COMPARISON BETWEEN DONOR IMPUTATION AND MODEL BASED WEIGHTING IN PRESENCE OF NON-RESPONSE AND RISK OF MODEL MISSPECIFICATION

Roberto Gismondi[1]

ISTAT, Italian National Statistical Institute, Business Short-term Statistics Directorate - Via Tuscolana 1788, 00173 Roma, Italy, gismondi@istat.it

Abstract

In this paper we deal with a statistical survey context, where some population units can not be observed, because of non responses or not inclusion in the theoretical sample. We formalise an estimation strategy based on non responses' deterministic donor imputation. Under a simple super-population model, we develop the mean squared error of the estimator based on donor imputation and alternative minimum distance donor's selection rules and compare its efficiency with respect to the optimal model-based weighting estimation. We discuss the conditions – depending on both the theoretical sampling rate and the response rate – for which donor imputation may improve the ordinary model based predictor. Finally, we present outcomes of an empirical comparison among donor imputation, model based prediction and calibration, where the target variable is wholesale trade turnover.

Keywords: Donor, Imputation, Nearest neighbour, Non-response, Weighting

1. INTRODUCTION

In this context, according to a model-based approach (Valiant *et al*., 2000; Kalton, 2002), we refer to a generic survey context where *n* observations of a target variable are available from a population including *N* units. This theoretical frame characterises quite common operational contexts, as:

1) census or cut/off surveys with non responses, or *late* responses playing the role of non responses if they are not available at an early estimation stage. In the field of official business statistics, two relevant examples are given by the yearly structural survey on firms with more than 99 persons employed (census) and the monthly industrial production index (cut/off), both manayded by ISTAT.

---

[1] *The opinions herein expressed do not involve ISTAT and must be addressed to the author only, as well as possible errors or omissions. Statistical tables derive from elaborations on ISTAT data.*

2) Pure sampling surveys where a preliminary sample of quick respondents (*n* units) is used in order to release flash estimates, the whole sample of respondents (*N* units) is used for releasing final estimates and the estimations strategy is aimed at containing as much as possible the expected difference between *preliminary* and *final* estimates (Gismondi, 2008).

Without loss of generality, we will define as *non respondents* all the units whose data are not available at the estimation stage in operational frameworks as 1) or 2).

In a statistical survey frame, adjustments for tackling non responses are aimed at reducing the potential non response bias (Billiet *et al*., 2007). It often depends on a model misspecification, for instance because respondents and not respondents follow different patterns. Late experiences (ISTAT, 2008) showed that in many real business surveys contexts the non response bias is not systematic, but could happen for some survey occasions and/or for some domains only. Performances of traditional strategies for reducing non response bias are often poor, for instance because too few auxiliary variables are available at the estimation stage (Rizzo *et al*., 1996). In particular, the most part of imputation techniques do not reduce bias enough to balance the increase of variance due to imputation (Copeland and Valiant, 2007).

In the follow, first we resume the basic rules for carrying out prediction of the total in presence of non responses through re-weighting of respondents, under the hypothesis that non response bias can be neglected (section 2). Afterwards (section 3) we formalise an estimation strategy based both on non responses' deterministic donor imputation and optimal model-based estimation through re-weighting of sample units. Donor imputation is frequently used in surveys, but very few variance estimations have been developed. In particular, nearest neighbour imputation is one of the hot deck methods frequently used to compensate for non-response in sample surveys. We derive the mean squared error of the estimator based on donor imputation and carry out a theoretical efficiency comparison with respect to re-weighting of respondent units without imputation. We point out the conditions – depending on both sampling rates and response rates – for which donor imputation may improve the ordinary pseudo-optimal re-weighting. In section 4, the efficiency of donor imputation is evaluated under the hyphotesis that data related to the population under study derive from more than one model.

Finally, we present the main outcomes of an empirical study (section 5) where donor imputation, model based prediction and calibration have been compared, and draw some conclusions in section 5. Empirical results show that, when response rates are quite large (around 90%), broadly speaking donor imputation does not lead to significant improvements of estimates' efficiency. However, for some domains

the donor procedure can be helpful and could be preferred to model based weighting.

2. WEIGHTING UNDER A MODEL BASED APPROACH

The basic foundations of model based weighting can be briefly resumed as follows. We suppose that the main target of the survey is the estimation of an unknown total $Y = N\bar{y}$) of a population *U*, where $\bar{y}$ is the mean of a quantitative variable *y* and *N* is the population size. Estimation is based on a theoretical sample *S* with size *n*, selected according to a given sampling design. We also suppose the following model (m_1) underlying observed data:

$$y_i = \beta x_i + \varepsilon_i \quad \text{were} \quad \begin{cases} E(\varepsilon_i) = 0 & \forall i \\ Var(\varepsilon_i) = \sigma^2 v_i & \forall i \\ Cov(\varepsilon_i, \varepsilon_r) = 0 \quad if \quad i \neq r \end{cases} \quad \text{for each } i \in U \tag{1}$$

where *x* is available for *all* the units in the population, *v* is a variable determining *y* variability and the parameters β and σ^2 are generally unknown. Under the model (1) the unbiasedness condition is given by: $E(T - N\bar{y}) = 0$. We also know that the optimal linear estimator – e.g. the one minimising the *Mse* with respect to the model, $E(T - N\bar{y})^2$ – is (Cicchitelli *et al*., 1992, 385-387):

$$T^* = N[f\bar{y}_S + (1-f)\bar{x}_{\bar{S}}\hat{\beta}^*] = N\hat{\bar{y}}^* \quad \text{where: } \hat{\beta}^* = \left(\sum_S x_i y_i v_i^{-1}\right)\left(\sum_S x_i^2 v_i^{-1}\right)^{-1} \tag{2}$$

with *f*=*n*/*N* and $\bar{y}_S$, $\bar{x}_{\bar{S}}$ equal, respectively, to the sample *y*-mean and the *x*-mean referred to the not observed units and $\hat{\bar{y}}^*$ is the optimal unbiased predictor of the population mean. From (2) it follows that the optimal estimation process is based on the following sampling weighting scheme:

$$T^* = \sum_S w_i^* y_i \quad \text{where: } w_i^* = 1 + (N-n)\bar{x}_{\bar{S}}(x_i v_i^{-1})\hat{\beta}^* \left(\sum_S x_i y_i v_i^{-1}\right)^{-1}. \tag{3}$$

If we label with *j* the generic not observed unit, the model *Mse* of the optimal estimator based on *n* observation is equal to:

$$Mse(T^*; m_1) = \sigma^2 \left[\left(\sum_{\bar{S}} x_j\right)^2 \Big/ \left(\sum_S x_i^2 v_i^{-1}\right) + \sum_{\bar{S}} v_j\right]. \tag{4}$$

One can note that the optimal predictor (2) is equivalent to a predictor which sums all over the population units *N* values – of which *n* are *true* values derived from the observed sample and (*N*-*n*) are *estimated* values – that can be written as:

$$T^* = \sum_S y_i + \sum_{\bar{S}} \hat{y}_j^* = \sum_S y_i + \sum_{\bar{S}} \hat{\beta}^* x_j. \qquad (5)$$

The optimal model estimation through the weighting scheme (3) is equivalent to the optimal model estimation through imputation defined by (5), where the link between optimal weights and optimal imputation rule is given by the second identity in (3). As a consequence, when only one target variable *y* is the main object of estimates, there is not any real contrast between weighting and imputation, as far as in both cases weights w^* or imputation coefficients based on $\hat{\beta}^*$ are the optimal ones with respect to the model. Moreover, if the model (1) is correctly specified, there is no need to search for linear predictors alternative to (2).

However, there could be the possibility to use other imputation rules – corresponding to alternative weighting systems – when:

1) even though the model (1) is correctly specified, a non linear estimator may be used;
2) all and only the units not included in the sample follow a model different than (1);
3) the observed sample includes some units which follow the model defined by (1), but also some units which follow another model (or other models) to be specified. That is coherent with the risk that non responses derive from an *informative* drop-out mechanism (Diggle and Kenward, 1994; Little, 1995).

The first option will be considered in section 3. If situations 2) or 3) occur, then the estimator (1) would be model biased.

3. DETERMINISTIC DONOR IMPUTATION

3.1 BASIC ESTIMATOR

When a donor imputation is used, for each $j \in \bar{S}$ we select a donor - identified by the particular label *i*(*j*) - selected among the *n* available labels $i \in S$ referred to respondents. An equivalent symbolism is d_j=*i*. The donor imputation process can be expressed through the estimation rule:

$$\hat{y}_{jd} = \hat{\beta}_j x_j \quad \text{where:} \quad \hat{\beta}_j = \sum_S a_i y_i = a_{i(j)} y_i \quad \text{and:}$$

$$\begin{cases} a_i = a_{i(j)} = 1 \Big/ x_i^{\alpha} x_j^{1-\alpha} & \text{if} \quad d_j = i \\ a_i = 0 & \text{if} \quad d_j \neq i \end{cases} \quad \text{for each } j \in \bar{S} \tag{6}$$

where α=1 or α=0. Under model (1), the donor imputation rule defined by (6) leads to estimates of each unknown y_j that are model unbiased if the donor is selected among respondent units belonging to the same sub-population to which the *j-th* non respondent belongs and if α=1. In this case, each estimate $y_{i(j)} x_j \big/ x_{i(j)}$ takes into account the different magnitude of the receiving and the donor units and we have: $E(\hat{y}_{jd}) = E(y_j)$. Furthermore, we can define the estimator conditioned to the donor imputation (6), that is given by:

$$T_d = \sum_S y_i + \sum_{\bar{S}} \hat{y}_{jd} = \sum_S y_i + \sum_{\bar{S}} \hat{\beta}_j x_j . \tag{7}$$

where $\hat{\beta}_j$ is defined through (6). The estimator (7) is *not linear* if the choice of donors – and, as a consequence, of coefficients $a_{i(j)}$ – depends on the knowledge of single labels in *S*, e.g. if it is *label dependent* (Cassel *et al.*, 1977, 21). As a consequence, it does not belong to the same family of estimators for which the estimator (2) is optimal and it is worthwhile to find conditions under which it may be preferred to the model-based estimation strategy dealt with in section 2.

In order to guarantee unbiasedness of the estimator (7) in the case α=0 as well, we can define a quite common technique used for selecting donors, based on the *nearest neighbour* method:

$$D_{i(j),j} = \left| x_{i(j)} - x_j \right| = \underset{i \in S}{Min} \left(D_{i,j} \right) \quad \text{for each unit } j \in \bar{S} \tag{8}$$

where $D_{i,j}$ is a distance operator between the couple of units (*i*,*j*). It can be easily shown that when α=0 then $E(\hat{y}_{jd}) = E(y_j)$: since for each couple of units *j* and *i* we can write $y_j = y_i + \beta(x_j - x_i) + (\varepsilon_j - \varepsilon_i)$, it follows $E(y_j - y_i) = \beta(x_j - x_i)$ and this expectation is near zero when $x_j \approx x_i$. As a consequence, if d_j=*i* then $E(y_j - y_i) = E(y_j - \hat{y}_{jd}) \approx 0$. Rule (8) is not necessary for guaranteeing unbiasedness when α=1, but it can be quite helpful when the individual model variability satisfies the position $v = x^{\lambda}$: in this case the choice of donor based on (8) is equivalent to the choice of the unit whose model variance (beyond its model expectation) is quite the same of the receiving unit. Moreover, rule (8) simplifies next *Mse* formulas as well.

The nearest neighbour imputation method has a long history of applications. They often concern census surveys, as the yearly ISTAT survey on enterprises with at least 100 persons employed (ISTAT, 2007), or the Canadian population census (Bankier, 2000). Proper software has been also developed as commented in ISTAT (2005). Although its wide use in practice, few theoretical properties of the nearest neighbour imputation method are known. Chen and Shao (2000) show that under some conditions the nearest neighbour imputation method provides asymptotically unbiased and consistent estimators of functions of population totals and also derive the asymptotic variances for estimators based on nearest neighbour imputation. Brick *et al*. (2004) just introduced and developed the idea of using a model based approach in order to derive sampling variance in presence of hot deck imputation. Beaumont and Bocci (2009) develop a variance estimator for donor imputation based on the assumption that the imputed estimator of a domain total is approximately unbiased under an imputation model.

On the basis of (8), the estimator (7) is not linear because the weight given to each respondent unit depends on the knowledge of labels in the sample (and their related *x*-values). Since under (1) $Bias(T_{(d)}) = 0$, from formula (50) in section 7 we obtain:

$$Mse(T_d / m_1) = Var\left(\sum_{\bar{S}} y_j\right) + Var\left(\sum_{\bar{S}} \hat{y}_{jd}\right). \tag{9}$$

We have: $Var(\hat{y}_{jd}) = Var(y_{i(j)} x_j / x_{i(j)}) = \sigma^2 (x_j / x_{i(j)})^2 v_{i(j)}$. It is quite realistic to guess also that $v_{i(j)} \approx v_j$ for each unit $j \in \bar{S}$, meaning that through the donor selection rule (8) the model variability expressed by the variable *v* is more or less the same as regards both the donor unit *i*(*j*) and the receiving one *j*. As a consequence:

$$\begin{aligned} Mse\left(T_d / m_1\right) &= \sigma^2 \sum_{\bar{S}} v_j + \sigma^2 \sum_{\bar{S}} \left(x_j / x_{i(j)}\right)^2 v_{i(j)} = \\ &= \sigma^2 \sum_{\bar{S}} v_j + \sigma^2 \sum_{i \in S} \left[\sum_{j \in \bar{S}, d_j = i} \left(x_j / x_i\right)^2 v_i \right]. \end{aligned} \tag{10}$$

The meaning of the last right term is that, for each fixed label *i* in the first sum, the sum into squared brackets includes a number of terms equal to the number of not respondent units *j* whose donor is given by that particular unit *i*. If one imposes the constraint that each respondent unit may be a donor for not more than one non respondent unit, then this sum will include (*N*-*n*) terms. Given the formulas (4) and

(10), we can verify under which conditions one can have:

$$Mse(T_d) \le Mse(T^*) . \tag{11}$$

3.2 PARTICULAR CASES

We suppose: v_i=x_i for each unit *i* in the population. That simplifies formula (10) where we can put $v_{i(j)} = x_{i(j)} \approx v_j = x_j$. As a consequence, we obtain:

$$Mse\left(T_{d,v=x} / m_1\right) = \sigma^2 \sum_{\bar{S}} x_j + \sigma^2 \sum_{\bar{S}} x_j = 2\sigma^2 \sum_{\bar{S}} x_j = 2\sigma^2 X_{\bar{S}} \tag{12}$$

From (4) we have also:

$$\begin{aligned} Mse\left(T^*_{v=x} / m_1\right) &= \sigma^2 \left[\left(\sum_{\bar{S}} x_j \right)^2 \Big/ \left(\sum_{S} x_i \right) + \sum_{\bar{S}} x_j \right] = \\ &= \sigma^2 \left(X_{\bar{S}}^2 X_S^{-1} + X_{\bar{S}} \right) = \sigma^2 X_{\bar{S}} \left(X_{\bar{S}} X_S^{-1} + 1 \right). \end{aligned} \tag{13}$$

From (12) and (13) we can verify that the condition (11) is satisfied if:

$$X_S \le X_{\bar{S}} \qquad \leftrightarrow \qquad X_S \le X_U / 2 . \tag{14}$$

The relation (14) implies that the donor estimation strategy improves the model based one if the weighted response rate is not larger than the whole *x*-total (*x*-coverage not larger than 50%).

If: $v_i = x_i^2$ for each unit *i* in the population, we obtain:

$$Mse\left(T_{d,v=x^2} / m_1\right) = 2\sigma^2 \sum_{\bar{S}} x_j^2 \qquad Mse\left(T^*_{v=x^2} / m_1\right) = \sigma^2 \left(X_{\bar{S}}^2 / n + \sum_{\bar{S}} x_j^2 \right) \tag{15}$$

and the condition (11) will be satisfied if:

$$X_{\bar{S}}^2 \ge n \sum_{\bar{S}} x_j^2 . . \tag{16}$$

If: v_i=1 for each unit *i* in the population, we obtain:

$$Mse\left(T_{d,v=1} / m_1\right) = 2\sigma^2 \left(N - n\right) \qquad Mse\left(T^*_{v=1} / m_1\right) = \sigma^2 \left[X_{\bar{S}}^2 / \sum_{S} x_i^2 + \left(N - n\right) \right] \tag{17}$$

and the condition (11) will be satisfied if:

$$X_{\bar{S}}^2 \geq (N-n)\sum_S x_i^2 . \tag{18}$$

Finally, if: $v_i = x_i = 1$ for each unit i in the population, the condition (11) will be satisfied if:

$$n \leq N/2 . \tag{19}$$

However, in this case the choice of donors could not be based on the rule (8), but according to other deterministic criteria, as the same rule (8) applied to another auxiliary variable z.

Broadly speaking, rules (14), (16), (18) and (19) suggest that estimation based on donor imputation can be preferred when the size of the observed sample – expressed through the number or sampling units or the amount of x depending on the hypotheses on v – is not particularly large and, in particular, is less than 50% of the whole population size.

3.3 LACK OF PRECISION IN THE DONOR IMPUTATION PROCESS

Results from (12) to (19) have been obtained supposing that $v_{i(j)} = x_{i(j)} \approx v_j = x_j$ for each unit $j \in \bar{S}$. However, the donor rule (8) may often identify donors whose x-value is different from that of the corresponding receiving unit. This situation is quite common when x is a continuous variable (for instance, turnover). We can suppose: $x_j = \alpha_j x_{i(j)}$ with $\alpha_j \neq 1$ for each unit $j \in \bar{S}$. As a consequence:

$$Mse\left(T_d / m_1\right) = \sigma^2 \sum_{\bar{S}} v_j + \sigma^2 \sum_{\bar{S}} \alpha_j^2 v_{i(j)} . \tag{20}$$

When $v_i = x_i$ for each unit i, we have: $v_j = \alpha_j v_{i(j)}$, so that $Mse(T_{d,v=x} / m_1) = \sigma^2 \sum_{\bar{S}} (1+\alpha_j) x_j$.

When $v_i = x_i^2$ for each unit i, we have: $v_j = \alpha_j^2 v_{i(j)}$, so that $Mse(T_{d,v=x^2} / m_1) = \sigma^2 \sum_{\bar{S}} (1+\alpha_j) x_j^2$.

When $v_i = 1$ for each unit i, we have: $Mse(T_{d,v=x=1} / m_1) = 2\sigma^2 \left(N-n\right)$.

When for each unit i, we have: $Mse(T_{d,v=x=1} / m_1) = 2\sigma^2 \left(N-n\right)$.

It follows that condition (11) will be satisfied:

when $v_i = x_i$ for each unit i, if: $\quad X_S \leq \dfrac{X_{\bar{S}}^2}{\sum_{\bar{S}} \alpha_j x_j}$; $\tag{21}$

when $v_i = x_i^2$ for each unit *i*, if: $X_{\bar{S}}^2 \geq n \sum_{\bar{S}} \alpha_j x_j^2$; (22)

when $v_i = 1$ for each unit *i*, if: $X_{\bar{S}}^2 \geq \sum_{\bar{S}} \alpha_j^2 \sum_{S} x_i^2$; (23)

when $v_i = x_i = 1$ for each unit *i*, if: $n \leq N/2$. (the same condition 19)

Broadly speaking, the other conditions being steady usefulness of estimation based on donor imputation decreases as coefficients α_j tend to be large in correspondence of units characterised by large *x*-values.

3.4 AN ALTERNATIVE DONOR ELECTION CRITERION

The donor estimation defined by (6) implies that for each unit *j* the donor imputation would lead to an imputation error near zero if $y_j \approx y_i x_j / x_i$. If one knew the true value y_j, then the previous condition would be satisfied if the donor's choice is based on the alternative rule:

$$D'_{i(j),j} = \left| \frac{y_j}{x_j} - \frac{y_{i(j)}}{x_{i(j)}} \right| = Min_{i \in S} \left(D'_{i,j} \right) \text{ for each unit } j \in \bar{S} \tag{24}$$

where $D'_{i,j}$ is an alternative distance rule based on ratios. Of course, rule (24) is tautological since the true value y_j is not available by definition; however, that can be reasonably substituted by the new fully operational rule:

$$\overset{\circ}{D}_{i(j),j} = \sum_{t=1}^{T-1} \left| \frac{y_{tj}}{x_{tj}} - \frac{y_{ti(j)}}{x_{ti(j)}} \right| = Min_{i \in S} (\overset{\circ}{D}_{i,j}) \text{ for each unit } j \in \bar{S}. \tag{25}$$

where *t* is a time label, $\overset{\circ}{D}_{i,j}$ is a new distance rule based on the sum of differences between ratios including (*T*-1) terms and *T* is the time of current estimates with $T \geq 2$. The basic rationale of (25) is the possibility to use longitudinal micro-data in order to find a donor which could approximately satisfy rule (24) as well. If *t* is a month, terms labelled with (*t*-*r*) may be referred to the month (*t*-*r*) of the same reference year or to the same month *t*, *r* years before that of current estimates. In particular:

- the use of (25) will lead to a donor which will approximately satisfy (24) as well only if the underlying model (1) remains quite steady along time (no relevant changes of parameters β and/or σ);

- taking into account the previous remark, the number of addenda in the sum in (25) should be limited to 1, 2 or 3;
- each absolute difference in (25) may be weighted through a time coefficient whose level decreases as *t* decreases towards one;
- the selection of donors based on (24) or (25) does not imply that the donor unit found has a model variance similar to that of the receiving unit: for instance, relation (24) may be satisfied by a certain unit $i(j)$ such that $x_{i(j)} \neq x_j$.

According to the last remark, two potential improvements of (25) may be based on the two following options:

a) the search for donors satisfying (25) may be restricted to the only potential donors labelled as *i* such that $D^{x}_{i,j} = \left| x_{Tj} - x_{Ti} \right| \leq \delta$, where δ is an arbitrary (small) constant > 0;

b) for each unit $j \in \bar{S}$ the donor will be given by the particular unit $i(j)$ which minimises the function: $\omega D^{\circ}_{i(j),j} + (1-\omega) D^{x}_{i(j),j}$, where ω is an arbitrary constant ranging in [0,1].

The use of (25) coupled with (6) guarantees that the estimator (7) is still unbiased if α=1. If α=0, the estimator (7) is biased, but it would be approximately unbiased if the above mentioned option a) is used, since in this case (time labels are omitted): $\left| E(y_j - \hat{y}_{jd}) \right| = \left| E(y_j - y_{i(j)}) \right| = \beta \left| x_j - x_{i(j)} \right| \leq \beta\delta$.

The general formula of the model *Mse* of the estimator T°_{d} based on (7) and (25) is given by:

$$Mse\left(T^{\circ}_{d}/\alpha, m_1\right) = Var\left(\sum_{\bar{S}} y_j\right) + Var\left(\sum_{\bar{S}} \hat{y}_{jd}\right) + Bias^2\left(T^{\circ}_{d}/\alpha, m_1\right) \qquad (26)$$

that can be also written as:

$$Mse\left(T^{\circ}_{d}/\alpha, m_1\right) = \sigma^2\left[\sum_{\bar{S}} v_j + \sum_{\bar{S}} (x_j / x_{i(j)})^2 v_{i(j)}\right] + (1-\alpha)\beta^2\left[\sum_{\bar{S}} (x_j - x_{i(j)})\right]^2 \qquad (27)$$

When α=1 the previous formula corresponds to the first equality in formula (10) since bias is null, while when α=0 there is the additional last squared term which depends on the overall difference between the *x* levels of each couple of receiving and donor units. Let's note that this sum may be near zero if positive and negative differences compensate each other.

4. MORE MODELS

4.1 LINKS BETWEEN DONOR IMPUTATION AND POST-STRATIFICATION

There is another interpretation of the donor imputation rule defined by (6) when α=1. Let's suppose that each respondent unit may be selected as donor once at most. Since $\hat{y}_{jd} = (y_{i(j)} / x_i)x_j = \hat{\beta}_j x_j$, the implicit assumption is that for each non respondent unit *j* imputation is based on a particular coefficient β_j whose estimate is $\hat{\beta}_j = y_{i(j)} / x_i$. This consideration steers to the evaluation of all the available ratios y_i/x_i for each $i \in S$, with the purpose of identifying *k* post-strata which may be re-conducted to *k* super-population models according to the following steps.

1) Let's suppose that each of the *n* ratios y_i / x_i satisfy the following empirical rule:

$$y_i / x_i \in (\hat{\beta}_h - \Delta_h; \hat{\beta}_h + \Delta_h) \text{ for each } i \in S \text{ and one } h=1,2,\ldots,k. \qquad (28)$$

   Of course (28) is always satisfied when $k=n$ and each $\Delta_h=0$, but we can suppose that it is still satisfied for some $k<n$ and some Δ_h reasonably small. In this case, we are implicitly supposing that the number of super-population models underlying respondent units' data is *k* instead of 1 (as it has been supposed until now).

2) For each non respondent unit $j \in \bar{S}$ a donor is selected according to rules (24) or (25) and estimation (6) is carried out. As a consequence, if the particular donor *i*(*j*) belongs to the post-stratum *h*, then the receiving unit *j* is assigned to the same post-stratum as well by definition, because $\hat{\beta}_j = y_{i(j)} / x_i$.

3) In the end, each post-stratum will contain both respondent and non respondent units, even though some post-strata may contain respondents only.

Implicitly, the procedure described through the three previous steps is based on the idea of clustering available data in order to achieve to better estimates and may be connected to the proposals by Bianchi *et al.* (2005). Even though step 2) may be carried out using rule (8) for donors' selection as well, the basic rationale deriving from (28) is that rules (24) or (25) should be preferred. Inside each post-stratum model based or donor estimation may be used according to the same strategies discussed in sections 2 and 3. We remark that, broadly speaking, definition of post-strata is normally based on models identification criteria more robust with respect to the couple of empirical rules 1) and 2) described above (see also section 4.3). However, more skilled criteria need for significant auxiliary

variables and time series not always available in real contexts. Moreover, one may trust more the simpler hypothesis that one only model exists instead of several ones. As a consequence, we should consider the case when no *ex-ante* post-stratification can be carried out, because available information are not sufficient in order to identify more than one model and/or one guesses that serious classification mistakes may occur. In other terms, we can suppose that sub-populations corresponding to post-strata depend on some latent factors underlying units under observation that can not be observed *before* drawing the sample.

In this framework, it is useful to compare the performance of a donor estimation strategy as that discussed along section 3 – which implies a post-stratification of respondent units on the basis of (28) – with respect to efficiency of a model based estimation strategy when one only model is supposed, even though k models exist.

4.2 ESTIMATION STRATEGIES UNDER *K* MODELS

The optimal model based estimation under k models can be briefly resumed. We suppose that the population U can be split into k separate sub-populations U_h with h=1,2,…,k, including each N_h units, with $U = \bigcup_{h=1}^{k} U_h$. For each sub-population this model (m_k) is supposed true:

$$y_{hi} = \beta_h x_i + \varepsilon_{hi} \text{ where } \begin{cases} E(\varepsilon_{hi}) = 0 & \forall h,i \\ VAR(\varepsilon_{hi}) = \sigma_h^2 v_i & \forall h,i \\ COV(\varepsilon_i, \varepsilon_j) = 0 \quad if \quad i \neq j \end{cases} \text{ for } h=1, 2, \ldots,k \qquad (29)$$

where all symbols keep the same logical meaning as for model (1). We have also $S = \bigcup_{h=1}^{k} S_h$ and $n = \sum_{h=1}^{k} n_h$. Prospect 1 supplies an overall resuming scheme when k=2.

Prospect 1: ***Different patterns for two sub-populations 1 and 2***

DOMAIN	STRUCTURE			SIZE		
	Sub-population			Sub-total		
	Population	1	2	**Total**	1	2
Universe	$\boldsymbol{U}$	$\boldsymbol{U_1}$	$\boldsymbol{U_2}$	$\boldsymbol{N}$	$\boldsymbol{N_1}$	$\boldsymbol{N_2}$
Observed	$\boldsymbol{S}$	$\boldsymbol{S_1}$	$\boldsymbol{S_2}$	$\boldsymbol{n}$	$\boldsymbol{n_1}$	$\boldsymbol{n_2}$
Not observed	$\boldsymbol{S}$	$\boldsymbol{\bar{S}_1}$	$\boldsymbol{\bar{S}_2}$	$\boldsymbol{N\text{-}n}$	$\boldsymbol{N_1 - n_2}$	$\boldsymbol{N_1 - n_2}$

If Y_h is the unknown *y*-total in the *h-th* sub-populations and X_h is the correspondent *x*-total, the unknown total will be given by:

$$Y = \sum_{h=1}^{k} Y_h \quad \text{where:} \quad E(Y) = \sum_{h=1}^{k} \beta_h X_h \,. \tag{30}$$

A general formula for the final predictor of the population mean is given by:

$$T_{(k)} = \sum_{h=1}^{k} T_h = \sum_{h=1}^{k} N_h \hat{\bar{y}}_h \tag{31}$$

where T_h is a predictor of Y_h and $\hat{\bar{y}}_h$ is a predictor of $\bar{y}_h$. According to (2), if the optimal predictors are used separately for each sub-population *h* we can put:

$$T^*_{(k)} = \sum_{h=1}^{k} T^*_h = \sum_{h=1}^{k} N_h \hat{\bar{y}}^*_h \tag{32a}$$

$$\hat{\bar{y}}^*_h = f_h \bar{y}_{S_h} + (1 - f_h)\bar{x}_{\bar{S}_h}\hat{\beta}^*_h \text{ with } \hat{\beta}^*_h = \left(\sum_{S_h} x_i y_i v_i^{-1}\right)\left(\sum_{S_h} x_i^2 v_i^{-1}\right)^{-1} \text{ for } h=1,2,\ldots,k. \tag{32b}$$

The mean squared error of the optimal unbiased predictor is:

$$Mse(T^*_{(k)} / m_k) = \sum_{h=1}^{k} Mse(T^*_h) = \sum_{h=1}^{k} \sigma_h^2 \left[\left(\sum_{\bar{S}_h} x_j\right)^2 \Big/ \left(\sum_{S_h} x_i^2 v_i^{-1}\right) + \sum_{\bar{S}_h} v_j\right]. \tag{33}$$

Whatever predictor can be written as (31) even when the split into *k* sub-populations is not formally introduced, as it happens when the final estimation is carried out mixing together units deriving from different sub-populations. In particular, if one supposes the model m_1 while the right working model is m_k, then the pseudo optimal predictor (2) is model biased under (29) and its model bias will derive from the following formula:

$$\begin{aligned} E(T^* / m_k) &= E(Y) + \left[X_{\bar{S}} \sum_{h=1}^{k} \left(\beta_h \sum_{S_h} x_i^2 v_i^{-1}\right)\left(\sum_{S} x_i^2 v_i^{-1}\right)^{-1} - \sum_{h=1}^{k} \beta_h X_{\bar{S}_h}\right] = \\ &= E(Y) + Bias\left(T^* / m_k\right) \end{aligned} \tag{34}$$

If the right model (29) is ignored, the model bias due to the use of the predictor

(2) will be negligible if these conditions hold:

$$\left(\sum_{S_h} x_i^2 v_i^{-1}\right)\left(\sum_{S} x_i^2 v_i^{-1}\right)^{-1} \approx \left(\sum_{\bar{S}_h} x_j\right)\left(\sum_{\bar{S}} x_j\right)^{-1} \quad \text{for } h=1,2,\ldots,k. \tag{35}$$

When $x=v$, the previous conditions will be approximately satisfied if each sample S_h is a *balanced sample* with respect to its reference sub-population U_h. On the basis of formula (50) in section 7, the *Mse* of T^* under the model (29) is:

$$Mse(T^*/m_k) = \sum_{h=1}^{k}\left(\sigma_h^2 \sum_{\bar{S}_h} v_j\right) + \left(\sum_{\bar{S}} x_j\right)^2 \left(\sum_{S} x_i^2 v_i^{-1}\right)^{-2}\left[\sum_{h=1}^{k}\left(\sigma_h^2 \sum_{S_h} x_i^2 v_i^{-1}\right)\right] + Bias^2(T^*/m_k) \tag{36}$$

It must be remarked that bias may not be equal to zero even when (31) is used, because some units belonging to the observed and/or the not observed population may have been miss-classified: the consequent bias would be due both to the biased estimation of totals in each sub-population and the wrong evaluation of N_h, that until now has been supposed as known.

If the right model is m_k, with obvious generalisations the estimator conditioned to the donor imputation (6) carried out *separately* in each sub-population h is given by:

$$T_{(k)d} = \sum_{h=1}^{k} T_{hd} \quad \text{where:} \quad T_{hd} = \sum_{S_h} y_i + \sum_{\bar{S}_h} \hat{y}_{jd} \tag{37}$$

with:

$$Mse\left(T_{(k)d}/m_k\right) = \sum_{h=1}^{k} Mse(T_{hd}) = \sum_{h=1}^{k} \sigma_h^2 \left[\sum_{\bar{S}_h} v_j + \sum_{\bar{S}_h} (x_j / x_{i(j)})^2 v_{i(j)}\right] \tag{38}$$

The comparison between donor imputation and model based estimation when in both cases the right model m_k is assumed, is equivalent to compare (38) and (33), on the basis of the same argumentations already seen in section 3.2 applied inside each post-stratum. However, we can also evaluate conditions under which (38) is not larger than (36), that is when model based estimation does not take into account the right model (29) properly. Formally, this condition is:

$$Mse\left(T_{(k)d}/m_k\right) \le Mse\left(T^*/m_k\right). \tag{39}$$

Two basic conditions will be implicitly supposed as satisfied:

1) the post-stratification derived from (28) can *correctly* identify the k sub-populations. However, in real practice mistakes can occur as regards both k and, given the number of post-strata, the correct attribution of each unit to post-strata.
2) For sake of simplicity, for each receiving unit j we have that $x_{i(j)} \approx x_j$ – e.g. that no lack of precision in the donor imputation process occurs (section 3.3).

We can suppose the donor imputation rule (8) such that $v_{i(j)} \approx v_j$ for each $j \in \bar{S}$ and take into account the following particular cases.

When: $v_i=x_i$ for each unit i in the population, from (36) and (38) we obtain:

$$Mse\left(T^*/m_k\right) = \sum_{h=1}^{k} \sigma_h^2 \left[X_{\bar{S}_h} + \left(\frac{X_{\bar{S}}}{X_S} \right)^2 X_{S_h} \right] + \left[\sum_{h=1}^{k} \beta_h \left(\frac{X_{\bar{S}}}{X_S} X_{S_h} - X_{\bar{S}_h} \right) \right]^2 \qquad (40)$$

$$Mse\left(T_{(k)d}/m_k\right) = 2\sum_{h=1}^{k} \sigma_h^2 X_{\bar{S}_h} . \qquad (41)$$

As a consequence, we obtain that (39) is satisfied if:

$$\sum_{h=1}^{k} \sigma_h^2 X_{\bar{S}_h} \le \left(\frac{X_{\bar{S}}}{X_S} \right)^2 \sum_{h=1}^{k} \sigma_h^2 X_{S_h} + \left[\sum_{h=1}^{k} \beta_h \left(\frac{X_{\bar{S}}}{X_S} X_{S_h} - X_{\bar{S}_h} \right) \right]^2 . \qquad (42)$$

Of course, condition (41) reduces to (14) when k=1.

When $v_i = x_i^2$ for each unit i, we obtain that (29) is satisfied if:

$$\sum_{h=1}^{k} \sigma_h^2 \sum_{\bar{S}_h} x_j^2 \le \left(\frac{X_{\bar{S}}}{n} \right)^2 \sum_{h=1}^{k} \sigma_h^2 n_h + \left[\sum_{h=1}^{k} \beta_h \left(\frac{X_{\bar{S}} n_h}{n} - X_{\bar{S}_h} \right) \right]^2 . \qquad (43)$$

Finally, when $v_i=x_i=1$ for each unit i in the population, we obtain that (29) is satisfied if:

$$\sum_{h=1}^{k} \sigma_h^2 (N_h - n_h) \le \left(\frac{N-n}{n} \right)^2 \sum_{h=1}^{k} \sigma_h^2 n_h + \left[\sum_{h=1}^{k} \beta_h \left(\frac{(N-n) n_h}{n} - (N_h - n_h) \right) \right]^2 \qquad (44)$$

4.3 IDENTIFYING AND TESTING SUB-POPULATIONS

The search for an efficient grouping of observed data in presence of non responses is not a new idea. Following the classical approach (Särndal and

Lundström, 2005, 94-96), the design bias of a post-stratified estimator can be strongly reduced if in each post-stratum the *y*-mean of respondents and non respondents are quite similar and different post-strata are characterised by different average response rates.

Before carrying out estimates, it can be useful to verify if a m_k model as (29) should be preferred to the simpler m_1 model (1). Unrealistic assumptions on super-populations which generate observed data may seriously affect efficiency of estimates even when robust estimation techniques are used (Hedlin *et al*., 2001). In particular, model choice is crucial especially when non response problems must be faced (Lehtonen *et al*., 2003; Ibrahim *et al*., 2008). Even though robust analytical principles may be used (Nishii, 1988), in the field of many current surveys constraints due to timeliness and auxiliary information availability steer the recourse to more heuristic techniques.

We have seen in section 4 how a donor imputation strategy implicitly corresponds to a post-stratification of population into *k* post-strata, with *k* ranging from 1 to *n*. However, this post-stratification technique could be compared with other techniques, especially if the availability of one or more auxiliary variables for all the population units lets the possibility to use more precise algorithms.

Broadly speaking, the basic idea is that structural differences could be tested evaluating different average levels of the *y*-variable and/or the *x*-variable, even when structural differences could concern variability as well: the implicit, but realistic underlying hypothesis justifying this approach is that different average levels imply a different average variability, and vice-versa.

Another quite simple procedure is based on the method proposed by Cochran (1977, 128-130) for stratifying a given population, in order to minimise the variance of estimates in a stratified random sampling context. If z_i is the value assumed on the *i-th* population unit by an auxiliary variable *z* correlated with *y*, and *i'* is the place occupied by the *i-th* unit in the decreasing ranking of *z*-values, the rule for identifying boundaries of *k* sub-populations is based on the following equality:

$$\sum_{i' \in U_h} \sqrt{z_{i'}} \approx \sum_{i' \in U} \sqrt{z_{i'}} / k \qquad (45)$$

where U_h is the *h-th* sub-population including N_h units. Rule (45) means that the sum of the square roots of the *z*-values in the *h-th* sub-population must be almost equal to the same sum calculated in each of the other (*k*-1) sub-populations. The variable *z* may be given by the same variable *x* in model (29): for instance, *z* may be equal to the *y*-variable delayed. Even though the Cochran method is quite simple, it does

not properly take into account individual variability.

An alternative procedure is based on a cluster analysis algorithm aimed at identifying *k* clusters. A first option consists in the use of a matrix including *N* rows and *r* auxiliary variables z_r available for all the units in the population, where when *r*=1 it may be *z*=*x*. A second option is based on three steps: 1) on the basis of the vector containing the *n y*-values observed in the sample, *k* clusters are identified; 2) on the basis of the same matrix of auxiliary variables introduced above, a discriminant analysis or a logistic model are used in order to identify the most significant variables which the previous split depends on; 3) on the basis of results derived from step 2), each not observed unit is assigned to one specific cluster among those identified at step 1).

Let's note that, while the first option is based on a (post-)stratification that could have been done also before drawing the sample, the second one corresponds more strictly to the logic of a post-stratification based on the observed *y*-values.

Results of the post-stratification can be tested verifying the statistical significance of the difference between parameters estimated separately into 2 post-strata, using the same tools available for testing rightness of a general linear model (Gismondi, 2008). For instance, as regards expected values, according to the model (29) one can test the null hypothesis: $\beta_{h_1} = \beta_{h_2}$ against the alternative; $\beta_{h_1} \neq \beta_{h_2}$ for any couple of post-strata (h_1, h_2). Testing can be based on the random variable:

$$\left(\hat{\beta}_{h_1} - \hat{\beta}_{h_2}\right) \Big/ \sqrt{\left[\left(n_{h_1}-2\right)Var\left(\hat{\beta}_{h_1}\right)+\left(n_{h_2}-2\right)Var\left(\hat{\beta}_{h_2}\right)\right] \Big/ \left(n_{h_1}+n_{h_2}-4\right)} \qquad (46)$$

that is approximately a Student's *t* with $(n_{h_1}+n_{h_2}-4)$ degrees of freedom[2], where n_{h_1} is the number of observation belonging to post-stratum h_1. The same test (46) can be also used for testing structural differences between respondent and non respondent units, provided that both respondents' and non respondents' data are considered as samples with size *n* and (*N*-*n*) drawn from the realised population[3].

---

[2] *In each post-stratum h 2 degrees of freedom are lost because of the need of estimating both* β_h *and* δ_h*, where the latter estimate is necessary since from model (1) we have:* $Var(\hat{\beta}_h^*) = \sigma_h^2 \sum_{S_h} x_i \Big/ \sum_{S_h} x_i^2 v_i^{-1}$.

[3] *As regards testing variance structure, we address to Gismondi (2008, 105).*

5. AN APPLICATION TO REAL WHOLE SALE TRADE DATA

5.1 OVERVIEW ON THE SURVEY

Data used for the empirical attempt derive from the quarterly wholesale trade sample survey carried out by ISTAT. It is based on a stratified random sample that includes about 7.500 enterprises with a yearly rotation of about 2.000 units. The original stratification is based on NACE Rev.1.1 economic activities, employment classes and geographic areas. Even though the main purpose of the survey is the estimation of quarterly turnover indexes with base 2005=100[4] – based on turnover data picked up quarterly – in this context we will focus on the estimation of the quarterly total turnover for the years 2005-2007. We have supposed the following framework:

1) the population (size N) is given by the sample of *final* respondents, which are the units that have responded within 180 days from the end of the reference quarter (in the real survey context, final estimates are just released after 6 months);
2) the sample (size n) is given by the sub-sample of *quick* respondents, which are those responding within 90 days from the end of the reference quarter.

Separate estimations have been carried in 14 strata obtained crossing each other 2 employment classes (1-19, >19) and the following 7 groups of economic activities: 1) *Wholesale on a fee or contract basis*; 2) *Agriculture raw materials and live animals*; 3) *Food, beverages and tobacco*; 4) *Household goods*; 5) *Non-agriculture intermediate products*; 6) *Machinery, equipment and supplies*; 7) *Other products*. Also the *Total wholesale trade* has been taken into account as weighted mean of the previous 7 economic activities[5]. Economic activities also correspond to the final estimation domains.

In order to apply the model (1), we have considered as y-variable the total turnover referred to the each quarter q of the year (Y), while the x-variable has been given by turnover of the same quarter q of the previous year (Y-1): both these data are asked for in the same questionnaire referred to the quarter q of the year (Y). On the basis of a preliminary analysis aimed at reducing the influence of outlier observations on the estimates, about the 2% of available data has been excluded from further analyses[6]. Moreover, for each quarter only units whose quarterly

---

[4] *For more details we address to ISTAT (2008, 89-103).*

[5] *Weights derive from structural business statistics data and are proportional to the yearly turnover of the year 2005.*

[6] *Basically, we have excluded units whose turnover of quarter q in (Y-1) was not available, or was lower than a fixed threshold (10.000 Euros).*

turnover referred to quarter *q* of the year (*Y*-2) have been used, in order to apply bivariate calibration estimation (section 5.2). On the average 2005-2007, for the total wholesale trade and each quarter we have *N*=4.395 and *n*=534, so that 100·*n*/*N*= 87,5. According to the theoretical issues derived from table 1, we should not obtain significant efficiency gains using minimum distance donor imputation.

5.2 COMPARED STRATEGIES

A first group of strategies is based on predictor (2) and optimal weighting under the model (1), with the options $v=x=1$, $v=1$ and $v=x$ (table 1).

A second group is based on the idea to apply optimal model based prediction in separate populations, supposing 2 populations and 2 corresponding models, where post-strata are defined through the rule (28). The predictor is given by (32a) with $k=2$, supposing $v=1$ or $v=x$.

The third group uses donor imputation, according to the predictor defined by (6) and (7), the rule (8) for donors' detection and the options $\alpha=1$ or $\alpha=0$. The *x*-variable used in (8) is turnover of the same quarter of the previous year. An additional constraint imposed is that each respondent unit can be selected as donor for not more than 3 times for each quarter and domain[7].

The fourth group is still identified by donor imputation according to the predictor defined by (6) and (7), but using the rule (24) for donors' detection, putting $\alpha=1$. The donor selection rule (25) has been applied putting $T=2$, e.g. so that for each quarter and each non respondent unit *j* the donor *i*(*j*) minimises: $\left| y_{(Y-1),j} / x_{(Y-1),j} - y_{(Y-1),i(j)} / x_{(Y-1),i(j)} \right|$.

The fifth group is based on calibration which, as well known, is a widely used strategy for managing non response bias (Lundström and Särndal, 1999). In this context, calibration estimation has been applied on the basis of the following estimator:

$$T_{cal} = \sum_{S} w_{i,cal} y_i \tag{47}$$

under the constraint:

$$\sum_{S} (w_{i,cal} - w_i)^2 = \underset{w_i}{Min} \left\{ \sum_{S} (w_i' - w_i)^2 ; \sum_{S} w_i' z_{li} = \sum_{U} z_{li}, l = 1,2,...,L \right\} \tag{48}$$

---

[7] *Of course, because of this constraint in some cases the* x*-values of donor and receiving units may be quite different, as commented in section 3.3.*

where $w_{i,cal}$ are the calibrated weights which guarantee the constraint (48) and z_l are auxiliary variables whose total amount is known, for l=1,2,…,L. In the context under study, two calibration approaches have been tested, both using original model based weights with v=1: a) calibration with L=1, where z_1 is the quarterly turnover of year (Y-1); b) calibration with L=2, where z_1 is defined as above and z_2 is the quarterly turnover of year (Y-2). Let's note that in order to apply approach b) it has been necessary to include in the database quarterly data referred to 2003 as well (that is the year (Y-2) as regards 2005 data).

Finally, a sixth group of strategies consists in the use, for each estimation domain, of the best strategy among those listed above. In particular, we have identified the (*ex-post*) composite best strategy separately for groups (1) and (2) (model based weighting) and groups (3) and (4) (donor imputation). This mixed strategy represents a tool for synthesising efficiency of weighting and donor based strategies.

The basic quality indicator used for comparing efficiency of the various strategies is the *MAPE*, that in this context, for each strategy *Str*, is given by the mean of the absolute percent errors of estimates evaluated on the 4 quarters of each year:

$$MAPE(T_{Str}) = 100 \cdot \sum_{Y=2005}^{2007} \sum_{q=1}^{4} \frac{\left| \hat{T}_{Str,(Y),q} - T_{(Y),q} \right|}{T_{(Y),q}} . \qquad (49)$$

Formula (49) has been computed for each of the 8 domains listed in section 5.1; that is the main input of tables 4 and 5, which report the average *MAPE* calculated as mean of 8 domains as well. The average *MAPE* is the simple arithmetic mean of *MAPEs* referred to 8 domains.

Even though the basic empirical attempt is based on the use of real data, in order to assess steadiness of results an additional simulation study has been carried out, through blanking of the 50% of responses for each quarter and each stratum (7 economic activities by 2 employment classes) replicated at random 1.000 times. Main results have been summarised in tables 6 and 7, which report *MAPEs* obtained as arithmetic means of 1.000 *MAPEs* concerning each estimation domain and each strategy.

Tab. 1: Compared estimation strategies.

Code	Definition	Options	Details
(1)	Optimal model based prediction (m_1)	v=x=1	Predictor (2) with v=1
(1)	Optimal model based prediction (m_1)	v=1	Predictor (2) with v=x
(1)	Optimal model based prediction (m_1)	v=x	Predictor (2) with v=x=1
(2)	Optimal model based prediction (m_2) - Method (28)	v=1	Predictor (32a) with k=2, v=1, method (28) for identifying sub-populations
(2)	Optimal model based prediction (m_2) - Method (28)	v=x	Predictor (32a) with k=2, v=x, method (28) for identifying sub-populations
(3)	Donor imputation (formulas (6), (7), (8))	α=1	Predictor defined by (6) and (7) with ·=1 and rule (8) for donors' detection
(3)	Donor imputation (formulas (6), (7), (8))	α=0	Predictor defined by (6) and (7) with ·=1 and rule (8) for donors' detection
(4)	Donor imputation (formulas (6), (7), (24))	α=1	Predictor defined by (6) and (7) with ·=1 and rule (24) for donors' detection
(5)	Calibration z_1 (1 constraint)	v=1	Calibration estimator (48), original weights w from (3) with v=1, L=1, z_1 is the quarterly turnover of year (Y-1)
(5)	Calibration z_1, z_2 (2 constraints)	v=1	Calibration estimator (48): as in the previous case but with L=2, z_2 is the quarterly turnover of year (Y-2)
(6)	Best composite model based prediction		Best strategy among (1) and (2) for each domain
(6)	Best composite donor imputation prediction		Best strategy among (3) and (4) for each domain
(6)	Best composite calibration prediction		Best strategy (5) for each domain

5.3 MAIN RESULTS

A preliminary step consisted in testing, in each domain, statistical significance of the difference between the regression coefficient β evaluated separately on respondent and non respondent units. The test function is given by (46) and the domains are the seven economic activities defined in section 5.1. Provided that a linear model as (1) fits quite well available data[8], testing structural differences between respondent and non respondents is a simple tool for assessing the potential presence of non response bias.

[8] *Results of these preliminary tests have been omitted.*

According to table 2, the only domain for which the *T* test is significant at 95% level is the first one (*Wholesale on a fee or contract basis*) when *v*=1. Statistical significance characterises also domains 2 and 6 at 90% level. As a matter of fact, these outcomes suggest that, in the most part of cases, non response bias should not seriously affect quality of estimates, at least as regards the model regression coefficient.

Tab. 2: Test (46) of the difference between regression coefficients of respondent and non respondent units by domain (Student's T at the 95% significance level - real data - average 2005-2007).

		***v*=1**				***v*=x**			
Domain	**Kind**				**Significance**				**Significance**
of units		***β***	***n***	***T test***	**95%**	***β***	***n***	***T test***	**95%**
1	*Respondents*	0,622	2.330			0,945	2.330		
	Non respondents	1,100	411			1,103	411		
				-4,528	Yes			-1,097	No
2	*Respondents*	1,194	1.060			1,148	1.060		
	Non respondents	0,969	157			1,022	157		
				1,285	No (yes 90%)			0,741	No
3	*Respondents*	1,130	2.164			1,062	2.164		
	Non respondents	1,333	311			1,184	311		
					-1,195	No		-0,762	No
4	*Respondents*	0,936	4.985			1,006	4.985		
	Non respondents	1,022	589			1,038	589		
				-0,604	No			-0,213	No
5	*Respondents*	1,045	3.593			1,061	3.593		
	Non respondents	1,289	328			1,170	328		
				-1,544	No (yes 90%)			-0,686	No
6	*Respondents*	1,010	2.209			1,047	2.209		
	Non respondents	0,997	212			1,012	212		
				0,082	No			0,226	No
7	*Respondents*	1,034	1.237				1,052	1.237	
	Non respondents	1,046	128				1,057	128	
				-0,076	No			-0,030	No

As regards the identification of the 2 post-strata defined through the rule (28) with *k*=2, for each quarter and each domain the following algorithm has been applied:

- given the set of ratios y/x referred to the n observed units, a cluster analysis has been carried out in order to identify 2 sub-groups such that the ratio (*between groups* variance)/(*within groups* variance) is maximised.
- Each non respondent unit has been assigned to the same sub-group to which its correspondent donor unit belongs.

We have verified that, in his way, for each domain and quarter, rule (28) is satisfied putting $\Delta_h \approx 0{,}25\beta_h$ for h=1,2 by not less than 95% of population ratios y/x.

Each regression coefficient in table 3 has been calculated as arithmetic mean of the 12 quarterly coefficients related to the same reference domain. On the whole, the average levels of coefficients in groups 1 and 2 are, respectively, 0,893 and 1,428. The difference between them is the lowest one for domain 3 (0,426) and the largest one for domain 1 (0,761).

Tab. 3: Regression coefficients by post-strata identified according to the rule (28) with $\approx 0{,}25 \cdot$ (real data - average 2005-2007).

	Domain							
Group	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Group 1	0,834	0,863	0,921	0,879	0,924	0,855	0,858	0,893
Group 2	1,599	1,636	1,347	1,428	1,353	1,541	1,599	1,428
Difference	0,765	0,773	0,426	0,549	0,429	0,686	0,741	0,535

Note: on the whole the 2 post-strata include at least the 95% of units.

Each regression coefficient in table 3 has been calculated as arithmetic mean of the 12 quarterly coefficients related to the same reference domain. On the whole, the average levels of coefficients in groups 1 and 2 are, respectively, 0,893 and 1,428. The difference between them is the lowest one for domain 3 (0,426) and the largest one for domain 1 (0,761).

In each domain post-stratification defined as above has been used in order to apply strategy (2), even though without testing the statistical significance of the difference between the 2 average β_h.

Main results have been summarised through the mean of absolute percent estimation errors (*MAPE*): figures in table 4 and 6 are arithmetic means of 12 quarterly estimation errors covering the period 2005-2007. The use of real data as regards respondents lead to the following conclusions:

a) on average of eight domains, the best strategy is (1) with $v=x$ (*MAPE*=0,85%), that is also the best one (bold) for two domains (3 and 7) and the second best

(underlined) for domain 6 and the total wholesale trade. The second best strategy is the donor based (3) with $\alpha=1$ and is characterised by an average *MAP*E quite similar (0,89%): that is the best one for domains 4, 5 and the total wholesale. As a consequence, even when sampling rates are large (92,0% on the average), donor imputation based on the minimum distance (8) may be quite useful.

b) Even though the model based strategy (1) with $v=x$ is the best one, the other model based strategies lead to rather larger average *MAPEs*, with a partial exception given by strategy (2) with $v=x$ (average *MAPE*: 1,28%). Since this strategy is based on a post-stratification inspired by donor imputation, one can conclude that a) a m_1 model based strategy may be dangerous if the true model is not correctly specified; b) donor imputation may improve quality of estimates, even though for some domains only. Empirical evidence suggests the potential existence of more than one model and that the position $v=x$ is more realistic than $v=1$.

c) Among donor imputation strategies, (3) performs significantly better than (4), meaning that the correct identification of donors based on (24) – e.g. on ratios instead of levels – is quite difficult.

d) Finally, calibration is helpful when $L=1$ and the only auxiliary variable given by turnover of the same quarter of the previous year is used (average *MAPE*: 0,98%).

Tab. 4: MAPE by domain and strategy (real data - average 2005-2007).

		Estimation strategies									
Domain	***n/N***	**(1)** ($v=x=1$)	**(1)** ($v=1$)	**(1)** ($v=x$)	**(2)** ($v=1$)	**(2)** ($v=x$)	**(3)** ($\alpha=1$)	**(3)** ($\alpha=0$)	**(4)** ($\alpha=1$)	**(4)** (z_1, $v=1$)	**(5)** (z_1z_2,$v=1$)
1	86,2	3,07	1,85	0,96	0,61	0,46	0,87	0,84	0,53	**0,36**	0,50
2	89,9	2,95	0,59	0,43	0,32	**0,29**	0,65	0,63	0,89	0,47	0,43
3	91,2	2,61	1,84	**0,59**	1,09	0,83	0,88	1,42	1,63	1,25	1,19
4	92,6	1,21	1,57	1,04	3,10	2,22	**0,39**	0,50	0,51	1,08	3,36
5	93,6	4,21	1,03	1,04	1,86	2,30	**0,74**	1,24	2,40	1,05	2,57
6	95,0	8,45	1,23	1,16	1,42	1,46	1,55	**1,14**	2,40	1,23	1,88
7	95,6	16,93	1,88	**1,16**	1,50	1,21	1,68	1,96	1,49	1,80	1,97
Total	92,1	3,18	0,84	0,44	1,37	1,49	**0,38**	0,75	1,55	0,61	1,88
Average	92,0	5,33	1,35	**0,85**	1,41	1,28	0,89	1,06	1,42	0,98	1,72

Bold: best; underlined: second best.

The use of the best estimation strategy for each domain would lead to the new composite strategies reported in table 5. As regards model based estimation

strategies, the average *MAPE* (bold) is equal to 0,771% and is due to the use of strategy (1) with $v=x$ for domains 3, 4, 6 and 7 (as well as for total wholesale if this domain is object of a separate estimation independent from the other domain estimations), strategy (2) with $v=x$ for domains 1 and 2 and strategy (1) with $v=1$ for domain 5.

Tab. 5: MAPE related to the best strategy for each domain and group of strategies (real data – average 2005-2007).

	Estimation strategies									
Domain	**(1)** ($v=x=1$)	**(1)** ($v=1$)	**(1)** ($v=x$)	**(2)** ($v=1$)	**(2)** ($v=x$)	**(3)** ($\alpha=1$)	**(3)** ($\alpha=0$)	**(4)** ($\alpha=1$)	**(5)** (z_1, $v=1$)	**(5)** (z_1z_2,$v=1$)
1					0,46			0,53	0,36	
2					0,29		0,63			0,43
3			0,59			0,88				1,19
4			1,04			0,39			1,08	
5		1,03				0,74			1,05	
6			1,16				1,14		1,23	
7			1,16					1,49	1,80	
Total			0,44			0,38			0,61	
Average			**0,771**				**0,773**			**0,969**

Bold: average MAPE related to the best strategy.

The use of donor imputation would lead to a best strategy whose average *MAPE* is similar to the previous one (0,773%) and is the result of the use of the following strategies: (3) with $\alpha=1$ for domains 3, 4 and 5 (plus the total wholesale), strategy (3) with $\alpha = 0$ for domains 2 and 6 and strategy (4) for domains 1 and 7. It is worthwhile to note that donor imputation could significantly improve model based estimation for domains 4 and 5: as regards domain 4 *MAPE* would decrease from 1,04% to 0,39%, while for domain 5 from 1,03% to 0,74%.

Finally, on the average calibration is the less efficient group of strategies, since the average *MAPE* obtained choosing the best calibration strategy for each domain is 0,969%; broadly speaking, univariate calibration performs better than the bivariate one.

On average, a lower response rate (around 50%) would enforce the better performance of some model based strategies with respect to donor imputation (table 6). Main outcomes can be summarised as follows:

a) on average of eight domains, the best strategy is still (1) with $v=x$ (*MAPE*=2,70%), that is also the best one for two domains (4 and 7) and the second best for domain 5. The second best strategy is the model based (2) with $v=x$ (*MAPE*=2,72%), that

is the best one for domain 3 and the second best for domain 4. The position $v=x$ confirms to be more realistic than $v=1$. Also in this case, the other model based strategies lead to rather larger average *MAPEs*.

b) Even though two model based strategies perform as the best and the second best, donor imputation has not to be neglected, since the model m_2 on the basis of which strategy (2) is founded, derives from the implicit post-stratification related to donor imputation (formula (28)). A negative aspect which affects donor imputation strategies – and that is enforced with respect to outcomes derived from real data – is that they may perform well or badly depending on the domain: for instance, the average *MAPE* obtained as regards domain 5 is quite large and ranges from 3,78% to 4,80%.

c) Finally, calibration is still more useful when $L=1$ and the only auxiliary variable is given by turnover of the same quarter of the previous year, not only because of a quite low average *MAPE* (2,97%), but also because this strategy is the best one for domain 5 and the second best for domain 7.

Tab. 6: MAPE by domain and strategy (1.000 random replications - average 2005-2007).

		Estimation strategies									
Domain	***n/N***	**(1)** ($v=x=1$)	**(1)** ($v=1$)	**(1)** ($v=x$)	**(2)** ($v=1$)	**(2)** ($v=x$)	**(3)** ($\alpha=1$)	**(3)** ($\alpha=0$)	**(4)** ($\alpha=1$)	**(5)** (z_1, $v=1$)	**(5)** ($z_1 z_2$,$v=1$)
1	50,0	5,21	4,85	1,15	1,16	0,81	**0,65**	1,75	2,56	1,88	0,79
2	50,0	15,66	9,58	5,38	6,90	4,55	**2,91**	3,14	6,62	6,40	7,22
3	50,0	7,17	2,50	3,24	3,76	**2,36**	5,26	3,33	5,89	2,99	3,08
4	50,0	8,73	4,29	**1,67**	4,76	1,81	2,32	2,99	2,30	2,25	2,40
5	50,0	17,34	4,77	1,88	4,96	3,67	4,80	3,78	4,67	**1,67**	3,31
6	50,0	6,52	8,10	5,78	3,89	4,91	2,34	2,77	**1,78**	5,95	5,75
7	50,0	17,58	1,38	**0,82**	2,20	1,45	3,70	2,69	2,58	1,38	2,30
Total	50,0	11,52	1,35	1,65	2,67	2,18	4,02	3,36	4,09	**1,28**	1,66
Average	50,0	11,22	4,60	**2,70**	3,79	2,72	3,25	2,98	3,81	2,97	3,31

Bold: best; underlined: second best

A direct consequence of results reported in table 6 is that the use of the best estimation strategy for each domain would lead to a model based composite strategy whose average efficiency is much more larger than the donor composite strategy (table 7).

As regards model based estimation strategies, the average *MAPE* is equal to 2,166% and is due especially to the use of strategies (1) or (2) with $v=x$.

The use of donor imputation would lead to a best strategy whose average *MAPE* is rather larger than the model based one (2,586%) and derives from the use of a mixture of all the 3 donor imputation strategies.

Finally, on the average calibration confirms to be the less efficient group of strategies (*MAPE*=2,814%) and should be applied using the univariate calibration only, with the only exception of domain 1.

Tab. 7: MAPE related to the best strategy for each domain and group of strategies (1.000 random replications – average 2005-2007).

	Estimation strategies									
Domain	**(1)** (v=x=1)	**(1)** (v=1)	**(1)** (v=x)	**(2)** (v=1)	**(2)** (α=x)	**(3)** (α=1)	**(3)** (α=0)	**(4)** (α=1)	**(5)** (z_1, v=1)	**(5)** (z_1z_2,v=1)
1					0,81	0,65				0,79
2					4,55	2,91			6,40	
3					2,36		3,33		2,99	
4			1,67					2,30	2,25	
5			1,88				3,78		1,67	
6				3,89				1,78		5,75
7			0,82					2,58	1,38	
Total		1,35					3,36		1,28	
Average			**2,166**				**2,586**		**2,814**	

Bold: average MAPE *related to the best strategy.*

6. CONCLUSIONS

Even though nearest neighbour imputation is a quite common empirical tool for tackling non responses' effects, few theoretical properties of this method are known. In particular, it is quite useful to investigate conditions for which estimation strategies based on donor imputation may improve efficiency of strategies founded on weighting. In this context, under a super-population model, we have developed the mean squared error of various estimators based on donor imputation and have carried out comparisons with respect to optimal model-based weighting estimation. According to outcomes of an empirical simulation referred to wholesale trade turnover, donor imputation may improve precision of estimates for some estimation domains, even though on the whole model based weighting seems to be more reliable. On the other hand, donor imputation could be preferred to calibration. Future additional efforts should be spent towards two directions:

1) theoretical improvements of the donor selection mechanism (for instance, donor selection may be based on more than one variable);

2) further empirical attempts referred to other kinds of variables (for instance: not continuous, as the number of job vacancies) and frameworks characterised by low response rates.

7. APPENDIX

Each estimator of a total can be written as: $T = \sum_S y_i + \sum_{\bar{S}} \hat{y}_j$. Since $Y = \sum_S y_i + \sum_{\bar{S}} y_j$, we have:

$$Mse(T) = E\left(T - Y\right)^2 = E\left(\sum_{\bar{S}} \hat{y}_j - \sum_{\bar{S}} y_j\right)^2 = E\left[\sum_{\bar{S}} \hat{y}_j - E\left(\sum_{\bar{S}} y_j\right) - \sum_{\bar{S}} y_j + E\left(\sum_{\bar{S}} y_j\right)\right]^2 =$$

$$= E\left[\sum_{\bar{S}} \hat{y}_j - E\left(\sum_{\bar{S}} y_j\right)\right]^2 + E\left[\sum_{\bar{S}} y_j - E\left(\sum_{\bar{S}} y_j\right)\right]^2 - 2E\left\{\left[\sum_{\bar{S}} \hat{y}_j - E\left(\sum_{\bar{S}} y_j\right)\right]\left[\sum_{\bar{S}} y_j - E\left(\sum_{\bar{S}} y_j\right)\right]\right\}.$$

Into the squared brackets of the first term we can add and subtract $E\left(\sum_{\bar{S}} \hat{y}_j\right)$, so that it can be written as $Var\left(\sum_{\bar{S}} \hat{y}_j\right) + Bias^2(T)$, where: $Bias^2(T) = \left[E\left(\sum_{\bar{S}} \hat{y}_j\right) - E\left(\sum_{\bar{S}} y_j\right)\right]^2$. The second term is equal to: $Var\left(\sum_{\bar{S}} y_j\right) = \sigma^2 \sum_{\bar{S}} v_j$, while the third term is null because it is equivalent to: $Cov\left(\sum_{\bar{S}} \hat{y}_j, \sum_{\bar{S}} y_j\right) = \sum_S Cov(\hat{y}_j, y_j)$, but each covariance in the sum is null because each estimated value is based on units belonging to the sample *S* only, which on the basis of the model (1) are not correlated with any unit not belonging to the sample. We get finally:

$$Mse(T) = Var\left(\sum_S y_j\right) + Var\left(\sum_{\bar{S}} \hat{y}_j\right) + Bias^2(T). \tag{50}$$

In particular, if:

$$\hat{y}_j = \hat{\beta} x_j \text{ with: } \hat{\beta} = \sum_S a_i y_i \text{ for each unit } j \notin S \tag{51}$$

where the *n* coefficients a_i must be specified, under the model (1) $Var(\hat{\beta}) = \sigma^2 \sum_S a_i^2 v_i$ and:

$$Mse(T) = \sigma^2 \sum_{\bar{S}} v_j + \sigma^2 \left(\sum_{\bar{S}} x_j \right)^2 \sum_{S} a_i^2 v_i + Bias^2(T) \text{ with:}$$
$$Bias^2(T) = \beta^2 \left(\sum_{\bar{S}} x_j \right)^2 \left(\sum_{S} a_i x_i - 1 \right)^2. \quad (52)$$

If $a_i = \frac{x_i}{v_i} \left(\sum_{S} x_i^2 / v_i \right)^{-1}$ we obtain the optimal unbiased predictor (2).

REFERENCES

BANKIER M. (2000), "Canadian Census Minimum Change Donor Imputation Methodology", *Proceedings of the Workshop on Data Editing - UN/ECE Work Session on Statistical Data Editing*, UK, Cardiff.

BEAUMONT J.F., BOCCI C. (2009), "Variance Estimation When Donor Imputation is Used to Fill in Missing Values", *Canadian Journal of Statistics*, on line publication at the address: http://www3.interscience.wiley.com/cgi-bin/fulltext/122466454/PDFSTART.

BIANCHI G., BRUNI R., NUCARA R., REALE A. (2005), "Data Clustering for Improving the Selection of Donor for Data Imputation", paper presented at the *Fifth Conference on Classification and Data Analysis (CLADAG)*, June 6-8, Parma, Italy.

BILLIET J., PHILIPPENS M., FITZGERALD R., STOOP I. (2007), "Estimation of Non-response Bias in the European Social Survey Using Information from Reluctant Respondents", *Journal of Official Statistics*, Vol.23, 2, 135-162.

BRICK J.M., KALTON G., KIM J.K. (2004), "Variance Estimation with Hot Deck Imputation using a Model", *Survey Methodology*, 30, 57-66.

CASSEL C., SÄRNDAL C.E., WRETMAN J. (1977), *Foundations of Inference in Survey Sampling*, J.Wiley & Sons, New York.

CHEN J., SHAO J. (2000), "Nearest Neighbour Imputation for Survey Data", *Journal of Official Statistics*, 2, 113-131.

CICCHITELLI G., HERZEL A., MONTANARI G.E. (1992), *Il campionamento statistico*, Il Mulino, Bologna.

COCHRAN W.G. (1977), *Sampling Techniques*, J.Wiley & Sons, New York.

COPELAND K.R., VALLIANT R. (2007), "Imputing for Late Reporting in the U.S. Current Employment Statistics Survey", *Journal of Official Statistics*, Vol.23, 1, 69-90.

DIGGLE P.J., KENWARD M.G. (1994), "Informative Drop-Out in Longitudinal Data Analysis", *Applied Statistics*, 43, 49-93.

GISMONDI R. (2008), "Reducing Revisions in Short-term Business Surveys", *Statistica*, Anno LXVIII, 1, 85-116, Clueb, Bologna.

HEDLIN D., FALVEY H., CHAMBERS R., KOKIC P. (2001), "Does the Model Matter for GREG Estimation? A Business Survey Example", *Journal of Official Statistics* Vol.17, 4, 527-544.

IBRAHIM J.G., ZHU H., TANG N. (2008), "Model Selection Criteria for Missing-Data Problems Using the EM Algorithm", *Journal of the American Statistical Association*, Vol.103, 484, 1648-1658.

ISTAT (2005), "Methods and Software for Editing and Imputation: Recent Advancements at ISTAT", paper presented at the *UN/ECE Work Session on Statistical Data Editing*, Ottawa, Canada.

ISTAT (2007), *Conti economici delle imprese – Anno 2003*, Informazioni, 8, Istat, Roma.

ISTAT (2008), "Seminario: strategie e metodi per il controllo e la correzione dei dati nelle indagini congiunturali sulle imprese: alcune esperienze nel settore delle statistiche congiunturali", *Contributi Istat*, 13/2008, Istat, Roma.

KALTON G. (2002), "Models in the Practice of Survey Sampling (Revisited)", *Journal of Official Statistics*, 18, 129-154.

LEHTONEN R., SÄRNDAL C.E., VEIJANEN A. (2003), "The Effect of Model Choice in Estimation for Domains, Including Small Domains", *Survey Methodology*, Vol.29, 1, 33-44.

LITTLE R.J.A. (1995), "Modelling the Drop Out Mechanism in Repeated-Measures Studies", *Journal of the American Statistical Association*, 90, 1112-1121.

LUNDSTRÖM S., SÄRNDAL C.E. (1999), "Calibration as a Standard Method for Treatment of Nonresponse", *Journal of Official Statistics*, Vol.15, 2, 305-327.

NISHII R. (1988), "Maximum Likelihood Principle and Model Selection When the True Model is Unspecified", *Journal of Multivariate Analysis*, 27, 392-403.

RIZZO L., KALTON G., BRICK M.J. (1996), "A Comparison of some Weighting Adjustment Methods for Panel Non-response", *Survey Methodology*, 22, 1, 43-53.

SÄRNDAL C.E., LUNDSTRÖM S. (2005), *Estimation in Surveys with Nonresponse*, J.Wiley & Sons, New York.

VALLIANT R., DORFMAN A.H., ROYALL R.M. (2000), *Finite Population Sampling and Inference – A Prediction Approach*, J.Wiley & Sons, New York.

UN CONFRONTO TRA IMPUTAZIONE CON DONATORE E PONDERAZIONE *MODEL BASED* IN PRESENZA DI NON RISPOSTE E RISCHIODI ERRATA SPECIFICAZIONE DEL MODELLO

Riassunto

In questo lavoro si analizza la strategia di stima campionaria basata sull'imputazione deterministica delle osservazioni non desumibili dall'indagine statistica, in quanto riferite a unità non rispondenti, o comunque non incluse nel campione da intervistare, con selezione del donatore basata su criteri di distanza minima tra unità donatrice ed unità ricevente. Supponendo un semplice modello di super-popolazione, si determina la media quadratica dell'errore dello stimatore basato sull'imputazione per donatore e si confronta la sua efficienza con quella dello stimatore ottimale derivato dal modello. Inoltre, si introduce il problema dovuto alla possibile errata specificazione del modello e si analizza anche il caso in cui il campione osservato (e la stessa popolazione finita di riferimento) possano contenere osservazioni derivate da modelli diversi. In entrambi i casi si determinano le condizioni tali che l'imputazione con donatore possa essere preferita alla ponderazione da modello. Infine, si presentano i principali risultati di un'applicazione comparativa basata sull'utilizzo dei dati raccolti ogni trimestre nell'ambito dell'indagine sul commercio all'ingrosso condotta correntemente dall'ISTAT.