

Imputation and Optimal Stratification: an Application to the Italian Building Permits

Imputazione e Stratificazione Ottimale: un'Applicazione alle Concessioni Edilizie Italiane

Fabio Bacchini, Roberto Iannaccone
Direzione Statistiche Congiunturali, Istat, Roma
e-mail: bacchini@istat.it, iannacco@istat.it

Edoardo Otranto
Dipartimento di Economia, Impresa e Regolamentazione,
Università degli Studi di Sassari, e-mail: eotranto@uniss.it

Riassunto: Tradizionalmente il trattamento delle mancate risposte prevede la formazione di celle di ponderazione nel caso di unit-nonresponse e imputazione nel caso di item-nonresponse. La distinzione tra i due metodi diventa, comunque, più sfumata quando dal classico caso di indagine cross-section si passa ad indagini longitudinali. In questo lavoro si propone un confronto tra i due metodi nel caso longitudinale. Particolarmente accanto alla tradizionale applicazione del metodo delle celle di ponderazione si propone una procedura euristica basata su un metodo di cross-validation ed imputazione con donatore, simulando vari pattern che rappresentano la struttura di dati mancanti del data set osservato. L'applicazione, riferita ai permessi di costruzione italiani per il periodo 1997-2000 verifica anche l'ipotesi di MAR rispetto alle variabili di stratificazione.

Keywords: Missing data; imputation cells; neighbour imputation.

1. Introduction

A typical statistical problem arises when a large data set contains many missing values. The treatment of missing values leads to the use of weighting or imputation methods depending on the type of non-response: unit non-response or item non-response. In both cases the first step is the stratification of the population into homogeneous groups, using auxiliary variables observable for all the individuals. Weighting adjustment methods provide the formation of cells by the estimation of the response mechanism using logistic regression models. At each respondent unit in a given cell is associated a uniform response probability. For imputation methods, the missing values are replaced using corresponding stratum information (donor, mean of the stratum, median of the stratum, etc.). The typical routines used to automatically perform the stratification procedure (e.g. CART, STRAT2D, CHAID) find the optimal stratification, combining auxiliary variables and evaluating some penalty functions (classification tree models). However, each data set contains peculiarities very difficult to generalize: for example, if an auxiliary variable is a territorial variable, the neighbouring could be a constraint for the aggregation of strata; in addition, the geographical area could represent sub-

populations of interest (study domains). In this case it becomes very difficult to suggest a standard procedure due to the phenomenon studied, the availability of auxiliary variables and the constraints derived by a priori knowledge. We propose as solution a cross-validation procedure, that works in the presence of spatial constraints. The imputation with different stratification is compared with loss functions and various statistics. In this way it is possible to choose the optimal stratification for the imputation. This procedure is very flexible and easily applied. The data set used (the series of Italian building permits from 1997 to 2000) shows the characteristics before discussed; it presents a large number of missing values with available auxiliary variables given by the population and the geographical area (clearly correlated with the number of buildings). It is well known that the particular building structure of Italy strongly depends on the geographical zone, so the neighboring constraint is important. It is not possible to show in a few pages all the results obtained during this study, so we will report just some summary tables and graphs. Further results are available on request, whereas the description of the data set and some preliminary results are shown in Bacchini *et al.* (2003). In the next section the cross-validation procedure is briefly described, whereas the reconstruction of the full data set for the period 1997- 2000 and a comparison with a weighting adjustment method are reported in section 3.

2. A Cross-Validation Procedure

Our data set is composed of the 7940 Italian cities with less than 50000 citizens. With the aim to test our cross-validation procedure, we simulate several times the pattern to be imputed, using the observed data; in this way the artificial data sets contain the same characteristics of the true data set, in terms of missing observations. Then, various possible stratifications (satisfying the spatial constraints) are applied to the various patterns and the imputation of missing data is performed. The alternative stratifications are obtained combining 5 intervals of the variable population, (0-3000], (3000-7000], (7000-13000], (13000-25000], (25000-50000], with the following combinations of geographical area⁽¹⁾: [NW, NE, C, S+I], [NW, NE, C+S+I], [NE, NW+C+S+I]. We call these alternative stratifications respectively strat20, strat15, strat10; in addition we consider the stratification based only on the variable population (strat5) and the case of no stratification (strat0). Given the 5 alternative stratifications, we consider for each year the subset containing the cities respondent for 12 months; on these 4 subsets (recall that the data are relative to 4 years) we perform the cross-validation experiments. In practice, for each subset we simulate 500 patterns containing a structure of missing data equivalent to that relative to the complete data set of each year; in this way we create artificial "holes" in the data set, with the advantage of knowing the true missing values, with which we will compare our imputed values. In this operation, we are supposing that the data are missing at random (MAR) and that respondents and non respondents have a similar behaviour and characteristics within the strata. We estimate the missing data with the criterion of the nearest neighbour donor; in practice, within the stratum, we select as donor the unit having, for the only months observed, the most similar structure to the missing unit. To avoid the strong dependence of the Euclidean metrics on outliers, we use the city block distance; anyway, in the following applications, we

(1) We use NW for North West, NE for North East, C for Center, S for South and I for Islands.

have observed that the final stratification does not depend on the chosen metric. Using the city block distance, the donor for the unit j' will be the unit j satisfying:

$$\min_{j \neq j'} |x_{ij} - x_{ij'}| \quad (1)$$

where x_{ij} and $x_{ij'}$ are the number of new buildings of cities j and j' referred to the stratum i , for the months observed. In this particular application, it is evident that for a collaboration of few months the nearest neighbour donor should be not efficient. So, we establish a lower threshold to apply (1); in practice, we evaluate the goodness of imputation (via MAPE) for each year and each classification; under this threshold, the imputation will be made randomly. In Table 1 the MAPE averages for 1999 using different lower thresholds are showed. Taking into account that the other years show a similar behaviour, the choice of 4 as threshold for imputation with donor and strat20 as most homogeneous stratification seems logical. Table 2 shows the results in terms of mean and variances of monthly MAPE's with a threshold equal to 4 for each year. The choice of strat20 as optimal stratification is confirmed. This stratification is sufficiently consistent with the hypothesis of MAR data within the strata. We have applied an analysis of variance for each strata, comparing the means of the populations of the groups of missing data and the group of respondents (see Howell, 2002). The analysis has been performed for each year from 1997 to 2000. In the 22,3% of the 960 cases analyzed (20 strata by 48 months), we reject the null of equal mean (equivalent to the MAR hypothesis) at a nominal size test of 0.05, and in 13.1% of cases at 0.01 (details are available on request).

3. The Reconstruction of the Italian Building Permits 1997-2000

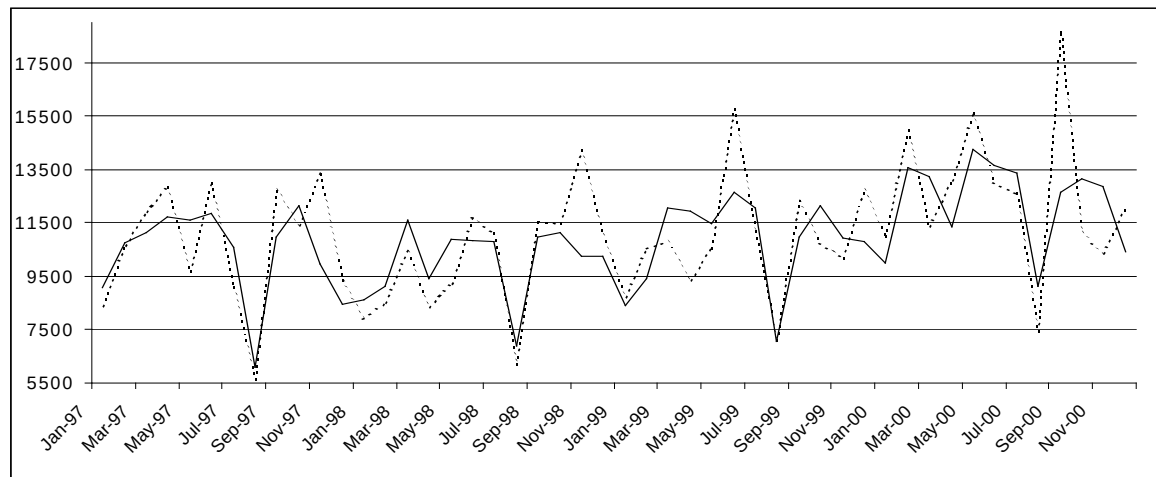
We apply the method of imputation with lower threshold, using the strat20 classification, to reconstruct the full data set of the Italian building permits from 1997 to 2000. In this case we do not have "true" values to use as a benchmark, so we compare the results with those obtained using the well known method of Eltinge and Yasaneh (1997). This method is based on adjustment cells to account for nonresponse; in this case we construct cells that are homogeneous in their estimated response probabilities ϕ_i . For each month in the period 1997-2000 the ϕ_i are estimated by a logit model using as explanatory variables the population size, geographical area and two indicator variables assuming values 0 or 1 if the municipality i has collaborated at time $t-12$ and at time $t-24$ respectively. In Figure 1 the series estimated with the nearest neighbour imputation (NNI) method and with the nonresponse adjustment cells (NAC) method are shown; they show a certain level of correlation (0.72), with a similar trend. The NNI series is smoother than the NAC series, which presents several peaks; both the series capture the seasonal behaviour, with troughs in correspondence of August (see Figure 1). Moreover; we have used also the CART routine to search for the optimal stratification finding similar boundaries for the population strata, aggregated without satisfying the constraint of neighbouring geographical area.

Table 1: Means of monthly MAPE for different lower thresholds - year 1999

Strat.\threshold	1	2	3	4	5
Strat20	9.3	8.3	8.7	7.9	7.9
Strat15	13.4	10.8	12.6	10.6	11.8
Strat10	27.5	21.9	27.0	24.4	26.4
Strat5	67.0	61.0	67.8	66.6	67.8
Strat0	208.0	199.6	208.4	207.3	212.3

Table 2: Means (standard errors in parentheses) of monthly MAPE for various stratifications

Strat.\year	1997	1998	1999	2000
Strat20	6.4 (4.3)	7.9 (5.9)	7.9 (6.0)	6.8 (4.8)
Strat15	10.9 (7.4)	13.9 (9.0)	10.6 (9.1)	10.6 (8.0)
Strat10	31.9 (11.4)	33.8 (13.1)	24.4 (12.4)	30.9 (11.1)
Strat5	79.1 (14.7)	78.0 (17.7)	66.6 (18.9)	78.2 (13.9)
Strat0	191.9 (21.2)	201.1 (26.4)	207.3 (28.1)	209.0 (22.0)

Figure 1: Reconstructed series with NNI (continuous line) and NAC (dot line) methods

References

- Bacchini F., Iannaccone R., Otranto E. (2003): The reconstruction of the number of Italian building permits in 1999, *Atti del Convegno Modelli Complessi e Metodi computazionali per la stima e la previsione*, Treviso 4-6 settembre 2003, Dipartimento di Statistica Università Ca' Foscari Venezia, 33-38.
- Eltine J., Yansaneh I. (1997): Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey, *Survey Methodology*, 23, 33-40.
- Howell D. (2002): Treatment of missing data, *mimeo*, University of Vermont.