

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 5 of the provisional agenda

IMPUTATION IN THE NEW DUTCH STRUCTURE OF EARNINGS SURVEY (SES)

Submitted by Statistics Netherlands¹

¹ Prepared by Eric Schulte Nordholt.

ABSTRACT

The Structure of Earnings Survey (SES) of Statistics Netherlands has been created by matching three data sources: the business survey on Employment and Wages, the registration system of the social security funds and the Labour Force Survey. The first reference year of the SES is 1995. The advantage of matching these three sources is that an enormous amount of records with detailed wages information becomes available. This information can be used to study wages by several variables such as economic sector, level of education and occupation. After matching the three sources we have the problem of missing values. For some variables this problem is solved by imputing and for other variables this problem is tackled in the weighting procedure. In this report the choices where to impute and where to weight are explained. It is important not to disturb the distribution of the variables too much and to preserve the covariances between the different variables as much as possible. As the number of records is too large to use the random hot deck method, the chosen imputation method is the sequential hot deck method.

Keywords: imputation, sequential hot deck method, Structure of Earnings Survey (SES)

I. INTRODUCTION

1. The new Structure of Earnings Survey (SES) of Statistics Netherlands has been created by matching three data sources: the new business Survey on Employment and Wages (SEW), the Registration System of the Social security funds (RSS) and the Labour Force Survey (LFS). The first reference year of the SES is 1995. The three sources are described briefly in section 2 of this report. The advantage of matching these three sources is that an enormous amount of records with detailed wages information becomes available. This information can be used to study wages by several variables such as economic sector, level of education and occupation.

2. After matching the three sources the problem of missing values was discovered. For some variables this problem is solved by imputing and for other variables this problem is tackled in the weighting procedure. In section III the choices where to impute and where to weight are explained. It is important not to disturb the distribution of the variables too much and to preserve the covariances between the different variables as much as possible. As the number of records is too large to use the random hot deck method the chosen imputation method is the sequential hot deck method that is described in section IV.

II. THE DATA

3. The new business Survey on Employment and Wages (SEW) contains a large number of records and a lot of information about wages. In particular, the public sector is well represented in the SEW. The Registration System of the Social security funds (RSS) contains an even larger number of records. In this source the private sector is very well represented,

but the number of variables is less than in the SEW. The Labour Force Survey (LFS) contains the necessary information for the SES about the level of education and the occupation of the employees. Only a limited number of the available variables in the three sources (SEW, RSS and LFS) is relevant for the SES and therefore a selection has been made of those variables and the variables that are used in the matching, imputation or weighting process. The reference year for both the SEW and the RSS is 1995. To get more records of the LFS that match with the other two sources, the Labour Force Surveys of 1994, 1995 and 1996 are combined. Of course the score on a variable is not constant over time, so the risk is introduced that we do not have the correct score on a variable from the LFS. Therefore combining three years must be considered as a compromise between using only the LFS of 1995 on the one hand and combining more than three years of the LFS on the other hand. To gain time, all records in the LFS with a missing score on one of the relevant variables for the SES are dropped. The number of dropped records is not too large and therefore it is efficient to compensate for this loss of data in the weighting process.

4. After finishing the data preparation, the matching process starts. The records of the SEW and RSS are matched with the LFS on the variables address, postal code, city, date of birth and gender. Only exact matches are allowed as the aim of the SES is to analyse the structure of earnings. A unique combination on the matching variables could be found in the sources that is not unique in the population; this could of course lead to a mismatch. Also missed matches could occur, e.g. caused by typing errors in the postal code. Experiments showed that in spite of these mismatches and missed matches a quite reasonable matching result was obtained. The intention to prevent synthetic matching could thus be maintained. As the data of the SEW form the basis of our SES, only the records from the RSS that match with the LFS are considered for inclusion in the SES data set. The added records from the RSS must belong to the population of the SEW, but must not yet be present in the SEW data set because of the sampling design of or nonresponse in the SEW. The SES records can be classified into five groups depending on in which data sources of the SES they are present. In this manner, the SES variables can be classified into three groups. For the records in the five groups of records it is indicated in Table 1 in which of the sources these records can be found, from which sources the scores on the variables in the three groups of variables will be taken and how many records the different groups contain:

Table 1. Outline of the different SES groups.

SES group	Records are present in the following data sources:			Variables that are available in both the SEW and the RSS	Variables that are only available in the SEW	Variables that are only available in the LFS	Number of records
	SEW	RSS	LFS				
1	yes	yes	yes	SEW data	SEW data	LFS data	21 105
2	yes	yes	no	SEW data	SEW data		806 489
3	yes	no	yes	SEW data	SEW data	LFS data	19 995
4	yes	no	no	SEW data	SEW data		733 084
5	no	yes	yes	RSS data		LFS data	84 977

III. THE IMPUTATION STRATEGY

5. In Table 1 we have for SES group 5 a cell with horizontal shading that indicates missing values. This problem can be solved by imputation. Auxiliary variables in the imputation process will be variables that are available in both the SEW and the RSS.

6. In sequencing the derivation of composite variables and imputation, several alternatives are possible. An alternative is first to impute all relevant variables in the data sources and then to deduce all composite variables. Though this alternative can be deduced without any problem, the disadvantage of this approach is that several data sets have to be treated separately, which means that a lot of variables have to be imputed. This is a lot of work and will not be possible within the tight time schedule of the SES. A second alternative is to start with the derivation of all composite variables and finish with the imputation. This is a realistic approach as all composite variables can be derived for all records in the SES groups 1-4 and then the imputation of the composite variables in SES group 5 can start. However, the disadvantage of the second alternative is that no optimal imputation will result when composite variables are imputed. Therefore a middle course is adopted: first some composite variables are deduced, then the imputation process takes place and finally the last composite variables are deduced based on the imputed data.

7. In Table 1 we find two cells with vertical shading that also indicates missing values. This problem is tackled in the weighting procedure as imputation is unwanted here. Firstly, not much auxiliary information is available for such a mass imputation. Secondly, a mass imputation would give dramatic results if not only the imputed variable is analysed but also the crossing of the imputed variable by other variables that were not taken into account in the imputation process. This investigation into the structure of this data set is the main aim of the SES and therefore it is wise to limit ourselves to the weighted and imputed data of the groups 1, 3 and 5. The imputed data set is raised to the population totals in the weighting process.

IV. THE SEQUENTIAL HOT DECK METHOD

8. The variables that are only available in the SEW are imputed using some auxiliary variables that are available in both the SEW and the RSS. Examples of the variables that have to be imputed are gross wages per month, gross wages for overtime per month and the number of holidays. The auxiliary variables for the imputation are gender, type of employment contract, age, gross wages per day and economic sector. All together, 26 labour variables were imputed using the 5 auxiliary variables mentioned before. The classes in which the auxiliary variables are categorised are chosen in such a way that homogeneous groups result that contain approximately the same number of records. As there is a big difference in the scores on the variables that have to be imputed between different economic sectors, this auxiliary variable is categorised in homogeneous groups that do not all contain approximately the same number of records. The auxiliary variables appeared to be of good quality and did not contain missing values themselves.

9. As deterministic imputations distort the distribution of the imputed variable and a distributions of variables are of major concern in the SES, stochastic imputations are necessary in the SES imputation process. The question is which stochastic imputation method is best suited for the imputation of the SES. An easy choice would be a stochastic regression imputation, but this does not always lead to imputed values that are feasible. Therefore, a hot deck method is a better alternative. As the number of records is too large to use the random hot deck method, the chosen imputation method is the sequential hot deck method. A random element is introduced in this method by sorting the unimputed data set randomly before the imputation process starts. For every combination of scores on the categorised auxiliary variables, a help array with so-called potential donor values is created.

10. The first time a missing value from a record with that combination of scores on the categorised auxiliary variables is found the first score of the relevant help array is copied. The second time the second score of the help array is copied and so on. If all records of an array have been used once, a second pass through the help array starts. The record from which the imputed value is copied is called the donor record. Although we have 1 580 673 potential donor records (SES groups 1-4) that could be included in the help arrays, we have to be careful with the number of categories of the auxiliary variables. If we create too many of these categories there is the risk that a record has to be imputed with an empty help array which is of course not possible.

11. If the help array contains a few values but a lot of records have to be imputed using this array, there is the problem of the multiple use of donors that will often lead to underestimation of the variance of the imputed variable. Therefore empty or almost empty help arrays have to be combined with other help arrays. Methodologically this corresponds with the introduction of a priority ordering of the help variables in the random hot deck method. Also, in that case we cannot impute all records using all auxiliary variables categorised in the finest categorisation we have available.

12. To avoid inconsistencies between related variables as a result of the imputations, the method of record matching is used for the imputation of the SES. This means that related variables are imputed simultaneously using the same imputation model. This way, we also seek to preserve covariances between imputed variables which is important for the analyses of the SES.