European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May-3 June 2016

The Use of Regression Models in Labour Market Flow Statistics

Hannah Kiiver,[1] Frank Espelage[2]

[1] *Eurostat, Luxembourg; Hannah.Kiiver@ec.europa.eu (corresponding author)*
[2] *Eurostat, Luxembourg; Frank.Espelage@ec.europa.eu*

Abstract

The publication of breakdowns of labour market flow statistics based on quarterly matched samples of the EU-LFS are hampered by small sample sizes. In this paper, we propose to use regression methods to extract as much of the desired information as possible from the data, keeping the methodology simple enough to present results to the general public alongside the traditional statistics already published by Eurostat.

Keywords: labour market flows data, regression analysis

1. Motivation

In October 2015, Eurostat published for the first time labour market flow statistics using individual country Labour Force Survey (LFS) data, covering the transitions between the three ILO labour market statuses of employment, unemployment and inactivity for the population aged 15-74, broken down by sex. This exercise, which is done by Eurostat in the absence of longitudinal weights calculated directly by countries themselves, requires the matching of quarterly LFS data at the level of the individual respondent to exploit the realised overlap in the sample. Consequently final weights are recalibrated in order to gross up the subsample to the population totals, while respecting the distribution of individuals over the labour market statuses in the two quarters. The size of the matched subsample ranges from below 40% to above 80% percent of the average quarterly sample size. Even at the current breakdown by sex only, recalibration of weights is marred by zeros in the transition matrices of small countries with small quarterly LFS samples. In addition, approximately 10% of data points so far published cannot be shown due to confidentiality rules related directly to the sample size of the derived statistic (see table 1 for an overview). This does not include data points which might be unreliable, based on available quarterly reliability limits. Further analysis, driven by high demand for additional breakdowns of this data, revealed that calibrating weights at higher levels of breakdown will not be feasible for smaller countries; this in itself might not necessarily be a problem, if requirements of matching flows to stocks for breakdowns were to be relaxed. However, even adding one further breakdown by age (15-24, 25-54, 55-74) or education (ISCED0-2, ISCED3_4, ISCED5_8) using the weights derived from matching only the marginal distributions for totals will result in a sizeable increase of data points which cannot be published due to low sample size. Crucially, these are usually the data points of interest, i.e. the ones for changes in the ILO labour status. In small countries with small sample sizes, the situation is considerably worse than for the average, as seen in the last column of table 1.

Table 1: Share of confidential data points of flows data, by breakdown, 2015Q1-Q2

No data is available for DE and BE; data for EE, LU and MT is shown separately to illustrate impact on data of small countries

Breakdown variable	**Total # of breakdowns**	**% confidential data points**		
		26 MS - total	**26 MS – change of ILO status**	**EE/LU/MT - total**
Age groups	3	12	17	50/41/56
Age groups, sex	6	21	30	71/61/57
Education, sex	6	25	36	67/69/59
Duration of unemployment, sex (sample restricted to flows from unemployment in initial period)	8	26	36	91/95/90

This constraint seriously hampers the work of analysts and policymakers who are generally interested in the circumstances of those individuals who do change status, as well as in much more rigorously defined subgroups, to understand, for example, the labour market transitions of the long-term unemployed, of individuals close to retirement, or of those young individuals who are neither in employment nor in any education nor training. Over time, solutions to this problem can and should be found at country level by e.g. optimising sample overlap or adopting appropriate modelling techniques; in fact, Eurostat has explicitly stated the long-term goal of producing micro-data including longitudinal information and longitudinal weights, which then should be made available for analysts to extract the relevant information directly. However, in the meantime, Eurostat is expected to deliver some information on the groups of interest and their relative transition probabilities for relevant flows. In the absence of relevant auxiliary information to directly estimate the flow statistics using e.g. small-area estimation, we propose in this paper to use simple regression techniques to extract as much as possible of the desired information from the data. This method also has limits, and results for small countries with small samples should be treated with extreme caution.

While the introduction of regressions can thus partially fill the current detailed information gap, a whole new set of issues arises. LFS based flows data compiled by Eurostat is currently presented in predefined data tables in levels, or as outflows of each status expressed as share of the initial quarter status. Explaining the meaning of log odds ratios, marginal effects or predicted probabilities derived from regression results, recalling their limits and indicating how to use them is a particular challenge for Eurostat, as results have to be presented in a digestible manner for all countries for which data is available.

This paper discusses these questions, and provides potential solutions found for the case of labour market flow statistics. Using the transitions from unemployment to employment as an example, the main results and their implications and interpretations are examined; furthermore, the paper proposes a format in which regression-based results can be presented alongside the existing statistics without neglecting either the cross-country comparability or the comparability over time.

2. General methodological issues

The use of modelling in statistics is neither new nor particularly controversial. In the case of survey data, the use of modelling in production may not be as obvious as it is e.g. in flash estimates of (macro-level) data; however it is used, as in many other domains, in several parts of the production process. This means that even if tabulated statistics may be perceived as being free from the influence of assumptions and decisions necessarily made in modelling, this is not actually true – one example might be the choice of a data producer to correct for non-response bias or not, and with which model. Even so, the use of regression analysis on survey data is usually the realm of researchers intending to answer specific questions; methods are selected specifically to answer these questions, and to establish causalities. Results are generally not meant to replace descriptive statistics.

While the use of modelling in itself is thus not as questionable as it might have appeared initially, it is important to stress to users the limits of the results: contrary to research, here regression is used instead of summary statistics; models are chosen for simplicity and, in the context of Eurostat's commitment to EU-wide comparable statistics, cannot be tailor-made for each country. Results may not be presented in the same way as in an academic paper, since both the aims of the analysis and the audiences are different.

Another generally relevant issue related to regression analysis on survey data is the use of weights. Generally speaking, weights are not strictly necessary in regression analysis of survey data, in particular if the variables used for calibration are part of the set of explanatory variables. However, using incorrect weights, or using weights incorrectly, may lead not only to non-negligible changes in estimated coefficients in comparison with an unweighted regression, but most importantly to an underestimation of standard errors. While most software packages provide solutions to take the whole sample design into account, thus applying correct weights in the correct fashion, this requires full information on the sample design as well as the availability of a large number of technical variables.

Related to both points mentioned above is the interpretation and presentation of regression results. Depending on the method and specification used, regression coefficients cannot easily be interpreted without being familiar with the method used; depending on the knowledge of users, diagnostics statistics on e.g. goodness of fit may confuse rather than enhance understanding; and finally, the full provision of standard errors, especially in the context of Eurostat's data offer, where none of the standard indicators derived from survey data is published with information on variance, may lead users to believe that tabulated data presented in the traditional fashion are "true" population figures rather than estimates with their own variance.

In the following sections, our approach to all three discussed issues is presented, using the flow from unemployment to employment as an example.

3. Method and results

The most basic flow statistic derived from the LFS, and published by Eurostat, is the 3x3 matrix of transitions between the three ILO statuses (see table 2, here expressed as transition probabilities).

Table 2: Transitions in labour market status in 26 EU MS, Q1-Q2 2015
(in % of initial status; population aged 15-74)

	Employment Q2 2015	Unemployment Q2 2015	Inactivity Q2 2015
Employment Q1 2015	97.1%	1.3%	1.6%
Unemployment Q1 2015	18.6%	64.6%	16.8%
Inactivity Q1 2015	3.0%	3.7%	93.3%

Following the reasoning in section 2, we refrain from modelling all transitions at the same time, using e.g. some version of a conditional logit; instead, and in parallel to what would usually be published using breakdowns, we look at each transition separately. In the example we present, we discuss the flow from unemployment in the initial period to employment in the second period. The main example presented is that for duration of unemployment, which is the main variable users think to be relevant and interesting for further breakdowns. Ideally, we would like to know in how far transition probabilities differ depending on the duration of unemployment, and if further variables like sex, education, and age matter. We use a simple logit model and apply it to the flow from unemployment to employment. The set of explanatory variables is chosen not depending on an underlying economic model, but instead is driven by the demand for specific breakdowns mentioned above as well as the availability of data in the LFS, i.e. there is no use made of external sources.[1] The relevant advantage of the regression approach is that it allows us to include a function of age into the regression, exploiting the fact that we have a continuous variable at our disposal. This allows us to "borrow strength" from the distribution of age over the dependent variable. Table 3 gives an overview of all estimated specifications for this paper, with a short definition of the explanatory variables. The full definition of the explanatory variables, descriptive statistics and full regression outputs are all available on request. The regressions are estimated separately by country using data from 2014, with the exception of specification (4)* where data from 2014 is matched with the same quarters from 2015 to compare the impact of duration of unemployment over time. In principle, instead of defining different specifications depending on the purpose, it would be possible to include all regressors in one. For small countries with small samples, this is however not feasible; we therefore opted instead for a solution where we include only one interaction in each. Even then, results for small countries are questionable in some cases.

The choice on whether or not to use weights in the regression, and in which fashion, was made based on practical considerations. Eurostat has only very limited technical information at its disposal, and it includes neither information on strata nor primary (or, where relevant, secondary) sampling units. The only weights available are the final survey weights after the calibration process used to gross up individual observations in the overlapping sample to the

[1] External sources a researcher might use to explain the transition probability of an individual could be the registered unemployment rate by NUTS region, past GDP growth by region, degree of urbanization or similar indicators.

population total. Including these weights as sampling weights may be questionable; however, while individual estimates and standard errors were not identical in results including and excluding these weights, general patterns and significance levels were comparable.

Table 3: Specifications estimated using logit, by country, 15-74, Q2-Q3 2014

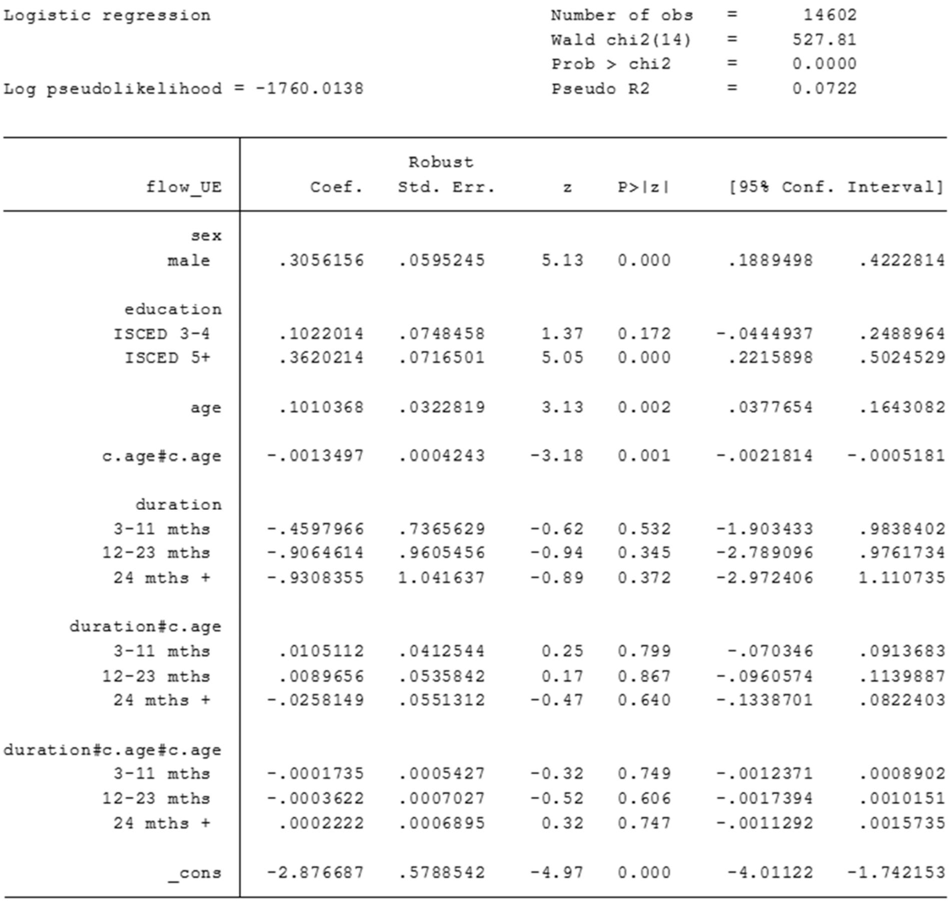
Dependent variable =1 if individual moved from unemployment to employment

Explanatory variables	Definition	(1)	(2)	(3)	(4)*
sex	0= female 1=male	x	x	x	x
age	continuous	x	x	x	x
age*age		x	x	x	x
sex*age		x			
sex*age*age		x			
duration of unemployment	0= less 3 mths 1= 3 to 11 mths 2= 12-23 mths 3= 24+ mths	x	x	x	
duration*age			x		
duration*age*age			x		
educational attainment	0= ISCED0_2 1=ISCED3_4 2=ISCED5_8	x	x	x	
education*age				x	
education*age*age				x	
year	0=2014 1=2015				x
year*age					x
year*age*age					x

The uncertainty concerning the correct choice of weighting in the context of regressions is one of several factors influencing the proposal on how to present the regression results. Even more important factors are the difficulty in communicating the meaning of coefficients expressed in log odds. To illustrate this issue, we present one example based on Spanish data using specification (3). The interaction terms make the interpretation of overall effects of any one variable on the probability of an individual becoming employed difficult in this form, as can be seen in table 4. Following the regression, we calculate predictive margins for 5-year age groups between 25 and 65 and the 4 groups for duration. This table (table 5) is easily understood. The margins could be shown directly in an online database, with appropriate flags to indicate estimates which are not significantly different from zero. Users can then download the information they are interested in; however, if results for all countries should be presented, we propose a graphical solution instead, as the mass of information otherwise is overwhelming.

Table 4: Logit results specification (3), Spain, Q2-Q3 2014

Dependent variable =1 if individual moved from unemployment to employment

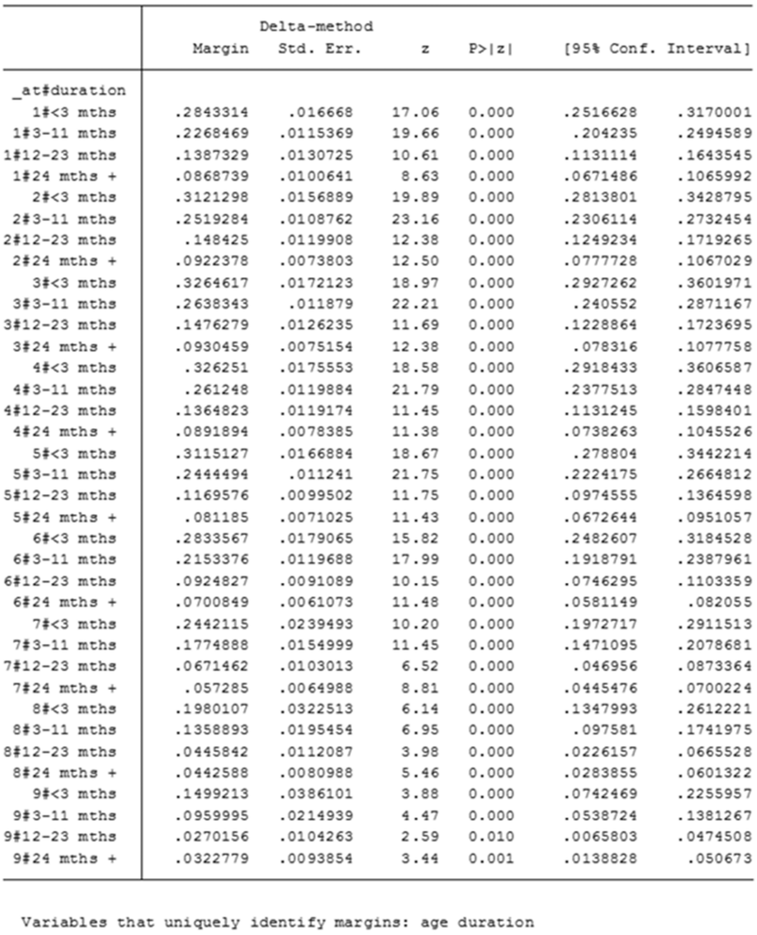


```
Logistic regression                             Number of obs     =      14602
                                                Wald chi2(14)     =     527.81
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -1760.0138               Pseudo R2         =     0.0722

------------------------------------------------------------------------------
                     |               Robust
             flow_UE |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+--------------------------------------------------------
                 sex |
               male  |   .3056156   .0595245     5.13   0.000     .1889498    .4222814
                     |
           education |
          ISCED 3-4  |   .1022014   .0748458     1.37   0.172    -.0444937    .2488964
          ISCED 5+   |   .3620214   .0716501     5.05   0.000     .2215898    .5024529
                     |
                 age |   .1010368   .0322819     3.13   0.002     .0377654    .1643082
                     |
         c.age#c.age |  -.0013497   .0004243    -3.18   0.001    -.0021814   -.0005181
                     |
            duration |
          3-11 mths  |  -.4597966   .7365629    -0.62   0.532    -1.903433    .9838402
         12-23 mths  |  -.9064614   .9605456    -0.94   0.345    -2.789096    .9761734
          24 mths +  |  -.9308355   1.041637    -0.89   0.372    -2.972406    1.110735
                     |
      duration#c.age |
          3-11 mths  |   .0105112   .0412544     0.25   0.799     -.070346    .0913683
         12-23 mths  |   .0089656   .0535842     0.17   0.867    -.0960574    .1139887
          24 mths +  |  -.0258149   .0551312    -0.47   0.640    -.1338701    .0822403
                     |
duration#c.age#c.age |
          3-11 mths  |  -.0001735   .0005427    -0.32   0.749    -.0012371    .0008902
         12-23 mths  |  -.0003622   .0007027    -0.52   0.606    -.0017394    .0010151
          24 mths +  |   .0002222   .0006895     0.32   0.747    -.0011292    .0015735
                     |
               _cons |  -2.876687   .5788542    -4.97   0.000     -4.01122   -1.742153
------------------------------------------------------------------------------
```

Figure 1 shows the predicted probabilities from table 5. Presented in this way, the main messages are immediately conveyed: the relationship between duration of unemployment, age and probability to transition from unemployment to employment can be read directly off the figure. Showing the same information for all countries for which data is available, the same type of figure for each country, presented on one sheet, using the same scale on the y-axis lets users compare the effects and discern whether certain patterns are country-specific. Especially in the context of Eurostat publications, we propose this type of visualization over the presentation in tables. In the annex, the results for specifications 2-5 are shown.

Table 5: Predictive margins following logit regression

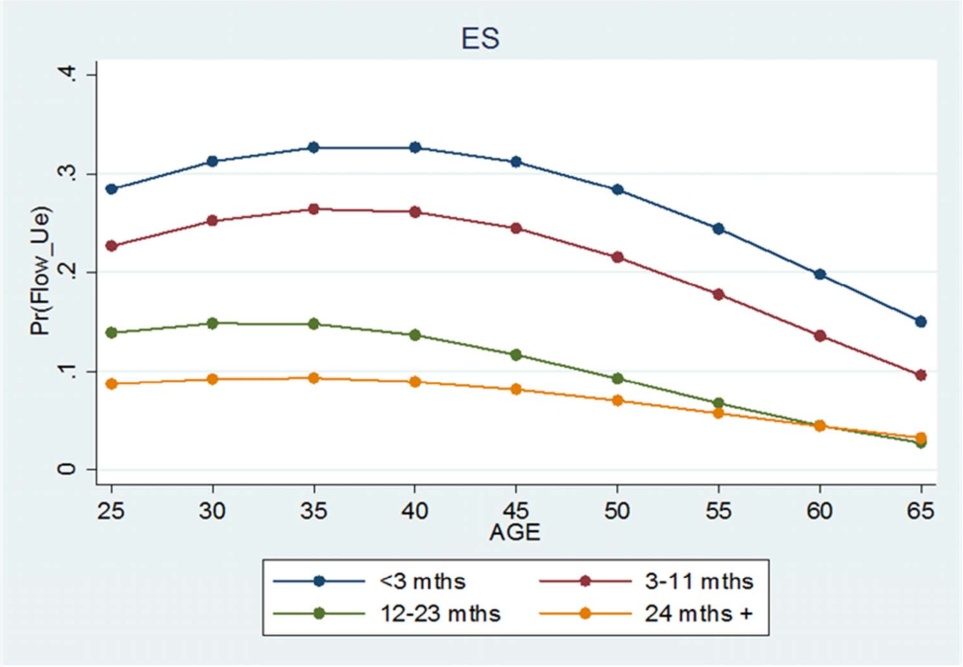


```
              |            Delta-method
              |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
  _at#duration|
    1#<3 mths |   .2843314    .016668    17.06   0.000     .2516628    .3170001
   1#3-11 mths|   .2268469   .0115369    19.66   0.000      .204235    .2494589
  1#12-23 mths|   .1387329   .0130725    10.61   0.000     .1131114    .1643545
   1#24 mths +|   .0868739   .0100641     8.63   0.000     .0671486    .1065992
    2#<3 mths |   .3121298   .0156889    19.89   0.000     .2813801    .3428795
   2#3-11 mths|   .2519284   .0108762    23.16   0.000     .2306114    .2732454
  2#12-23 mths|    .148425   .0119908    12.38   0.000     .1249234    .1719265
   2#24 mths +|   .0922378   .0073803    12.50   0.000     .0777728    .1067029
    3#<3 mths |   .3264617   .0172123    18.97   0.000     .2927262    .3601971
   3#3-11 mths|   .2638343    .011879    22.21   0.000      .240552    .2871167
  3#12-23 mths|   .1476279   .0126235    11.69   0.000     .1228864    .1723695
   3#24 mths +|   .0930459   .0075154    12.38   0.000      .078316    .1077758
    4#<3 mths |    .326251   .0175553    18.58   0.000     .2918433    .3606587
   4#3-11 mths|    .261248   .0119884    21.79   0.000     .2377513    .2847448
  4#12-23 mths|   .1364823   .0119174    11.45   0.000     .1131245    .1598401
   4#24 mths +|   .0891894   .0078385    11.38   0.000     .0738263    .1045526
    5#<3 mths |   .3115127   .0166884    18.67   0.000      .278804    .3442214
   5#3-11 mths|   .2444494    .011241    21.75   0.000     .2224175    .2664812
  5#12-23 mths|   .1169576   .0099502    11.75   0.000     .0974555    .1364598
   5#24 mths +|    .081185   .0071025    11.43   0.000     .0672644    .0951057
    6#<3 mths |   .2833567   .0179065    15.82   0.000     .2482607    .3184528
   6#3-11 mths|   .2153376   .0119688    17.99   0.000     .1918791    .2387961
  6#12-23 mths|   .0924827   .0091089    10.15   0.000     .0746295    .1103359
   6#24 mths +|   .0700849   .0061073    11.48   0.000     .0581149     .082055
    7#<3 mths |   .2442115   .0239493    10.20   0.000     .1972717    .2911513
   7#3-11 mths|   .1774888   .0154999    11.45   0.000     .1471095    .2078681
  7#12-23 mths|   .0671462   .0103013     6.52   0.000      .046956    .0873364
   7#24 mths +|    .057285   .0064988     8.81   0.000     .0445476    .0700224
    8#<3 mths |   .1980107   .0322513     6.14   0.000     .1347993    .2612221
   8#3-11 mths|   .1358893   .0195454     6.95   0.000      .097581    .1741975
  8#12-23 mths|   .0445842   .0112087     3.98   0.000     .0226157    .0665528
   8#24 mths +|   .0442588   .0080988     5.46   0.000     .0283855    .0601322
    9#<3 mths |   .1499213   .0386101     3.88   0.000     .0742469    .2255957
   9#3-11 mths|   .0959995   .0214939     4.47   0.000     .0538724    .1381267
  9#12-23 mths|   .0270156   .0104263     2.59   0.010     .0065803    .0474508
   9#24 mths +|   .0322779   .0093854     3.44   0.001     .0138828     .050673
-------------------------------------------------------------------------------

  Variables that uniquely identify margins: age duration
```

Figure 1: Predicted probability for transition from unemployment to employment, Q2-Q3 2014, Spain

Marginal effects of duration evaluated for 5-year age groups covering ages 25 to 65



4. Conclusion

From the presented exercise we conclude that the use of regression to enhance the offer of breakdowns for indicators derived from small sample sizes, and in the absence of micro-data which could directly be exploited, is feasible. Results can be presented in a fashion that is easily understood by users of data who are not experts in regression techniques. While the long-term goal in the specific case of flow statistics should be to increase the sample size of the matched samples and to produce matched data with longitudinal weights so that users can calculate their own statistics, we find that using simple models should be considered as a viable option in the production of more detailed indicators in the meantime.