UNIVERSITÀ POLITECNICA DELLE MARCHE

Research notes
n.296

Microsimulation Models.
An Integrated Approach with Real Data.

Francesco Marchionne
Faculty of Economics – Marche Polythecnic University
Piazzale Martelli, 8 – Ancona, Italy
Email: markionne@tin.it)

August 2007

Abstract

Due to the increasing the calculus power of computers, a growing number of economic phenomena are being studied through microsimulation. However, this methodology is not updated, and the basic structure of these models is till tied to outdated technologies where economic hypothesis was used to simplify computational complexity as the calculus power of computers was not strong enough. In this paper, I suggest an innovative approach to the microsimulation. From a theoretical point of view, it should be more efficient than the traditional approach. The paper is divided in three section. In the first section, I explain this innovative approach and show its operating differences respect to the traditional approach. In the second section, I explore the difficulties of its concrete implementation through a case study and recommend some technical solutions in order to overcome them. In the last section, I summarize the main results.

Methodology

Definition and classifications

Microsimulation is the most power-intensive technique in terms of calculus resource and data required in economic analyses that use personal computers extensively. Introduced for the first time in 1957 [Orcutt 1957], microsimulation was used poorly for over 25 years because of the absence of datasets and of limited calculus power of computers. However, starting from the 80s these obstacles have been overcome. Its applications have increased in number and have become more interesting extending from the field of economics [Mertz and others 1991], to sociology [Orcutt and others 1986], and until recently, also to geography [Clarke 1986]. Also in the field of economics, there is diversity in the topics taken up. At the beginning, in fact, research was focused on taxation and income distribution [Atkinson and others 1988, Monteduro-Di Nicola 2004], in particular income from jobs [De Lathouwer 1996]. Then they extended to education on the one hand [Bourguignon and others 2003] and to the pension system on the other [Miniaci 1998, Atkinson and others 2002]. With the increase in computer power, microsimulation found wider fields of application. So, there was a need to go deeper into the implementation method so as to make clear the behaviour, potentialities, and limits of microsimulation and avoiding that it becomes a semi-unknown black box.

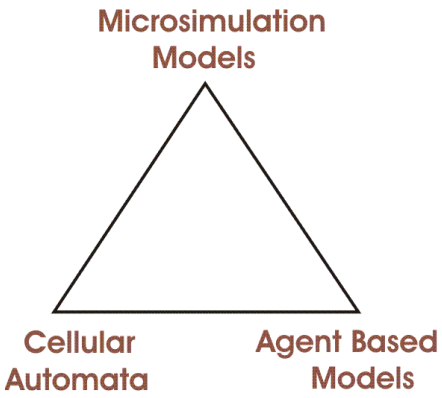
In brief, Microsimulation (an acronym for "Microanalytical Simulation") is a modelling technique based on a computer. It operates at the level of individual units such as people, households or firms. In this model, each unit is represented by a record containing a unique identifier and a set of associated attributes (e.g. a list of people with known age, sex, marital and employment status). A set of rules (transition probabilities) are then applied to these units leading to simulated changes in state and behaviour. These rules may be deterministic (probability = 1), such as changes in tax liability resulting from changes in tax regulations, or stochastic (probability ≤ 1), such as chance of dying, marrying, giving birth or moving within a given time period. In either case the final result is an estimate of the outcome after applying these rules, possibly over many time periods, including both overall aggregate change and, more significantly, the distributional nature of any change. The last aspect is crucial as it marks the distinctive characteristic of this technique. Given the emphasis on changes in distribution, in fact, microsimulation models are often used to investigate the impact of fiscal and demographic changes (and their interactions) [1] on social equity.

Microsimulation (MSM) comes very close to two other individual-level modelling approaches: Cellular Automata (CAs) and Agent Based Models (ABMs). In their originally conceived forms these three approaches may be regarded as those that represent the three corners of a continuum of individual level modelling approaches.

In a pure CA all entities are spatially located within a grid of cells, and all entities have only one attribute (alive or dead), with behaviours deterministically dependent upon the state of neighbouring cells [Shelling 1971]. In a pure ABM, the emphasis is on the interaction between individuals, with the main attribute of each individual being his operating characteristics (behavioural rules), which evolve stochastically over time in response to the success or failure of interactions with other individuals [Terna 2002]. In

[1] Microsimulations are also applied in other scientific fields such as in road traffic density analyses or in biology.

a pure MSM, transition probabilities lack evolutionary and spatial dimensions [Bourguignon-Spadaro 2005].



When microsimulation models add more behavioural and spatial interaction between individual units, when CAs add a growing range of individual attributes and start to incorporate aspatial behaviours, and when ABMs add both space and fiscal/demographic characteristics to their agents, the three approaches move towards a common ground. Thus, from an empirical point of view, the main characteristic of microsimulation is to use a sample of microdata built by single economic units (microeconomic panel), instead of dealing with information at an aggregated level according to macroeconometric models. It brings about some advantages, but it also imposes some limits. So, the building of these models essentially depends on the phenomenon under investigation.

Conceptually, microsimulation models can be divided into static and dynamic models [Caldwell 1990]. The latter are once again subdivided into dynamic population models (dynamic *cross-section*) and dynamic cohort models (dynamic longitudinal). The distinction affects the type of used data and the projection method of the starting sample in the future [Niccodemi 1998]. In static models, the structure and the unit number of the starting sample remain unchanged throughout the simulation and the updating procedure consists only in the re-weighting of sample data (*static ageing*[2]). In dynamic models, instead, there is continuous updating. So, the sample units in the former period become the starting point in the latter one. In this way, the unit number is not constant as in static models, but changes in time through stochastic Montecarlo procedures based on defined transition probability (*dynamic ageing*). Within dynamic models, instead, the distinction depends on the procedure of simulation [Vagliasindi 2004]. In dynamic population models, all the individuals in one period are simulated before moving to all the individuals in the next period (proceeding by periods). In dynamic cohort models, all the periods for the same individual are simulated before moving to all the periods for the next individual (proceeding by individuals). Therefore, a cohort dynamic model does not allow interaction between individuals as the latter unit is defined only when the former unit finishes its loop (implicit independency between units).

As mentioned before, the choice for the model basically depends not only on the topic of analysis but also on the amount and quality of available data. In economic matters, microsimulation is used to study the impact of demographic transformation, but it is

[2] After re-weighing, every microunit represents a share of population different from the initial one but the total unit number is unchanged.

also particularly fitting to analyse the phenomena tied more or less directly to personal or family income such as fiscal or pension reforms. The choice for the models also depends on the length of the simulation period. In a medium-long term period, a dynamic model is better than a static one because its evolutionary structure enables to focus on events that characterise the life of an individual unit. This allows a better specification of the model. Among these kinds of models, the cohort dynamic model is more appropriate to life-cycle problems even if it gives rise to less accurate outcomes than does a dynamic population model that, allowing interaction between units, achieves better performances but is more resource-intensive.

Limits and potentialities

Microsimulation provides remarkable advantages in terms of the traditional aggregated analysis. The main advantage is the ability to highlight not only the total affect but also the effect at each economic unit level. In this way, it will be possible later to easily combine units in limited subject groups to better highlight some particular effects.
There are also other elements that consider this technique preferable to other ones. First of all, a microsimulation model, not aggregating data, does not suffer from loss of information that affects macroeconomic models and is not forced to assume restricted hypothesis on individual behaviours and on the distribution of interest variables such as income. Secondly, it allows to overcome some difficulties that arise using macroeconometric models of equilibrium or models based on limited groups of "typical" individuals [Vagliasindi 2004]. For instance, as regards income, macroeconometric models reproduce "functional" distribution, but not "personal" distribution. Instead models with "typical" units cover a limited part of the economic agents. So they study only a restricted part of heterogeneity in microdata. Thirdly, microsimulation allows to compare alternative policies within the same demographic and socio-economic scenario or different scenarios for the same economic policy. Fourthly, working with microdata, a better representation of heterogeneity and interdependence can be obtained [Niccodemi 1998]. In the end, its modular structure allows wide flexibility because every aspect of realty is simulated by a different section. However, in spite of the powerful calculators the complexity of realty makes any virtual copy of the world impossible and therefore simplified hypotheses are always necessary to isolate and focus on the phenomenon that is to be analysed[3].
Nonetheless, a microsimulation model has some weak points also. First of all, during a simulation, it does not take into account behavioural reactions after modifications of the surrounding environment [Ballas-Clarke-Turton 1999]. Generally, this distorts the estimates as individuals continue to behave in the same way also in the face of significant changes in the context in which they take their decisions. Secondly, the construction of a microsimulation model requires heavy investment in human capital and in calculators as it has to project, program and update the model continually until a malleable and efficient instrument is obtained. Thirdly, increase in the reliability of a model raises its complexity in an exponential manner because it is mode difficult to identify errors and to control the results. This leads to a considerable waste of energy which is faced with simplified assumptions. These in turn can lead to negative consequences of not having precise results and of their not being comparable with other models (as regards interest variables and accuracy of results). Fourthly, being a

[3] In general, this happens almost in the same way as in all economic models.

stochastic procedure, the results obtained are not unambiguous. They are however more probable than the others and give important indications on the studied phenomena [Niccodemi 1998]. Finally, last but not least, a big quantity of information is required even though data are often limited in quantity and quality or their use is expensive and inconvenient.

General structure of a microsimulation model

The structure of an economic microsimulation is generally traditional and standard. However, there is no explicit codification of the procedure. In fact, the steps to follow to build this kind of models are so simple to be universally accepted [Vagliasindi 2004].
At the beginning, an appropriate phenomenon is focused on. Generally, when it is not a demographic phenomenon, it is linked to income because income is a good economic variable for this kind of analysis. The next step is to choose an individual unit. For example, if the evaluation of a new production tax is the matter, the firm will be adopted as unit. Otherwise, if the aim is to understand the impact of a benefit on family income, the family will be the unit. Lastly, if the aim is to analyze pension benefits, the unit will be merely a single individual. When the unit to be analysed is established, the characteristics that affect the main interest variable (income) are identified. Economic literature usually provides a strong theoretical approach to search for these control variables. Finally, the characteristics are made to "evolve" on the basis of implicit rules in transitional probabilities. After building a developmental path for the variables we are interested in, the last step is to analyze the examined phenomenon using simulated data.
There are two weak point in this procedure: i) starting data and ii) evolutionary phase for the control variables. Initial distribution of a variable has an essential importance in microsimulation. We can compare the starting data of a microsimulation model with the initial conditions of a differential equation. Theoretical solution is the simplest because it assigns starting values of each variable artificially by extracting them from a defined analytical distribution which we assume to represent the real distribution. In this way, the starting point is built ad hoc without sample distortions, missing values or excessive anomalies. Therefore, it turns out to be the ideal substrate for the implementation of microsimulation. It is clear that this procedure marks a discontinuity point with real data and puts microsimulation in a totally virtual environment binding it to be almost purely theoretical.
The empirical solution that starts from real available data eliminates this difficulty but introduces a lot of other problems. Real data, in fact, suffer from errors of measure, are often incomplete, and sustain sample distortions [Baltagi 1995]. All this can affect negatively on microsimulation and empirical works generally prefer the first solution [Klevmarken 1997, Borella 2004, ecc.], fitting, as much as possible, a theoretical density function into real data distribution. This procedure involves the assumption of specific distribution hypotheses that are difficult to verify empirically and to justify theoretically.
The choice for the solution depends on the quality and the quantity of available data and it often pushes the researcher towards taking forced decisions. As information at unitary level are necessary, the data should have maximum level of detail for each control variable. These high requirements are supplied only by few datasets and this limits the application of microsimulation very much. In the empirical section of this work, I have attempted to solve the questions linked to the input of real data through two original and

innovative procedures that apply more flexible and less tightening hypotheses than assumptions generally used starting with artificial data.
The second weak point of the implementation procedure in microsimulation models is identifiable in the evolutionary phase of control variables. The set of these variables specifies a unit state that is the overall condition in which the unit occurs at a precise moment. In formal terms, the unit state $\boldsymbol{X}$ is the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ of control random variable used to describe the attributes of the units. Defining the evolutionary path for these variables means assigning them a value at each moment during microsimulation. This is done by using the probabilistic "rules" embodied in the transition matrices. When the "n" variables in time "t" are given and the state $\boldsymbol{x_t} = (x_1, x_2, \ldots, x_n)$ is defined, the transition matrix is the $n \times n$ matrix of cumulate conditional probabilities. Each probability is that of moving from state $\boldsymbol{x_t}$ at the time t to state $\boldsymbol{x_{t+1}}$ at time $t+1$. Using the Montecarlo method and the transition matrices, state of the latter periods can be detected starting from the state of the former ones, so that all control variables are defined. However, transition matrices can be implemented in different ways but empirical works are standardized on sequential approach [Niccodemi 1998, Ando-Nicoletti Altimari 1999, ecc] which has now become "traditional". This solution is not optimal and can appreciably affect the outcomes achieved on the basis of the situation.

Traditional sequential approach.

In the traditional approach [Vagliasindi 2004, Borella 2004, ecc.] control variables are simulated sequentially one by one. Practically, after an order is defined, the program proceeds to simulate the attributes of the units on the basis of that order. Every characteristic has its transition matrix to which it will refer to when the sequence arrives to it. For example, only two characteristics for the units is hypothesised. They are A and B. Attribute A can assume a value equal to 1 or 0 with probability p(A) and 1-p(A) while attribute B will have probability p(B) to be 1 and 1-p(B) to value 0. If order A-B is established, the simulation will determine before the attribute A through transition matrix p(A) and 1-p(a), and then attribute B with matrix p(B) and 1-p(B). The following matrices can be used to represent cumulate probabilities of A and B.

A = 1	A = 0
p(A=1)	1

B = 1	B = 0
p(B=1)	1

Obviously, if the characteristics are not a random-walk in the period but path-sensitive (such as the great part of the cases), the conditional probability of previous period situation will be important and not the simple frequency distribution. It means that transition matrices will have a form like:

	A = 1	A = 0
A = 1	p(A=1|A=1)	1
A = 0	p(A=0|A=0)	1

	B = 1	B = 0
B = 1	p(B=1|B=1)	1
B = 0	p(B=0|B=0)	1

To simulate the sequence A-B or B-A is indifferent if A and B are independent. If it is not, however, the order will impact on transition matrices introducing an arbitrary element. The transition matrices, in fact, will depend not only on the same characteristic of the previous period but also on the other variable. It will be intuitively as follows:

	A = 1	A = 0
A = 1	p(A=1|A ∩ B)	1
A = 0	p(A=0|A ∩ B)	1

	B = 1	B = 0
B = 1	p(B=1|B ∩ A)	1
B = 0	P(B=0|B ∩ A)	1

The correlation between unitary characteristics creates two problems in the sequential approach. The first is a time-series difficulty in determining the same variable. The second is a cross-section difficulty in determining the other variables. From a temporal point of view, to define some attributes (and not others) makes conditional probability of examined variable dependent on sequence order because in the conditioning characteristics there will be some attributes that will already have been simulated (the former in the sequence) and others not (the latter in the sequence). Practically, the simulation of A in the current period affects the probability to be used for B at the same moment, "*t*".
In literature, the solution adopted in these cases is to isolate each period avoiding the intermediate update of unit state [Niccodemi 1999, ecc.]. The new state is assigned together to all the variables only when the period is finished. From a spatial point of view, the correlation between characteristics behaves in analogous way. To be B in the former period changes the probability to be A in the latter period. In realty, if the variables are path-sensitive, then a given B in time *t* would affect all transition probabilities of all variables in residual periods. There would be a serial correlation problem also in this case. Here, the conditioning of present B to future A only, is considered in order not to complicate the explanation too much. A sequential conditioning is used to account for this situation in the model. Each characteristic is simulated taking a transition matrix "conditioned" by the situations of other characteristics (that are "B" in our case). Continuing with the previous example, therefore, we will have a transition matrix for A in case of B = 0 and another one in case of B = 1. Obviously, in this way, a "tree" of conditional transition matrices is created. The "tree" divides itself more and more when simulated characteristics increase (vertical direction) or when the modalities for each attribute increase (horizontal direction).

	A = 1	A = 0
A = 1	p(A=1|A=1)	1
A = 0	p(A=0|A=0)	1

A = 0	B = 1	B = 0
B = 1	p(B=1|B=1 ∩ A=0)	1
B = 0	p(B=0|B=0 ∩ A=0)	1

A = 1	B = 1	B = 0
B = 1	p(B=1|B=1 ∩ A=1)	1
B = 0	p(B=1|B=0 ∩ A=1)	1

Sequential conditioning does not theoretically lead to distortions because order A-B should yield outcomes that are not substantially different compared to order B-A. However, to increase building of conditional transition matrices means having a lot of detailed data available because the reference "sample" becomes considerably thinner during each conditioning.
A lot of works use sequential conditioning in a partial way [Borella 2004, etc] introducing practically discretionary elements. If the probability of the "*k*" characteristic

is not conditioned with respect to all the previous ones, then the hypothesis that the "*k*" probability is independent of the "*k-1*" characteristics is implicitly assumed. If it is conditioned with respect to only some characteristics but not to other relevant ones, a distorted transition matrix is used. Lack of data moves many works towards partial conditioning. It is not wrong *a priori* but it becomes an error when independence is hypothesised which does not exist. It could happen that if the connection between the two variables is unknown. All this generates ambiguous models in which the simulation mechanism is a black box with a well-known final output but also with quite an unknown internal operation.

The traditional structure has clear inefficiencies also from a technical programming point of view. Building of a sequential evolutionary model involves an enormous waste of energy because each new characteristic added to the model must be put into practice in all its conditionings extending difficulties of its implementation in an exponential way. It complicates debugging making the model more difficult to control. As such, it becomes more and more a victim of eventual errors.

However, sequential conditioning was never born casually as an irrational approach. It comes from an adaptive conception of microsimulation structure. It starts from a base model developed on a given software-hardware technology. In the course of time, the structure of microsimulation expanded, new characteristics being added to it. Also, the burden of programming increased exponentially. As such, it is split in modules and each module is developed by a team (LaborSIM, DYNAMITE, EUROMOD, XEcon, etc.). Every team adds a little part to the model and uses the entire model exploiting the economies of scale that are essential in such an extensive work. This is also one of the reasons for which microsimulation models are essentially few and altogether very similar. However, the basic structure is still thought to be valid even though it is technologically outdated. In the light of new developments in information, it is possible to think of a different model structure and theoretically more efficient and easier way to implement it.

New Integrated Approach

The idea of an integrated approach is to create a unique transition matrix with all possible states of the world that a single unit could assume in input and in output. Practically, all combinations of control variables are determined. The "*n*" possible states of the world are obtained. They are equal to the product of the characteristic number for the modality number that each characteristic could assume.

The size of the transition matrix will be a $n \times n$. Coming back to the our example, therefore, we will have 4 states of the world for mixed variable AB. They are identified with the codes deriving from the union of A value and B value, namely 11, 10, 01 and 00. The resulting transition matrix will be the following:

	AB = 11	AB = 10	AB = 01	AB = 00
AB = 11	p(AB=11 | AB=11)	p(AB=10 | AB=11)	p(AB=01 | AB=11)	p(AB=00 | AB=11)
AB = 10	p(AB=11 | AB=10)	p(AB=10 | AB=10)	p(AB=01 | AB=10)	p(AB=00 | AB=10)
AB = 01	p(AB=11 | AB=01)	p(AB=10 | AB=01)	p(AB=01 | AB=01)	p(AB=00 | AB=01)
AB = 00	p(AB=11 | AB=00)	p(AB=10 | AB=00)	p(AB=01 | AB=00)	p(AB=00 | AB=00)

From a theoretical point of view, the integrated approach does not put any independence hypothesis and the identification of all states of the world is the only necessary and

sufficient condition. Also, the eventual extension of individual characteristics does not complicate the model. If a new event C with "*k*" modalities has to be inserted to extend the model, it is sufficient to determine the $n \times k$ new states of the world that derives from the previous states combination with "*k*" new characteristic modalities and to widen the transition matrix proportionally. The states of the world will be identified with the new variable ABC and the logical operation will be absolutely identical to the previous one made in the case of AB. In an extreme situation, time variable could also be considered as a characteristic made by as many modalities as the years. Therefore, in this approach, it is not theoretically necessary to further implement work to extend the model to new characteristics or to new modalities for the current variable. However, from an empirical point of view, it is more complex. Increasing the states of the world, some of them could be "impossible to exit". In fact, the transition matrix could show many zeroes. This creates many difficulties in the simulation. In the next paragraph, the problems will be analysed in detail and will be solved with two complementary solutions. However, at this moment I will not focus on this matter.

The other side of the coin in the integrated approach is a greater effort required for elaboration. In fact, it is heavier to manage an $n \times k$ squared matrix than to manage "*k*" squared matrices of "*n*" sizes. It affects the simulation negatively because the elaboration becomes longer. However, if the integrated transition matrix does not become too big, the penalization in general performance of the simulation is unimportant and if the programming phase is added, the model can be implemented in shorter total time with lower probabilities of errors. In conclusion, this new approach is more simple, more scalable, and most of all, it has the advantages of directing the power of the elaborators towards reducing economic hypotheses and towards making the construction of the model easier.

Case Study

A model for pension analysis.

A model to analyze pension reforms in Italy [Marchionne 2003] was built to apply an integrated approach to microsimulation. This is my case study. In fact, in every pension system, the main interest variable is individual income that responds well to this kind of analysis [Marchionne 2004]. In every period, the amount of paid contributions and the level of given benefits depend on gross individual income directly (defined benefits) or indirectly (defined contributions). Thus, an individual as reference unit is the natural consequential choice for the purpose of our study. Therefore, the aim of the model will be to simulate individual income level. Now, I will not go into detail in the subject of social security because I plan to focus only on methodology.

In the economic theory, individual income depends on some personal characteristics, in particular, on sex, geographical area, education level, job qualification, work sector and experience [Mirrless 1971]. The Italian pension system structure merges job qualification and work sector in one only distinctive element [INPS 1994, 2004]. For this reason, the two latter variables have been encoded again in order to obtain a better match between data and the necessary input for eventually calculating pension.

The structure of microsimulation is modular. From a logic-operational point of view, it is subdivided into 3 sections: Dataset, Demography, Income. I used data as the starting point for simulation and also for obtaining the estimated equation of individual income. Therefore, different from other studies, population was generated starting from a real situation.

I would like to say a few words on data before passing on to every single module. As data have a particularly important role in the construction of a microsimulation model, they need to have well-defined characteristics and must set up a panel of "*i*" individuals observed for "*t*" years. This kind of data is usually obtained through surveys and is generally characterized by a wide "*i*" and a small "*t*". All this can cause some consistency problems on estimated coefficients. However, there are other difficulties. In fact, microdata are typically affected by distortions. In particular, they suffer of errors in measurement and selection problems (self-selectivity and non-response). They are also subject to a significant phenomenon of attrition, that is, the reduction in the unit number due to moving out of some individuals from the sample for whatever reasons [Wooldridge, 2002].

Presently, there are 3 sources available that are potentially appropriate for our analysis: i) INPS archives, ii) ISTAT's survey on the Italian work force and iii) Bank of Italy's Survey on Italian Householder Income and Wealth (SHIW). The archive of INPS is the largest and more complete than others, but their data are not good for scientific use because they do not cover all individuals and do not have information on individual education which is an essential variable for income analysis. The survey on work force by ISTAT is representative, wide and detailed, but does not have information on income, so it is unusable. The survey on householder income and wealth made by the Bank of Italy is the only one to show not only information on an individual's characteristics and personal income but also maintains a good representation of the population. However, SHIW data present a lot of problems. First of all, they are a pseudo-panel because the panel rotates the units to maintain the representation of the population. The panel was brought back to a balanced form to facilitate the

implementation of data in the model. The second problem is that income is measured net of contributions and taxes. As I mentioned before, in each pension system the main interest variable is gross personal income. It means that SHIW data have to be elaborated in order to pass the income from net to gross. Finally, bi-yearly surveys (on average) cause remarkable difficulties in the building of transition matrices because some units could "jump" more than once but leave only one trace[4] in the data. Even if in this analysis I avoided, in every way (which was sometimes very hard to do), to take from different sources because of problems of incoherence on data, SHIW data do not have demographic information on fertility and death rates. So, they have necessarily been taken from ISTAT's yearbook.

Dataset Module

The Dataset module converts raw data in input for the model. Raw data are the SHIW data of the Bank of Italy [2002] and ISTAT's[5] [2004] mortality and fertility charts. The module is divided into 5 sub-sections: Panel, Net-to-Gross, Transition, Survivorship and Fertility. The Panel section deals with the creation of a balanced panel of individuals. The years included are 1993, 1995, 1998, 2000, and 2002. The average number of observations is 22864 yearly. Panel building reduces to 2703 units observed per year. It is obtained with a complex procedure. At first, the relevant variables in different archives are combined in a unique cross-section for each year, then income recipients, dependent family members (partner, children, rather familiar member) and other situations of fiscal relevance (handicapped or under 36 months year old children) are identified. A unique identification code is built starting from time-invariant individual characteristics for the last year. Then, this code is applied to the former period coming back to latter one through 2 by 2 connection. So, the archive is purified from all superfluous observations, unnecessary variables, and units that were incoherently connected (changes of sex, instruction levels decreasing in time, impossible job transfers, anomalous income variability, etc.).

The Net-to-Gross section determines gross individual income starting from net ones and considering the different fiscal brackets and all individual/family detractions attributable to the taxed subject. The procedure does not check for financial income and limits itself to earnings from work and transfers. Even if it could seem simple from a theoretical point of view, the inversion of tax function with detractions is not univocal so a numeric procedure is necessary. I use a goalseeking routine that minimizes the loss function given by the relative difference between real and estimated net income. Practically, a control variable (gross income) is modified to reach the optimal point of an objective function (minimum of a relative loss function) through numeric optimization. The procedure keeps under consideration the fiscal laws and tax rates in force in each year of the survey [Herr 2001, Targetti Lenti 2004, Frizzeri et al. 2004]. Net-to-Gross is a very expensive routine in terms of resources and time. However, the outcomes are excellent: at the end, an error over only ±3% is verified in 130 situations out of 13515 panel observations.

The third section deals with the creation of transition matrices. For every individual a 6 digit code concatenating values of interest SHIW variables (*area3 sesso studio settp7 qualp7n nonoc*) is built. The routine SGEE transforms the 6 digit code into a 4 digit

[4] In reality this phenomenon is always present but widens more when the period between two near waves increases.

[5] Both archives are downloadable from web sites of the Bank of Italy and ISTAT.

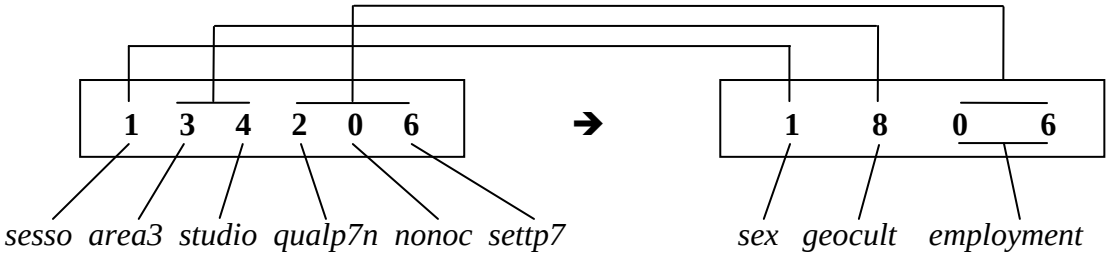
one. The first digit shows sex (*sex*), the second is a synthesis of geographical area and education (*geocult*), and the last two digits identify the kind of work (*employment*)[6] a subject does. The education level is reduced from 6 to 3 categories and job qualification and work sectors[7] have been arranged in 12 modalities according to pension setting[8] and the main activities of non working subjects. The following table shows the classification of SHIW and SGEE codes:

Table – SHIW and SGEE codes

Variable	SHIW	Variable	SGEE
Sesso	1 = Male 2 = Female	Sex	1 = Male 2 = Female
Area3	1 = North 2 = Center 3 = South	Geocult	1 = North – Low Education 2 = North – Average Education 3 = North – High Education 4 = Center – Low Education 5 = Center – Average Education 6 = Center – High Education 7 = South – Low Education 8 = South – Average Education 9 = South – High Education
Studio	1 = No 2 = Primary School 3 = Secondary School 4 = High School 5 = Degree 6 = Master post-lauream		
Qualp7n	1 = Blue Collar 2 = White Collar – Teacher 3 = Executive 4 = Manager 5 = Freelancer 6 = Self-Employed 7 = Unemployed	Employment	1 = Dep. Employed – Public Workers, Teachers 2 = Dep. Employed – Private Workers, Teachers 3 = Dep. Employed – Public Manager 4 = Dep. Employed – Private Manager 5 = Freelancer 6 = Farmer – Settler – Cropper 7 = Craftsman 8 = Retailer – Wholesaler 9 = Unemployed 10 = Student 11 = Retired 12 = Other
Settp7	1 = Agricolture 2 = Industry – Buildings 3 = Commerce – Tourism 4 = Transport – Comunication 5 = Fin. Interm. – Insurance 6 = Pub. Amm. – Services 7 = Unemployed		
Nonoc	0 = Employed 1 = Searching first job 2 = Housewife 3 = Well-Off 4 = Retireted 5 = Unemplyed 6 = Student 7 = Other		

[6] SGEE is Sex Geocult Employment. Letter E is doubled because of 2-digits for Employment.
[7] Sector 4 and 5 are unified to sectors 2 and 3 respectively (see table).
[8] Pension set-up is a combination of job qualification and sector so it is an unusual classification for these variables from an economic point of view. See the Law regarding this topic for details.

In SGEE, “Unemployed” merges into individuals “looking for first job” whereas “Other” includes “Housewife”, “Well-Off” and “Other” of SHIW *code*.
An example of SGEE routine working is shown in the following figure:



The SGEE routine has an essential role in the functioning of the model because it is the mechanism that allows the adoption of the innovative approach previously described. The compression of the different individual characteristics in a single variable, in fact, allows summarization of all states of the world in only one variable and simulation of the events all together according to an integrated approach (avoiding the sequential one). The price paid for this gain in efficacy is the loss in efficiency because of the increase in simulation time. After the SGEE codes are assigned to the units, $K_{status} \times K_{status}$ transition matrices (216×216) are built. For each row, they show the probability of passing from a given SGEE code in time “*t*” to any SGEE code at time “*t+1*”. The values are cumulative, so it means that the totals of the row are all equal to 1. As the transition probabilities are a function of age, 12 different matrices are created dividing the working age (15-75 years old) in half-decade. The hypothesis that nobody works after 75 years old is assumed. A further hypothesis concerns individuals under 16 years of age. If they are under 5 years old, then they are “Other” but if they are between 6 and 15 years old, then they are “Student” in abidance to mandatory education laws. Even if it would be conceptually more correct, to go into detail through the yearly period, it was impossible for lack of available data. For the same reason, the other two hypotheses are made.
The fourth and fifth sections deal with the determination of model demographic input using ISTAT data (until now SHIW data had only been used). In the fourth section Survivorship Tables are generated starting from Tables of Death. They are meant to determine the probability to survive the following year for each age and sex [Livi Bacci 1990]. The fifth section determines total fertility index given in the Fertility Tables. It is the average number of children per woman [Terra Abrami 1998, ISTAT 2003]. This index is specified by age and is not a summarized total value in order to take into account the population structure during the simulation and to have a good tool for demographic hypotheses.

Demography Module

The Demography module deals with the determination, for each period (year), the new entries in population (births), the state of every unit (SGEE) and dead individuals (deaths). The module is divided into two sections. The former is strictly “accounting” aimed to account for births and deaths. The result is a huge $N^+ \times T^+$ matrix where N^+ indicates the total number of simulated individuals (total units) and T^+ is the overall time that goes from the birth year of the oldest individual present in the panel from 1993-2002 to the last year simulated (total periods). Every row represents the whole life

of a single unit. In every period, the unit can be active or "sleeping". It is active when the individual is alive and sleeping otherwise (individual was not yet born or is already dead). An active unit always has a positive value. It is equal to 1 for male, 2 for female.
The second section assigns a valid SGEE code to each active unit. Given the geographical area to which it belongs[9], the minimum education level and the hypothesized employment ("Other" or "Student") is assigned to every unit younger than 16 years. Then, for each unit a random number "*x*" between 0 and 1 is extracted (Montecarlo Method). Given the SGEE code in time "*t*", it allows to set up the SGEE code in time "*t+1*" comparing the extracted value with the cumulative transition probability conditioned by age.
The operating mechanism is not complex. For reasons of simplification, it is assumed that there are only 3 modalities of the variable for which evolution is simulated (SGEE) and that the transition matrices in age 1 and age 2 are the following:

Modality for Age 1	A	B	C
A	0.4	0.7	1
B	0.3	0.6	1
C	0.2	0.9	1

Modality for Age 2	A	B	C
A	0.4	0.8	1
B	0.4	0.7	1
C	0.3	0.8	1

At time "*t*", age 1 and modality B for the simulated individual is hypothesized. A random number "*x*" from a uniform distribution between 0 and 1 is extracted. Its value is 0.68. The simulated individual will pass to modality C in time "*t+1*", "*x*" being higher than B value (0.6) and lower than C value (1) in row B of the matrix for age 1. Similarly, another individual with modality B, an extracted value "*x*" always equal to 0.68, but in age 2 is considered. In this case, at time "*t+1*", the individual will still remain in modality B because 0.68 is higher than 0.4 (A) but lower than 0.7 (B).
Even though the ambiguity of outcomes is not a desirable property, it is especially the presence of structural all-zero rows in the transition matrix to cause greater problems to the model. A structural zero is a conditional probability equal to 0 in the non-cumulative transition matrix. It means that it is impossible to achieve a given state in the output starting from a given state in the input. Generally, all these matrices have structural zeroes because some transitions can never come about[10]. When data used for conditional probability determination are too scarce or the observations are too concentrated in few modalities, structural zeroes can fill a whole row. This does not happen for natural causes, but rather for sample reasons. This phenomenon increases when the waves are spaced out in time (as SHIW data) because some individuals could change their situation more than once without leaving any trace in the data. When this happens, looking at cumulative probabilities, a status with positive probability in the input and zero in the output is created[11]. If an individual enters this state, the model generates an error because it is not able to assign a code to the simulated unit in the next period.
Two different ways were followed to solve this problem. The first way simulates the last "*e*" periods again with an increasing "*e*" going back to as far as it enters a different

[9] Average values of long period extracted from 1977-2002 SHIW sample are used as geographical distribution percentiles: North 40,34%– Center 21,15%– South 38,51%.
[10] For example, a transition from being a pensioner to becoming a student .
[11] The last modality counts 0 instead of 1. Given the cumulative nature of the matrix, this means that all the modalities have zero probabilities.

evolutionary path that allows it to go around the error[12]. From a theoretical point of view, this procedure is equivalent to redistributing conditional probabilities towards a structural all-zero row among the other probabilities proportional to their values. The transition matrices of the previous example are modified for age 2 in the following way:

Modality for Age 1	A	B	C
A	0.4	0.7	1.0
B	0.3	0.6	1.0
C	0.2	0.9	1.0

Modality for Age 2	A	B	C
A	0.4	0.8	1.0
B	0.4	0.7	1.0
C	0.0	0.0	0.0

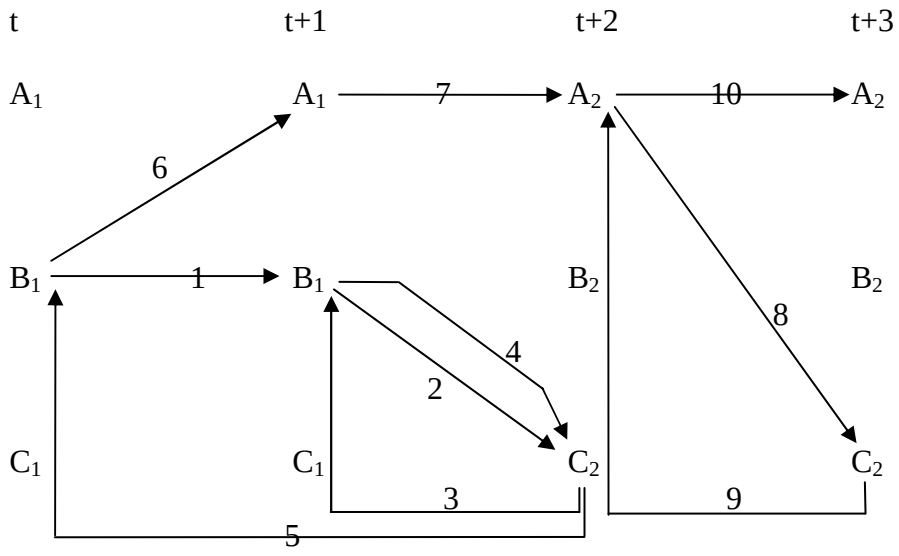
Now, the last row of the second matrix is a structural all-zero row. It is hypothesized that the simulated individual has age 1 and modality B at time "*t*". Random extraction generates an $x = 0.51$ so there is no transition (*step 1*). It is assumed that in time "*t+1*" the individual does not achieve age 2. His transition matrix, therefore, is the same as that of age 1. An $x = 0.87$ is extracted and there is a shift from modality B to C (*step 2*). At time "*t+2*" the individual achieves age 2 and changes the transition matrix. However, his present modality, which is C, has a structural all-zero row. The random number extraction does not assign a modality to exit from the transition matrix. So, an error comes up. At this point, the correction mechanism starts. It scales down by $e=1$ in time and repeats the simulation in period "*t+1*" starting from modality B and using the transition matrix of age 1 again (*step 3*). An $x = 0.97$ extraction leads to the same identical error (*step 4*). This time, the period goes back to *t* by $e=2$ (*step 5*). An $x = 0.27$ is extracted with the move towards modality A (*step 6*). In "*t+1*", an extraction $x = 0.13$ does not mark any change with the exception of the transition matrix (*step 7*). An $x = 0.83$ leads towards a new entry in C when time is "*t+2*" (*step 8*). An error is generated and the correction mechanism is brought back to action. However, this error is different from the previous one. So, the counter of error "*e*" is reinitialized which starts from $e=1$ again (*step* 9). In *t+2*, an $x = 0.39$ is extracted. The individual can close his simulation in modality A at the time "*t*+3" (*step* 10). This regressive error correction mechanism is expensive in terms of time and resources used, but despite that it behaves neatly.
The following table summarizes the sequence of random extractions (Montecarlo) and of the regressive error correction mechanism starting (RECM):

Montecarlo Sequence (Start in B_1)

01: $x = 0.51$ B_1	06 : $x = 0.27$ A_1
02: $x = 0.87$ C_2	07 : $x = 0.13$ A_2
03: RECM e=1	08 : $x = 0.83$ C_2
04: $x = 0.97$ C_2	09 : RECM e=1
05: RECM e=2	10 : x = 0.39 A_2

and the following diagram summarizes the *iter* described. The indexes of modalities A, B, and C indicate the ages of the individual. It reflect the transition matrix to use for evolution.

[12] Practically, when a transition error is due to a structural all-zero row, the time counter *t* is scaled by *e=1* in order to re-simulate the last period. If the same error comes up again, *t* is reduced by *e*=2. At every repetition, *e* increases by 1 and therefore last *e* periods are simulated again.



The second solution adopted in order to avoid structural all-zero rows is the partially transition matrix closure. A transition matrix is "closed" when every modality in the input is able to generate a modality in the output. This happens if the last column of the matrix is full of 1 only. To understand the logic under this original solution, a hypothetical rule to face structural all-zero rows is imagined. It assigns the same modality of the previous period to the next one. In probabilistic terms, it means to put 1 in the main diagonal of the transition matrix when it crosses a structural all-zero row. So, it is certain that the modalities in the input and output are the same. Now, closing the matrix with a positive value lower than 1 is imagined. What does it mean? What is the effect? Below a total closure (left) and a partial closure (right) applied on the transition matrices used in previous examples is shown.

Modality for Age 2	A	B	C
A	0.4	0.8	1.0
B	0.4	0.7	1.0
C	0.0	0.0	1.0

Modality for Age 2	A	B	C
A	0.4	0.8	1.0
B	0.4	0.7	1.0
C	0.0	0.0	0.5

Even if the event is not certain, the meaning of closure does not change. Facing a structural all-zero row, the individual has some probabilities to maintain the same condition of the previous period. However, in the case of partial closure, random extraction can lead to the refusal of the hypothesis leading to an error similar to the initial situation. Therefore, partial closure is less incisive than total closure, but combined to the control mechanism described previously, it allows to make more acceptable hypotheses on transition matrices. For these reasons both methods were implemented.

Income Module

The Income module takes care of estimating individual income based on a set of control variables in order to obtain an equation to be used for forecasting. It uses the SHIW data re-elaborated by the Dataset module. An individual's characteristics, therefore, are recoded according to the SGEE code. At the same time, incomes are transformed from net to gross, carried back to lire of the year 2000 (real values) and expressed in

logarithmic form. The usual way to represent a panel is a double entry table in which spatial and temporal sizes are spread sequentially one after the other. Choosing only the first two units, SHIW takes up the following shape:

id	year	birth	area3	sesso	studio	settp7	qualp7n	nonoc	YL
1	1993	1945	2	2	4	6	2	0	37000
1	1995	1945	2	2	4	6	2	0	37400
1	1998	1945	2	2	4	6	2	0	37800
1	2000	1945	2	2	4	6	2	0	38200
1	2002	1945	2	2	4	0	0	4	27800
2	1993	1975	2	1	4	0	0	6	0
2	1995	1975	2	1	4	0	0	6	0
2	1998	1975	2	1	4	0	0	6	0
2	2000	1975	2	1	4	0	0	6	0
2	2002	1975	1	1	5	4	2	0	22000

Id = identifier; Year = time variable; birth = birth year of individual; Area3 = geographic area (north – center – south); Sesso = sex; Studio = Education; Settp7 = work sector; Qualp7n = job qualification; Nonoc = non working activities; YL = gross income (real value in thousand of lire 2000)

In the example, the first individual (id=1) is a woman (sex=2) with high school diploma (studio=4) and who worked for Public Administration (settp7=6 and qualp7n=2) and lived in the Center of Italy (area3=2). She retired (nonoc=4) when she was 57 years old (age = 2002 – 1945). The second unit is a student (nonoc=6), a boy (sex=1) from the same geographical area (area3=2). He graduated (studio=5) and then went to the North (area3=1) to work (qualp7n=2).
The SHIW panel re-elaborated in Data module, instead, has the following form:

id	year	birth	sgee	YL
1	1993	1945	2501	37000
1	1995	1945	2501	37400
1	1998	1945	2501	37800
1	2000	1945	2501	38200
1	2002	1945	2511	27800
2	1993	1975	1510	0
2	1995	1975	1510	0
2	1998	1975	1510	0
2	2000	1975	1510	0
2	2002	1975	1602	22000
…	…	…	…	…

Id = identifier; Year = time variable; birth = birth year of individual; sgee = sex-geocult-employment code; YL = gross income (real value in thousand of lire 2000)

The data compression is well-evident. The importance of SGEE routine is that it really allows a two-dimensional representation of data much more manageable than the sequential one through the synthesis of control variables. The whole panel can be shown using 2 relatively very small tables. In both x-axis and y-axis there are several units and periods.

SGEE		t				
id	birth	1993	1995	1998	2000	2002
1	1945	2501	2501	2501	2501	2511
2	1975	1510	1510	1510	1510	1602
…	…	…	…	…	…	…

YL		t				
id	birth	1993	1995	1998	2000	2002
1	1945	37000	37400	37800	38200	27800
2	1975	0	0	0	0	22000
…	…	…	…	…	…	…

Id = identifier; birth = birth year of individual; SGEE = sex-geocult-employment code; YL = gross income (real value in thousand of lire 2000); t = time variable

To simulate income, therefore, it is sufficient to have an equation that determines the income in relation to the variables as summarized in the SGEE code. In fact, after the code is expanded and inserted in this equation, the calculation of the next individual income becomes a matter of simple computation. The problem is to identify the equation. In literature, the variables typically used as regressors in econometric models to estimate incomes are sex, geographical area, educational qualification, occupational place, affiliation sector and experience. As it is difficult to find a good measure for this last variable, generally, job seniority or, more roughly (but always good), age is used as proxy for experience. However, in reality, there are other 2 variables that should be inserted fully among regressors, but whose measuring is very complicated. The first is individual skill. Each individual, in fact, has an ability regardless of education, environment and experience he or she has gone through. This is called innate ability. The second variable is the variability as a result of the cohort effect that arises when, ceteris paribus, being born in different periods determines different incomes among individuals.
However, in its complete form, the estimate model is as follows:

$$y_{i,t} = \alpha + \beta \times x_{i,t} + v_i + z_t + \varepsilon_{i,t}$$

where α is a fixed constant for all "*i*" and "*t*", β is the vector of parameters, $x_{i,t}$ are the control variables, v_i is an individual effect with $E(v_i)=0$ and $var(v_i)=\sigma^2_v$, z_t is a time-effect with $E(z_t)=0$ and $var(z_t)=\sigma^2_z$ and $\varepsilon_{i,t}$ is an classic idiosyncratic normal error with $\varepsilon_{i,t} \rightarrow N(0;\sigma^2_\varepsilon)$. If variables able to capture v_i e z_t are found, the model could be reduced. With reference to time-effect, the level of GDP is an acceptable proxy of z_t even if it is very limited. With reference to innate abilities v_i, it is impossible to identify such an appropriate variable. Moreover, being unobservable, its effect is merged with all residual effects raising a problem of collinearity between v_i and the constant α. It was overcome substituting them with the generic individual effect $\alpha_i = \alpha + v_i$.

Analytically, the final model is the following:

$$y_{i,t} = \alpha_i + \beta \times x_{i,t} + \varepsilon_{i,t}$$

where the only remaining specific effect is the individual effect and GDP is among the regressors. Depending on the nature of α_i, the model is a fixed effects model (FE) or a random effects model (RE). It is possible to detect the model that is more accurate with some tests [Baltagi 1995]. In fact, the presence of RE can be detected through Breusch and Pagan Lagrangian Multiplier Test (BPLM) that verifies the null hypothesis H_0: $\sigma^2_\alpha = 0$. Rejecting the test indicates that individual effects are random and not fixed because of no-zero variance. Alternatively, a comparison FE against RE can be attained through the Hausman Test. It is based on the difference between an unbiased and consistent estimator subjected to the null hypothesis (H_0) and to the alternative one (H_a) as is FE, and an unbiased and efficient estimator subjected to H_0 but biased subjected to H_a as is RE. Passing the test indicates the presence of a random effect.
Even though the latter estimators work well in many models of panel data, FE and RE are static models and are not able to capture any dynamic evolution. In particular, any lags of dependent variables capable of stressing serial correlation phenomena are not between the regressors. Individual income shows a persistent trend. So, lags of $y_{i,t-s}$ should be included in the regressors or else notable problems of accuracy and consistency come up.
The estimate problem of dynamic model such as[13]:

$$y_{i,t} = \varphi \times y_{i,t-1} + \alpha_i + \beta \times x_{i,t} + \varepsilon_{i,t}$$

subjected to strictly exogeneity hypothesis of x[14], has been faced in econometrics transforming the equation in first differences before in order to eliminate individual effects, and then using a GMM model to estimate the lagged differences of dependent variable $\Delta\, y_{i,t-1}$ through appropriate instrument variables. Arellano and Bond have suggested the lagged level of regressors including the lag dependent variable (1991) or, alternatively, the set of levels and differences of the same variables (1995) as instruments [Arellano 2003]. The number of instruments increases very quickly leading to an illusory improvement in estimates due to the quantity of regressors more than their quality [Baltagi 1995]. Using the Arellano-Bond estimator, a dynamic model has to show autocorrelation of the first order that justifies the move to differences, but at the same time must not show autocorrelation of the second order because otherwise the quasi-differentiation improvement declines [Wooldridge 2002]. These assumptions can be verified through specific tests.
However, there are also intermediate models between static and dynamic ones. Here is a specification structure:

$$y_{i,t} = \alpha_i + \beta \times x_{i,t} + \varepsilon_{i,t}$$

in which an error follows an autoregressive process of the first order:

$$\varepsilon_{i,t} = \rho \times \varepsilon_{i,t-1} + z_{i,t} \qquad \text{with } |\rho| < 1 \text{ and } z_{i,t} \sim i.i.d.(0\,;\,\sigma^2_z)$$

[13] The number of lags in dependent variable can be greater than 1.
[14] $E(\varepsilon_{i,t} \mid x_{i,1}, \ldots, x_{i,T}, \alpha_i) = 0$ with $t = 1, \ldots, T$

In 1978, Lillard and Willis applied a transformation that was able to convert the AR(1) errors in a classic serially non-correlated error to estimate the model above. The typical element $y^{*}_{i,t}$ of this transformation is similar to the familiar $y_{i,t} - y_{i,.}$ of FE and can be obtained subtracting a pseudo-average from a modified $y_{i,t}$ value [Baltagi, 1995].

Estimates of individual income

lnY	FE	RE	FE+AR1	RE+AR1	AB95
lnY[-1]	-	-	-	-	0,227599
					(6,49)**
Age	-0.0127005	0.0160723	-0.0254883	0.0242296	-0,04115
	(-0.98)	(5.78)***	(1.57)	(8.75)***	(6,38)***
Age^2	0.000348	-0.0000827	-0.0007308	-0.0001619	0,000549
	(8.42)***	(2.97)***	(9.38)***	(5.81)***	(8,13)***
Sex=1	-	0.4023947	0.1389501	0.4010123	0,32934
		(19.51)***	(2.21)**	(21.35)***	(11,36)**
Geo=1	-	0.1860222	0.267261	0.1817556	0,123243
		(8.09)***	(1.51)	(8.69)***	(4,35)**
Geo=2	-	0.1716919	0.0394127	0.1684787	0,089819
		(5.95)***	(1.29)	(6.43)***	(2,62)**
Cult=2	0.0973535	0.2749432	0.1150634	0.2903409	0,28381
	(2.64)***	(13.69)***	(1.23)	(15.08)***	(4,6)*
Cult=3	0.1298448	0.4457165	0.0460925	0.4543573	0,480343
	(1.32)	(12.84)***	(0.45)	(14.05)***	(4,95)*
Employ=1	0.0188143	0.0308236	0.2293707	0.0248601	-0,01838
	(0.88)	(1.67)*	(2.77)***	(1.35)	(0,35)*
Employ=3	0.1025256	0.1787553	0.1582549	0.1942732	0,211277
	(1.4)	(2.62)***	(1.42)	(2.83)***	(2,09)
Employ=4	0.0915888	0.1856926	0.1244862	0.2130514	0,368943
	(1.21)	(2.52)**	(1.89)*	(2.82)***	(2,06)
Employ=5	0.2954979	0.2415076	0.2547533	0.2073933	0,42406
	(4.95)***	(4.90)**	(4.49)***	(4.27)***	(4,43)*
Employ=6	0.0066192	-0.0596803	-0.0159765	-0.0528658	0,284515
	(0.09)	(0.95)	(0.4)	(0.84)	(1,38)
Employ=7	0.1978347	0.2149156	-3.0862751	0.2132616	0,487511
	(4.68)***	(6.03)***	(7.42)***	(5.99)***	(4,96)*
Employ=8	0.2400032	0.2116516	0	0.20887	0,528151
	(6.50)***	(7.04)***	(7.43)***	(6.96)***	(5,31)*
Employ=11	-0.1461913	-0.1945516		-0.2006354	0,011254
	(5.84)***	(8.73)***		(8.68)***	(0,19)***
Employ=12	-2.442705	-1.9020919		-1.8620027	-5,42622
	(13.86)***	(14.42)***		(14.32)***	(1,99)***
GDP per capite	3.21e-12	2.34e-12		2.28e-12	3.59e-12
	(4.68)***	(30.48)***		(27.14)***	(20,76)***
Constant	13.0505715	11.5131163	8.5973148	11.2689738	11,53229
	(101.56)***	(117.22)***	(50.97)***	(107.43)***	(26,48)***
Observations	7750	7750	5895	7750	5543
R-squared	0.18				
Number of idkey	1855	1855	1692	1855	1602

Robust t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

TEST			
Model	Type	Statistic	p-value
FE	All u_i=0:	F(1854,5881) = 5.85	Prob>F = 0.000
RE	BPLM	chi2(1) = 2400.00	Prob>chi2 = 0.000
RE	Hausman	chi2(13) = 353.25	Prob>chi2 = 0.000
FE + AR1	All u_i=0:	F(1691,4189) = 2.01	Prob>F = 0.000
AB95	Hansen	chi2(43) = 316.06	Prob>chi2 = 0.000
AB95	AB AR(1)	z = -8.44	Pr>z = 0.000
AB95	AB AR(2)	z = 0.69	Pr>z = 0.488

In our specific case, the estimates of all models overall yield quite significant coefficients and their signs agree with the expectations. Therefore, the choice is made on the basis of the results of the tests. Test F on individual errors indicates that FE model is better than the simple OLS. Instead, between FE and RE, BPLM and Hausman Test turns out to be divergent. If on the one hand the BPLM test moves towards the RE model because the hypothesis of no variability on individual effects is rejected, on the other hand, Hausman Test marks a better performance of FE. Moreover, the FE estimates show a non-null correlation of individual effects as regard error terms. This situation is in contradiction with the RE hypothesis. The incoherence of tests is attributable to a bad specification of regressors and in particular to the lack of a variable able to capture the dynamic structure of the model. If the model forecasted, for instance, lagged income between regressors, the pro FE Hausman Test result (RE is not the most efficient and consistent model subjected to H_1, so the test was badly built) and the pro RE Breusch-Pagan result (there is anyway a positive variance of individual effects) would be justified. The shift towards a dynamic structure through Arellano-Bond (1995)[15] leads to an improvement of the estimates, but also to greater complexity[16]. From an inferential point of view, Arellano-Bond Test confirms the presence of the first order serial correlation and rejects the second order correlation as expected. The coefficients are significant on average. The reduction in the number of instruments from 187[17] to 62 involved a decisive improvement in the estimates.

Finally, the estimates of FE and RE models with AR(1) error structure are considered. The two-step method leads to more consistent estimates of serial correlation coefficient. Coefficients of FE with AR(1) are not very significant while those of RE are clearly better. They are similar to previous FE and RE models. The coefficients come back towards signs (but not values) and significance similar to OLS model.

In the choice for forecasting models, computational considerations are added to econometric ones. Using the Arellano-Bond model allows to assign an income to each unit still in the work force, but there is the question of determining the gross income of

[15] *Number of groups* is not equal to 2703 sample individuals because only individuals with income are taken in account.

[16] Using instruments in levels and in differences, their number increases rapidly. It is important to keep their numbers down as much as possible removing all those that are unnecessary. All dummy variables that have low variability in time, similar to time-invariant variables (null first differences), have been included as instruments only in levels because most of them will have null differences. Other variables, instead, have been joined in order to create only one instrument for each variable and lag size, instead of one for each period, variable and lag size. Secondly, the dynamic structure of both models makes the exclusion of outsiders necessary because they lead to conspicuous distortions in estimates. Individuals with an average income under the first or above the last percentile are considered outsiders. It leads to the exclusion of 16 units.

[17] Estimates are not reported here.

the new entries and of all individuals that have moved from an inactive state to an active one ($y_{i,t-1} = 0$) for whatever reason. The problem can be solved assuming an initial distribution for individual income. Avoiding pure theoretical hypotheses, it means to estimate a second model with a static structure capable of forecasting the first job income after an inactive state. The best candidate is the RE model despite its limitations. However, managing 2 independent interactive forecasting structures creates complications. Abandoning the dynamic structure is not an attractive idea because time incoherent situations could be generated. In forecasting, it could happen that there could be individuals that have unnatural decreasing careers because their values are fixed without reference to past ones. A good compromising solution, that maintains dynamicity without excessive increases in complexity, is the RE model with AR(1) errors. When an individual enters the work force, the model simulates individual effect α_i and assumes a lagged error $\varepsilon_{i,t-1}$ equal to zero[18]. In the following periods, the AR(1) error structure is fully implemented. Thus, the RE with AR(1) solution is not optimal from an econometric point of view. However, given the instrumental and functional nature that it has in the whole work, it seems a right compromise between the opposite requirements of precision in estimates and the facility of implementation.
Finally, the Income module simulates the weeks passed in every state for each year. If the individual does not change his condition, it will hypothesised that he had the same job for all 52 weeks of the year. Instead, if he makes a transition, then a random number in [-26;+26] is extracted to establish in which week of previous [-26;0] or next [0;+26] year it will happen. To implement the model and not to excessively complicate the calculation, the income for these weeks (determined as the proportion out of the total income of the same year) moves from the previous to the following year in order to have an income from only one job for every year. The distortion introduced in this way (no indexation of at the most 26 weeks) is small and allows an easier implementation of income management in various categories. In conclusion, the forecasting model used is:

$\log y_{i,t} =$	$+ 11.26897$	(Costant)
	$+ 0.0242296 * eta$	(Age in years)
	$- 0.0001619 * eta^2$	(Age^2 in years)
	$+ 0.4010123 * sex1$	(*dummy* Male)
	$+ 0.0000000 * sex2$	(*dummy* Female)
	$+ 0.1817556 * geo1$	(*dummy* North Italy)
	$+ 0.1684787 * geo2$	(*dummy* Center Italy)
	$+ 0.0000000 * geo3$	(*dummy* South Italy)
	$+ 0.0000000 * cult1$	(*dummy* Low Education)
	$+ 0.2903409 * cult2$	(*dummy* Middle Education)
	$+ 0.4543573 * cult3$	(*dummy* High Education)
	$+ 0.0248601 * employ1$	(*dummy* Private Employeed)
	$+ 0.0000000 * employ2$	(*dummy* Public Employeed)
	$+ 0.1942732 * employ3$	(*dummy* Public Manager)
	$+ 0.2130514 * employ4$	(*dummy* Private Manager)
	$+ 0.2073933 * employ5$	(*dummy* Freelancer)
	$- 0.0528658 * employ6$	(*dummy* CD/CM[19])
	$+ 0.2132616 * employ7$	(*dummy* Craftman)
	$+ 0.2088700 * employ8$	(*dummy* Retailer)

[18] In theory, RE models can be estimated through AR(2) or mode, but every lag means abandoning an observation wave. Given that there are only 5 periods, AR(1) seems to be the most appropriate choice.
[19] CD/CM is INPS acronym for farmer, settler and cropper.

*– 0.2006354 * employ11*	(*dummy* Pensioner)
*– 1.8620030 * employ12*	(*dummy* Well-Off, Housewife and Others)
*+ 2.28e-12 * GDP*	(GDP pro capite in lire 2000)

in which reference individual base unit for all coefficients is SGEE=2902[20]. Let me recall that *employ9* (unemployeed) and *employ10* (student) do not receive income by definition. After the equation has been established, it is applied to the SGEE code of each individual for every simulated year in order to establish the future income.

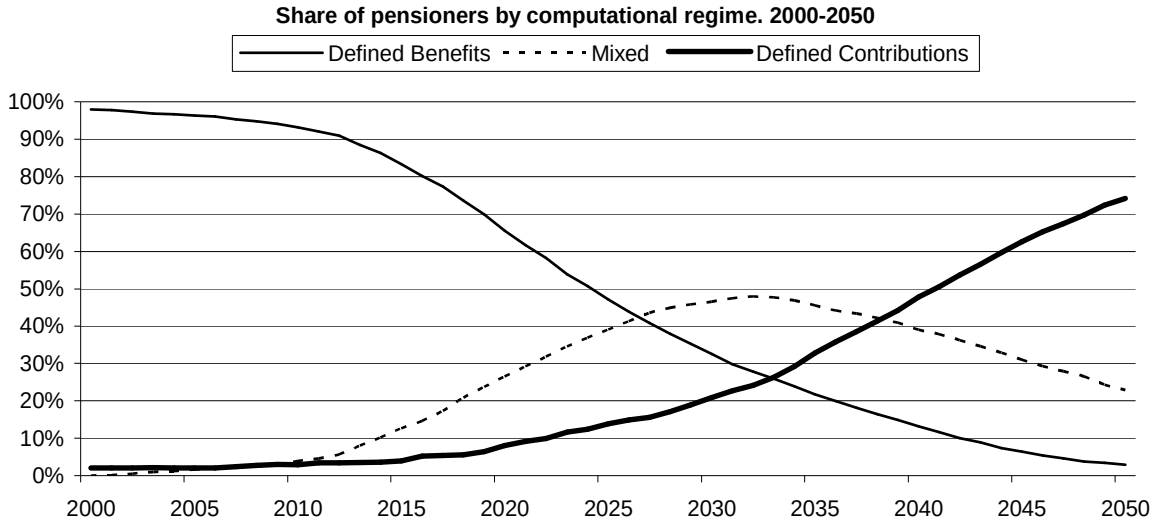
Outcomes

It is always difficult to evaluate the performance of microsimulation. All the more so it is difficult to judge its particular implementation arising from an integrated approach. It is possible to control if the model forecasts correctly the years for which data are available. With reference to SHIW 2004 data, I can measure the frequency with which forecasting errors happened for each characteristic considered. Age and sex are perfectly predictable for every unit in the 2002 data. However, the error was due to sample rotation. Broadly speaking, these mismatches can been considered as the benchmark to evaluate other variables.

Frequency	Age	Sex	Geo		Cult		Employ											
(in %)		1	1	2	2	3	1	2	3	4	5	6	7	8	9	10	11	12
Errors>5%	1,7	0,7	2,0	2,7	5,8	4,0	4,3	5,1	3,1	2,4	4,2	2,9	3,3	4,7	3,9	5,4	5,4	3,4
Errors>3%	2,3	1,3	2,9	4,1	6,6	5,2	5,1	6,6	3,9	2,8	4,9	4,2	3,8	5,5	4,8	6,4	6,1	4,6
Errors>1%	3,6	3,3	4,1	4,5	8,0	7,4	7,5	8,3	5,3	3,8	6,1	6,0	6,1	6,3	6,7	9,1	7,9	6,3

The results seem very good. However, they should be taken into account carefully because the evaluation period is of only 1 wave and the rotation mechanism introduces an error that is difficult to evaluate. So the benchmark could be not so much reliable. Moreover, in microsimulation models errors tend to propagate more than how much they offset themself [Bourguignon-Spadaro 2005]. So long term predictions become increasingly uncertain and are subject to distortions. Obviously, the normal uncertainty of every prediction is added. When comparing an integrated approach with a traditional one, it is necessary to refer to other models. Also in this direction, evaluation should be taken into account carefully for different reasons. Firstly, model specification is generally not so detailed enough to allow a real technical comparison. Only a global comparison is possible. Secondly, the models do not always account for the same variables and this makes their comparison practically impossible. However, the figure below shows the quota of pensioners according to calculus rules. This is one of the few variables that allows for matching with other models.
The results obtained are similar to the ones achieved by INPS, Ragioneria Generale dello Stato, Ministero del Tesoro and other works on the topic [Ferraresi 2000, OECD 1998, ecc.]. Generally, the performance of the model seems by and large good. Unfortunately the comparison is almost only demographic but it was the best available. As pointed out by the graph, the defined contribution system starts working fully only after a very long transition period in which there is a slow decrease in defined benefits till 2015 (starting from 87.4%) and a peak of mixed schemes, around 2033 (47.9%).

[20] The choice is made in order to get coefficients of *dummy sex*, *geo* and *cult* that are always positive and increasing (the lower category is the base), while the biggest group is the reference for *employ*.



These are the same results obtained by all the other models. More interesting would be to verify the distortion due to the partial conditioning of the sequential approach on employment composition, but it would be necessary to have detailed data of other models and their whole implementation structure, which is very difficult. The matching of pensioner quota in this model with other models is not a trivial result because it shows how, even by getting similar overall predictions, it is possible to build microsimulations with different distributions of employment. It generates models that are seemingly unbiased but in reality they hide significant distortions that could affect very hard the outcomes in more complex situations.

Summary of module contents

0. SOURCES:
    a. SHIW (historical) from 1993 to 2002 for Panel
    b. ISTAT 2003
    c. SHIW (historical) from 1998 to 2002 for Transition
1. DATASET Module:
    a. Panel (*attrition*)
    b. Net-to-Gross for IRPEF (*Goalseek*)
    c. Transition (SGEE compression)
    d. Survivorship (by sex)
    e. Fertility (fertility index by age)
2. DEMOGRAPHY Module:
    a. Births (Fertility)
    b. Deaths (Survivorship)
    c. Evolution (Transition)
        i. Regressive error correction mechanism
        ii. Partial transition matrix closure
3. INCOME Module:
    a. OLS FE ➔ Test F
    b. FE RE ➔ BPLM and Hausman
    c. AB_{91} ➔ AB AR(1) and AB AR(2)
    d. RE AR(1) ➔ *Compromise solution*

Conclusions

In this work, implementation problems of a microsimulation model are faced from a methodological and empirical point of view. The aim was to overcome the weaknesses of the traditional sequential approach on the one hand and to build a model on real data on the other. The sequential approach, in fact, is often used without appropriate considerations on theoretical aspects. This leads to implicit hypotheses that are usually more restricted than as much as one believes. A partial sequential conditioning, in fact, is equivalent to hypothesizing independency of events. This approach, moreover, becomes exponentially more expensive in programming and debugging when characteristics to simulate increase. So this limits expandability and scalability of the model. It is possible to simplify model building through an integrated approach. This innovative method is based on a compression mechanism that summarizes all the states of the world in only one variable. So a unique transition matrix is built to jointly determine all the characteristics of a unit. However, lack of data generates technical difficulties. In particular, structural all-zero rows can arise. They avoid to assign a state of the world to some units. I have overcome this difficulty through a regressive error correction mechanism and a partial transition matrix closure. The former reflects on the hypothesis of redistributing error probability on remaining events. The latter assumes to maintain a status quo. All this makes it possible to build a model for carrying out analyses on pension systems. It is similar to a sequential model (as far as possible to compare them) but has better performance in theoretical terms. Therefore, this work shows that there is more than one approach to microsimulation, and especially, that through more power elaborators it is possible to imagine a more efficient structure in order to implement these kinds of models.

Bibliography

Ando A. and Nicoletti Altimari S. (1999), *A Microsimulation Model of the Italian Household Sector*, Convegno Banca d'Italia – CIDE

Arellano M. (2003), *Panel Data Econometrics*, Oxford University Press, New York

Atkinson A., Bourguignon F., Chiappori P.A., (1988) "What Do We Learn About Tax Reforms from International Comparisons? France and Britain", *European Economic Review*, Vol 32.

Atkinson A., F. Bourguignon, C. O'Donoghue, H. Sutherland and F. Utili (2002), "Microsimulation of Social Policy in the European Union: Case Study of a European Minimum Pension", *Economica*, nº 69, pp. 229-243.

Banca d'Italia (2004), *I Bilanci delle famiglie italiane nell'anno 2002*, in *Supplementi al Bollettino Statistico – note metodologiche e informazioni statistiche*, Roma

Ballas D., Clarke G., Turton I [1999], Exploring Microsimulation Methodologies for the Estimation of Household Attributes, 4th International Comference on GeoComputation, Mary Washington Collage, Virginia, USA.

Baltagi Badi H. (1995), *Econometric Analysis of Panel Data*, Wiley, Chichester.

Borella M. [2004] - "The Distributional Impact of Pension System Reforms: an Application to the Italian Case", *Fiscal Studies* (2004) vol. 25, n. 4, pp. 415-437

Bourguignon F, Ferreira F., Leite P., (2003), "Conditional Cash Transfers, Schooling and Child Labour : Micro-Simulating Bolsa Escola". DELTA WP-07/2003

Bourguignon, F., Spadaro A. (2005), "*Microsimulation as a Tool for Evaluating Redistribution Policies*", Paris-Jourdan Sciences Economiques, WP-02/2005.

Caldwell, S.B. (1990) "Static, Dynamic and Mixed Microsimulation" Working Paper Department of Sociology, Cornell University.

Clarke M. (1986) *Demographic processes and household dynamics: a microsimulation approach*, in R.Woods and P.H.Rees (eds*,)* Population structures and models: developments in spatial demography, Allan & Unwin, London, 245-272

De Lathouwer L., (1996), "A Case Study of Unemployment Scheme for Belgium and The Netherlands", in *Microsimulation and Public Policy*, Harding editor, North Holland, Amsterdam.

Ferraresi P.M. and Fornero E. (2000), *Social Security transition in Italy: costs, distorsions and (some) possible corrections*, CeRP, Torino

Frizzeri, Gobbi and Postal (2004), *Testo Unico Imposte sui Redditi*, ilSole24ore

Herr U. (2001), *L'evoluzione della struttura dell'IRPEF: un'analisi attraverso le dichiarazioni*, Bari

Klevmarken N.A., (1997) "Modelling Behavioural Response in EUROMOD", *Microsimulation Unit WP* MU9702. Cambridge University

INPS (1994), *Atti Ufficiali*, Roma

INPS (2004), *Atti Ufficiali*, Roma

ISTAT (2004), *Annuario Statistico 2004*, ISTAT, Roma.

ISTAT (2003), *Previsioni della popolazione residente per sesso, età e regione*, Roma.

Livi Bacci M. (1990), *Introduzione alla demografia*, Loescher, Torino, pp. 350-380.

Marchionne F. (2003) - L'invecchiamento della popolazione mette in crisi i sistemi pensionistici: quali interventi si possono attuare?, Economia Marche, 2/2003

Marchionne F. (2004), *Una proposta di riforma per il sistema pensionistico italiano*, in Moneta e Credito, giugno

Mertz, J. (1991), *Microsimulation - A survey of principles developments andapplications*, International Journal of Forecasting, 7, pp.77-104

Miniaci R. (1998), *Microeconometric analysis of the retirement decision: Italy*, Working Paper AWP 1.7, OECD, Parigi.

Mirrlees J. A. (1971), *An Exploration in the Thoery of Optimum income Taxation* in *Review of Economic Studies*, 38(114), aprile

Monteduro M.T. e Di Nicola F. (2004), *Un modello di microsimulazione delle imposte sulle persone fisiche*, SECIT

Niccodemi A. (1998), *La microsimulazione dinamica nell'analisi delle politiche redistributive: un'applicazione al caso italiano* (Tesi di laurea), Università di Pisa

OECD (1998), *Microsimulation analysis of the retirement decision: Italy*, OECD Publications, Parigi.

Orcutt G, (1957), "A New Type of Socio-Economic System", *Review of Economic and Statistics*, n. 58. pp. 773-797.

Orcutt, G.H., Mertz J., Quinke, H (Eds.), (1986), *Microanalytic Simulation Models to Support Social and Financial Policy*, North-Holland, Amsterdam

RGS – Ragioneria Generale dello Stato (2004), *Tendenze di medio-lungo periodo del sistema pensionistico e sanitario*, Rapporto n. 5, maggio.

Schelling, T. (1971) *Dynamic Models of Segregation*, Journal of Mathematical Sociology, vol. 1, n°1,. pp.143-186

Targetti Lenti R. (2004), *Distribuzione del reddito primario e secondario delle famiglie nella Sam relativa al 1998 per le quattro macroregioni italiane*, pg. 55-58, SIEP, Pavia

Terna P. [2002], Economic Simulations in Swarm: Agent-Based Modelling and Object Oriented Programming, in The Electronic Journal of Evolutionary Modeling and Economic Dynamics, n° 1013, http://www.e-jemed.org/1013/index.php

Terra Abrami V. (1998), *Le previsioni demografiche*, Il Mulino, Bologna

Vagliasindi P. A. (2004), *Effetti redistributivi dell'intervento pubblico. Esperimenti di microsimulazione per l'Italia*, Giappichelli, Torino

Wooldridge J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Massachusetts Institute of Technology, Massachusetts.