Chapter 18
Processing Initial Data for the Agent-Based Model of the Russian Federation Spatial Development

Aleksandra L. Mashkova

Abstract In our research, the preparation of initial data for the simulation model of the spatial development of the Russian Federation is being conducted. Within the model database, information about population, economy, and social institutions from the official open sources: Web sites of the ministries, federal, and regional statistical yearbooks, is integrated. While information about population and education is quite detailed, referring to the production system it is disaggregated and needs processing to be converted into required structure. In this paper, structure of initial data for simulating production in the model and methods of matching it with available open data are presented. Iterative proportional fitting technique is implemented for aggregating information about employment and equipment of organizations in different regions. Within the procedure of generation, the model database is filled with objects created on the basis of initial data.

18.1 Introduction

Our research is aimed at developing a tool for evaluating alternative managerial decisions by simulating different scenarios of the spatial development of the Russian Federation. Creation of such a model requires integration of population, economic structures, and social institutions to reflect a wide range of control actions, including tax, monetary, and investment policies. Significance of practical application of the model directly depends on completeness and detailing of its input data, which ensure that the model environment corresponds to the real world and the forecast estimates obtained on its basis are adequate.

Initial modeling data on the population, production, infrastructure, financial state of organizations, and households can be found in federal and regional statistical

A. L. Mashkova (✉)
Orel State University named after I.S. Turgenev, Komsomolskaja St. 95, 302026 Orel, Russian Federation
e-mail: aleks.savina@gmail.com

Central Economics and Mathematics Institute, Russian Academy of Sciences, Nakhimovsky Av. 47, 117418 Moscow, Russian Federation

S.-L. Peng et al. (eds.), *Sensor Networks and Signal Processing*, Smart Innovation, Systems and Technologies 176, https://doi.org/10.1007/978-981-15-4917-5_18

collections [1, 2], open data of ministries [3–5], and results of sociological surveys [6, 7]. Since this information was collected by different sources and for different reasons, it needs processing to be converted into required structure of initial modeling data tables.

Of particular difficulty is the search for information in a spatial context. To analyze demographic processes, one can rely on census data [1], which provides a full range of necessary information about population in regional, urban, and rural context. Detailed information on higher and secondary vocational education is presented in the reports of the Ministry of Education and Science [5]. For economic structures, obtaining information in a spatial context is fraught with considerable difficulties. The main source of information is the Federal State Statistics Service Web site [2] and its annual collections, but information provided in them is significantly generalized. More detailed information is available in regional collections, but there are a few difficulties that determine the need for initial data preprocessing:

1. Detailing of sector structure of gross regional product is presented in collections for only 14 regions out of 82.
2. In these 14 collections, some economic activities are not detailed; for example, economic activity ‘personal services’ is detailed only in collections of Karelia and Arkhangelsk.
3. Often, 2–3 related sectors are combined into a complementary set (e.g., food and tobacco industry); moreover, composition of complementary sets differs from one regional collection to another.

The purpose of the presented work is integration of data from different sources, which retains high detail for the regions that presented the most complete information and uses proportional fitting techniques for others. The solution of the indicated problems is only the first step in the study, since in the future it is planned to integrate data on individual organizations and corporations within the model database. The task of statistical data preprocessing, however, is urgent for the development of the demo version of the model and debugging of its algorithms.

18.2 Research Methodology

For modeling spatial development of the Russian Federation agent-based approach was chosen [8], which includes heterogeneity, bounded rationality, and global dynamics as a result of micro-level agents’ interactions. Based on these assumptions, agent-based models become computer laboratories to test the effects of policies on macroeconomic and spatial dynamics [9]. Agent-based models have already been implemented in different policy areas such as fiscal [10], monetary [11], financial [12], and labor market policy [13].

Loading real data into agent-based models remains an urgent task. There are examples of solving this task for different regions, for example, for Leeds, the UK [14], and Saint Petersburg, Russia [15]. Due to the regional scale of these models, only regional management measures are simulated in these models, while macroeconomic effects are treated as the external environment.

Within creation of the model of spatial development of the Russian Federation sex-age structure of the population, infrastructure, production, administration, and educational institutions in each region are reconstructed. Dynamics of the system is simulated through behavior of agents and decisions of organizations.

In this paper, methods of processing initial modeling data used for reconstructing production system of the Russian Federation in the spatial context are considered. For data processing, iterative proportional fitting (IPF) is used, which is a procedure for assigning values to internal cells based on known marginal totals in a multidimensional matrix [16]. In different branches of academic research, IPF is called bi-proportional fitting or RAS algorithm [17]. IPF is widely used in spatial microsimulation studies for integrating geographically aggregated data sources [18].

18.3 Structure of the Agent-Based Model

In our research, the agent-based model is being developed, which simulates demography, production, employment: financial, education, and budgetary system (see Fig. 18.1). Population and organizations are related to regions of the Russian Federation [19].

In the module, 'population' demographic processes are reflected, including birth and death, as well as creation of new households connected with marriages and divorces. Within other modules, agents act as students, employees and employers, taxpayers and consumers.

Organizations in the model are aggregated: one organization in the model responds to a set of organizations of one economic sector in the region [20]. There are commercial, budgetary, and financial organizations in the model. Commercial organizations interact within 'production and service' module. Financial institutions in the model

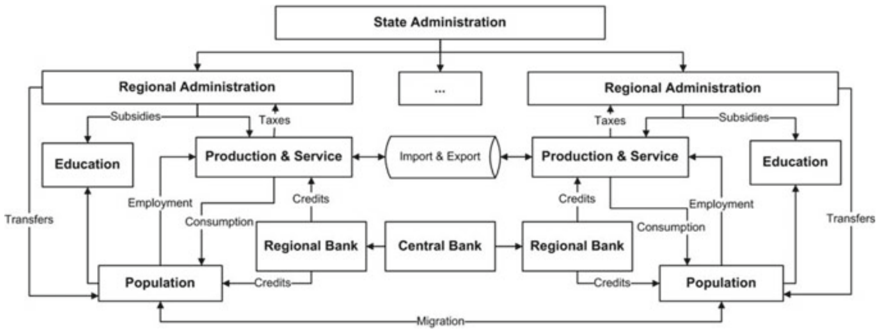

Fig. 18.1 Main processes in the agent-based model of spatial development of the Russian Federation

are the Central Bank (that sets the interest rate and issues bonds) and regional banks, which deal with deposits and credits of households and organizations. The state administration is responsible for tax collection, social transfers, and federal investment programs. The regional administration implements their functions through budgetary organizations in social security, education, and medicine [19].

18.4 Data Processing Algorithms

In the further paragraphs, methods of processing data from the portal of the Federal State Statistics Service [2], federal and regional statistical yearbooks are considered. Information on regional production, employment, property, and equipment is presented in the economic activities scale, which is much less detailed than the sector scale used in the federal input–output table [19]. Organizations in the model, however, represent sector structure of regional economies. Table 18.1 shows that available data requires preprocessing to match initial modeling data structure.

18.4.1 Production and Service

Firstly, production volume of each sector in the complemented set $P_r^{\sum_{i=1}^{k} s_i}$, presented in the regional statistical yearbook, is calculated.

$$\widehat{d_r^s} = \frac{P^s}{\sum_{i=1}^{k} P^i} \tag{18.1}$$

$\widehat{d_r^s}$—share of sector s in the complemented set in region r; P^s, P^i—gross product of sectors in Russia, k—number of sectors in the complemented set.

$$P_r^s = P_r^{\sum_{i=1}^{k} s_i} * \widehat{d_r^s} \tag{18.2}$$

P_r^s—product of sector s in region r; $P_r^{\sum_{i=1}^{k} s_i}$—product of the complemented set of sectors.

Results of processing data about production in economic activity 'transport and communication' in region 1 (Belgorod) are presented in Table 18.2. Supplementary sets are: 'land transport service' and 'water transport service'; 'additional transport service' and 'communication'.

Table 18.1 Initial modeling data for simulating spatial distribution of production

Initial modeling data tables	Table content	Available open data tables	Table content
Production	P_r^s—product of sector s in region r	Input–output table	P^s—product of sector s, billion RUR
		Aggregated gross regional product structure	P_r^e—product of economic activity e in region r
		Detailed gross regional product structure	P_r^s—product of sector s in region r
Export and import	e^s; i^s—share of export and import in the output of sector s	Input–output table	E^s; I^s—volume of export and import in sector s
Supply	a_{ij}—volume of sector i product used in production of a unit of sector j	Input–output table	x_{ij}—cost of sector i product used in total production of sector j
		Product of sectors in physical terms	Pr^s—product of sector s in physical terms (standard units)
Property, plants and equipment	PE_r^s—volume of property, plant, and equipment in sector s in region r	Property, plant, and equipment in different regions	PE_r—volume of property, plant, and equipment in region r
		Property, plant, and equipment of organizations of different economic activities	PE^e—volume of property, plant, and equipment in economic activity e
Employment	L_r^s—number of employees in sector s in region r	Number of employees in different regions	L_r—number of employees in region r
		Number of employees of organizations of different economic activities	L^e—number of employees in economic activity e

Total product of sector s in regions from 1 to u with detailed statistical data is:

$$V_s^{\det} = \sum_{r=1}^{u} P_r^s \tag{18.3}$$

$V_s^{\det}$—total product of sector s in regions from 1 to u; u—number of regions with detailed statistical data about sector s; P_r^s—product of sector s in region r.

For different economic activities u accepts values from 2 (‘personal services’) to 14 (manufacturing). In economic activity ‘transport and communication’ $u = 11$.

Table 18.2 Results of data processing (transport and communication in Belgorod region)

Sectors in economic activity 'transport and communication'	$P_1^{\sum_{i=1}^{2} s_i}$, billion RUR	P^s, billion RUR	$\hat{d}_r^s$	P_1^s, billion RUR
Land transport service	25.6	5596.5	0.97	24.8
Water transport service		191.3	0.03	0.8
Air and space transport service	0	928.9	0	0
Additional transport service and communication	13.8	2669.5	0.58	8.0
Communication		1971.0	0.42	5.8

For each sector its total product in other regions is:

$$V_s^{\text{ost}} = V_s - V_s^{\text{det}} \tag{18.4}$$

V_s^{ost}—total product of sector s in regions from u to 82.

For regions from u to 82, approximate share of sector s in the corresponding economic activity type is calculated:

$$\widehat{d^{s-e}} = V_s^{\text{ost}} / \sum_{r=u+1}^{82} P_r^e \tag{18.5}$$

$\widehat{d^{s-e}}$—approximate share of sector s belonging to economic activity e; P_r^e—product of economic activity e in region r.

For further operations, share of sectors in economic activities in different regions is calculated:

$$d_r^{s-e} = \frac{P_r^s}{P_r^e} \tag{18.6}$$

d_r^{s-e}—share of sector s in economic activity e in region r, P_r^s—product of sector s in region r, P_r^e—product of economic activity e in region r.

18.4.2 Export and Import

Share of export in the output of sector s is:

$$e^s = \frac{E^s}{P^s} \tag{18.7}$$

e^s—share of export in the output of sector s; E^s—volume of export in sector s; P^s—product of sector s.

Share of import in consumption of product sector s is:

$$i^s = \frac{I^s}{P^s} \tag{18.8}$$

i^s—share of import in sector s; I^s—volume of import in sector s; P^s—total consumption of product of sector s.

Since more detailed information is not available, it is assumed that share of export in output and import in supply of each organization is equal to the average share for the sector, which the organization belongs to.

18.4.3 Supply

In the federal input–output tables, deliveries between sectors are presented in monetary terms. In order to preserve quantitative relationships in the real economy in the period of changes in currency exchange rates and inflation, accounting of production in standard units for each sector is used. For this task, the price of a standard unit of each sector is calculated:

$$\text{Price}^s = \frac{P^s}{Pr^s} \tag{18.9}$$

Price^s—price of a standard unit of sector s; P^s—product of sector s in monetary terms; Pr^s—product of sector s in physical terms (standard units)

Supply matrix in the model should be presented in physical terms as well:

$$a_{ij} = \frac{x_{ij}}{Pr^j * \text{Price}^i} \tag{18.10}$$

a_{ij}—element of supply matrix (volume of sector i product used in production of a unit of sector j); x_{ij}—cost of sector i product used in total production of sector j (presented in the federal input–output table); Pr^j—product of sector j in physical terms; Price^s—price of a standard unit of sector s.

18.4.4 Property, Plants, and Equipment

For aggregating information about property, plant, and equipment of organizations, iterative proportional fitting algorithm (IPF) is used. Initial data includes PE_r—total volume of property, plant, and equipment in region r (table 'property, plant, and equipment of organizations in different regions' presented on Federal State Statistics Service Web site) and PE^e—total volume of property, plant, and equipment in economic activity e (table 'property, plant, and equipment of organizations of different economic activities' presented on Federal State Statistics Service Web site).

The aim of calculations is defining PE_r^e—total volume of property, plant, and equipment in economic activity *e* in region *r*, so that:

$$PE^e = \sum_{r=1}^{82} PE_r^e \tag{18.11}$$

$$PE_r = \sum_{e=1}^{15} PE_r^e \tag{18.12}$$

$$PE = \sum_{r=1}^{82} PE_r = \sum_{e=1}^{15} PE^e \tag{18.13}$$

where $PE = 74{,}662.4$ billion RUR—total volume of property, plant, and equipment in Russia [7], PE^e and PE_r are presented in appropriate line and column in Table 18.3.

At first, preliminary volume of property, plant, and equipment in economic activities in each region is calculated:

$$PE_r^e(\text{preliminary}) = PE^e * P_r^e \tag{18.14}$$

PE_r^e(preliminary)—preliminary volume of property, plant, and equipment in economic activity *e* in region *r*, PE^e—total volume of property, plant, and equipment in economic activity *e*, P_r^e—share of region *r* in production volume of economic activity *e*.

Iteration 1 is calculating variation of preliminary volume of property, plant, and equipment in regional scale:

$$v_r(i1) = \frac{PE_r}{\sum_{e=1}^{15} PE_r^e(\text{preliminary})} \tag{18.15}$$

PE_r—baseline volume of property, plant, and equipment in region *r*; PE_r^e(preliminary)—preliminary volume of property, plant, and equipment in economic activity *e* in region *r*.

Table 18.3 Initial data and results of IPF algorithm for calculating property, plant, and equipment in different regions and economic activities

Parameter	PE_r, billion RUR	$\sum_{e=1}^{15} PE_r^e(i4)$ billion RUR	$PE_r^e(i4)$, billion RUR					
			Agriculture and forestry	Fishery	Mining	Manufacturing	…	Personal service
PE_1 (Belgorod)	651.4	651.4	93.1	0.0	66.6	88.4	…	11.6
PE_2 (Bryansk)	288.7	288.7	27.0	0.0	0.2	34.2	…	5.0
PE_3 (Vladimir)	376.7	376.7	13.8	0.0	1.2	72.9	…	12.5
PE_4(Voronezh)	673.9	673.9	57.0	0.0	2.6	57.6	…	10.6
…	…	…	…	…	…	…	…	…
PE_{82}(Chukotka)	54.8	54.8	0.6	0.0	20.3	0.1	…	2.0
PE^e	74,662.4	–	2209.9	61.0	6950.9	7199.4	…	1567.6
$\sum_{r=1}^{82} PE_r^e(i4)$	–	74,662.4	2221.0	62.4	7174.1	7173.9	…	1560.0
$v^e(i4)$, %	–	0	0.5	2.2	3.1	0.4	…	0.5

Maximal variation $v_r(i1)$ is 127% in Chechen Republic. Since variation is very high, the volume of property, plant, and equipment in economic activity e in region r using $v_r(i1)$ is calculated and used as correcting coefficient:

$$PE_r^e(i1) = PE_r^e(\text{preliminary}) * v_r(i1) \tag{18.16}$$

$PE_r^e(i1)$—volume of property, plant, and equipment in economic activity e in region r at the first iteration.

Iteration 2 Calculation of variation of volume of property, plant, and equipment in the scale of economic activities:

$$v^e(i2) = \frac{PE^e}{\sum_{r=1}^{82} PE_r^e(i1)} \tag{18.17}$$

PE^e—baseline volume of property, plant, and equipment in economic activity e; $PE_r^e(i1)$—volume of property, plant, and equipment in economic activity e in region r at the first iteration.

Maximal variation $v^e(i2)$ is 19.1% in economic activity 'fishery,' so the data needs further processing:

$$PE_r^e(i2) = PE_r^e(i1) * v^e(i2) \tag{18.18}$$

$PE_r^e(i2)$—volume of property, plant, and equipment in economic activity e in region r at the first iteration.

Repeating iterations, on the fourth the deviation in regional scale goes down to zero and maximum 3.1% in economic activity scale (Table 18.3).

Further information about property, plant, and equipment of organizations is detailed to the sector scale:

$$PE_r^s = PE_r^e(i4) * d_r^{s-e} \tag{18.19}$$

PE_r^s—property, plant, and equipment of an organization, representing sector s in region r; $PE_r^e(i4)$—volume of property, plant, and equipment in economic activity e in region r at the fourth iteration.

18.4.5 Employment

IPF is also applied for aggregating information about employment, which is presented separately in regional scale and in scale of economic activities. Starting from maximum variation 232% in regional scale and 22% in economic activity scale, on the fourth iteration a zero and 2.5% variation accordingly is reached.

Number of employees of organizations in different sectors is also calculated:

$$L_r^s = L_r^e(i4) * d_r^{s-e} \tag{18.20}$$

L_r^s—number employees of an organization, representing sector s in region r; $L_r^e(i4)$—number employees in economic activity e in region r at the fourth iteration.

18.5 Further Data Processing

The aim of the first stage of our research methodology is to prepare and process initial data into detailed and interconnected objects of the database. The population and organizations at the base year of modeling are reconstructed by filling the model database step by step. Structure of the model database is presented in [19].

Algorithm of artificial society reconstruction is considered in [20]; it includes the following steps:

1. Set geographical structure of the Russian Federation (82 regions).
2. Create the original generation of agents in accordance with the sex-age structure of population and structure of households in each region (on the basis of All Russian Population Census data).
3. Generation of aggregated organizations in the regions. After that the property, plant, and equipment, financial state (credits and deposits) and accounting of each organization are initialized, as well as its supplies. Each organization has employees, which are assigned to it through workplaces with different qualification and salaries
4. Generation of educational places and assigning agents of the appropriate age to them. Number of educational places in higher and secondary vocational education is presented in the reports of the Ministry of Education and Science [5].

Procedure of synthetic population, organizations, and institutions generation was programmed on C# in Microsoft Visual Studio 2015. Generated objects are stored in the model database for later access via SQL-queries. Modeling results are presented in [20].

18.6 Results and Discussion

In this paper, methods of data preprocessing for the agent-based model of spatial development of the Russian Federation have been presented. At the first stage of the research, the population, production, and social institutions in the regions of Russia were reconstructed. This stage required collection of large amounts of information from various sources: All Russian Population Census, federal and regional statistical yearbooks, official information on the portals of the ministries and survey results.

While information on demographic situation and educational system in the regional context is quite fully represented in open official sources [1, 5], information on regional production and service requires processing. In official collections, information on employment and equipment is presented separately in the regional scale and in the scale of economic activities, while for simulating spatial development both of these aspects are crucial. The iterative proportional fitting algorithm was used to integrate such data arrays. Another problem was incompleteness of information in regional collections; for its solution, a method of sequential detailing was used, which maintained high accuracy for the regions that provided complete information.

The resulting arrays of initial data were converted using the generation algorithm [20] to the synthetic society that represents population of the Russian Federation in 2014, production, property, and equipment of organizations, employment, and educational system.

The task of loading real data is a crucial part of creating the model of the Russian Federation spatial development, since it gives the opportunity to evaluate impact of federal and regional policy on different categories of population, taking into account age, education, income and status in marriage, as well as their preferences and beliefs. The concept of integration large data arrays can be implemented in different countries which would seriously improve prognostic capabilities of the social simulation models.

Acknowledgements The reported study was funded by the Russian Foundation for Basic Research within the research project № 18-29-03049.

References

1. All-russian population census 2010 official website. http://www.gks.ru. Last accessed 22 Mar 2019
2. Russian Federation Federal State Statistics Service. http://www.gks.ru. Last accessed 15 Mar 2019
3. The Central Bank of the Russian Federation official site. http://www.cbr.ru/eng/. Last accessed 22 Mar 2019
4. Ministry of Economic Development of the Russian Federation official website. http://economy.gov.ru/minec/main. Last accessed 22 Mar 2019
5. Ministry of science and higher education Russian Federation official website. https://minobrnauki.gov.ru/. Last accessed 21 Feb 2019
6. Internet recruitment portal Head Hunter, https://hh.ru/article/research, last accessed 2019/02/06
7. National credit bureau official website. https://www.nbki.ru/company/news/. Last accessed 16 Mar 2019
8. Lebaron, B., Tesfatsion, L.: Modeling macroeconomies as open-ended dynamic systems of interacting agents. Am. Econ. Rev. **98**, 246–250 (April, 2008). https://doi.org/10.1257/aer.98.2.246
9. Fagiolo, G., Roventini, A.: Macroeconomic policy in DSGE and agent-based models redux: New developments and challenges ahead. J. Artif. Soc. Social Simul. **20**(1), 1 (2017). https://doi.org/10.18564/jasss.3280. http://jasss.soc.surrey.ac.uk/20/1/1.html

10. Napoletano, M., Roventini, A., Gaffard, J.: Time-varying fiscal multipliers in an agent-based model with credit rationing. Economics Discussion Papers 112, Kiel (2017). http://hdl.handle.net/10419/172323
11. Gatti, D., Desiderio, S.J.: Monetary policy experiments in an agent-based model with financial frictions. J. Econ. Interact. Coord. **10**(2), 265–286 (2015). https://doi.org/10.1007/s11403-014-0123-7
12. Ashraf, Q., Gershman, B., Howitt, P.: Banks, market organization, and macroeconomic performance: an agent-based computational analysis. In: Working Paper 17102, National Bureau of Economic Research (June 2011). https://doi.org/10.3386/w17102. http://www.nber.org/papers/w17102
13. Napoletano, M., Dosi, G., Fagiolo, G., Roventini, A.: Wage formation, investment behavior and growth regimes: an agent-based analysis. Revue de l'OFCE **124**(5), 235–261 (2012). https://doi.org/10.3917/reof.124.0235
14. Ballas, D., Kingston, R., Stillwell, J.: Using a spatial microsimulation decision support system for policy scenario analysis. In: Recent Advances in Design and Decision Support Systems in Architecture and Urban Planning. Springer, Dordrecht (2004)
15. Sushko, E.: Multi-agent model of the region: concept, design and implementation. Tech. Rep. Preprint WP/2012/292, CEMI RAS, in Russian (2012)
16. Cleave, N., Brown, P.J., Payne, C.D.: Evaluation of methods for ecological inference. J. Ro. Stat. Soc. Series A (Stat. Soc.) **158**(1), 55–72 (1995). https://doi.org/10.2307/2983403. https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2983403
17. Lahr, M.L., de Mesnard, L.: Biproportional techniques in input-output analysis: table updating and structural analysis. Econ. Sys. Res. **16**(2), 115–134 (2004). https://doi.org/10.1080/0953531042000219259
18. Lovelace, R., Birkin, M., Ballas, D., van Leeuwen, E.: Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an established technique. J. Artif. Soc. Social Simul. **18**(2), 21 (2015). https://doi.org/10.18564/jasss.2768. http://jasss.soc.surrey.ac.uk/18/2/21.html
19. Mashkova, A.L., Savina, O.A., Banchuk, Y.A., Mashkov, E.A.: Using open data for information support of simulation model of the russian federation spatial development. In: Chugunov, A., Misnikov, Y., Roshchin, E., Trutnev, D. (eds.) Electronic Governance and Open Society: Challenges in Eurasia, pp. 401–414. Springer International Publishing, Cham (2019)
20. Mashkova, A.L.: Reconstructing an artificial society on the basis of big open data. In: Supplementary Proceedings of the 7th International Conference on Analysis of Images, Social Networks and Texts (AIST-SUP 2018), vol. 2268, pp. 241–251. Moscow, Russia, 5–7 July 2018