Improving of Household Sample Surveys Data Quality on Base of Statistical Matching Approaches

Ganna Tereshchenko
Institute for Demography and Social Studies
of the National Academy of Sciences of Ukraine,
26, Panasa Myrnogo str., Kyiv, 01011, Ukraine
a_tereschenko@ukr.net

1. Introduction

Ukrain's labour force indicators are defined on the basis of the information obtained from different sources – State Sample Survey of Population Economic Activity, State Service of Employment, Sample Survey of Households Living Conditions and All-Ukrainian Population Census of 2001.

The different sources of information vary in essence because they reflect different aspects of the social phenomena and are characterized by different structure of the information and periodicity of its actualization. The population surveys are also carried out on samples of different design. As the result, when calculating the statistical indicators, the data from these sources is used mainly independently and isn't mutually coordinated, which results in essential reduction of information use efficiency. It has especially negative consequences for the state statistics, the main task of which isn't the measurement of separate aspects of society life but creation of the relevant, accurate, full and timely information for political and public organizations, scientific institutions, etc. The problems of development of methodological maintenance of the Statistical matching from different sources and estimation of matching data are important first of all for the state statistics.

2. Main objectives of the work

In this work the main attention is given to the question of LFS data Quality improving obtained on the on the basis of two samples with different design.

3. Matching of LFS data, received on the basis of two samples with different design

Labour force indicators in Ukraine are measured by results of households sample survey of population economic activity (LFS). Since 1995 LFS is conducted by State Statistics Committee of Ukraine: in 1995–1998 once a year, in 1999–2003 – quarterly, since 2004 – monthly. In LFS the population surveyed is aged 15–70. LFS sample covers all regions of Ukraine by the type of settlements: urban area (cities, towns) and rural area. The size of a monthly LFS sample is 32,5 thousand of surveyed households. Indicators of economic activity, employment and unemployment are measured in survey according to the international standards by ILO methodology.

Reliability of the unemployment rate in rural area by regions until 2004 was much lower than for the regions on the whole.

On fig. 1 the variation coefficients of unemployment rate annual estimates of Ukraine for the year 2003 are resulted.

To improve reliability of employment and unemployment estimates by regions of Ukraine in rural area since 2004 the survey is carried out on the of two probability stratified two stage samples: sample of LFS and sample of household agricultural activity survey (AAS). Interview of households on questions of economic activity in

the AAS is carried out under the identical programmer, as in the LFS, but the sample design in AAS is differ from LFS. In AAS households are selected in the second stage with probability proportionally to their area of agricultural allotment, in LFS - on the basis of the procedure of systematic selection.

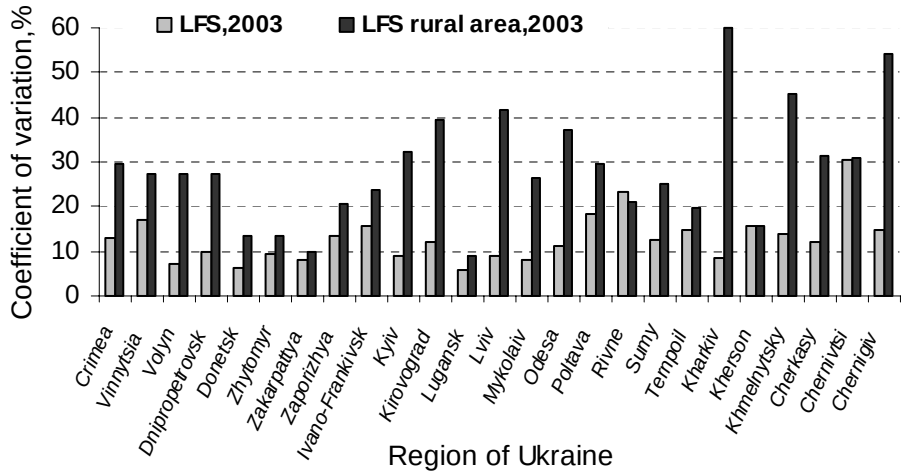

Fig. 1. Reliability of unemployment rate annual estimates

The size of monthly LFS sample in the rural area makes approximately 3,6 thousand households, and the size of AAS sample of households which have to be interviewed under LFS questionnaire is 7,4 thousand households. Total size of a monthly sample for interview under LFS questionnaire in the rural area due to AAS has increased three times and is equal to 11,1 thousand households.

Based on the results of the primary data analysis of LFS and AAS we can establish, that surveys data is mutually coordinated; in the February of 2007 in particular, the coefficient of correlation for employment rate is equal 0,79, and for unemployment rate – 0,77. The similar picture is observed throughout all months of 2007.

Based on the results of the primary data analysis of LFS and AAS ??? have established, that surveys data is mutually coordinated (fig. 2); in the February of 2007 in particular, the coefficient of correlation for employment rate is equal 0,79, and for unemployment rate - 0,77. The similar picture is observed throughout all months of 2007.

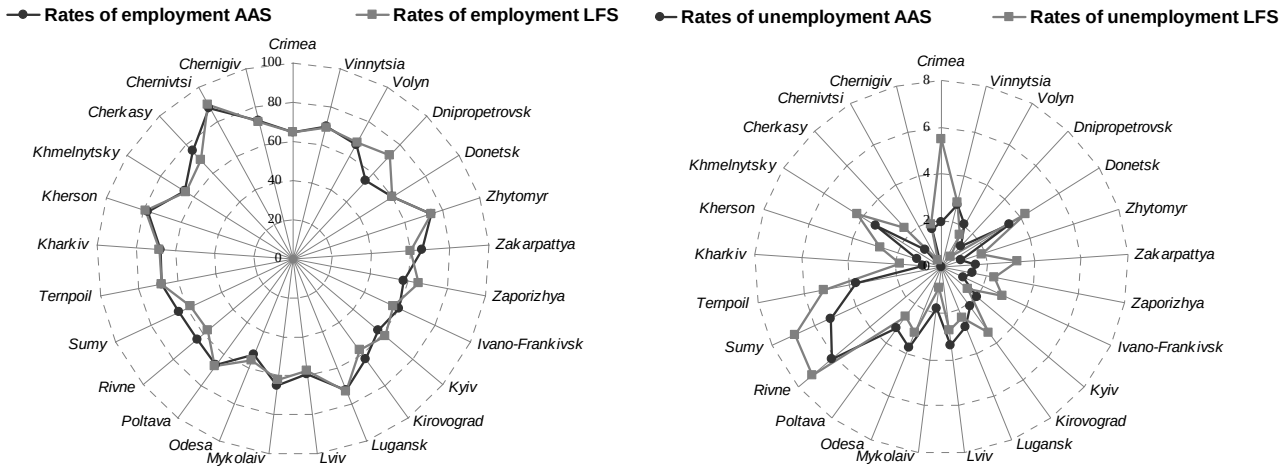

Fig. 2. Rates of employment and unemployment by regions of Ukraine, February, 2007

Matching of LFS data, obtained on the basis of two different samples is carried out with the use of data merging at the microlevel method (Marcello D'Orazio, Marco Di Zio, Mauro Scanu, 2006). This approach is used in conditions when, from different sources the data is obtained in units of an identical level (in this case – on household members aged 15–70) and by identical attributes (under identical questionnaires). The result of data merging is the file where we have observations from both - LFS file and AAS file.

The employed and unemployed population is calculated by formula for composite estimation:

$$\hat{Y}_{em} = \hat{\phi}_{em}\hat{Y}_{em}^{(LFS)} + (1-\hat{\phi}_{em})\hat{Y}_{em}^{(AAS)}$$

$$\hat{Y}_{un} = \hat{\phi}_{un}\hat{Y}_{un}^{(LFS)} + (1-\hat{\phi}_{un})\hat{Y}_{un}^{(AAS)} \quad (1)$$

where – $\hat{Y}_{em}^{(LFS)}$ – estimate of employed population number on LFS sample;

$\hat{Y}_{em}^{(AAS)}$ – estimate of employed population number on AAS sample;

$\hat{Y}_{un}^{(LFS)}$ – estimate of unemployed population number on LFS sample;

$\hat{Y}_{un}^{(AAS)}$ – estimate of unemployed population number on AAS sample.

Based on the performed researches it has been established that direct estimates of the employed and unemployed population number, which are obtained on the basis of AAS sample are biased. The estimates of bias for the employed and unemployed population number for the current month in this work were defined as an average bias for current month and the previous two.

During file matching of AAS with LFS unit weights, calculated separately on each file are corrected for rural area on each region with the use of coefficients ϕ_{em} for employed population and ϕ_{un} – for unemployed (Sarioglo, 2005):

$$\hat{\phi}_{em} = \frac{SE^2(\hat{Y}_{em}^{(AAS)}) + \overline{B}^2(\hat{Y}_{em})}{SE^2(\hat{Y}_{em}^{(LFS)}) + SE^2(\hat{Y}_{em}^{(AAS)}) + \overline{B}^2(\hat{Y}_{em})} \; for\ employed\ \ persons$$

$$\hat{\phi}_{un} = \frac{SE^2(\hat{Y}_{un}^{(AAS)}) + \overline{B}^2(\hat{Y}_{un})}{SE^2(\hat{Y}_{un}^{(LFS)}) + SE^2(\hat{Y}_{un}^{(AAS)}) + \overline{B}^2(\hat{Y}_{un})} \; for\ unemployed\ \ persons \quad (2)$$

where $SE(\hat{Y}_{em}^{(LFS)})$ – standard error of estimate of employed population number $\hat{Y}_{em}^{(LFS)}$ on LFS sample;

$SE(\hat{Y}_{em}^{(AAS)})$ – standard error of estimate of employed population number $\hat{Y}_{em}^{(AAS)}$ on AAS sample;

$SE(\hat{Y}_{un}^{(LFS)})$ – standard error of estimate of unemployed population number $\hat{Y}_{un}^{(LFS)}$ on LFS sample;

$SE(\hat{Y}_{un}^{(AAS)})$ – standard error of estimate of unemployed population number $\hat{Y}_{un}^{(AAS)}$ on AAS sample;

$\overline{B}(\hat{Y}_{em})$ – the bias of estimate of number of employed population;

$\overline{B}(\hat{Y}_{un})$ – the bias of estimate of number of unemployed population.

The corrected statistical weights of employed and unemployed persons in rural area of each region are calculated by the formula:

$$w_i' = w_i \cdot k_i \quad (3)$$

Correction coefficient of statistical weights system k_i is calculated for each region in two stages.

On the first stage the value of k_i is calculated for employed and unemployed persons in rural area:

$$k_i = \begin{cases} \hat{\phi}_{em} \ \ for\ employed\ person\ on\ LFS\ sample \\ \hat{\phi}_{un} \ \ for\ unemployed\ person\ on\ LFS\ sample \\ (1-\hat{\phi}_{em}) \quad for\ employed\ person\ on\ AAS\ sample \\ (1-\hat{\phi}_{un}) \quad for\ unemployed\ person\ on\ AAS\ sample \end{cases} \quad (4)$$

On the second stage the value of k_i is calculated for economically inactive persons in rural area for each region by the formula:

$$k_i = \frac{N_{15-70} - (\hat{Y}_{em}^{'(LFS)} + \hat{Y}_{em}^{'(AAS)} + \hat{Y}_{un}^{'(LFS)} + \hat{Y}_{un}^{'(AAS)})}{\hat{Y}_{ei}^{(LFS)} + \hat{Y}_{ei}^{(AAS)}} \quad (5)$$

where N_{15-70} – total number of able-bodied population in rural area of region, calculated on external data;

$\hat{Y}_{em}^{'(LFS)}$ – estimate of employed population number on LFS sample in view of corrected statistical weights w_i';

$\hat{Y}_{em}^{'(AAS)}$ – estimate of employed population number on AAS sample in view of corrected statistical weights w_i';

$\hat{Y}_{un}^{'(LFS)}$ – estimate of unemployed population number on LFS sample in view of corrected statistical weights w_i';

$\hat{Y}_{un}^{'(AAS)}$ – estimate of employed population number on AAS sample in view of corrected statistical weights w_i';

$\hat{Y}_{ei}^{'(LFS)}$ – estimate of economically inactive population number on LFSP sample in view of corrected statistical weights w_i';

$\hat{Y}_{ei}^{'(AAS)}$ – estimate of economically inactive population number on AAS sample in view of corrected statistical weights w_i'.

It should be noted, that described procedure is used only for sample units in rural area.

4. Quality of matching data

If we want to compare reliability characteristics of employment and unemployment rates estimates for regions of Ukraine on the matched data file and LFS data file, it is necessary to draw the conclusion that data matching allows essential improvement in the reliability characteristics of employment and unemployment rates estimates in rural area.

As it is clear from fig. 3, for the majority of regions CV of employment rate estimates became less than 5 percent.

It should also be noted that for 8 regions CV of unemployment rate estimates has decreased twice and for 3 regions - three times (fig. 4). But data matching does not improve the reliability for all regions. For 6 regions variation coefficient of unemployment rate estimates in February of 2007 has decreased by less then 5 per cent, whereas for other regions variation coefficient has decreased on average by 23 per cent. At the same time when using statistical matching, the amount of calculation increases dramatically which can have a negative impact on the timeliness and accessibility of the data. If we abolish implying statistical matching for the mentioned 6 regions, where the reliability increase is not essentiall than the effectiveness of the

method on the whole will increase, but a rather complicated problem of the specific explanation of the obtained estimates comparability will arise.

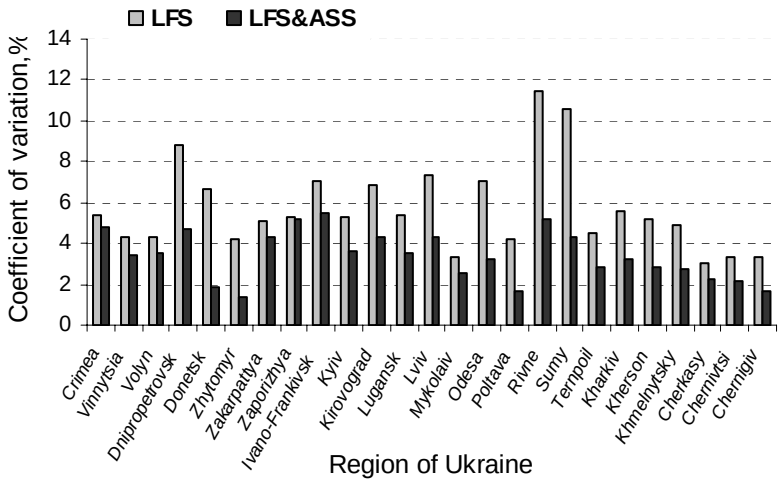

Fig. 3. Reliability of employment rate monthly estimates in rural area before and after statistical matching of the LFS data, February, 2007

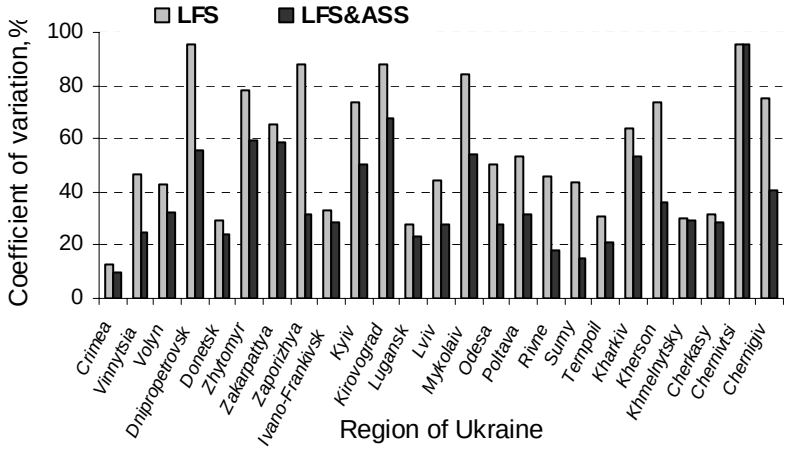

Fig. 4. Reliability of unemployment rate monthly estimates in rural area before and after statistical matching of the LFS data, February, 2007

The main factor for comparability infringements in this case is the differences of statistical methodology. In this case it's recommended to calculate and analyze special indicator – the share of incomparable estimates. This indicator calculated separately for each region is equal "0" if the estimates obtained on the basis of data matching, is equal "1" if the estimates obtained on the basis of data from only one of surveys (Tabl.1).

Share of incomparable estimates:

$$R_c = \frac{\sum_{i=1}^{N} Di}{N} = \frac{6}{25} = 0{,}24$$

where N – number of regions,

$$D_i = \begin{cases} 0, \ if \ \ estimates \ \ for \ \ region \ i \\ are \ received \ by \ \ LFS \ \& \ ASS \ matched \ \ data; \\ \\ 1, \ if \ \ estimates \ \ for \ \ region \ i \\ are \ received \ by \ \ LFS \ data \ \ only. \end{cases}$$

Table 1. Share of incomparable unemployment rate estimates by regions of Ukraine

Region		Type of data			Type of data
Crimea	1	LFS	Mykolaiv	0	LFS&ASS
Vinnytsia	0	LFS&ASS	Odesa	0	LFS&ASS
Volyn	0	LFS&ASS	Poltava	0	LFS&ASS
Dnipropetrovsk	0	LFS&ASS	Rivne	0	LFS&ASS
Donetsk	0	LFS&ASS	Sumy	0	LFS&ASS
Zhytomyr	0	LFS&ASS	Ternpoil	0	LFS&ASS
Zakarpattya	0	LFS&ASS	Kharkiv	0	LFS&ASS
Zaporizhya	0	LFS&ASS	Kherson	0	LFS&ASS
Ivano-Frankivsk	1	LFS	Khmelnytsky	1	LFS
Kyiv	0	LFS&ASS	Cherkasy	1	LFS
Kirovograd	0	LFS&ASS	Chernivtsi	1	LFS
Lugansk	1	LFS	Chernigiv	0	LFS&ASS
Lviv	0	LFS&ASS			

A special analysis of relative efficiency was carried out (fig. 5). For example in February, 2007 the average efficiency of statistical matching procedure by regions of Ukraine was 0,41 for estimates of employed population number

$$(reff_{em} = \frac{V(\hat{Y}_{em}^{(LFS\&AAS)})}{V(\hat{Y}_{em}^{(LFS)})} = 0{,}41)$$

and 0,48 for estimates f employed population number $(reff_{un} = \frac{V(\hat{Y}_{un}^{(LFS\&AAS)})}{V(\hat{Y}_{un}^{(LFS)})} = 0{,}48)$.

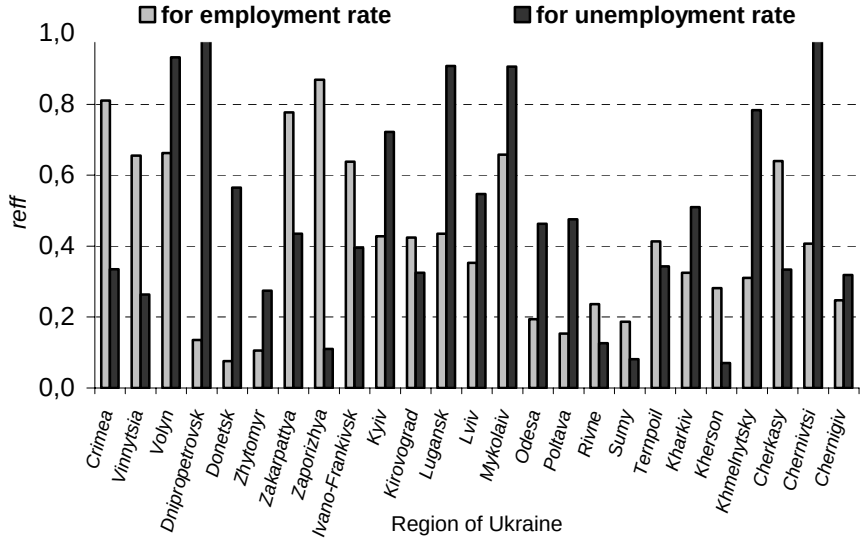

Fig. 5. Relative efficiency of matching procedure by regions, February, 2007

5. Conclusion

Statistical matching of the labour force survey data, obtained on samples with different design has allowed improving the reliability level of employment and unemployment indicators estimation in rural area of Ukraine. The methodological problem of adequate determination of indicators estimates bias value is revealed; it is

connected with differences of populations and with features of the separate surveys organization.
At the same time there is a potential problem with providing data comparability.
It is necessary to take into account that the volume of information for processing grows and estimation procedures are becoming complicated.
The developed approaches are implemented into the state statistics system of Ukraine. Researches in these directions are still continuing.

References

1. Marcello D'Orazio, Marco Di Zio, Mauro Scanu (2006). Statistical Matching: Theory and Practice / Marcello D'Orazio, Marco Di Zio, Mauro Scanu. – John Wiley&Sons, 2006.– 256 p.

2. Sarioglo V.G. Problems of the sample data statistical weighting. – Kyiv: State Statistics Committee of Ukraine, 2005. – 264 p.