WP.26
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and**
**ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Vienna, Austria, 21-23 April 2008)

Topic (iv): New and emerging methods

# ESTIMATION OF PRELIMINARY UNEMPLOYMENT RATES BY MEANS OF MULTIPLE IMPUTATION

Prepared by Thomas Burg, Statistics Austria

***Abstract***

The Austrian Labour Force Survey is performed quarterly. Its main estimates are unemployment rates. Field work is time consuming and evidently final results are available only after all data records have been collected. Due to requirements by users and to provide the public with useful key information about the labour market as soon as possible it is desirable to deliver preliminary results. To come to these provisional figures several methods were tested. Besides some grossing up procedures some approaches using multiple imputation were tested. Multiple Imputation (MI) seems not so wide spread in official statistics because it focuses rather on certain estimators than receiving an authentic database. However having the situation of estimating the contribution of a certain amount of records not available at an early point of time to some key estimators gives a good opportunity to investigate how MI-methods work in comparison to another more traditional method (like weighting).

The paper describes the work flow of the MI-method how it was implemented. Some problems during the implementation process are envisaged. Finally the results of the MI and the grossing up methods are compared with final data actually published.

## I. INTRODUCTION

1. Nowadays receiving results of statistical surveys as early as possible is one of the key demands to official statistics especially when the political interest of the resulting figures is very high. However if you have the situation that data collection during the fieldwork is a continuous process distributed over a certain time period you can calculate the estimators only after this process is finalised. On the other hand the need of quick results forces data producers to develop methods for preliminary results. Due to the high impact of these figures on decision makers it is inevitable to ensure a high quality level of the estimators.

2. In the Austrian Labour Force survey we were confronted just with this situation described above. Of course the key figure of this survey is the number of unemployed persons in comparison to the number who are in labour force (which means they are potential employees) - the classical unemployment rate. Along with that indicator some other figures which should illustrate the situation on the Austrian labour market all more or less dealing with the number of employees were considered as candidates for preliminary estimation. The table below gives a brief description of the indicators we looked upon.

| Indicator | Description |
|---|---|
| | |
| Unemployment rate | Number of employed persons divided by all people in labour force |
| Activity rate | Number of economically active persons divided by all people of the population |
| Rate of part time workers | Number of part time workers divided by total of working persons |

Some of these figures were also considered broken down roughly by sex, age, and citizenship.

3. Looking at the problem of preliminary estimation of figures like the ones described above you can think over several possible methods. To select between various methods it is necessary to have a criterion on which you can base your decision. Of course this depends highly on the concrete situation regarding the availability of data you use for your estimation. In the following the situation of the Austrian Labour Force Survey and the concrete estimation problem will be described.

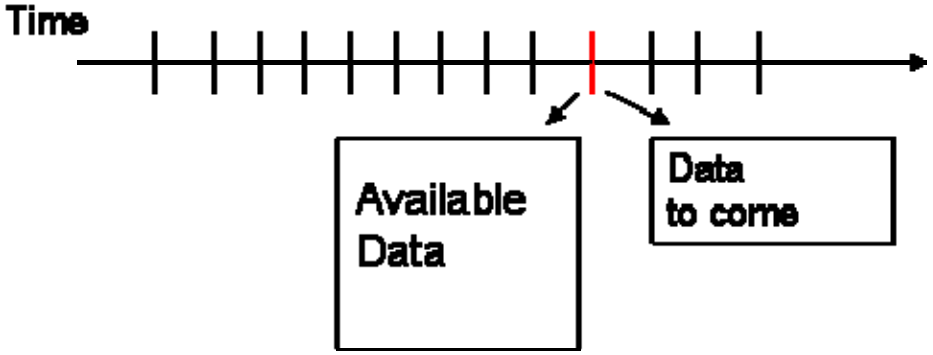
## II. THE AUSTRIAN LABOUR FORCE SURVEY AND THE ESTIMATION OF PRELIMINARY RESULTS

4. The Austrian Labour Force Survey is based on a rotating sample of households. It is performed quarterly. Every quarter one fifth of the sample is replaced by newly drawn households. Within one household every person is eligible for an interview. The interview at the first contact is held face to face. All follow ups are conducted by the telephone studio at Statistics Austria via CATI.

5. For sure the most important purpose of the LFS is to measure the working status of persons belonging to the population living in Austria. As you may remember the drawn sampling units are households (or dwellings). Along with the kernel of the labour status data you receive a large amount of information concerning the dwelling and the whole set of socio demographic basic information (Sex, Age, Marital Status etc…..) on personal level.

6. An important aspect is the distribution of the households to so called reference weeks. The questions of the LFS are related to the working situation of a reference week. Therefore the sample of each quarter is uniformly distributed to the 13 weeks belonging to the quarter. Normally the household is interviewed in the week which lies directly after the reference week but in exceptional situations the gap between interview and reference week can be stretched up to five weeks.

7. The consequence of the timely distribution of the sample units is that the receiving of incoming data records is a continuous process and the calculation of final estimators can only be done after the fieldwork is completed. Starting with editing and imputation there are fixed procedures for data processing and the final estimation is based on an iterative weighting procedure. The picture below shows the situation concerning data availability at a certain time point during the field work.



8. In the picture you see the situation during a relative late stage of the fieldwork and that is basically spoken the core of our problem. Is it possible to estimate the indicators listed in paragraph 2 based only on the available data records? And accompanying this problem: how good is such a preliminary nowcast?

9. The most desirable time point you want to have your preliminary estimates is exactly on the first working day in the new quarter for instance the first labour day in April for the fist quarter of a year. This lies 5 weeks before the real end of the field work and normally we have then approximately 70% of all the data available.

10. On the data you have not received the situation is not that you have no information at all. If they are follow ups you have of course the values from the last quarter(s). For the newly entering sample units the information you have is of course very rudimental and restricted to variables provided by the sampling frame.

11. The task to do was to find out suitable methods for this estimation problem. It has to be said that at the moment no decision has been taken yet on which method will finally be selected. Even the question if there will be any preliminary results, has not yet been answered. In the following, the methods which were tested are described with a strong focus on the multiple imputation approach.

## III. METHODS OF ESTIMATION

12. There seems to be two different kinds of methods with which you can approach the problem. First you can estimate by using only the available data. On the other hand it would be for sure useful to incorporate the characteristics of the missing data into your estimates as much as possible. To achieve this it could be a proper method to impute the items on which the indicators are based.

13. Based on this two principal approaches we decided to test two methods. As told before the LFS as a sample survey is weighted via a raking procedure according to marginal distributions of the Austrian population. So it seems natural that you are able to use simply the same procedure for the reduced set of data records. The second method should be based on imputing the relevant variables on the data records still to come. However since there are only few variables involved we saw there a good opportunity for a multiple imputation approach.

14. Multiple imputation is a not so common method in official statistics. Maybe this has the reason that national statistical agencies mainly have the goal to produce final data sets often called authentical databases from which a lot of tables are generated. Imputation in its classical form guarantees that all items are non-missing with correct values with respect to given edit rules. On the other hand multiple imputation focuses rather on concrete estimation problems when there is no need to store the single values imputed at every imputation run.

15. To select methods is one thing but an important issue is how to test the results you receive. In our situation it suggested almost itself to use already finalised quarters. Since the date of receiving of records is known we can simulate the exact situation for any time point we wish. Having this the evaluation criterion is simply the differences from the preliminary results to the final estimates.

16. There was also an idea to use a method based on time series. However some basic tests showed that this approach delivered not very useful results. One of the reasons was that the ingoing time series are not sufficiently long. It also seemed to be the fact that short term effects on unemployment are not pictured well enough by the time series approach.

## IV. PRELIMINARY ESTIMATION USING MULTIPLE IMPUTATION

17. The most relevant item which is the basis of most of the indicators is ‘Labour Status’. This qualitative variable has 4 possible values (1=’employed’, 2=’unemployed’, 3=’not relevant for

employment', 4='military person'). Although there are other items involved we will concentrate on this one during our considerations. The values '3' and '4' are in some sense very stable over time whereas the relevant fluctuations concerning our problem are in the movements between '1' and '2'.

18. Before implementing multiple imputation it must be decided how the imputation at a single run should be performed. To work this out we analyzed the relevant variables and their dependencies to other items along with the correlation to the time of receiving the answers.
The basic idea was to use probabilistic imputation based on distributional assumptions gained by the analysis of the given data.

19. It seems evident that 'labour status' has some correlation to classical socio demographic items. So ´first of all we included 'Sex', 'Age' (in Groups) and 'Citizenship' (Austrian, non Austrian). Of course there seems to be other items worth to investigate their connection to labour activity of a person. But due to the fact that we want to advance towards results quickly we decided not to put further resources to that part of the problem.

20. One of the most crucial aspects was the question if the pattern for labour status of the respondents is dependent from the time point during fieldwork when the answer of the respondent is received. In other words: Is there a significant difference of the distribution of the item "labour status" between the known records and the interviews still to come? It was not surprising that this was really the case. What was indeed surprising was the fact that the differences were not even equal for all quarters. As a consequence this had to be built into the imputation algorithm. So based on results of totally processed quarters we estimated differences in every single stratum for each of the four quarters separately. Post-stratification was initially on the variables 'Austrian – 'Non -Austrian') ,'Sex' and 'Age group' ('15-24', '25-64', '65 and older'). Using this stratification it turned out that the results in the end where not satisfactory at all. This led to the assumption that there must be another factor which has a decisive impact to the labour status. First we thought it is sufficient to rarefy the age classes. Neither that nor the inclusion of other socio-demographic items led to resounding success.

21. Deepening our analysis we came behind that the weight of a person is the key factor for the problems we had. So what turned out was the fact that depending on the weight of a person the differences of the distribution of labour status between the known values and that which laid beyond the time point of interest changed significantly.

22. Due to this conclusion we included a variable called "weighting class" into the poststratification. This variable has 5 possible values depending on the value of the weight. (1=smaller than 100, 2 = 100 and < 200, 3 = 200 and < 300, 4 = 300 and < 400, 5 = 500 or more). All together we had now 60 strata. For each stratum we produced an estimation of the difference of the distribution for the item 'labour status'. These estimations for the expected distribution differences were performed separately for each quarter.

23. One disadvantage in the estimation of the distribution lies in the fact that the new continuous labour force was implemented in the beginning in 2004 which is not very long ago. Our investigations started in the end of 2007. At this time we had only two completed datasets for the first quarter and even only one for the $2^{nd}$, $3^{rd}$ and $4^{th}$. So the database for estimating the distributional differences was very thin.

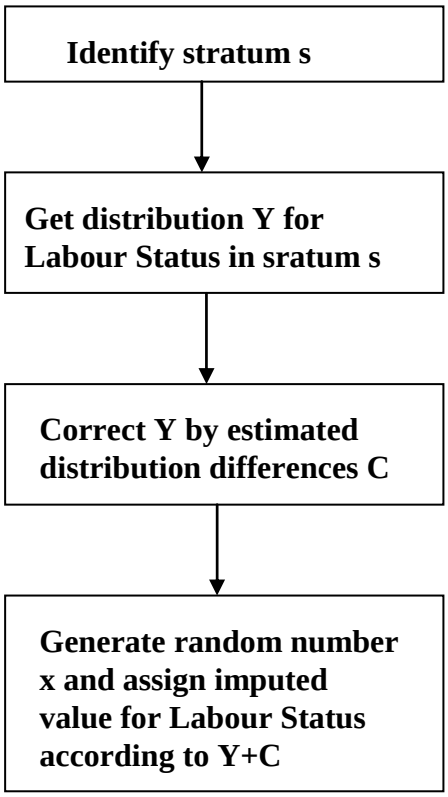
24. Another problem I want to mention lies in the fact that the records you have to impute can be of two different types. With the follow up the handling is not that difficult because you have all of the relevant items for poststratification stored from prior waves. Quite different is the situation for households newly entering the sample. Here you have only a certain assumption regarding to the composition of the household. Even this as our experiences show can differ from the real situation a lot. So here the stochastic element does not come only from the imputation alone but there exists also uncertainty because you can find quite different persons during field work. The problem should also be seen in the context of including "weighting class" as additional stratification variable. To do this you have to calibrate your sample before you run the imputation. The weighting process involves socio-demographic items from the respondents. Again for the follow ups the situation allows to take the those

values from prior waves but for the latest rotation you are restricted to the information of the sampling frame.

25. It has to be stated that the work described in this paper did not further investigate the effects due to this problem. For simplicity and to evaluate the functionality of the multiple imputation it was assumed that all the relevant information for weighting is given. To move the method from prototype to real production some consideration of this problem is required.

26. We can now summarize what is happening at a single imputation step. For every record for which you like to impute labour status you identify the stratum to which the person belongs. Then you take the distribution of labour status in this stratum from the records you classified as known and correct it by the differences estimated for the quarter concerned as described in paragraph 22. After calculating a random number you assign the imputed value according to the resulting distribution. The graphics below shows how the imputation is running.

**Single Imputation for a record not received**




27. Doing this our tests showed quite a huge variability in our results. Therefore we decided to use multiple imputation here to come to smoothed figures. The single imputation described above was performed 25 times and the estimators of interest were calculated at every imputation run. Finally the average of the 25 imputation runs was taken as an estimator for the preliminary result. So to sum up we can identify the following working steps for estimating for instance the preliminary unemployment rate in a certain quarter.

| |
|---|
| Estimation of differences of distribution between known records and expected records of the specified quarter on the basis of quarters already processed. |
| Weighing of the dataset based on certain socio-demographic assumptions concerning the records still to come. |
| Computation of the distribution of labour status of the already known records |
| 25 times single imputation of the item labour status according to the algorithm above and calculation of the unemployment rate on the basis of imputed and non-imputed values for every single imputation. |
| Final estimation of the unemployment rate by the mean value over all imputation runs. |
| |

28. In the following chapter we will show the results of the estimation by this multiple imputation procedure and compares it to the original results as well as to the results of the method using grossing up.

## V. RESULTS

29. As already mentioned we are currently in a testing phase. We had the opportunity to simulate the situation of missing records. To evaluate the results of the quality of the estimates we could easily compare the figures to the final values.

30. The second reference with which the results estimated by the MI-procedure are compared is a completely different estimation method. Here the grossing up procedure is performed simply on the records already available and the corresponding indicators (unemployment rate). One big advantage of this method is that you don't have that problems with assigning weights to records which are not existing in the moment of calculation. You simply treat the non-available records as non-response and calibrate the data material you have already collected. On the other hand by having a smaller amount of units you will have a higher sampling error or more scattered weights. However in this paper we won't give a final conclusion which method should be preferred.

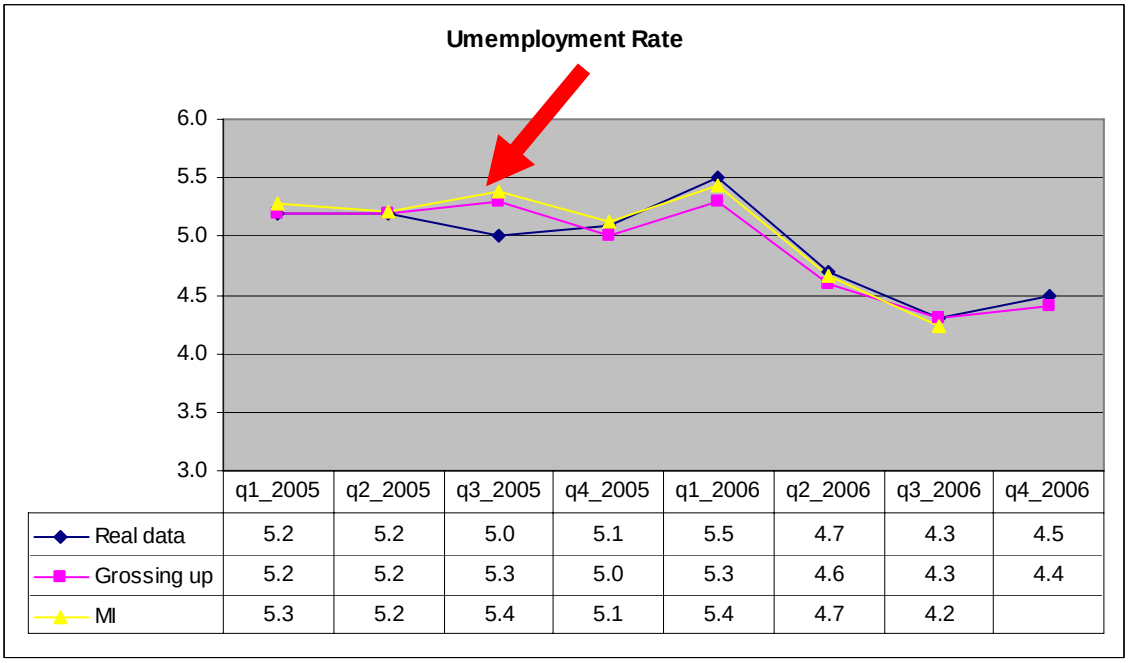
***Results for the MI-Estimation of preliminary figures compared to the real data***

| Quarter | 2005/1 | | 2005/2 | | 2005/3 | | 2005/4 | | 2006/1 | | 2006/2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Estimation | Real Data | Estimation | Real Data | Estimation | Real Data | Estimation | Real Data | Estimation | Real Data | Estimation | Real Data |
| **Employment** | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| Employed Persons | 3781865 | 3757717 | 3799157 | 3800391 | 3858961 | 3881549 | 3845805 | 3845182 | 3804836 | 3818155 | 3923042 | 3916880 |
| Employed Men | 2056107 | 2038631 | 2081284 | 2081762 | 2120423 | 2134838 | 2110580 | 2115249 | 2063437 | 2070893 | 2153478 | 2149284 |
| Employed Women | 1725758 | 1719086 | 1717873 | 1718629 | 1738538 | 1746710 | 1735225 | 1729933 | 1741399 | 1747262 | 1769565 | 1767596 |
| | | | | | | | | | | | | |
| **Unemploment** | | | | | | | | | | | | |
| | | | | | | | | | | | | |
| Unemployed Persons | 210666 | 206996 | 208835 | 209758 | 219420 | 205441 | 207940 | 207250 | 218295 | 223665 | 192142 | 194000 |
| Aunemployed men | 112013 | 111786 | 112052 | 111701 | 110437 | 99478 | 109416 | 106742 | 120993 | 120305 | 97202 | 95494 |
| Uanemployed Women | 98653 | 95209 | 96783 | 98057 | 108983 | 105963 | 98523 | 100508 | 97302 | 103359 | 94940 | 98505 |
| Unemployment rate overall | 5.28 | 5.22 | 5.21 | 5.23 | 5.38 | 5.03 | 5.13 | 5.11 | 5.43 | 5.53 | 4.67 | 4.72 |
| Unemployment rate men | 5.17 | 5.20 | 5.11 | 5.09 | 4.95 | 4.45 | 4.93 | 4.80 | 5.54 | 5.49 | 4.32 | 4.25 |
| Unemployment rate women | 5.41 | 5.25 | 5.33 | 5.40 | 5.90 | 5.72 | 5.37 | 5.49 | 5.29 | 5.59 | 5.09 | 5.28 |
| Umemployment rate foreigners | 12.51 | 11.65 | 11.70 | 12.26 | 10.99 | 10.15 | 12.68 | 12.10 | 12.98 | 13.41 | 10.38 | 9.92 |

31. In the table above you can observe the differences between the estimated figures and the real data. If you take the unemployment rate as the most relevant indicator you see that there are reasonable results for most of the quarters. Only in the third quarter 2005 there seem to be a bigger discrepancy.

32. Seeing that the results for totals are quite reasonable, you can observe that the discrepancies are raising if you move to subgroups. In the last line of the table the unemployment rate for Non-Austrians is presented. With the differences quite high it seems not arguable to publish provisional values based on the estimates.

***Comparison of estimation of unemployment rate – MI, Grossing up and Real data***




| | q1_2005 | q2_2005 | q3_2005 | q4_2005 | q1_2006 | q2_2006 | q3_2006 | q4_2006 |
|---|---|---|---|---|---|---|---|---|
| Real data | 5.2 | 5.2 | 5.0 | 5.1 | 5.5 | 4.7 | 4.3 | 4.5 |
| Grossing up | 5.2 | 5.2 | 5.3 | 5.0 | 5.3 | 4.6 | 4.3 | 4.4 |
| MI | 5.3 | 5.2 | 5.4 | 5.1 | 5.4 | 4.7 | 4.2 | |

33. Visualizing the results as done in the graphics shown you see that the methods of grossing up and multiple imputation most of the time deliver reasonable results and are not afar from each other. Looking at the charts it is evident that there exists a problem in the 3rd quarter 2005. Using preliminary estimation there would lead to an overestimation of the unemployment rate by 0.3 %.

34. Trying to interpret this leads to the assumption that for some reason the status “unemployed” is assigned too often. Going back how the imputation flows one possible explanation is that the pattern of labour status of the 3rd quarter of 2004 from which the distributional differences are estimated is not representative for the corresponding one in 2005. This is possibly also caused by the fact that the learning phase for estimating the correct distribution of the records not brought in was too short (see paragraph 23).

## VI. CONCLUDING REMARK

35. As already stated the methods presented are of a purely prototypal nature. We are now in a discussion process with the subject matter directorate. It is not decided if there will be a publication of provisional figures and if there is a decision in favour of doing that which method will be used finally. As seen by the results, further testing with processed quarters can be useful. Another aspect is that the problem of assigning weights for newly interviewed households beyond the time point of interest has to be investigated deeper.