

The Labour Force Survey in the Municipality of Florence: Technical Innovations and Methods for Small Area Estimation

La Rilevazione delle Forze di Lavoro nel Comune di Firenze: Innovazioni Tecniche e Organizzative e Metodi di Stima per Piccole Aree

Andrea Giommi (*), Riccardo Innocenti (**), Emilia Rocco (*)

(*) Dipartimento di Statistica “G. Parenti” Università di Firenze

(**) Ufficio di Statistica del Comune di Firenze

Riassunto: l’obiettivo del lavoro è la valutazione di stimatori per piccole aree applicati ai dati della Rilevazione Trimestrale sulle Forze di Lavoro svolta dal Comune di Firenze in parallelo a quella ISTAT. Il disegno di campionamento dell’indagine comunale prevede la selezione trimestrale di un campione stratificato proporzionale di singoli individui che vengono parzialmente rinnovati con un piano di rotazione identico a quello dell’ISTAT. Le piccole aree sono rappresentate dai cinque quartieri in cui si suddivide il territorio comunale. Lo stimatore di quartiere che si propone nel lavoro è di tipo combinato e diretto, mutuando stabilità dai dati dello stesso quartiere rilevati in due occasioni precedenti. La valutazione è svolta mediante la simulazione di un’ipotetica popolazione di quartiere con caratteristiche simili a quelle reali

Keywords: Small area estimation, Combined estimators, Difference estimators.

1. Introduction

The quarterly Labour Force Survey (LF) of ISTAT has been designed to warrant accuracy at regional level for the estimates of each wave and at province level for yearly estimates. No data are published at sub-province level, for example communes (municipalities). To cope with this problem, the Municipality of Florence has designed a LF survey, parallel to that of ISTAT, which can produce several reliable estimates for the whole town area, at least on yearly basis. However, the estimates at town level cannot be easily split into sub-town areas reliable estimates. For instance, the sample was not been designed to produce reliable estimates for districts. Estimation at district level especially for waves represents a typical small area estimation problem in which in order to improve the estimates we must “borrow strength” either from the same sample data across time or space, or from data external to the sample, that is coming from census or from other non statistical sources.

Our objective in this work is to estimate the total number of people searching for a job in a district of the municipality at each survey wave. The nature of the sampling design allows us to resort to an estimation method that borrows strength from the information collected across time, namely in two previous survey waves carried out three months and nine months before the one which we are interested in. In fact, due to the rotating plan, each sample wave may be divided into four parts: (a) people never interviewed before; (b) people interviewed for the second time who were interviewed the first time

three months before; (c) people interviewed for the third time who were interviewed the first time nine months before; and finally (d) people interviewed for the fourth time who were interviewed the third time three months before. Suppose, for example, we are computing estimates for the April 2004. Some of the people in the wave of April 2004, namely those at (b) and (d), are also in the wave of January 2004, while those at (c) are also in the wave of July 2003. Finally, people at (a) are only in the wave of April 2004. We name ‘wave 2’ the one for which we actually estimate the number of unemployed people, April 2004 in the example, while we name ‘wave 1’ that obtained collapsing the two preceding waves as they were one single wave (January and July in the example). Furthermore, waves 1 and 2 together give us a total sample of size, say n , which may be divided into sub-samples. Notice that the sub-samples may be partially or completely overlapped since two sub-samples may contain the same units observed in two different times. There are several ways to combine information split in different sub-samples. Here we propose to use some of the sub-samples to build a difference estimator and then to combine the difference estimate with the direct estimate obtained from an other part of the total sample. The proposed estimator is formally described in section 2, while in section three we present an application to simulated data. In section 4 we briefly discuss the results of the application.

2. Small Area Combined Estimation

Let Y_k be a binary study variable defining the unemployment state of a population individual k ($k = 1, \dots, N$) in a specific district, which takes value 1 if the individual is searching for a job and 0 otherwise. We need to estimate the proportion of unemployed people, say Y^p , in the district at the time of the present wave, that is wave 2, from a sample of size n , defined in the previous section. Now we set:

n_1 = size of all the sample units observed in wave 1;

$m_1 = m_2$ = size of the sample units observed the first time in wave 1 and the second in wave 2;

n_2 = size of the sample units observed only in wave 2.

Thus 1 and 2 are wave subscripts while m indicates the sample units that are observed twice.

Then let y_1 , y_1' , y_2' and y_2'' be unbiased direct estimates of the proportion respectively from sub-samples of size n_1 , m_1 , m_2 and n_2 . The difference estimator:

$$y_{Diff} = y_2' + B(y_1 - y_1'), \quad (1)$$

where B is an opportune constant, is unbiased for Y at the present wave.

Even the direct estimator y_2'' is unbiased and therefore also the combined estimator:

$$y_C = w y_{Diff} + (1 - w) y_2'' \quad (2)$$

with $0 \leq w \leq 1$, is of course an unbiased estimator of the same total. Estimators (1) and (2) are well known (see for example Särndal et al., 1992). Under the simple hypothesis that people are selected by simple random sampling their variances respectively are:

$$V(y_{Diff}) = \left(1 - \frac{n_1}{N}\right) \frac{S_2^2}{n_1} + \left(1 - \frac{m_1}{n_1}\right) \frac{S_D^2}{m_1}; \quad (3)$$

$$V(y_C) = w^2 V(y_{Diff}) + (1-w)^2 V(y_2'') - 2w(1-w) S_2^2 / N \quad (4)$$

where S_2^2 is the elementary variance of the population at wave 2 and S_D^2 is the variance of the difference between the population values at wave 2 and those at wave 1, multiplied by B . Of course (3) and (4), in real applications, must be estimated from sample.

Though our samples are actually formed by proportional stratified sampling, we assume the simplest hypothesis of simple random sampling and use four artificial population to evaluate the performance of estimator (4) compared with the direct estimator based on the sample of wave two.

At wave 1, a population P1 has been formed which reproduce a situation similar to that of a district in the Municipality of Florence. The total size of the district P1 is 60.000 and 2.000 are the unemployed people. For wave 2 we built six populations. The first three, called P2, P3, and P4, have respectively 1.800, 2.000 and 2.200 unemployed people, that is a true proportion of 0,03 0,033 and 0,037. The correlation coefficient for the study variable between P1 and the other populations are:

$$\rho_{(P1,P2)} = 0,62 \quad \rho_{(P1,P3)} = 0,59 \quad \rho_{(P1,P4)} = 0,56$$

The other three, say P5, P6, and P7 have again respectively 1.800, 2.000 and 2.200 unemployed people, but with the following correlation coefficients:

$$\rho_{(P1,P5)} = 0,92 \quad \rho_{(P1,P6)} = 0,90 \quad \rho_{(P1,P7)} = 0,93$$

Both series of correlation coefficients may reflect real situations. In fact, the level of the correlation coefficient depends on the dynamics of the labour market between two or three survey waves.

Concerning the value of B in (1) we assume that past knowledge on the study variable suggests a value between 0,7 and 1. Thus we compare the efficiency of the combined estimator using both these values. The results of the simulation are summarized in the following table 1.

3. Conclusive remarks

Table 1 shows that not surprisingly the combined estimator is always better than the direct estimator. Furthermore the efficiency of the combined estimator is associated with that of the difference estimator which in turn depends mostly on the correlation coefficient between the populations at the two waves and on the choice of B . Concerning B , it is known that the best situations would be those in which it equals the regression coefficient between the study variable at wave 2 and the same variable at wave 1. In our simulation we used two boundary values but, in practice, past knowledge

on the LF survey may suggest a value very near the optimum one. Anyway, irrespective of B, it is apparent that the gain in efficiency is substantially larger for P5, P6 and P7, that is for the populations that are highly correlated with population P1. However, the level of efficiency of the combined estimator is not so high to make it completely satisfactory in every practical situation, since correlations such as those between P1 and P2, P3 or P4 are not unusual in the LF survey. Nevertheless this approach seems promising for at least two main reasons. A direct small area estimation method: (a) does not suffer of biases which are typical of the synthetic methods of estimation; (b) does not require auxiliary information typical of the model approach that, coming from censuses or administrative data, are often dated and of low quality.

Table 1: Standard error (s.e.) comparisons between combined (y_C) and direct estimator (y_{Dir}).

Wave two $B=0.7$	Wave one: P1; $n_1 = 480$; $m_1 = m_2 = 170$; $m_2 + n_2 = 240$		
	$s.e.(y_C)$	$s.e.(y_{Dir})$	$1 - \frac{s.e.(y_C)}{s.e.(y_{Dir})} \times 100$
P2	0,01016	0,01099	7,6%
P3	0,01065	0,01156	7,9%
P4	0,01112	0,01211	8,2%
P5	0,00887	0,01099	19,3%
P6	0,00919	0,01156	20,5%
P7	0,00950	0,01211	21,6%
Wave two $B=1$	Wave one: P1; $n_1 = 480$; $m_1 = m_2 = 170$; $m_2 + n_2 = 240$		
	$s.e.(y_C)$	$s.e.(y_{Dir})$	$1 - \frac{s.e.(y_C)}{s.e.(y_{Dir})} \times 100$
P2	0,01048	0,01099	4,7%
P3	0,01110	0,01156	4,0%
P4	0,01168	0,01211	3,5%
P5	0,00865	0,01099	21,3%
P6	0,00905	0,01156	21,8%
P7	0,00942	0,01211	22,2%

References

- Falorsi P. D., Falorsi S., Russo A. (1994) Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey. *Survey Methodology*, 20, 2, 171-177.
- Giommi A., Rocco E. (2003) Small Area Estimation for the Labour Force Survey in the Municipality of Florence, Atti del convegno *S.CO.2003: Modelli complessi e metodi computazionali intensive per la stima e la previsione*, Treviso 4-6 settembre.
- Ghosh M. and Rao J. N. K. (1994) Small Area Estimation: An Appraisal. *Statistical science*, 9, 1, 55-93.
- Rao J. N. K. (2002) *Small Area Estimation*. Wiley – Interscience
- Särndal C. E., Swensson B., Wretman J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York, Inc.