ECONOMICS EDUCATION AND RESEARCH CONSORTIUM RUSSIA

РОССИЙСКАЯ ПРОГРАММА ЭКОНОМИЧЕСКИХ ИССЛЕДОВАНИЙ

S. A. Aivazian, S. O. Kolenikov

POVERTY AND EXPENDITURE DIFFERENTIATION OF RUSSIAN POPULATION

Final report, August 2000

Contents

1 ABSTRACT

The problem of poverty and inequality measurement in contemporary Russian society is considered in the framework of the general problem of social tension reduction via the efficient organization of the social assistance system. We argue that features specific to Russian transition stipulate poverty indicators (e.g. Foster-Greer-Thorbecke family) to be calculated on the basis of *expenditure* rather than income as it is usually done. These features are also accustomed for in the proposed econometric model of per capita expenditure distribution. The model includes special methods to calibrate, or to adjust, the distributions obtained from the official budget surveys' statistics. The results of the empirical approbation of the technique are reported which use the RLMS (Rounds 5–8) statistical data as well as budget surveys of Komi Republic, Volgograd and Omsk oblasts.

2 MOTIVATION AND STATEMENT OF THE PROBLEM

Various measures of poverty and expenditure inequality act as the key indicators of the quality of social policy and are used, in particular, to target social assistance, with the distant aim to reduce the social tension in the society.

The indicators and estimation procedures used nowadays by Russian statistical authorities ([1]–[3]), as well as those proposed by other researchers ([4]–[6]), are based on the household budget survey data and suffer from certain drawbacks, even after correction for the macroeconomic balance of income and consumption[1)] and/or equivalence scales.

We see the following reasons to explain those distortions:

(i) The specific features of Russian transition economy suggest that **expenditure** rather than *income* is to be used for the purposes of poverty and inequality evaluation as well as for the dichotomy of the households into poor or non-poor. We would like to note that if expenditure is used,

a) the problem of wage arrears in a household is resolved;

b) intentionally or non-intentionally hidden income, including income from shadow economy, is accounted for;

c) the concept of household welfare is appropriately generalized to include land (subsidiary plot) and property (real estate, private transportation means, jewelry, etc.) the household possess.

(ii) The two-parameter lognormal income distribution model used by the statistical authorities (State Committee in Statistics, or Goskomstat) for modeling regional and Russian income distribution is inadequate. The main distortions of the model fall to the tails of the distributions, while, evidently, the main contribution to inequality and poverty indicators are due to the tails of the distribution.

(iii) The calibration of the lognormal model used by statistical authorities does not eliminate the sample bias. (The calibration is to adjust sample weights so that the social and demographic structure of the sample complies with that of the population. Also, the level of average household per capita income is aligned with the one obtained from macroeconomic income and expenditure balance [3]. The (lognormal) shape and the parameters of the distribution (in particular, the mode) are assumed to be retained under the transformation which is also questionable.)

(iv) Distribution approximation and weighting (calibration) techniques proposed by other researchers (e.g. [4], [5]) also tend to lead to substantial distortions. They do not allow for estimation of neither the share in the population nor the structure of the *unobserved* range of "rich" and "ultra rich" households as weighting only re-weights the observed households, but *does not generate observations from the latent part* of distribution.

(v) Head-count ratio, or the proportion of households with per capita expenditure below subsistence level, is usually used as an appropriate poverty measure despite what the goal of

---

[1)] Some estimations (e.g., [3], [6], [16]) show that the ratio of the average income in the top quintile to mean income in the bottom quintile is biased downwards by the factor of at least 2, while the proportion of households with per capita income below subsistence level, as obtained by methods described in [1]–[6] and [9], might differ by a factor of 1.5–2. A similar conjecture was obtained in this study, as well. See below section 4.5.

the analysis is ([7]–[9]). However, the choice of poverty indicator (or criteria to classify a household as poor) is to be determined by the goal of economic analysis, i.e., by the application. In particular, Foster-Greer-Thorbecke family of indices is known to be better compliant with the targeted assistance goals.

(vi) The problem of the optimal, in terms of the specific poverty indicator (see (v) above), allocation of resources addressed to targeted assistance has never been stated, let alone solved, in Russian economic theory and policy.

2.1 *The aim and the main tasks of the project*

The goals of the project are determined by the desire of the project participants to overcome the aforementioned drawbacks (i)–(vi). In particular, we are aiming at: the development of the methodology for econometric analysis of per capita expenditure distribution based on Russian budget survey data; construction of the main characteristics of poverty and welfare inequality of Russian population and their statistical assessment; and formulation and solution to the problem of optimal allocation of the limited amount of resource dedicated to targeted assistance to the poor.

In general, the research problem statements are necessitated by the above goals. In their aggregated formulation, the two main problems are as follows.

The main task is to obtain from theory and approve empirically an interpretable econometric model of the regional/national per capita expenditure distribution. This also implies the development of identification methodology based on the sample budget surveys and macroeconomic balance of income and expenditure.

The solution to this task is to be linked to the specific features of Russian economy and the way these specificities are reflected in household behavior. In particular, the intentional refusal of the household to participate in the survey (unit non-response, or truncation) plays an important role in the analysis of expenditure distribution, as the truncation that leads to the deterioration of the sample representativeness.

In the analysis of the survey results, the heterogeneity of the households in terms of their probability to refuse to participate in the survey should be accounted for. We find it reasonable

to assume that there are households escaping surveys with probability of one. It is likely that the rich households (i.e. those with per capita expenditure above a certain value) would be the members of this category, as high income is quite often associated with legal or semi-legal economic activities.

Apparently, any econometric model of income / expenditure distribution that would aim at elimination (or at least attenuation) of the data quality problems must be based on explicitly formulated (and, if possible, substantiated and proved with the statistics) additional working hypotheses and assumptions. In this study, these (verifiable) hypotheses are widely defined as follows:

- The first hypothesis H_1 concerns the shape of the distribution function;
- The second hypothesis H_2 concerns the probability of the unit non-response, i.e. the refusal of a household to participate in the budget survey, conditional on its welfare (expenditure), as well as some other social and economic characteristics;

We also formulate, without any proofs, the following additional assumptions:

- The model assumption H_3 states that the coefficient of variation of per capita expenditures (or the variance of log expenditure) is constant across all strata;
- The model assumption H_4 deals with the shape of the distribution of household per capita expenditure within the *unobserved* range of expenditures (right distribution tail, the richest population strata).

The hypothesis H_1 is deep-seated in the salient transition features of Russia (see below section 4.1). Statistical testing and further exploitation of this hypothesis is essential for the formulation of the interpretable model of per capita expenditure. Statistical testing and further use of the hypothesis H_2 is aimed at elimination of the unit non-response bias. The assumptions H_3 and H_4 are purely technical and mainly deal with mitigation of the truncation of the super-rich stratum.

The detailed description and foundation for all these hypotheses will be given in the main part of the report.

Task 2 is an auxiliary one and serves as an example of application of the proposed methodology in the field work. We shall aim to consider a broad class of poverty indices based on the per capita expenditure distribution, and formulate the problem of optimal allocation of a limited resource S devoted to targeted social assistance to the poor, based on the objective function from this class.

The following family of poverty indices would be considered:

$$I\big(w(x), f(x)\big) = \int_0^{z_0} w(x)\ f(x)\ dx\ , \tag{1}$$

where $f(x)$ is the per capita expenditure density function, z_0, poverty line, and weighting function $w(x)$ is supposed to be differentiable, decreasing and convex at $[0, z_0)$ (the latter property is due to the transfer principle). Apparently, the family (1) such popular measures as Foster-Greer-Thorbecke family of indices (with $w(x) = \left(\frac{z_0 - x}{z_0}\right)^{\alpha}$; referred to as FGT($\alpha$) further in this work), Dalton class indicators, and poverty-line-discontinuous measures [13]–[15].

Let S is the amount given for targeted assistance. Let S is less than the poverty gap. Denote the rule of allocation of this resource among population with per capita expenditure $x < z_0$ (e.g. distribution density) as $\varphi(x|S)$, and the population per capita expenditure distribution density observed *after* the realization of social assistance according to $\varphi(x|S)$, as $\tilde{f}(x|\varphi, S)$. The ex post indicator value would thus be:

$$I\big(w(x), \tilde{f}(x|\varphi; S)\big) = \int_0^{z_0} w(x)\ \tilde{f}(x|\varphi; S)\ dx\ . \tag{1′}$$

Task 2 is then reduced to the identification of the $\varphi_0(x|S)$ such that (1') achieves its minimum, given $w(x)$ and S:

$$\varphi_0(x|S) = \arg\min_{\varphi} \int_0^{z_0} w(x)\ \tilde{f}(x|\varphi; S)\ dx\ . \tag{2}$$

It is worth noting that the problem 2 is considered within the framework of a specific project of the long-term poverty alleviation [8]–[9]. The implications of this context are twofold. First, the argument for relatively high income mobility [40] is not fully applicable to this population category. Second, the main instruments of the long-term poverty alleviation are the direct transfers to the needy households rather than creation of incentives schemes (which is the way relevant for temporarily poor, e.g. unemployed).

Problem 3 is also auxiliary. By using the solution to the main problem (i.e. the estimates of the per capita expenditure distribution for Russia and the three regions), we shall calculate the estimates of inequality indices, such as Gini index and the funds ratio (the ratio of the total expenditure in the top decile to that in the bottom decile); compare the figures with the officially reported ones (by Goskomstat); and find the analogies among other countries.

Apparently, the truncation of the super-rich cannot affect the poverty indices that form the framework for the 2nd problem. In fact, the poverty analysis focuses on the left tail of the expenditure distribution, while the use of the model assumption H_4 is aimed to fit the right tail of the distribution.

The account of the super-rich stratum, however, does affect inequality indices[1]. This correction is viewed as important one by us, as inequality and polarization indices characterize the social tension in the population, i.e. the nuclei of the potential conflict between population groups.

3 LITERATURE OVERVIEW

Let us first discuss the sources where problems close to our main task (see above) were posed.

The model of per capita expenditure distribution developed in this project is supposed to develop and modify the basic model of population per capita *income* distribution pioneered in

[1] The calculations in [16] show that after the similar calibration of 1995–1996 data, the Gini index rises from 0.376 to 0.531, while the funds ratio, from 12.9 to 22.8.

[16]. The *modification* includes i) introduction and statistical estimation of budget survey unit non-response probability (see H_2 above); ii) replacement of income by *expenditure* in the log-normal mixture model; and iii) calibration of the existent observations followed by Monte Carlo generation (parametric bootstrap) of additional data. The latter are unobserved in the sample and resampled on the basis of the known macroeconomic balance of household expenditure as supplemented by hypotheses $H_2 – H_4$.

The papers [4]–[6], [16] contain arguments which prove the validity of our critique (i)~(iv) in the introduction. Velikanova *et al.* in [2] describe an approach which is also based on the mixture of lognormal distributions, but this source neither provide econometric tools to analyze this mixture nor proposes any ways to reconstruct the unobserved data. The approach by Ershov and Mayer in [5] is based on polynomial density approximation and seems to be too formal. It does not allow for establishment of an interpretable model of the phenomenon studied and does not account for the latent expenditure range.

The main drawback of the approach by Suvorov and Ulyanova in [6] is inadequacy of the basic assumption on lognormality of income distribution though the authors do study a three parameter model, as opposed to the biparametric Goskomstat model. Nevertheless, the authors a) analyze income, not expenditure; b) do not provide any convincing arguments in favor of the basic assumption on the adequacy of the estimate of the *modal* income out of the Goskomstat budget survey sample (which is considered substantially biased even by Goskomstat specialists, let alone independent experts); c) propose a formal approximation technique of unknown parameters fitting. While the *economic analysis* of stylized facts on income redistribution processes in Russia during transition does clarify the mechanism of formation of the right distribution tail (the one that remains unobserved in the Goskomstat budget surveys, the drawbacks of the approach can be quite heavily criticized.

Special attention needs to be paid to the work of Shevyakov and Kiruta [4], especially to the differences of the approach of theirs from the one proposed in our project. Their work is currently the most serious attempt to describe *realistically* the regional per capita income distribution with the information contained in the Goskomstat budget survey data and macroeconomic

"Population Income and Expenditure Balance" (a special balance of monetary flows on both regional and national levels routinely calculated by Goskomstat). The attempt is based on the non-parametric approach to density estimation and a technique to eliminate the Goskomstat sample bias. It also describes the procedure to aggregate the regional data corrected for regional deflators and equivalence scales. To our view, the main drawbacks of Shevyakov & Kiruta approach are as follows:

a) The weighting (calibration) technique proposed in [4], in fact, ignores the population beyond the maximum income observed. The right tail of the distribution remains not accounted for, and censuring problem is not addressed. In our model, the tail is recovered by using the set of hypotheses H_1–H_4.

b) The immediate consequence of the previous critique point is a principally erroneous inference that "the excessive economic inequality is in whole caused by the excessive poverty". Given that the authors ignore the right tail, there cannot be any other result.

c) A seemingly attractive "non-parametricity" of the approach has, in fact, two serious drawbacks. First, the estimate of the per capita income distribution obtained in this way is a *purely formal approximation* of the unknown distribution analyzed and *cannot be interpreted in understandable terms*. Second, the model is not at all suitable for prediction purposes.

d) To estimate the poverty rate, wealth inequality and other welfare indicators, expenditure is more appealing in Russian situation than income, as long as it removes inconsistencies related to wage arrears, hidden income, etc.

Let us now focus on the works related to the Task 2. First of all, worth mentioning are the World Bank project [8] and pilot programs [9]. They do accomplish a rightful attempt to assess poverty according to re-estimation of realistic household per capita income (termed 'potential consumption expenditures' in [8]). Both approaches, however, still suffer from significant drawbacks analyzed by Aivazian in [17]. Besides, the only poverty index used is again the head-

count ratio (i.e. (1) with $w(x) \equiv 1$), and the problem of optimal allocation of social assistance is not stated (i.e. problem (2) is not solved for).

A comprehensive overview of poverty indicators is given in [18]. This work discusses, in particular, a special case of criterion (1), i.e., Foster-Greer-Thorbecke set of indices, and reports the sample statistics of quarterly budget surveys as of 1996. Still, the index calculation relies on income distribution and, which is more important, is not related to targeted assistance optimization.

Thus, to our knowledge, neither economic theory nor practice in Russia states or solves Task 2. Nevertheless, various aspects of this problem are addressed in the Western literature though most authors still rely on income rather than expenditure distributions ([15], [19]–[24]). In particular, [23] proves that under FGT indices with

$$w(x) = \left(\frac{z_0 - x}{z_0} \right)^{\alpha}, \quad 0 \le x < z_0, \quad \alpha > 1, \tag{3}$$

the optimal solution to (2) is the pure strategy of transferring the poorest people enough money to raise their income to the threshold $\bar{z}_0 < z_0$, where $\bar{z}_0$ is found from the government budget

$$N \left(\int_0^{\bar{z}_0} f(x)\, dx \right) \left(\int_0^{\bar{z}_0} (\bar{z}_0 - x)\, f(x)\, dx \right) = S \tag{4}$$

where N is the total population. This strategy is referred to as “allocation of p-type” in [15] and [23] and implies that each person with income below $x < \bar{z}_0$ is to receive a subsidy $\bar{z}_0 - x$. An alternative option is the allocation of mixed-type when a portion S_1 of S is used to raise the incomes of the poorest up to $\bar{z}_0$. With this strategy, S_1 substitutes S in the RHS of (4), and the rest of S is used to raise the incomes of the richest among the poor to z_0. It is proved in [23] that the mixed strategy can only be optimal if $w(z_0) = \delta > 0$, i.e. if the underlying poverty index is discontinuous. These type of indices are referred to as ‘poverty-line-discontinuous, or PLD, measures’ in [15].

As for the analysis of the third problem, we would like to mention Esteban-Ray polarization index proposed in [25]. This index crucially depends on the knowledge of the tail strata of

the distribution and is effectively used along with Gini coefficient (which is a special case of Esteban-Ray index with the value of a certain self-identification parameter being zero) in empirical works as a factor of crime [26].

4 MODEL SPECIFICATION AND PRELIMINARY ESTIMATION RESULTS

4.1 Discussion of the basic working research hypotheses and model assumptions

The solution to the above stated Task 1 is based on the theoretical inference and/or empirical testing of a number of working hypotheses.

- **Hypothesis H_1** states that the distribution of Russian households by per capita expenditures can be adequately described by a *mixture of lognormal distributions.* This hypothesis can be verified by a fit criteria. An example for 1996 data is [16].

  Theoretical reasoning for this hypothesis is as follows.

  a) Per capita expenditure ξ distribution within a homogeneous strata follows lognormal distribution with parameters $a = \mathbf{E}(\ln \xi(a))$ and $\sigma^2(a) = \mathbf{D}(\ln \xi(a))$. Here, homogeneity refers to similar income sources, geographical, social, demographic, and professional characteristics of its representatives.

  b) If society as a whole can be represented by a continuous (in terms of the average log expenditures a) spectrum of such strata, then under a certain though natural shape of the mixing function $q(a)$, the population distribution by per capita expenditures is reproduced to be lognormal.

  c) If continuity of the spectrum is violated (i.e. some strata are eliminated, or crowded out), or $q(a)$ is not monotonically decreasing as its argument a increases from the global average a_0, then the population lognormality holds no longer, and the distribution is transformed into a discrete-type mixture.

  Let us now discuss each of the postulates.

The first statement is quite widespread in income distribution studies and results from multiplicative shocks to expenditure (income, wages) within the strata. The data generating mechanism is described in [27] and applied to wages of workers in Soviet Union.

The second postulate follows from the fact that if the within-strata-average log expenditures $a = \mathbf{E}(\ln \xi)$ are distributed normally with parameters $(a_0; \Delta^2)$ (i.e. if $q(a)$ is normal), then the resulting distribution of expenditure logarithms $(\zeta = \ln \xi)$

$$\varphi(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma(a)} e^{-\frac{(z-a)^2}{2\sigma^2(a)}} q(a)\, da$$

is a composition of normal distributions and thus normal itself. If $\sigma^2(a) = \sigma^2 = const$, then the parameters of the resulting distribution are $a_0 = \mathbf{E}(\ln \xi)$ and $\sigma_0^2 = \sigma^2 + \Delta^2$. This fact is mentioned and proved in [25].

The third statement is apparent in a degenerate situation when the number of points where the mixing function $q(a)$ is different from zero is finite: $a_1, a_2, \ldots, a_k$. The realistic distribution of expenditures in Russian economy is, of course, more complicated. But it nevertheless is characterized by a significant transformation of the mixing function $q(a)$. The transition period do not abolish the a) and b) postulates though affected the shape of $q(a)$.

- **Hypothesis H_2** states that the probability of the household to refuse to participate in the official budget survey is a function of its social, economic, and geographical characteristics. This hypothesis can also be verified against the data such as RLMS ([28]) and some additional information from Goskomstat. This hypothesis was prompted by the Head of Living Standards Department of Goskomstat E. B. Frolova and was apparently implied by the field experience.
- **Assumption H_3** states that the *coefficient of variation* of the household per capita expenditures is constant across the social strata, i.e., is independent of the strata number. This hypothesis can also be verified by criteria of variance homogeneity ([16]). As long as income and expenditure $\xi(j)$ of population of *j*-th homogeneous strata are distributed lognormally

with the parameters $a(j) = \mathbf{E}(\ln \xi(j))$ and $\sigma^2(j) = \mathbf{D}(\ln \xi(j))$ (e.g. [29]), the hypothesis H_3 is equivalent to:

H_3': Var[ln ξ (j)]= σ^2=const

The equivalence of H_3 and H_3' follows from the relation between the moments of the lognormal distribution:

$$\frac{[Var(\xi(j))]^{\frac{1}{2}}}{\mathbf{E}\xi(j)} = \left(e^{\sigma^2} - 1\right)^{\frac{1}{2}}$$

- **Assumption H_4** states that the population per capita expenditures x in the latent range of $x > \max_{1 \le i \le n}\{x_i\}$, where x_i is per capita expenditures in the i-th household surveyed, and n, total number of households, can be approximated by three parameter lognormal distribution with a shift parameter $x_{(n)} = \max_{1 \le i \le n}\{x_i\}$ and variance of logarithms $Var(\xi(k) = \sigma^2$ where σ^2 is independent of strata and estimated from the observed strata (see hypothesis H_3 above).

  Strictly speaking, this model assumption is not a statistical hypothesis, as it cannot be directly verified against the data available with any statistical criteria since the necessary data cannot be observed. It can be established *ex ante* by some economic argument, and *ex post*, by matching the levels of the observed characteristics with the model output. To support this hypothesis, let us mention some stylized facts related to Russian transition.

One of the real conseqences of the USSR and it economic system disintegration is the formation of "new Russian" group from the communist, state bureaucracy and managerial elites. By using the privatization for their own benefit, they managed to get access to the rent flows in the form of elements of national wealth which could (and was) sold on the domestic and world markets. Some calculations (e.g. [6]) show that market intervention of Russian national wealth per annum amounting to 0.2–0.3% is equivalent to the increase of gross population income by 10–20%. Evidently, the major part of this income is distributed into this novo riche group of population, which can be classified as a separate stratum as long as its representatives are ho-

mogeneous by their social status and power. It is this stratum which is referred to in the assumption H_4.

Usually, the right tail of income / expenditure distribution beyond (high enough) x_0 is approximated by Pareto distribution. This assumption, however, is only valid if the density function decreases monotonically for all $x \geq x_0$ (as it is the case in a well-functioning economy). In our case, we cannot rule out a local maximum in the unobserved richest strata to the right of x_0.

By using the hypotheses and model assumptions H_1-H_4, a non-formal (i.e., an interpretable) model of Russian population per capita household expenditure distribution can be developed. Further in the project, the statistical methodology will be described to estimate poverty and inequality indicators from the budget survey data, plus some additional macroeconomic characteristics of social and demographic family structure and population expenditures.

4.2 *The main variables and information sources*

1) Gross (rescaled to monthly window) per capita expenditures ξ of a randomly sampled (surveyed) household x_i.

Following Goskomstat methodology from [7], we shall define (with the time quantum of a quarter) gross pecuniary expenditures of a household as the sum of:

- $\xi^{(1)}$ — *quarterly consumption expenditures*, which is the sum of food products expenditures, alcohol, non-food private consumption goods and private services;
- $\xi^{(2)}$ — *interim consumption expenditures* (household expenditures for subsidiary land plot);
- $\xi^{(3)}$ — *the quarterly average of the net household capital accumulation* (acquisition of land and property, jewelry, construction and dwelling maintenance expenditures);
- $\xi^{(4)}$ — *the quarterly total of taxes paid and other obligatory payments* (including alimony, debt, club and public payments);
- $\xi^{(5)}$ — *cash in hands and net savings increase* (including currency and stock accumulation, bank deposits);
- $\xi^{(6)}$ — *estimate of monetary equivalent of the household produced products*.

All in all,

$$\xi = \frac{1}{3m_\xi}\sum_{l=1}^{6}\xi^{(l)},$$

where $\xi^{(l)}$, l=1,2,…,6 are as defined above, and m_ξ is the effective number of the consumers in the households, and the factor of 3 is introduced to reduce the quarterly data, as in Goskomstat budget surveys, to the monthly data that most readers are likely to be accustomed to. The observed values of x_i, $x_i^{(1)}$, $x_i^{(2)}$, ..., $x_i^{(6)}$ of the random variables $\xi, \xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(6)}$ are the results of the survey in the i-th household.

In fact, the definition of the scale factor m, known as the equivalence scale, is a discussable issue. Russian statistical authorities use an implicit equivalence scale with the equivalence factors of 0.9 and 0.6 for children and pensioners, respectively[2], on the basis of nutritional requirements. The OECD equivalence scale is based on the economy of scale argument rather than nutritional scheme. According to this scale, the first adult is given a factor of 1, all other adults, 0.7, and each child, 0.5. One of the most comprehensive discussions of the various equivalence scales can be found in [45] where several dozens of equivalence scales are analyzed and reduced to a simple parametric scheme.

It need not be apparent, but a number of theoretical assumptions concerning household preferences and the shape of the equivalence scale need to be made to construct an easy-to-deal-with equivalence scale (see e.g. [46] and [47]). It is not at all clear, however, whether these assumptions are actually satisfied, and it is not even clear how these assumptions might be tested. In this light, we view equivalence scales as a technical correction that can be incorporated with relatively low computational cost. We are not aware, however, of any convincing argument in favor of any equivalence scale that it should be the equivalence scale for Russia. Thus, we stick to our basic assumption that per capita calculations are good enough for expenditure analysis in this country. The preliminary analysis convinced us in robustness of our findings with respect to the use of equivalence scales.

2) Regional/national average per capita expenditures μ_{macro} defined from macroeconomic characteristics, namely, quarterly Goskomstat “The Population Income and Expenditure

[2] Rather than deflating the observed household characteristics by these factors, Goskomstat calculates the poverty lines separately for each population group using the above factors.

Balances" [30]. μ_{macro} has the same structure as ξ but is defined from regional trade, tax, bank and security market statistics rather than surveys.

3) The proportion of households $p(x)$ with per capita expenditure level x who refused to participate in the survey in the given period. The sources of information are supposed to be Goskomstat and RLMS.

4) Social and demographic composition of the region (regional averages on household size, number/proportion of children, retired, etc.).

Let us now describe in some detail the RLMS and Goskomstat budget survey data that comprise the information base of our research.

(1) RLMS data, Rounds V–VIII [28]. The RLMS questionnaire contains expenditures for a large number of goods and services. This data can be aggregated to large groups of goods and services, and to total expenditures.

Expenditure data include a wide range of categories, though the time spans in each category might be different. The expenditures for food (~60 items) are based on weekly reports; fuel, services (with a breakdown to about 10 items), rent, club payments, insurance premia, savings and credits have one month window; non-food consumer goods and durables expenditures are found on the quarterly basis. RLMS also traces home production on the annual basis, as well as intermediate expenditures for the subsistence plot. All those data are rescaled towards monthly basis and published in **`r#heexpd`** RLMS data files. Currently, we have used variables **`totexpr*`** from these data files. These data have been verified by RLMS staff and include the appropriately scaled data.

Of course, the quality of the results cannot be higher than the quality of the data, and we must make this reservation before proceeding any further. For instance, it can be argued that the family welfare (measured here as consumption expenditures) should also include the depreciation of durables, property and vehicles that have been inherited from earlier periods, might as well from Soviet times. This correction, however, remains, to our knowledge, a purely theoretical argument that has never been implemented in the applied work.

(2) Household budget survey data as of Q3 1998 on three regions of Russia, namely, Komi epublic, Volgograd and Omsk oblasts, with a supplementary questionnaire [11]. According to Goskomstat methodology [7], the sample is constructed to be representative of the household types, except the collective households (e.g. hospitals, military units, etc.), on the basis of 1994 microcensus. During the quarterly budget survey, a household fills in twice in the quarter a two-week daily log of expenditures, two bi-weekly logs, and is exposed to a intermediate monthly survey. From this primary data, Goskomstat infers the following aggregate indicators: pecuniary expenditures ("**denras**" variable in the Goskomstat survey datasets; the sum of actual expenditures made by household members in the period of account; includes consumption and non-consumption expenditures); consumption expenditures ("**potras**" variable; the proportion of pecuniary expenditures directed to acquisition of consumption goods and services); final household consumption expenditures ("**konpot**" variable; consumption expenditures sans food products transferred outside the household, plus the in-kind household income, i.e. the sum of non-cash and natural intakes of food products and subsidies); household disposable resources ("**rasres**" variable; the sum of pecuniary resources, "**denres**" variable, i.e., pecuniary expenditures and nominal savings by the end of the period, and natural intakes, "**natdox**" variable). The budget surveys referred to were supplemented with the questionnaire on quality of life [11].

4.3 *Model description and parameter interpretation*

Denote ξ (ths. rub.) the yearly average expenditure of a randomly selected representative of Russian population, and ξ_j (ths. rub.), per capita expenditures of the representative of j th homogeneous stratum. According to hypotheses H_1 and H_4, the distribution density of the random variable ξ is described by the model of lognormal mixture:

$$f(x|\Theta) = \sum_{j=1}^{k} q_j \frac{1}{\sqrt{2\pi}\,\sigma_j x} e^{-\frac{(\ln x - a_j)^2}{2\sigma_j^2}} + q_{k+1} \frac{1}{\sqrt{2\pi}\,\sigma_{k+1}\cdot(x - x_0)} e^{-\frac{(\ln(x-x_0) - a_{k+1})^2}{2\sigma_{k+1}{}^2}}, \qquad (5)$$

where $\Theta = (k; q_1,\ldots,q_{k+1};\ a_1,\ldots,a_{k+1};\ x_0;\ \sigma_1^2,\ldots,\sigma_{k+1}^2)$ are the model parameters interpreted as follows:

$k+1$ is the number of mixture components, or homogeneous strata;

q_j $(j = 1,2,\ldots,k+1)$ is the *ex ante* probability of the *j*-th mixture component, or the share of the respective stratum in the population;

x_0 is the threshold separating observed expenditures $(x \le x_0)$ from the unobserved ones $(x > x_0)$;

$a_j = \mathbf{E}(\ln \xi_j)$ $(j = 1,2,\ldots,k+1)$ are the model averages of logarithms within *j*-th stratum;

$\sigma_j^2 = \mathbf{D}(\ln \xi_j)$ $(j = 1,2,\ldots,k+1)$ are the respective expenditure logarithms variance.

We assume that per capita expenditures of the population of the richest $k+1$-th stratum exceed the threshold x_0, and that they always refuse to participate in surveys. The rest households are available to statistical investigation, thought also can escape from the survey with probability $p(x)$ monotonically increasing with x (see hypothesis H_2 above).

Econometric analysis of the model (5) implies estimation of the parameter vector Θ by survey data, as well as some social and demographic population characteristics necessary to derive individual distribution from the household one.

4.4 Econometric analysis methodology

4.4.1. *Estimation of the dependence p(x) of refusal probability its social and economic characteristics*

The following variables are considered as covariates of the refusal probability p:

$z^{(1)} = \ln \xi$ is the logarithm (in base e) of the total per capita household expenditure;

$z^{(2)}$ is the settlement type, with categories metropolitan areas, urban and rural areas, settlement of city type (PGT, “poselok gorodskogo tipa”;

$z^{(3)}$ is the education of the primary income earner (below secondary, secondary, vocational school, technical school, higher).

In terms of these variables, the dependence of p on $Z = (1, z^{(1)}, z^{(2)}, z^{(3)})^{\mathrm{T}}$ is supposed to follow the logistic model:

$$p(Z) = P\{\eta_i = 0 \mid Z\} = \frac{1}{1 + e^{\beta^{\mathrm{T}} Z}}, \tag{6}$$

where

$$\eta_i = \begin{cases} 0, \, i\text{ - th household refused from participation in the survey;} \\ 1, \, i\text{ - th household participated in the survey,} \end{cases}$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$ is the vector of the coefficients to be estimated. Geographical and education factors enter the model as dummy variables, while expenditure elasticity is assumed to be the same for all population categories. Thus, the model (6) gives a set of 4*5=20 models (one for each population category) describing the dependence of refusal probability p on per capita expenditure:

$$p_{kl}(z) = p(z \mid z_k^{(2)}, z_l^{(3)}) = P\{\eta_i = 0 \mid z^{(1)} = z, z^{(2)} = z_k^{(2)}, z^{(3)} = z_l^{(3)}\}, \tag{6'}$$

$$k = 1, 2, 3, 4; \qquad l = 1, 2, 3, 4, 5$$

In fact, a wider set of regressors was used initially in the analysis that included also social and demographic structure of the household and the characteristics of the household head, beside the log per capita expenditure, the settlement type, and household head education. The subsequent analysis showed statistical insignificance of some characteristics, and the selection of the logistic regression model led us to the result reported above.

The results of the model estimation (i.e., the estimates of β) by using RLMS data (Rounds V–VIII) are given in the Appendix 3. These results assert the monotonic dependence of the refusal probability p upon the level of expenditure. For comparison, a simplified model was also estimated that only includes the (log of) expenditure $z = z^{(1)} = \ln \xi$:

$$p(z) = P\{\eta_i = 0 \mid z^{(1)} = z\} = \frac{1}{1+e^{\beta z}}. \qquad (6'')$$

4.4.2. Calibration (weighting) of the existing observations

The analysis of the models (6) and (6'') is, of course, interesting per se. In our study, however, this is only a by-product used for the calibration of the existing observations. By using the weights obtained as the inverse of the probability to participate in the survey, we re-estimate per capita expenditure regional / national distribution. When the information is sufficient (the categories k_i, l_i corresponding to $z_{k_i}^{(2)}, z_{l_i}^{(3)}$ variables are known for i-th household, as in RLMS), the "fine" weights according to (6) are used. Otherwise, if only per capita expenditure is available (as with our regional data), weights (6') are used. We would unite notations as $p(z)$ when referring to the *logs* of observed expenditure, and as $p(x)$ when referring to the *initial* observations, or *levels* (ths. rubs).

Let $f(x)$ be the density function of the per capita expenditure distribution of population of a Russian region. If n is the total size of the survey sample and x^* is a certain value of per capita expenditures, then the number $\nu(x^*)$ of observations in the Δ-neighborhood of the point x^* on the condition that no one escapes from the survey, is given by

$$\nu(x^*) \approx nf(x^*)\boldsymbol{\Delta} \ . \qquad (7)$$

The actual number of observations, however, would be adjusted for the probability of refusal $p(x)$:

$$\tilde{\nu}(x^*) \approx nf(x^*)\left(1 - p(x^*)\right)\boldsymbol{\Delta} \ . \qquad (8)$$

(7) and (8) imply that

$$\nu(x^*) = \tilde{\nu}(x^*) \cdot \frac{1}{1 - p(x^*)} \ . \qquad (9)$$

In particular, by choosing the actually observed data on per capita expenditure as the x^* and taking small enough Δ, we would have:

$$\tilde{\nu}(x_i) = 1$$
$$\nu(x_i) = \frac{1}{1 - p(x_i)}.$$

It means that if we want to estimate the underlying density $f(x)$ from the existing sample

Observed x	x_1	x_2	$\cdots$	x_n
Observation weights	$\frac{1}{n}$	$\frac{1}{n}$		$\frac{1}{n}$

(10)

then we should recalibrate, or re-weight, the sample in the following way:

Observed x	x_1	x_2	$\cdots$	x_n
Observation weights	ω_1	ω_2	$\cdots$	ω_n

(11)

where ω_i are found from

$$\omega_i = \frac{\frac{1}{1 - p(x_i)}}{\sum_{j=1}^{n}\left(\frac{1}{1 - p(x_j)}\right)}.$$

It is worth noting that ω_i increase with the refusal probability $p(x_i)$, and $\sum_{i=1}^{n}\omega_i = 1$

4.4.3. Estimation of the observed mixture components parameters

At this stage we solve the problem of estimation from the sample (11) of the mixture parameters $k, \tilde{q}, \ldots, \tilde{q}_k,\ a_1, \ldots, a_k,\ \sigma_1^2, \ldots, \sigma_k^2$ driving the distribution shape:

$$\tilde{f}(x) = \sum_{j=1}^{k}\tilde{q}_j \frac{1}{\sqrt{2\pi}\,\sigma_j x} e^{-\frac{(\ln x - a_j)^2}{2\sigma_j^2}} \qquad (12)$$

The problem is in fact reduced to that of parameter estimation of the mixture of *normal* distributions:

$$\tilde{\varphi}(y) = \sum_{j=1}^{k} \tilde{q}_j \frac{1}{\sqrt{2\pi}\,\sigma_j} e^{-\frac{(y-a)^2}{2\sigma_j^2}} \qquad (13)$$

by the sample

Observed y	y_1	y_2	$\cdots$	y_n
Observation weights	ω_1	ω_2	$\cdots$	ω_n

, $(8')$

with $y_i = \ln x_i \ \ (i = 1,2,\ldots,n)$.

The results of the estimation of the mixture model for the RLMS and regional data (2Q 1998) are given in the following section. The numerical methods used in estimation are briefly described in Appendix 3; for more detail, see [31]–[35]. The software implementations are CLASSMASTER software developed at CEMI and denormix STATA module developed by S. Kolenikov.

4.4.4. Estimation of the unobserved mixture component and distribution as a whole

Let the (relative) weight of the unobserved $\hat{k}+1$-th mixture component is $q_{\hat{k}+1}$, and the mean logarithm of per capita expenditures is $a_{\hat{k}+1}$. Then the regional average μ from the model (5) based on the parameter estimates $\hat{k};\ \hat{\tilde{q}}_1,\ldots,\hat{\tilde{q}}_{\hat{k}};\ \hat{a}_1,\ldots,\hat{a}_{\hat{k}};\ \hat{\sigma}_1^2,\ldots,\hat{\sigma}_{\hat{k}}^2$ obtained earlier is given by

$$\mu = \int_0^{\infty} x \left(\sum_{j=1}^{\hat{k}} \hat{q}_j \frac{1}{\sqrt{2\pi}\,\hat{\sigma}_j x} e^{-\frac{(\ln x - \hat{a}_j)^2}{2\hat{\sigma}_j^2}} + q_{\hat{k}+1} \frac{1}{\sqrt{2\pi}\,\sigma_{\hat{k}+1}(x - x_0)} e^{-\frac{(\ln(x-x_0) - a_{\hat{k}+1})^2}{2\sigma_{\hat{k}+1}^2}} \right) dx, \qquad (14)$$

where

$$\hat{q}_j = \hat{\tilde{q}}_j(1-q_{\hat{k}+1}), \quad j=1,2,\ldots,\hat{k}\,. \tag{15}$$

Given the properties of lognormal distribution,

$$\mu = \sum_{j=1}^{\hat{k}} \hat{q}_j \; e^{\frac{1}{2}\hat{\sigma}_j^2+\hat{a}_j} + q_{\hat{k}+1}\left(x_0 + e^{\frac{1}{2}\sigma_{\hat{k}+1}^2+a_{\hat{k}+1}}\right). \tag{14′}$$

The value of μ from (14') depends on the unknown $q_{\hat{k}+1}, a_{\hat{k}+1}$, as well as on x_0 and $\sigma_{\hat{k}+1}^2$. By construction, x_0 is taken to be the maximum of the observed expenditure:

$$x_0 = \max_{1\le i\le n}\{x_i\} \tag{16}$$

Under H_3' (see section 4.1 above), the overall estimate $\hat{\sigma}^2$ of the variance of logarithms is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{\hat{k}} n_j \hat{\sigma}_j^2\,, \tag{17}$$

and then $\sigma_{\hat{k}+1}^2$ is taken to be equal $\hat{\sigma}^2$.

We can then graph the level line in the plane $(q_{\hat{k}+1}, a_{\hat{k}+1})$:

$$\mu(q_{\hat{k}+1}, a_{\hat{k}+1}) = \mu^{\text{macro}}\,, \tag{18}$$

where the model value $\mu(q_{\hat{k}+1}, a_{\hat{k}+1})$ is calculated by (14') with $x_0 = \max_{1\le i\le n}\{x_i\}$ and $\sigma_{\hat{k}+1}^2 = \hat{\sigma}^2$, while μ^{macro} is obtained from the macroeconomic Balance of Population Incomes and Expenditures for the relevant region and time point.

The final selection of the point $(\hat{q}_{\hat{k}+1}, \hat{a}_{k+1})$ on the line (18) requires some additional conditions, assumptions, or expert information.

When constructing the line (18), it is worth considering that:

(i) Apparently,

$$q_{\hat{k}+1} << \min_{1\le j\le \hat{k}}\{q_j\}$$

where the sign << means “much less”, i.e. that q_{k+1} in about an order of magnitude less than $\min\limits_{1\le j\le \hat{k}} \{q_j\}$.

(ii) The level line (18) may be represented by a table with the values of $q_{\hat{k}+1}$ as input and $a_{\hat{k}+1}$ from (14’)–(18) as output. A possible range of values $q_{\hat{k}+1}$ could be chosen as follows (with $\min\limits_{1\le j\le \hat{k}} \{q_j\} = m\cdot 10^{-2},\ \ 1\le m\le 9$, i.e. if the least of the stratum shares is at the level of several per cent):

$$q_{\hat{k}+1} = \begin{cases} \nu\cdot 10^{-2}, & \nu = m-1,\, m-2,\ldots,1; \\ \nu\cdot 10^{-3}, & \nu = 9,8,\ldots,1; \\ \nu\cdot 10^{-4}, & \nu = 9,8,\ldots,1. \end{cases}$$

(iii) by using (14’), the following limit from above for the share of the unobserved stratum can be calculated:

$$q_5 < \frac{1}{x_0}\left(\mu^{\text{макро}} - \sum_{j=1}^{\hat{k}} \tilde{q}_j e^{\hat{a}_j + \frac{\sigma_j^2}{2}}\right) \qquad (19)$$

4.4.5. Poverty indices and targeted assistance to the poor

If we restrict the class of weighting functions $w(x)$ in (1) to the functions like (3), then we can use results of [23] on the optimal allocation of the financial aid to the poor. By combining those with the estimates of the per capita expenditure density function $f(x)$, we can formulate the following rule of targeted assistance:

(i) For given inputs of the model (such as the population size N, poverty line z_0, total resource S for targeted assistance, density function $f(x)$ describing the population per capita expenditures, and Foster-Greer-Thorbecke index parameter $\alpha > 1$), the threshold value $\bar{z}_0$ can be found from

$$N\cdot I_0^{(\bar{z}_0)}(f)\cdot \bar{z}_0 \cdot I_1^{(\bar{z}_0)}(f) = S; \qquad (4')$$

(ii) Each inhabitant of the region whose per capita expenditure x is below the threshold, $x < \bar{z}_0$, is then eligible to the lump sum transfer $\bar{z}_0 - x$.

Apparently, for each weighting function $w(x)$ there a corresponding optimal allocation.

In this study, the share of poor (head count ratio, FGT(0)) and poverty depth (FGT(2) sensible to the extreme poverty and thus interpretable as the social tension indicator) are calculated for each data set (the three regions and RLMS) in the following ways: i) immediate (non-parametric) sample statistics; ii) by using the estimates of the lognormal expenditure distribution model mimicking Goskomstat; ii) by using the estimates of the lognormal mixture model. The results follow in section 4.5.

4.5 *The results of econometric estimation*

The main and auxiliary tasks formulated in the motivation section and the methodological issues outlined in the previous paragraph have determined the following steps of the econometric analysis.

1. The analysis of sample distributions of per capita expenditure is conducted by using the Goskomstat budget surveys data on Komi Republic, Volgograd and Omsk oblasts (Q2 1998), as well as RLMS Round VIII data (Q4 1998). In particular, sample statistics and histograms are obtained as the output of this step (see Appendix 1).
2. By using RLMS panel data Rounds V–VIII and additional refusal data[3], the multiple logit model was estimated to relate the probability of a household with particular characteristics to refuse to participate in a budget survey.
3. According to the methodology described in section 4.4.2, either rough (with the logit model (6'')) or fine (with the logit model (6)) calibration (re-weighting) of the existing data was performed to eliminate truncation bias.
4. Sample distributions are re-analyzed accounting for weights estimated at the previous step. The results are compared to those obtained at step 1.

[3] The authors are grateful to P. M. Kozyreva and E. Artamonova from RAS Institute of Sociology who kindly provided this data to us.

5. The mixture model parameters for the three regional and the national data sets are estimated with the observed range of per capita expenditure (see the methodology in paragraph 4.4.3).
6. According to the methodology described in paragraph 4.4.4, the unobserved component parameters are produced for each of the four data sets by using the estimates of the mixture components obtained at the previous step. The goal here is to eliminate the truncation bias (see Motivation).
7. With the estimates of the distribution functions from steps 4 and 6, the poverty and inequality indices are calculated and analyzed for the three regions as of Q2 1998 and for Russia as a whole as of Q4 1998.

4.5.1. Statistical analysis and calibration of the per capita expenditure distributions

In this section, some evidence from Figures A.1–A.4 and Tables A.1–A.8 will be discussed.

First, the per capita distributions cannot be adequately described by the simple lognormal model (neither within any of the regions nor within the country as a whole). The p-value of χ^2 goodness of fit statistic is less than 10^{-6} for all data sets (see columns 5 and 6 of Tables A.5 – A.8). Two-parameter lognormal mixture model does not perform very well, too. The p-value in this case does not exceed 0.001, except for Omsk oblast where it amounts to 0.085. The three component model for Omsk demonstrate better performance in terms of LR statistics, AIC and ICOMP information criteria, and χ^2 statistic, though.

Second, the algorithms of both CLASSMASTER and Stata of the automatic search for the unknown number of the mixture components k typically lead to the estimates $\hat{k} = 3$ or $\hat{k} = 4$, i.e. the per capita expenditure distribution of a region / country can be represented as a mixture of three or four homogeneous socio-economic strata. This would not necessarily mean that there exist three or four local density maxima.

Volgograd region was the only exemption. While for all other cases increasing the number of mixture components beyond four leads to serious deterioration of the identification quality (multiple maxima of the likelihood function, flat regions that the algorithms stumble upon, coinciding components, etc.), the five components model for Volgograd was the most parcimonious model accepted by goodness of fit criteria.

Third, as compared to the 1996 picture [16], the stratification of population is less manifested. This complies with the tendency of the per capita expenditure distribution to return to its "normal" lognormal shape as economic transition proceeds.

The figures from [16] describing the per capita income distribution of Russian population as of Fall 1995 suggest local maxima of the density function. Population strata are then well-defined, which allows for sensible classification of population by the strata with the following analysis of social and economic characteristics of each stratum. The similar analysis for the 1998 data is hampered by the fact that most population is classified into the central, or modal, strata preventing us from conducting a similar study within this project.

Fourth, the share of unobserved strata is relatively small and varies about 0.1–0.01%. Nevertheless, it has crucial influence on the mean income of the population and inequality indices. The parameters of the hidden stratum are estimated up to the level curve relating $q_{\hat{k}+1}$ and $a_{\hat{k}+1}$ under certain restrictions (see (14)–(18) in section 4.4.4). The example of such line for RLMS data is given below at Fig. 1.

It turns out, however, that the indicators of our interests (mean expenditure, poverty characteristics, Gini index of inequality) do not crucially depend on the choice of a particular point $(\hat{\mu}^{(\hat{k}+1)}, \hat{q}_{\hat{k}+1})$ on this level line. In fact, poverty indices are focused on the left tail of the distribution. Inequality and polarization measures do depend upon the unobserved stratum, but, for instance, Lorenz curve on which Gini index is based is not very sensitive to the particular choice of the parameters couple (though it is sensitive to the very fact of inclusion or omission of the hidden stratum). The relative insensitivity of Gini index to the tails of the distribution was discussed in [20]. Various estimates of the share of hidden income range from 25% to 40% (see

[16] and references in it). In this study, calibration effect is to increase the mean of observed expenditure by some 2–3%, while introduction of the hidden strata is responsible for the most of the 20–30% difference. In particular, the increase of the mean expenditure due to the hidden stratum is (1211—913)/913=0,326=32,6%.

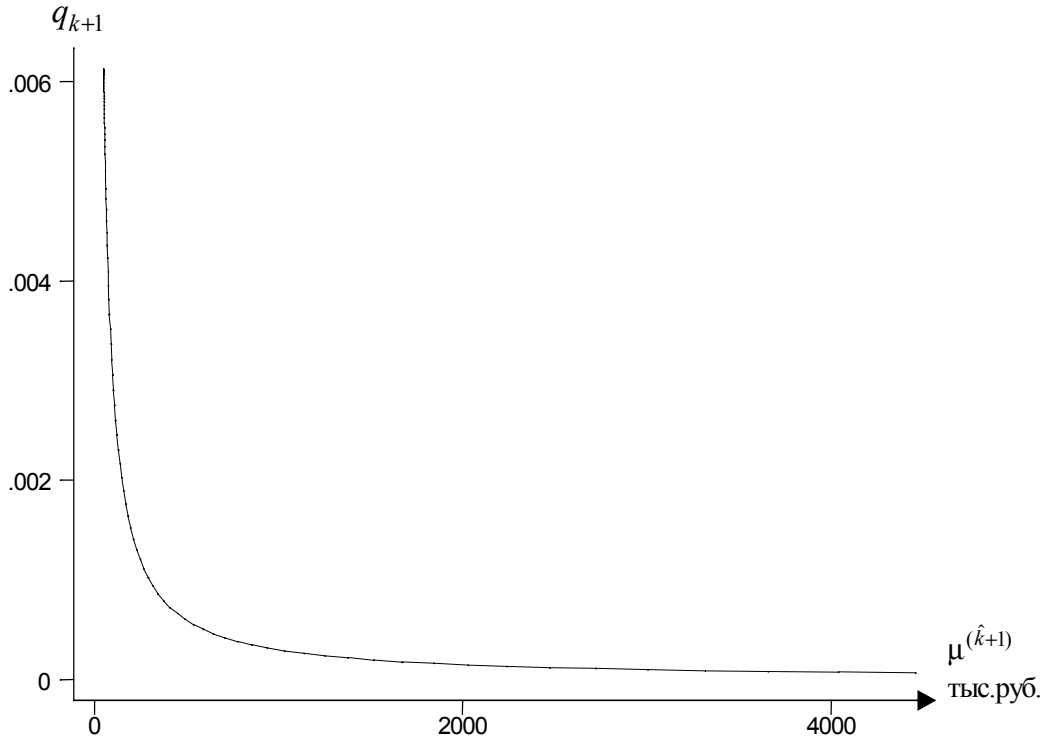

Figure 1. The relation between the share and the mean per capita expenditure used in the estimation of the latent population strata parameters.

Fifth, the observation re-weighting (used here to adjust for truncation) and Monte Carlo simulation modeling of the unobserved stratum help explaining the 40% difference between the official (i.e. registered by the statistical bodies) and actual (i.e. observed in budget surveys) income / expenditure of population.

4.5.2. Poverty, inequality and social tension indices estimation

Table 1 reports the estimates of poverty and social tension indicators. In terms of Foster-Greer-Thorbecke family of indices $I_{\alpha}^{(z_0)}(f)$ (see [14] and (1)–(1') in the motivation section), these are FGT(0), or the head count ratio, and FGT(2), the indicator of poverty depth (and

hence social tension caused by the existence of the poorest people). The table includes the official Goskomstat data (column 4); the data from the World Bank targeted assistance pilot projects [9] (column 5 for the regions participated in these projects), direct weighted sample estimates of the indices (columns 8 and 9), and the FGT(0) and FGT(2) estimates from the lognormal model (columns 6 and 10) and the lognormal mixture model (columns 7 and 11).

Table 1. Poverty and social tension indicators.

	Region	Poverty line (ths. rub.)	Poverty rate					Poverty depth (social tension, FGT(2))		
			Official Goskomstat figure	Pilot projects estimates	Lognormal model	Mixture model	Sample	Sample	Lognormal model	Mixture model
1	Russia	0,636	28,4	—	52,5	52,8	53,9	0,137	0,139	0,130
2	Komi Republic	0,466	20,6	26,7	53,8	56,2	56,7	0,127	0,130	0,128
3	Volgograd oblast	0,368	31,5	49,2	62,0	62,7	63,0	0,177	0,177	0,177
4	Omsk oblast	0,372	25,2	—	42,6	43,2	44,2	0,082	0,089	0,081

Sources: Columns 3 and 4 are due to [37]–[39]. Column 5 is due to [9]. The rest are authors' estimates based on the regional datasets (Q2 1998) and RLMS Round VIII (October–November 1998).

Table 2. The results of the distribution calibration and inequality comparisons

No.	Region, data source, sample size	Mean expenditure, ths. rbs.			Gini index		Funds ratio	
		Raw	Calibrated (+Δ, %)	With the latent stratum (+Δ, %)	Raw data	Model (5) with latent stratum	Raw data	Model (5) with latent stratum
1	2	3	4	5	6	7	8	9
1	Russia, RLMS VIII, n=11397	0.913	0.932 (2%)	1.211 (29%)	0.478, 0.380*	0.599	13.5*	22.3
2	Komi Republic, HBS, n=1089	0.633	0.686 (8%)	1.159 (83.1%)	0.395	0.667	15.6	43.7
3	Volgograd oblast, HBS, n=1263	0.412	0.433 (5%)	0.642 (55.6%)	0.389	0.590	14.0	32.0
4	Omsk oblast, HBS, n=1244	0.611	0.641 (5%)	0.699 (14.4%)	0.357	0.442	10.5	14.8

Russia: Q4 1998; the regions: Q2 1998.
Funds ratio is the ratio of the total income / expenditure in the top decile to the one in the bottom decile

Table 2 contains the results of each of the calibration stages: weighting of the existent observations, and introduction and estimation of the unobserved mixture component. The inequality characteristics such as Gini index and funds ratio are also reported. Goskomstat does not report the regional figures for these indices, so we provide the direct sample estimates.

The analysis of the tables leads to the following conclusions.

1) There exists a significant dispersion of the indicators, both between regions and (for each region) between the estimation methods. We believe that the weighted sample estimates are the most precise (columns 8 and 9). It does not re-rank the regions by poverty rates and depths, but result in higher poverty rates than the official statistics asserts. For Komi Republic and Volgograd oblast the discrepancy is twofold! On the other hand, the mixture model estimates produce results much closer to the sample estimates than the official values. This is not surprising given a satisfactory quality of fit evidenced by the statistical tests.
2) Although the share of the unobserved super-rich stratum is relatively low (tenth or hundredth of the percentage point), it crucially affects the main characteristics of inequality and polarization. For instance, Gini index as officially reported by Go-

skomstat for November 1998 was 0.372. The RLMS sample value (with weights) was 0.488, and after the inclusion of the latent stratum, it further raises to 0.610. The same behavior can be observed for the funds ratio (i.e. the ratio of the mean incomes in the top and bottom deciles). It might also be noted that the discrepancy is largest for Komi republic which is a resource rich region. This fact is supported by the rent seeking theory, i.e., that the rent seeking behavior emerge in the economic environments with substantial rent flows, natural resource rent being the most typical example.

The key characteristics of expenditure inequality have changed substantially once the latent stratum is taken into account. In particular, Gini index for Russia in Q3 1998 was reported to be 0.380; the sample estimate from the RLMS data is however 0.478, while the estimate based on the latent stratum model gives 0.599. The similar pattern of the increase in Gini values is observed for the regional data, too (except may be for the Omsk oblast). The magnitude of changes in the funds ratio is also really large, 50% to 200%.

How the revealed differences in figures can be explained, and should the results based on the model (5) be trusted?

Table 1 reports poverty indices estimates based on the left tail of the distribution. As it should have been expected, the differences between the results from the lognormal model and the mixture model (5) (see columns 6 vs. 7, and 10 vs. 11), are small though systematic, as all “mixture” estimates are greater than the respective “lognormal” estimates). We cannot provide a good explanation for the differences between the lognormal model-based indices and the official figures, as those seem to be based on the same methodology. In fact, the data sources for the two figures are different, as we were using the RLMS data, and Goskomstat used its HBS data for Q4 1998. Besides, the way Goskomstat treats those budget data is different from what is usually done by researchers.

Table 2 shows the expenditure inequality characteristics that require the knowledge of the whole distribution, including both tails. As one of the most prominent features of the mixture

model (5) is the modification of the right tail approximation, the differences in the inequality figures from those reported by Goskomstat are quite striking (compare columns 7 and 9 vs. 6 and 8). One might even say that the inequality indices obtained by using the model (5) are too large[4]. To provide some explanation, we need to note that in (5), all the discrepancy between the macroeconomic figure for the mean expenditure and the sample mean from the RLMS / HBS is assigned to this latent stratum (see columns 4 and 5 for Table 2). If this assumption is too strong, and the discrepancy is only partially explained by the latent stratum (and partially, due to misreporting in the observed ranges), then the estimates of the inequality indices given by (5) are biased upward. On the other hand, in earlier studies, the discrepancy was compensated for only by calibration of the existent observations, i.e., the latent stratum was ignored. It is likely that the truth is somewhere in between.

This question is addressed in more detail in the following subsection.

4.5.3. The sensitivity analysis of Gini index estimate with respect to misreporting

The overstatement of the latent stratum importance in explaining the discrepancy between the macro and micro averages in the model (5) might be caused by the systematic bias of the sample data due to misreporting (see above footnote 2 on page 6). In other words, if the individuals surveyed intentionally underreport their income and expenditure, then the sample mean is biased downwards, and the aforementioned discrepancy, upward.

To investigate the sensitivity of model (5) based Gini index to misreporting in the RLMS data, the following framework was adopted.

Let the distortion due to the misreporting is measured as

$$\lambda = \frac{\mu_{act} - \mu_{cal}}{\mu_{cal}} \times 100\%$$

[4] The cross country comparison of Gini indices suggests that some figures in Table 2 might be overstated. The lowest values of about 0.25–0.30 are observed in Nordic countries; the figure for US is about 0.35–0.36; and the countries that are known to have high inequality are Brazil, Mexico, or South Africa, but even in these countries, the value of Gini is estimated to be about 0.45–0.6.

where μ_{cal} is the sample mean expenditure (possibly biased due to misreporting) calibrated according to the methodology described in section 4.4.2 (here, μ_{cal} = 0.932, see column 4 of Table 2); and μ_{act} is the actual average expenditure of the households in the sample.

Evidently, in the preceding analysis (and the Tables 1 and 2, as its result) it was assumed that λ=0, and thus the discrepancy between the μ_{macro} and μ_{cal} in the observed expenditure range is about 30% ((1211-932)/932 = 0.299). The international practice of budget studies suggests that the discrepancy of several percentage points is inevitable, but figures larger than 10% should signal serious problems with the sample quality. Still, some fraction of the discrepancy can be attributed to the households in the observed expenditure ranges.

The estimates of Gini index in the framework of the model (5) with correction for the misreporting for a number of λ's are given in the Table 3.

Table 3. The sensitivity analysis of Gini index estimate with respect to misreporting

Relative distortion λ	0%	5%	10%	15%
Estimates of Gini index based on (5) and corrected for misreporting	0.608	0.592	0.569	0.554

The simplest model of misreporting was adopted, namely, that each households understates its true expenditure by the factor of $(1+\lambda)^{-1}$ (in other words, that all reported figures should be increased by λ per cent). The reported figures are the results of the parametric bootstrap that used the underlying distribution (5) with the estimates of the mixture parameters obtained earlier. The parameters of the latent stratum were estimated as described in section 4.4.3, 4.4.4, and Appendix 2. For each λ, 20 bootstrap samples of size 400000 were created. The sample size was chosen to guarantee the adequate representation of the latent stratum with the share of the stratum in the population $q_{k+1}<0.1\%$. Typically, about a hundred households from the latent stratum were present in each bootstrap sample. The number of bootstrap samples

(20) allows to interpret the observed range as the approximate 95% confidence interval for the true Gini (see box-whisker plots on Fig. 2).

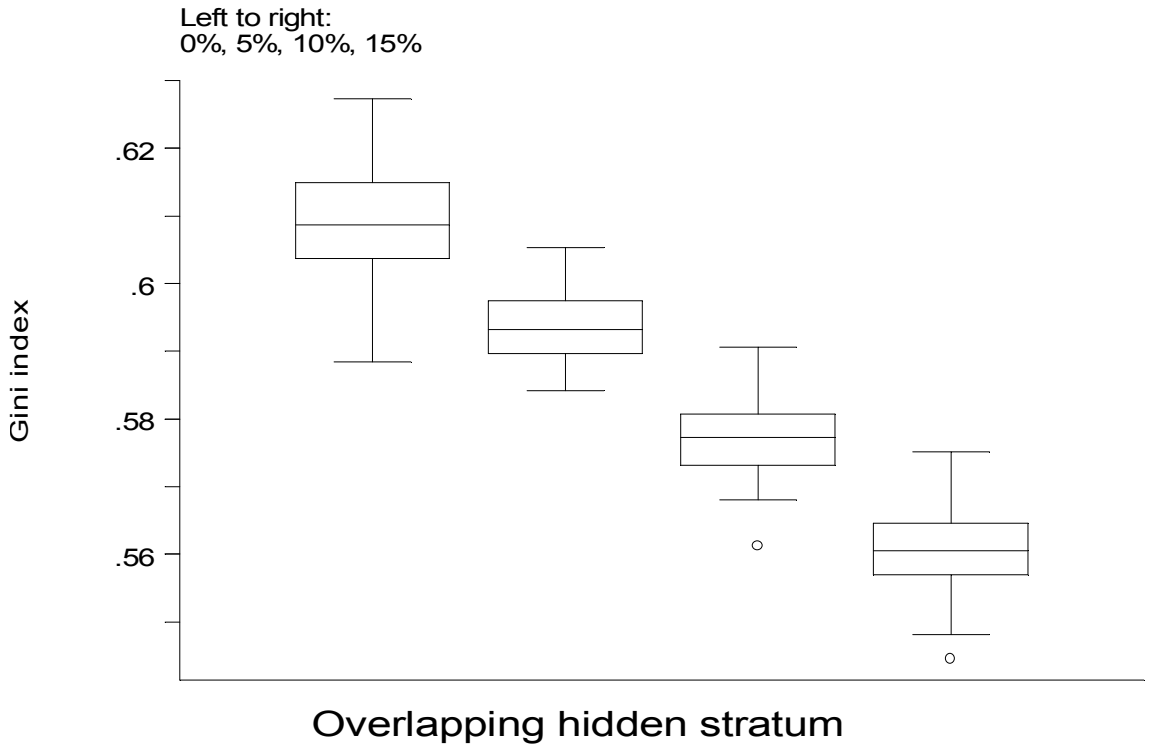

Fig. 2a. Box-whisker plots for Gini indices obtained at various levels of λ.

The results of this sensitivity analysis suggest that the estimates based on model (5) are still higher than the sample values even if half of the total discrepancy is attributed to the misreporting factor. Assuming that the realistic values of λ for RLMS sample range from 10% to 15%, *the most viable range of Gini index is 0.55–0.57*.

5 CONCLUSIONS

Uncertainties and ambiguities on almost all levels characterize the economic environment of Russian transition. At the level we are interested in, namely, the microeconomics of households, one of the major ambiguities is the level of the household welfare. Uncertainty about the welfare suggests that **expenditure** rather than *income* is to be used for the purposes

of poverty and inequality evaluation as well as for the dichotomy of the households into poor or non-poor. We would like to note that if expenditure is used,

a) the problem of wage arrears in a household is resolved;
b) intentionally or non-intentionally hidden income, including income from shadow economy, is accounted for;
c) the concept of household welfare is appropriately generalized to include land (subsidiary plot) and property (real estate, private transportation means, jewelry, etc.) the household possess.

When gross expenditure of the household is calculated, all sorts of expenditure are added up, including expenses for consumer goods, intermediate goods (including the subsistence plot operations), net savings in all assets (including bank deposits and foreign currency operations), fixed capital growth, taxes and other obligatory payments, cash, and home production. In this work, this total expenditure was simply divided by the household size, i.e. the simplest equivalence scale was used. More complicated equivalence scales might have been used, but we view these as technicalities that can be easily accommodated into the research, but that would hardly affect any of the qualitative results.

The specific features of Russian transition did not cancel the lognormal model of income/expenditure distribution, though they did affect the mixing function $q(a)$. The genesis of the *discrete* lognormal mixture (instead of *continuous* mixture of special form reproducing the lognormal distribution typical for stable economies) is explained by the structural labor, human capital and skills demand shifts during the transition. These changes have crowded out the "Soviet middle class", i.e. relatively qualified workers, who has to seek other, as a rule, less profitable, income sources. This search has been adversely affected by low labor mobility (primarily, geographical mobility) typical for Russia. At the same time, new «extra rich» population groups have acquired substantial rent flows. Thus, a well-defined pattern of groups of income earners has developed which has led to the discrete character of distribution mixture, the distribution being lognormal within each group. Hence, it is natural to try to model the underlying distribution

by a discrete lognormal mixture. It is worth noting that as transition draws to a close, i.e. the Russian economy evolves towards its steady state, the shape of the mixing function $q(a)$ (and, consequently, of the whole expenditure distribution) would tend to resemble a usual two parameter lognormal distribution. The comparison of the estimation results based on 1998 and 1996 data confirms this tendency.

The econometric analysis of the proposed model includes: **a)** per capita expenditures denisty idenitification via lognormal finite mixture parameters $\Theta = (k; a_1, \ldots, a_k;\ \sigma_1^2, \ldots, \sigma_k^2)$ estimation by the appropriate statistical procedures (see [31]–[35]); **b)** re-weighting of the distribution accounting for the probability of unit non-response as a function of per capita expenditures; **c)** reconstruction of the unobserved $(\hat{k}+1)$-th stratum with the second re-calibration of the model based on partially verifiable working hypotheses and macroeconomic income and expenditure balances.

The proposed per capita expenditure distribution model includes two stage calibration procedure and allows to compensate for truncatoin bias by adjusting the parameters of the model to comply with the macroeconomic statistical data. Thus, a better estimate of the main poverty and inequality figures is obtained as compared to the methods currently used in the official practice ([1]–[3], [7], [28]) and by other researchers ([4]–[6]). In particular, the model can be used for computations related to the establishment of the targeted social assistance framework.

It should be noted that the techniques developed in this study are only applicable at the regional level. Regional results can only be aggregated if the appropriate deflators and coefficients are used that would account for interregional price differentials, purchasing power, the subsistence basket composition, etc.

6 REFERENCES

[1] Ê ìåòîäèêå îöåíêè ðàñïðåäåëåíèÿ ãåíåðàëüíîé ñîâîêóïíîñòè íàñåëåíèÿ ïî âåëè÷èíå ñðåäíåäóøåâîãî äîõîäà íà îñíîâàíèè ýìïèðè÷åñêèõ äàííûõ. - Ìàòåðèàëû Ãîñêîìñòàòà ÐÔ (òåçèñû ê äîêëàäó íà

Ó÷åíîì ñîâåòå â ÖÓÆå), 1999.

[2] *Âåëèêàíîâà Ò., Êîëìàêîâ È., Ôðîëîâà Å.* Ñîâåðøåíñòâîâàíèå ìåòîäèêè è ìîäåëåé ðàñïðåäåëåíèÿ íàñåëåíèÿ ïî ñðåäíåäóøåâîìó äîõîäó. - «*Âîïðîñû ñòàòèñòèêè*», 1996, ¹ 5.

[3] *Âåëèêàíîâà Ò.Á., Ôðîëîâà Å.Á.* Ìåòîäîëîãèÿ ïðÿìîé îöåíêè âåëè÷èíû óðîâíÿ áåäíîñòè, îñíîâàííàÿ íà ðàñïðîñòðàíåíèè ðåçóëüòàòîâ âûáîðî÷íîãî îáñëåäîâàíèÿ íà ãåîðàñïðåäåëåííîì ìàòåðèàëå Óïðàâëåíèÿ ñòàòèñòèêè óðîâíÿ æèçíè íàñåëåíèÿ Ãîñêîìñòàòà ÐÔ, 1999.

[4] *Øåâÿêîâ À.Þ., Êèðóòà À.ß.* Ýêîíîìè÷åñêîå íåðàâåíñòâî, óðîâåíü æèçíè è áåäíîñòü íàñåëåíèÿ Ðîññèè è åå ðåãèîíîâ â ïðîöåññå ðåôîðì: ìåòîäû èçìåðåíèÿ è àíàëèç ïðè÷èííûõ çàâèñèìîñòåé. - Ôèíàëüíûé îò÷åò ïî ïðîåêòó EERC, èþëü 1999.

[5] *Åðøîâ Ý.Á., Ìàéåð Â.Ô.* Ìåòîäîëîãè÷åñêèå è ìåòîä÷åñêèå ïðîáëåìû îïðåäåëåíèÿ óðîâíÿ, îáúåìà è äèôôåðåíöèàöèè äîõîäîâ íàñåëåíèÿ. - Ìàòåðèàëû ê çàñåäàíèþ ÂÖÓÆ 28 äåêàáðÿ 1998 ã.

[6] *Ñóâîðîâ À.Â., Óëüÿíîâà Å.À.* Äåíåæíûå äîõîäû íàñåëåíèÿ Ðîññèè: 1992-1996 ãã. - «Ïðîáëåìû ïðîãíîçèðîâàíèÿ», 1997 ã.

[7] Îïðåäåëåíèå ñîîòíîøåíèÿ äàðåãèðîâàíèÿ äàííûõ îáñëåäîâàíèÿ áþäæåòîâ äîìàøíèõ õîçÿéñòâ. - Ìàòåðèàëû Ãîñêîìñòàòà ÐÔ, 1999.

[8] *Áðåéòóéò Äæ.* Àäðåñíîñòü è îòíîñèòåëüíî äëèòåëüíàÿ áåäíîñòü â Ðîññèè. Ôåðàëü 1999. Äîêëàä íà ñåìèíàðå Âñåìèðíîãî áàíêà 19 àïðåëÿ 1999 ã.

[9] Ïèëîòíûå ïðîãðàììû ïî âûâåäåíèþ àäðåñíîé ñîöèàëüíîé ïîääåðæêè ìàëîèìóùèõ ñåìåé â Ðåñïóáëèêå Êîìè, Âîðîíåæñêîé è Âîëãîãðàäñêîé îáëàñòÿõ. Ïðåäâàðèòåëüíûå èòîãè. Ìèíèñòåðñòâî òðóäà è ñîöèàëüíîãî ðàçâèòèÿ ÌÔ. - Ì.: 1999. - 104 ñ.

[10] *Àéâàçÿí Ñ.À.* Èíòåãðàëüíûå èíäèêàòîðû êà÷åñòâà æèçíè íàñåëåíèÿ: èõ ïîñòðîåíèå è èñïîëüçîâàíèå â ñîöèàëüíî-ýêîíîìè÷åñêîì óïðàâëåíèè è ìåæðåãèîíàëüíûõ ñîïîñòàâëåíèÿõ. - Ìîñêâà: ÖÝÌÈ ÐÀÍ, 2000. - 117 ñ.

[11] *Àéâàçÿí Ñ.À., Ãåðàñèìîâà È.À.* Ñîöèàëüíàÿ ñòðóêòóðà è ñîöèàëüíîå ðàññëîåíèå íàñåëåíèÿ Ðîññèéñêîé Ôåäåðàöèè (ïî ìàòåðèàëàì âûáîðî÷íîãî îáñëåäîâàíèÿ íàñåëåíèÿ òðåõ ðåãèîíîâ ÐÔ). - Ì.: ÖÝÌÈ ÐÀÍ, 1998.

[12] *Êîëåíêîâ Ñ.Î.* Ìåòîäû àíàëèçà êà÷åñòâà æèçíè. Ñåðèÿ ÐÝØ «Ëó÷øèå ñòóäåí÷åñêèå ðàáîòû», 1999 (â ïå÷àòè).

[13] *Hagenaars A.* A Class of Poverty Indices. - «Interational Economic Review», 28, 1987, pp. 583-607.

[14] *Foster J., Greer J., Thorbeck E.* A Class of Decomposable Poverty Measures. - «Econometrica», 52(3), 1984, pp. 761-766.

[15] *Bourguignon F., Fields G.* Discontinuous loss from poverty, generalized P_α measures, and optimal transfers to the poor. - XI-th World Congress of the International Economic Association. Tunis, December 1995.

[16] *Àéâàçÿí Ñ.À.* Ìîäåëü ôîðìèðîâàíèÿ ðàñïðåäåëåíèÿ íàñåëåíèÿ Ðîññèè ïî âåëè÷èíå ñðåäíåäóøåâîãî äîõîäà. - «*Ýêîíîìèêà è ìàòåìàòè÷åñêèå ìåòîäû*», òîì 33 (1997), ¹ 4, ñ. 74-86.

[17] Íàöèîíàëüíàÿ îöåíêà è ðàñïðîñòðàíåíèå èíôîðìàöèè. Ïðåäâàðèòåëüíûé îò÷åò â ðàìêàõ ïðîåêòà Âñåìèðíîãî áàíêà ïî òåìå «Ñòðóêòóðíàÿ ïåðåñòðîéêà ñèſòåìû ñîöèàëüíîé çàùèòû íàñåëåíèÿ»

¹SPIL-2.2.2/11. ÃÓ-ÂØÝ, Ìîñêâà, 1999.

[18] *Êîð÷àãèíà È., Îâ÷àðîâà Ë., Òóðóíöåâ Å.* Ñèñòåìà èíäèêàòîðîâ óðîâíÿ áåäíîñòè â ïåðåõîäíûé ïåðèîä â Ðîññèè. - Íàó÷íûé äîêëàä ¹98/04 ïî ñåêöèè «Ìèêðîýêîíîìèêà-2 (ïîâåäåíèå äîìîõîçÿéñòâ)» Ðîññèéñêîé ïðîãðàììû Êîíñîðöèóìà ýêîíîìè÷åñêèõ èññëåäîâàíèé. ÐÏÝÈ/Ôîíä Åâðàçèÿ, àïðåëü 1999.

[19] *Sen A.K.* A Sociological Approach to Measurement of Poverty. - Oxford Economic Papers, 37, pp. 669-667.

[20] *Atkinson A.B.* On the Measurement of Poverty. - Econometrica, vol.55 (1987), N4, pp.749-764.

[21] *Kanbur S.M.R.* Measurment and Alleviation of Poverty. - «IMF Staff Papers», 34, 1987, pp. 60-85.

[22] *Foster J.E., Shorroks A.F.* Poverty Orderings. - «Econometrica», 56(1), 1988, pp. 173-177.

[23] *Bourguignon F., Fields G.S.* Poverty Measures and Anti-Poverty Policy. - «Recherches Economiques de Louvain», 56(3-4), 1990, pp. 409-427.

[24] *Ravallion M.* Poverty Comparisons. - Chur, Switzerland, Harwood Academic Publishers, 1994.

[25] *Esteban J.-M., Ray D.* On the Measurement of Polarization. - Econometrica, 62, No.4, pp. 819-851.

[26] *Fajnzulber P., Lederman D., Loayza N.* Inequality and Violent Crime. - The research project «Crime in Latin America» of the World Bank, 1999.

[27] *Àéâàçÿí Ñ.À., Ðàáêèíà Í.Å., Ðèàøåâñêàÿ Í.Ì.* Ìåòîäèêà ðàñ÷åòà îæèäàåìîãî ðàñïðåäåëåíèÿ ðàáî÷èõ è ñëóæàùèõ ïî ðàçìåðàì çàðàáîòíîé ïëàòû. -Ì.: ÍÈÈ òðóäà ÃÊÑÌÑÑÐ ïî âîïðîñàì òðóäà è çàðàáîòíîé ïëàòû, 1967.

[28] *Mroz T., Popkin B., Mancini D., Glinskaya T., Lokshin V.* Monitoring Economic Conditions in the Russian Federation: The Russia Longuitudinal Monitoring Survey 1992-1996. - Report submitted to the U.S. Agency for International Development. Carolina population Center. University of North Caroline at Chapel Hill. February, 1997.

[29] *Aivazian S.A.* Probabilistic-Statistical Modelling of the Distributary Relations in Society. - In: «Private and Enlarged Consumption», North-Holland Publ. Comp., 1976.

[30] Ìåòîäîëîãè÷åñêèå ïîëîæåíèÿ ïî ñòàòèñòèêå. Âûïóñê 1. - Ì., Ãîñêîìñòàò ÐÔ, 1996.

[31] *Day N.E.* Estimating the Components of a Mixture of normal distributions. - Biometrika, vol. 56 (1969), N3, pp. 463-474.

[32] *Dempster A., Laird G., Rubin J.* Maximum Likelihood from Incomplete Data via the EM-algorithm. - J.R. Statist. Soc., B.39 (1977), pp. 1-38.

[33] *Aivazian S.A.* Mixture-Model Cluster Analysis Using the Projection Pursuit Method. - In: «Computational Learning and Probabilistic Reasoning», John Wiley and Sons Ltd, 1996, pp. 278-286.

[34] *Rudzkis R., Radavicius M.* Statistical Estimation of a Mixture of Gaussian Distributions. - In: «Acta applicandae Mathematical», vol. 38, N1, 1995.

[35] *Jakimauskas G., Sushinkas J.* Computational aspects of statistical analysis of gaussian mixture combining EM algorithm with non-parametric estimation (one-dimensional case). - Preprintas N96-6, Matematicos ir Informaticos Institutas, Vilnius, Lietuva, 1996.

[36] The Russia Longitudinal Monitoring Survey: «Family questionnaire» and «Sample of Russian Federation». Rounds V and VI. Technical Report. August-October 1996.

[37] Ñîöèàëüíî-ýêîíîìè÷åñêîå ïîëîæåíèå Ðîññèè. - Ãîñóäàðñòâåííûé êîìèòåò Ðîññèéñêîé Ôåäåðàöèè ïî ñòàòèñòèêå, 1998, I-XII.

[38] Êðàòêîñðî÷íûå ñîöèàëüíî-ýêîíîìè÷åñêèå ïîêàçàòåëè. Ãîñêîìñòàò ÐÔ, àïðåëü 1999.

[39] Ñîöèàëüíîå ïîëîæåíèå è óðîâåíü æèçíè íàñåëåíèÿ Ðîññèè. Îôèöèàëüíîå èçäàíèå. - Ì.: Ãîñêîìñòàò Ðîññèè, 1999.

[40] Áîãîìîëîâà Ò.Þ., Òàïèëèíà Â.Ñ., Ðîñòîâöåâ Ï.Ñ. Ðîñò ìîáèëüíîñòè ïî äîõîäàì â èçìåíåíèè íåðàâåíñòâà â ðàñïðåäåëåíèè äîõîäîâ. Íîâûé ïðîåêò, ÐÏÝÈ, äåêàáðü 1999.

[41] Deaton, Angus. Understanding consumption, Clarendon Lectures in Economics. Oxford University Press, Clarendon Press, 1992.

[42] Stata Reference Manual. Release 6. Stata Press, 1999.

[43] Little, R.J.A and D.B. Rubin. Statistical Analysis with Missing Data. Wiley, 1987.

[44] Gould, W., and W. Sribney. Maximum Likelihood Estimation with Stata. Stata Corp., 1999.

[45] Buhmann, B., L. Rainwater, G. Schmaus, T. M. Smeeding. Equivalence Scales, Well-being, Inequality and Poverty: Sensitivity estimates across ten countries using the Luxembourg Income Study (LIS) database. The Review of Income and Wealth, 34, 115–142, 1988.

[46] Coulter, F.A.E., F.A.Cowell, S.P.Jenkins. Differences in Needs and Assessment of Income Distributions. Bulletin of Economic Research, 44 (2), 77–124, 1992.

[47] Cowell, F.A., M.Mercader-Prats. Equivalence of Scales and Inequality. DARP Discussion Paper, No. 27. STICERD, LSE, 1997.

APPENDICES

APPENDIX 1. THE ANALYSIS OF THE SAMPLE DISTRIBUTIONS OF PER CAPITA EXPENDITURE FOR PARTICULAR REGIONS AND RUSSIA AS A WHOLE

A1.1. Russian Federation

RLMS data set, Round VIII, October–November 1998, 3618 observations.

Table A.1 Sample statistics of the essential characteristics of Russian per capita expenditure distribution.

№№ пп.	Indicator (ths. rub.)	Sample value	
		Raw data	Weighted data
1	Mean per capita expenditure $(\hat{\mu})$	0,914	0,932
2	Standard deviation (S)	1,765	1,682
3	Minimal expenditure $(x_{\min})$	0,008	0,008
4	Maximal expenditure $(x_{\max})$	49,344	49,344

5	Bottom decile $(\hat{x}_{0,1})$	0,200	0,209
6	Top decile $(\hat{x}_{0,9})$	1,699	1,763

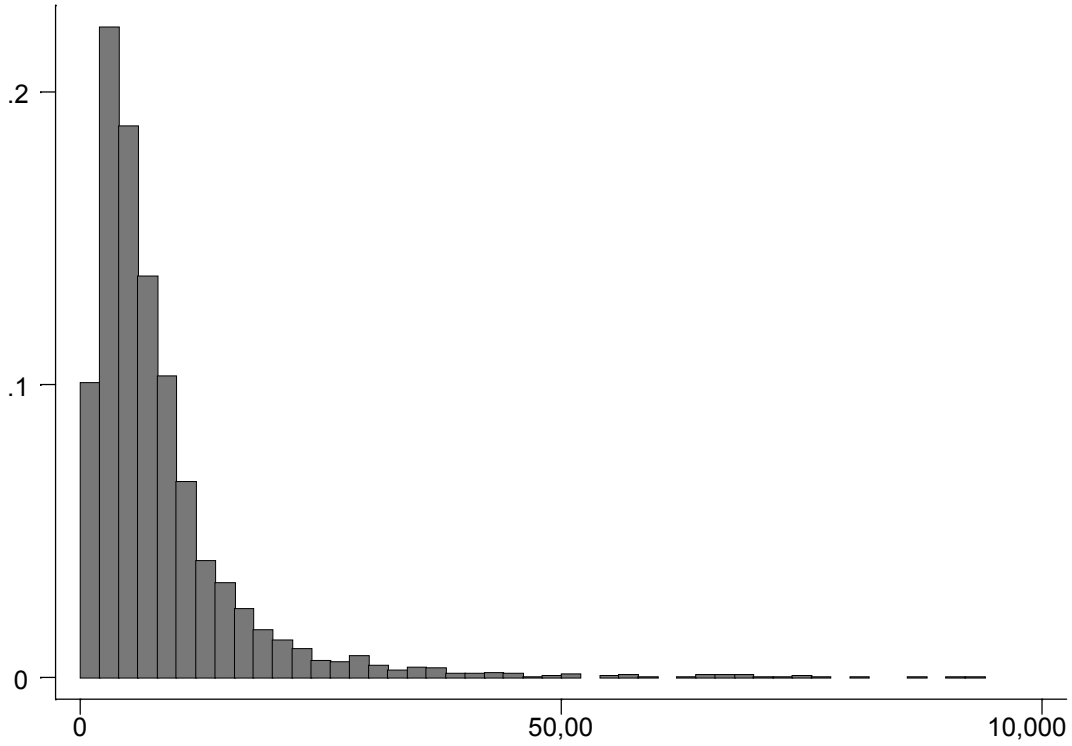

Fig. A.1a.

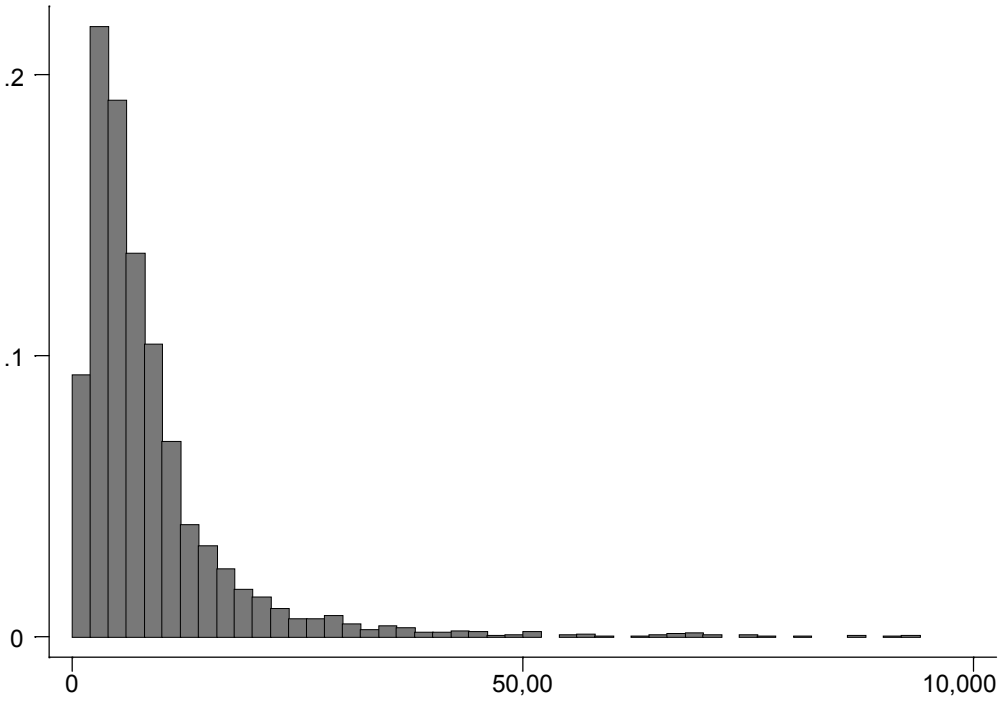

Fig. A.1b. The histogram of the per capita expenditure distribution of Russian population (weighted data).

A1.2. Komi Republic

Budget survey sample of 1089 individuals, Q2 1998.

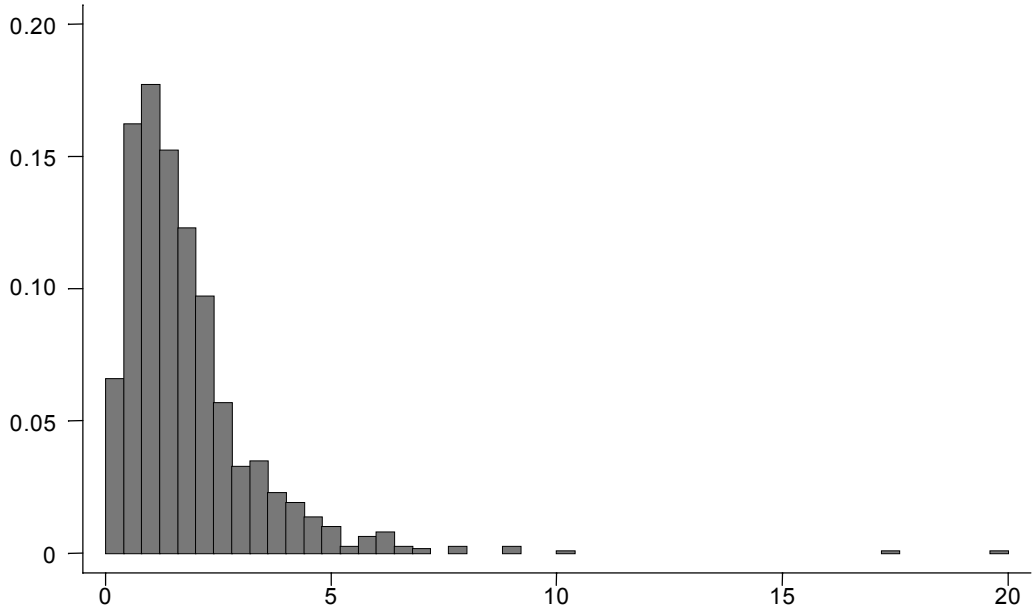

Fig. A.2a. The histogram of the per capita expenditure distribution of population of Komi republic (raw data).

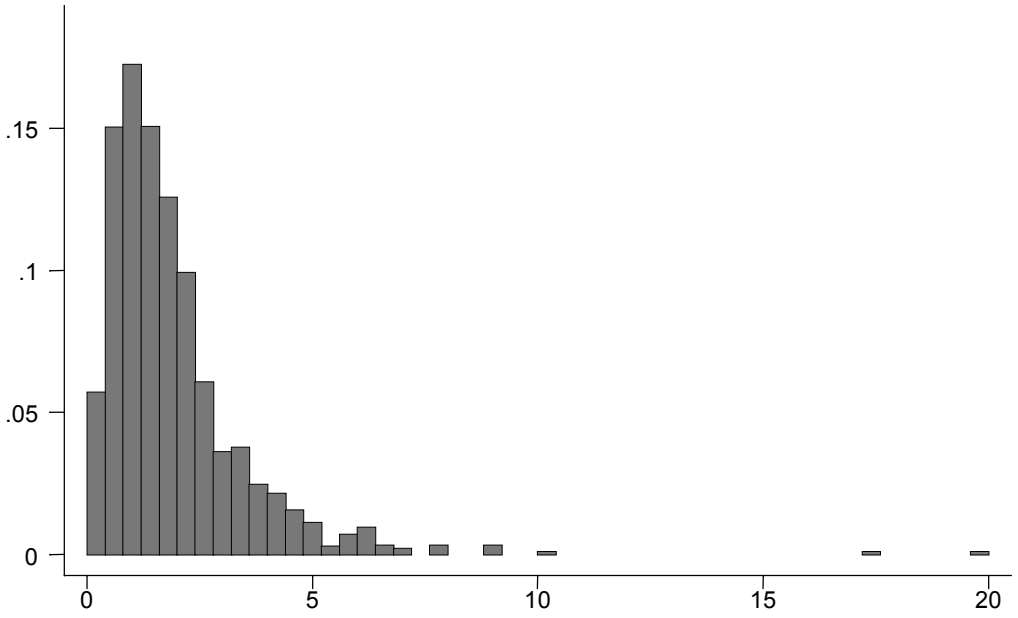

Fig. A.2b. The histogram of the per capita expenditure distribution of population of Komi republic (weighted data).

Table A.2. Sample statistics of the essential characteristics of Komi republic per capita expenditure distribution.

	Indicator (ths. rub.)	Sample value	
		Raw data	Weighted data
1	Mean per capita expenditure $(\hat{\mu})$	0,633	0,686
2	Standard deviation (S)	1,087	1,249
3	Minimal expenditure $(x_{\min})$	0,054	0,054
4	Maximal expenditure $(x_{\max})$	24,797	24,797
5	Bottom decile $(\hat{x}_{0,1})$	0,154	0,163
6	Top decile $(\hat{x}_{0,9})$	1,208	1,302

A1.3. Volgograd oblast

Budget survey sample of 1263 individuals, Q2 1998.

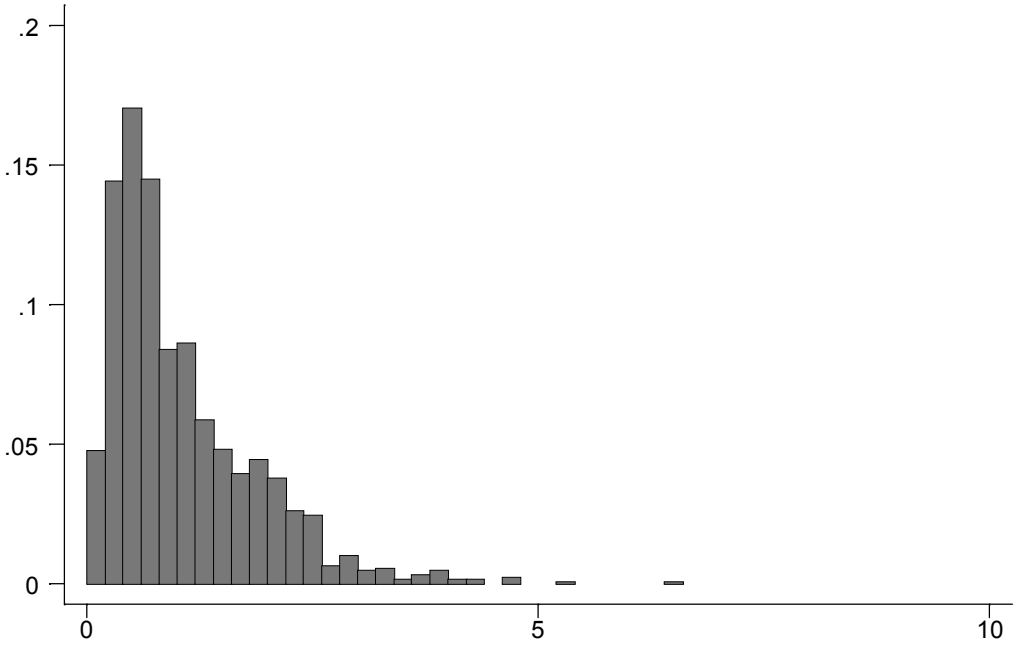

Fig. A.3a. The histogram of the per capita expenditure distribution of population of Volgograd oblast (raw data).

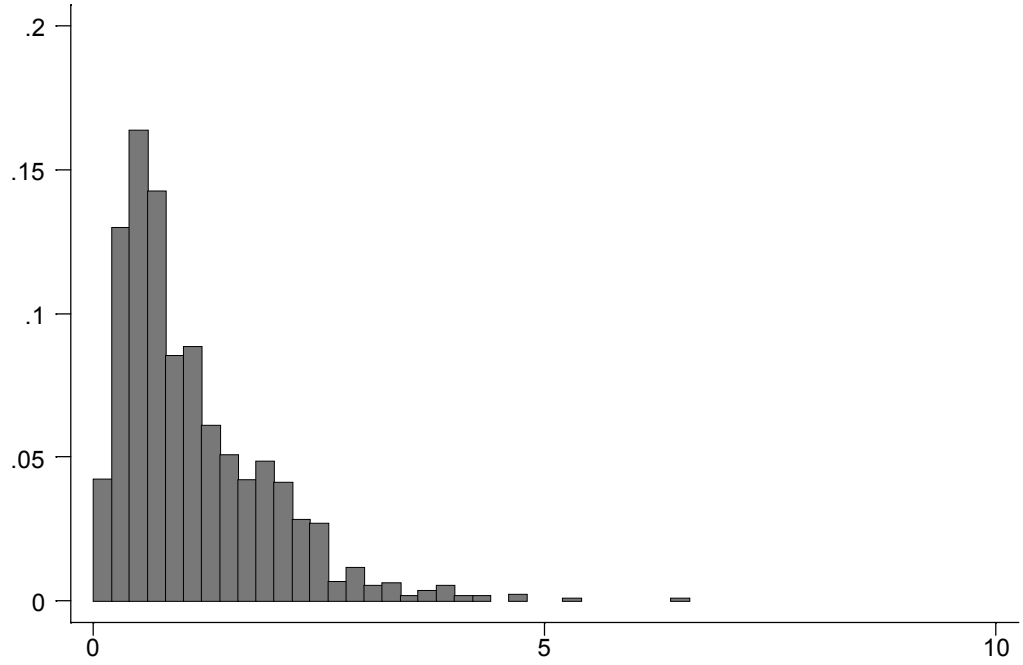

Fig. A.3a. The histogram of the per capita expenditure distribution of population of Volgograd oblast (raw data).

Table A.3. Sample statistics of the essential characteristics of Volgograd oblast per capita expenditure distribution.

	Indicator (ths. rub.)	Sample value	
		Raw data	Weighted data
1	Mean per capita expenditure $(\hat{\mu})$	0,412	0,433
2	Standard deviation (S)	0,458	0,479
3	Minimal expenditure $(x_{\min})$	0,017	0,017
4	Maximal expenditure $(x_{\max})$	6,101	6,101
5	Bottom decile $(\hat{x}_{0,1})$	0,101	0,110
6	Top decile $(\hat{x}_{0,9})$	0,766	0,794

A1.4. Omsk oblast

Budget survey sample of 1244 individuals, Q2 1998.

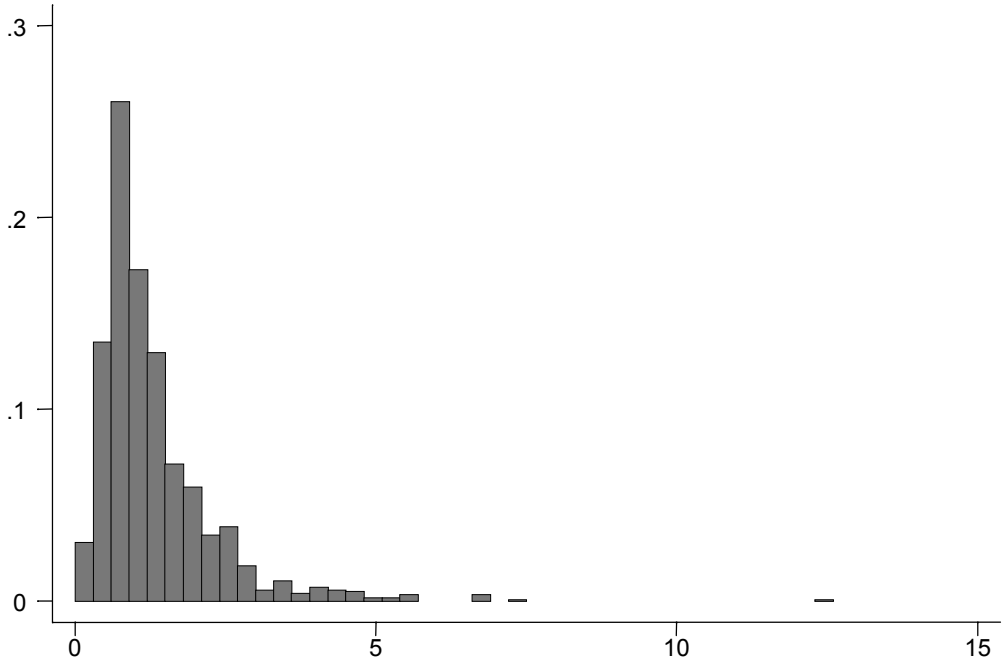

Fig. A.4a. The histogram of the per capita expenditure distribution of population of Omsk oblast (raw data).

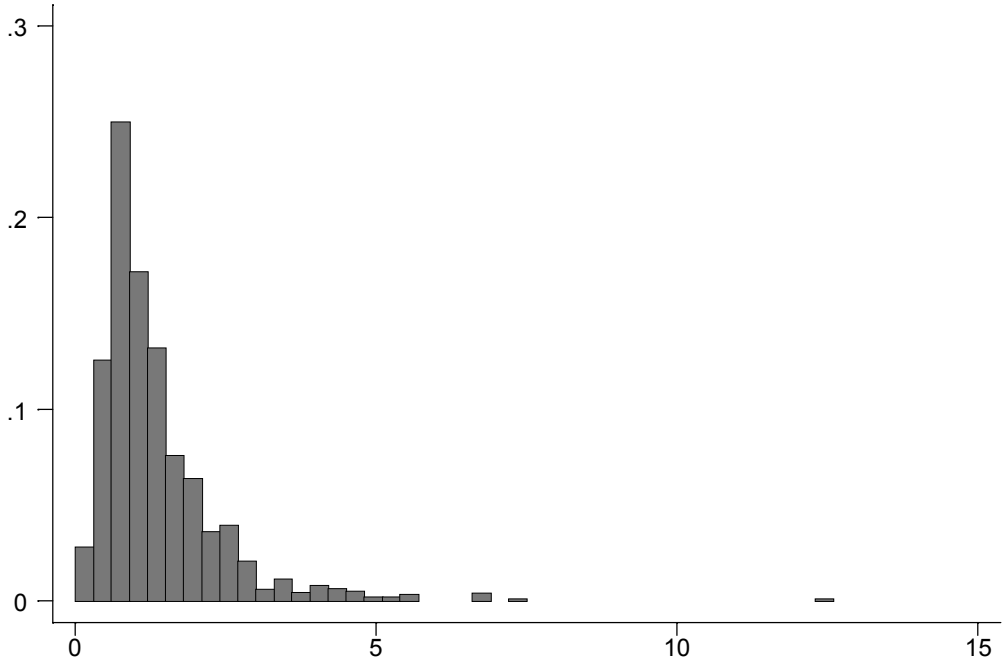

Fig. A.4b. The histogram of the per capita expenditure distribution of population of Omsk oblast (weighted data).

Table A.4. Sample statistics of the essential characteristics of Volgograd oblast per capita expenditure distribution.

	Indicator (ths. rub.)	Sample value	
		Raw data	Weighted data
1	Mean per capita expenditure $(\hat{\mu})$	0,611	0,641
2	Standard deviation (S)	0,708	0,761
3	Minimal expenditure $(x_{\min})$	0,034	0,034
4	Maximal expenditure $(x_{\max})$	11,809	11,809
5	Bottom decile $(\hat{x}_{0,1})$	0,160	0,163
6	Top decile $(\hat{x}_{0,9})$	1,211	1,238

APPENDIX 2. THE ESTIMATION RESULTS FOR THE MIXTURE MODEL IN THE OBSERVED PER CAPITA EXPENDITURE RANGE

A2.1. The methodology of estimation

In this section, the methods of statistical estimation of the mixture model by EM algorithm and its modification will be described. The problem is to estimate the vector of parameters

$$\Theta(k) = \left(\widetilde{q}_1,\ldots,\widetilde{q}_k;\ a_1,\ldots,a_k;\ \sigma_1^2,\ldots,\sigma_k^2\right) \tag{A.1}$$

of the density function (П.1) (A

$$\widetilde{\varphi}_k\left(z|\Theta\right) = \sum_{j=1}^{k} \widetilde{q}_j \varphi\left(z|a_j;\ \sigma_j^2\right) \tag{A.2}$$

by using the random sample (8') data via maximum likelihood when the number of components k is fixed. Here, $\varphi(z|a_j;\ \sigma_j^2)$ is the density function of a normal distribution with mean a_j and variance σ_j^2. I.e., the problem is to find such

$$\hat{\Theta}(k) = \left(\hat{\widetilde{q}}_1,\ldots,\hat{\widetilde{q}}_k;\ \hat{a}_1,\ldots,\hat{a}_k;\ \hat{\sigma}_1^2,\ldots,\hat{\sigma}_k^2\right), \tag{A.3}$$

that the log likelihood function ((

$$l_k(\Theta(k)) = \sum_{i=1}^{n} \omega_i \left[\ln \sum_{j=1}^{k} \widetilde{q}_j \varphi\left(z_i|a_j;\sigma_j^2\right)\right] \tag{A.4}$$

would attain its maximum over θ:

$$\hat{\Theta}(k) = \arg\max_{\Theta(k)} l_k(\Theta(k)) \tag{A.5}$$

In (A.4), z_i are the sample (observed) values, ω_i, the weights assigned to the observations by (11'), and n, the sample size.

Iterative EM (Expectation – Maximization) algorithm solves the problem (A.5) in the following way.

(i) Log likelihood function (A.4) is decomposed as

$$l_k(\Theta(k)) = \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij} \ln \tilde{q}_j + \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij} \ln \varphi(z_i | a_j; \sigma_j^2) - \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij}\,, \tag{A.6}$$

where

$$g_{ij} = \frac{\tilde{q}_j \varphi(z_i | a_j; \sigma_j^2)}{\tilde{\varphi}_k(z_i | \Theta(k))} \tag{A.7}$$

are the *a posteriori* probabilities to have observed the class j conditionally on the observed z_i.

(ii) The expectation step is to calculate, by using (A.7), the $g_{ij}^{(t)}$ conditionally on the parameters estimates

$$\hat{\Theta}^{(t)}(k) = \left(\hat{\tilde{q}}_1^{(t)}, \ldots, \hat{\tilde{q}}_k^{(t)};\ \hat{a}_1^{(t)}, \ldots, \hat{a}_k^{(t)};\ \left(\hat{\sigma}_1^2\right)^{(t)}, \ldots, \left(\hat{\sigma}_k^2\right)^{(t)} \right) \tag{A.8}$$

obtained at t-th iteration. The $g_{ij}^{(t)}$ are then plugged into (A.6) as estimates of g_{ij}.

(iii) The maximization step is to maximize over $\hat{\Theta}^{(t)}(k)$ with fixed $g_{ij}^{(t)}$ the log likelihood

$$l_k\left(\hat{\Theta}^{(t)}(k)\right) = \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij}^{(t)} \ln \hat{\tilde{q}}_j^{(t)} + \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij}^{(t)} \ln \varphi\left(z_i | \hat{a}_j^{(t)}; (\hat{\sigma}_j^2)^{(t)}\right) - \sum_{i=1}^{n} \omega_i \sum_{j=1}^{k} g_{ij}^{(t)} \tag{A.9}$$

The solutions are:

$$\begin{aligned}
\hat{\tilde{q}}_j^{(t+1)} &= \sum_{i=1}^{n} \omega_i g_{ij}^{(t)}, \\
\hat{a}_j^{(t+1)} &= \frac{1}{\hat{\tilde{q}}_j^{(t+1)}} \sum_{i=1}^{n} \omega_i g_{ij}^{(t)} z_i, \\
\left(\hat{\sigma}_j^2\right)^{(t+1)} &= \frac{1}{\hat{\tilde{q}}_j^{(t+1)}} \sum_{i=1}^{n} \omega_i g_{ij}^{(t)} \left(z_i - \hat{a}_j^{(t+1)}\right)^2, \\
& j = 1,2,\ldots,k.
\end{aligned} \tag{A.10}$$

Here, the iteration ends, and the expectation step is repeated with the updated $\hat{\tilde{q}}_j^{(t+1)}, \hat{a}_j^{(t+1)}$ and $(\hat{\sigma}_j^2)^{(t+1)}$ $(j = 1,2,\ldots,k)$. [32] and later works[4] prove, under some general assumptions, of which the most restrictive one being the requirement of bounded log likelihood, that EM algorithms have some useful properties. In particular, they converge in probability to the solution of (A.5).

Some technical modifications of this general scheme were used in our study. The observations z_i were given weights ω_i. Also, a background cluster was used at the early stages of the algorithm to account for the insufficient number of components. Roughly speaking, the data points in this background cluster are supposed to be uniformly distributed over the whole range of observed values. Detailed description of the EM algorithm version implemented in CLASSMASTER software can be found in [35].

Let us now turn to the problem of estimation of the number of components k that was supposed to be known in the above procedures. In other words, the question is to be asked, what the number of components that can be reliably discovered in the data (per capita expenditure) is.

The procedure of the k estimation is to sequentially test simple nested hypotheses

H_0: $k = j$

against the alternative

H_1: $k = j+1$, $-j = 1,2,\ldots,$ —

by using the standard likelihood ratio statistic

$$\gamma(j) = -2\ln\frac{l_j(\hat{\Theta}(j))}{l_{j+1}(\hat{\Theta}(j+1))}.$$

The first value $j = \hat{k}$ such that the hypothesis H_0 is not rejected was taken to be the estimate of the number of components in (A.2). This procedure was supplemented by the technique of the number of clusters estimation via projection pursuit [33].

There are however other options to proceed. One of them is to use information criteria instead of likelihood ratio tests. In this framework, the model is preferred which has the optimal value of information criteria (such as Akaike information criteria or ICOMP information complexity index) that serves as an estimate of the amount of information captured to the model as opposed to its dimension. Another way to choose the "best" model is to use goodness of fit criteria (e.g. χ^2) to test whether the model distribution function resembles the sample CDF. The range of the observed values is divided into m bins (it is recommended that the number of these bins be $\log_2 N$ where N is the total sample size), and the theoretical frequency is confronted with the empirical one. It is known that the distribution of the test statistics is asymptotically $\chi^2(m\text{-}p\text{-}1)$ where m is the number of bins and p is the number of parameters to be estimated.

In parallel to the modified EM-algorithm as implemented in CLASSMASTER software, a Stata program was developed that performs maximum likelihood estimation by using built-in Stata **`ml`** maximizer [44]. Stata maximization algorithm can be described as follows.

1 feasible initial values are found by random search, if reasonable starting values are not provided externally by the user;
2 search for the better values is performed in the neighborhood of the feasible starting values;
3 unidimensional optimization is performed for each of the model parameters;

---

[4] The general framework of the algorithms that later were given the name "EM-algorithms" seems to have been pioneered by M. Shlesinger in *Шлезингер М.И.* О самопроизвольном различении образов. — «Читающие автоматы», Киев, Наукова думка, 1965, с. 38—45. The properties of these algorithms were also studied there. This work is however not easily accessible in the West and thus it is not known among the Western statisticians.

4 multidimensional iterative optimizer is launched:
   4.1 the log likelihood derivatives of the first and second order are found numerically;
   4.2 if the log likelihood is found to be concave, Newton-Raphson step is performed;
   4.3 otherwise, gradient based steepest ascent method is used.
5 the algorithm terminates if any of the following happens:
   5.1 log likelihood has stabilized (by default, the change at the last iteration less than 10^{-6});
   5.2 the estimates of the coefficients have stabilized (by default, relative change less than 10^{-7});
   5.3 the gradient of the log likelihood is small enough (the value 10^{-3} is used in some of the program runs);
   5.4 too many iterations are performed (by default, 16000. Some runs resulted in 3000+ iterations which took about a day to compute on Pentium II 333 MHz 128 M RAM, in parallel with a couple of other Stata sessions);
   5.5 critical error is issued if Stata cannot calculate the numerical derivatives. It might happen if there is a plateau, a sharp pike or a sharp (multidimensional) ridge of the log likelihood.

If the maximization was successful (in terms of the above criteria), Stata outputs the table of the coefficient estimates along with their standard deviations and confidence intervals. Some other statistics were added to the output such as goodness of fit tests (information criteria AIC, ICOMP, and χ^2 test), as well as the inequality and poverty indices computed for the current mixture model. According to the above stated hypothesis H_3 (or, rather, H_3' as in section 4.1), the estimation is performed under a simplifying constraint $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$ where $\sigma_j^2 = Var(\ln \xi)$.

A2.2. Estimation results

The estimation results for the RLMS Round VIII data as well as for the regional data sets (Komi Republic, Volgograd and Omsk oblasts) are reported in Tables A.5–A.8.

Table A.5. The results of maximum likelihood estimation of the normal mixture parameters for the log of per capita expenditure (Russia).

Number of mixture compo-nents (k)	Log likeli-hood at maximum	Akaike criterion (AIC)	Bozdo-gan's ICOMP index	Goodness of fit $\chi^2(N)$ test	p-value	$\hat{\sigma}^2$	$\hat{a}_1$	$\hat{\tilde{q}}_1$ %	$\hat{a}_2$	$\hat{\tilde{q}}_2$ %	$\hat{a}_3$	$\hat{\tilde{q}}_3$ %	$\hat{a}_4$	$\hat{\tilde{q}}_4$ %	$\mu_{.\text{model}}$ ths. rub.
1	-4244,8	8489,5	8490,0	53,23	$< 5 \cdot 10^{-5}$	0,790	6,400	100	—	—	—	—	—	—	0,893
2	-4222,6	8453,1	8461,7	34,35	$3 \cdot 10^{-4}$	0,722	6,430	98,71	4,118	1,29	—	—	—	—	0,879
3	-4192,4	8384,8	8398,0	13,72	0,248	0,600	6,412	95,60	4,397	2,57	8,578	1,83	—	—	0,920
4	**-4189,1**	**8378,0**	**8393,0**	**12,12**	**0,354**	**0,538**	**6,376**	**91,09**	**4,469**	**3,19**	**7,759**	**5,44**	**9,816**	**0,27**	**0,937**

***Four** components model is selected*

Table A.6. The results of maximum likelihood estimation of the normal mixture parameters for the log of per capita expenditure (Komi Republic)

Number of mixture components (k)	Log likelihood at maximum	Akaike criterion (AIC)	Bozdogan's ICOMP index	Goodness of fit $\chi^2(N)$ test	p-value	$\hat{\sigma}^2$	$\hat{a}_1$	$\hat{\tilde{q}}_1$ %	$\hat{a}_2$	$\hat{\tilde{q}}_2$ %	$\hat{a}_3$	$\hat{\tilde{q}}_3$ %	$\hat{a}_4$	$\hat{\tilde{q}}_4$ %	$\mu_{.\text{model}}$ ths. rub.	$\hat{\tilde{q}}_5$ %	$\hat{\mu}_{\text{мод.}}$ (тыс. руб.)
1	-1313,74	2627,48	2627,78	32,08	0,000	0,654	-0,842	—	—	—	—	—	—	—	—	—	0,598
2	-1302,37	2604,75	2615,06	16,54	0,085	0,610	2,114	0,50	-0,857	99,50	—	—	—	—	—	—	0,630
3	-1300,69	2601,38	2613,91	23,67	0,009	0,475	-0,162	22,88	-1,058	76,79	2,527	0,33	—	—	—	—	0,636
4	**-1299,34**	**2598,69**	**2615,72**	**9,77**	**0,461**	**0,285**	**-1,976**	**8,99**	**0,010**	**25,41**	**-1,038**	**65,17**	**2,338**	**0,43**	**—**	**—**	**0,628**
5	-1294,49	2588,97	2609,11	5,90	0,824	0,168	0,890	3,07	-1,082	55,02	2,756	0,27	-0,115	29,47	-2,029	12,17	0,633

***Four** components model is selected.*

Table A. 7. The results of maximum likelihood estimation of the normal mixture parameters for the log of per capita expenditure (Volgograd oblast).

Number of mix-ture compo-nents (k)	Log likeli-hood at maximum	Akaike criterion (AIC)	Bozdo-gan's ICOMP index	Good-ness of fit $\chi^2(N)$ test	p-value	$\hat{\sigma}^2$	$\hat{a}_1$	$\hat{\tilde{q}}_1$ %	$\hat{a}_2$	$\hat{\tilde{q}}_2$ %	$\hat{a}_3$	$\hat{\tilde{q}}_3$ %	$\hat{a}_4$	$\hat{\tilde{q}}_4$ %	$\mu_{.\text{model}}$ ths. rub.	$\hat{\tilde{q}}_5$ %	$\hat{\mu}_{\text{мод.}}$ (тыс.руб.)
1	-1587,56	3175,11	3175,36	43,63	0,000	0,723	-1,259	100	—	—	—	—	—	—	—	—	0,408
2	-1586,26	3172,52	3185,81	42,78	0,000	0,673	0,116	2,58	-1,295	97,42	—	—	—	—	—	—	0,414
2	-1587,49	3174,98	3191,68	42,72	0,000	0,716	-1,254	99,77	-3,044	0,23	—	—	—	—	—	—	0,407
3	-1585,25	3170,50	3191,55	37,53	0,000	0,577	-2,456	4,10	-1,280	90,57	0,021	5,34	—	—	—	—	0,414
4	-1573,11	3146,23	3157,29	32,694	0,000	0,180	0,613	4,10	-2,833	7,00	-0,647	38,56	-1,661	50,34	—	—	0,413
5	**-1568,44**	**3136,88**	**3149,21**	**12,03**	**0,283**	**0,099**	**-2,943**	**6,16**	**-1,927**	**28,58**	**-,481**	**27,89**	**0,650**	**4,21**	**-1,266**	**33,16**	**0,411**

Five components model is selected

Table A. 8. The results of maximum likelihood estimation of the normal mixture parameters for the log of per capita expenditure (Omsk oblast).

Number of mix-ture compo-nents (k)	Log likeli-hood at maximum	Akaike criterion (AIC)	Bozdo-gan's ICOMP index	Good-ness of fit $\chi^2(N)$ test	p-value	$\hat{\sigma}^2$	$\hat{a}_1$	$\hat{\tilde{q}}_1$ %	$\hat{a}_2$	$\hat{\tilde{q}}_2$ %	$\hat{a}_3$	$\hat{\tilde{q}}_3$ %	$\hat{a}_4$	$\hat{\tilde{q}}_4$ %	$\mu_{.\text{model}}$ ths. rub.	$\hat{\tilde{q}}_5$ %	$\hat{\mu}_{\text{мод.}}$ (тыс. руб.)
1	-1503,17	3006,34	3006,63	23,83	0,008	0,656	-0,838	—	—	—	—	—	—	—	—	—	0,600
2	-1499,93	2999,85	3013,09	13,75	0,184	0,602	0,671	2,34	-0,875	97,66	—	—	—	—	—	—	0,612
2	-1497,50	2995,00	3002,55	25,63	0,004	0,591	-2,915	1,48	-0,807	98,52	—	—	—	—	—	—	0,592
3	**-1482,10**	**2964,20**	**2972,02**	**13,47**	**0,198**	**0,382**	**-0,960**	**84,72**	**-2,911**	**2,20**	**0,294**	**13,09**	**—**	**—**	**—**	**—**	**0,607**
4	-1479,72	2959,44	2979,77	23,30	0,010	0,351	2,192	0,18	0,165	16,95	-2,908	2,28	-0,998	80,59	—	—	0,613
4	-1481,45	2962,90	2977,45	14,77	0,141	0,278	-3,003	2,06	-1,260	46,09	0,548	8,67	-0,563	43,18	—	—	0,606
5	-1476,16	2952,33	2970,11	18,42	0,048	0,211	-3,047	2,01	0,436	11,87	-1,433	34,27	-0,662	51,68	2,302	0,18	0,612

Five components model is selected.

APPENDIX 3. PROBABILITY OF HOUSEHOLD REFUSAL TO PARTICIPATE IN A SURVEY AS A FUNCTION OF ITS CHARACTERISTICS

In this section, the results of the analysis of the logit model for the unit non-response probability conditional on social and economic characteristics of the household are reported. The definition of the model of the dependence of the probability (p) to refuse to participate in a survey on the log of the household per capita expenditure $(z^{(1)})$, settlement type $(z^{(2)})$ and the primary income earner education $(z^{(3)})$ is written down in section 4.4.1.

RLMS panel data were used to study the probability of a household to refuse to participate in a sociological survey. For each of the 4718 households in the RLMS sample (Rounds V-VIII), interviewers wrote down whether the household participated in the survey, and, if not, why. The codes registered (i.e., most typical responses) are reproduced in the Table A.9.

Table A.9. Visit result codes

01	Survey conducted
	Objective failure reasons
02	Uninhabited premises
03	No one lives in the house (apartment) at the moment
04	Apartment cannot be reached
05	Apartment is rented by foreigners
06	No one is at home
07	They neither open the door nor communicate
08	Survey impossible because of illness
09	Survey impossible because of handicap
10	No adults at home
11	Person opened the door is drunk
14	Family is absent during the whole period of the survey
15	Family is present only late in the evenings
16	Family actually lives at another location
18	Other
	Refusals
30	*Refused to participate*
	Communication circumstances
21	Refusal with the door closed
22	Refusal of the person opened the door
23	Refusal of the respondent
24	Refusal of another family member
25	Refusal when being interviewed
26	Refusal with lies
27	Action against interviewer
28	Other
	Refusal reasons
41	Unmotivated refusal
42	“Too busy”
43	“Have no time”
44	“I never open the door”
45	“These surveys change nothing”
46	“Don’t want to tell about my life to anyone”
47	“I have a right not to answer”
48	“I want to have rest”
49	“I do not want to be in a computer”
50	“Participated in a sociological survey recently”
51	“We are temporarily here”
52	Family reasons
53	Not interested in the survey topic
54	Bored with politics
55	Refusal out of protest
56	Reluctant to release information on political views
57	*Reluctant to release information on family welfare level*
58	Do not trust the interviewer
59	Other

Table A.10 reports the refusal rates in Rounds V–VIII.

Table A10. Rates of refusal to participate in the survey

	Round 5	Round 6	Round 7	Round 8
Survey not conducted	743	963	1118	1254
Number of refusals	410	539	489	701
Refusal because of unwillingness to inform about family welfare			17	19
Survey conducted	3973	3781	3750	3831

Source: RLMS data, additional RLMS refusal data, authors' calculations.

The final goal of the analysis is the answer to the question: "Does the probability to refuse to participate in a sociological survey depend on the welfare and other characteristics of the household?" or, in more general form, "Is truncation random?" By using the above data on refusals combined with appropriate household data on expenditure level and settlement type in the RLMS household data, and individual incomes and education in the RLMS individual data, a binary dependent variable econometric model for unit non-response probability (6) can be estimated.

Apparently, if the household had refused to participate in the survey in a given round, the data on its expenditure are not observable. However, as the data we use are of panel type, the same households have been visited, and information from other rounds can be used to assess the level of welfare of this household. Here we assume that the welfare is approximately constant over time. This assumption may be subject to critique as long as income mobility is often considered to be high (e.g. [40]). We think however that income mobility does not crucially affect our analysis. The within-unit (between years) variance of log expenditure ranges from 0.018 to 1.32 (thus, the level changes are between 2% and a factor of 3.7). The average variance is 0.25, and the variance is 0.21, so that the magnitude of the expenditure fluctuations is about 25%.

To adjust for income mobility, we use the average, for all available years, log of appropriately deflated expenditure[5] to smooth out these fluctuations. The analogy can be drawn here to Friedman's lifecycle permanent income hypothesis [41]. Experimentation with other welfare

[5] The deflator from Russian Economic Trends is used in RLMS to make nominal figures comparable across years. The figures indicated as "real" in the (derived) RLMS data are to be interpreted as "in 1992 prices".

measures such as median of expenditure for available years, imputed expenditures[6], or principal components did not affect results qualitatively, and even the estimates of the coefficients were quite alike. We report results of the logit model estimation for both mean and median of log expenditures as a covariate of the unit non-response probability. It is the mean log of expenditures that would be used in application of the logit model to the distribution calibration, as a clearly interpretable characteristic.

The basic RLMS variables used for the analysis of the refusal probability were per capita expenditures deflated to the same period (1992 prices), namely, **totexpr*** ; settlement type, or urbanization level of the household residence $(z^{(2)})$; and the education level of the primary income earner ($z^{(3)}$). The dependent variable η here is the indicator whether the household has ever refused to participate in RLMS. The analysis of the indicator that the household reported reluctance to provide information on income as a dependent variable was also performed. We did not find it relevant to report the results, however, as this category of refusals is not numerous (29 out of 4239, i.e. about 0.5%), while the logit model is known to perform well if the share of successes is within 10–90% range. The situation is satisfactory for the "all reasons" formulation with this respect as its share in the total number of households ever participated in RLMS is 795/4239 = 18.8%.

The estimates for several logit model specifications are reported in Table A.11. Along with the mean expenditure, urbanization level and the head education level dummies are used in the analysis. The base category for the settlement type is "city" (denoted as U in the graph below); other categories include "metropolitan areas" (M), "town-type settlement" (P), and "rural area" (R). The educational categories are based on the accumulative scheme. The base category is "education lower than secondary" (L); the dummy for secondary education (S) measures the difference between those two. The vocational school (P) and technical school (T) dummies

[6] Stata software has a built-in routine for imputation by using (a set of) linear regression models, in our case, for the household expenditure. For each pattern of the missing data, the most comprehensive regression model is estimated, and then prediction for the missing data is performed [42], [43]. In other words, for each missing value of interest, a regression model with all non-missing in this observation variables is constructed and estimated with the data available, and prediction is made that serves as an estimate of the mean of the missing variable conditional on all other observed characteristics. It should be noted, however, that if imputed values are then used as regressors, the estimates of the corresponding coefficients tend to be biased (usually, towards zero) which is a known effect of measurement error.

do not rule out the possibility of having secondary education (moreover, those schools base on the secondary education), so the respective coefficients measure the difference of those two categories from the secondary education. Finally, the high education category (H) covers all other educational categories, in the sense that one can go to the university after secondary, vocational, or technical school. So the interpretation of the coefficient is what difference does it make to have a higher school diploma.

For further calibration, model (4) is used that have the highest LR per one degree of freedom.

Table A.11. The estimates of the multivariate logit model for the survey refusal probability.

	(1)	(2)	(3)	(4)
Median expenditure	0.396 (0.084)**	0.355 (0.075)**		
Mean expenditure			0.429 (0.089)**	0.399 (0.079)**
Metropolitan areas (M)		1.052 (0.206)**		1.043 (0.203)**
Rural areas (R)		-1.583 (0.292)**		-1.576 (0.291)**
Town-type settlement (P)		-0.876 (0.310)**		-0.878 (0.308)**
Secondary education (S)		-0.862 (0.156)**		-0.868 (0.156)**
Vocational school (P)		-1.826 (0.184)**		-1.825 (0.182)**
Technical school (T)		-1.268 (0.212)**		-1.277 (0.213)**
Higher education (H)		-0.857 (0.142)**		-0.880 (0.142)**
Constant	-4.532 (0.653)**	-3.140 (0.588)**	-4.788 (0.691)**	-3.464 (0.632)**
No. of observations	4239	4239	4239	4239
Wald test (d.f.) p-value	Wald(1)= 22.05 0.00	Wald(8)= 317.86 0.00	Wald(1)= 23.39 0.00	Wald(8)= 334.78 0.00

Source: RLMS data, additional RLMS refusal data, authors' calculations.
Standard errors corrected for clusterisation on PSU (sample stratification) are in parentheses.
* denotes significance on 5% level, **, on 1% level.

The predicted values of the refusal probability are shown on figure A.5 for several household categories. The horizontal axis is the log scale of the deflated expenditure. As there are 4 geographical and 5 educational categories, the total number of partial logistic curves for each combination of the dummy variables should be 20. Drawing all them on the same graph is likely to hamper readability, so the graph shows several most populated and representative of them.

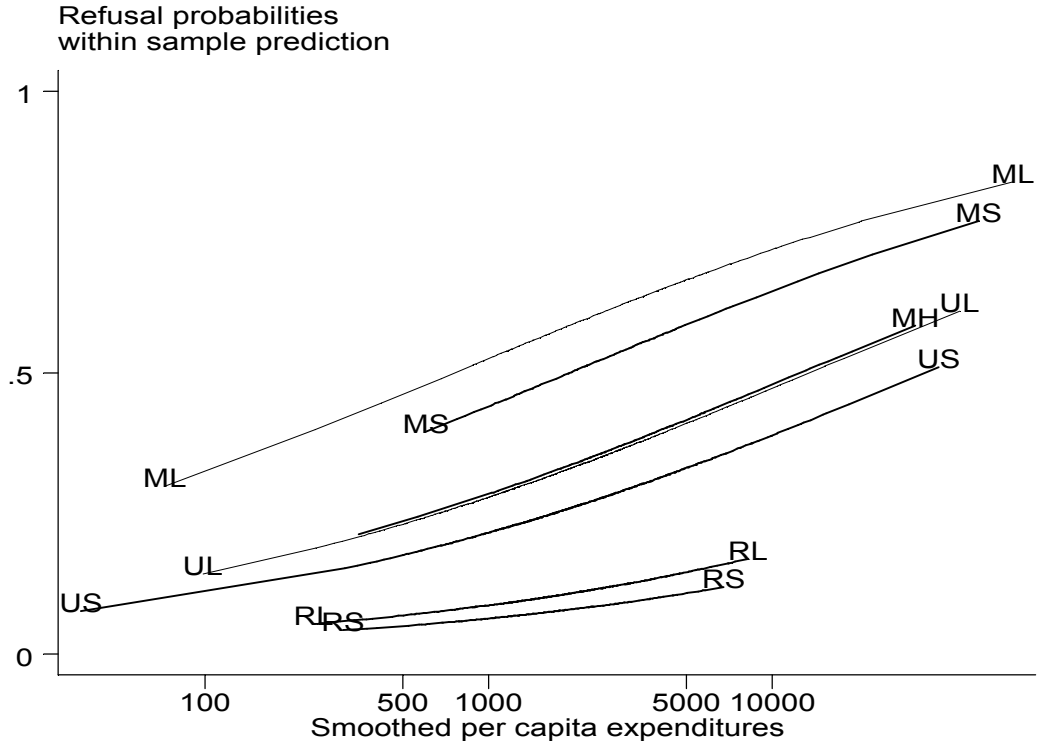

Fig. A.5. The family of the curves describing the dependence of the refusal probability on the mean expenditure, in 1992 rubles.

The results obtained in this section, though interesting *per se*, are only used to calculate the household weights to adjust for truncation bias. A bivariate model was used in the interim report linking the refusal probability with the mean expenditure only. As there is an apparent improvement in the log likelihood of the model due to introduction of the additional covariates, the precision of weighting should improve as compared to the one that uses the bivariate model. The fact that all confidence intervals for the welfare proxy (mean or median expenditure) overlap for all four reported models can be considered as an additional evidence for the strong and consistently verified link between the level of welfare and propensity to disclose the information on individual or household wealth to the third parties.