Business Informatics
Doctoral School

COLLECTION OF THESES

Dávid Burka

Supporting demographic forecasts and pension research through the application of microsimulation methodology

Ph.D. dissertation

Tutors:

Attila Tasnádi dr.
professor

László Mohácsi dr.
senior lecturer

Budapest, 2019

Department of Computer Science

COLLECTION OF THESES

Dávid Burka

Supporting demographic forecasts and pension research through the application of microsimulation methodology

Ph.D. dissertation

Tutors:

Attila Tasnádi dr.
professor

László Mohácsi dr.
senior lecturer

Contents

1. Former research, justification for the topic

In the last few decades the economical and social processes in developed countries went through such significant changes that have not happened since the Industrial revolution. Digitalization completely changed the structure of the markets, most industries are dominated by intellectual products instead of material ones. The improved living conditions caused by the development of health care and technology have in turn lead to the ageing of modern societies and the continuous decline of the population. The complex demographic changes create never previously experienced challenges for the political decision-makers. One of the most remarkable of these challenges is to ensure the financial sustainability of the pension system while securing the adequate living conditions of the elderly at all times.

Securing the proper living conditions for the elderly is necessary in an aging society. In countries where the expenses of the pension system are covered by the contributions of the active population, the financial sustainability of the system can only be ensured by major reforms. For historical reasons Hungary belongs to these group of countries as well.

The effects of starting data errors on the results of long term forecasts are increased by an order of magnitude compared to short term ones. In addition, it is necessary to analyze the distribution of the populations based on different properties in order to track the sociological changes.

Before the regime change of 1990 most of the employees were continuously employed, however now days a fragmented carrier path is increasingly common. This change will have a significant effect on the pension system in the next decades, thus the forecasts have to consider this process by all means as well. However, this phenomenon cannot be described by an analytically manageable function, thus it cannot be modelled by methodologies which work with indicators that are aggregated on the group level.

Because of the problems described above, “traditional” methods like the cohort-component method or the multi-agent based modelling are not appropriate. These usually cannot be described by a continuous function, but can be modelled on the individual level for example by a Markov-chain or a complex system of rules. Forecasts based on the microsimulation methodology are capable of managing theses kind of phenomena and they can be expanded without significant limitations. Above all this – compared to other methods – microsimulation can be used to answer questions regarding the distribution of the population. Thus with the help of this methodology the effect of the different suggestions for changes regarding the pension system become comparable.

The goal of this research is to create an analytical environment for demographic and economical forecasts based on microsimulation and to verify the connected methodologies. The goal is not to answer the actual social and economical problems, but to develop a methodology and demonstrate it through detailed impact assessments. A simulation framework capable of implementing the models connected to the research questions and performing the impact assessments has been designed and developed in order to support the research. Based on the above, this dissertation is looking to answer the following set of questions:

K1. Are there appropriate data sources in Hungary that are necessary to create long term, microsimulation based, demographic forecasts?

K2. Is it possible to create an easy to use, flexible and fast microsimulation framework that is capable of performing the impact assessment of the outlined economic and social challenges without exceeding the resource frame of standard office computer?

K3. Is it possible to examine the composition of the population regarding the connection between educational attainment and child bearing based on the statistical data sources available in Hungary?

K4. Is it possible to analyze the adequacy of the different pension systems along the distribution of incomes?

K5. Is it possible to analyze the financial sustainability of the different pension systems while considering external impacts?

2. Applied methods

This chapter will introduce the main characteristics and benefits of microsimulation, the methodology this dissertation focuses on. Furthermore, the methods used to overcome its drawbacks will also be described.

2.1. Microsimulation

The different forecasting methodologies differ from each other primarily in the nature of their compromise between the complexity of the models and the resource requirements (i.e.: runtime, computing capacity, invested workhours). The simplest estimations are based on macro time series – like the size of the population –, and these indicators are usually forecasted through trend fitting or regression models. Traditional demographic forecasts like the cohort-component model create groups, and the yearly iteration happens through the change of the states or counts of the groups. Beside transition probabilities, the agent based models simulate the connections and interactions between the agents (groups or entities) with the help of rules and algorithms based on empirical data.

Microsimulation follows the life path of every single entity – instead of grouping them. Iteration steps happen on the entity level, and the properties of entities are modified in every step according to the given rules, statistical probabilities or algorithms based on empirical data (Csicsman/Fényes, 2012). The foundation of this approach is the Monte Carlo method which is only effective if the appropriate sample size is available, thus a microsimulation model often contains millions of entities, unlike the case of "representative agents" (Dekkers et al., 2015).

Basically, a microsimulation model forecasts the properties of the individuals in time based on their other properties and different exogenous parameters of the system. (Csicsman/Fényes, 2012; Freudenberg/Berki/Reiff, 2016) Steps taken to ensure the sustainability of the pension system can significantly influence the course of demographic trends, furthermore some changes or reforms can have an effect on each other in a direct or indirect way. Thus the connections between some indicators have to be considered as well. As the number of properties used to describe the individuals increase it is less likely for the cohort-component method to be suitable for the given task.

There are phenomena that cannot be described by analytically manageable functions. In a function that describes a phenomenon through the life path of an individual a tear – or even multiple one – can happen at any given time. From the standpoint of the service period the maternal leave is good example for this. Managing these cases would require such a group breakdown that approximates the resolution of a microsimulation.

Regarding the comparability of models, the need to compare indicators on the level of groups emerges. In case of an agent model only group means and deviations can be determined, but proper distributions functions cannot. However, for example from the standpoint of the adequacy of the pension system the distribution of pension allowances is an important question. In apropos of the current child support program of the government the distribution of the mothers based on the number of their children could be interesting as well. With the use of microsimulation questions like these can be answered.

Microsimulation based solutions can manage the connections between the properties of individuals and the tears in the process, the appropriate breakdown of the forecasts is ensured, and questions regarding distribution can be answered as well. Thus, if the appropriately detailed data sources and the required computing capacity is available, the microsimulation is the best method to implement the demographic forecasts related to the impact assessments of the pension system.

2.2. Qualitative, exploratory research

One of the main restrictions of the realization of microsimulation is the lack of appropriately detailed datasets. For this reason, along the quantitative analyzes the qualitative, thorough examination of the available Hungarian data sources based on the legal framework is also a part of the dissertation. Analyzes include the data collection procedures, the privacy policy restrictions and the question of data quality as well.

2.3. System development

A main part of the research is the development of a simulation framework which allows the implementation of the outlined models.

The dissertation demonstrates the complete development lifecycle from the specification, through the special technical solutions, till the practical use. The actual version of the framework is available for everyone on GitHub (dburka001/SimulationFramework).

2.4. Statistical methods

Creating the model inputs and the related parameters requires the use of different, widely used, general statistical methods (i.e.: data cleaning, standardizing). Along these the connection between the different datasets is established through the statistical matching procedure.

Mortality and fertility rates are forecasted with the Lee-Carter model (Lee/Carter, 1992). It is not a goal of the dissertation to accurately estimate the indicators, this well-known model was chosen for the sake of transparency – regardless of it being criticized in many studies (Vékás, 2016).

In some cases, such historical data of the individuals is required that is not available for every year. For example, the complete income and contribution history of an individual is necessary to calculate their pension. The missing values of the starting population are procreated with the synthetic approach (Spielauer, 2009), thus the unknown properties of the individuals are determined by using a reverse simulation with the help of the model parameters.

3. Results of the dissertation

This chapter presents the results of the dissertation along the research questions introduced in the first chapter. The first two subchapters concentrate on overcoming the difficulties of microsimulation methodology. The following parts focus on the practical possibilities of the methodology, and their statements are confirmed by the models implemented with the help of the realized simulation framework.

3.1. Hungarian data sources

Microsimulation based demographic forecasts require exceptionally detailed data collections. Practically the methodology itself has no limitations on the number of examined properties, however the absence of appropriately detailed data still sets some boundaries for the simulations. During the planning of the forecasts understanding these limitations and the available data sources as accurately as possible is essential.

In Hungary, practically all the data collections that are relevant from a demographic standpoint are regulated by laws or decrees. Because of this, most of the indicators are available in the form of long term time series which facilitate the forecasting of the models through statistical methods.

The breakdown of the available data sources is not always suitable for use in a microsimulation. The datasets required for building the starting population is only available from the years of cross-sectional data collections (i.e.: census, micro census). In addition, part of the registered medical data cannot be considered accurate because of the way it was gathered. Many people – probably from the minorities – refused to answer the since 1992 optional questions about ideology and nationality, thus there is no accurate data collection from the distribution of these groups either.

There is no unified national data warehouse in Hungary, and the privacy policies prohibit publishing datasets that are detailed enough that an individual could be identified from them. Thus the datasets stored at the different data administrators (i.e.: HCSO, NHIF) can only be merged through statistical methods (statistical matching), and so can only be used by considering the prevailing estimation error.

3.2. Microsimulation framework

The microsimulation methodology is fading into the background compared to other forecasting methodologies because of the absence of comparability between different models and the implementation difficulties. An easy to use and fast microsimulation framework could overcome these problems. However, the available generalized solutions are not flexible enough, furthermore they are usually considered more like a collection of useful features, than a framework.

Mohácsi examined the fields of application of parallel architectures, and among others he displayed the possibilities of the methodology with a simple demographic simulation. (Mohácsi, 2014) The framework described in this dissertation grew from upgrading his model.

In order to meet the objectives stated for the framework it is necessary for the solution to be flexible, fast and easy to use at once. However, these are some conflicting conditions, and satisfying one can only be done at the expense of the other two. The foundation of frameworks realization was to find the balance between the three conditions while keeping the resource requirements low enough not to obstruct application.

Ensuring **flexibility** is logically essential in a basically generalized framework. A high level programming language is the most flexible solution, in which the whole simulation is built by the users and the system only supports them through pre-implemented auxiliary functions. Naturally, the more freedom is given to the user the higher level software development skills are required to operate the framework.

The **speed** of the finished simulation can also be significantly influenced by the development skills of the user. In case of the Hungarian example the demographic forecast requires the simulation of nearly ten million individuals. Thus, it is important to be careful during the development, otherwise a simulation could easily take multiple days to finish. Because of the projective nature of microsimulation, the main goal is to compare the impact assessments of multiple different scenarios, so it is necessary to run multiple different configurations. Because of the number of configurations speed becomes even more important, since a properly developed system can be multiple magnitudes faster, and its runtime can be measured in hours instead of weeks.

Ease of use – in addition to making it simple to implement the models examined in this dissertation – allows other researchers and demographers to utilize the features of the framework. Developing a generalized system is obviously more complicated than developing a specialized solution for a given task. The more flexible a system, and the more freedom it gives to the user, the more difficult it becomes to keep runtimes at a reasonable level.

It is necessary to fulfill and balance these three conditions together, in order for the framework to satisfy the previous expectations. The biggest problem is that a user less proficient in software development cannot be expected to ensure proper speed while designing a model. From a technological standpoint, the greatest challenge while implementing a demographic simulation are the modules responsible for iteration and in connection the parts of the simulation in which new entities are created or deleted, since these are worth parallelizing. Thus the framework solves the implementation of these features in the background, and the only thing it leaves for the user is the – from a software development standpoint simple – algorithm that executes the iteration step which is responsible for the change of the properties of a given individual.

Building the iteration step algorithm is possible in the framework through a graphical programming interface – similar to the ones used in the education of younglings (i.e.: Scratch, Google Blockly). This interface contains simple blocks that correspond to basic configuration options from which the iteration step algorithm can be built. The particular blocks are often hiding exceptionally complex program code from the user. At the start of the simulation based on the algorithm created by the user the framework builds, interprets and runs the code of the simulation. With this approach the system stays flexible, since the most relevant part of the models, the iteration step can be modified without limitations, yet a simple procedural code is enough to implement it. Meanwhile the core of the model runs on a parallel architecture, furthermore the code generated at runtime makes it possible to remove most of the unnecessary logical checks – further improving the runtime of the simulation.

3.3. Analysis of the population composition

Depending on educational attainment and the number of their children, there is a significant difference between the child bearing trends of women, so for the forecasts it is advisable to break the population down according to these variables. The analysis of the distribution of families based on the number of children is essential for the decision-making process regarding the child support program, while the composition of the population is relevant for questions regarding the pension system.

In the following such a microsimulation model will be presented that considers the carrier path of the individuals new to the traditional properties (age, gender). The social position of individuals is strongly influenced by their income level which is connected to their educational attainment. For this reason, the current model assumes that the income – and along that the prosperity – of individuals is exclusively dependent on their educational attainment, however the social status influences every other aspect of their lives as well.

Considering that according to data collections the educational attainment of children is strongly correlated to the one of their parents (Andor/Liskó, 2000), the education of an individual in the model is decided – in a deterministic way – at the time of their birth. From the models standpoint the education path of a child is determined by a probability distribution based on the educational attainment of their mother.

The so called carrier model got to be implemented in three different ways: The **base** version is based on actual data. In the **pessimistic** version the children do not try to exceed their parents so with a high probability they do not choose a higher educational attainment. However, in the **optimistic** model the children will almost certainly reach a higher level than their parent.

Presuppositions of the carrier model

- The probability of death is dependent on the age, the educational attainment and the gender.
- The probability of giving birth is dependent on the age, the number of children and the educational attainment.
- The dependency of death and birth probabilities from educational attainment is time invariant.
- The income – and consequently the prosperity – is exclusively dependent on the gender and educational attainment of the individuals

- The educational attainment of an individual is not influenced by events during their lifetime, thus it is predetermined at the time of their birth.
- The educational attainment of an individual is dependent on their gender and the educational attainment of their mother.
- Mothers do not continue their education after their first child.

The different carrier versions change the composition of the simulated population. The effect of the application of different probability distributions is infinitesimal at the start of the simulation, but after fifty iteration significant changes appear between the individual carrier variants (Figure 1). It is clearly visible that the three model versions reacted to the change of the carrier probability distributions in an expected way, so the effect of the changes is likely comparable for the other properties as well.

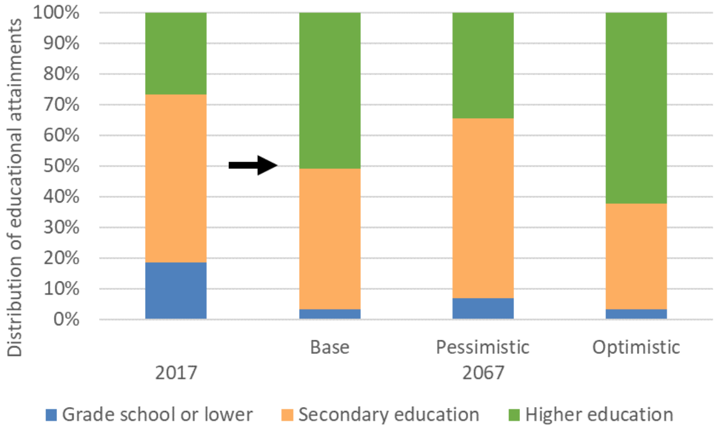

Figure 1 - Average distribution of educational attainments

The results of the carrier models support the theory that the available data is suitable for building such microsimulations through which the distribution of the population can be forecasted along multiple properties. Methodologies that are based on the grouping of the population can only roughly estimate such indicators.

3.4. Adequacy of the pension system

The difference between social groups is implemented via the results of the carrier models. The educational attainment of the majority is determined by the base carrier model. Such convergence is typical for the minority groups in society that make it difficult for the younglings to leave the group. According to this in case of the minorities – based on the pessimistic carrier scenario – the model assumes that the children will choose a similar carrier path to their parent whit a high probability. In case of immigrants a similar convergence can be observed, however a community that left their usual environment is probably more ambitious than the other minorities. Thus, the educational attainment of the immigrants will rather be represented by the optimistic model.

Presuppositions regarding demography

- Every presupposition outlined in the carrier model is also valid here.
- Social affiliation does not influence the birth probabilities. The different birth patterns of the groups can be traced back to the different living – or in this case educational – patterns.
- Social mobility is very limited in smaller communities thus in minority groups it is unlikely that a child reaches higher educational attainment than their parents.
- Immigrant groups are more prone to changes thus their children have a higher probability of exceeding the educational attainment of their parents.

- Social mobility is restricted to the social groups, there is no movement between individual groups. There is no assimilation.

One of the goals of this dissertation is to show that the long term effects of pension reforms can be analyzed with the help of the microsimulation methodology. Developing actual reforms or pension models would be outside of the scope of this dissertation by far. The work of András Simonovits titled “Pension models” (Simonovits, 2007) serves the purpose of familiarizing people without an economist degree with the results of pension research through a few simple model. The following pension models and presuppositions are all based on this article. Three different pension models will be introduced: one proportional to service time, one with a fix base pension and a hybrid solution that combines the base pension with the half of the proportional. All three models extend the above described social group model with the appropriate regulations.

Presuppositions regarding income and pension

- There is no inflation and the pension is based on non-valorized values.
- Incomes and their growth rate is mainly based on the individuals educational attainment.
- Except in the year of start of the carrier employment status is only dependent on the status of last year.
- Women always go to a year of maternal leave after giving birth.
- Incomes increase according to the years spent in employment.
- There is a fixed contribution rate for every individual.
- Every individual can work from the end of their education until the pension age limit, so there is no early retirement.

- The goal of the pension system is to maintain the individuals' standard of living compared to the one before retirement.

In the models the income is based on the educational attainment and the time spent in employment. The changes in employment since the regime change resulted in often heavily fragmented service periods. Thus while forecasting the incomes – and consequently the pensions – it is essential to include the service period as well.

The modelling of the service period is based on the microsimulation model introduced in the 2010 report of "Nyugdíj és Időskori Kerekasztal" (Holtzer, 2010). The employment time of a given individual in a given year is determined by a probability distribution based on their last employment status.

At the start of the simulation (2017) the proportional pension system has the highest rate of low pensions. During the simulation the distribution is continuously changing thus the result will be different (Figure 2). With the years in the case of the base pension the income of the pensioners increasingly falls behind of the working population. This happens because even without inflation the average incomes increase as the active population increases their educational attainment. In case of the proportional model the pensions follow the changes of the working population. Thus the adequacy of the system is hard to sustain in the base pension model.

Examining the adequacy rate that is calculated as the ratio of average income and the average pension (Freudenberg/Berki/Reiff, 2016) the base pension model has the worst performance as well. However, the graphs describing the individual models are practically parallel, but the analysis of the distribution showed that the base model was only forced back with time.

Thus distributions recovered from the microsimulation approach could contain information that would be lost if only the aggregated indicators would be available.

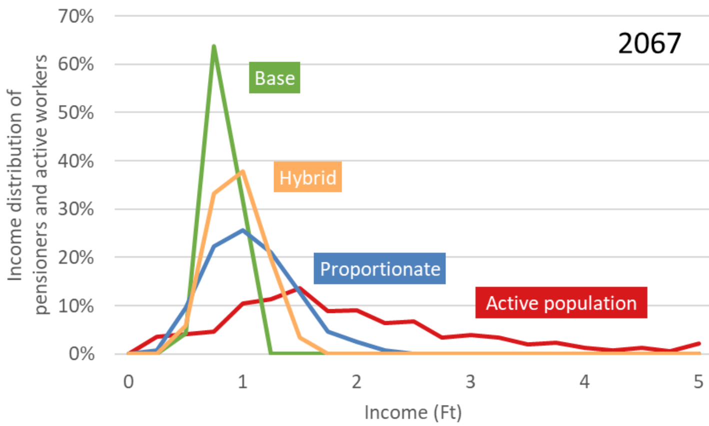

Figure 2 - Distribution of income at the end of the simulation

3.5. Financial sustainability of the pension system

In addition to ensuring adequacy it is necessary to preserve the financial sustainability of the pension system as wall. The net income of the individual pension models is continuously decreasing even with the yearly 10.000 working aged immigrants (Figure 3). Not surprisingly the least adequate base model has the most positive budget estimation. Since the majority of the population is at least on the secondary education level at the start of the simulation the models containing proportional parts result in higher pension benefits, thus these reach the level of contribution incomes much faster. The end of the simulation shows a slight improvement but it is impossible to determine whether this is a lasting improvement or not.

Some – from a model standpoint exogenous – phenomena may arise that could have a positive effect on the budget of the pension systems (i.e.: current child support program). The exogenous phenomena cannot be accurately forecasted, only expert estimations can appear in the simulations. Thus it is essential to run multiple different scenarios.

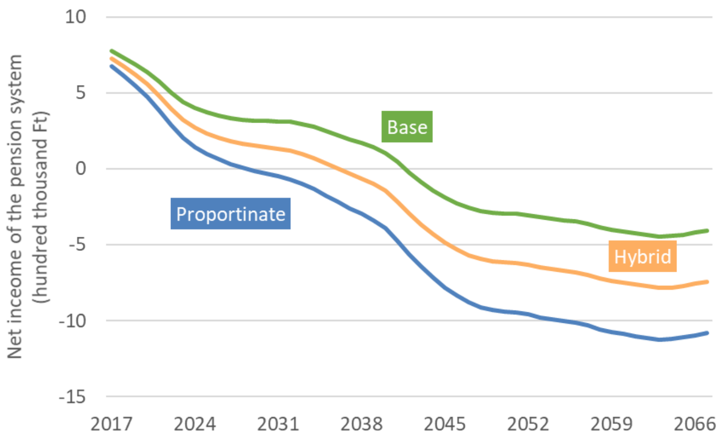

Figure 3 - Net income of the pension systems

In the examples examined in the dissertation a short non-lasting economic growth resulted in the improvement of the financial sustainability of the pension system, but – in the proportionate models – this effect reversed with time. The increase of the age limit obviously improved the state of the budget, but based on the pension system a significantly different scale of improvement could be measured. The explanation of the changes in the budget is not trivial even in these simple models, and in a more complex model that contains multiple feedbacks it would be even harder to see the effects ahead. Thanks to its speed the realized framework makes it possible to compare multiple different scenarios, thus microsimulation is suitable to accomplish impact assessments related to problems like securing the financial sustainability of the pension system.

3.6. Conclusion

The pension systems introduced in the dissertation are from the capacity limitations of a microsimulation based demographic – or pension – forecast. Practically a model of any complexity can be implemented with this method, and with the help of the framework many different scenarios can be examined in a reasonable time.

All of the featured models had a starting population of about 10 million individuals based on Hungarian statistical data of 2017, and simulations ran till the year 2067. For the sake of comparability of the runtimes (Table 1) every simulation was executed on the same personal computer (Intel Core i7-8750H CPU @ 2,2 GHz; 16 GB RAM; 64 bites Windows 10 operation system).

Model	**Runtime**	**Nbr of simulations**
Demographic base	0:06:05	1
Carrier	0:21:01	3
Social group	0:07:02	1
Pension systems	0:31:52	3
Economic growth	0:32:15	3
Age limit increase	0:32:16	3
Total	2:10:31	14

Table 1 - Runtime of the simulations

The two main drawbacks of a microsimulation based demographic forecast is that it requires appropriately detailed datasets and high level software development skills. Based on the results of the dissertation the data sources available in Hungary – despite their flaws – are suitable for realizing microsimulation based models, and the technical difficulties can be overcome with a proper simulation framework.

With the help of the realized framework it is possible to build simulations containing complex logical connections and feedbacks in a short time. It is possible to analyze the distribution of the forecasted population based on different properties, or even the complete life path of an individual. Thus such complex models can be realized that are capable of answering the questions arising from the different theories.

4. Main references

Andor, Mihály/Liskó, Ilona (2000): Iskolaválasztás és mobilitás. Iskolakultúra

Csicsman, Josef/Fényes, Cecília (2012): Developing the Microsimulation Service System. Hungarian Electronic Journal of Sciences

Dekkers, Gijs et al. (2015): On using dynamic microsimulation models to assess the consequences of the AWG projections and hypotheses on pension adequacy: Simulation results for Belgium, Sweden and Hungary. Federal Planning Bureau - Central Administration of National Pension Insurance

Freudenberg, Christoph/Berki, Tamás/Reiff, Ádám (2016): A long-term evaluation of recent Hungarian pension reforms. MNB Working Papers – Technical report

Holtzer, Péter (2010): Jelentés a Nyugdíj és Idoskor Kerekasztal tevékenységérol. Budapest: Miniszterelnöki Hivatal

Lee, Ronald D/Carter, Lawrence R (1992): Modeling and forecasting US mortality. Journal of the American statistical association, 87, Nr. 419, 659–671

Mohácsi, László (2014): Gazdasági alkalmazások párhuzamos architektúrákon. Ph.D thesis, Budapesti Corvinus Egyetem

Simonovits, András (2007): Nyugdíjmodellek. Magyar Tudomány, 168, Nr. 12, 1527–1533

Spielauer, Martin (2009): Microsimulation Approaches. Statistics Canada working paper

Vékás, Péter (2016): Az élettartam-kockázat modellezése. Ph.D thesis, Corvinus University of Budapest

5. Personal publications regarding the topic

Burka, Dávid (2016): Supporting the Hungarian Demographic Pre-calculations with Microsimulation Methods. SEFBIS Journal, 10, 13–23

Burka, Dávid (2017a): Comparing different scenarios of Hungarian demographic processes. SEFBIS Journal, 11, 17–26

Burka, Dávid (2017b): Developing a complex microsimulation system to project the Hungarian demographic processes. Mediterranean Issues, 1, 367–378

Burka, Dávid et al. (2017): Supporting Pension Pre-Calculation With Dynamic Microsimulation Technologies. ECMS 2017 Proceedings, 31, 562–569

Ágoston, Kolos Csaba et al. (2019): Klaszterelemzési eljárások halandósági adatokra. Statisztikai Szemle (Megjelenés folyamatban)

Csernai, Márton et al. (2019): Dynamics of sleep oscillations is coupled to brain temperature on multiple scales. The Journal of Physiology (Megjelenés folyamatban)