Economics Education and Research Consortium
Working Paper Series

Poverty and Expenditure Differentiation of the Russian Population

Serguei Aivazian
Stanislav Kolenikov

Working Paper No 01/01

This project (No 99-113) was supported
by the Economics Education and Research Consortium

Research area: **Labor Markets and Social Policy**

JEL Classification: C13, C15, D31, I32, P29

Aivazian S.A., Kolenikov S.O. Poverty and Expenditure Differentiation of the Russian Population — Moscow: EERC, 2001. — pp 1–63.

The problem of poverty and inequality measurement in contemporary Russian society is considered within the framework of the general problem of social tension reduction via efficient organization of the social assistance system. We argue that features specific to Russian transition stipulate that poverty indicators (*e.g.*, Foster-Greer-Thorbecke family) be calculated on the basis of *expenditure* rather than income as it is usually done. These features are also accounted for in the 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 regions.

Acknowledgements. The authors are grateful to John Earle, Michel Sollogoub, Michael Beenstock, Anthony Shorroks, Constantin Colonescu, Tom Mroz, and Klara Sabirianova for substantial comments and references; to all participants of EERC workshops for useful discussions; to Elena Frolova for discussions and help with data collection; and to Carolina Population Center and Polina Kozyreva for access to the RLMS data. Support to this project has been extended by the Russian Humanitarian Scientific Foundation (Grant №99-02-00270) and by the Moscow Public Science Foundation out of the funds provided by the U.S. Agency for International Development-USAID (Grant №020/1-01-O). Views reflected in this document and by the authors may not coincide with the views of the U.S. Agency for International Development or the Moscow Public Science Foundation.

Keywords: Russia, economic inequality, per capita expenditure distribution of population, indicators of poverty, survey, unit non-response (truncation), censoring, misreporting, mixture model, optimal allocation of social assistance, transition, missing data.

Serguei Aivazian, Stanislav Kolenikov

Central Economics and Mathematics Institute RAS

47 Nakhimovsky pr., 117418 Moscow, Russia.

Tel. +7 (095) 129 13 00

Fax: +7 (095) 719 96 15

E-mail: aivazian@cemi.rssi.ru, skolenik@yahoo.com

CONTENTS

NON-TECHNICAL SUMMARY

Upon what kind of information should a government base its decisions concerning the country's social policies? One of the main guidelines when selecting among different political schemes is the knowledge of the population's distribution according to per capita income and expenditure. Certain key indicators of the degree of success of a government's policy may be computed from the sample distribution of these two variables. Some of the well-known empirical measures that relate to the potency of the state's social policy are the proportion of the poor, the Foster–Greer–Thorbecke indices, the Gini index, and the funds ratio, all of which show the extent and the degree of poverty and social inequality. The values of these descriptive measures may be established based on the two quantitative distributions mentioned above.

The methodology of evaluating these indicators employed by official Russian statistical agencies has proved to produce utterly incoherent results. The methods of estimation used in the framework of developed western economies as well as the methods proposed by various Russian authors also fail to give an adequate picture of our home situation. There are two principal reasons why these approaches fail to work.

First of all, the established methods neglect the *recent radical changes in the socioeconomic stratification of Russian society*, which include a reduction in the size of the middle class, a significant increase of the proportion of very poor as well as very rich, and the rapid change of the whole spectrum of social classes in Russia. The second reason for the bias is the *non-representativity of the samples customarily used for estimation*. Virtually all households in the highest income group avoid interviews with regard to their financial standing, therefore keeping a certain part of the information about the distribution masked (which technically speaking results in *censoring* of the sample data). Other categories of households also show increasing tendency to refrain from participating in relevant surveys (that leads to the so-called *truncation effect*). A large number of households understate their income since at least part of it comes from shady sources (which results in *misreporting* errors).

In this work we've attempted to account for these peculiarities of the data gathering process. Our primary objective was to reduce the bias imposed on the distribution of the Russian population by the level of wealth (poverty). In order to compensate for this bias, we apply a certain calibration procedure. This procedure involves the following basic concepts:

- the above-mentioned indicators of poverty (prosperity) and indicators of inequality should be computed from the distribution of population expenditure rather than income;
- instead of the traditional lognormal model (conventionally used to describe such kinds of distributions), we use a mixture of lognormal distributions, which includes a component describing the expenditure distribution within the latent (unobservable) strata of the highest-income households (the *super-rich*);
- the probability of refusal to enter a survey is incorporated into the model. This probability is represented as a function of the household's per capita expenditure, place of residence, and the education level of the head of the household (or the particular respondent).

The purpose of our method is to adjust, re-weight, and calibrate the available distributions. We illustrate our theoretical reasoning with an application of the data provided by the Russia Longitudinal Monitoring Survey V – VIII (a state-wide survey). We also use our technique to analyze the sample data obtained in other surveys in three regions of the Russian Federation (Komi, Volgograd, and Omsk) pertaining to the second quarter of 1998. Here are some important points that we want to make with reference to these applications and their results. The largest discrepancies between our method and the official statistics are observed in the analysis and interpretation of measures of social differentiation like the Gini index and funds ratio. Particularly, our method estimates the state Gini index and funds ratio to lie in the range of 0.55 – 0.57 and 36 – 39, correspondingly, whereas official statistics claims their point estimates to be 0.38 and 13.5. The estimates of population poverty and inequality indicators obtained by the application of our model have shown robustness with respect to relaxation or modification of the model assumption.

The newly obtained estimates of the population poverty and inequality indicators such as the Foster–Greer–Thorbecke indices may be considered with regard to possible alterations in the government's social policy. One of the policy implications of our research is formulated in terms of a rule for optimal distribution of social aid to the *long-term poor* class in Russian society.

1. INTRODUCTION

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 long-run aim to reduce social tension in the society.

The indicators are based on household budget survey data estimation procedures used nowadays by Russian statistical agencies (Proceedings of Goskomstat R.F., 1999a; Velikanova *et al*., 1999; Velikanova and Frolova 1999), as well as those proposed by other researchers (Shevyakov and Kiruta, 1999; Yershov and Maier, 1998; Suvorov and Ul'anjva, 1997) are burdened by certain drawbacks, even after correcting for the macroeconomic balance of the population's income and expenditure and/or equivalence scales. These drawbacks yield significant distortions in the values of the appropiate characteristics.[1]

We see the following reasons to explain those distortions:

(i) The specific features of Russia's transition economy suggest that **expenditure** rather than *income should* be used for the purposes of poverty and inequality evaluation as well as for dichotomizing 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; and

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

(ii) The two-parameter lognormal income distribution model used by the statistics (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

[1] Some estimations (e.g., Velikanova and Frolova, 1999; Suvorov, and Ulyanova, 1997; Aivazian, 1997) show that the ratio of the average income in the top decile to mean income in the bottom decile is biased downwards by a factor of at least 2, while the proportion of households with per capita income below the poverty line, as obtained by the methods described in the above-mentioned papers, might differ by a factor of 1.5 – 2. A similar result was obtained in this study, as well. See below Section 3.

main contribution to inequality and poverty indicators are due to the tails of the distribution.

(iii) The calibration of the lognormal model used by official Russian statistical agencies does not eliminate the sample bias. Our calibration will 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 the macroeconomic income and expenditure balance (Velikanova and Frolova, 1999). The (lognormal) shape and the parameters of the distribution (in particular, the mode) are assumed to be preserved under the transformation, which is also questionable.

(iv) Distribution approximation and weighting (calibration) techniques proposed by other researchers (*e.g.*, Shevyakov and Kiruta, 1999; Yershov and Maier, 1998) also tend to lead to substantial distortions. These approximations do not allow us to estimate the share of "rich" and "ultra rich" households from the unobserved part of the expenditure range.

(v) The head-count ratio, which is the proportion of households with per capita expenditure below the poverty line, is usually used as an appropriate poverty measure no matter what the goal of the analysis is (Proceedings of Goskomstat R.F., 1999b; Braithwaite, 1999; Ministry of Labour of Russia, 1999). However, the choice of poverty indicator (or criteria to classify a household as poor) is determined by the goal of economic analysis, *i.e.*, by the particular application. In particular, the Foster–Greer–Thorbecke family of indices is known to be more sensitive with the targeted assistance goals.

(vi) The problem of the optimally allocating resources as targeted assistance has never been stated l*et al*one solved in Russian economic theory and policy, **when** the optimality is determined in the mean of minimization of a certain social tension indicator.

The goals of the project are to overcome the aforementioned drawbacks (i) – (vi). In particular, we aim at developing a methodology for econometric analysis of per capita expenditure distribution based on Russian budget survey data, analysing the main characteristics of poverty and welfare inequality of the Russian population and their statistical assessment, and formulating and solving the problem of optimal allocation of resources dedicated to targeted assistance for the poor.

The main objective of this study is to construct a meaningful econometric model of the regional/national per capita expenditure distribution. This also implies developing an identification methodology based on sample budget surveys and macroeconomic balance of income and expenditure.

The solution to this task will be linked to the specific features of the Russian economy and to the way these are reflected in household behavior. In particular, refusal of a household to participate in a survey (unit nonresponse, or truncation) plays an important role in the analysis of expenditure distribution, leading to deterioration of the sample's 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 the probability of one. It is likely that the rich households (*i.e.*, those with per capita expenditure above a certain value) would belong to this category, as high income is often associated with illegal or semi-legal economic activities.[2]

Apparently, any econometric model of income/expenditure distribution that would aim at eliminating (or at least attenuating) the data quality problems must be based on explicitly formu-lated (and, if possible, substantiated and proved with the statistics) additional working hypotheses and assumptions. In this study, such hypotheses are 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 nonresponse, *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 proof, the following additional assumptions:

- Working assumption A_1, which states that the coefficient of variation of per capita expenditures (or the variance of log expenditure) is constant across all strata;
- Working assumption A_2, wich deals with the shape of the distribution of per capita expenditure within the *unobserved* range of expenditures (right distribution tail, the richest population strata).

Hypothesis H_1 is based on the salient transition features of Russia (see Section 2.1 below). Statistical testing and further use of this hypothesis is essential for the formulation of a meaningful model of per capita expenditure. Statistical testing and further use of hypothesis H_2 is aimed at

[2] The adjustment for another source of sample bias, namely, misreporting (*e.g.*, in order to conceal true income) is largely beyond the scope of this paper. Some aspects are touched upon in Section 3.3.

eliminating the unit non-response bias. Assumptions A_1 and A_2 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.

The second objective of this study serves as an example of the application of the proposed methodology in the fieldwork. We shall aim to consider a broad class of poverty indices based on the per capita expenditure distribution and formulate the problem of the optimal allocation of a limited resource S devoted to targeted social assistance for the poor, based on an objective function from this class.

The following family of poverty indices will be considered:

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

where $f(x)$ is the per capita expenditure density function, z_0 is the 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) encompasses such popular measures as the Foster–Greer–Thorbecke family of indices (FGT(α)), Dalton class indicators, and Poverty-Line-Discontinuous measures (Bourguignon and Fields, 1995; Foster *et al*., 1984; Hagenaars, 1987).

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

$$I(w, \tilde{f}) = \int_0^{z_0} w(x) \tilde{f}(x \mid \varphi; S) dx . \qquad (1')$$

Our second objective is then reduced to the identification of $\varphi_0(x \mid S)$ such that (1') achieves its minimum, given $w(x)$ and S:

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

It is worth noting that our second objective is considered within the framework of a specific project of long-term poverty alleviation (Braithwaite, 1999; Ministy of Labour R.F., 1999). The implications of this context are twofold. First, the argument for relatively high income mobility (Bogomolova *et al*., 1999) is not fully applicable to this population category. Second, the main instruments of long-term poverty alleviation are direct transfers to needy households rather than the creation of incentive schemes (which is most relevant for the temporarily poor, *e.g*., the unemployed).

Our third objective 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 the 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 try to find countries which have similar levels of the above indicators.

Apparently, the truncation of the super-rich cannot noticeably affect the poverty indices that form the framework for the problem of social aid distribution. In fact, the poverty analysis focuses on the left tail of the expenditure distribution, while the use of working assumption A_2 is aimed to fit the right tail of the distribution.

Introduction of the super-rich stratum into the model, however, does affect inequality indices.[3] This correction is viewed as an important one by us, as inequality and polarization indices characterize the social tension in the population. Let us discuss the sources where problems similar to our main task were addressed.

The model of per capita expenditure distribution developed in this project is supposed to enhance and modify the basic model of population per capita *income* distribution pioneered by Aivazian (1997). The *modification* includes i) introduction and statistical estimation of the budget survey unit non-response probability (see H_2 above); ii) replacement of income by *expenditure* in the lognormal 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 hypothesis H_2 and working assumptions A_1 and A_2.

[3] The calculations in Aivazian (1997) 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.

The papers by Shevyakov and Kiruta (1999), Yershov and Maier (1998), Suvorov and Ul'anova (1997) and Aivazian (1997) contain arguments which prove the validity of our critique (i) – (iv) in the introduction. Velikanova, T., Kolmakov, I., and Frolova, E. (1996) describe an approach which is also based on the mixture of lognormal distributions, but this source neither provides econometric tools to analyze this mixture nor proposes any way to reconstruct the unobserved data. The approach by Yershov and Mayer (1998) is based on polynomial density approximation and seems to be too formal. It does not allow for the 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 Ul'anova (1997) is inadequacy of the basic assumption on the 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 about the adequacy of the model's estimate of income based on the Goskomstat budget survey sample (which is considered substantially biased even by Goskomstat specialists, l*et al*one independent experts); c) propose a formal approximation technique of unknown parameter fitting. While economic analysis of the 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 (1999), especially to the differences of their approach from the one proposed in our project. Their work is currently the most serious attempt to describe *realistically* the regional per capita income distribution using 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. In our opinion, the main drawbacks of Shevyakov and Kiruta's approach are as follows:

a) The proposed weighting (calibration) technique, in fact, ignores the population beyond the maximum income observed. The right tail of the distribution remains unaccounted for and the censuring problem is not addressed. In our model, the tail is recovered by using assumption A_2.

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 analyzed unknown distribution 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 the 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 Task 2. First of all, worth mentioning are the World Bank project (Braithwaite, 1999) and pilot programs (Ministry of Labour R.F., 1999). They do accomplish a rightful attempt to assess poverty according to the re-estimation of realistic household per capita income (termed "*potential consumption expenditures*" by Braithwaite, 1999). Both approaches, however, still suffer from significant drawbacks analyzed by Aivazian in Proceeding of the Higher School of Economics (1999). 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).

A comprehensive overview of poverty indicators is given by Korchagina, Ovcharova, and Turuncev (1999). 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 solves the problem of optimizating targeted social assistance for the poor. Nevertheless, various aspects of this problem are addressed in Western literature though most authors still rely on income rather than expenditure distributions (Bourguignon and Fields, 1995; Sen, 1985; Atkinson, 1987; Kanbur, 1987; Foster and Shorroks, 1988; Ravallion, 1994). In particular, Bourguignon and Fields (1990) prove 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 enough money to the poorest people to raise their income to the threshold $\bar{z}_0 < z_0$, where $\bar{z}_0$ is found from the government budget and

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

where N is the total population. This strategy is referred to as "allocation of p-type" in Bourguignon and Fields (1990 and 1995) 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 a 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 by Bourguignon and Fields (1990) 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' by Bourguignon and Fields (1995). The transaction costs related to the distribution of the government subsidies are rarely accounted for, however.

As for the analysis of the third problem, we would like to mention the Esteban–Ray polarization index proposed by Esteban and Ray (1994). This index crucially depends on the knowledge of the tail strata of the distribution and is effectively used along with the Gini coefficient (which is a special case of the Esteban–Ray index with the value of a certain self-identification parameter being zero) in empirical works as a factor of crime (Fajnzulber *et al*., 1999). This measure, however, is only defined for discrete income groups, and the continuous extension is not straightforward.

2. THE MODEL AND METHODOLOGY OF ITS ECONOMETRIC ANALYSIS

2.1. Discussion of the basic working research hypotheses and model assumptions

The building of the population per capita expenditure distribution model is based on the theoretical inference and/or empirical testing of a number of working hypotheses.

- **Hypothesis H_1** states that the distribution of the Russian population 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 provided by Aivazian (1997).

The 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 $\sigma^2(a) = \mathbf{Var}\,(\ln\xi(a))$. Here, homogeneity refers to similar income sources as well as similar geographical, social, demographic, and professional characteristics of its representatives.

(b) If society as a whole can be represented by a spectrum of such strata (continuous in terms of the average log expenditures a), 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 these propositions. 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 by Aivazian, Rabkina, and Rimashevskaya (1967) and applied to the wages of workers in the Soviet Union. This distributional assumption is closely related to Mincer-type earning equations with normal errors.

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.

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

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 by Aivazian *et al.*, (1967).

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 the 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 does not abolish the a) and b) postulates, though it affects the shape of $q(a)$.

- **Hypothesis H_2** states that the probability that a household refuses 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 (Mroz *et al*., 1997) and some additional information from Goskomstat. This hypothesis was prompted by Mrs. Frolova (the Head of Living Standards Department of Goskomstat).
- **Assumption A_1** states that the *coefficient of variation* of the household per capita expenditures is constant across the social strata, *i.e*., it is independent of the strata number. This assumption can also be verified by criteria of variance homogeneity (Aivazian, 1997). As long as income and expenditure $\xi(j)$ of the population of j-th homogeneous strata are distributed lognormally with the parameters $a(j) = \mathbf{E}(\ln\xi(j))$ and $\sigma^2(j) = \mathbf{Var}(\ln\xi(j))$ (*e.g*., Aivazian, 1976), the assumption A_1 is equivalent to A_1' : $\mathbf{Var}\,(\ln\xi\,(j)) = \sigma^2 = \text{const}$.

The equivalence of A_1 and A_1' follows from the relation between the moments of the lognormal distribution:

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

- **Assumption A_2** states that the population per capita expenditures x in the latent range of $x > \max\limits_{1 \le i \le n}\{x_i\}$, where x_i is per capita expenditures in the i-th household surveyed, and n, total number of households, can be approximated by a three-parameter lognormal distribution with a shift parameter $x_0 = \max\limits_{1 \le i \le n}\{x_i\}$ and variance of logarithms $\mathbf{Var}\,(\ln\xi\,(k)) = \sigma^2$ where σ^2 is independent of strata and estimated from the observed strata (see assumption A_1 above).

  Strictly speaking, this working 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 assumption, let us mention some stylized facts related to the Russian transition.

One of the important consequences of the rapid disintegration of the USSR and demolition of its socioeconomic structure is the evolution of a new social elite. This narrow class of highly prosperous individuals comprises mainly those who come from the recent partisan and bureaucratic elite, former higher executives, and finally members of organized crime. This "select" group employing certain oblique methods of privatising national wealth now has the opportunity of trading it on internal and external markets, whether openly or by underhand means.

Some specialists in the field (cf. Suvorov and Ul'anova, 1997) estimate that an increase in the sale of the country's natural resources of 0.2 – 0.3 per cent per annum results in an increase of gross population income by 10 – 20 per cent. It is obvious that most of this growth can be attributed to this "select" class that may be, considering the uniformity of the social ranking and social background of its members, categorized as a separate socioeconomic stratum. This is why the population distribution by per capita expenditure, which is referred to in assumption A_2, pertains specifically to this stratum.

Usually, the right tail of the 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 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 H_1, H_2 and working assumptions A_1, A_2, a non-formal (*i.e.*, an interpretable) model of the 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.

2.2. The main variables and information sources

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

Following Goskomstat's methodology from Proceedings of Goskomstat of R.F. (1999b), we shall define (quarterly) gross pecuniary expenditures of a household as the sum of:

- $\xi^{(1)}$ — *quarterly consumption expenditures*, which is the sum of food product expenditures, alcohol, private non-food 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 the 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 $x_i, x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(6)}$ 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,[4] on the basis of nutritional re-

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

quirements. 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 Buhmann *et al*. (1988) 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*., Coulter, Cowell, and Jenkins, 1992; Cowell and Mercader-Prats, 1997). 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. However, we are not aware of any convincing argument in favor of any equivalence scale that 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 robustness of the findings of our preliminary analysis convinces us they are good enough.

2) Regional/national average per capita expenditures μ_{macro} defined from macroeconomic characteristics, namely, the quarterly Goskomstat publication "The Population Income and Expenditure Balances" (Goskomstat R.F., 1996) 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 (Mroz *et al*., 1997). The RLMS questionnaire contains expenditures for a large number of goods and services. This data can be aggregated into large groups of goods and services, and into 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 break-

down to about 10 items), rent, club payments, insurance premia, savings and credits have a one-month window; non-food consumer goods and durables expenditures are calculated 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 to a 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, as well as from Soviet times. This correction, however, remains, to our knowledge, a purely theoretical argument that has never been implemented in applied work.

2. Household budget survey data as of the Q3 1998 on three regions of Russia, namely, Komi Republic, Volgograd and Omsk , with a supplementary questionnaire (Aivazian and Gerasimova, 1998). According to Goskomstat methodology (Proceedings of Goskomstat R.F., 1999b), the sample is constructed to be representative of household types, except collective households (*e.g.*, hospitals, military units, *etc.*), on the basis of the 1994 microcensus. During the quarterly budget survey, a household fills a two-week daily log of expenditures twice during the quarter, 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 being accounted for, including consumption and non-consumption expenditures); consumption expenditures ("**potras**" variable; the proportion of pecuniary expenditures directed to the acquisition of consumption goods and services); final household consumption expenditures ("**konpot**" variable; consumption expenditures excluding food products transferred outside the household, plus 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 a questionnaire on quality of life (Aivazian and Gerasimova, 1998).

2.3. Model description and parameter interpretation

Let ξ denote (in thousands of rubles) the yearly average expenditure of a randomly selected representative of the Russian population, and let ξ_j (in ths. of rub.) denote per capita expenditures of the representative of the 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:

$$f(x \mid \Theta) = \sum_{j=1}^{k} q_j \frac{1}{\sqrt{2\pi}\sigma_j x} \exp\left(- \frac{(\ln x - a_j)^2}{2\sigma_j^2} \right) + I_{(x_0, +\infty)}(x) \times$$
$$\times \frac{q_{k+1}}{\sqrt{2\pi}\sigma_{k+1} \cdot (x - x_0)} \exp\left(- \frac{[\ln(x - x_0) - a_{k+1}]^2}{2\sigma_{k+1}2} \right), \quad (5)$$

where $I_{(x_0, +\infty)}(x)$ is the indicator function of set $(x_0, +\infty)$ (*i.e.* $I_{(x_0, +\infty)}(x) = 0$ for $x \le x_0$ and $I_{(x_0, +\infty)}(x) = 1$ for $x > x_0$), and $\Theta = (k; q_1, \ldots, q_{k+1}; a_1, \ldots, a_{k+1}; x_0; \sigma_1^2, \ldots, \sigma_{k+1}^2)$ 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 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 the 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 richest $k + 1$-th stratum of the population exceed the threshold x_0, and that these individuals always refuse to participate in surveys. The rest of the households are available for statistical investigation, although they can also escape from the survey with probability $p(x)$, which is monotonically increasing with x (see hypothesis H_2 above).

Econometric analysis of 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. Note that the parameter k also needs to be estimated. This poses some additional problems, as the behavior of the usual estimators when the true parameter is on the boundary (which would be the case if k is overestimated) might be rather strange.

2.4. Econometric analysis methodology

2.4.1. Estimation of the dependence of refusal probability p(x) on 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 of 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 assumed to follow the logistic model:

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

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, model (6) gives a set of $4 \times 5 = 20$ models (one for each population category) describing the dependence of refusal probability p on per capita expenditure:

$$p_{kl}(z) = p(z \mid z_k^{(2)}, z_l^{(3)}) = P\{\eta_i = 0 \mid z^{(1)} = z,\ z^{(2)} = z_k^{(2)},\ z^{(3)} = z_l^{(3)}\}, \quad k = 1, 2, 3, 4;\quad l = 1, 2, 3, 4, 5. \tag{6'}$$

In fact, a wider set of regressors was used initially in the analysis that also included the social and demographic structure of the household and the characteristics of the household head, besides the log per capita expenditure, the settlement type, and household head's education. The subsequent analysis shows statistical insignificance of some characteristics, and the selection of the logistic regression model leads 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 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 $z = z^{(1)} = \ln \xi$:

$$p(z) = P\{\eta_i = 0 \mid z^{(1)} = z\} = \frac{e^{\beta_0 + \beta_1 z}}{1 + e^{\beta_0 + \beta_1 z}}. \qquad (6")$$

2.4.2. Calibration (weighting) of the existing observations. The analysis of 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 regional/national per capita expenditure distribution. When the information is sufficient (categories k_i, l_i corresponding to $z^{(2)}_{k_i}$, $z^{(3)}_{l_i}$ 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 use 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. rub.).

Let $f(x)$ be the density function of the per capita expenditure distribution of the 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

$$\nu(x^*) \approx n f(x^*) \Delta . \qquad (7)$$

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

$$\bar{\nu}(x^*) \approx n f(x^*)[1 - p(x^*)] \Delta . \qquad (8)$$

From (7) and (8) it follows that

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

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

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

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

$$\left(x_1;\ \frac{1}{n}\right), \left(x_2;\ \frac{1}{n}\right), \ldots, \left(x_n;\ \frac{1}{n}\right), \tag{10}$$

in which each observation x_i $(i = 1, 2, \ldots, n)$ has *the same weight* $1/n$, then we should recalibrate, or *re-weight*, the sample in the following way:

$$(x_1;\ \omega_1),\ (x_2;\ \omega_2), \ldots, (x_n;\ \omega_n), \tag{11}$$

where ω_i are found from

$$\omega_i = \frac{1/[1-p(x_i)]}{\sum_{j=1}^{n} 1/[1-p(x_j)]}. \tag{11'}$$

It is worth noting that ω_i increases with the refusal probability $p(x_i)$, and

$$\sum_{i=1}^{n} \omega_i = 1$$

2.4.3. Estimation of the observed mixture component parameters. At this stage we solve the problem of estimation from the sample (11) of the parameters $k, \tilde{q}, \ldots, \tilde{q}_k, a_1, \ldots, a_k, \sigma_1^2, \ldots, \sigma_k^2$ in the mixture of distributions:

$$\tilde{f}(x) = \sum_{j=1}^{k} \tilde{q}_j \frac{1}{\sqrt{2\pi}\sigma_j x} \exp\left(-\frac{(\ln x - a_j)^2}{2\sigma_j^2}\right). \tag{12}$$

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

$$\tilde{\varphi}(z) = \sum_{j=1}^{k} \tilde{q}_j \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(- \frac{(z - a_j)^2}{2\sigma_j^2} \right) \tag{13}$$

by the sample

$$(z_1; \omega_1), (z_2; \omega_2), \ldots, (z_n; \omega_n), \tag{8'}$$

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

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 the estimation are briefly described in Appendix 3 (for more detail, see Day, 1969; Dempster, Laird, and Rubin, 1977; Aivazian, 1996; Rudzkis and Radavicius, 1995; Jakimauskas and Sushinkas, 1996). The software implementations are CLASSMASTER software developed at CEMI and denormix STATA module developed by S. Kolenikov (available from his web page, http://www.komkon.org/~tacik/stata).

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

$$\mu = \sum_{j=1}^{\bar{k}} \hat{q}_j \int_0^{\infty} x \frac{1}{\sqrt{2\pi}\hat{\sigma}_j x} \exp\left(- \frac{(\ln x - \hat{a}_j)^2}{2\hat{\sigma}_j^2} \right) dx +$$
$$+ \int_0^{\infty} q_{\bar{k}+1} \int_{x_0}^{\infty} x \frac{1}{\sqrt{2\pi}\hat{\sigma}_{k+1}(x - x_0)} \exp\left(- \frac{(\ln(x - x_0) - \hat{a}_{k+1})^2}{2\hat{\sigma}_{k+1}^2} \right) dx, \tag{14}$$

where

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

Because of the properties of lognormal distribution,

$$\mu = \sum_{j=1}^{\bar{k}} q_j \exp\left(\frac{1}{2}\hat{\sigma}_j^2 + \hat{a}_j \right) + q_{\bar{k}+1} \left[x_0 + \exp\left(\frac{1}{2}\sigma_{\bar{k}+1}^2 + a_{\bar{k}+1} \right) \right]. \tag{14'}$$

The value of μ from (14') depends on the unknown $q_{\bar{k}+1}, a_{\bar{k}+1}$, as well as on x_0 and $\sigma^2_{\bar{k}+1}$. 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 assumption A_1' (see Section 2.1 above), the overall estimate $\hat{\sigma}^2$ of the variance of logarithms is

$$\hat{\sigma}^2 = \sum_{j=1}^{\bar{k}} \hat{q}_j \hat{\sigma}_j^2 \tag{17}$$

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

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

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

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

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

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

It is resonable to assume that

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

where the sign << means "much less," *i.e.*, that q_{k+1} is about an order of magnitude less than $\min_{1 \le j \le \bar{k}} \{q_j\}$.

Level line (18) may be represented by a table with the values of $q_{\bar{k}+1}$ as input and $a_{\bar{k}+1}$ from (14') – (18) as output. A possible range of values $q_{\bar{k}+1}$ could be chosen as follows (with $\min_{1 \le j \le k} \{q_j\} = m \times 10^{-2}$, $1 \le m \le 9$,

i.e., if the least of the stratum shares is at the level of several per cent):

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

By using (14'), the following limit from above for the share of the unobserved stratum can be calculated:

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

2.4.5. Poverty indices and targeted assistance for the poor. If we restrict the class of weighting functions $w(x)$ in (1) to functions like (3), then we can use the results of Bourguignon and Fields (1990) on the optimal allocation of 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

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

where $\bar{z}_0 < z_0$ and $I_\alpha^{(z_0)}(f) = \int_0^{z_0} \left(\frac{z_0 - x}{z_0}\right)^\alpha f(x)\,dx$;

(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 is a corresponding optimal allocation.

In this study, the share of poor (head count ratio, FGT(0)) and poverty depth (FGT(2) sensitive to 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 3.

3. EMPIRICAL ANALYSIS RESULTS

For conducting our empirical analysis we used the following steps.

S t e p 1. The analysis of the sample distributions of per capita expenditure has been conducted by using the Goskomstat budget survey data on the Komi Republic, Volgograd and Omsk regions (Q2 1998), as well as the RLMS Round VIII data (Q4 1998). In particular, sample statistics and histograms are obtained as the output of this step (see Appendix 1).

S t e p 2. By using RLMS panel data Rounds V–VIII and additional refusal data,[5] the multiple logit model was estimated to relate the probability of a household with particular characteristics to refuse to participate in a budget survey.

S t e p 3. According to the methodology described in Section 2.4.2, either rough (with logit model (6")) or fine (with logit model (6)) calibration (re-weighting) of the existing data was performed to eliminate truncation bias.

S t e p 4. Sample distributions are re-analyzed accounting for the weights estimated at the previous step. The results are compared to those obtained in step 1.

S t e p 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 Section 2.4.3).

S t e p 6. According to the methodology described in Section 2.4.4, the unobserved component parameters are produced for each of the four data sets by using the estimates of the mixture components obtained in the previous step. The goal here is to eliminate the censoring bias.

S t e p 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 the Q4 1998.

---

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

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

The estimation results are reported in Appendix A. Some evidence based on Figs A.1 – A.4 and Tables A.1 – A.8 will be discussed in this section.

First, per capita distributions cannot be adequately described by the simple lognormal model (either within any of the regions or within the country as a whole). Columns 2 and 3 of Tables A.5 – A.8 suggest that the two-homoskedastic-component model does not describe the data well, either. The fact that the mixture model fits the RLMS data demonstrates very high values of the χ^2 fit statistic for any number of components (and very low p-value, respectively) is probably related to the "large sample curse.[6] In fact, we have 9.716 observations in the RLMS data, while there are about a thousand observations in any of the regional data sets.

Second, the algorithms of both CLASSMASTER and Stata of the automatic search for the unknown number of the mixture components *k in the observed expenditure range* 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. In fact, the population share of the modal strata is more than 90%, which effectively masks all other components.

Volgograd region was the only exception. While for all other cases an increase in the number of mixture components beyond four would lead to a serious deterioration of the fit criteria (AIC, SBIC, ICOMP) and the very identification quality (multiple maxima of the likelihood function, flat regions that the algorithms stumble upon, coinciding components, *etc.*), The five-component model for Volgograd was the most parcimonious model accepted by the goodness of fit criteria.

Third, as compared to 1996, the stratification of the 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.

---

[6] This concept can be briefly described as follows. All real data are produced by data generating mechanisms that are in fact very complex. On the other hand, it is known that the power of a test increases with the sample size. Thus, with several thousand observations, the tests would likely reject relatively unsophisticated hypotheses.

The figures from Aivazian (1997) describing the per capita income distribution of the Russian population as of the fall of 1995 suggest local maxima of the density function. Population strata are then well-defined, which allows for sensible classification of the population by the strata with the following analysis of social and economic characteristics of each stratum. A similar analysis for the 1998 data is hampered by the fact that most of the 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 $q_{\bar{k}+1}$ and $a_{\bar{k}+1}$ under certain restrictions (see (14) – (18) in Section 2.4.4). The example of such a line for RLMS data is given below in 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 $(\bar{\mu}^{(\bar{k}+1)}, \hat{q}_{\bar{k}+1})$ on this level line. In

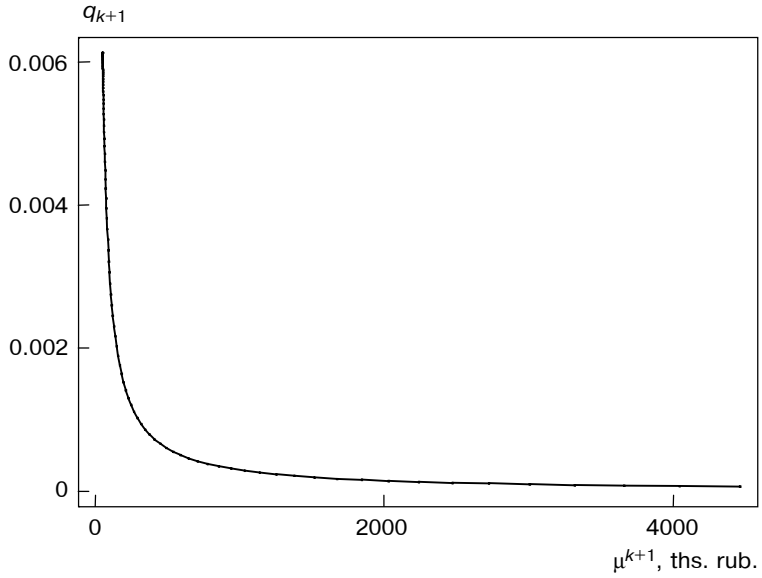

Fig. 1. The relation between the share and the mean per capita expenditure used in the estimation of the latent population strata parameters.

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, the Lorenz curve on which the Gini index is based is not very sensitive to the particular choice of the parameter couple (though it is sensitive to the very fact of inclusion or omission of the hidden stratum). Various estimates of the share of hidden income range from 25% to 40% (see Aivazian, 1997). In this study, the 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 – 830)/830 = 0.459 = 45.9%.

Fifth, the observation re-weighting (used here to adjust for truncation) and Monte Carlo simulation modeling of the unobserved stratum help explain 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.

3.2. Estimation of poverty, inequality and social tension indices

Table 1 reports the estimates of poverty and social tension indicators. In terms of the Foster–Greer–Thorbecke family of indices $I_\alpha^{(z_0)}(f)$ (see Foster *et al*., 1984, 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 (Ministry of Labour R.S., 1999) (column 5 for the regions that 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 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 the 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.

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

Table 1. Poverty and social tension indicators.

No	Region	Poverty line, ths. rub. [1]	Poverty rate				
			[1]	[2]	[3]	[4]	[5]
1	2	3	4	5	6	7	8
1	Russia	0.636	28.4	—	55.2	55.6	57.1
2	Komi Republic	0.466	20.6	26.7	53.8	56.2	56.7
3	Volgograd region	0.368	31.5	49.2	62.0	62.7	63.0
4	Omsk region	0.372	25.2	—	42.6	43.2	44.2

No	Region	Poverty depth (social tension) FGT(2)		
		[5]	[3]	[4]
		9	10	11
1	Russia	0.143	0.137	0.139
2	Komi Republic	0.140	0.134	0.139
3	Volgograd region	0.177	0.175	0.176
4	Omsk region	0.089	0.082	0.085

[1] Official Data of Goskomstat (Goskomstat R.F., 1998, 1999a, 1999b);
[2] Estimates from (Ministry of Labour R.F., 1999);
[3] Estimates from the lognormal model;
[4] Estimates from the mixture model (5);
[5] Direct weighed sample estimates.

8 and 9). This method gives higher poverty rates than the official statistics do, as long as the left tail of the distribution turns out to be heavier than the lognormal model could give. 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.

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

No	Region, data source, sample size	Mean expenditure, ths. rub.			Gini index		Funds ratio	
		A	B	C	D	E	D	E
1	2	3	4	5	6	7	8	9
1	Russia, RLMS VIII, n = 9716	0.913	0.932 (2%)	1.211 (32.6%)	0.478 0.380*	0.599	13.5*	45.8
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 region, HBS, n = 1263	0.412	0.433 (5%)	0.641 (55.6%)	0.389	0.590	14.0	32.0
4	Omsk region, 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.

* Goskomstat estimate as of Q4 1998.

A — Raw.

B — Calibrated (+Δ, %).

C — With the latent stratum (+Δ, %).

D — Raw data

E — Model (5) with latent stratum.

2) Although the share of the unobserved super-rich stratum is relatively low (one tenth or hundredth of the percentage point), it crucially affects the main characteristics of inequality and polarization. In particular, the Gini index for Russia in the 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. A similar pattern of increase in the Gini values is observed for the regional data, too (except maybe for the Omsk region). The magnitude of changes in the funds ratio is also really large, 50% to 200%. It might also be noted that the discrepancy is the largest for the Komi Republic, which is a resource rich region. This fact is supported by the rent seeking theory, *i.e.*, that rent seeking behavior emerges in economic environments with substantial rent flows, natural resource rent being the most typical example.

How can the revealed differences in the figures be explained, and should the results based on 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 the 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 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 model (5) are too large.[7] To provide some explanation, we need to note that in (5), all 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 lies somewhere in between. This question is addressed in more detail in the following subsection.

3.3. The sensitivity analysis of the Gini index and funds ratio estimates 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

[7] 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 the 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.

(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 on the Gini index to misreporting in the RLMS data, the following framework was adopted.

Let the distortion due to the misreporting be measured as

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

where μ_{cal} is the sample mean expenditure (possibly biased due to misreporting) calibrated according to the methodology described in Section 2.4.2 (here, $\mu_{cal} = 0.932$, 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 in 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 45% ((1211 − 830)/830 = 0.459). 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 the Gini index in the framework of model (5) with correction for the misreporting for a number of λ's are given in Table 3.

Table 3. The sensitivity analysis of the Gini index and funds ratio estimates with respect to misreporting.

Relative distortion λ	0%	5%	10%	15%
Estimates of Gini index based on (5) and corrected for misreporting	0.599	0.592	0.569	0.554
Estimates of funds ratio based on (5) and corrected for misreporting	45.8	42.5	39.5	35.6

The simplest model of misreporting was adopted, namely that each household understates its true expenditure by a 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 Sections 2.4.3, 2.4.4, and Appendix A2. For each λ, 20 bootstrap samples of size 400,000 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 $q_{k+1} < 0.1\%$.

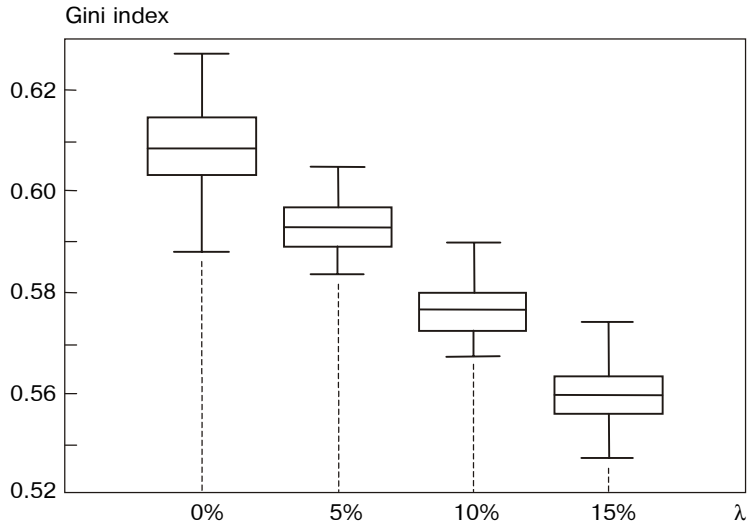

Fig. 2. Box-whisker plots for Gini indices obtained at various levels of λ.

Typically, about a hundred households from the latent stratum were present in each bootstrap sample. The number of bootstrap samples (20) allows us to interpret the observed range as the approximate 95% confidence interval for the true Gini index (see box-whisker plots on Fig. 2).

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 the RLMS sample range from 10% to 15%, *the most viable range of the Gini index is 0.55 – 0.57*.

4. CONCLUSIONS

1. The specifics of modern Russian economy induce one to use the level of per capita expenditure (instead of per capita income, as is done in other studies) for estimating poverty level and degree of inequality, and for constructing testing procedures that would help to classify households according to their level of wealth. 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 the shadow economy, is accounted for;

c) the concept of household welfare inflow is appropriately generalized to include home production, land (subsidiary plot) and property (real estate, vehicles, jewelry, *etc*.) that the household possesses.

2. When gross expenditure of the household is calculated, all sorts of expenditures are added up, including expenses for consumer goods, intermediate goods (*e.g*., tools and materials for the subsistence plot operations), net savings in all assets (including bank deposits and foreign currency), fixed capital growth, taxes and other obligatory payments, cash, and home production. In this work, 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, although would hardly affect any of the qualitative results.

3. The peculiar situation of the transition period in Russia, though not rejecting the general scheme of the lognormal mixture model of income/expenditure distribution, leads to certain changes in the nature of the mixing function $q(a)$. The genesis of the *discrete* lognormal mixture (instead of *continuous* mixture of a 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 have had 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 the 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.

4. The econometric analysis of our model includes: a) per capita expenditure density identification via lognormal finite mixture parameters

$$\Theta = (k; q_1, q_2, \ldots, q_k; a_1, a_2, \ldots, a_k; \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$$

estimated by the appropriate statistical procedures; b) re-weighting of the distribution accounting for the probability of unit non-response as a function of per capita expenditures and other household characteristics; c) reconstruction of the unobserved $(\hat{k}+1)$ -th stratum via the second recalibration of the model based on partially verifiable working hypotheses and macroeconomic income and expenditure balances.

5. Unlike the approaches used for official statistics and by other Russian researchers, our model of per capita expenditure distribution and the corresponding estimation methods do adjust for the households that refused to participate in the survey. The sample weights for the existing households are calculated by using the estimates of the probability of the household to avoid being surveyed, which is believed to depend upon the household economic, demographic, and geographic characteristics, and the distribution in the unobserved expenditure range is estimated with special hypotheses.

6. Further applications of the per capita expenditure model included:

- the estimation of the poverty indicators within the problem of the optimal allocation of the bounded targeted assistance for the poor;
- the estimation of the main characteristics of wealth inequality that serve as indicators of the level of social tension.

7. Most of the corrections proposed in our model tend to improve the fit of the upper tail of the distribution. The poverty indicators (Foster–Greer–Thorbecke family) are insensitive to the right tail, so they are not affected much by the methodology we propose. The second objective of this inquiry that was stated in the beginning has been achieved by formulation in Section 2.4.5. of the optimal allocation rule for targeted public relief.

8. The substantial innovation that our model introduces is thus related to the right tail of the distribution, and hence, the inequality indicators (Gini

index, funds ratio), estimation and analysis. This paper shows that the sensitivity of the estimates with respect to the variations in the model assumptions, in particular, in the ways to account for misreporting, are not too high. Upon analyzing the results obtained for realistic model variations, we suggest the following estimates: for the Gini index, 0.55 – 0.57, and for the funds ratio, 36 – 39 (*c.f.* the official figures of 0.38 and 13.5).

9. 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*.

APPENDICES

A.1. The analysis of the sample distributions of per capita expenditure for particular regions and Russia as a whole

A.1.1. Russian Federation. RLMS data set, Round VIII, October–November 1998, 9716 observations.

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

No	Indicator, thousands of rubles	Sample value	
		Raw data	Weighted data
1	Mean per capita expenditure ($\hat{\mu}$)	0.814	0.830
2	Standard deviation (S)	1.318	1.386
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

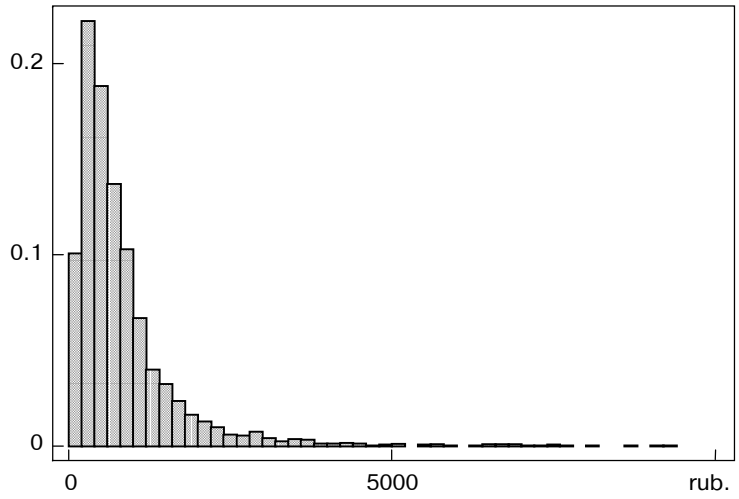

Fig. A.1a. The histogram of the per capita expenditure distribution of the Russian population (initial data).

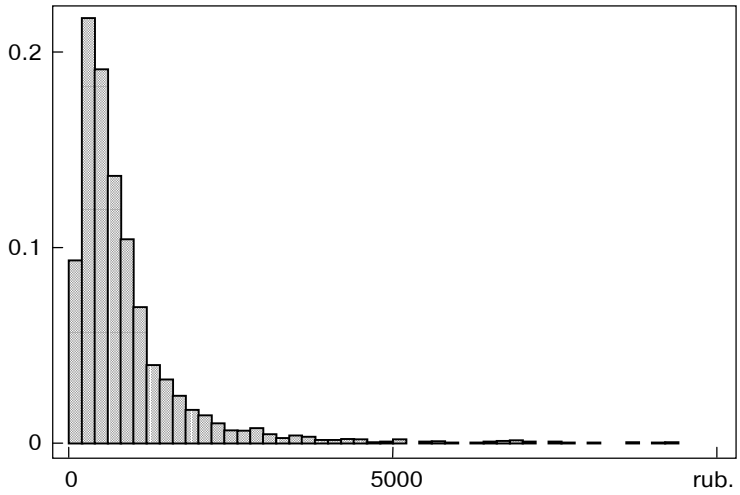

Fig. A.1b. The histogram of the per capita expenditure distribution of the Russian population (weighted data).

A.1.2. Komi Republic. Budget survey sample of 1089 individuals, Q2 1998.

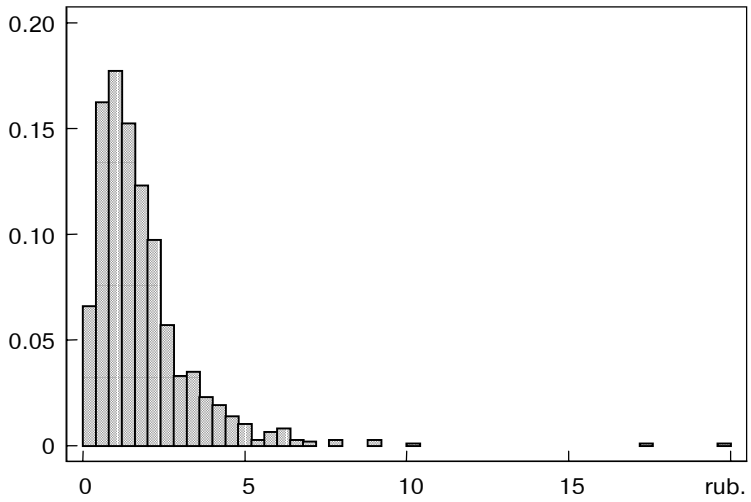

Fig. A.2a. The histogram of the per capita expenditure distribution of population of Komi republic (raw data).

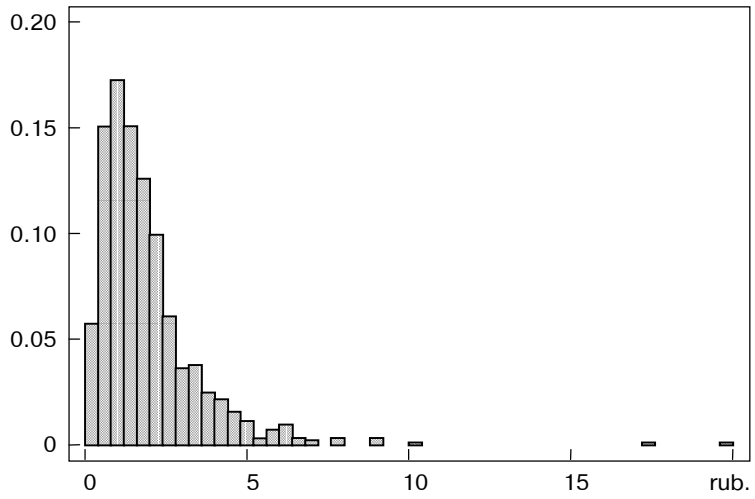

Fig. A.2b. The histogram of the per capita expenditure distribution of the population of the Komi Republic (weighted data).

Table A.2. Sample statistics of the essential characteristics of Komi Republic's per capita expenditure distribution.

No	Indicator, thousands of rubles	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 ($\bar{x}_{0,1}$)	0.154	0.163
6	Top decile ($\bar{x}_{0,9}$)	1.208	1.302

A.1.3. Volgograd region. Budget survey sample of 1263 individuals, Q2 1998.

Table A.3. Sample statistics of the essential characteristics of Volgograd region's per capita expenditure distribution.

No	Indicator thousands of rubles	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 ($\bar{x}_{0,1}$)	0.101	0.110
6	Top decile ($\bar{x}_{0,9}$)	0.766	0.794

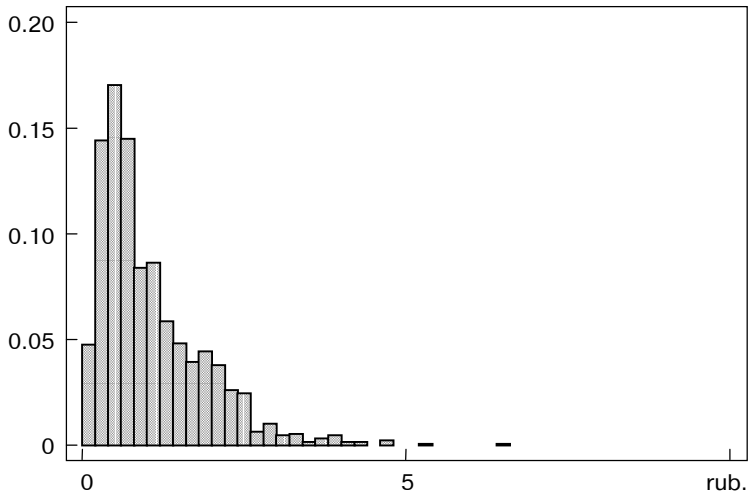

Fig. A.3a. The histogram of the per capita expenditure distribution of the population of the Volgograd region (raw data).

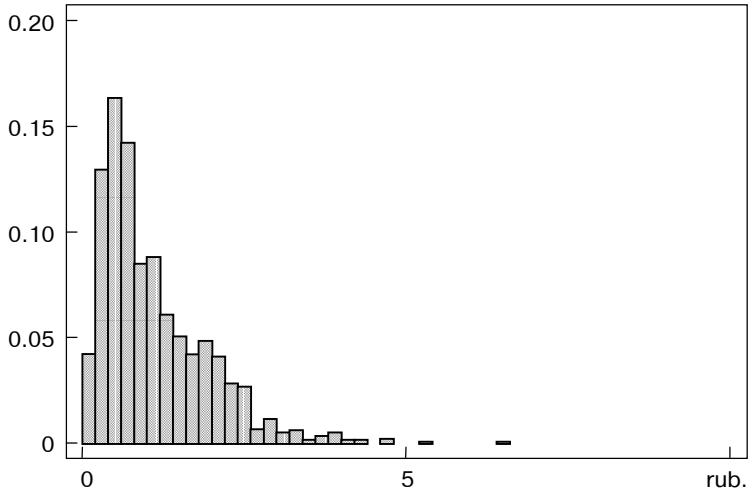

Fig. A.3b. The histogram of the per capita expenditure distribution of the population of the Volgograd region (raw data).

A.1.4. Omsk region. Budget survey sample of 1244 individuals, Q2 1998.

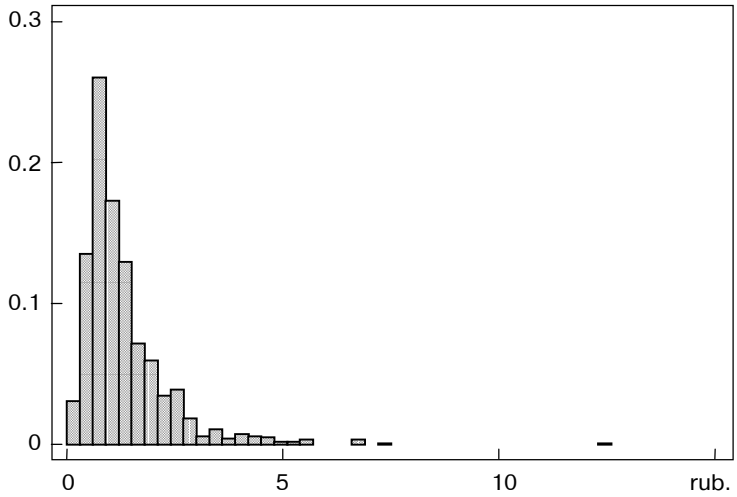

Fig. A.4a. The histogram of the per capita expenditure distribution of the population of the Omsk region (raw data).

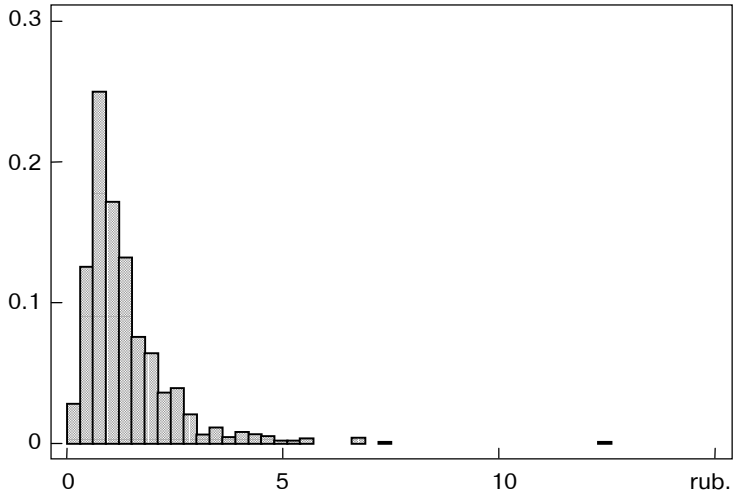

Fig. A.4b. The histogram of the per capita expenditure distribution of the population of the Omsk region (weighted data).

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

No	Indicator, thousands of rubles	Sample value	
		Raw data	Weighted data
1	Mean per capita expenditure ($\bar{\mu}$)	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 ($\tilde{x}_{0,9}$)	1.211	1.238

A.2. The estimation results for the mixture model in the observed per capita expenditure range

A.2.1. Estimation methodology. 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

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

of the density function

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

by using the random sample (8') data via the maximum likelihood when the number of components k is fixed. Here, $\varphi(z \mid a_j;\ \sigma_j^2)$ 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

$$\Theta(k) = (\tilde{q}_1, \ldots, \tilde{q}_k;\ a_1, \ldots, a_k;\ \sigma_1^2, \ldots, \sigma_k^2)\,, \tag{A.3}$$

that the log likelihood function

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

would attain its maximum over θ:

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

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 (Day, 1969; Dempster *et al*., 1977):

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

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

where

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

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 parameter estimates

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

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

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

The solutions are:

$$\tilde{q}_j^{(t+1)} = \sum_{i=1}^{n} \omega_i g_{ij}^{(t)},$$

$$\hat{a}_j^{(t+1)} = \frac{1}{\tilde{q}_j^{(t+1)}} \sum_{i=1}^{n} \omega_i g_{ij}^{(t)} z_i,$$

$$(\sigma_j^2)^{(t+1)} = \frac{1}{\tilde{q}_j^{(t+1)}} \sum_{i=1}^{n} \omega_i g_{ij}^{(t)} (z_i - \hat{a}_j^{(t+1)})^2, \qquad \text{(A.10)}$$

$$j = 1, 2, \ldots, k.$$

Here the iteration ends, and the expectation step is repeated with the updated $\tilde{q}_j^{(t+1)}, \hat{a}_j^{(t+1)}$ and $(\tilde{\sigma}_j^2)^{(t+1)}$ $(j = 1, 2, \ldots, k)$. Dempster *et al*. (1977) and others in later works[8] prove, under some general assumptions, of which the most restrictive one is the requirement of the 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 Jakimauskas and Sushinkas (1996).

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

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

$$H_0 : k = j$$

[8] The general framework of the algorithms that later were given the name "EM-algorithms" seems to have been pioneered by Shlesinger (1965). The properties of these algorithms were also studied in this work; however, this work is not easily accessible in the West and thus it is not known among Western statisticians.

against the alternative

$$H_1 : k = j + 1, \quad - j = 1, 2, \ldots,$$

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 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 (Aivazian, 1996).

There are, however, other options to proceed to. 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 by 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 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 $\chi^2(m - p - 1)$ 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 (Gould and Sribney, 1999). The 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. 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, the Newton–Raphson step is performed;

   4.3. otherwise, the 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 is less than 10^{-6});

   5.2. the estimates of the coefficients have stabilized (by default, relative change is 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, 16,000. Some runs resulted in 3000 + iterations which took about a day to compute on a 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 $\sigma_j^2 = Var(\ln \xi)$.

A.2.2. Estimation results. The estimation results for the RLMS Round VIII data as well as for the regional data sets (Komi Republic, Volgograd and Omsk regions) 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; $n = 9716$, $m = 14$).

k	Goodness of fit $\chi^2(\nu(k))$ test	p-value	$\hat{\sigma}^2$	$\hat{a}_1$	$\hat{\tilde{q}}_1$, %	$\hat{a}_2$
1	152.5	$<10^{-7}$	0.865	6.343	100	—
2	96.4	$<10^{-6}$	0.826	6.370	98.90	3.914
3	**58.3**	**$<10^{-5}$**	**0.756**	**6.340**	**95.80**	**8.282**
4	58.4	$<10^{-5}$	0.716	6.297	90.96	7.618

k	$\hat{\tilde{q}}_2$ %	$\hat{a}_3$	$\hat{\tilde{q}}_3$, %	$\hat{a}_4$	$\hat{\tilde{q}}_4$, %	μ_{model}, ths. rub.
1	—	—	—	—	—	0.876
2	1.1	—	—	—	—	0.874
3	**2.3**	**4.159**	**1.80**	—	—	**0.927**
4	6.54	4.235	2.29	9.790	0.21	0.953

Three-component model is selected.

k is the number of mixture components;

$\nu(k)$ is the degree of freedom number;

$p(k)$ is the number of the estimated model parameters;

m is the number of the grouping intervals;

$\nu(k) = m - p(k) - 1$.

Table A.6. The results of maximum likelihood estimation of the normal mixture parameters for the log of per capita expenditure (Komi Republic; n = 1089, m = 12).

k	Goodness of fit $\chi^2(\nu(k))$ test	p-value	$\hat{\sigma}^2$	$\hat{a}_1$	$\hat{\tilde{q}}_1$, %	$\hat{a}_2$
1	32.08	0.0002	0.654	−0.842	–	–
2	16.54	0.0208	0.610	2.114	0.50	−0.857
3	23.67	0.0003	0.475	−0.162	22.88	−1.058
4	**9.77**	**0.4615**	**0.285**	**−1.976**	**8.99**	**0.010**
5	5.90	0.0150	0.168	0.890	3.07	−1.082

k	$\hat{\tilde{q}}_2$ %	$\hat{a}_3$	$\hat{\tilde{q}}_3$, %	$\hat{a}_4$	$\hat{\tilde{q}}_4$, %	$\hat{a}_5$	$\hat{\tilde{q}}_5$, %	μ_{model}, ths. rub.
1	—	—	—	—	—	—	—	0.598
2	99.50	—	—	—	—	—	—	0.630
3	76.79	2.527	0.33	—	—	—	—	0.636
4	**25.41**	**−1.038**	**65.17**	**2.338**	**0.43**	—	—	**0.628**
5	55.02	2.756	0.27	−0.115	29.47	−2.029	12.17	0.633

Four-component model is selected.

k is the number of mixture components;

$\nu(k)$ is the degree of freedom number;

$p(k)$ is the number of the estimated model parameters;

m is the number of the grouping intervals;

$\nu(k) = m - p(k) - 1$.

Table A.7. The results of maximum likelihood estimation of the normal mixture parameters for the log of per capita expenditure (Volgograd region; $n = 1263$, $m = 12$).

k	Goodness of fit $\chi^2(\nu(k))$ test	p-value	$\hat{\sigma}^2$	$\hat{a}_1$	$\hat{\tilde{q}}_1$, %	$\hat{a}_2$
1	43.63	$<10^{-5}$	0.723	−1.259	100	–
2	42.78	$<10^{-5}$	0.673	0.116	2.58	−1.295
2	42.72	$<10^{-5}$	0.716	−1.254	99.77	−3.044
3	37.53	$<10^{-5}$	0.577	−2.456	4.10	−1.280
4	32.694	$<10^{-5}$	0.180	0.613	4.10	−2.833
5	**12.03**	**0.0005**	**0.099**	**−2.943**	**6.16**	**−1.927**

k	$\hat{\tilde{q}}_2$ %	$\hat{a}_3$	$\hat{\tilde{q}}_3$, %	$\hat{a}_4$	$\hat{\tilde{q}}_4$, %	$\hat{a}_5$	$\hat{\tilde{q}}_5$, %	$\hat{\mu}_{\text{model}}$, ths. rub.
1	—	—	—	—	—	—	—	0.408
2	97.42	—	—	—	—	—	—	0.414
2	0.23	—	—	—	—	—	—	0.407
3	90.57	0.021	5.34	—	—	—	—	0.414
4	7.00	−0.647	38.56	−1.661	50.34	—	—	0.413
5	**28.58**	**−0.481**	**27.89**	**0.650**	**4.21**	**−1.266**	**33.16**	**0.411**

Five-component model is selected.

k is the number of mixture components;

$\nu(k)$ is the degree of freedom number;

$p(k)$ is the number of the estimated model parameters;

m is the number of the estimated model parameters;

$\nu(k) = m - p(k) - 1$.

Table A.8. The results of maximum likelihood estimation of the normal mixture parameters for the log of per capita expenditure (Omsk region; n = 1244, m = 12).

k	Goodness of fit $\chi^2(\nu(k))$ test	p-value	$\hat{\sigma}^2$	$\hat{a}_1$	$\hat{\tilde{q}}_1$, %	$\hat{a}_2$
1	23.83	0.0050	0.656	−0.838	—	—
2	13.75	0.0550	0.602	0.671	2.34	−0.875
2	25.63	0.0006	0.591	−2.915	1.48	−0.807
3	**13.47**	**0.0190**	**0.382**	**−0.960**	**84.72**	**−2.911**
4	23.30	<0.0001	0.351	2.192	0.18	0.165
4	14.77	0.0020	0.278	−3.003	2.06	−1.260
5	18.42	<0.0001	0.211	−3.047	2.01	0.436

k	$\hat{\tilde{q}}_2$ %	$\hat{a}_3$	$\hat{\tilde{q}}_3$, %	$\hat{a}_4$	$\hat{\tilde{q}}_4$, %	$\hat{a}_5$	$\hat{\tilde{q}}_5$, %	$\hat{\mu}_{\text{model}}$, ths. rub.
1	—	—	—	—	—	—	—	0.600
2	97.66	—	—	—	—	—	—	0.612
2	98.52	—	—	—	—	—	—	0.592
3	**2.20**	**0.294**	**13.09**	—	—	—	—	**0.607**
4	16.95	−2.908	2.28	−0.998	80.59	—	—	0.613
4	46.09	0.548	8.67	−0.563	43.18	—	—	0.606
5	11.87	−1.433	34.27	−0.662	51.68	2.302	0.18	0.612

Three-component model is selected.

k is the number of mixture components;

$\nu(k)$ is the degree of freedom number;

$p(k)$ is the number of the estimated model parameters;

m is the number of the grouping intervals;

$\nu(k) = m - p(k) - 1$.

A.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 2.4.1.

RLMS panel data were used to study the probability that a household refuses to participate in a sociological survey. For each of the 4.718 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

Continued from p. 55

	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 A.10. 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 to answer the question: "Does the probability of refusing to participate in a sociological survey depend on the welfare and other characteristics of the household?" or, in more general form, "Is truncation random?" In terms of Little and Rubin (1987), the question is whether the data are MAR, MCAR or anything eise.

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 nonresponse 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.*, Bogomolova, Topilina and Rostovcev, 1999). 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, so that the magnitude of the expenditure fluctuations is about 25%.

To adjust for income mobility, we use the average log, for all available years, of the appropriately deflated expenditure[9] to smooth out these fluctuations. The analogy can be drawn here to Friedman's lifecycle permanent income hypothesis (Deaton, 1992). Experimentation with other welfare measures such as the median expenditure for available years, imputed expenditures[10], 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 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. 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 success is within the 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

[9] 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".

[10] 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 (STATA, 1999; Little and Rubin, 1987) 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.

of household's 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).

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.)	Wald(1)= = 22.05	Wald(8) = = 317.86	Wald(1) = = 23.39	Wald(8)= = 334.78
p-value	0.00	0.00	0.00	0.00

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

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 do not rule out the possibility of having secondary education (moreover, those schools are based on secondary education), so the respective coefficients measure the difference of those two categories from secondary education. Finally, the higher 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 which has the highest LR per one degree of freedom.

The predicted values of the refusal probability are shown in Fig. 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 of the most populated and representative curves.

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 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 additional evidence for a strong and consistently verified link between the level of welfare and propensity to disclose information on individual or household wealth to third parties.

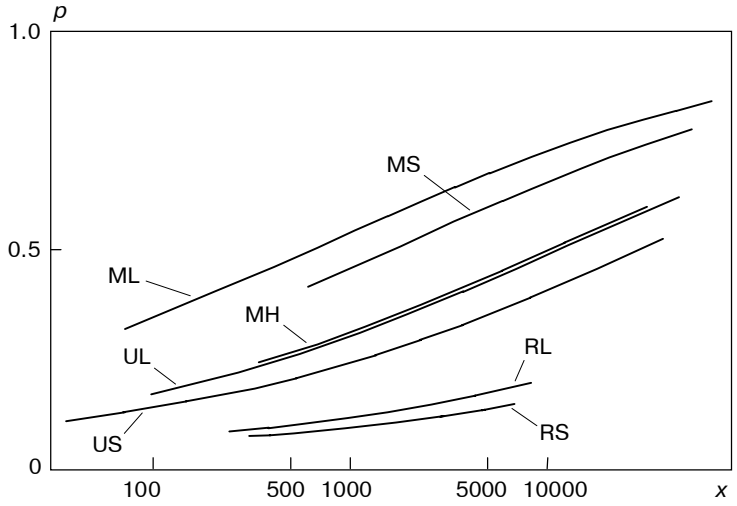

Fig. A.5. The family of the curves describing the dependence of the refusal probability on mean expenditure, in 1992 rubles.

REFERENCES

Aivazian, S.A. (1976) Probabilistic-Statistical Modelling of the Distributary Relations in Society, in: *Private and Enlarged Consumption* (North-Holland Publ. Comp, Amsterdam).

Aivazian, S.A. (1996) Mixture-Model Cluster Analysis Using the Projection Pursuit Method, in: *Computational Learning and Probabilistic Reasoning* (John Wiley and Sons Ltd.) 278–286.

Aivazian, S.A. (1997) Model for the mechanism generating distribution of Russian population by per capita income, *Economica i matematicheskie metody* **33**(4), 74–86, in Russian.

Aivazian, S.A. and I.A. Gerasimova (1998) Social structure and stratification of *Russian population* (CEMI RAS Press, Moscow), in Russian.

Aivazian, S.A., N.E. Rabkina and N.M. Rimashevskaya (1967) Predicting distribution of employees according to the salary level, *NII of labor of GKSMSSSR on labor and salary*, in Russian.

Atkinson, A.B. (1987) On the Measurement of Poverty, *Econometrica* **55**(4), 749–764.

Bogomolova, T.U., V.S. Tapilina, and P.S. Rostovcev (1999) Growing mobility of income level in the dynamics of population's distribution by income, *New Project*, in Russian (EERC Press, Moscow).

Bourguignon, F. and G.S. Fields (1990) Poverty Measures and Anti-Poverty Policy, *Recherches Economiques de Louvain* **56**(3–4), 409–427.

Bourguignon, F. and G. Fields (1995) 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).

Braithwaite, J. (1999) Casework and long-term poor in Russia. Report presented at the *World Bank Seminar*, April 19.

Buhmann, B., L. Rainwater, G. Schmaus, and T. Smeeding (1988) Equivalence Scales, Well-being, Inequality and Poverty: Sensivity. Estimates Across Ten Countries Using the Luxemburg Income Study Database, *Review of Income and Wealth* **34**, 115–142.

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

Cowell, F.A. and M. Mercader-Prats (1997) Equivalence of Scales and Inequality. *DARP Discussion Paper*, No 27 (STICERD, LSE).

Day, N.E. (1969) Estimating the Components of a Mixture of Normal Distributions, *Biometrika* **56**(3), 463–474.

Deaton, A. Angus (1992) *Understanding consumption*, *Clarendon Lectures in Economics(*Oxford University Press, Clarendon Press).

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

Ershov, E.B. and V.F. Maier (1998) Methodological problems of estimating level of income, revenue, and income differentiation, *Proceedings of the Russian Center of Quality of Life Statistics*, in Russian.

Esteban, J.-M., and D. Ray (1994) On the Measurement of Polarization, *Econometrica* **62**(4), 819–851.

Fajnzulber, P., D. Lederman, and N. Loayza (1999) Inequality and Violent Crime, The research project "*Crime in Latin America*" of the World Bank.

Foster, J., J. Greer, and E. Thorbeck (1984) A Class of Decomposable Poverty Measures, *Econometrica* **52**(3), 761–766.

Foster, J.E., and A.F. Shorroks (1988) Poverty Orderings, *Econometrica* **56**(1), 173–177.

Goscomstat R.F. (1996) Working Paper 1: Statistical policy and standards, *Goskomstat R.F. Printing Office*, in Russian.

Goscomstat R.F. (1998) Socioeconomic status of Russia, *Federal Commitee on Statistical Methodology*, I–XII, in Russian.

Goscomstat R.F. (1999a) Short-term socioeconomic indicators, *Goscomstat R.F. Printing Office*, in Russian.

Goscomstat R.F. (1999b) Welfare and quality of life in Russia, *Goskomstat R.F. Printing Office*, in Russian.

Gould, W. and W. Sribney (1999) Maximum Likelihood Estimation with STATA, *Stata Corp.*

Hagenaars, A. (1987). A Class of Poverty Indices, *International Economic Review* **28**, 583–607.

Jakimauskas, G. and J. Sushinkas (1996) Computational aspects of statistical analysis of gaussian mixture combining EM algorithm with non-parametric estimation (one-dimensional case), *Matematicos ir Informaticos Institutas* (Preprintas 96–6, Vilnius, Lithuania).

Kanbur, S.M.R. (1987) Measurment and Alleviation of Poverty, *IMF Staff Papers*, **34**, 60–85.

Kolenikov, S.O. (1999) Methods of Quality of Life Analysis, *Best Student Papers*, in Russian National School of Economics Series, Moscow.

Korchagina, I., L. Ovcharova and E. Turuncev (1999) System of indicators of poverty level in transition period. Report 98/04, in Russian. Section "Micro-2 (Households)," *Russian Program of Consortium of Economic Research* (EERC/Eurasia Foundation).

Little, R.J.A and D.B. Rubin (1987) *Statistical Analysis with Missing Data* (Wiley).

Ministry of Labor R.F. (1999) *Pilot Programs of Public Relief for Low-income Families in the Rebublic of Komi, Voronezh and Volgograd Regions*, Preliminary results (Moscow).

Mroz, T., B. Popkin, D. Mancini, T. Glinskaya, and V. Lokshin (1997) *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).

Proceedings of Goscomstat of Russian Federation (1999a) *Comments on methodology of estimation of per capita income distribution*, Report Presented at the Scientific Council session at the Center of Quality of Life Statistics, in Russian.

Proceedings of Goscomstat of Russian Federation (1999b) *Definition of integral indexes based on household budget survey data*, (Moscow).

Proceedings of the Higher School of Economics (1999) National evaluation and dissemination of information, *Structural reorganization of the system of public relief*, World Bank Project, Preliminary Report SPIL-2.2.2/11, in Russian.

Ravallion, M. (1994) *Poverty Comparisons* (Harwood Academic Publishers, Chur, Switzerland).

RLMS (1996) *The Russia Longitudinal Monitoring Survey: "Family Questionnaire" and "Sample of Russian Federation"*, Rounds V and VI. Technical Report (August–October).

Rudzkis, R. and M. Radavicius (1995) Statistical Estimation of a Mixture of Gaussian Distributions, in: *Acta applicandae Mathematical* **38**(1).

Sen, A.K. (1995) A Sociological Approach to Measurement of Poverty, *Oxford Economic Papers* **37**, 669–667.

Shevyakov, A.U. and A.Y. Kiruta (1999) *Economic Inequality, Welfare, and Poverty of Russian Population and Russia's Regions in Transition: Methods of Estimation and Analysis of Causal Relationships*, a EERC final report (Moscow).

Shlesinger M. (1965). About pattern recognition.- In: "*Reading Automatic Devices*" (Kiev, Naukova Dumka) 38–45, in Russian.

Suvorov, A.V. and E.A. Ul'anova (1997) Revenue of the Russian population: 1992–1996. *Problemy prognozirovaniya* (Moscow).