MULTI-SOURCE DATA APPROACH IN EVALUATIONS OF HOUSEHOLD STANDARD OF LIVING

The paper deals with the approach of multi-source data use to evaluate the indicators of standard of living of households including high income households which are usually associated with higher unit non-response rate in sample surveys. Two approaches that are based on complex use of several sources of information are presented. The first approach implies the formation of cumulative microdata set which is based on the results of Household Budget Survey for few periods. Such approach serves to assess the social and demographic profile of high income households. It is shown that the second approach presupposes the formation of a corrected microdata set with the use of the results of Household Budget Survey and additional source of information either as the data of System of National Accounts or the results of Salary Survey or the data of Fiscal Office. With the second approach it is possible to obtain more accurate evaluations of expenditures and incomes of households and of high income ones as well.

Keywords: multi-source data approach, data of Household Budget Survey, data of System of National Accounts, data of Salary Survey, fiscal data, high income households.
запропонованих підходів передбачає формування кумулятивного масиву даних мікрорівня, що грунтується на результататах вибіркового обстеження бюджетів домогосподарств за кілька періодів. Підхід використовується для оцінки соціально-демографічного профілю домогосподарств із високими доходами. Показано, що другий запропонований підхід передбачає формування скоригованого масиву даних мікрорівня на основі результатів вибіркового обстеження бюджетів домогосподарств і додаткового джерела інформації, зокрема даних системи національних рахунків, або результатів вибіркового обстеження відомостей про заробітну плату працівників підприємств, або даних фіскальної служби. Застосування другого підходу умовлюється отримання точніших оцінок показників рівня життя домогосподарств/населення, в тому числі домогосподарств із високими доходами.

Ключові слова: комплексне використання даних із декількох джерел, дані вибіркового обстеження бюджетів домогосподарств, дані системи національних рахунків, дані вибіркового обстеження відомостей про заробітну плату працівників, дані фіскальної служби, домогосподарства із високим рівнем доходів.

Н.Н. Романчук
канд. экон. наук, науч. сотруд.
Институт демографии и социальных исследований имени М.В. Птухи НАН Украины
01032, г. Киев, бул. Т. Шевченко, 60
e-mail: Romanchuk_Natalia@ukr.net

ИСПОЛЬЗОВАНИЕ НЕСКОЛЬКИХ ИСТОЧНИКОВ ДАННЫХ ДЛЯ ОЦЕНКИ ПОКАЗАТЕЛЕЙ УРОВНЯ ЖИЗНИ ДОМОХОЗЯЙСТВ

Рассмотрены основные подходы к использованию нескольких источников информации для оценивания показателей уровня жизни домохозяйств, в том числе домохозяйств со сравнительно высоким уровнем доходов, которые, как правило, отказываются участвовать в выборочных обследованиях. Изложено два подхода, основанные на комплексном использовании данных из нескольких источников информации разных уровней агрегации. Первый из предложенных подходов предполагает формирование кумулятивного массива данных микроуровня, на основе результатов выборочного обследования бюджетов домохозяйств. Метод предназначен для оценки социально-демографического профиля домохозяйств с высоким уровнем доходов. Показано, что второй предложенный подход предполагает формирование скорректированного массива данных мікрорівня, на основе результатов выборочного обследования бюджетов домохозяйств и дополнительного источника информации, таких как данные системы национальных счетов, или результаты выборочного обследования данных о заработной плате работников предприятий, или данные фискальной службы. Применение второго подхода делает возможным получение более точных оценок показателей уровня жизни домохозяйств/населения, в том числе, домохозяйств с высоким уровнем доходов.

Ключевые слова: комплексное использование данных из нескольких источников, данные выборочного обследования бюджетов домохозяйств, данные системы национальных счетов, данные выборочного обследования данных о заработной плате работников предприятий, данные фискальной службы, домохозяйства с высоким уровнем доходов.

Introduction. The majority of studies concerning standards of living of population and its various income groups are based on the analysis of household budget sample surveys. However, the general issue of the survey conducting is the constantly growing unit non-response rate, which significantly affects the quality of these survey results. Besides, the higher non-response rate is typical of high income households/individuals [1]. Hence, one of the problems of evaluation of these households’ standard of living is the lack of relevant and reliable information concerning their incomes and expenditures, welfare, etc. On this basis, the evaluations of population which are based on the results of the mentioned surveys are obviously biased towards the poorer households. The following problem is topical from the point of view of reliable and actual evaluations of whole population’s standard of living.

110

ISSN 2072-9480. Demography and Social Economy, 2017, № 1 (29)
**Analysis of recent publications** highlights the main approaches aimed at improving the reliability of data based on sample surveys results, in terms of both unit non-response bias reduction and approaches to evaluating indicators of household/population standards of living with the use of multi-source data.

The approach which is capable of improving the reliability of estimates and decreasing non-response rate implies oversampling of potentially high income households/individuals [2, p. 34–35]. Also, complex sample design, i.e. few sampling frames, could be applied for this purpose [3, p.147].

Additional widely used approach for verification of income evaluations based on results of budget sample surveys is the record linkage. It involves merging the survey data and tax records for the same period and the same person to verify the data of household incomes (Manfred, A. et al. (2014) [4].

The approach that uses multi-source data is applied for evaluation of the well-off’s expenditures as presented in Ogay, M. (2011) [5]. The paper shows the results of study with different data sources: results of Household Living Condition Survey, data of System of National Accounts, the information of Kyiv city household electricity consumption. In [5] the results of approach, the key indicators of standard of living are shown and compared but the author does not specify the methodology of implementation and limitations of such approach in detail.

Theoretical aspects of using data from various sources, including results of sample surveys are presented in Sarioglo, V. (2005) [6]. However, the paper does not refer to the use of multi-source data while studying small groups of households/populations, in particular the use of the most appropriate sources of information.

The lack of technical procedures of implementing the developed approaches which are presented in most publications significantly limits the possibility of their replication in other studies. The actual paper deals with presentation of detailed procedures for the implementation of the suggested methodological approaches in studies of household groups, which are characterized by a significant non-response rate in household budget sample surveys.

Thus, the **aim of this paper** is to present the methodological approaches in order to increase of reliability of assessment of indicators of household standards of living with the use of multi-source data. These approaches are developed to minimize the impact of high income households’ non-response when conducting the sample survey and they are based on disposable data from various sources of different levels of aggregation.

**Main results.** Based on the analysis of the recent publication, the additional information can be applied in order to increase the reliability of sample survey data and to reduce the unit non-response. It can be used at the following stages:

- data collection (e.g. when conducting a sample survey) to provide the correct representation of households/individuals with the required characteristics [1; 2; 3];
- during the study/evaluation with the use of disposable information (data which has been obtained already). It means that the data from various sources concerning the same subject (i.e. household incomes) could possibly improve the reliability of final evaluations [4; 5].

This paper aims to present the approaches to the more reliable assessment of indicators of household standards of living including high income households, which are based on the simultaneous and complex usage of disposable data from different sources of information. The specific feature of the complex use of multi-source data lies in the fact that additional data is used only for adjustment and improvement of the reliability of the main information. The necessary data are:
microdata (main source of information) — the results of household budget sample survey (hereinafter — HBS);
macrodata (additional source of information), for instance:
- fiscal data — data of State fiscal offices, e.g. centile distribution of the employed by their labour incomes;
- information concerning salaries of the employed based on the results of sample surveys of enterprises (salary sample surveys);
- data of household sector of the System of National Accounts — information concerning final consumption expenditures of households.

Employing the relevant statistical tools provides simultaneous use of main and additional sources of information which results in adjusted microdata sets. These datasets are used for more adequate complex assessment of indicators of standards of living of households/population (including high income households).

So, the use of multi-source data of different levels of aggregation provides the formation of new (adjusted) datasets. It can be achieved with two methodological ways (Fig. 1). They could be used separately or simultaneously, depending on a study’s aim.

The first approach presupposes the formation of cumulative microdata set by joining the results of HBS of several years. The formation of cumulative microdata set provides ad hoc increase of high income households’ records in dataset of HBS of the certain year through their accumulation within few periods. Increasing the number of observation units in a cumulative

---

**Fig. 1.** Methodological approaches for formation of the adjusted microdata set based on multi-source data

*Source: compiled by the author.*
dataset requires correction of system of statistical weights to ensure the correspondence of the results of survey to the population. Besides, the cumulative dataset is based on information of different time periods, so there is a necessity to adjust monetary variables, i.e. incomes, expenditures, etc., whose values are exposed to the impact of inflation.

The major reason to develop and apply this approach is the usually small number of households/individuals with high incomes in HBS dataset of a single period (year). The share of these households assessed by various criteria is 0.1%–1.4% of total number of households of Ukraine (evaluations are based on the results of the Household Living Condition Sample Survey, 2012). This means that for reliable evaluations of demographic, social and economic characteristics of these households the formation of cumulative dataset with enlarged number of their records is quite beneficial. On methodological basis such procedure is possible because of:

- annual sample rotation which makes it impossible to interview the same households repeatedly in a few year period;
- the consumption behavior of households for a few years could be defined as inertia;
- households’ standards of living have relatively low probability of significant changes during relatively short period of time (in a case of stable economic and political conditions).

The suggested methodological approach includes:

- increasing number of records of high income households in the dataset by the means of their accumulation in a single HBS microdata set;
- value adjustment of monetary indicators for a few periods within one dataset;
- correction of statistical weights of cumulative microdata set by applying special adjusting procedures.

The adjustment of monetary indicators implies the harmonization of amounts of household incomes, expenditures and other indicators which have monetary value. This adjustment procedure presupposes the harmonization of all monetary values with the baseline.

Depending on what the research is aimed at the monetary variables could be adjusted by the means of:

1. Consumer Price Index (hereinafter – CPI);
2. relative units.

The first type of adjustment should be used in an analysis, which requires the consideration of inflation impact on household incomes (i.e. their purchasing power). The leveling of inflation impact on the values of goods and services consumed by households during the studied periods as well as the purchasing power of household incomes needs additional information, such as CPI that reflects the changes of the monetary value of the consumer basket which includes a fixed set of consumption goods and services.

The procedure for adjustment with the use of relative units is simplified so there is no need for additional information, and it could be used in rough analysis (see Fig. 2).

Algorithm for formation of the cumulative dataset is based on three adjacent periods and presupposes the adjustment of monetary variables with the use of CPI includes the following stages:

1. Harmonization of the monetary variables for the base period prices.

At this stage the baseline values of household incomes, expenditures, etc. are determined with the use of CPI. For the three periods the calculations are as follows:
Fig. 2. Scheme of formation of cumulative microdata set with the use of A) CPI; B) relative units

Source: compiled by the author.

*X1, X2, ..., Xn — qualitative variables
**Y1, Y2, ..., Yn — monetary variables
Multi-Source Data Approach in Evaluations of Household Standard of Living

\[ h^*_t = h^{t-2}_i \ast \bar{NP}I_i^{t-1} \ast CPI_i^t \]  
\[ h'^*_t = h'^{t-1}_i \ast CPI_i^t \]  

where \( t \) – baseline period;  
\( t - 1, t - 2 \) – prior periods for baseline \( t \);  
\( h^*_i \) – monetary variable of \( t \)-th household which is expressed in the base period prices;  
\( h^{t-1}_i, h^{t-2}_i \) – the values of monetary variable of \( t \)-th household which is expressed in the prices of \( t - 1, t - 2 \) periods;  
\( CPI_i^{t-1}, CPI_i^t \) – CPI of baseline and \( t - 1 \) period;

2. Identification and extraction of high income households records from the microdata sets.

The application of this approach involves prior selection of a criterion for high income household identification in order to define the target population group. Such criterion reflects the level of income, expenditures, etc. or the share of population which, according to the researcher, identifies high income households. Accordingly, the records of high income households/individuals identified by the use of this criterion are extracted from datasets for all periods (but not from the baseline one).

3. Increase of the number of records within dataset of the base period.

Information, which is extracted at the previous stage and represents the households/individuals who meet the defined criterion, should be attached to the microdata set of the base period.

4. Adjustment of statistical weights of base period dataset.

The increased number of records on high income households/individuals in the baseline dataset presupposes adjusting the statistical weights of the other records. This stage substantiates that the assessments of cumulative dataset (the number of households/individuals, their distribution by region, type of area) meet the population estimates of the base year. The method of iterative approximation or the calibration of dataset weights can be used for the adjustment procedure (see Fig. 2A).

In order to use variables expressed in relative units, the relative values (ratios) of monetary indicators of each record (household/individual) are calculated to make the use of information possible for different periods in one cumulative dataset (see Fig. 2B).

Algorithm for construction of cumulative microdata set with the use of relative units is provided by the following stages:

1. Calculation of mean values for each of the monetary variables.

Within the microdata sets of each studied period the mean value of the monetary indicator for all the records (households or individuals) should be defined.

2. Ratios construction.

Mean values calculated at the previous stage are used for calculation of relative values for each record. The mentioned calculations are as follows:

\[ h'_{i,rl} = \frac{h_{i,rl}}{m_{rl}} \]
\[ h'_{i,t2} = h'_{i,t1} \cdot g_{m_{t2}} \]

where \( h'_{i,t1} \) — ratio of \( h^t \) indicator for the \( i \)-th household in \( t1 \) period;

\( h^t \) — value of \( h^t \) indicator value for the \( i \)-th household in \( t \) period;

\( m_{t1}, m_{t2} \) — mean values of \( h^t \) indicator in \( t1, t2 \) periods, respectively;

\( h'^{*,t2} \) — reconstructed value of \( h^t \) indicator for the \( i \)-th household for \( t2 \) period.

3. The identification and extraction of high income households from the microdata sets. Methodologically, the stage corresponds to p. 2 of the previous algorithm.

4. The information on households which are extracted at the previous stage is joined to the base period microdata set. The stage reflects the p. 3 of the previous algorithm.

5. The reconstruction of absolute values of monetary indicators in the cumulative dataset. They are calculated as multiplication of the relative values (ratios) and mean value defined for base period (see Fig. 2B).

Application of the above-mentioned approach is appropriate when analyzing characteristics of small groups of population, including high income households, when characteristic evaluations of these groups, e.g. the region of residence, the size and composition of the household, type of residence area, employment, level of education etc., vary substantially which are assessed on the basis of sample survey data.

The second approach provides the formation of adjusted microdata set based on harmonization of multi-source data: information from main data source (microdata) and additional information (see Fig. 1). The implementation of this approach is aimed at harmonizing data from different sources and different levels of aggregation with the use of special statistical tools. The use of calibration procedure of statistical weights is the most beneficial. In general, the idea of the procedure is to transform the statistical weights of HBS dataset using external (additional) information.

As already mentioned, data of SNA, fiscal information or results of salary sample survey may be used as additional information. Selection of one of these information sources depends on the purposes of the research, the reliability of the information source, the coverage of the studied population. Besides, sources must have the same or similar period and regularity of publication and be available for users etc.

According to the results of previous studies, the most accessible data source for this approach is from SNA. However, the use of this source is often doubtful in terms of methodological comparability of the HBS and SNA data in case they are the sources of information which differs substantially in methodological nature. Sample surveys data are received from respondents (household members), SNA data are aggregate information from various branches of statistics that is balanced within the whole economy. In order to maximize the compatibility of these information sources, imputed rent, subsidies of population should be excluded from SNA data. In case of the SNA as is used additional information source, the system of statistical weights of the HBS microdataset is calibrated according to the structure of household final consumption expenditures of SNA.

The practical implementation of the approach with the use of SNA as additional information during the evaluations of households welfare including incomes and expenditures of
high income households [8] show a significant decline of the quality of statistical weights, provided when the household expenditures’ structure of the SNA and one of the HBS do not substantially correspond each other because of several reasons mentioned above. So, the gap between main data and additional information can be «artificially» reduced to improve the quality of weights after calibration.

The results of the salary sample survey can be also used as a source of additional data as well. In this case, workers’ salary data from survey results need to be adjusted for the amount of tax payments and deductions, because HBS records contain net income of individuals (employees). When it is necessary to use the results of salary survey, the calibration procedure of statistical weights of HBS microdata set agrees with the decile employees’ salary distributions which are evaluated on the base of the results of two surveys.

It is beneficial to use fiscal data as additional information such as centile income distribution of tax payers (defined by place of their registration) grouped by region. Amounts of fiscal data on salary taxes and fees must be reduced to harmonize two types of data methodologically. The application of calibration procedure adjusts the employees’ salary distributions derived from fiscal data and HBS results.

When the latter two sources of information are used, the application simultaneously of two sources of data (HBS results and salary survey data or HBS results and fiscal data) that are characterized by the different statistical unit – household and individual – is possible. For the purpose of leveling of the mentioned methodological inconsistencies of data sources the estimations of employees’ salary distribution (deciles) of additional data are used. These estimates are compared with the distribution of employed household members (individuals) that are aggregated into households whose weights are calibrated at the next stage.

It is sufficient to note that the use of the mentioned data sources as additional information gives a close estimation results. In particular, the table below shows the comparison of the Gini index estimates obtained by means of adjusted microdata set and the results of Household Living Conditions Sample Survey (hereinafter — HLCS).

Table. Gini index evaluated by the adjusted data set and the results of HLCS

<table>
<thead>
<tr>
<th>Additional data source used in calibration procedure</th>
<th>Data source for Gini index evaluation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adjusted dataset</td>
<td>HLCS results</td>
</tr>
<tr>
<td>SNA*</td>
<td>0.27</td>
<td>0.23</td>
</tr>
<tr>
<td>Salary Survey*</td>
<td>0.26</td>
<td>0.23</td>
</tr>
<tr>
<td>Fiscal data**</td>
<td>0.24</td>
<td>0.23</td>
</tr>
</tbody>
</table>

*Y2012 data  
**Y2015 data  
Source: calculated by the author

Calibration procedure is defined methodologically as procedure of re-weighting used to incorporate auxiliary information. The calibrated weights are obtained by minimizing the distance between the calibration weights and weights that are calibrated, provided that the calibration weights give more reliable estimates of the auxiliary variables for population [7].

When the results of salary survey data as additional information are used, the problem of calibration could be defined as the follows:
where \( d_i \) — statistical weights of \( i \)th household before calibration, \( i=1,2,\ldots, n \); 
\( n \) — population; 
\( w_i \) — statistical weights of \( i \)th household after calibration, \( w_i = g_i d_i \); 
\( \text{dec}_1, \text{dec}_2, \ldots, \text{dec}_{10} \) — salary of employed household members, defined by 1st — 10th deciles of employees’ salary distribution which is defined using HBS results, where \( i \) — the number of household; 
\( \text{Dec}_1, \text{Dec}_2, \ldots, \text{Dec}_{10} \) — amounts of salary of household members grouped by decile which are defined by employees’ salary distribution based on the results of salary sample survey.

It is sufficient to note that some aspects of the second approach should be highlighted. It must be emphasized that during the evaluation procedure, initial data (records) of the sample survey are not changed; procedure implies only modification of statistical weights of surveyed households. The shortcoming of the approach application is the potential increase of variation of weights and, therefore, the decreasing of reliability of evaluations. Calibration
procedure changes high income households’ weights to the larger ones in order to «increase» the number of high income households in population. But it does not increase the amount of information concerning high income households (i.e. number of their records is unchangeable after calibration). Thus, any social and demographic characteristics of this group of households assessed for the population can vary considerably because of their relatively small number in the survey sample. Accordingly, the reliability of conclusions based on such evaluations is usually insufficient.

So, the procedure of harmonization of multi-source data should be conducted with mandatory quality control of statistical weights which takes place at every stage of gradual implementation of procedure. Evaluation of the quality of statistical weights at each stage of procedure implementation makes it possible to stop calibration if their quality is reduced significantly.

In total, the construction of adjusted microdata set by means of this approach provides more accurate evaluations of basic indicators of household standards of living including high income households.

To sum up, to evaluate some indicators of standard of living of households and also some groups of households which are difficult to survey it is beneficial to use multi-source data. Depending on what the study is aimed, the approaches, presented above as well as the data sources to be used in the evaluation are to be selected. However, it is obligatory to take into account the shortcomings and restrictions imposed by the use of the mentioned approaches.

Conclusions. The lack of reliable information for the research and evaluation of various social and economic phenomena is the problem for major statistical studies. The information concerning the study of population/household standard of living is microdata such as results of household budget sample survey, administrative etc.

One of the approaches referring to evaluating of complex social and economic structures and processes, including incomes and expenditures of high-income households, is the use multi-source data of different levels of aggregation.

Two methodological approaches for the assessment of incomes and expenditures are presented. Their application depends on the research aims and disposable data sources.

In order to determine social and demographic characteristics of high income households or to analyze their specific expenditures the cumulative microdata set based on household budget survey results of a few years is used. The formation of the microdata set presupposes the adjustment of monetary variables of different periods (with the use of Consumer Price Index or relative units). The main disadvantage of such approach is «artificial» increase in the number of records of high income households in a dataset for a period. This implies the addition of information (records) into dataset which is not taken during the survey of a particular year. This procedure makes the study more subjective because it is the researcher who determines the criteria and reasons whether to choose and add information to the dataset.

The approach of cumulative microdata set formation is suggested to evaluate and analyze incomes and expenditures of households/population and their groups as well. Such approach implies the harmonization of multi-source data of different levels of aggregation: main source of information — results of household budget survey (microdata) and additional (auxiliary) information — aggregated macro level data (e.g. data of System of National Accounts or results of sample survey of employees’ salary, fiscal data) with the use of calibration procedure of statistical weights of survey microdata set. The advantage of application of the mentioned procedure is the absence of changes for the microdata set records — the only modification is applied to the system of statistical weights. On the contrary, the calibration procedure can
lead to increase of variance of weights which reduce the reliability of final estimates that requires mandatory quality control of statistical weights for every stage of procedure implementation. Besides, the application of the mentioned additional data source gives similar results of assessments of the standard of living indicators of households which are evaluated on the base of cumulative microdata set.

The application of the suggested approaches for studying of incomes and expenditures of households and their groups (e.g. high income households) demands careful examination of all data sources to be used and statistical tools to be applied to clearly understand the final results of research and main restrictions which presuppose the analysis.

LITERATURE


REFERENCES


