Reconstructing an Artificial Society on the basis of Big Open Data

Aleksandra L. Mashkova[1,2][0000−0003−1701−5324]

[1] Orel State University, Orel, Russian Federation
[2] Central Economics and Mathematics Institute Russian Academy of Sciences, Moscow, Russian Federation
aleks.savina@gmail.com

Abstract. In this paper we integrate big data into the computer model of an artificial society. The model is agent-based and consists of several modules, representing demographic, economic, financial processes, employment and consumption, educational and administrative institutions. In order to create an artificial society that would simulate Russian Federation in 2014 year we use big open data, including Federal State Statistics Service yearbooks and official information on the websites of the ministries. Algorithm of an artificial society reconstruction includes creation of agents and organizations; distributing them among geographical regions; setting interrelations between agents, households and organizations. For the verification of the artificial society model we propose a DES-analysis method, which compares demographic, economic and social indicators of the simulation output data with the real values of these indicators in the base year. We present statistical analysis of modeling results variation in retrospective period.

Keywords: Artificial society · Agent-based model · Computational experiment · Verification · Big data · Statistical analysis.

1 Introduction

In the practice of socio-economic systems' management, there is a need to create and implement new methods and tools for making prognosis and planning. Sustainable economic and technological growth requires development of infrastructure, production capacities, human resources and living standards of the population. To take into account multiple factors, it is necessary to analyze big data, including both open statistical information, results of sociological surveys, monitoring of federal programs, and private data from ministries, departments, social networks and search systems. In order to use these data we need special methods and tools. Particularly, we integrate big data analysis methods into computer models of socio-economic processes.

We have chosen agent-based modeling as a main method in this study, since it allows to reflect dynamics of a macro-system as a result of the interaction of microlevel objects. The concept of agent-based modeling was proposed in the

1990s [8] and since then has been widely applied in analysis of economic, financial, social and environmental processes [5, 7, 11, 16, 19, 12]. Complexity of agent-based models has risen along with advances in computing power and information resources, resulting in larger models with complex interactions, whose inputs require sophisticated analytical approaches. Similarly, the increasing use of agent-based models data has further enhanced the complexity of their outputs [10].

The aim of our research is to construct an agent-based computer model of an artificial society, which reflects age and sex structure and regional resettlement of population, composition of households, economic structures, administrative and educational institutions. For the information content of the model, federal statistical yearbooks and official information on the websites of the ministries of the Russian Federation are used [1, 3, 2].

2 Structure of the Model of an Artificial Society

The developed model includes 7 interconnected modules reflecting various aspects of an artificial society: Demographics, Education, Employment, Production, Consumption, Financial System and Public Administration (see Fig. 1). Each module corresponds to a set of information objects and events that change their state.

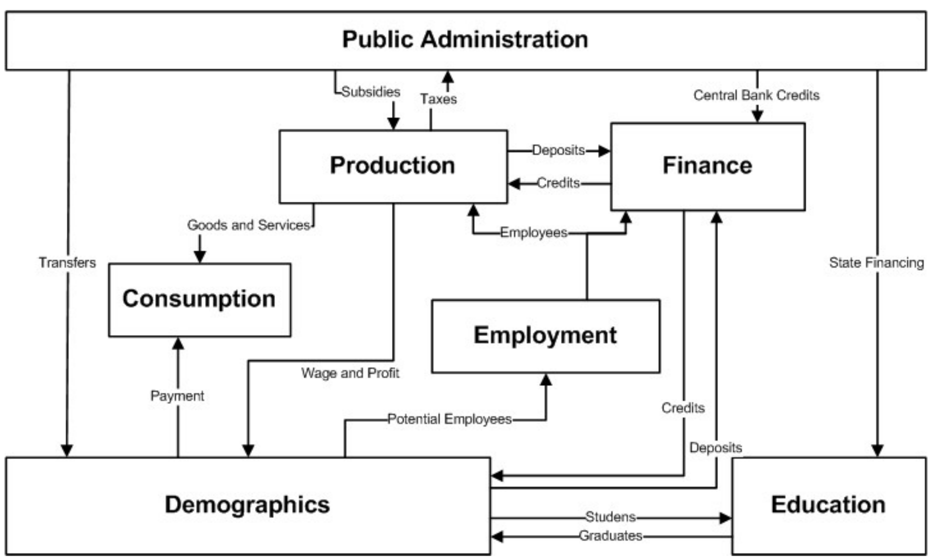

Fig. 1. Interrelation between modules of an artificial society.

The module "Demographics" reflects birth, maturation and death of agents. New households are formed after marriages and divorces. Agents can act as labor, tax-payers, consumers, creditors and students, thus interacting with the environment and with each other [13]. We set network connections between the agents and determine closeness of the connection in the range from 0 to 1 using the following rules: 0 – not acquainted; 0.1..0.3 – neighbors, colleagues; 0.4..0.6 – friends; 0.7..0.9 – relatives. Relationships among agents are reflected in a square matrix (Table 1). Closeness of the relationship 1 agent has only to himself (the main diagonal of the matrix). Connections are transitive: if the first agent does not have direct connection (closeness is 0) with the third agent, but has connection with the second agent (e.g. 0.5) and the second agent is connected with the third agent (e.g. 0.8), closeness of the indirect connection between the first and the third agent will be equal to the product of the intermediate connections (described in example 0.5 * 0.8 = 0.4). Indirect connections are important for disseminating information process. Information about the agent will be available for another agent, if the relationship between them (direct or indirect) is higher than the threshold [14].

Table 1. Network connections between the agents.

Agents	A_1	A_2	...	A_j	...	A_n
A_1	1	0.6	...	0	...	0.3
A_2	0.2	1	...	0	...	0.1
...	...	...	...	...	...	...
A_i	0	0	...	0.8	...	0.1
...	...	...	...	...	...	...
A_n	0.3	0.1	...	0.2	...	1

The model includes three types of organizations: commercial, financial and budgetary. Organizations interact with individual agents within hiring employees, paying wage or firing them. Beyond that, educational organizations recruit students, promote them to the following courses and graduate them (module "Education"). Financial organizations give credits to agents and commercial organizations and take deposits from them (module "Finance").

Organizations permanently interact with counterparties within sales, deliveries and financial settlements for them; each operation is reflected in the accounting, which is a simplified version of the system adopted in the Russian Federation. Operations are written-off into the table "Accounting entries" as a set of the following structure:

<Date, Deb_acc, Deb_start, Cred_acc, Cred_start, Sum, Deb_fin, Cred_fin>, where Date – date of the operation, Deb_acc – debit account of the entry, Deb_start – value of the debit account before the operation, Cred_acc – credit account of the entry, Cred_start – value of the credit account before the operation, Sum – sum of the operation, Deb_fin – value of the debit account after the operation, Cred_fin – value of the credit account after the operation [15].

For reproducing economic dynamics in the model we implement an algorithm of organizational decision making instead of a production function that is widely used in macroeconomic models [4], [18]. Trading agents in each region compare stocks of products that remained at the end of the year with the stocks that were available at the beginning. Exceeding the volume of current stocks over the initial level means an increase in demand for final products compared to the previous year; as a result volume of wholesale orders for the final product is growing. This causes a loop of positive feedback in the form of increased production and supply of materials and components. In case of a shortage of production capacity, enterprises implement investment programs, if necessary attracting credit resources from financial organizations. Decrease of demand for final products causes decrease of production and employment, which in turn leads to a reduction in deliveries and investment programs.

The public administration determines structure of the budget, taxation scale, transfer payments, the interest rate and other parameters. Administrative functions are implemented through educational, medical, social security and defense budgetary organizations.

At the current research stage we generate information objects of each module and their standard functions, excluding procedures of decision making of agents and organizations that determine dynamics of the system.

3 Initial Modeling Data Structure

Initial modeling data that reflects socio-economic structures at the base year is presented in Excel tables. The tables contain information on the demographic structure of the population, organizations and their economic interrelations, production, import, export, employment, financial characteristics of organizations and households, tax rates, transfer payments and other. Information content of the tables is based on the collections of Federal State Statistics Service [3], All-Russian Population Census, Economic Development Ministry's reports, Bank of Russia [1] and Ministry of Finance [2] open data. Table 2 presents the basic tables of input data used in the creation of objects in different modules.

Information presented in statistical resources requires preprocessing to match ini-tial modeling data structure. For example, to reproduce sector structure of economy in each region we need an aggregated table "Organizations". In the official statistics sector structure of the economy, cross-sector interrelations, export and import are presented in the input-output tables; regional production structure is presented in the table "Gross added value of the regions by sectors of the economy" in the statistical yearbook. Direct comparison of data of these two tables is impossible for two reasons. First, calculation of the gross regional product differs from calculation of the gross domestic product, as a result of which sum of gross regional product in all regions is less than the gross domestic product. Secondly, information on regional production is presented in the form of economic activity types, which implies less detail in comparison with the sector structure (for example, 37 sectors are classified as one economic activity type "manufacturing activities"). Thus, regional representation of production in

Table 2. Initial modeling data structure.

Module	Initial data tables
Demographics	Population by age groups Age-sex composition and status in marriage
Production	GDP branch structure GDP regional structure Export structure Import structure
Finance	Credit structure Deposit structure
Employment	Labor force size and composition Unemployment by age groups and educational attainment Accrued average monthly nominal wages of employees of organizations by economic activity
Consumption	Structure of money income and expenditures of population Subsistence minimum level
Education	Organizations carrying out training under education programs
Public Administration	Consolidated budget of the Russian Federation

the model requires matching statistical data from different sources. Necessary calculations are presented in steps:

1. Calculation of share of each sector in the corresponding type of economic activity on the basis of the input-output table.

$$d_s a = V_s / V_a \tag{1}$$

$d_s a$ - share of sector s in economic activity type a, V_s - gross product of sector s, calculated by method of added value, V_a - gross product of economic activity type a, calculated by method of added value; sector s belongs to economic activity type a.

2. Correction of the table of output of economic activities in regions, taking into account difference in the domestic product and total amount of regional products in separate economic activities:

$$k_a = V_a / \sum_{r=1}^{90} V_{r,a} \tag{2}$$

$$v_{r,a}^k = v_{r,a} \cdot k_a \tag{3}$$

k_a – correction coefficient of economic activity type a; V_a – gross product of economic activity type a; $v_{r,a}$ – product of economic activity type a in region r, presented in statistical tables; $v_{r,a}^k$ – corrected product of economic activity type a in region r.

3. Completion of the table of output of sectors in regions:

$$v_{s,r} = v_{r,a}^{k} \cdot d_{s,a} \quad (4)$$

$v_{s,r}$ – product of sector s in region r; sector s belongs to economic activity type a.

Preprocessing of initial data is required for all modules except from Demographics, since the recent information about composition of households in each region is reflected in results of All-Russian Population Census [3].

4 Algorithm of Artificial Society Reconstruction

Reconstruction of an artificial society is carried out in the base year of modeling. The first step is to set geographical structure of the Russian Federation (see Fig. 2). After that the original generation of agents is created, distributed among households and resettled by regions; the composition of households is determined by data of All-Russian Population Census of 2010 [3]. After that, the financial state of households is initialized.

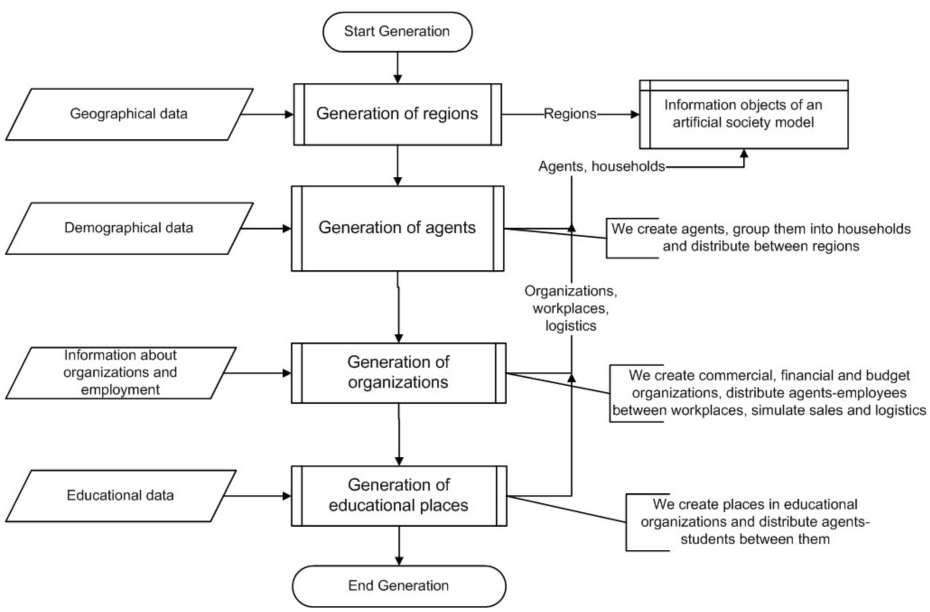

Fig. 2. Algorithm of reconstructing an artificial society using initial data.

Organizations in the model are aggregated: one organization in the model re-sponds to a set of organizations of one economic sector in the region. Generation of organizations is based on the input-output table, which determines the gross output of each economic sector, and the regional distribution of production table [3]. After generating organizations we set their type – commercial, financial or budgetary, and initialize values of their accounts. Economic interrelations between organizations, including sales and logistics, are set by the first quadrant of the input-output table. Agents are distributed to workplaces in accordance with their qualifications and employment structure in each economic sector.

Educational institutions are associated with sets of educational places for various groups of specialties and levels of education: school, secondary professional educa-tion; bachelor's, master's or postgraduate courses. Agents of the appropriate age are assigned to educational places.

The generated artificial society is stored in a database for later use in a series of scenario calculations.

5 Program realization

Model of an artificial society was programmed on C# in Microsoft Visual Studio 2015, which is free available for scientific research. Figure 3 shows the sequence of data processing in the model. The initial modeling data is loaded in the form of Excel tables, after that it is checked for completeness and consistency. In the module of an artificial society generation, the initial modeling data is transformed to information objects of the model (agents, households and organizations).

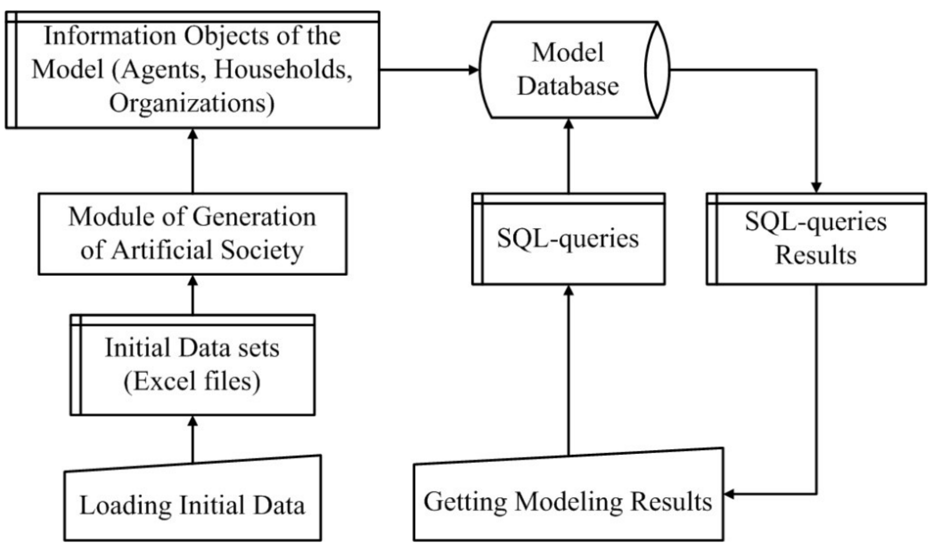

Fig. 3. Data processing in the model.

Results of the generation procedure are stored in the model database, access to them is provided by SQL-queries. The main resulting tables are accounting entries of organizations, households and state administration. Thus, GDP is calculated as sum of profits, wages and taxes paid by commercial, financial and budget organizations (we use method of calculating GDP by added value). Using grouping queries, it is possible to present gross output in the context of sectors and regions; flows of imports and exports in different sectors; household incomes and expenditures; savings and credits of households and organizations.

Interface of the software application of the model is presented in Figure 4 ("Start Modeling" button launches algorithms of dynamics that are being developed at the moment).

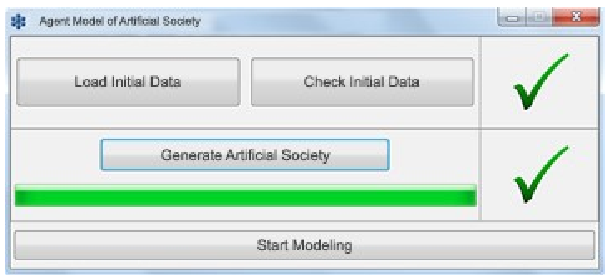

Fig. 4. Screenshot of the model interface.

6 DES-analysis for Verification of the Model

Verification of the artificial society model was carried out in two stages. The first step is validation of algorithms through the test data set. At this stage, the procedures for entering, conversion, storage, retrieval of data and algorithms for generating model objects were checked. At the second stage, the model was verified on the basis of retrospective data. For verification of the model on retrospective data, we propose a method, which compares demographic, economic and social indicators of the simulation output data with the real values of these indicators (DES-analysis).

The verifying simulation was carried on a time period of one year (12 clock ticks). Population of the Russian Federation was represented by 1.5 million agents, that is, one agent in the model corresponds to 100 residents; for the convenience of further interpretation of the output data, at the end of modeling inverse scaling was performed. Taking into account the accepted assumption that one organization in the model responds to a set of organizations of one economic sector in the region, about 4.5 thousand organizations were created in the model (58 sectors in input-output tables and 90 regions, some sectors are not presented in certain regions). We conducted a series of 10 experiments; the following Table 3 gives a summary of verification results.

Table 3. DES-analysis parameters variation.

Modules	Parameter	Real Value	Averaged modeling results	Variation, %
Demographics	Population, thousand person	143667	143667	0.0
	Men age 5 and younger, thousand	4569	4562	0.15
	... (other sex-age groups)	...	...	...
	Women age 70 and older, thousand	9630	9651	0.22
	Number of households, thousand	54560	54560	0.00
	Single households, thousand	14019	14019	0.00
	... (other household types)	...	...	...
Economics	GDP, billion RUR	70975	69059	2.70
	Agriculture output, billion RUR	4764	4752	0.19
	... (other sectors' output)	...	...	...
	Import, billion RUR	16530	16603	0.44
	Agriculture import, billion RUR	584	586	0.31
	... (other sectors' import)	...	...	...
	Export, billion RUR	16212	16077	0.83
	Agriculture export, billion RUR	277	279	0.74
	... (other sectors' export)	...	...	...
	Average wage, thousand RUR	32.5	32.2	0.92
Social situation	Unemployed, thousand person	1026	1028	0.65
	Poor, thousand person	16091	16229	0.86

The set of demographic indicators for comparison includes population number by sex and age groups, number and composition of households. As economic indicators we have chosen GDP, gross output, imports and exports of economic sectors. Social indicators are the average wage, number of unemployed and number of people below the poverty line. Variation of the observed parameters is determined by stochastic elements of generation procedures, including age of agents, composition of households, deviation of wage values in different economic sectors and regions from the average value. To analyze the adequacy of the model, the output of the generation procedures was averaged over 10 runs [6], [9].

Statistical analysis of modeling results for the base year (2014) showed deviation of the selected parameters within 1%, excluding GDP, which variation is about 3%. It is connected mostly with simplifying the tax system in the model: for organizations we set 13% rate, while some sectors of the economy pay additional taxes and it makes significant contribution in GDP. Taking this assumption into consideration, we can conclude that calculated variation of modeling parameters indicates a sufficient accuracy of artificial society reconstruction for the base year. However, we should take into consideration that models performing well on data sets available at the time of their publication might perform less well or badly when applied to post-publication data [17]. Due to this reason we plan to verify the model at a few input datasets (years 2015-2017) after they become available in official statistics.

For a short period of one year, we estimate deviation of modeling results from retrospective values, which is sufficient for assessing similarity of real society and reproduced artificial society for the base year of modeling. Within calibration of the model on longer time series (at least 5 years), it would be possible to determine autocorrelation and mutual influence of various factors using methods of regression analysis.

7 Conclusions

The computer model of an artificial society is designed for reproducing geographical distribution of population of the Russian Federation and its current socio-economic situation. In this paper we describe structure of the initial parameters of the simulation and the procedure for their verification. Since the aim of our research is forecasting economic development of Russia and assessing the impact of the state economic policy on this process in our future work we are going to add algorithms of economic dynamics to the model and calibrate them on retrospective data for 2014-2017 period. After official registration of the program model, access to the source code would be available through Team Foundation Server of Microsoft Visual Studio 2015. However, to reproduce presented computational results, preprocessed initial data sets should be entered to the model. In order to increase the accuracy of the prognosis we plan to specify the initial modeling data by adding sociological surveys and results of social networks monitoring, which would help to reflect subjective parameters and social moods that are not obvious within standard statistical methods.

Acknowledgments. The reported study was funded by RFBR according to the research project 18-310-00185.

References

1. The Central Bank of the Russian Federation official site, http://www.cbr.ru/eng/, last accessed 2018/03/22
2. Ministry of Finance of the Russian Federation official site homepage, http://old.minfin.ru/en/statistics/, last accessed 2018/03/18
3. Russian Federation Federal State Statistics Service, http://www.gks.ru, last accessed 2018/03/26
4. Bakhtizin, A.: Agent-based models of economy. Ekonomika, Moscow (2008), (in Russian)
5. Barros, J.: Agent-based models of geographical systems. chap. Exploring Urban Dynamics in Latin American Cities Using an Agent-Based Simulation Approach., pp. 571–589. Springer, Dordrecht (2012)
6. Baucells, M., Borgonovo, E.: Invariant probabilistic sensitivity analysis. Management Science (59(11)), 25362549 (2013)
7. Bonabeau, E.: Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences **99**(suppl 3), 7280–7287 (2002)

8. Epstein, J., Axtell, R.: Growing Artificial Societies: Social Science From the Bottom Up. MIT Press, Brookings Institution (1996)
9. Fonoberova, M., Fonoberov, V.A., Mezi, I.: Global sensitivity/uncertainty analysis for agent-based models. Reliability Engineering & System Safety **118**, 8 – 17 (2013). https://doi.org/10.1016/j.ress.2013.04.004, http://www.sciencedirect.com/science/article/pii/S0951832013000999
10. Lee, J.S., Filatova, T., Ligmann-Zielinska, A., Hassani-Mahmooei, B., Stonedahl, F., Lorscheid, I., Voinov, A., Polhill, G., Sun, Z., Parker, D.: The complexities of agent-based modeling output analysis. JASSS : the journal of artificial societies and social simulation **18**(4) (2015)
11. Macy, M.W., Willer, R.: From factors to factors: Computational sociology and agent-based modelling. Annual Review of Sociology **28**, 143–166 (2002)
12. Makarov, V.L., Bakhtizin, A.R., Sushko, E.D., Vasenin, V.A., Borisov, V.A., Roganov, V.A.: Supercomputer technologies in social sciences: agent-oriented demographic models. Herald of the Russian Academy of Sciences **86**(3), 248–257 (2016)
13. Mashkova, A.L., Demidov, A.V., Savina, O.A., Koskin, A.V., Mashkov, E.A.: Developing a complex model of experimental economy based on agent approach and open government data in distributed information-computational environment. In: eGose '17: Proceedings of the Internationsl Conference on Electronic Governance and Open Society: Challenges in Eurasia. pp. 27–31. ACM international conference proceedings series, ACM, New York, NY, USA (2017)
14. Mashkova, A.L., Novikova, E.V., Savina, O.A.: Agent model for evaluating influence of tax policy on political preferences. In: EGOSE '16: Proceedings of the International Conference on Electronic Governance and Open Society: Challenges in Eurasia. pp. 258–261. ACM, New York, USA (2016)
15. Mashkova, A.L., Savina, O.A.: Management of financial flows of organizations in the agent model of the experimental economy. Upravlencheskiy uchet **12**, 89–98 (2015), (in Russian)
16. Mashkova, A.L., Savina, O.A., Lazarev, S.A.: Agent model for evaluating efficiency of socially oriented federal programs. In: 11th IEEE International Conference on Application of Information and Communication Technologies (AICT). vol. 2, pp. 217–221. Institute of Control Sciences of Russian Academy of Sciences, Moscow (2017)
17. Moss, S.: Alternative approaches to the empirical validation of agent-based models. Journal of Artificial Societies and Social Simulation **11**(1), 5 (2008)
18. Ogibayashi, S., Takashima, K.: Influence of the corporation tax rate on gdp in an agent-based artificial economic system. In: Chen, S., Terano, T., Yamamoto, R., Tai, C. (eds.) Advances in Computational Social Science. Agent-Based Social Systems, vol. 11, pp. 147–161. Springer, Tokyo (2014)
19. Tesfatsion, L.: Agent-based computational economics: Growing economies from the bottom up. Artificial Life **8**(1), 55–82 (2002)