleting difficult cases such as children whose parents could not be identified, imputing the unknown information according to known information such as children’ age specification from group classification, this could be imputed from their education levels, and the last but not the least, reweighting is automatically changed depending on the deletion and addition.

3.3. Simulation procedure:

It is the core in microsimulation model. In this step, the relevant demographic and economic events will be simulated. The common statistical methods used in this step are linear regression, generalized linear regression (GLM), mixed model (MIX) and generalized linear mixed model (GLMM).

In the simulation procedure, 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](image-url)

Figure 2. Describes us the life-cycle structure in the normal dynamic microsimulation model.

4. Application

In this section, the practical work is done by assessing the Finnish static model-JUTTA Model, also certain experimental statistical models will be discussed to improve the JUTTA model.

4.1. Description of the 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. 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)

4.2. Practical work design
After setting up the JUTTA model, an assessment system is necessarily built by evaluating the accuracy of the model in micro-level: personal level and household level, which depends on the unit type in particular sub-model. The practical framework has been divided into two parts: one is measuring the accuracy of the JUTTA model (including sub-models) and finding out the most inaccurate model, the other one is improving this model by using experimental statistical strategy.

4.3. 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 interval 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, health 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.

4.4. Experimental Innovation methods in TOIMTUKI
In order to improve Toimtuki’s accuracy, two experimental statistical methods are applied.

4.4.1. Method one: Regression method. The experiment step:
1. Giving the estimated value zero in case the real data is zero.
2. Fit the regression with the rest of the individual, and estimate the values.
3. Comparing this model with the JUTTA model by assessment the results.

Step 1: By SAS, directly give the estimated value 0 when the real value is 0.
Step 2:
The hypothesis variables that would affect the real value are: htyotper tyot tyotseu martul04 tyotkmuu lpaktyva tyotpr palkm vvvmk1 vvvpvt1 svatva maksvuok jasenia desmod lapsia. Where the meanings of variables are:

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.
martul04: An earning related unemployment allowance which is awarded in November.
tyotkmuu: Other unemployment compensation.
tyotpr: Unemployment allowance.
lpaktyva: Employee compulsory unemployment insurance.
vvvmk1: Paid earnings-related unemployment allowance.
vvvpt1: Paid earning-related unemployment allowance days in total.
svatva: Personal annual income.
maksvuok: Household rent for last month.
jasenia: Number of people in the household.
lapsia: Number of children in the household.
desmod: Decile, from 0 to 9. According to OECD, the average household consumption, each decile group has 10% people.

However after fitting the regression, the significant effecting variables are: tyot, ttyotpr, svatva, maksvuok, jasenia, with their p-values under 0.0001, we say that variable is statistically highly significant. So, in this model, tyot, ttyotpr, svatva, maksvuok, jasenia should be kept, hence, the table below is carried out:

| Variable  | Parameter Estimate | Standard Error | t Value | Pr>|t| |
|-----------|--------------------|----------------|---------|-------|
| Intercept | 195.67209          | 230.39284      | 0.85    | 0.3960 |
| tyot      | 158.40442          | 21.28756       | 7.44    | <.0001 |
| ttyotpr   | -0.17366           | 0.03029        | -5.73   | <.0001 |
| svatva    | -0.10823           | 0.01709        | -9.86   | <.0001 |
| maksvuok  | 2.02230            | 0.31497        | 6.42    | <.0001 |
| jasenia   | 1308.86787         | 183.66519      | 7.13    | <.0001 |
| lapsia    | -786.69217         | 198.34392      | -3.97   | <.0001 |
| desmod    | 520.13951          | 74.80554       | 6.95    | <.0001 |

Now estimates the values by the equation:
\[ \hat{y} = X\hat{\beta}, \]
where X is the characteristics vector and \( \hat{\beta} \) is the vector of estimated coefficients listed in table 7.

Now we plug the numeric in this equation:
\[ \hat{y} = 195.67209 + 158.40442 \times \text{tyot} - 0.17366 \times \text{ttyotpr} - 0.10823 \times \text{svatva} + 2.02230 \times \text{maksvuok} + 1308.86787 \times \text{jasenia} - 786.69217 \times \text{lapsia} + 520.13951 \times \text{desmod}; (R-Square=0.3744) \]

Step3:
The final results seem to be improved to some extent, see the table two:

| Model | Variable | [0, 1) | [1, 10) | [10, 100) | [100, 1000) | [1000, ∞) | [0, 0.1%) | [0.1%, 1%) | [1%, 10%) | [10%, ∞) |
|-------|----------|--------|---------|-----------|-------------|------------|------------|----------|----------|----------|
| JUTTA TUKI | 0       | 0.00296 | 0.02569 | 0.33103   | 0.64032     | 0          | 0.00494    | 0.03458   | 0.96047  |
| M1 TUKI  | 0.00161 | 0.01288 | 0.05314 | 0.37037   | 0.56200     | 0.00322    | 0.01771    | 0.08535   | 0.89372  |

So, from this method we see that the moth of unemployment, unemployment allowance, household monthly income, household rent, number of persons in the household, number of children and household decile number, they play significant roles in estimating the benefit for TOIMTUKI.

4.4.2. Method two: 2 Stage Least Squares Method. The experiment step:

1. 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 \( \pi_i \) is the probability of being status=1. Finally, generating
random value from the uniform distribution, and compare this value with $\pi_i$, the probability, if the random value is larger than the probability, giving status value 0, if not, giving value 1.

2. 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.

3. Comparing this model with the JUTTA model by assessment the results.

Step1:
The logistic regression is shown in table three:

| 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   |
| vvmk1     | -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 $\beta$ is the vector and its estimation has been shown above. 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 $\pi_i$, we give the estimated status 1 and 0.

Step2:
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. The regression is shown table 4:

| Variable | Parameter Estimate | Standard Error | t Value | Pr>|t| |
|----------|--------------------|----------------|---------|-------|
| Intercept| -152.15070         | 177.19949      | -0.86   | 0.3909|
| tyot     | 202.66061          | 17.99137       | 11.26   | <.0001|
| tyyotpr  | -0.13827           | 0.02625        | -5.27   | <.0001|
| svatva   | -0.07511           | 0.00867        | -8.67   | <.0001|
| maksvuok | 1.78485            | 0.26905        | 6.63    | <.0001|
| jasenia  | 394.85153          | 70.85511       | 5.57    | <.0001|
| desmod   | 341.67733          | 64.15313       | 5.33    | <.0001|

Then, plug the estimator to the equation:

$$\hat{y} = -152.15070 + 202.66061 \cdot tyot - 0.07511 \cdot svatva + 1.78485 \cdot maksvuok - 0.13827 \cdot tyyotpr + 394.85153 \cdot jasenia + 341.67733 \cdot desmod; (R-Square = 0.3930)$$

where $\hat{y}$ is the estimated TUKI. However, in this equation, $\hat{y}$ could be negative value, but in real case, it could not be, so when it is negative, we change it to value 0.

Step3:
Comparing the real TUKI and estimated one, see table 5:

**Tab. 5** Comparison of TOIMTUKI and Method 2

| Model      | Variable | Number of Observation | Absolute Error Percentage | Relative Error Percentage |
|------------|----------|-----------------------|----------------------------|----------------------------|
|            |          | [0, 1)                | [1, 10)                   | [10, 100)                  | [100, ∞)                  | [0%, 0.1%)     | [0.1%, 1%)    | [1%, 10%)     | [10%, ∞)     |
| TOIMTUKI   | TUKI     | 1012                  | 0.00296                   | 0.02569                    | 0.33103                   | 0.64032       | 0.00494       | 0.03458       | 0.96047       |
| MEHOD2     | TUKI     | 894                   | 0.00224                   | 0.02685                    | 0.42953                   | 0.54139       | 0.00224       | 0.01454       | 0.98322       |

From the table, 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.

Also, below table 6 shows us the comparison among TOIMTUKI, METHOD1 and METHOD2.

**Tab. 6** Comparisons of TOIMTUKI, Method 1 and Method 2

| Model      | Variable | Number of Observation | Absolute Error Percentage | Relative Error Percentage |
|------------|----------|-----------------------|----------------------------|----------------------------|
|            |          | [0, 1)                | [1, 10)                   | [10, 100)                  | [100, ∞)                  | [0%, 0.1%)     | [0.1%, 1%)    | [1%, 10%)     | [10%, ∞)     |
| TOIMTUKI   | TUKI     | 1012                  | 0.00296                   | 0.02569                    | 0.33103                   | 0.64032       | 0.00494       | 0.03458       | 0.96047       |
| TOIMTUKI(N1) | TUKI     | 621                   | 0.00161                   | 0.0128                    | 0.05154                   | 0.37037       | 0.56200       | 0.00322       | 0.01771       | 0.08555       | 0.89372       |
| TOIMTUKI(N2) | TUKI     | 894                   | 0.00224                   | 0.02685                    | 0.42953                   | 0.54139       | 0.00224       | 0.01454       | 0.98322       |

5. Conclusion

The study above firstly introduced static and dynamic microsimulation. Then, a Finnish static model-JUTTA has been assessed, the results demonstrated that JUTTA microsimulation model performed quite well in all sub-models, only except for the “residual” model-Toimtuki(income-related supplementary benefit). For Toimtuki, two experimental statistical methods: Linear Regression model and 2SLS model were applied, and they both improved the accuracy the model to some extent, especially in the absolute difference percentage point of view. Also, there might be more potential significant variables need to be found in future that could help TOIMTUKI to be more accurate.

Acknowledgement

This research is jointed supported by Finnish Social Insurance Institution and Science Department, Southwest Petroleum University, China. The numerical calculation in this article has been done on the super-computing system in Super-computing Center for science and engineering of Southwest Petroleum University.

References

[1] TRIM3. The urban institute of US, 2012a. Available at: <http://trim3.urban.org/T3IntroMicrosimulation.php>. Accessed 2011.
[2] TRIM3. The urban institute of US, 2012b. Available at: <http://trim3.urban.org/documentation/Static%20versus%20Dynamic%20Microsimulation.html>. Accessed 2011.
[3] Random Utility Models. : Mathematical Marketing: Chapter 13.
[4] Anthony King, Hans Barkgaard, Martin Robinson (December 1999). DYNAMOD-2: AN OVERVIEW. Technical Paper no. 19, ISSN 1443-5098, ISBN 0858898004, in NATSEM, University of Canberra, Australia.
[5] Azizur Rahman (2009). Small Area Estimation Through Spatial Microsimulation Models: Some Methodological issues. Second International Microsimulation Association Conference (2009), Ottawa, Canada.
[6] Azizur Rahman, Ann Harding, Robert Tanton, Shuangzhe Liu. Methodological Issues in Spatial
Microsimulation Modelling for Small Area Estimation. International Journal of Microsimulation 2010, 3(2) 3-22.

[7] Cathl O'Donoghue (2001). Dynamic microsimulation: A methodological survey. Brazilian Electronic Journal of Economics, 2001.

[8] Francesco Figari, Maria Iacovou, Alexandra Skew, Holly Sutherland (2010). Approximation to the truth: comparing survey and microsimulation approaches to measuring income for social indicators. Working paper in Institute For Social and Economic Research, UK.

[9] Honkanen Pertti, edited. JUTTA-käsikirja. Tulonsiirtojen ja verotuksen mikrosimulointijärjestelmä. Kela, Helsinki, 2010, Finland.

[10] Honkanen Pertti. JUTTA-käsikirja. KEIA, 2009, Helsinki, Finland.

[11] Kornstad Tom, Thoresen Thor O (2007). A discrete choice model for labor supply and childcare. Journal of Population Economics, 20: 781-803.

[12] Kornstad Tom, Thoresen Thor O (2006). Effects of family policy reforms in Norway: results from a joint labour supply and childcare choice microsimulation analysis. Discussion Paper No.450, March 2006. Statistics Norway, Research Department.

[13] 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.

[14] Marcia Keegan, Simon Kelly (2009). APPSIM-Dynamic microsimulation modelling of social security and taxation. Working Paper 14 in NATSEM, University of Canberra, Australia.

[15] McFadden, Daniel (1995). Computing willingness-to-pay in random utility models. Working paper from University of California at Berkely. Econometrics Laboratory Software Archive.

[16] 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 Canada. Slightly adapted by Bernard Dussault(March 2000) for inclusion on the IAA website.

[17] Steve Gribble, Chantal Hicks, Geoff Rowe (2003). The LifePaths Microtsimulation Model. International Microsimulation Conference on Population, Canberra, Australia, 2003.

[18] Zaidi Asghar, Harding Ann, Williamson Paul, eds. New frontiers in microsimulation modelling. Ashgate, 2009, ISBN 978-0-7546-7647-8.

[19] Lehtonen Risto. Practical methods for design and analysis of complex surveys. 2ndChichester, Wiley, Statistics in practice, 2004.

[20] Lehtonen Risto, Veijanen Ari. Design-based Methods of Estimation for Domains and Small Areas. In: D. Pfeffermann and C. R. Rao, ed. Hanbook of Statistics- Sample Surveys, Inference and Analysis, Volume 29B., 2009, 219-249, Amsterdam.