Sample data calibration and actualization for socio-economic microsimulations.

Ter-Akopov Sergey
Higher School of Economics, Moscow, Russia
Zharomskiy Vladislav
Higher School of Economics, Moscow, Russia

Background of STEND

- Developed in 2003
- Based on NOBUS (National Survey of Household Welfare and Participation in Social Programs) carried out by Russian Federal State Statistics Service (Rosstat) and World Bank
- Main purpose – to justify progressive tax system in Russia

Available income surveys

1. National Survey of Household Welfare and Participation in Social Programs (NOBUS)
2. Generation and Gender Survey (GGS)
3. Statistical Survey of Income and Participation in Social Programs (SSIPSP)
4. Russia Longitudinal Monitoring Survey of HSE (RLMS-HSE)

Sample data

Sample Data	Households	Individuals	Regions covered
RLMS-HSE	≈8 000	≈21 000	≈40
SSIPSP	≈10 000	≈25 000	All (85)

Data

Uses data on

- Household formation
- Expenditures
- **Income structure**
- Education
- Different socio-economic characteristics
- etc

STEND structure

Population structure reweighting

Calculating additional weights which together with initial weights will make share of main population groups, which are considered to be important (e.g. pensioners, children under some age, unemployed or employed, teachers working in private schools etc.), to have minimal difference with the official statistics.

Income weights calculating

Income weights are used so that the income characteristics of the sample do not differ a lot from the official statistics. Not only main income characteristics (average income, wage rate, pensions, social transfers, etc.), but also their distributions should be the same or as close as possible to official statistics. Additionally if non-response rate in one of the income groups (most often in the highest and lowest income groups) is too high it is possible to correct income weights so that it would solve the problem.

Simulation

Sample data corrected on the 1st two steps than used in the tax benefit model to simulate any policy.

What is it capable of ?

1. Simulates individual and household tax liabilities and benefit entitlements based on a policy or some economic shock.
2. Examine the effects of actual policy over time if periodic survey is available
3. Evaluates proposed or hypothetical policy effects on regional and federal levels
4. Evaluates budget expenditures
5. Uses population structure re-weighting and income balancing mechanism

Population structure

	Official statistics		Sample data		Difference
Population	142 865 000	100,0%	143 749 365	100,0%	-0,6%
including					
Men	66 162 460	46,3%	65 188 219	45,3%	1,5%
Women	76 702 540	53,7%	78 561 146	54,7%	-2,4%
Children (under 18)	26 251 699	18,4%	28 275 835	19,7%	-7,7%
Pensioners	41 819 000	29,3%	38 081 219	26,5%	8,9%

Population structure

		Official statistics		Sample data		Difference
Employed		70 856 000	49,6%	76 319 720	53,1%	-15,6%
including						
	Male employed	36 032 000	25,2%	37 365 116	26,0%	-3,7%
	Female employed	34 824 000	24,4%	38 954 604	27,1%	-11,9%

Population structure

	Official statistics		Sample data		Difference
Employed	70 856 000	49,6%	76 319 720	53,1%	-15,6%
including working in					
Education	5 364 354	3,8%	6 343 153	4,4%	-18,2%
Healthcare	4 455 331	3,1%	5 700 558	4,0%	-27,9%

Income balancing (IPFP)

a_{11}	…	a_{1j}	…	a_{1m}	x_1	$\boldsymbol{w_1}$
…	…	…	…	…	…	…
a_{i1}	…	a_{ij}	…	a_{im}	x_i	$\boldsymbol{w_i}$
…	…	…	…	…	…	…
a_{1n}	…	a_{in}	…	$a_{n,m}$	x_n	$\boldsymbol{w_n}$
y_1	…	y_j	…	y_m		
$\boldsymbol{v_1}$	…	$\boldsymbol{v_j}$	…	$\boldsymbol{v_m}$		

Income balancing (IPFP)

x_i - total income of household i

y_j - share of component j in per capita income

a_{ij} - component j in per capita income of household j

Step 1.

$$w_i = \frac{x_i}{\sum_{j=1}^{m} v_j a_{ij}}$$

Step 2.

$$v_j = \frac{y_j}{\sum_{i=1}^{n} w_i a_{ij}}$$

Go to Step 1.

Population structure re-weighting

	Y_1	…	Y_j	…	Y_m		
w_1	a_{11}	…	a_{1j}	…	a_{1m}	x_1	X_1
…	…	…	…	…	…	…	…
w_i	a_{i1}	…	a_{ij}	…	a_{im}	x_i	X_i
…	…	…	…	…	…	…	…
w_n	a_{1n}	…	a_{in}	…	$a_{n,m}$	x_n	X_n
	y_1	…	y_j	…	y_m		
	v_1	…	v_j	…	v_m		

Population structure re-weighting

	Population	Men	Women	Employed	Children (U18)		
$Y_j =$	142 865 000	66 162 460	76 702 540	70 856 000	26 251 699		
w_1	4	1	3	1	2	x_1=11	X_1
…	…	…	…	…	…	…	…
w_i	1	1	0	1	1	x_i=4	X_i
…	…	…	…	…	…	…	…
w_n	5	1	3	1	2	x_n=12	X_n
$y_j =$	143 749 365	65 188 219	78 561 146	76 319 720	28 275 835		
	v_1	…	v_j	…	v_m		

Population structure re-weighting (step 1)

Calculating weighted sum y_j for each column

$$\begin{cases} \sum_{i=1}^{n} w_i a_{i1} = y_1 \\ \dots\dots\dots \\ \sum_{i=1}^{n} w_i a_{ij} = y_j \\ \dots\dots\dots \\ \sum_{i=1}^{n} w_i a_{im} = y_m \end{cases}$$

Evaluating difference between macro statistic (Y_i) and sampled data (y_i) calculated on the previous step.

$$v_j = \frac{Y_j}{y_j}$$

Population structure re-weighting (step 2)

New modified weights w_i^* would be equal to a sum with socio-economic group weights (v_j).

$$w_i^* = \frac{w_i \sum_{j=1}^{m} a_{ij} v_j}{\sum_{j=1}^{m} a_{ij}}$$

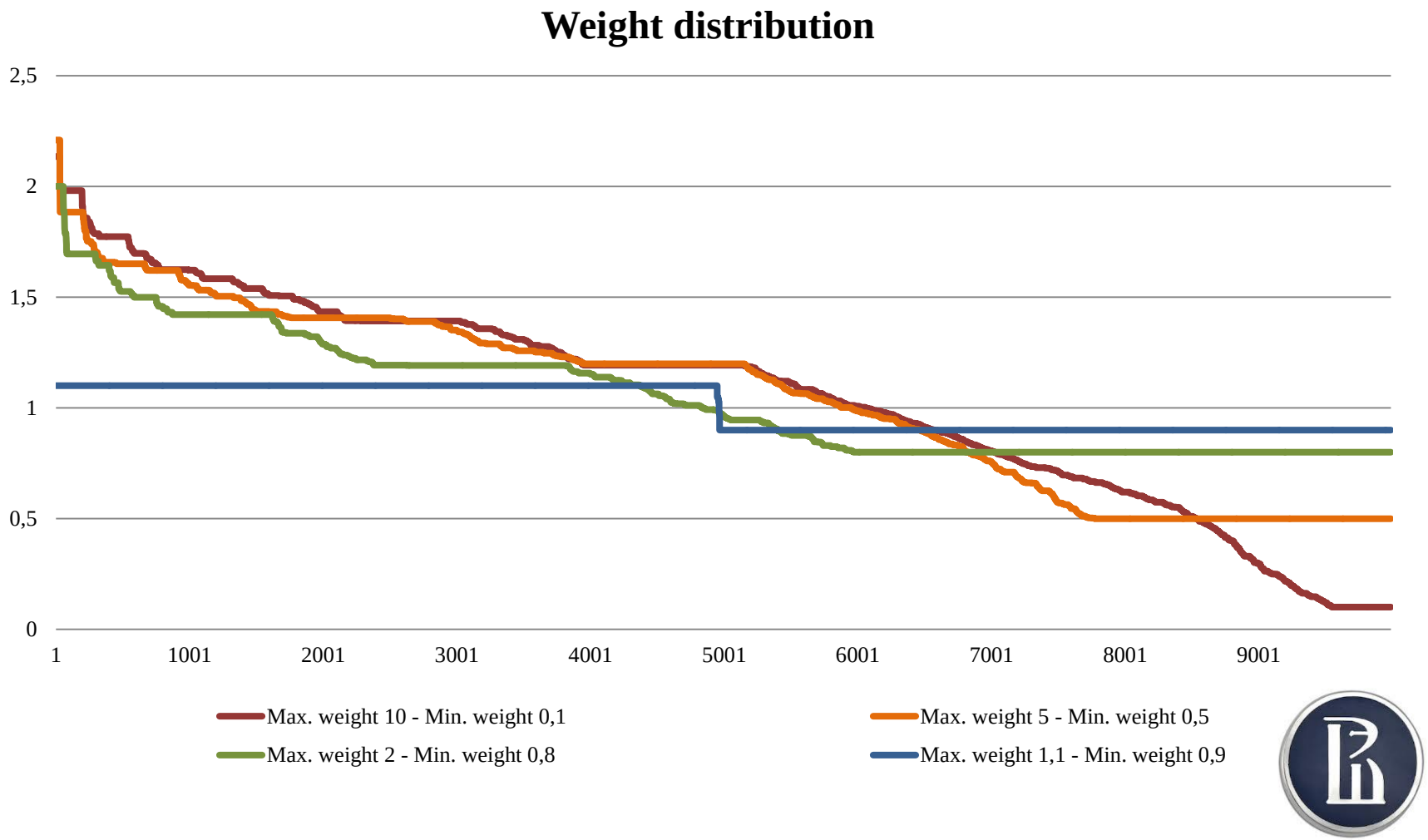
where

$$\min(v_i) \leq \frac{w_i^*}{w_i} \leq \max(v_i)$$

Constraints

$$w_i^* = \begin{cases} \overline{w}, & \text{if } \dfrac{w_i \sum_{j=1}^{m} a_{ij} v_j}{\sum_{j=1}^{m} a_{ij}} > \overline{w} \\ \dfrac{w_i \sum_{j=1}^{m} a_{ij} v_j}{\sum_{j=1}^{m} a_{ij}}, & \text{if } \dfrac{w_i \sum_{j=1}^{m} a_{ij} v_j}{\sum_{j=1}^{m} a_{ij}} \in [\underline{w}; \overline{w}] \\ \underline{w}, & \text{if } \dfrac{w_i \sum_{j=1}^{m} a_{ij} v_j}{\sum_{j=1}^{m} a_{ij}} < \underline{w} \end{cases}$$

Re-weighting results



Re-weighting results

		Official statistics		Calibrated sample data					
				$w_i \in [5; 0{,}5]$		$w_i \in [2; 0{,}8]$		$w_i \in [1{,}1; 0{,}9]$	
				%%	Diff.	%%	Diff.	%%	Diff.
Population		142 865 000	100%	100%	0,0%	100%	0,0%	100%	0,0%
including									
	Men	66 162 460	46,3%	46,3%	0,0%	46,3%	0,0%	45,5%	1,8%
	Women	76 702 540	53,7%	53,7%	0,0%	53,7%	0,0%	54,5%	-1,5%
Children (under 18)		26 251 699	18,4%	18,4%	0,0%	18,4%	0,0%	19,3%	-5,0%
Pensioners		41 819 000	29,3%	29,3%	0,0%	29,3%	0,0%	27,9%	4,8%

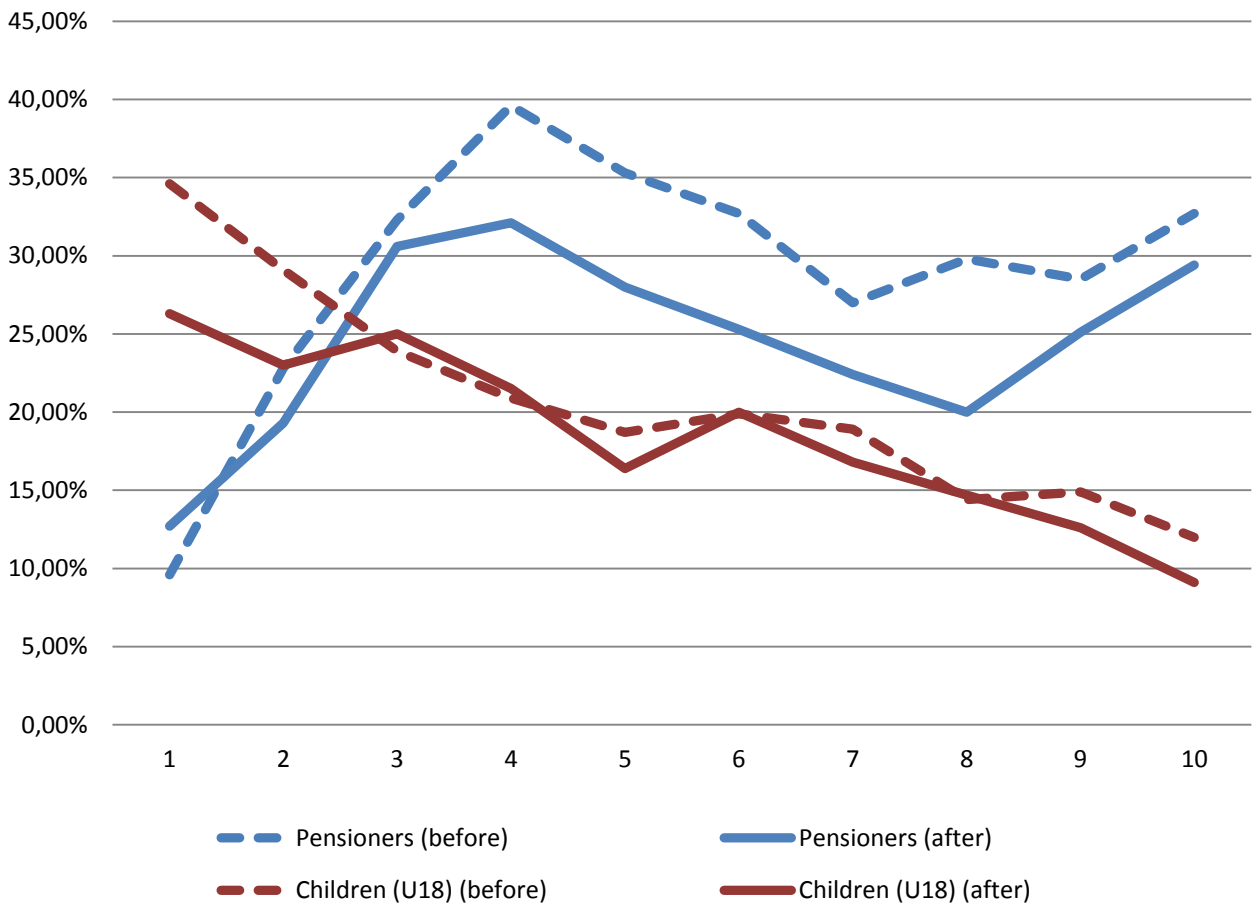
Re-weighting results

		Official statistics		Calibrated sample data					
				$w_i \in [5; 0{,}5]$		$w_i \in [2; 0{,}8]$		$w_i \in [1{,}1; 0{,}9]$	
				%%	Diff.	%%	Diff.	%%	Diff.
Employed		70 856 000	49,6%	49,6%	0,0%	49,6%	0,0%	51,8%	-4,4%
including									
	Male employed	36 032 000	25,2%	25,2%	0,0%	25,2%	0,0%	25,7%	-1,8%
	Female employed	34 824 000	24,4%	24,4%	0,0%	24,4%	0,0%	26,1%	-7,0%

Re-weighting results

		Official statistics		Calibrated sample data					
				$w_i \in [5; 0{,}5]$		$w_i \in [2; 0{,}8]$		$w_i \in [1{,}1; 0{,}9]$	
				%%	Diff.	%%	Diff.	%%	Diff.
Employed		70 856 000	49,6%	49,6%	0,0%	49,6%	0,0%	51,8%	-4,4%
including working in									
	Education	5 364 354	3,8%	3,8%	0,0%	3,8%	0,0%	4,2%	-11,3%
	Healthcare	4 455 331	3,1%	3,1%	0,0%	3,2%	-2,4%	3,7%	-17,7%

Re-weighting results



Re-weighting results

Re-weighting results

	Initial data	Calibrated data
P10/P90	16,10	14,13
Gini coefficient	0,373	0,327
Relative Poverty	14,86%	14,91%
Relative Poverty (Children U18)	24,43%	20,17%

Thank you for your attention!