

Longitudinal imputation for the quarterly Labour Force Survey

Imputazione Longitudinale: il Caso della Rilevazione Trimestrale sulle Forze di Lavoro

Simona Rosati

Istituto nazionale di statistica, Via A. Ravà 150 – Roma, e-mail: sirosati@istat.it

Riassunto: La Rilevazione trimestrale sulle forze di lavoro (RTFL) è un'indagine campionaria il cui scopo principale è quello di fornire stime *cross-section* (cioè riferite ad un dato istante di tempo) sulla misura e sulle caratteristiche dell'occupazione e della disoccupazione. Il particolare disegno campionario della RTFL consente altresì di effettuare analisi longitudinali e di flusso, importanti per lo studio delle transizioni nel mercato del lavoro. L'utilizzo delle informazioni ricavate dalla struttura longitudinale della RTFL pone tuttavia diversi problemi di carattere metodologico, tra cui il trattamento delle mancate risposte parziali nei dati longitudinali. Questo lavoro ha lo scopo di presentare il metodo di correzione dei dati longitudinali della RTFL che risulta in gran parte un metodo generalizzato, cioè estendibile a qualsiasi indagine.

Keywords: item non-response, longitudinal imputation.

1. Introduction

The quarterly Labour Force Survey (LFS) is a rotating-panel survey which provides data on labour market activity. Although the main objective of the survey is to produce quarterly and annual estimates, longitudinal analysis is obviously an aim of great consequence. The survey's longitudinal dimension, in fact, makes it possible to understanding the structure of labour market in terms of its dynamic components. Thus it is possible to make a deeper analysis of the incidence of conditions and events such as employment and unemployment over time. Nevertheless the time dimension makes it more difficult to develop a strategy for handling item non-response as well as unit non-response. Unit non-response results in the lack of any information on a given observation. If only a subset of information is missing from a responding observation this failure is referred to as item non-response. The latter also includes the case where the values for two or more variables are logically inconsistent and the case of a response that is out of an accepted range of values. Unit non-response is generally compensated for by weighting (this issue is not dealt with in this paper), whereas item non-response is often handled by imputation. An imputation technique, widely used in surveys, was developed by Fellegi and Holt (1976). The method, which is currently applied to cross-sectional data of the LFS, was adapted for categorical longitudinal data.

This paper is in particular devoted to the longitudinal imputation applied to replace item non-response on single questions in the LFS. In Section 2 are briefly introduced the two main approaches for imputation and the Fellegi and Holt methodology. Section 3 presents a description of the imputation procedure as it is applied to correct the item non-response in longitudinal data of the LFS. The paper concludes with a discussion about the main outcomes and some issues about future research in the new Labour Force Survey.

2. The Fellegi and Holt methodology

The literature proposed several classification schemes for imputation methods. As Kalton and Kasprzyk (1982) discuss, two main categories of imputation can be distinguished: deterministic and probabilistic. Deterministic imputation assigns only one value, *a priori* determined, on the base of other values considered “true” by experts. On the contrary, probabilistic imputation assigns a value according to a stochastic model (e.g. a regression model) or using a donor unit which is similar to record to be imputed. A *distance function* is generally used to define similarity. The authors showed that deterministic imputation methods distort the distributions and attenuate variances, whereas probabilistic methods yield approximately unbiased estimates of distributions and element variances.

The Fellegi and Holt methodology can be counted among the probabilistic methods⁽¹⁾. It is concerned primarily with the stages of data editing and of imputation. In editing process some edits (rules) are specified by experts with a knowledge of the subject matter of the survey. An edit expresses the judgment of some experts that certain combinations of values, corresponding to the different questions on the questionnaire, are unacceptable. The set of such edit rules is referred to as *explicit edits*. A record which fails any of the edits does need to be corrected. Conversely, a record which passes all the stated edits does not need to be corrected. In order to identify which values must be changed to correct the record (change it in such a fashion that the changed record passes all the edits), it is essential to obtain all edits logically implied by the explicit edits. All edits, explicit and implied (with no redundancies), form a *complete set of edits*.

Imputation is preceded by the identification of a *minimal set* of variables (the smallest set) whose values can be changed in such a fashion that the resulting record would pass all the edits. Imputation will provide suitable values for those variables.

One of the major objectives of the Fellegi and Holt methodology is to retain the structure of the data. This means that univariate and multivariate distributions of survey data reflect as nearly as possible the distributions in the population. This goal is achieved by the use of *hot-deck* imputation. Hot-deck imputation consists of imputing for a variable of the current record the value recorded in the same field of some other record, which passed all the edits and which is similar in all respects to the current record. This can be done one variable at time (*sequential imputation*) or all variables at once (*joint imputation*). The former aims to preserve the univariate distributions of the variables, whereas the latter preserves the multivariate distributions.

3. Longitudinal imputation applied to the LFS

The imputation of item non-response relating to longitudinal data in the LFS follows a two step procedure. The first step consists of a probabilistic module entirely based on the Fellegi and Holt methodology. It is essentially made of a set of longitudinal edits involving the categorical variables (i.e., variables which are not subject to a meaningful metric) of the questionnaire.

A longitudinal edit is a logical expression which expresses the judgement of some experts that certain combination of values (corresponding to the different questions of the

⁽¹⁾ Many arguments reported in this paragraph were extracted from the original paper of Fellegi and Holt published in 1976. A revised version was presented at the Conference of European Statisticians, Statistical Standards and Studies in 1994 by the same authors (Fellegi and Holt, 1994).

questionnaire) is unacceptable regards to the time. For example, a longitudinal edit may be

$$\text{Marital Status}_{t1} (\text{married}) \cap \text{Age}_{t2} (0-14),$$

meaning that a longitudinal record (i.e., containing individual data observed at two different waves) will be incorrect if it has the values Marital status=Married, at the wave $t1$, and Age=0-14 at the wave $t2$. Those records, that need to be correct, are imputed by an algorithm implemented in SCIA (System of Control and Automatic Imputation) according to the methods of Fellegi and Holt. More exactly, the version of SCIA included in CONCORD (CONtrol and Data CORrection), a system designed for Windows environment, was used ((Riccini Margarucci and Floris, 2000).

Since several LFS data are continuous and concerning retrospective questions (e.g., number of months for searching a job, number of months from the ending of the last job, number of months from the beginning of the present job, etc.), a deterministic imputation is carried out in the next step. Retrospective questions are indeed subject to problem of recalling, and of panel conditioning (Kalton *et al.*, 1989). A continuous variable which is missing and/or inconsistent is imputed using the value given by the non-proxy respondent either in the current wave or in the previous wave; otherwise the value in the current wave is imputed from the value in the previous wave. Except for this case, longitudinal imputation consists in using data from the previous wave to impute the missing and/or inconsistent data in the current wave. Although this strategy places additional constraints to the method, it ensures a basic requirement for the LFS that is to obtain definitive estimates for each wave.

Given that the dissemination of longitudinal files starts again in 2003, after almost thirty years, it is necessary to preserve the consistency with the cross-sectional estimates. For this reason longitudinal imputation, both probabilistic and deterministic, is preceded by a cross-sectional imputation as it is currently adopted in the LFS.

As result of the impact of imputation in Table 1 are reported some findings related to the records which have been corrected by sequential imputation or joint imputation. Finally, Table 2 shows the number of records per number of imputations at the end of the imputation procedure (for brevity these results refer to the longitudinal file 2001-2002).

Table 1: *Records per type of imputation (percent values)*

	1998-1999	1999-2000	2000-2001	2001-2002
Sequential	3.3	1.3	2.3	2.2
Joint	96.7	98.7	97.7	97.8

Table 2: *Records per number of imputations*

Number of imputations	Percentage of records
0	65.9
1	16.5
2	13.7
3	2.9
4	0.7
5 or more	0.4

4. Concluding remarks

The experience reveals that the Fellegi and Holt methodology can be adapted for correcting longitudinal data. Although it is a method, which could require considerable computational resources, it can be even implemented for large survey data such as the LFS.

The longitudinal imputation is made taking into account consistency of cross-sectional data. This aim could be achieved imputing both cross-sectional and longitudinal data at the same time. On the other hand, this approach would have been improper for the purpose, given that the cross-sectional estimates has been published for several years before those longitudinal. A simultaneous imputation for cross-sectional and longitudinal data seems to be more suited for a regular functioning of the imputation procedure. This could be the case of the new LFS, which has started at the beginning of this year. Owing to the surveying technique (Computer Assisted Interviewing) some errors will be eliminate, such as incorrect followed skip patterns and values out of range. The remaining errors, including variables which are logically inconsistent, could be corrected by an adjustment of the probabilistic method described here for imputing cross-sectional and longitudinal data together.

However, an attempt of imputing simultaneously cross-sectional and longitudinal data, using the Fellegi and Holt methodology, was previously made for the quarterly LFS (Rosati, 2000).

References

- Fellegi I.P., Holt D. (1976) A Systematic Approach to Automatic Edit and Imputation, *Journal of the American Statistical Association*, 71, 17-35.
- Fellegi I.P., Holt D. (1994) A Systematic Approach to Automatic Edit and Imputation, in: *Statistical Data Editing, Methods and Techniques, Conference of European Statisticians, Statistical Standards and Studies – No. 44*, United Nations, New York and Geneva, 1, 69-110.
- Kalton G., Kasprzyk D. (1982) Imputing for Missing Survey Responses, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 22-31.
- Kalton G., Kasprzyk D., Mc Millen D.B. (1989) Nonsampling Errors in Panel Surveys, in: *Panel Surveys*, Wiley, New York, 249-270.
- Riccini Margarucci E., Floris P. (2000) *Controllo e correzione dati - Manuale utente*, Istat, Roma.
- Rosati S. (2000) La correzione di dati longitudinali nell'indagine sulle forze di lavoro, *Rivista di Statistica Ufficiale, Quaderni di Ricerca*, Istat, 3, 5-29.